

Jun Liu · Yongcai Wang · Bin Wu ·
Zehao Jiang · Yao Xiao *Editors*✉

The Proceedings of 2024 International Conference on Artificial Intelligence and Autonomous Transportation

Volume V

Series Editors

Leopoldo Angrisani, *Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Napoli, Italy*

Marco Arteaga, *Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico*

Samarjit Chakraborty, *Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany*

Shanben Chen, *School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China*

Tan Kay Chen, *Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore*

Rüdiger Dillmann, *University of Karlsruhe (TH) IAIM, Karlsruhe, Germany*

Haibin Duan, *Beijing University of Aeronautics and Astronautics, Beijing, China*

Gianluigi Ferrari, *Dipartimento di Ingegneria dell'Informazione, Sede Scientifica Università degli Studi di Parma, Parma, Italy*

Manuel Ferre, *Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain*

Faryar Jabbari, *Department of Mechanical and Aerospace Engineering, University of California, Irvine, USA*

Limin Jia, *State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China*

Janusz Kacprzyk, *Intelligent Systems Laboratory, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Alaa Khamis, *Department of Mechatronics Engineering, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt*

Torsten Kroeger, *Intrinsic Innovation, Mountain View, USA*

Yong Li, *College of Electrical and Information Engineering, Hunan University, Changsha, China*

Qilian Liang, *Department of Electrical Engineering, University of Texas at Arlington, Arlington, USA*

Ferran Martín, *Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Spain*

Tan Cher Ming, *College of Engineering, Nanyang Technological University, Singapore, Singapore*

Wolfgang Minker, *Institute of Information Technology, University of Ulm, Ulm, Germany*

Pradeep Misra, *Department of Electrical Engineering, Wright State University, Dayton, USA*

Subhas Mukhopadhyay, *School of Engineering, Macquarie University, Sydney, NSW, Australia*

Cun-Zheng Ning, *Department of Electrical Engineering, Arizona State University, Tempe, AZ, USA*

Toyoaki Nishida, *Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan*

Luca Oneto, *Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genova, Genova, Italy*

Bijaya Ketan Panigrahi, *Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India*

Federica Pascucci, *Department di Ingegneria, Università degli Studi Roma Tre, Rome, Italy*

Yong Qin, *State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China*

Gan Woon Seng, *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore*

Joachim Speidel, *Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany*

Germano Veiga, *FEUP Campus, INESC Porto, Porto, Portugal*

Haitao Wu, *Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China*

Walter Zamboni, *Department of Computer Engineering, Electrical Engineering and Applied Mathematics, DIEM—Università degli studi di Salerno, Fisciano, Italy*

Kay Chen Tan, *Department of Computing, Hong Kong Polytechnic University, Hong Kong, Hong Kong*

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Jun Liu · Yongcai Wang · Bin Wu · Zehao Jiang ·
Yao Xiao
Editors

The Proceedings of 2024 International Conference on Artificial Intelligence and Autonomous Transportation

Volume V

Editors

Jun Liu
North China University of Technology
Beijing, China

Yongcai Wang
Renmin University of China
Beijing, China

Bin Wu
Wuhan University of Technology
Wuhan, Hubei, China

Zehao Jiang
Huazhong University of Science
and Technology
Wuhan, Hubei, China

Yao Xiao
Sun Yat-sen University
Shenzhen, Guangdong, China

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-96-3972-4

ISBN 978-981-96-3973-1 (eBook)

<https://doi.org/10.1007/978-981-96-3973-1>

© Beijing Paike Culture Commu. Co., Ltd. 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

Contents

Research on Calibration of Dynamic Simulation Models for Rail Vehicles	1
<i>Junyi Lan and Dao Gong</i>	
Survey on Current Situation and Development Trend of Measurement Quality Assurance Technology	9
<i>Daiwen Peng, Sha Li, Hu Liu, and Ming Deng</i>	
Snowy Urumqi County Postal Logistics Road Transportation Route Planning Study	18
<i>Mengxi Wu, Bing Wang, and Yingqi Li</i>	
A Dual-Stream Convolutional Network for Visible and Infrared Image Fusion in Pedestrian Detection	27
<i>Xinyan Xiong, Jie Yang, Yanxuan Jiang, and Xinkai Hu</i>	
Comparative Analysis of Lane Change Intention Recognition Models Based on the CitySim Dataset	35
<i>Hanfei Wang, Huanfeng Liu, Ziyang Wu, Anning Song, and Zishuo Zhang</i>	
Enhanced SlowFast Networks with Dark Channel Prior for Nighttime Pedestrian Detection in Port Surveillance	46
<i>Han Zhang, Xinqiang Chen, Dezhi Han, Yiwen Zheng, Bing Han, and Yongsheng Yang</i>	
Research on the Application of Privacy Computing Technology in the Scene of Railway Data Rights Confirmation and Authorization	53
<i>Peiran Wang, Xiaoning Ma, Dan Zou, Siqu Sun, Jiang Wu, and Kaopeng Li</i>	
A Fair Traffic Management Approach for Electric and Conventional Vehicles in Urban Areas	61
<i>Liang Zhang, Huijuan Zhou, and Xiaoming Liu</i>	
A Cooperative Control Model and Simulation Framework for Unmanned Aircraft Considering Multi-subject Communication	69
<i>Zheng Zhang, Jingyuan Chen, Endian Wang, Ziqi Wang, Zihua Li, Gege Jiang, and Yao Xiao</i>	
The Influence of Parameter Optimization of Rubber Nodes in High-Speed Train Turning Arms on Suppressing Hunting Motion	78
<i>Tao Guo, Ya Li, and Yongqiang Liu</i>	

Multi-object Tracking Algorithm of Moving Scene Based on Bidirectional Buffering of Lost Trajectory	86
<i>Chanzhen Xiong and Yuanlong Wu</i>	
A Cost-Sensitive Multi-scale Feature Multi-order Fusion Network for Bearing Fault Diagnosis Under Data Imbalance Conditions	94
<i>Shuaiqing Deng, Zihao Lei, Guangrui Wen, Yu Su, Zimin Liu, Zhangxuan Meng, and Zhifen Zhang</i>	
Research on Secure Storage Architecture of Railway Data Based on Data Lake	107
<i>Jiang Wu, Dan Zou, Siqi Sun, Peiran Wang, and Xiaoning Ma</i>	
Research on Unplanned Skip-Stop Scheme for Interconnected Suburban Railway Based on Particle Swarm Multi-objective Optimization Algorithm	117
<i>Haiwei Jin, Wenqian Huang, Changyu Chen, Ming Xu, and Rudong Yang</i>	
Adaptive Fault-Tolerant Control for Over-Actuated Ships Based on Safe Reinforcement Learning	127
<i>Zongdi Tang, Yamin Huang, Yuanqiao Wen, Jie Liu, and Linying Chen</i>	
Parallel Path-Following Control of Fully-Actuated Autonomous Surface Vehicles Based on a High-Order Learning Extended State Observer	139
<i>Jiaxue Xu, Haodong Liu, Nan Gu, Lu Liu, and Zhouhua Peng</i>	
The Development and Challenges of Intelligent Ship	148
<i>Peng Zhang, Yong Ma, Huihui Chen, Chengli Liu, and Zhongdai Wu</i>	
Functional Safety Monitoring Strategy for Motor Control Unit of Autonomous Navigation Domain Controller in Unmanned Surface Vehicle	155
<i>Ahui Xu, Yamin Huang, Yuanqiao Wen, Jie Liu, and Lihang Song</i>	
Research on Uplink Signal Processing of Balise Based on VMD Optimization	171
<i>Qianshu Xue, Jing Xun, Minxue Fu, and Chenghao Cui</i>	
Real-Time Classification of Freeway Traffic Congestion Levels Based on Surveillance Video	179
<i>Chuanyun Fu, Jinzhao Liu, Zhaoyou Lu, and Wei Bai</i>	
Numerical Analysis of the Influence of Free Surface on the Propulsive Efficiency of Biomimetic Fish Bodies	188
<i>Chaoming Bao, Kang Ren, Fei Guo, Daiyu Zhang, and Jiyang Li</i>	

Real-Time LiDAR Point Cloud Semantic Segmentation for Unstructured Road Autonomous Driving	198
<i>Huazhi Li, Guizhen Yu, Zhangyu Wang, Yang Chen, and Fei Zhao</i>	
A Review of the Following Behavior of Human-Driven Vehicles and Automated Vehicles in Adverse Weather Conditions	209
<i>Qiong Lu, Haoda Fang, Xinzhao Jia, and Bing Wang</i>	
Optimization of Arrival and Departure Track Application Considering In-App Reconnection Decomposition	221
<i>Mengjie Cui and Lei Nie</i>	
Optimization of Vehicle Body Modal Recognition Measuring Point Based on Particle Swarm Optimization	229
<i>Xiao Sun and Gong Dao</i>	
High-Precision Belt Material Flow Tracking Research in Bulk Cargo Ports	239
<i>Changan Li, Baodong Sun, Huiqiang Li, Baoxin Qin, and Xiaoshu Qiao</i>	
Enhancing Online HD Map Construction with Trajectory Guidance in Challenging Conditions	251
<i>Ziying Yao, Zhongxia Xiong, and Xinkai Wu</i>	
Traffic Signal Timing Optimization Based on D3QLight	259
<i>Haoyue Wang, Zundong Zhang, and Shilin Zhang</i>	
Research on the Transportation Organization Mode of Low-Speed Maglev System	268
<i>Zinuo Li, Jun Liu, Wu Bai, Yu Liu, and Qingying Lai</i>	
Downstream Intersection Entrance Flow Arrival Rate Estimation Using Sampled Trajectory Data	277
<i>Yingjie Bai, Shuyan Jiang, Ke Lin, Zhijie Zhou, and Kai Lu</i>	
Multi-spatial Scales Delay Time Index Calculation Method Based on Digital Road Networks	289
<i>Kai Lu, Kun Zeng, Yaoting Wu, Jingfeng Wei, and Xingdong Deng</i>	
Behavioral Experimental Study on Multi-mode Combined Travel Route Choice Considering Information Acceptance	303
<i>Xiaohui Sun, Jiaxin Zhang, and Min Chen</i>	
Distributed Adaptive Coordination Control for Autonomous Road Traffic at Multiple Intersections	312
<i>Honghai Ji, Qirui Shi, Shida Liu, Lingling Fan, and Li wang</i>	

Optimization Method Based on Improved PointNet++ Network Parameters	319
<i>Lei Kou, Hongzheng Zhao, Ruiyu Tang, Junsheng Su, Xuqiang Zhao, Wei Yan, and Feng Guo</i>	
Research on Coverage Path Planning of Unmanned Surface Vehicle Swarm for Maritime Search and Rescue	329
<i>Lootha Loserian Muriatoy, Gaihui Wang, Yixuan Dong, Dingxiang Sun, Zhiqiu Huang, and Haoliang Wang</i>	
Research on the Method of Mooring Position Selection for Ships Riding to Single Anchor Based on Nautical Practice	337
<i>Xiao Liu, Hua Deng, Bing Wang, Ke Zhang, Yong Wu, and Xingya Zhao</i>	
Study on the Scale of Urban Rail Station Connection Facilities Based on the Concept of Integration	349
<i>Xiaohui Sun, Jingqi Zheng, and Yuling Fu</i>	
Optimization Method of the Turnover Plan of Rolling Stocks of Urban Rail Transit Considering Service Life and Maintenance	360
<i>Xianghao Kong, Dan Wei, Jinduo Chen, Wanting Jia, and Hui Li</i>	
Short-Term Passenger Flow Prediction of Urban Rail Transit Based on Station Clustering and ConvLSTM	369
<i>Zhengda Tian, Dan Wei, Lei Huang, Yunzhou Tan, and Lvqi Yao</i>	
Study and Implementation of File Transparency Encryption Technology for Intelligent Connected Vehicles Based on SM4	378
<i>Qi Wang, Yongzhong Li, and Yuheng Li</i>	
Cognition and Modeling of Urban Low-Altitude Airspace	387
<i>Sidao Chen, Xuejun Zhang, and Weidong Zhang</i>	
Simulation Analysis of Mixed Traffic Flow Safety in Risk Scenarios on Highways Under Foggy Conditions	396
<i>Ying Zhao, Haijun Li, Jianqiang Wang, and Yan Huang</i>	
Research on the Innovation and Environmental Sustainability Development of Coal Storage Facilities in Chinese Ports	404
<i>Yuhua Wei</i>	
A Smart Tourism Recommendation Algorithm Based on Profiling Technology and Bipartite Graph	414
<i>Zexian Guo, Yuemin Lin, and Zhouyang Luo</i>	

Trajectory Optimization for Mining Robots Based on an Improved Chicken Swarm Optimization Algorithm	422
<i>Jianjian Yang, Teng Zhang, Guanghui Zhao, Qi Cheng, Kangshuai Chen, Zhixiang Jiao, Shaowen Zheng, and Zifan Zhao</i>	
Evaluation of the Energy-Saving Ability of Pulse-and-Glide Mode in Vehicles: A Coupled Simulation Approach	435
<i>Honggang Li, Changhui Yan, and Jinjun Sun</i>	
Research on High-Confidence Simulation Testing Methods for AI-Based End-to-End Autonomous Driving Vehicles	444
<i>Zhangjie Meng, Deng Pan, Qiang Zhang, Bin Lu, Jinying Zhou, Pan Xiao, Bing He, and YiChao Tang</i>	
A Star Structure Algorithm Based on GNSS/CORS and Experimental Analysis	454
<i>Zhu Chao</i>	
Research on Test Scenarios Coverage for the Behaviour Safety of Autonomous Driving Vehicles	462
<i>Yu Tang, Yingzhi Xiong, Junfu Huang, Qiang Zhang, Jinying Zhou, Xiaona He, and Kan Yi</i>	
Research on Charging Location of New Energy Vehicles and Mobile Charging Vehicles Under Multi-objective Coordination	477
<i>Yongqi Diao and Ying Lv</i>	
Analyzing User Needs of Autonomous Transportation Systems with Activity Theory and Knowledge Graph	486
<i>Lai Zheng, Chunrong Bao, and Wei Wei</i>	
Dynamic Performance Prediction and Fluctuation Analysis Using Momentum Modeling in Competitive Environments	495
<i>Enxin Gu</i>	
Behavior Prediction of Inland River Ferries Based on Encounter Scenarios	505
<i>Jian Li, Yuanqiao Wen, Liang Huang, Ahui Xu, and Yamin Huang</i>	
A Real-Time Railway Intrusion Detection and Multi-factor Collaborative Risk Assessment Method	521
<i>Genwang Peng, Yang Gao, Zhiwei Cao, Yong Qin, and Jianyuan Guo</i>	
Author Index	531



Research on Calibration of Dynamic Simulation Models for Rail Vehicles

Junyi Lan and Dao Gong^(✉)

College of Transportation Engineering, Tongji University, Shanghai 201804, China
gongdao@tongji.edu.cn

Abstract. The dynamic model of rail vehicles is an important basis for vehicle design, dynamic analysis and optimization. However, due to factors such as the differences between the actual physical parameters of rail vehicles and theoretical parameters, and the inconvenience of obtaining line input excitations, there is often a large error in the dynamic response between the established dynamic model of rail vehicles and the actual vehicle. This study attempts to use optimization algorithms to calibrate and correct the dynamic simulation model of rail vehicles for the above problems. First, a dynamic model is established based on SIMAPCK multi-body dynamics software. Then, various optimization algorithms are used to calibrate the model parameters. Finally, dynamic simulations are performed based on the model before and after calibration, and its dynamic performance is compared with the measured data.

Keywords: Dynamic simulation model · calibration correction · optimization algorithm · dynamic performance

1 Introduction

Through the dynamic model of rail vehicles, researchers can simulate the actual operation process of vehicles in a virtual environment, predict the performance and safety of vehicles, and provide strong support for railway engineering design, vehicle operation control, optimization, and fault diagnosis.

In order to establish a more accurate and reliable dynamic model of rail vehicles, it is necessary to calibrate and correct the model. Parameter optimization algorithm is the core technology of calibration and correction of simulation model. These algorithms continuously iterate and adjust model parameters to achieve the best match between simulation results and actual system data. WANG Z et al. [1] explored the impact of hollow-worn wheels on the stability of rail vehicles using the SIMPACK dynamics model. LIU G et al. [2] established a dynamics model of high-speed electric multiple unit vehicle system and studied the evolution law of modal parameters of vehicle system as running speed changes. GONG D et al. [3, 4] employed a dynamics model to put forward a multi-modal vibration control method to lower the multi-modal vibration of the car body of high-speed electric multiple unit trains and enhance the running quality of trains. By integrating methods such as field testing, wheel and track profile testing, modal

parameter testing, and root locus analysis, they deeply probed into the causes of car body jitter and proposed corresponding control measures to improve the running stability and ride quality of trains. Yang Dandan [5] took a certain type of train and its scaled model as the research object, and studied the modeling and correction of the car body and the modeling and correction of the rigid-flexible coupling model of the whole vehicle based on SIMPACK. Akiyama et al. [6] used an iterative method based on particle swarm optimization algorithm to calibrate the model of the Shinkansen passenger car body, and proposed a three-dimensional analytical model of the elastic vibration of the car body and a method for determining parameters. After model calibration, the difference between the measured results and the calculated results is significantly reduced. Bragança et al. [7] Based on dynamic tests under actual operating conditions, a method for calibrating and validating the numerical model of freight cars using genetic algorithms is proposed. After model calibration, excellent consistency is obtained between experiments and numerical values. Ribeiro et al. [8] The numerical model of freight railway vehicles was experimentally calibrated based on test modal parameters. The calibration depends on an iterative method based on genetic algorithms. The response obtained from the calibrated model has a very good correlation with the experimental response.

Based on the relevant applications of optimization algorithms, this paper calibrates the dynamic simulation model of a certain type of intercity train. By comparing the error between the simulation results before and after model calibration and the measurement results of the actual vehicle, the reliability of the calibration method is verified.

2 Modeling and Simulation Verification

2.1 Modeling of Simulation Model

In this paper, a model is established according to the vehicle dynamics parameters of a certain type of intercity train.

Considering the need for subsequent dynamic simulation of the model and calibration using optimization algorithms, the following settings and assumptions are made when establishing the model:

- (1) Parameterization of design variables. In order to facilitate obtaining a large number of samples later, in SIMPACK, multiple parameters such as vehicle suspension parameters, vehicle body mass, and inertia are set as replacement variables to achieve parametric modeling.
- (2) Processing of nonlinear relationships. During the operation of rail vehicles, there are nonlinear relationships in key parts such as wheel-rail contact, creep, suspension system, side bearings and stops.

The dynamic simulation model of a certain type of intercity train established based on SIMPACK according to the above content is shown in Fig. 1.

2.2 Modeling of Simulation Model

According to "GB/T 5599-2019 Specification for evaluation and test approval of dynamic performance of locomotives and rolling stock" [9], the Sperling value and the maximum

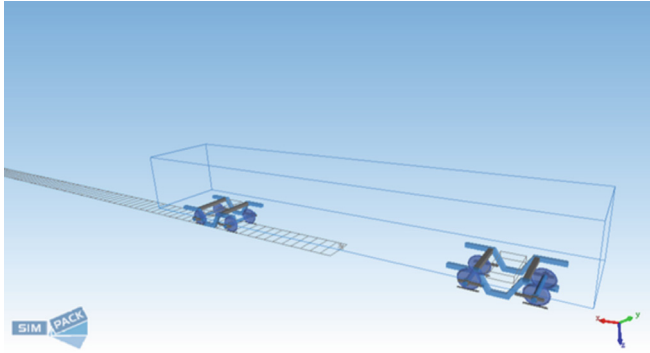


Fig. 1. Dynamic simulation model of rail vehicles.

values of lateral and vertical vibration accelerations of the car body obtained by simulating the established rail vehicle dynamics model on straight lines, large curves ($R = 450$ m), and small curves ($R = 250$ m) are shown in Table 1.

Table 1. Simulation result table of dynamic model.

Indicator	Straight line	Large curve ($R = 450$ m)	Small curve ($R = 250$ m)
W_{y1}	1.58754	1.85353	1.82717
W_{y2}	1.4655	1.71683	1.84464
W_{z1}	1.81187	1.72728	2.19435
W_{z2}	1.86403	1.74482	2.25675
$a_{ymax}(m/s^2)$	0.129195	0.477954	0.652597
$a_{zmax}(m/s^2)$	0.259475	0.425957	0.765808

In the table, W is the Sperling value, a is the car body vibration acceleration, Y represents the lateral direction, and Z represents the vertical direction. Through analyzing the data in the table, it can be found that all the indicators of the three working conditions in the table comply with the GB/T 5599–2019 specification, which proves the reliability of the established dynamic simulation model. As the track radius decreases, the values of various evaluation indicators gradually increase.

2.3 Vehicle Test Data Processing

The calibration of the dynamic simulation model needs to be compared with the test data of the actual vehicle. In this paper, the DASP software of the large-capacity data automatic acquisition and signal processing system is used to process the collected vibration acceleration data. Based on the previous simulation results, the measured point data at three positions (front, middle, and rear) of the car body on the small curve line

under the working condition with an operating speed of 90 km/h are selected for data processing. The waveform diagrams of the lateral and vertical data in the middle of the selected car body are shown in Figs. 2 and 3.

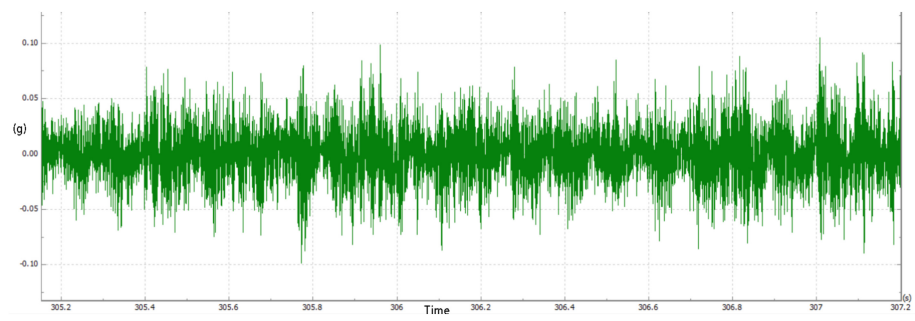


Fig. 2. Data of the lateral part in the middle of the car body.

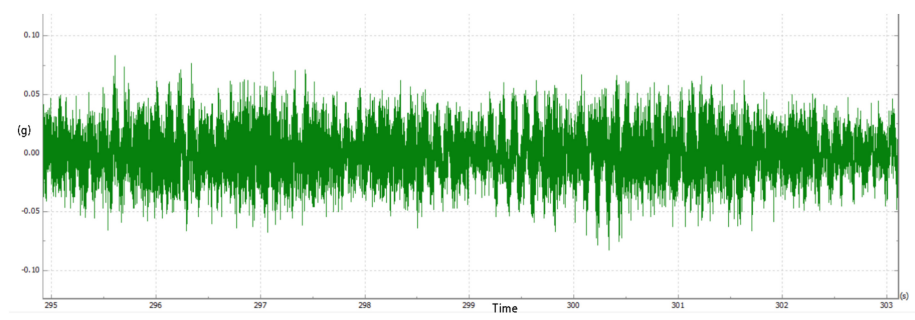


Fig. 3. Data of the vertical part in the middle of the car body.

The data is imported into MATLAB for calculation and processing. The car body stability index and the maximum value of the car body vibration acceleration are shown in Table 2.

Table 2. Data processing results.

Indicator	numerical value
W_{y1}	1.86945
W_{y2}	1.88324
W_{z1}	2.23241
W_{z2}	2.29473
$a_{ymax}(m/s^2)$	0.774324
$a_{zmax}(m/s^2)$	0.912343

Observing the data in the table, it can be found that the values obtained according to the test data are all slightly larger than the values obtained by the model simulation.

3 Model Calibration Based on Optimization Algorithm

3.1 Model Calibration Results

After the surrogate model is established, optimization algorithms are used to calibrate the model. Since calibration involves multiple parameters, multi-objective optimization methods and global exploration methods are adopted for calibration, that is, four optimization algorithms including Multi-Island Genetic Algorithm (MIGA), Non-dominated Sorting Genetic Algorithm-II (NSGA-II), Multi-Objective Particle Swarm Optimization (MOPSO), and Adaptive Simulated Annealing (ASA) are used to calibrate the model. In this section, taking the NSGA-II algorithm as an example:

In NSGA-II [10, 11], the SBX (Simulated Binary crossover) method is used for crossover and mutation. The crossover operation is shown in formula (1):

$$\begin{cases} x_i^{(1,t+1)} = \frac{1+\beta_{qi}}{2}x_i^{(1,t)} + \frac{1-\beta_{qi}}{2}x_i^{(2,t)} \\ x_i^{(2,t+1)} = \frac{1-\beta_{qi}}{2}x_i^{(1,t)} + \frac{1+\beta_{qi}}{2}x_i^{(2,t)} \end{cases} \quad (1)$$

According to the SBX method, the sudden mutation operation for generating child individuals is shown in formula (2):

$$\begin{aligned} x_i^{(1,t+1)} &= x_i^{(1,t)} + \delta_i(x_i^{UB} - x_i^{LB}) \\ \delta_q &= \begin{cases} [2u + (1-2u)(1-\delta)^{\eta_m+1}]_{\eta_m+1}^1 - 1 & u \leq 0.5 \\ 1 - [2(1-u) + 2(u-0.5)(1-\delta)^{\eta_m+1}]_{\eta_m+1}^1 & u > 0.5 \end{cases} \\ \delta &= \frac{\min(x_i - x_i^{LB}, x_i^{UB} - x_i)}{x_i^{UB} - x_i^{LB}}, \quad u \in [1, 0] \end{aligned} \quad (2)$$

The scatter plot of individual responses in the optimization history of the maximum value of car body vibration acceleration obtained by the NSGA-II algorithm is shown in Figs. 4 and 5.

The optimization results of the four optimization algorithms are shown in Table 3.

3.2 Comparison of Simulation Results Before and After Calibration

In order to verify the reliability of the calibrated model, the dynamic performance simulation results before and after model calibration are compared with the test data. The comparison of the absolute values of the relative errors between the model simulation results before and after calibration based on the four optimization algorithms and the test data is shown in Fig. 6.

It can be seen from the comparison chart that the models obtained after calibration based on four optimization algorithms have responses that are closer to the responses obtained from processing the test data, which proves the feasibility of using optimization algorithms to calibrate the dynamic simulation model.

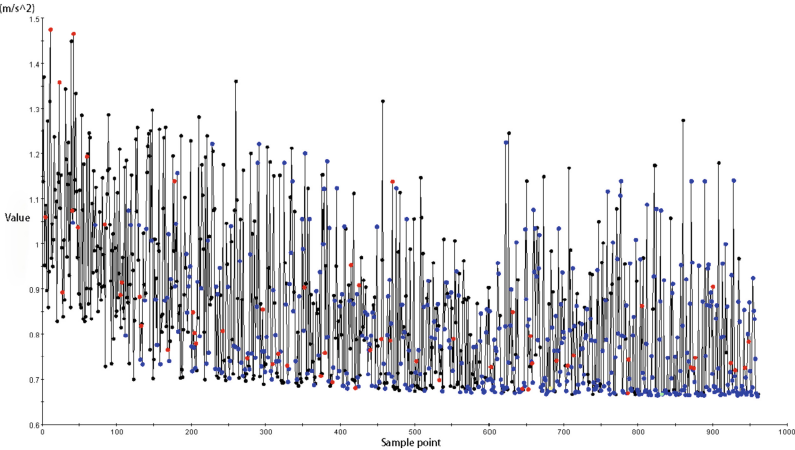


Fig. 4. Maximum value of car body lateral vibration acceleration.

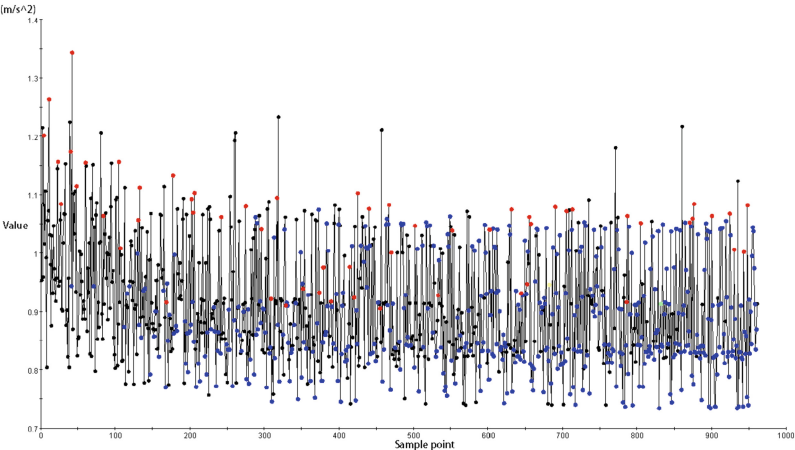


Fig. 5. Maximum value of vehicle body vertical vibration acceleration.

Table 3. Table of simulation results of dynamic model.

Indicator	MIGA	NSGA-II	MOPSO	ASA
W_{y1}	1.845588255	1.874948988	1.836881447	1.878449249
W_{y2}	1.841228094	1.88053422	1.85106	1.840329466
W_{z1}	2.223314476	2.258552504	2.235910606	2.214681101
W_{z2}	2.279686928	2.315391398	2.332603169	2.262213879
a_{ymax}	0.798494959	0.730076659	0.773678124	0.713810492
a_{zmax}	0.966740847	0.854134321	0.892451859	0.858840132

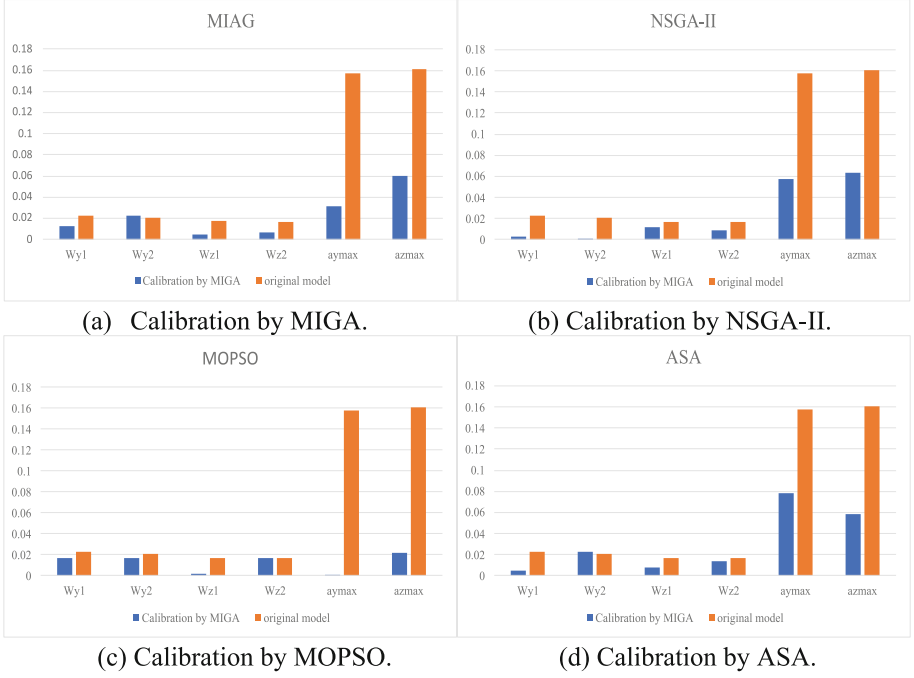


Fig. 6. Comparison before and after calibration using MIGA, NSGA-II, MOPSO, and ASA.

4 Summary

This paper conducts research on the calibration of the rail vehicle dynamics simulation model of a certain type of intercity train based on optimization algorithms. First, a rail vehicle dynamics simulation model is established and simulation verification is carried out based on SIMAPCK multi-body dynamics software. Subsequently, four optimization algorithms, MIGA, NSGA-II, MOPSO, and ASA, are used to calibrate the suspension parameters, mass, and moment of inertia of the model. The feasibility of the calibration method is verified by comparing the dynamic performance of the model before and after calibration with the response obtained from test data. Through the above work, the obtained results show that calibrating the rail vehicle dynamics simulation model based on optimization algorithms. Compared with the model before calibration, the calibrated model has dynamic performance that is closer to the results obtained from processing real test data, indicating the feasibility of the method.

Acknowledgments. The authors gratefully acknowledge the financial support from National Natural Science Foundation of China (No. 52375115), National Natural Science Foundation of China (No. 52475122) and Supported by the Fundamental Research Funds for the Central Universities (No. 22120240338).

References

1. Wang, Z., Gong, D.: A survey on countermeasures of railway vehicle stability decrease caused by the evolution of hollow-worn wheels. *J. Vib. Control* (2024). <https://doi.org/10.1177/10775463241234726>
2. Liu, G., et al.: Frequency veering of railway vehicle systems and its mapping to vibration characteristics. *Multibody Syst. Dyn.* (2024). <https://doi.org/10.1007/s11044-024-09989-x>
3. Gong, D., Liu, G., Zhou, J.: Research on mechanism and control methods of carbody chattering of an electric multiple-unit train. *Multibody Syst. Dyn.* **53**(2), 135–172 (2021)
4. Gong, D., Zhou, J., Sun, W., Sun, Y., Xia, Z.: Method of multi-mode vibration control for the carbody of high-speed electric multiple unit trains. *J. Sound Vib.* **409**, 94–111 (2017). <https://doi.org/10.1016/j.jsv.2017.05.010>
5. Yang, D.: Rigid-flexible coupling modeling and model correction of a certain type of train based on SIMPACK. Nanjing University of Aeronautics and Astronautics (2016). (in Chinese)
6. Akiyama, Y., Tomioka, T., Takigami, T., et al.: A three-dimensional analytical model and parameter determination method of the elastic vibration of a railway vehicle carbody. *Veh. Syst. Dyn.* **58**(4), 545–568 (2019)
7. Bragança, C., Neto, J., Pinto, N., et al.: Calibration and validation of a freight wagon dynamic model in operating conditions based on limited experimental data. *Veh. Syst. Dyn.* **60**(9), 3024–3050 (2022)
8. Ribeiro, D., Bragança, C., Costa, C., et al.: Calibration of the numerical model of a freight railway vehicle based on experimental modal parameters. *Structures* **38**, 108–122 (2022)
9. GB/T 5599-2019. Specification for evaluation and test verification of dynamic performance of locomotives and rolling stock (2019). (in Chinese)
10. Xie, G., Cheng, G., Liang, H., et al.: Software diversification combination method based on multi-objective optimization algorithm NSGA-II [J/OL]. *Comput. Sci.*, 1–15 (2024). (in Chinese)
11. Deb, K., Pratap, A., Agarwal, S., et al.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 182–197 (2002)



Survey on Current Situation and Development Trend of Measurement Quality Assurance Technology

Daiwen Peng^(✉), Sha Li, Hu Liu, and Ming Deng

State Grid Intelligent Vehicle Networking Technology Co., LTD, Beijing, China
2981822050@qq.com

Abstract. Metrological quality assurance techniques are essential in modern society. At present, the accuracy of some measuring equipment is limited, and it is difficult to meet the demand of high precision. The data processing technology is uneven, which affects the accuracy of the results. Industry standards are not unified, resulting in difficulties in cross-regional cooperation. This paper elaborates on the achievements achieved at the present stage, including the continuous introduction of high-precision measuring instruments, the strengthening of international cooperation, and the progress of data processing and analysis technology, which have improved the quality assurance capabilities. At the same time, the paper plans the later development of measurement quality assurance technology, mainly including the research and development of more intelligent measurement equipment and automatic testing system. Using big data and artificial intelligence to optimize metrological data analysis and quality prediction; Promote the unification of international standards, strengthen global cooperation, and realize the sustainable development of measurement quality assurance technology.

Keywords: Metrology quality assurance technology · intelligent metrology · big data · artificial intelligence

1 Introduction

1.1 Research Background

With the rapid development of the global economy and the continuous progress of science and technology, the importance of metrology quality assurance technology in modern society has become increasingly prominent. As the basis of science and technology and the guarantee of social development, the accuracy and reliability of measurement are directly related to production, scientific research and trade activities in various fields.

In today's era of globalization, cross-regional cooperation is becoming more and more frequent [1], which puts forward higher requirements for measurement quality assurance technology. However, current metrology quality assurance techniques still face some challenges. Some measuring equipment has limited accuracy, which is difficult to

meet the increasing demand for high precision. For example, in the fields of microelectronics and nanotechnology, the requirements for measurement accuracy have reached the nanometer or even subnanometer level, while some existing metrology devices cannot achieve such accuracy requirements [2]. At the same time, the data processing technology is uneven, and the data analysis methods and software adopted by different measurement institutions and enterprises are different, which directly affects the accuracy and comparability of measurement results [3]. In addition, non-uniform industry standards are also a prominent problem, and there are differences in measurement standards in different countries and regions, which bring great difficulties to cross-regional cooperation and hinder the smooth progress of international trade and technical exchanges.

In order to solve these problems, domestic and foreign scholars and metrological institutions have carried out a lot of research and practice. In recent years, high-precision measuring instruments have been continuously introduced [4], which provides strong support for improving measurement accuracy. For example, the emergence of nanoscale measurement devices makes it possible to perform accurate measurements on the nanoscale [5]. At the same time, international cooperation is also constantly strengthened. Metrological institutions of various countries jointly promote the unification and improvement of metrological standards through cooperative research and comparison tests [6]. In addition, data processing and analysis technology is also advancing, and new algorithms and software are constantly emerging to improve the efficiency and accuracy of measurement data processing.

1.2 Research Purpose

This paper analyzes the current situation of metrological quality assurance technology, clarifies the existing problems and challenges, summarizes the achievements at the present stage, and makes a reasonable plan for the future development trend of metrological quality assurance technology. It provides theoretical support and practical guidance for promoting the progress of measurement quality assurance technology, improving the accuracy and reliability of measurement, promoting cross-regional cooperation and realizing sustainable development in this field.

2 Current Situation of Measurement Quality Assurance Technology

2.1 Limited Accuracy of Measuring Equipment

2.1.1 Specific Performance

1) Limited measurement range

Some measuring devices can only guarantee accuracy within a specific measurement range, and the error increases sharply beyond the range. For some scenarios that need to measure extremely small or extremely large physical quantities, the existing metering equipment cannot meet the requirements. For example, in the micro field, for the measurement of nanoscale size [7], traditional metrology equipment is often difficult to achieve the required accuracy. However, in the macroscopic domain, for the measurement of super-sized objects, there may also be insufficient accuracy.

2) Low resolution

The resolution of a metrological device determines the minimum change of physical quantities it can distinguish [8]. Some metrological devices have low resolution and cannot accurately detect small physical quantity changes. The low resolution can also lead to spurious readings during the measurement, that is, the measurement results appear stable, but in fact cannot truly reflect the actual changes in the physical quantities due to the insufficient resolution.

3) Poor stability

The measurement accuracy of metering equipment may change during long time use. Some devices have high accuracy at the initial stage of use, but with the passing of time, due to the aging and wear of internal components, the accuracy gradually decreases. Environmental factors have a great impact on the stability of metering equipment [9]. Environmental changes such as temperature, humidity, and electromagnetic fields will cause fluctuations in the measurement results of metering equipment.

2.1.2 Related Cases

In the process of electronic chip manufacturing, the size of the chip is getting smaller and smaller, and the accuracy requirement of the metering equipment is also getting higher and higher [10]. For example, for the fabrication of nanoscale chips, metrological devices such as lithography machines and electron microscopes need to be used with high precision. If the accuracy of these devices is limited, it will lead to deviations in the chip circuit, which will affect the performance and reliability of the chip. In the production process of an electronic manufacturing enterprise, due to the lack of accuracy of the measurement equipment used, a number of chips have quality problems. After testing, it was found that the size deviation of the chip exceeded the design requirements, because the metering equipment was affected by the ambient temperature during the measurement process, resulting in inaccurate measurement results. The company had to rework this batch of chips, resulting in huge economic losses.

In the laboratory department of the hospital, biochemical analyzer is one of the commonly used measurement equipment [11]. If the accuracy of the biochemical analyzer is limited, it will lead to inaccurate test results and affect the doctor's judgment of the patient's condition. Due to the lack of accuracy of the biochemical analyzer used in the liver function test of a patient in a hospital, the test results were biased. Doctors made treatment plans based on erroneous test results, causing pain and financial burden to patients.

2.2 Data Processing Techniques Are Uneven

1) Traditional data processing technology

Based on statistical methods: such as the calculation of mean, variance, standard deviation and other statistics, the data can be preliminarily described and analyzed. It is characterized by simple and intuitive calculation, but its ability to analyze complex data is limited.

Linear regression analysis: Used to study the linear relationship between two or more variables. It can establish a prediction model, but it is weak in dealing with nonlinear relationships [12].

Fourier transform: The time domain signal is converted into a frequency domain signal, which facilitates the analysis of the frequency component of the signal. It is widely used in the field of signal processing, but the effect of processing non-stationary signals is not good.

Due to the limitations of traditional data processing techniques, the information in the data cannot be fully mined, resulting in limited accuracy of the results. Traditional techniques have weak ability to deal with outliers and noise, which will interfere with the results.

2) Modern data processing technology

Machine learning algorithms: including support vector machines, random forests, neural networks, etc. It can automatically learn patterns and rules in data, and has powerful classification, regression and prediction capabilities. However, a large amount of data is required for training, and the interpretability of the model is relatively weak.

Big data processing technology: such as Hadoop, Spark and other distributed computing frameworks. It can process large-scale data with high throughput and scalability. However, it requires high hardware and technical requirements, and data security and privacy protection are challenges.

Deep learning algorithms: specifically convolutional neural networks and recurrent neural networks, etc. It has achieved great success in image recognition, speech processing, natural language processing and other fields. It has strong feature extraction and modeling capabilities, but requires large computational resources and requires long training time.

Modern data processing techniques are able to deal with complex data more effectively and improve the accuracy of the results.

2.3 Industry Standards are not Uniform

Different countries and regions have different development processes in the field of metrology, which leads to different emphasis and investment in metrology quality assurance technology. Some developed countries have established relatively complete measurement systems in the early stage, while some developing countries have started this aspect late. Different levels of economic development make countries have gaps in the upgrading of measuring equipment, technology research and development, and then lead to industry standards are difficult to unify.

At present, although there are some international metrological organizations, the efforts in coordinating the industry standards of various countries are not enough. Various countries and regions often give priority to their own interests and needs when formulating standards, and lack of unified understanding and implementation of international standards. The imperfection of international coordination mechanism is also reflected in the lack of adequate communication and consultation in the process of standard formulation. Countries often go their own way when setting standards, and no effective cooperation mechanism has been formed.

There are differences in the research and development and application of metrology technology in different countries and regions. Some advanced countries may have mastered high-precision measurement technology, while some backward regions are still using more traditional measurement methods. This imbalance in technical development makes it difficult to unify industry standards.

The continuous emergence of new technologies also brings challenges to the unification of industry standards. Different countries and regions have different acceptance degrees and application speeds of new technologies, which makes it difficult to synchronize the formulation and update of standards.

3 Achievements Achieved at the Current Stage of Measurement Quality Assurance Technology

3.1 High-Precision Measuring Instruments Continue to Be Introduced

1) Nanoscale measuring instruments

Scanning probe microscopy (SPM): A sharp probe is used to scan the surface of a sample, and the morphological information of the sample surface is obtained by detecting the interaction force between the probe and the sample. Its resolution can reach the atomic level, which enables accurate measurement of nanoscale structures.

Atomic force microscopy (AFM): is a measurement instrument based on inter-atomic forces, which enables high-resolution imaging by detecting weak forces between a probe and a sample. It can work in different environments such as atmosphere, liquid and vacuum, and is suitable for measuring biological samples, polymer films, etc. [13].

Scanning Electron microscopy (SEM): An electron beam is used to scan the surface of a sample to produce secondary electron images with high resolution and magnification. It can observe the microstructure of nanometer scale, and has a wide range of applications in materials science, biology and other fields.

2) Optical measuring instruments

Interferometer: By measuring the interference fringes of two beams of light to determine the optical path difference, so as to achieve high-precision measurement of object size, shape and other parameters.

Spectrometers: Analyze the absorption, emission, or scattering properties of a substance to different wavelengths of light, thereby determining the composition and structure of the substance. High-resolution spectrometers can accurately measure small spectral changes, and have important applications in chemical analysis, environmental monitoring and other fields.

Optical profiler: The use of optical principles to measure the three-dimensional shape of the object surface, with the characteristics of non-contact, high precision. It can be used to measure the roughness and shape of the surface of semiconductor chips, optical components and so on.

3) Electronic measuring instruments

Digital oscilloscope: used to measure the waveform, frequency, amplitude and other parameters of electrical signals. Digital oscilloscopes with high bandwidth and high sampling rate can accurately measure high-speed electronic signals, which are widely used in electronic engineering, communication and other fields.

Logic analyzer: Used to analyze the signal logic relationship in digital circuits. Multiple signal channels can be monitored at the same time, and the timing and logic state of digital signals can be accurately analyzed.

High precision electronic scale: The use of advanced sensor technology and digital signal processing technology, can achieve high precision weight measurement. It has important applications in laboratory, industrial production and other fields.

3.2 International Cooperation Has Been Strengthened

CGPM is organized by BIPM, where representatives from various countries participate to discuss and formulate international measurement standards [14].

The European Organization for Metrological Cooperation (EURAMET) and the Asia Pacific Metrological Program (APMP) promote metrological cooperation and coordination among countries in the region. These organizations carry out measurement comparison, technical exchange and training and other activities to improve the level of measurement in the region.

International metrological Mutual Recognition Agreement. By signing a metrological mutual recognition agreement, countries recognize each other's metrological results and reduce trade barriers. The International Recommendations on Metrological Law (MRA) provides a framework for the mutual recognition of measurement results among countries.

3.3 Data Processing and Analysis Technology Progress

Machine learning algorithms can automatically learn patterns and regularities in data, detect and process outliers and noise, and improve the accuracy of data. In metrology data, if there are outliers, machine learning algorithms can be used to identify and correct them, so as to improve the quality of data. Machine learning algorithms can analyze historical measurement data to establish quality prediction models, predict the quality trend of products in advance, and provide decision support for quality control. Through the analysis of historical production data, machine learning algorithms can be used to predict the qualified rate of products, adjust production processes and parameters in advance, and improve product quality.

Big data technology can analyze large-scale measurement data, find potential errors and inconsistencies in the data, and improve the accuracy of the data. By comparing and verifying the data from multiple data sources, errors and contradictions in the data can be found and corrected in time.

4 Development Trend of Measurement Quality Assurance Technology

Intelligence and automation: measuring instruments will have stronger intelligent perception and data processing capabilities. Intelligent sensors can automatically collect, process and transmit metering data to achieve real-time monitoring and remote control. At the same time, the device can automatically adjust the measurement parameters

according to the changes of the measurement environment and the object, and improve the accuracy and reliability of the measurement. In industrial production, quality testing and other fields, the measurement process will be more and more automated. Through the integration with the production line, the real-time measurement and monitoring of the key parameters in the production process are realized, and the errors and uncertainties caused by human intervention are reduced. Moreover, the automatic measurement system can automatically judge and analyze the data according to the preset standards and procedures, and find the quality problems in time and issue an early warning.

High precision and high stability: With the continuous progress of science and technology, more advanced measurement technology will be applied to the field of metrology, such as laser interferometry, nanometer measurement, quantum measurement, etc. These technologies can provide higher measurement accuracy and resolution to meet the metrology demand for micro-sized, high-precision products. In semiconductor manufacturing, precision machining and other industries, the demand for measurement accuracy is increasing, and advanced measurement technology will play an important role.

Datamation and informatization: a large amount of measurement data will be collected and stored, and through the deep mining and analysis of these data through big data analysis, artificial intelligence and other technologies, potential quality problems and rules can be found. Information technology will promote the sharing and collaboration of measurement information. Measurement institutions in different regions and departments can realize the sharing and exchange of measurement data through the network platform to improve the utilization efficiency of measurement resources. At the same time, the internal measurement department and production, quality, research and development departments will also strengthen information collaboration, to achieve efficient flow and application of measurement data within the enterprise, to provide strong support for the quality management of enterprises.

Internationalization and standardization: International cooperation in measurement technology will be continuously strengthened, and countries will jointly promote the development and application of measurement technology. Under the coordination of the International Metrology Organization, countries will strengthen the unification and mutual recognition of measurement standards and improve the international comparability of measurement results. At the same time, the international exchange and cooperation of measurement technology will promote the dissemination and application of advanced measurement technology, and promote the improvement of global measurement technology level. Standardization is an important means of quality assurance measurement, the future will continue to improve the measurement standard system. On the one hand, we should strengthen the revision and update of existing measurement standards to meet the needs of scientific and technological development and industrial upgrading. On the other hand, the formulation of measurement standards in emerging fields, such as new energy, intelligent manufacturing and digital economy, will be accelerated to provide technical support for the development of these fields.

5 Conclusion

Currently, metrological quality assurance technology is facing a series of challenges, including accuracy limitations of metrological equipment, inconsistency of data processing technology, and inconsistent industry standards. These problems affect, to varying degrees, the accuracy of measurement results, the efficiency of cross-regional cooperation, and the development process in various fields.

However, with the continuous emergence of high-precision measurement instruments, such as nanoscale measurement devices, scanning probe microscopes, and atomic force microscopes, we have more powerful tools to cope with the needs of microscopic and high-precision measurements. These new instruments play a vital role in several fields such as material science, biomedicine, electronic engineering, and aerospace due to their high resolution, high accuracy, versatility, and automation and intelligence.

The strengthening of international cooperation has also played a role in promoting the development of metrology technology. Through the coordination of international metrology organizations, transnational scientific research cooperation projects and international cooperation between enterprises, we are moving toward the unification of measurement standards, the innovation of measurement technology and the improvement of the quality of measurement personnel.

In terms of data processing and analysis technology, the application of machine learning algorithms, big data technology and cloud computing technology has significantly improved the accuracy of data, enhanced the fault diagnosis ability, and optimized the quality control process.

In addition, by developing more intelligent metering equipment and automated inspection systems, we can further improve the accuracy and reliability of measurement while reducing human errors by realizing intelligent perception, automatic parameter adjustment, automated metering process and seamless integration with the production line.

Finally, promoting the unification of international standards and strengthening global cooperation can not only improve the international comparability of measurement results, but also promote international technical exchange and cooperation. At the same time, it is also a key task for us to continuously improve the measurement standard system to adapt to the development needs of emerging fields.

Fund. Project Supported by Science and Technology of Project State Grid Corporation of China (No.5700-202321624A-3-2-ZN).

References

1. Wu, Z., Wen, H.: Analysis on Great power diplomacy with Chinese characteristics from independence to rejuvenation and self-strength. *Northeast Asia Forum* **28**(05), 3–16+127 (2019). (in Chinese)
2. Xing, J., Li, J.Y., Wang, P.S., et al.: Research on intelligent prediction method of laboratory measurement equipment verification error. *Mach. Electron.* **41**(10), 49–53 (2023). (in Chinese)

3. Gao, J., Zhai, Y., Li, M., et al.: Research status of health care big data processing in the field of precision medicine. *Chin. Hosp. Manage.* **41**(05), 8–13 (2021). (in Chinese)
4. Pei, P., Cao, B., Ge, X., et al.: Research on high-precision measurement technology of Track Spacing of Electromagnetic Gun. *J. Baltics* **36**(2), 105–110 (2024). (in Chinese)
5. Jin, H., Rao, Z., Qin, K.: Online magnification calibration of feature size scanning electron Microscopy Based on wafer grid standard samples. *J. Meas. Sci. Technol.* **67**(02), 29–35 (2023). (in Chinese)
6. Discussion on the establishment of measurement standardization system and evaluation criteria. *Public Standardization* (16), 1–3 (2024). (in Chinese)
7. Bazo, A., Fernandez, B.E., Ibarz, R.A., et al.: Intensity- and time-based strategies for micro/nano-sizing via single-particle ICP-mass spectrometry: a comparative assessment using Au and SiO₂ as model particles. *Analytica Chimica Acta* (2024). 1331343305-343305
8. Zhang, F.: Research on high resolution demodulation method of full spectrum fiber FP sensor. University of electronic science and technology (2024). <https://doi.org/10.27005/dcnki.Gdzku.2024.002537>. (in Chinese)
9. Ma, X., Jin, X.: Improving the stability of measuring equipment and consolidating the foundation of measurement integrity system. *Industrial Metrol.* **27**(02), 79 (2017). (in Chinese)
10. Jasper, M.: DARPA Announces Microelectronics Program to Boost Domestic Manufacturing of Certain Chips. *Nextgov.com* (Online) (2021)
11. Zilka, S., Wei, R., Payto, D., et al.: Analytical and operational considerations of measuring glucose 6-phosphate dehydrogenase (G6PD) activity using a fully automated assay. *Am. J. Clin. Pathol.* **163**, 153–160 (2024)
12. Cerdo, T., Moral, T.T.: Potential risk genes for primary Sjogren’s syndrome from a meta-analysis by linear regression and random forest classification. *Genes Diseases* **11**(3), 101016 (2024)
13. Liang, Z., Jiang, Z., Tang, X., et al.: Mechanical properties of shale during pyrolysis: atomic force microscopy and nano-indentation study. *Int. J. Rock Mech. Mining Sci.* (2024). 2024183059, 29-105929
14. The 27th International Metrological Conference passed 7 important resolutions. *Railway Technical Supervision* **50**(12), 38 (2022). (in Chinese)



Snowy Urumqi County Postal Logistics Road Transportation Route Planning Study

Mengxi Wu¹, Bing Wang^{1(✉)}, and Yingqi Li²

¹ School of Transportation Engineering, Xinjiang University, Urumqi 830047, China
451619037@qq.com

² School of Mechanical Engineering, Xinjiang University, Urumqi 830047, China

Abstract. Logistics and transportation route planning is crucial in snowy winter weather conditions. In this paper, we study the path planning of postal logistics transportation in Urumqi County in snowy days, firstly, we establish the optimization model with the objective of shortest time, which covers the path impedance, the speed of the graded highway and the mileage of the different graded highway, and secondly, we solve the model according to the dynamic planning method, and the simulation verification based on MATLAB shows that the obtained scheme can effectively reduce the logistics transportation by the interference of abnormal weather, reduce the transportation time, and improve the efficiency and reliability of logistics transportation, which is a reference for the transportation decision-making under other abnormal weather conditions. The efficiency and reliability of logistics transportation can provide reference for postal logistics transportation decision-making in other abnormal weather.

Keywords: Snowy · Logistics · Route Planning · Dynamic optimization

1 Introduction

Conventional logistics transportation and emergency response in Xinjiang basically rely on highway transportation, Xinjiang is a vast area with a low density of highway distribution, which is characterized by the distribution of highways in a wide area with a sparse road network. However, in recent years, the weather in Xinjiang has become worse and worse, not only the common rain, snow, fog and other phenomena, but also encountered wind-blown snow and other diversified meteorological conditions, which seriously threaten the road traffic safety, thus affecting the transportation of logistics.

Vehicle path problem is one of the most classic optimization problems in the field of logistics, which has been widely concerned by scholars at home and abroad for many years, and has made great progress and rich research results so far [1]. Li Yi [2] studied the urban cold chain logistics distribution problem, the model is divided into the path optimization under static demand and increased consideration of dynamic demand in two parts, making the minimum additional distribution costs, and finally the results of the example analysis shows that the fusion co-distribution is conducive to reducing the cost of

enterprise distribution. Chen Cheng et al. [3] for the consideration of vehicle speed spatio-temporal dynamics of the vehicle path problem, the establishment of the distribution network speed matrix so as to calculate the travel time, the model to the operating cost and carbon emission cost as the goal, through the example analysis shows that in the vehicle path model to incorporate the speed of the spatial and temporal consideration can be better portrayed in the urban logistics and distribution. Zhichao Ma et al. [4] took the minimization of distribution cost and maximization of customer satisfaction as the goal, took the actual data of an enterprise as an example, used MATLAB software to find the optimal distribution path, and locally optimized the distribution path according to the immediate customer demand. Yunxiao [5] studied the cold chain logistics distribution problem under the F2B2C mode, constructed an optimization model, which considered the constraints of time-varying road network, and took into account the cost of cargo damage and refrigeration cost in distribution, and used the real data of an enterprise to get the reliable results in the arithmetic experiment. The current research on the vehicle path problem has a fairly long history, and the field has matured, but most of the literature does not consider the effect of path impedance generated under abnormal weather.

2 Optimization Study of Transportation Path Based on Minimum Time

2.1 Description of the Problem

Abnormal weather in Xinjiang not only has the common phenomena of rain, snow, fog, etc., but also wind-blown snow and other diversified meteorological conditions, which seriously threaten road traffic safety, through the review of the literature [6] to understand that the general transportation vehicles in all types of meteorological and road icing conditions under the highest safe speed as shown in Table 1.

Table 1. Table of Maximum Safe Driving Speeds for Special Climatic and Meteorological Conditions and Special Situations.

snowfall	velocity
neutral snow	90 km/h
snowstorms	81 km/h
blizzard	45 km/h
great blizzard	27 km/h
extremely heavy snowstorm	18 km/h

2.2 Modeling Assumptions

Construct a traveler's problem model [7] for township transportation, with the goal of the shortest total logistics transportation time [8], and make the following assumptions

about the model: Assumption 1: there is and only one transportation center; Assumption 2: the geographic location of both the transportation center and the transportation node is known and does not change, and there is an exact location of the transportation node for each transportation task; Assumption 3: the transportation paths from the transportation center to each node, and from node to node, and before nodes There is a straight line transport, there is no curve transport trajectory; Assumption 4: the transport process, the transport task can not be interrupted, and has been executed until the completion of the transport task; Assumption 5: the transport time is equal to the sum of the path time; Assumption 6: there is only one type of cargo model in the transport center and the vehicle carrying capacity is clear, and the sum of the distribution of the transport route can not exceed the vehicle's clear carrying capacity; Assumption 7: the use of the vehicle from the logistics center and returns to the logistics center after delivering goods to all nodes, and the demand of each node is satisfied and served once.

2.3 Parameter Description

In this paper, we study the node-to-node transportation problem under snowy conditions, where the round trip time between any two nodes is asymmetric [9], and the relevant parameters are shown in the following Table 2.

Table 2. Definitions of Variables and Parameters in the Model.

Relevant parameters	define
X_{ijk}	Decision variable, whether to choose the k th path from node i to node j , $= 1$ if chosen, 0 otherwise
T_{ijk}	denotes the time required for the k th path from node i to node j
m_{ij}	denotes the number of different paths from location i to location j
Subsets S	A subset of the set of locations, excluding the center node. By constraining the different subsets, it is ensured that no subloops are formed
$ S $	denotes the size of the subset S
d_{ijk}	denotes the road length of the k th path from node i to node j
v_{ijk}	denotes the speed of the k th path from node i to node j

2.4 Objective Function Construction

In the actual situation of logistics transportation, since there are often several different routes between various nodes, the highway grades involved in these routes may differ, and the specified travel speeds of different grades of highways are also different, thus obtaining Eq. 1, which represents the transportation time of the k th route from node i to

node j . The objective of the shortest logistics transportation function is as follows, Eq. 2 represents the choice of the total time used to traverse all the nodes of the post office.

$$T_{ijk} = \frac{d_{ijk}}{v_{ijk}} \quad (1)$$

$$\min T = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^{m_{ij}} T_{ijk} X_{ijk} \quad (2)$$

2.5 Constraint Function Construction

In order to determine the shortest transportation time for logistics, find an overall relatively optimal constraint, and solve the actual optimal solution for different routes according to the constraints. According to the modeling assumptions in the previous section, set Eqs. 3 to 5 as the constraints of the model as.

$$\sum_{j=1}^n \sum_{k=1}^{m_{ij}} X_{ijk} = \sum_{j=1}^n \sum_{k=1}^{m_{ij}} X_{jik} \quad (3)$$

$$\sum_{j=2}^n \sum_{k=1}^{m_{ij}} X_{ijk} = 1 \quad (4)$$

$$\sum_{j=1}^n \sum_{k=1}^{m_{ji}} X_{jik} = 1 \quad (5)$$

Equation 3 indicates that each node is visited once and only once, i.e., the total number of paths into the node is equal to the total number of paths out of the node; and Eq. 4 and Eq. 5 indicate the final return to the central node after traversing all nodes.

In the traveler problem, a complete loop should return to the starting node after traversing all nodes exactly once. However, there are one or more small closed loops that appear in the planned paths that do not cover all nodes, so a subloop elimination constraint is set, and for any subset $S \subset \{2, 3, \dots, n\}$ (that does not contain the center node), there is $\sum_{i \in S} \sum_{j \in S} \sum_{k=1}^{m_{ij}} X_{ijk} \leq |S| - 1$, this constraint ensures that the number of paths chosen in any subset does not exceed the size of the subset minus one, thus avoiding the formation of subloops.

3 Case Study

3.1 Data Collection

This paper takes the postal logistics in Urumqi County as the research object, the townships under the jurisdiction of Urumqi County [10] are ShuiXiGou Township, AnNingQiu Township, BanFangGou Township, YongFeng Township, GanGou Township, Tori Township, according to the statistics of the distribution of Urumqi County postal office outlets as shown in Fig. 1.



Fig. 1. Distribution of postal logistics outlets in Urumqi County.

In this paper, the Banfanggou postal branch office is the center point to the other seven business outlets transported, according to the actual situation in Urumqi County, the coordinate system is established with the Xibaiyanggou Natural Scenic Area as the origin, as shown in Fig. 2.



Fig. 2. Urumqi County Coordinate System.

The specific coordinates of each business outlet are measured according to the latitude and longitude of each outlet, and the latitude and longitude as well as the coordinates of each business outlet are shown in Table 3 below.

Through the research to understand that the common abnormal weather in Urumqi County is snowy days, so this paper takes snowy days as an example, the closer to the center of the snowfall, the greater the amount of snowfall is usually, as the distance extends outward, the amount of snowfall generally decreases gradually [11], from the center of the 20–30 km or so, it may be from the blizzard to a large or medium snow; to 40–50 km, it may be changed to a small snow or sporadic snowflakes falling. In this paper, the slate ditch postal branch as the center of the snowfall point that is blizzard, 10 km outward snowfall from large to small gradual transition, according to the formula (1) combined with according to the above various types of meteorological and road icing conditions under the maximum safe speed table, calculated between the post office business outlets in normal weather and the snow day the shortest time data table shown in Table 4 (brackets for the time data of the snow day).

Table 3. Latitude, longitude and coordinates of postal outlets in Urumqi County.

Serial number	Business network	longitudes	Longitude (i.e. angle)	X-axis coordinate	Y-axis coordinate
1	Banfangou Post Office	87.418531	43.480229	2.2	0.6
2	Yongfengqu Post Office	87.353541	43.617575	1.8	2
3	Shuisigou Post Office	87.485171	43.473584	2.7	0.5
4	Gangou Township Post Office	87.18343	43.520397	0.5	1
5	Xiaodonggou Post Office	87.530843	43.450624	3.1	0.3
6	Xishan Farm Post Office	87.384544	43.686523	2	2.8
7	Saldasaka Township Post Office	87.274009	43.472829	1.2	0.4
8	Tori Township Post Office	87.661516	43.467415	4.1	0.5

Table 4. Data table of the shortest time for postal outlets in Urumqi County during normal weather and snowy days (minutes).

Names of business outlets and their corresponding numbers		1	2	3	4	5	6	7	8
		BF	Yfq	Shxg	GgT	Xdg	XshF	Saldasaka	TTs
1	BF	0	29(51)	13(19)	44(69)	21(35)	40(63)	27(38)	32(44)
2	Yfq	30(51)	0	38(50)	22(39)	49(68)	14(21)	21(36)	50(85)
3	Shxg	12(18)	37(51)	0	52(62)	13(20)	47(62)	34(46)	24(33)
4	GgT	41(62)	20(40)	50(67)	0	63(70)	30(48)	30(48)	69(94)
5	Xdg	24(34)	49(69)	13(20)	64(73)	0	51(83)	44(57)	21(37)
6	XshF	44(64)	17(20)	48(63)	33(52)	53(84)	0	30(43)	48(96)
7	Saldasaka	26(38)	20(36)	33(46)	28(48)	44(58)	28(44)	0	60(96)
8	TTs	34(44)	53(85)	24(33)	74(103)	21(35)	46(97)	60(109)	0

3.2 Model Solution

Based on the shortest logistics transportation time determination through the dynamic planning algorithm [12] for model solving, through the overall relative optimal constraints on each post office node iterative calculation of the respective optimal route, and ultimately get the optimal route, the solution process is as follows [13].

Step 1: Firstly, all the post offices within the regional scope of Urumqi County are labeled S_i ($i = 1, 2, 3, \dots, 19$), Then mark the roads between the post offices X_{ijk} ($k = 1, 2, 3, \dots, n$), used $(1, S)$ to denote that it starts from the center point 1 of Banfangou postal branch, passes through each post office in the set S once, and finally returns to the center point 1 of Banfangou postal branch, where the set S is the set of post offices, and the initial state is that when S is the empty set, that is to say, it represents that all the nodes have been accessed, and at this time, if the current node is the node 1, then the length of the path is 0, That is, $D(1, \text{empty set}) = 0$.

Step 2: List the state transfer equations [14] $D(1, S) = \min\{d_{1jk} + D(j, S - \{j\})\}$, The equation represents the shortest time to start from node 1, pass through each of the nodes of the set S once, and finally return to the starting node.

Step 3: Initialize the number of selected path segments L_k from node S_1 to node S_j and the design speed of the road class corresponding to each segment of L_k , and calculate the time from S_1 to each road node S_j according to Eq. (6) T_{1jk} ($k = 1, 2, 3, \dots$).

$$T_{1jk} = \sum_{k=1}^N \frac{L_k}{v_k} \quad (6)$$

Step 4: Create a set N_j , compare the T_{1jk} obtained in step 3, put the path of the shortest T_{1jk} into the set N_j , and add 1 to j or 1 to 1 to continue the computation after each step.

$$N_i = \{X_{ijk}\} \quad (7)$$

Step 5: Repeat steps 3 and 4 to determine whether X_{ijk} has computed all the node to node roads, and repeat steps 3 and 4 above when X_{ijk} has not computed all the roads.

Step 6: Solve the optimal value from the bottom up, calculate the optimal paths from node 1 to all the nodes in the set S in turn, when the last position $j-1$ is calculated, it is the optimal value of the whole problem.

3.3 Case Simulation Verification

Using the steps in Sect. 3.2, the shortest route under normal weather conditions is $1 \rightarrow 7 \rightarrow 4 \rightarrow 2 \rightarrow 6 \rightarrow 8 \rightarrow 5 \rightarrow 3 \rightarrow 1$, i.e., from Banfanggou Postal Sub-office $\rightarrow$ Sardasan Township Post Office $\rightarrow$ Gangou Township Post Office $\rightarrow$ Yongfengqu Postal Sub-office $\rightarrow$ Xishan Farms Post Office $\rightarrow$ Tori Township Post Office $\rightarrow$ Xiaodonggou Post Office $\rightarrow$ Shuixigou Post Office $\rightarrow$ Banfanggou Postal Sub-office, and the shortest route time is 183 min. The optimal transportation route map is shown in Fig. 3.

Similarly, the shortest time route is $1 \rightarrow 7 \rightarrow 6 \rightarrow 5 \rightarrow 2 \rightarrow 4 \rightarrow 8 \rightarrow 3 \rightarrow 1$, i.e., from Banfanggou Postal Branch Office $\rightarrow$ Sardazhan Township Post Office $\rightarrow$ Xishan Farm Post

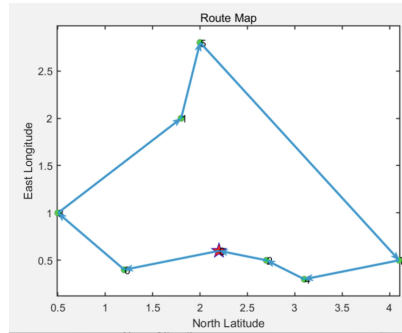


Fig. 3. Optimal Distribution Roadmap in Normal Weather.

Office→ Xiaodonggou Post Office→ Yongfengqu Postal Branch Office→ Gangou Township Post Office→ Tori Township Post Office→ Shuixigou Post Office→ Banfanggou Postal Branch Office, with a shortest time of 299 min, and the optimal distribution The path diagram is shown in Fig. 4.

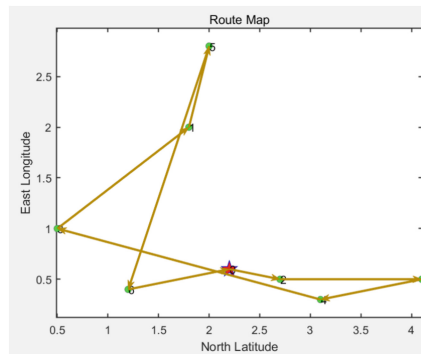


Fig. 4. Optimal Distribution Roadmap for Snowy Days.

4 Conclusion

4.1 Analysis of Results

In order to highlight the effect of the improvement, the logistics transportation is carried out in snowy days according to the optimal route obtained in normal weather, i.e., from Banfanggou postal branch office→Sardazhan Township post office→Gangou Township post office→Yongfengqu postal branch office→Xishan Farm post office→Toli Township post office→Xiaodonggou post office→Shuisigou postal branch office→Banfanggou post office, which takes a total of 316 min, whereas the shortest time of the improved route is 299 min, thus the improved route time is 17 min shorter than the previous route time.

4.2 Evaluation of Optimized Transport Routes

Optimized routes reduce the amount of time cargo is in transit, making it less dangerous for vehicles to travel in snowy weather. Optimized routes can try to avoid areas susceptible to snowy weather or choose safer roads, which not only protects the driver's life, but also reduces the possibility of cargo damage. For logistics companies, it can also reduce the cost of compensation due to accidents and the loss of business interruption, thus improving the reliability and stability of logistics services.

Acknowledgments. This work was supported by the project of Natural Science Foundation of the Autonomous Region - Safe Driving Speed Modeling on Ice and Snow Roads and Its Application in Dynamic Speed Limiting VMS (2018D01C054).

References

1. Li, M.: Optimization of cooperative distribution of UAVs and vehicles considering the impact of abnormal weather. Dalian Maritime University (2021). (in Chinese)
2. Li, Y.: Research on the optimization of urban cold chain logistics co-distribution path under the background of new retail. Beijing Jiaotong University (2020). (in Chinese)
3. Chen, C., Liu, Y., Lin, Q., et al.: An urban delivery vehicle path problem considering spatio-temporal dynamics of vehicle speed. *Ind. Eng. Manage.* **26**(03), 56–62 (2021). (in Chinese)
4. Ma, Z., Zhang, J., Wang, H., et al.: Optimization of sustainable bi-objective cold-chain logistics route considering carbon emissions and customers' immediate demands in China. *Sustainability* **15**(7) (2023)
5. Yun, X.: Research on cold Chain Kaban multi-temperature co-distribution vehicle path problem under F2B2C model. Dalian Maritime University (2021). (in Chinese)
6. Shi, X.: Research on first aid quantitative dispatch system platform based on Xinjiang highway traffic characteristics. Xinjiang University (2017). (in Chinese)
7. Zhang, Y., Zhang, G., He, N., et al.: Multi-objective distribution route optimization simulation based on improved genetic algorithm. *Logist. Eng. Manage.* **45**(10), 33–37+3 (2023). (in Chinese)
8. Meilita, S.T., Steven, C.: Ant colony optimization implementation on traveling salesman problem to achieve the shortest logistic route. *IOP Conf. Ser. Mater. Sci. Eng.* **1003**(1), 012045 (2020)
9. Aguayo, M.M., Avilés, F.N., Sarin, S.C., Sherali, H.D., et al.: A two-index formulation for the fixed-destination multi-depot asymmetric travelling salesman problem and some extensions. *Informatica* **33**(4), 671–692 (2022)
10. Zhu, S.: Research on the current situation of farmers' household finance management in Urumqi County. Xinjiang Agricultural University (2012). (in Chinese)
11. Tao, Y.-W., Zhang, F.-H., Yu, B.-X., et al.: Causes and forecasting analysis of extreme snowfall in Altay that triggered avalanches in January 2024. *Heavy Rainfall Disasters* **43**(04), 407–418 (2024). (in Chinese)
12. Wan, Y., Li, X.: Dynamic planning method of product multimodal logistics transportation path. *Int. J. Manuf. Technol. Manage.* **36**(2–3–4), 241–256 (2022)
13. Tian, T.: Optimized configuration of emergency disposal equipment for natural disasters on Xinjiang highways. Xinjiang University (2019). (in Chinese)
14. Li, J.: Application of dynamic programming in optimizing transportation routes. *Logist. Eng. Manage.* **46**(04), 32–35 (2024). (in Chinese)



A Dual-Stream Convolutional Network for Visible and Infrared Image Fusion in Pedestrian Detection

Xinyan Xiong^{1,2}, Jie Yang^{1,2}(✉), Yanxuan Jiang^{1,2}, and Xinkai Hu^{1,2}

¹ School of Electrical Engineering and Automation, Jiangxi University of Science
and Technology, Ganzhou 341000, Jiangxi, China
yangjie@jxust.edu.cn

² Jiangxi Provincial Key Laboratory of Maglev Rail Transit Equipment, Ganzhou, China

Abstract. In complex environments, pedestrian detection faces problems such as low visibility, poor image contrast and noise interference, which seriously affect the accuracy and safety of detection. In this paper, a pedestrian detection method based on improved dual-stream convolution channel based on visible and infrared image fusion is proposed. Firstly, taking advantage of the thermal infrared image under low light conditions, a dual-stream feature extraction network was constructed to enhance the fusion effect of visible light and infrared features. Then, the lightweight YOLOv8n is used as the basic network, and the features are extracted through the parallel improved CSPDarknet53 to ensure real-time performance in complex scenes. Finally, we conduct experiments on the LLVIP dataset, and the results show that the proposed method achieves 0.984 in average accuracy (mAP@50), which is 4.9% higher than that of the basic network, and significantly reduces the number of parameters by 20.1% and the amount of computation by 17.8%. This method has superior performance in the accuracy, robustness and real-time performance of image detection, and provides a new solution for pedestrian detection.

Keywords: Traffic safety · visible and infrared fusion · object detection · pedestrian detection

1 Introduction

Pedestrian detection is an important research topic in the field of computer vision [1], with widespread applications in autonomous driving, intelligent security, and traffic safety. However, the limitations of a single RGB sensor [2] significantly impact detection performance in complex scenarios such as rainy days, foggy weather, and low-light conditions [3]. As shown in Fig. 1, images captured under weak light conditions often suffer from insufficient information [4], making it difficult to recognize pedestrian intrusions and potentially compromising traffic safety.

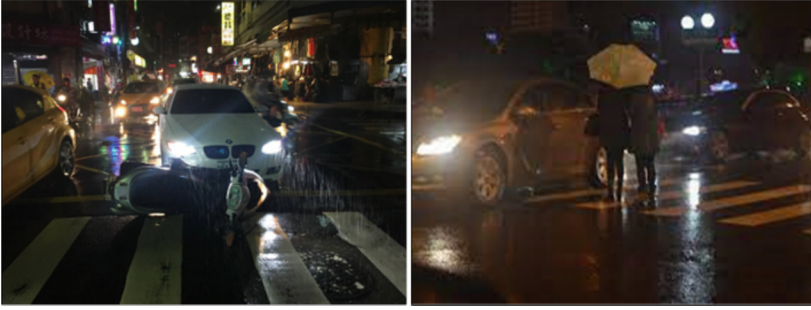


Fig. 1. Traffic safety under low-light conditions.

To address this challenge, some studies [5–7] have attempted to incorporate additional sensors to assist RGB sensors. Among these, methods that combine infrared sensors with RGB sensors have demonstrated particularly impressive performance, especially in low-light environments, significantly enhancing detection accuracy.

Infrared imaging and RGB images possess different physical characteristics. Infrared sensors excel at capturing the thermal radiation and temperature information of objects in low-light environments, while RGB sensors are better suited for obtaining color information. Combining both can fully leverage their respective advantages, effectively enhancing the accuracy and robustness of nighttime pedestrian detection.

Traditional fusion methods often rely on Aggregated Channel Features (ACF) [8] or spatial domain pixel transformations to directly process images. However, these methods overly depend on manual feature extraction [9], making it difficult to fully utilize the features of both modalities. Therefore, more refined fusion strategies are needed. In recent years, deep learning-based fusion methods have shown remarkable results. For example, Gao [10] et al. extracted deep features from images using residual networks and quantified the information of the original images, thereby more effectively guiding the training of the fusion network. Wang [11] et al. employed an adaptive weight fusion algorithm to achieve multimodal fusion, ensuring high fusion quality while reducing computational costs. However, these studies do not address the issues of information loss during fusion and the high complexity of models, leading to inadequate detection performance in complex environments [12].

To tackle this issue, this paper proposes an improved dual-stream convolution channel network, Dual-YOLO, specifically designed for the extraction of visible and infrared features while reducing the number of channels. This method effectively enhances the integrity of feature fusion and the performance of the model, providing a new solution for pedestrian detection in complex scenarios. Specifically, the contributions of our work include the following:

- (1) A pedestrian detection method based on dual-stream convolution channels for visible and infrared image fusion is proposed, aiming to enhance detection performance in complex environments.

- (2) A lightweight YOLOv8n is used as the base network, and a dual-stream convolution channel is constructed through parallel improved CSPDarknet53 to extract visible and infrared features, achieving effective information fusion.
- (3) Experiments on the LLVIP dataset demonstrate that the improved model achieves an average precision (mAP@50) of 0.984, which is a 4.9% improvement over the base network, while significantly reducing parameter count and computational complexity. This reflects the method's superiority and application potential in various complex environments.

2 Methodology

2.1 Constructing a Two-Stream Convolutional Channel

Thermal infrared images have stronger contrast and target recognition capabilities in low-light environments such as nighttime, effectively compensating for the limitations of visible light images. A pedestrian detection system using visible light and infrared fusion can provide more stable detection results in various complex environments, enhancing overall detection performance. YOLOv8n is a lightweight model suitable for operation in resource-constrained environments [13], with lower parameter count and complexity compared to other variants of YOLOv8. This study constructs a dual-stream convolution channel based on YOLOv8n for feature extraction of visible light and infrared images, termed Dual-YOLO. The dual-stream convolution channel consists of two parallel improved CSPDarknet53 structures, forming the backbone for feature extraction in the entire network. Through continuous convolution operations, it extracts visible light features and infrared features from the original images, allowing the model to merge information from different sources at the same level, thereby enhancing the richness and expressive capability of the features and improving the model's understanding of complex scenes. In the improved network, the convolution layers use smaller kernels, reducing from the original 64 and 128 to 32 and 64, thereby lowering the model's complexity and computational cost while maintaining basic feature extraction capabilities.

The inputs to the network are visible and infrared images in the same state space with an input size of 640×640 and the outputs are the classification target, the detection frame position and the confidence level of the target. The visible and infrared images are firstly passed through the dual-stream feature extraction channels P1 to P5 respectively to get five feature layers, four of the feature layers are fused through the fusion Concat module, the fused features are fused and enhanced through the feature pyramid network (FPN) of the neck network and adaptive feature aggregation for feature fusion and enhancement, to achieve multi-scale fusion of the features, and the final fused feature maps are sent into the detection cast Obtain the target information, position information and confidence information in the source image.

In Fig. 2, F2, F3, F4, and F5 are the output feature maps of modules P2, P3, P4, and P5, respectively, while UP denotes the upsampling structure. Its output is shown in Eq:

$$F2 = \text{Conv}(\text{Conv}(C2F(x))) \quad (1)$$

$$F2' = \text{Conv}(\text{Conv}(C2F(x'))) \quad (2)$$

$$F_{i+1} = \text{Conv}(C2F(F_i \oplus F_i')) \quad i = 3, 4, 5 \quad (3)$$

$$F_{i+1}' = \text{Conv}(C2F(F_i')) \quad i = 3, 4, 5 \quad (4)$$

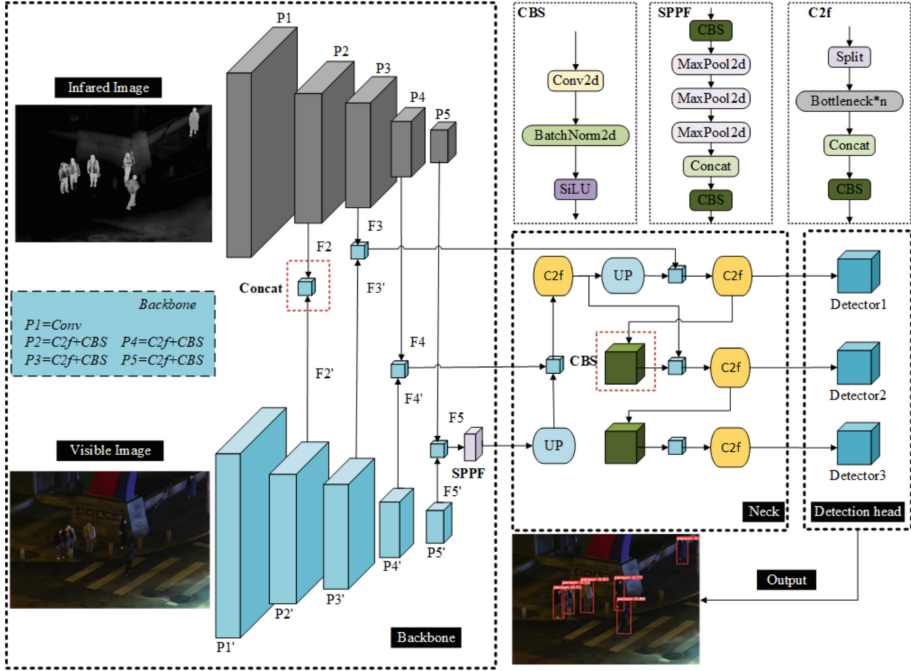


Fig. 2. Dual-YOLO network structure diagram.

3 Experiments

3.1 Dataset

Low-light pedestrian detection uses the public dataset LLVIP for model training. LLVIP stands for A Visible-infrared Paired Dataset for Low-light Vision [14], which is a dataset of visible and infrared images related to low-light vision. It contains 30,976 images, or 15,488 pairs, across 24 nighttime scenes and 2 daytime scenes, with a training period of 200 epochs. The images in the LLVIP dataset include various lighting and weather conditions, such as night-time with varying levels of artificial lighting, as well as challenging outdoor conditions that may include shadows, reflections, and varying degrees of illumination from streetlights or headlights. It is specifically designed for low-light vision tasks, with the images including both visible and infrared modalities. A portion of the dataset is shown in Fig. 3.



Fig. 3. Images from the LLVIP subset of the dataset.

3.2 Experimental Environment

The algorithm in this study is implemented in Visual Studio Code on a CentOS 7 operating system, using the PyTorch deep learning framework and an Nvidia TITAN RTX GPU (24G) for network training. The training employs the stochastic gradient descent (SGD) method with a momentum of 0.9, for a total of 200 training epochs, with eight images processed per batch and a learning rate of 0.01. The evaluation metrics used in this study are mean average precision (mAP), parameter count (parameters), and floating point operations per second (FLOPs). The average precision mAP_50 represents the mean precision for all object categories at an intersection over union (IoU) threshold of 0.5, while mAP_95 indicates the mean detection precision for all object categories at an IoU threshold of 0.95.

3.3 Experimental Results

After determining the improved network structure, this paper conducts comparative experiments of the current mainstream target detection models on the LLVIP dataset by using target detection algorithms including YOLOv5n, YOLOv5s, YOLOv6n, YOLOv8n, YOLOv8s, YOLOv10n, and the improved Dual-YOLO. The experimental results are shown in Table 1.

In Table 1, it can be observed that among the YOLO series models, YOLOv8s achieves the highest accuracy, with an mAP@50 value of 0.948. However, its parameter count reaches 11.14M, and the computational load is 28.6 GFLOPs. In contrast, the base network used in this study, YOLOv8n, has an average precision mAP@50 of 0.938, with a parameter count of 3.00M and a computational load of 8.1 GFLOPs. By constructing

a dual-stream convolution channel, the average precision of the model is improved to 0.984, which is an increase of 4.9% compared to the base network and 3.8% compared to YOLOv8s. In terms of parameter count, this model has 2.44M parameters, representing a reduction of 20.1% compared to the base network and 78.1% compared to YOLOv8s. The computational load is 6.66 GFLOPs, which is a decrease of 17.8% from the base network and 76.7% from YOLOv8s. These results indicate that the dual-stream convolution channel significantly enhances the model’s performance while effectively reducing both parameter count and computational complexity.

Table 1. Experimental structural parameters.

Model	Parameters/M	mAP@50	mAP@95	GFLOPs	Layer
YOLOv5n	2.50	0.932	0.552	7.10	193
YOLOv5s	9.12	0.945	0.598	24.00	262
YOLOv6n	4.23	0.936	0.557	11.80	142
YOLOv8n	3.01	0.938	0.565	8.10	168
YOLOv8s	11.14	0.948	0.603	28.60	225
YOLOv10n	2.69	0.934	0.556	8.20	285
Dual-YOLO	2.44	0.984	0.706	6.66	239

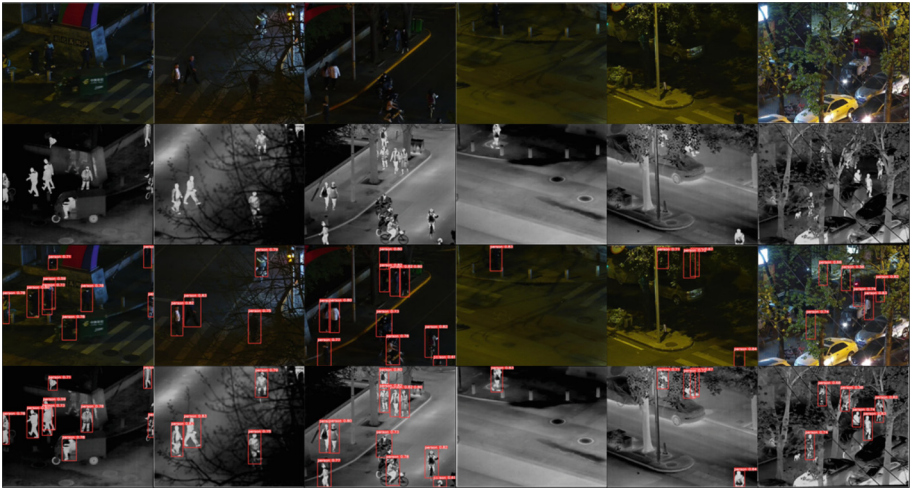


Fig. 4. Visible-Infrared Pedestrian Detection Visualisation Results.

Figure 4 shows the visualisation of the test results of this paper on some data from the LLVIP dataset. Our proposed network demonstrates excellent detection performance. The network effectively adapts to various pedestrian scales, achieving high detection accuracy across small and large targets. Furthermore, it maintains robust performance

under diverse lighting conditions, such as daylight, shadows, and low-light nighttime scenarios, showcasing strong resilience to brightness variations.

The enhanced visible-infrared fusion strategy significantly improves the model's adaptability in complex environments, particularly under low-light and high-contrast scenes, where detection accuracy is markedly improved. Whether identifying subtle target contours in dark settings or capturing fine target features in high-contrast environments, the network exhibits remarkable stability and accuracy. This improved strategy effectively mitigates the impact of lighting and contrast variations, allowing the model to better handle real-world complexities. These results validate the effectiveness and practicality of our approach, underscoring the network's potential in pedestrian detection tasks.

4 Conclusion

This paper proposes a pedestrian detection system based on YOLOv8n with visible and infrared image fusion, designed to enhance detection performance in complex environments. The limitations of visible light images are effectively compensated by exploiting the advantages of thermal infrared images in low light conditions. We constructed a dual-stream convolutional channel and used the improved CSPDarknet53 as the feature extraction backbone to achieve parallel extraction and fusion of visible and infrared features. Experimental results show that the improved model achieves 0.984 in mean accuracy (mAP@50), which is 4.9% higher than the base network YOLOv8n, while significantly reducing the number of parameters and computational complexity. Specifically, the number of parameters of the model is 2.44M and the GFLOPs are 6.66, which are 20.1% and 17.8% lower than the base network, respectively. These results validate the effectiveness and robustness of the proposed method in multiple complex environments and provide new ideas for pedestrian detection.

Acknowledgement. This work was supported by a grant from the National Key Research and Development Project (2023YFB4302100), Key Research and Development Project of Jiangxi Province (No. 20232ACE01011) and Independent Deployment Project of Ganjiang Innovation Research Institute, Chinese Academy of Sciences (E255J001).

References

1. Chen, Z., Guo, H., Yang, J., et al.: Fast vehicle detection algorithm in traffic scene based on improved SSD. *Measurement* **201**, 111655 (2022)
2. Monno, Y., Teranaka, H., Yoshizaki, K., et al.: Single-sensor RGB-NIR imaging: high-quality system design and prototype implementation. *IEEE Sens. J.* **19**(2), 497–507 (2018)
3. Chen, Z., Yang, J., Chen, L., et al.: Efficient railway track region segmentation algorithm based on lightweight neural network and cross-fusion decoder. *Autom. Constr.* **155**, 105069 (2023)
4. Guo, X., Li, Y., Ling, H.: LIME: low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* **26**(2), 982–993 (2016)

5. Liang, M., Hu, J., Bao, C., et al.: Explicit attention-enhanced fusion for RGB-thermal perception tasks. *IEEE Robot. Autom. Lett.* **8**(7), 4060–4067 (2023)
6. Li, G., Liu, Z., Chen, M., et al.: Hierarchical alternate interaction network for RGB-D salient object detection. *IEEE Trans. Image Process.* **30**, 3528–3542 (2021)
7. Shi, C., Lai, G., Yu, Y., et al.: Real-time multi-modal active vision for object detection on UAVs equipped with limited field of view LiDAR and camera. *IEEE Robot. Autom. Lett.*, 6571–6578 (2023)
8. Nguyen, T.N., Ly, N.Q.: Abnormal activity detection based on dense spatial-temporal features and improved one-class learning. In: *Proceedings of the 8th International Symposium on Information and Communication Technology*, pp. 370–377 (2017)
9. Cheng, C., Wang, Q., Nikitin, Y., et al.: A data-driven distributed fault detection scheme based on subspace identification technique for dynamic systems. *Int. J. Robust Nonlinear Control* **33**(5), 3107–3128 (2023)
10. Gao, Y., Cheng, Z., Su, H., et al.: Infrared and visible image fusion method based on residual network. In: *2023 4th International Conference on Computer Engineering and Intelligent Control (ICCEIC)*, pp. 366–370. IEEE (2023)
11. Chen, Z., Yang, J., Feng, Z., et al.: RailFOD23: a dataset for foreign object detection on railroad transmission lines. *Sci. Data* **11**(1), 72 (2024)
12. Wang, Y., Zhu, Z., Li, J., et al.: Feature enhancement method for multimodal nighttime pedestrian detection. In: *2023 7th International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE)*, pp. 1028–1032. IEEE (2023)
13. Terven, J., Córdova-Esparza, D.M., Romero-González, J.A.: A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extract.* **5**(4), 1680–1716 (2023)
14. Jia, X., Zhu, C., Li, M., et al.: LLVIP: a visible-infrared paired dataset for low-light vision. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3496–3504 (2021)



Comparative Analysis of Lane Change Intention Recognition Models Based on the CitySim Dataset

Hanfei Wang^{1,2}(✉) , Huanfeng Liu^{1,2} , Ziyang Wu^{1,2} , Anning Song^{1,2} ,
and Zishuo Zhang^{1,2}

¹ School of Transportation, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
{1202206094, 1202306026, 1202306054,

1202206124}@student.stdu.edu.cn, huanfengliu@stdu.edu.cn

² Hebei Key Laboratory of Traffic Safety and Control, Shijiazhuang 050043, China

Abstract. To accurately identify vehicle lane change intentions and enhance driving safety, this paper utilizes the CitySim dataset to conduct research on lane change intention recognition. First, the driving intentions of target vehicles are categorized into three types: going straight (TS), lane changing to the left (TL), and lane changing to the right (TR). A dimensionality reduction strategy is then employed to eliminate redundant data. Next, lane change intention recognition algorithms are designed based on fundamental principles of K-Nearest Neighbors (KNN), Random Forest (RF), Convolutional Neural Networks (CNN), and Multi-Layer Perceptrons (MLP). Subsequently, five-fold cross-validation is used for hyperparameter optimization of the KNN and RF models, employing techniques such as grid search. For the CNN and MLP models, relevant functions are introduced to dynamically adjust the learning rate and increase the number of training epochs. Finally, to prevent overfitting in the field, Synthetic Minority Over-sampling Technique (SMOTE) is utilized to balance the samples. The results indicate that all four models achieve high accuracy in lane change intention recognition, with the accuracy for TS exceeding 97%. Compared to the other three models, the accuracy of the KNN model is higher by 1, 2, and 3 percentage points, making it the most accurate overall.

Keywords: Lane change intention · CitySim dataset · K-Nearest Neighbors · vehicle trajectory

1 Introduction

In the field of transportation, trajectory prediction is crucial for traffic safety and operational efficiency, making the study of lane change trajectories highly necessary. Currently, research on lane change trajectories primarily focuses on three aspects: lane change intention recognition, lane change behavior prediction, and lane change trajectory planning [1]. Common datasets used in this research include CitySim [2], NGSIM [3], HighD [4], and Cityscapes [5], among others. Selecting appropriate datasets is vital

for improving urban traffic conditions, enhancing road safety, and promoting the development of intelligent transportation technologies. Yang Da [6] and others conducted research on vehicle lane change trajectory prediction using the HighD dataset, achieving high accuracy. Zhu Sunke [7] focused on the influence of spatiotemporal diversity characteristics and hyperparameter sensitivity on the accuracy of vehicle behavior intention recognition, conducting detailed research on lane change intention recognition and verifying models using the NGSIM dataset, with recognition accuracy for 2.5 s before lane change exceeding 91%. Zhao Jiandong and others [8] researched vehicle lane change intention recognition, validating the model's performance using the NGSIM dataset and achieving a recognition accuracy of 97.37%. Yuan Renteng [9] and others selected the CitySim dataset for their research, achieving a recognition accuracy of 96.67%. Datasets are indispensable for the study of lane change trajectories, and using machine learning for research is currently a popular direction.

Machine learning is widely used to solve classification, regression, clustering, and other problems [10], with its applications increasingly impacting urban traffic management and intelligent driving [11]. Most scholars now integrate machine learning with datasets to study lane change trajectories. Traditional machine learning methods for driving intention recognition are characterized by their relatively small data requirements. Support Vector Machines [12] (SVM) are a classic classification algorithm that has gained widespread attention in vehicle lane change intention recognition research. Chen Liang [11] and others designed a vehicle lane change recognition model based on multi-class SVM using the NGSIM dataset, providing support for research on vehicle lane change phases. Subsequently, Vallon [13] and Do [14] also used SVM to complete lane change trajectory predictions. Kumar and others [15] proposed a prediction method combining SVM and Bayesian filtering to enhance recognition effectiveness, achieving shorter average intention prediction times with better robustness. In addition to traditional machine learning methods, deep learning research on lane changes is also a current hotspot. Han Lei and others [16] proposed an end-to-end connected autonomous vehicle lane change decision model based on multi-source information fusion in a mixed traffic flow environment, which enables vehicles to achieve higher lane change success and merging rates. Zhao et al. [17] proposed the Direction Convolution LSTM (DCLSTM) lane change intention recognition method, achieving a recognition rate of 94.10% on the NGSIM US 101 and I-80 datasets. Qasemabadi and others [18] established a lane change intention recognition model based on LSTM, comparing the impacts of different feature parameters and surrounding vehicle quantities on the model, achieving a lane change recognition rate of 92.43% on the HighD dataset.

Currently, there is considerable research on the application of machine learning in lane change intention recognition. However, most existing studies tend to adopt a single model for predicting and recognizing lane change intentions. While this approach can validate the effectiveness of specific algorithms to some extent, it seldom delves into comparative analysis of different models and their respective advantages and disadvantages, overlooking the importance of comprehensively understanding lane change intention recognition mechanisms through comparison of various methods.

Considering this, this paper utilizes the open-source CitySim dataset to conduct in-depth research on vehicle lane change intention recognition, employing traditional

machine learning methods such as KNN and Random Forest (RF), along with deep learning techniques like CNN and MLP. A systematic comparative analysis of different models is performed, scientifically evaluating their performance effects through a controlled variable method.

2 Trajectory Data Extraction

Vehicle trajectory data, by analyzing changes in vehicle positions and times, can not only identify specific lane change actions but also predict future lane change trends based on historical data. This is important for the development of autonomous driving technology and optimization of intelligent transportation systems.

2.1 CitySim Dataset

This study selects the CitySim dataset for lane change intention recognition. This dataset is extracted from drone video footage captured at 1140 different locations over a total duration of 12 min, covering a wide area. Additionally, this experiment utilizes a sub-dataset from a specific segment of freeway-B in China, which not only reflects domestic traffic conditions well but also enhances the accuracy and effectiveness of lane change behavior recognition.

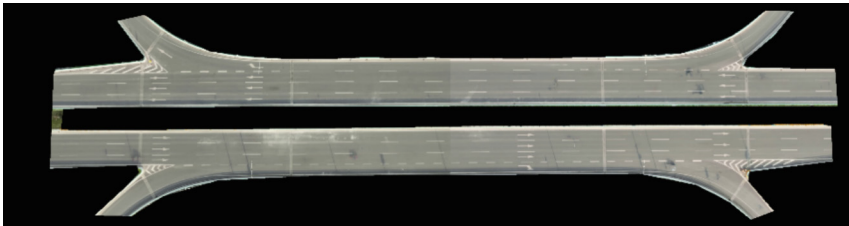


Fig. 1. Road Testing Scene of CitySim

Figure 1 shows the road-testing scene of the freeway-B sub-dataset in CitySim, designed with six bi-directional lanes, each lane being 3.75 m wide, and a total length of 680 m. Two drones collected data in this area at a rate of 30 frames per second over a continuous 60-min period, recording a total of 4225 vehicle trajectories. The original dataset contains pixel coordinates of each vehicle at seven testing points and information about their respective lanes, with the specific layout of testing points illustrated in Fig. 2.

In the CitySim dataset, in addition to providing the actual latitude and longitude values of each point, a coordinate system defined within the group is also used for labeling. This approach offers multiple options for subsequent researchers but results in a significant amount of duplicated data. Therefore, this paper selects only key nodes for analysis, excluding those with excessive duplication rates. The following table introduces the relevant point information used in the experiment (Table 1):

Table 1. Character Encoding and Explanation of Key Points

Character	Meaning
CarID	Vehicle ID
FrameNum	Data frame number: 30 frames collected per second
CarCenterX	X-coordinate of the vehicle bounding box center pixel
CarCenterY	Y-coordinate of the vehicle bounding box center pixel
Speed	Vehicle speed
Heading	Heading angle, used to determine if the vehicle is changing lanes
LaneID	The ID of lane

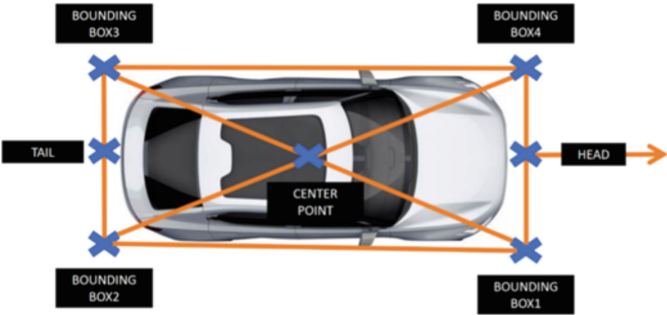


Fig. 2. Schematic of Vehicle Testing Point Locations in CitySim

2.2 Data Preprocessing

To ensure the accuracy of the results, the data undergoes preprocessing, which includes three parts: removal of anomalous data, filtering, and data normalization:

- (1) **Removal of Anomalous Data:** The original data, merged from the synchronized footage of two drones, contains missing trajectories or interruptions between frames. Therefore, in the early stages, IDs of the same vehicle appearing more than once in adjacent data frames are excluded from the study.
- (2) **Data Filtering:** First, the laneId column in the dataset is used to determine the lane difference (lane_diff) between adjacent lanes. Based on this difference, lane change behaviors are classified: if $\text{lane_diff} < 0$, it indicates a lane change to the left (TL), marked as 1; if $\text{lane_diff} > 0$, it indicates a lane change to the right (TR), marked as 2; otherwise, the vehicle is classified as going straight (TS), marked as 0. Additionally, for continuous left or right lane changes, only the first lane change marking (1 or 2) is retained, while subsequent markings are reset to 0. Moreover, if the vehicle maintains its lane for an extended period, a down-sampling strategy is adopted, collecting data every 30 steps during that continuous interval. If no lane change occurs, it is marked as NaN. Finally, all rows marked as NaN are removed

to prevent duplicate records of TS. The classification and data processing flow for vehicle trajectories are illustrated in Fig. 3.

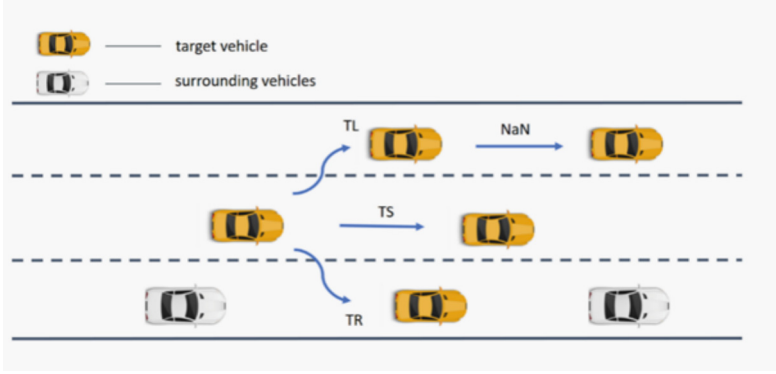


Fig. 3. Vehicle Trajectory Classification and Data Processing

- (3) **Data Normalization:** To eliminate dimensional influences and accelerate model convergence, all metrics are standardized to facilitate faster achievement of optimal solutions.

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

In the formula, x represents the original feature value, μ is the meaning of the feature, σ is the standard deviation of the feature, and x' is the standardized feature value.

3 Lane Change Intention Recognition

This study utilizes the CitySim dataset to predict vehicle lane change intentions through KNN, RF, CNN, and MLP algorithms. A 5-fold cross-validation method is employed, with 80% of the data used for training and the remaining 20% for testing. During the training of the CNN and MLP models, a TQDM progress bar callback function is introduced to dynamically adjust the learning rate and increase training cycles. KNN and RF undergo hyperparameter optimization through techniques such as grid search. Additionally, to ensure a balanced distribution of samples across categories and avoid overfitting during model training, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to balance the number of samples in the TS, TL, and TR categories, enhancing the overall classification performance of the models. Figure 4 illustrates the overall experimental framework.

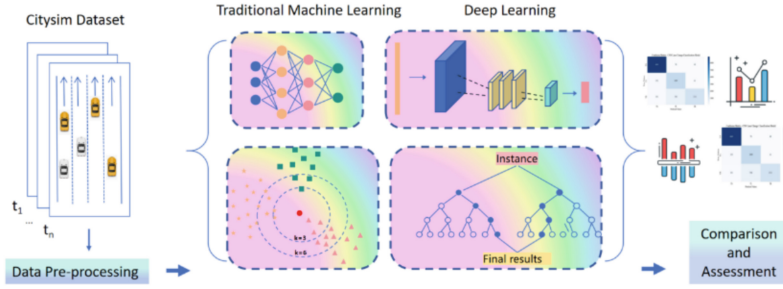


Fig. 4. Lane Change Intention Recognition and Comparative Analysis

3.1 Lane Change Intention Recognition Models

- (1) **KNN:** The K-Nearest Neighbors (KNN) algorithm is one of the classic algorithms in traditional machine learning. In an effective KNN traffic flow prediction model, when predicting the class of a new sample, the model determines the class of the new sample based on the classes of the K nearest training samples. In this study, the `n_neighbors` parameter indicates the number of neighbors and is set to [3, 5, 7, 9] in GridSearchCV to test different neighbor counts. The `weights` parameter defines the weighting type, set to ['uniform', 'distance'], where uniform treats all neighbors equally, and distance assigns weights based on proximity. The `metric` parameter defines the distance metric, set to 'euclidean' for Euclidean distance. In the parameter settings of GridSearchCV, estimator is set to the used classifier, configured as `KNeighborsClassifier()`.
- (2) **Random Forest (RF):** Random Forest generates new training sample sets by repeatedly and randomly sampling k samples from the original training sample set N with replacement. It then constructs k decision trees based on the bootstrap sample sets and aggregates their results through voting or averaging to obtain the final prediction. In this study, the `n_estimators` parameter specifies the number of deciding trees in the forest, set to 200. The `max_depth` parameter indicates the maximum depth of the trees, set to 15. The `min_samples_split` parameter is set to 5, indicating that if the number of samples in a node is less than 5, further splitting will not occur. The `min_samples_leaf` parameter is set to 2, indicating the minimum number of samples required in a leaf node. The `random_state` parameter controls the seed for random number generation to ensure consistent results across runs.
- (3) **CNN:** Convolutional Neural Networks (CNN) are deep neural networks specifically designed to handle data with a grid-like structure, primarily comprising convolutional layers, activation functions, pooling layers, and fully connected layers. CNN captures spatial structural information of input data through convolutional layers, reduces data dimensions via pooling layers, and maps the extracted features to specific task outputs through fully connected layers. The CNN model in this study consists of three convolutional layers, each followed by a pooling layer, utilizing 64, 128, and 256 filters with a size of 2×1 and employing the ReLU activation function. The fully connected layer contains 256 nodes, with a dropout rate of 0.5, and the output layer comprises three nodes using the Softmax activation function.

- (4) **MLP:** The Multi-Layer Perceptron (MLP) is a fundamental artificial neural network model structured with multiple layers of neurons. Its basic structure includes an input layer, an output layer, and at least one or more hidden layers. Each layer consists of multiple neurons, each generating output by summing weighted input values and passing through an activation function. During training, MLP updates the weights and biases between neurons through backpropagation to minimize the error between predicted and actual outputs.

The MLP model in this study features multiple fully connected layers, with 512, 256, 128, 64, and 32 nodes, and incorporates batch normalization and dropout (with dropout rates of 0.4, 0.4, 0.3, 0.3, and 0.2) between each layer. Each hidden layer uses the ReLU activation function, and L2 regularization is applied to prevent overfitting.

Moreover, both CNN and MLP utilize the Adam optimizer, with the loss function set to sparse_categorical_crossentropy and a dynamic learning rate adjustment strategy to optimize the training process. The training epochs for CNN are set at 150, while those for MLP are set at 200.

3.2 Experimental Result Analysis

To understand the basic conditions of the dataset, providing background information for subsequent processing and model training, and to ensure the uniqueness and accuracy of the data, the sample sizes and the number of unique carId entries are checked. A total of 35,040 vehicle trajectory records are extracted from the database, including 34,724 TS trajectories and 316 lane change trajectories, comprising 152 TL and 164 TR vehicle trajectories. After data filtering and removal of duplicates, the final confirmed effective vehicles number 791, with a total of 10,945 trajectories.

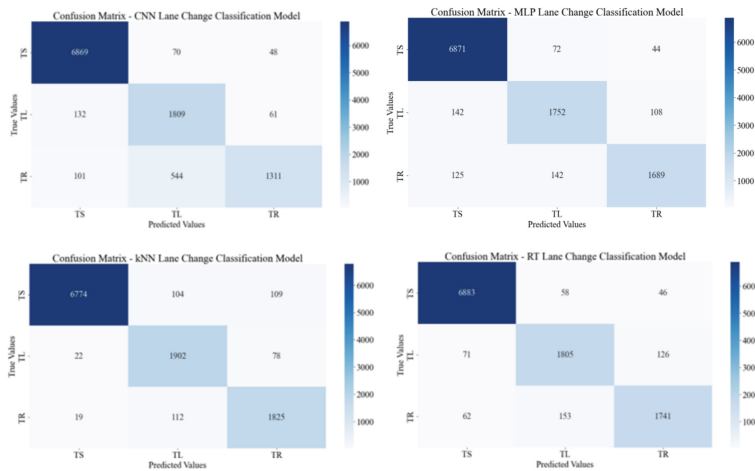


Fig. 5. Comparison of Confusion Matrices

- (1) **Visualization Analysis:** Using seaborn, a heatmap of the confusion matrix is drawn to showcase the correspondence between the model's prediction results and the true labels. The confusion matrices of each model are shown in Fig. 5.

As illustrated, the horizontal axis represents predicted values, and the vertical axis represents true values. The darkest square among the nine indicates the overlap between true and predicted values, representing correct predictions. The second and third squares of the first row indicate instances where true TS values are incorrectly predicted as TL and TR, respectively. The model demonstrates relatively accurate predictions for TS, TL, and TR, particularly for TS, as its larger sample proportion provides more learning opportunities during training, leading to higher prediction accuracy. Furthermore, compared to TL and TR, TS actions are likely easier to differentiate based on basic kinematic features (such as changes in speed, acceleration, and direction angle). Notably, CNN shows larger recognition errors for TL and TR, suggesting they were not adequately trained.

- (2) **Evaluation Metric Analysis:** Lane change intention recognition is a multi-class problem. This paper uses F1 score and accuracy as evaluation metrics for lane changes.

① **F1 Score:** The F1 score is the harmonic mean of precision and recall, providing a single value to evaluate classifier performance. Precision refers to the proportion of true positive predictions among all predicted positive data, while recall refers to the proportion of true positive samples among all actual positive samples. The formulas are as follows:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100\% \quad (2)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100\% \quad (3)$$

"In the formula, P represents precision, R represents recall, N_{TP} denotes the number of true positive samples that are correctly classified as positive, N_{FP} denotes the number of false positive samples that are incorrectly classified as positive, and N_{FN} denotes the number of false negative samples that are incorrectly classified as negative.

The formula for calculating the F1 score combines the above metrics:

$$F1 = 2 \times \frac{PR}{P + R} \times 100\% \quad (4)$$

② **Accuracy:** Accuracy measures the proportion of correctly classified samples by the classifier, calculated as the ratio of correctly predicted samples to the total number of samples:

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{FN} + N_{TN}} \quad (5)$$

In the formula, A represents the accuracy, N_{TN} is the number of samples that are negative and correctly classified as negative. The accuracy and F1 score comparison results for the four models are shown in Fig. 6.

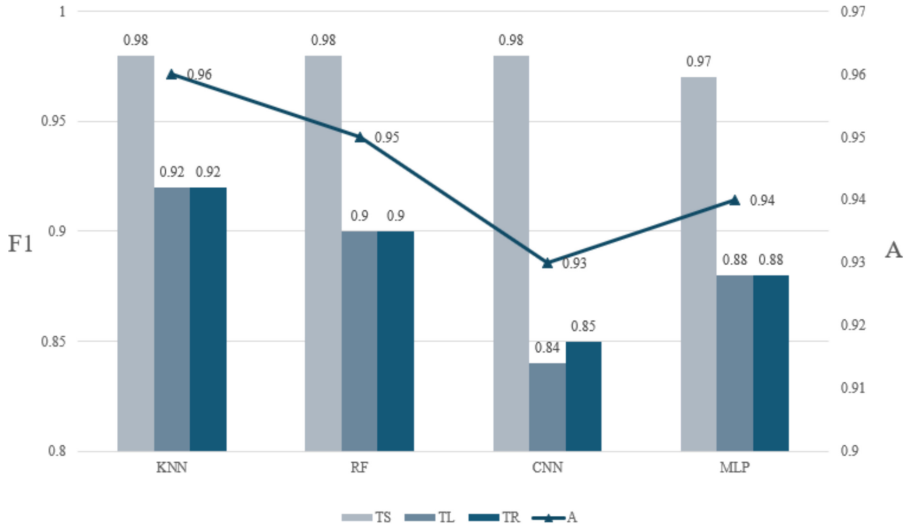


Fig. 6. Comparative Analysis of Prediction Abilities of Different Models

Figure 6 utilizes accuracy and F1 scores to analyze the lane change intention recognition performance of the four models. Overall, the accuracy of the KNN model is 0.96, while the accuracies of RF, MLP, and CNN models are 0.95, 0.94, and 0.93, respectively. Compared to the other three models, the accuracy of the KNN model is higher by 1, 2, and 3 percentage points. In terms of F1 scores, the KNN model achieves 0.92 for the TL category and 0.90 for the TR category, both higher than those of the other three models. The F1 score for the TS category is 0.98, with little difference compared to the other three models. Overall, the KNN model, with its excellent performance in specific categories and the highest overall accuracy, appears to be the best choice.

Additionally, both CNN and MLP achieve high F1 scores for the TS category, at 0.98 and 0.97, respectively, indicating that deep learning models possess strong capabilities for this category. However, for TL and TR categories, the performance of deep learning models is less stable than that of traditional machine learning models, particularly in the TL category, where the F1 score of CNN is only 0.84, while KNN and RF exceed 0.9. Thus, in this lane change intention recognition task, traditional machine learning algorithms outperform deep learning models in most categories, especially in the TL category.

4 Conclusion

Lane change intention recognition can predict whether drivers will change lanes by analyzing dynamic vehicle data and environmental information, thereby enhancing driving safety, improving traffic efficiency, and supporting Connected Autonomous Vehicles (CAVs) in making appropriate decisions. This study employs the open-source CitySim dataset to compare and analyze the effectiveness and performance of K-Nearest

Neighbors (KNN), Random Forest (RF), Convolutional Neural Networks (CNN), and Multi-Layer Perceptron (MLP) models.

- (1) The four models in this study indicate that traditional machine learning methods are more effective for small datasets and fewer features when addressing lane change intention recognition issues, allowing for rapid training without requiring extensive computational resources. In contrast, while deep learning models can handle complex patterns, their performance is inferior to that of traditional machine learning models when feature quantities are limited, and data is insufficient.
- (2) The comparison of the four classic models reveals that both CNN and MLP achieve high F1 scores for the TS category, at 0.98 and 0.97, respectively. However, for the TL and TR categories, particularly in the TL category, the F1 score of CNN is only 0.84, while KNN and RF exceed 0.9. This indicates that deep learning models are less stable in performance compared to traditional machine learning models.
- (3) Compared to the other three models, the KNN model's accuracy is higher by 1, 2, and 3 percentage points. In terms of F1 scores, the KNN model outperforms the other three models in both the TL and TR categories, with minimal differences in the TS category. Overall, the KNN model demonstrates excellent performance in certain specific categories and achieves the highest overall accuracy.
- (4) Among the four models, the predictions for the TS category are the most accurate, primarily because the TS category comprises a larger proportion of samples within the dataset. This provides the model with more opportunities to learn how to correctly classify these samples during training, thereby enhancing prediction accuracy for this category. Moreover, compared to TL and TR, the actions associated with TS may be easier to differentiate based on fundamental kinematic features. Accurately recognizing the lane change intentions of drivers is of significant importance, especially in the future mixed traffic flow environment comprising connected autonomous vehicles and traditional manually driven vehicles. Real-time identification of drivers' lane change intentions can assist connected autonomous vehicles and autonomous vehicles in making timely and correct decisions, which is crucial for enhancing road safety. The CitySim dataset, as an open-source repository of real human driving data, offers valuable research resources. In the future, the availability of more high-quality open datasets for autonomous driving will further support intelligent systems in learning human driving habits and enhance model precision to some extent.

Acknowledgments. This work was funded by Science and Technology Project of Hebei Education Department, project No. QN2020151.

References

1. Song, R., Li, B.: Surrounding vehicles' lane change maneuver prediction and detection for intelligent vehicles: a comprehensive review. *IEEE Trans. Intell. Transp. Syst.* **23**(7), 6046–6062 (2021)

2. Zheng, O., Abdel-Aty, M., Yue, L.: CitySim: a drone-based vehicle trajectory dataset for safety-oriented research and digital twins. *Transp. Res. Rec.* **2678**(4), 606–621 (2023)
3. Coifman, B., Li, L.: A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset. *Transp. Res. Part B Methodol.* **105**, 362–377 (2017)
4. Krajewski, R., Bock, J., Kloecker, L., Eckstein, L.: The High D dataset: a drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In: 21st international conference on intelligent transportation systems (ITSC), pp. 2118–2125. IEEE (2018)
5. Cordts, M., et al.: The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 3213–3223. IEEE (2016)
6. Yang, D., Liu, J.W., Zheng, B.: A vehicle lane change trajectory prediction model based on time domain convolution network and attention mechanism. *J. Transp. Syst. Eng. Inf.* **24**(02), 114–126 (2024). (in Chinese)
7. Zhu, S.K., Yan, J.R., Xiong, K., A, B.: Whale algorithm optimized CNN-BiGRU-ATTENTION model for vehicle lane change intention recognition. *J. Chongqing Univ. Technol. (Nat. Sci.)* **38**(06), 73–80 (2024). (in Chinese)
8. Zhao, J.D., Zhao, Z.M., Qu, Y.C., Xie, D.F., Sun, H.J.: Research on vehicle lane change intention recognition driven by trajectory data. *J. Transp. Syst. Eng. Inf.* **22**(04), 63–71 (2022). (in Chinese)
9. Yuan, R.T., Wang, C.Z., Xiang, Q.J., Zheng, O., Ding, S.X.: Trajectory data-driven vehicle lane change intention recognition model. *J. South China Univ. Technol. (Nat. Sci. Edn.)* **52**(06), 34–44 (2024). (in Chinese)
10. Tian, S.J., Zhang, Y.M.: Overview of machine learning algorithms and their applications. *Software* **44**(07), 70–75 (2023). (in Chinese)
11. Chen, L., Feng, Y.C., Li, Q.: Research on vehicle lane change behavior recognition model based on multi-class SVM. *J. Saf. Environ.* **20**(01), 193–199 (2020). (in Chinese)
12. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995). <https://doi.org/10.1007/BF00994018>
13. Vallon, C., Ercan, Z., Carvalho, A., Borrelli, F.: A machine learning approach for personalized autonomous lane change initiation and control. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 1590–1595. IEEE (2017)
14. Do, Q.H., Tehrani, H., Mita, S., Egawa, M., Muto, K., Yoneda, K.: Human drivers based active-passive model for automated lane change. *IEEE Intell. Transp. Syst. Mag.* **9**(1), 42–56 (2017)
15. Kumar, P., Perrollaz, M., Lefevre, S., Laugier, C.: Learning-based approach for online lane change intention prediction. In: 2013 IEEE Intelligent Vehicles Symposium (IV), pp. 797–802. IEEE (2013)
16. Han, L., Zhang, L., Guo, W.A.: Connected autonomous vehicle lane change decision model based on MSIF-DRL in mixed traffic flow environment. *J. Beijing Jiaotong Univ.* **47**(05), 148–161 (2023). (in Chinese)
17. Zhao, N., Wang, B., Lu, Y., Rong, S.: Direction convolutional LSTM network: Prediction network for drivers' lane-changing behaviours. In: 17th International Conference on Control & Automation (ICCA), pp. 752–757. IEEE (2022)
18. Nejadhossein Qasemabadi, A., Mozaffari, S., Rezaei, M., Ahmadi, M., Alirezaee, S.: A novel model for driver lane change prediction in cooperative adaptive cruise control systems. *arXiv e-prints*, pp arXiv-2305 (2023)



Enhanced SlowFast Networks with Dark Channel Prior for Nighttime Pedestrian Detection in Port Surveillance

Han Zhang¹, Xinqiang Chen²(✉), Dezhi Han³, Yiwen Zheng¹, Bing Han⁴,
and Yongsheng Yang²

¹ College of Ocean Science and Engineering,
Shanghai Maritime University, Shanghai 201306, China

{202440410003, 202440412001}@stu.shmtu.edu.cn

² Institute of Logistics Science and Engineering, Shanghai Maritime University,
Shanghai 201306, China

chenxinqiang@stu.shmtu.edu.cn, yangys@shmtu.edu.cn

³ Shanghai Ship and Shipping Research Institute Co., Ltd., Shanghai 201306, China
dzhan@shmtu.edu.cn

⁴ College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
han.bing@coscoshipping.com

Abstract. This paper proposes a sophisticated methodology for detecting pedestrian behaviors in nighttime surveillance videos in port environments. Given that ports operate continuously, reliable nighttime monitoring is critical, particularly under suboptimal lighting conditions. This research leverages advanced computer vision and deep learning techniques to enhance detection accuracy in low-visibility scenarios. Utilizing a dataset compiled from real port surveillance footage, we assess the efficacy of various deep learning models, particularly the integration of the SlowFast network with dark channel enhancement methods, such as dark channel prior and adaptive transmission refinement. These enhancements address the issues associated with low illumination. Our results demonstrate substantial improvements in detection accuracy compared to baseline approaches, thereby presenting a robust framework for augmenting safety and operational efficiency in ports.

Keywords: Pedestrian Detection · Port Surveillance · Nighttime Video · Dark Channel Enhancement · SlowFast Network · Deep Learning

1 Introduction

Ports are pivotal nodes in the global trade and logistics infrastructure, operating continuously throughout the day and night. Effective nighttime monitoring is vital to ensure safety and operational continuity, particularly as reduced illumination and unique environmental factors complicate pedestrian detection. Accurate detection of pedestrian activities in these low-light conditions remains challenging, as conventional pedestrian detection algorithms often struggle under such circumstances due to fluctuating lighting, reflective surfaces, and other environmental complexities.

Pedestrian detection under nighttime or low-light conditions has been a prominent area of research. A range of methods have been explored to improve detection accuracy under these challenging conditions. Approaches such as infrared and thermal imaging have been employed to capture clearer visuals of pedestrians despite the lack of ambient light [1]. The combination of visible and infrared images has also yielded promising results, enhancing detection accuracy by fusing complementary data sources [2, 3]. However, thermal cameras have several limitations, including their higher cost and reduced resolution when compared to visible light cameras [4].

To mitigate these issues, several image enhancement techniques have been specifically designed for nighttime video analysis. Among these, the dark channel prior (DCP) and adaptive enhancement methods have proven effective in improving image clarity and contrast in low-light environments. Originally devised for haze removal, the DCP method enhances nighttime images by refining contrast and reducing noise, thereby enabling more effective object detection and segmentation [5]. These enhancements not only improve visual clarity but also facilitate better training of convolutional neural networks (CNNs) for pedestrian detection [6].

Recent advances in deep learning have introduced specialized architectures, such as the Faster R-CNN and YOLO models, which are commonly used for real-time pedestrian detection due to their robustness and computational efficiency. However, their performance can degrade in low-light conditions without appropriate pre-processing. Several studies have proposed integrating image enhancement techniques like DCP with CNN-based detectors, leading to significant performance improvements in low-light scenarios [3, 7]. Enhanced YOLO versions, such as the TIFEEY model, combine thermal imaging and CNN-based feature extraction to further improve detection performance, demonstrating high efficacy in low-light surveillance settings [8].

This study builds on these methodologies by integrating dark channel enhancement techniques with the SlowFast network for pedestrian detection in nighttime port environments. The SlowFast network, which effectively captures temporal features in video data, offers an advantage in monitoring pedestrian behaviors across sequential frames, particularly in complex surveillance scenarios with moving objects and fluctuating lighting conditions. By leveraging the SlowFast network's capabilities in conjunction with dark channel enhancement methods, this approach significantly improves detection accuracy and reliability, establishing a comprehensive framework that can be implemented in port surveillance systems for better nighttime monitoring.

2 Dark Channel Enhancement and SlowFast Networks for Nighttime Pedestrian Detection

For the detection of pedestrian behaviors in nighttime port surveillance videos, leveraging advanced model architectures and enhancement techniques is indispensable given the unique challenges of low-light environments. The SlowFast network, designed to capture both spatial semantics and temporal dynamics, is particularly effective for video-based behavior analysis. Its dual-pathway design, which includes a Slow pathway for capturing coarse semantic information and a Fast pathway for capturing rapid temporal dynamics, makes it particularly adept at identifying complex pedestrian movements even under challenging visibility conditions [9].

Specifically, the Slow pathway processes input at a lower frame rate τ , capturing rich spatial details:

$$F_{\text{slow}}(t) = \text{SlowPath}(X(t \cdot \tau)) \quad (1)$$

The Fast pathway processes input at the native frame rate, focusing on temporal dynamics:

$$F_{\text{fast}}(t) = \text{FastPath}(X(t)) \quad (2)$$

The features from both pathways are fused to obtain a comprehensive representation:

$$F_{\text{fusion}}(t) = F_{\text{slow}}(t) + \beta \cdot F_{\text{fast}}(t) \quad (3)$$

where β is a scaling factor that balances the contributions of the Fast pathway.

To enhance the efficacy of detection in low-light scenarios, image enhancement techniques such as dark channel prior (DCP) and its adaptive variations are integrated to improve visibility. Integrating DCP-based de-fogging within YOLO-based detection models has demonstrated potential for improving object classification under poor visibility, largely due to attention-based mechanisms [10]. These mechanisms effectively emphasize pertinent visual features while suppressing irrelevant background noise, thereby enhancing the model's ability to detect pedestrians. Additional cross-task alignment techniques have been introduced to improve the feature representation of deep learning models in low-light conditions, thereby further refining detection capabilities [11]. Figure 1 illustrates the proposed network architecture for pedestrian behavior detection in nighttime low-light conditions. This model integrates various image enhancement techniques with advanced neural networks to improve detection performance.

Furthermore, studies focusing on low-illumination image enhancement for surveillance have consistently shown that improving image quality leads to substantial gains in model accuracy in low-light applications, such as UAV surveillance. Techniques like brightness enhancement using hyperbolic tangent curves, coupled with image denoising, have yielded robust improvements in detection performance [12]. The brightness enhancement can be mathematically expressed as:

$$J_{\text{enhanced}}(x) = \tanh(\alpha \cdot J_{\text{input}}(x)) \quad (4)$$

where $J_{\text{input}}(x)$ is the input image intensity at pixel x , and α is a parameter controlling the enhancement strength. Complementary to this, attention-based methods, such as the bright channel prior attention mechanism, have demonstrated the capability to dynamically focus on relevant features, thereby enhancing recognition accuracy across thermal and visible light spectrums [13].

3 Experiment

3.1 Model Training

The training phase involved optimizing the proposed enhanced SlowFast network using a custom dataset compiled from nighttime port surveillance cameras. The dataset was meticulously annotated to ensure high-quality labels, encompassing a range of conditions such as fog, rain, and variable illumination. The model was trained using the

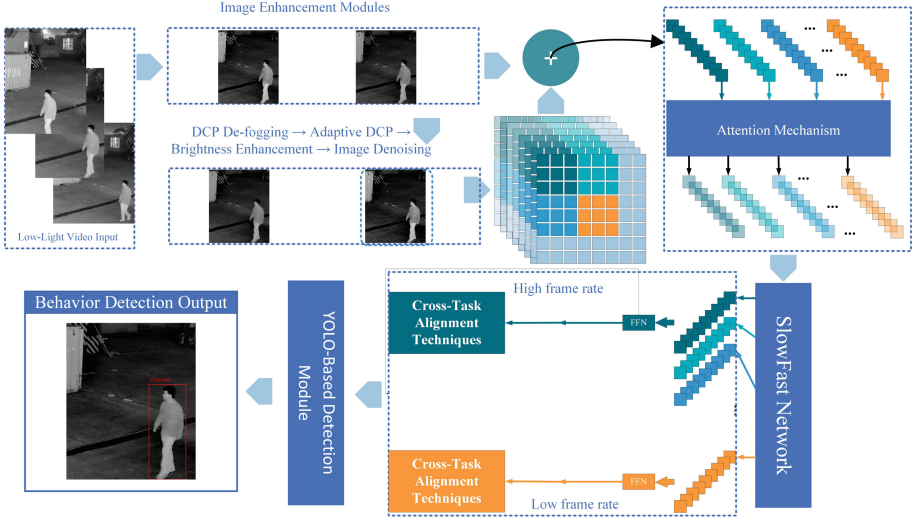


Fig. 1. Network Architecture for Pedestrian Behavior Detection in Nighttime Port Surveillance Video.

Adam optimizer with a learning rate of 0.001 and a batch size of 16. Data augmentation techniques—including random cropping and horizontal flipping—were employed to improve the model’s robustness against diverse environmental conditions.

All experiments were conducted on a workstation equipped with an AMD Ryzen 9 5900HX CPU, 16GB RAM, and an NVIDIA GeForce RTX 3070 GPU. The operating system was Windows 10, and the deep learning framework used was PyTorch.

3.2 Experimental Results

To evaluate the effectiveness of the proposed method, we compared it with the traditional SlowFast network under identical training conditions and hardware configurations. Key performance metrics including precision, recall, and F1-score were used to assess the detection performance, which are critical for pedestrian detection systems, especially in low-visibility conditions.

The proposed method demonstrated significant performance gains over the traditional SlowFast network. Specifically, it achieved a higher F1-score of 0.87 compared to 0.76 for the traditional method. The integration of dark channel enhancement methods, such as the dark channel prior and adaptive transmission refinement, effectively mitigated the challenges associated with low illumination. This resulted in an improved recall rate of 0.90, which is crucial for ensuring reliable pedestrian detection in low-visibility conditions.

The results in Table 2 illustrate that the proposed method maintains high detection performance across various challenging environmental conditions, further highlighting its robustness and effectiveness.

Table 1. Performance Comparison between Proposed Method and Traditional SlowFast Network

Method	Precision	Recall	F1-score	Training Time (hours)	Inference Time per Frame (ms)
Proposed Method	0.85	0.9	0.87	12	42
Traditional SlowFast	0.75	0.78	0.76	10	40

Table 2. Performance under Different Environmental Conditions

Environmental Condition	Precision	Recall	F1-score
Foggy	0.83	0.88	0.85
Rainy	0.82	0.87	0.84
Low Illumination	0.85	0.90	0.87

3.3 Computational Efficiency



Fig. 2. Sample detection results under different environmental conditions. The proposed method (left) demonstrates superior detection accuracy compared to the traditional SlowFast network (right), especially in scenarios with low illumination and high confusion.

Despite the additional computational steps introduced by the image enhancement techniques, the proposed method maintained a reasonable inference time per frame, making it suitable for real-time applications. The inference time per frame was approximately 42 ms, which is acceptable for surveillance systems requiring prompt detection

and response. The training time was comparable to the traditional SlowFast network, indicating that the enhancements did not introduce significant overhead.

3.4 Analysis of Results

The improved performance can be attributed to the combination of image enhancement techniques and attention mechanisms, which enhanced the visibility of pedestrians and suppressed background noise. The use of adaptive DCP and brightness enhancement using hyperbolic tangent curves contributed to better feature representation in low-light conditions. Additionally, the cross-task alignment techniques further refined the model's ability to detect pedestrians under diverse nighttime conditions.

To qualitatively assess the improvements, Fig. 2 shows sample detection results under low illumination conditions. The proposed method consistently detects pedestrians more accurately than the traditional SlowFast network, particularly in areas with low illumination and high confusion.

4 Conclusions and Future Work

This research presents a robust and adaptive framework for nighttime pedestrian behavior detection specifically tailored to port environments. By integrating advanced image enhancement techniques—including the dark channel prior and adaptive transmission refinement—into the preprocessing pipeline and leveraging the capabilities of the SlowFast network, the proposed framework significantly improves visibility and detection accuracy in low-light conditions. These enhancements resulted in notable performance gains in the deep learning-based pedestrian detection model, particularly in scenarios with challenging visibility, achieving an F1-score of **0.87** compared to **0.76** from traditional methods.

The experimental results underscore the effectiveness of this approach in enhancing safety at ports by providing reliable and accurate detection capabilities during nighttime operations. The framework not only addresses the limitations of existing methods in low-illumination environments but also demonstrates adaptability to various adverse weather conditions such as fog and rain.

Future work will focus on real-time implementation, further integration of thermal imaging to address extreme weather conditions, and the optimization of computational efficiency to enable practical deployment in real-world port environments. Additionally, exploring supplementary modalities such as LiDAR and radar will be pivotal to improving detection accuracy in increasingly complex port settings.

Acknowledgments. This work was jointly supported by National Natural Science Foundation of China (52331012, 52102397, 52472347), Shanghai Science and Program of Shanghai Academic/Technology Research Leader (23XD1431000).

References

1. Chen, Y.Y., Jhong, S.Y., Li, G.Y., Chen, P.H.: Thermal-based pedestrian detection using faster r-cnn and region decomposition branch. In: International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp. 1–2 (2019)
2. Kang, J., Hong, H.G., Park, K.: Pedestrian detection based on adaptive selection of visible light or far-infrared light camera image. *Sensors* **17**(7), 1598 (2017)
3. Li, G., Zhang, S., Yang, J.: Nighttime pedestrian detection based on feature attention and transformation. In: 2020 25th International Conference on Pattern Recognition (ICPR), 9180–9187 (2021)
4. Ge, J., Luo, Y., Tei, G.: Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Trans. Intell. Transp. Syst.* **10**(2), 283–298 (2009)
5. Abirami, G., Padmavathi, S.: Differential illumination enhancement technique for a nighttime video. In: Dash, S., Bhaskar, M., Panigrahi, B., Das, S. (eds.) *Artificial Intelligence and Evolutionary Computations in Engineering Systems. AAISC*, vol. 394. Springer, New Delhi (2016). https://doi.org/10.1007/978-81-322-2656-7_11
6. Shi, J., Zhang, G., Yuan, J., Zhang, Y.: Improved YOLOv3 Infrared image pedestrian detection algorithm. In: Zeng, J., Jing, W., Song, X., Lu, Z. (eds.) *Data Science. ICPCSEE 2020. CCIS*, vol. 1257. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-7981-3_37
7. Wang, X., et al.: When pedestrian detection meets nighttime surveillance: a new benchmark. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 509–515 (2020)
8. Tian, L., Chin, C.T., Cheng, C., Huang, R.P.: Night pedestrian detection using thermal image feature extraction enhanced YOLOv3 (TIFEEY). In: International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), pp. 80–89 (2022)
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6201–6210 (2018)
10. Liu, X., Lin, Y.: YOLO-GW: quickly and accurately detecting pedestrians in a foggy traffic environment. *Sensors* **23**(12), 5539 (2023). <https://doi.org/10.3390/s23125539>
11. Wang, Y.-Z., Lu, T., Zhang, Y., Fang, W., Wu, Y., Wang, Z.: Cross-task feature alignment for seeing pedestrians in the dark. *Neurocomputing* **462**, 282–293 (2021). <https://doi.org/10.1016/j.neucom.2021.07.096>
12. Wang, W., Peng, Y., Cao, G., Guo, X., Kwok, N.: Low-illumination image enhancement for night-time UAV pedestrian detection. *IEEE Trans. Industr. Inf.* **17**, 5208–5217 (2020)
13. Cui, C., Xie, J., Yang, Y.: Bright Channel Prior Attention for Multispectral Pedestrian Detection. *ArXiv* (2023)



Research on the Application of Privacy Computing Technology in the Scene of Railway Data Rights Confirmation and Authorization

Peiran Wang^(✉), Xiaoning Ma, Dan Zou, Siqi Sun, Jiang Wu, and Kaopeng Li

China Academy of Railway Sciences Corporation Limited, Beijing 100044, China
wpr_cars@163.com

Abstract. The “Twenty Articles on Data” proposes to establish a property rights operation mechanism for the equal distribution of data resource ownership, data processing and use rights, and data product operation rights, providing new ideas for the exploration of China’s data property rights system. The exploration of data property rights separation in enterprise data should be based on the protection of individual and organizational privacy, and safeguard their rights and interests related to data decision-making. By analyzing the relevant rights and interests of data resource ownership, data processing and usage rights, and data product management rights, exploring the operation path of data three rights separation, proposing a data authorization mechanism based on railway data sharing scenarios, and clarifying the responsibilities and obligations of all parties involved in railway data circulation. Using privacy computing technology to propose new solutions for railway authorization scenarios, thereby promoting railway data sharing and sharing.

Keywords: Railway data enterprise data property rights · data resource ownership rights · data processing and use rights · data product operation rights · privacy computing

1 Introduction

In December 2022, the Central Committee of the Communist Party of China and the State Council issued the “Opinions on Building a Data Infrastructure System to Better Play the Role of Data Elements” (hereinafter referred to as the “20 Articles on Data”), which systematically laid out the “four beams and eight pillars” of China’s data infrastructure system from four dimensions: digital property rights, circulation and transactions, data distribution, and data governance, which has milestone significance [1]. The “Article 20 on Data” proposes: “Based on the sources and characteristics of data generation, define the legitimate rights enjoyed by each participant in the process of data production, circulation, and use, and establish a property rights operation mechanism for the division of data resource ownership, data processing and use rights, and data product management rights [2].

It is generally believed that enterprise data can be divided into two categories: one is data collected from various aspects of its own production and operation, and the other is data obtained through purchase, authorization, and delegation [3]. As a transportation enterprise with railway passenger and freight transportation as its main business and diversified operations, China Railway collects a large amount of data through information technology and digital infrastructure in the production and operation process of engineering construction, transportation production, business development, resource management, comprehensive coordination, strategic decision-making, and other fields [4]. On the other hand, the enterprise also obtains external data such as weather, earthquake monitoring, national government information, etc. through purchasing, authorization, and commissioning. What property rights does the enterprise have for the above data and how to manage relevant data authorization are crucial for railway data sharing, utilization, and data circulation.

The passenger and freight business system of China Railway has also collected a lot of passenger information, shipper information, and other personal and organizational related data, involving personal privacy, organizational secrets, and other privacy data. How to protect and authorize rights for enterprises is also a key research direction.

This article will explore the operation path of data three rights separation by analyzing the relevant rights and interests of data resource ownership, data processing and use rights, and data product operation rights. It will propose a data authorization mechanism based on railway data sharing scenarios, and clarify the responsibilities and obligations of all parties involved in railway data circulation. By analyzing the impact of privacy computing on data ownership authorization, a new approach for railway authorization is proposed.

2 Data Ownership

Ownership usually manifests as the right to possess, use, benefit from, and dispose of property, with exclusive possession being the fundamental right of property rights [5]. Data elements have replicability, non competitiveness, and exclusivity, so data property rights differ from traditional production factor property rights.

There are usually two types of views on the ownership of data in the legal field. The first type is the theory of personality rights, which holds that personal data, as a carrier of personal information, should be owned by the user personally. However, this viewpoint only emphasizes the protection of personal interests and is not conducive to data circulation and transactions. The second type is the labor empowerment theory, which believes that data should be owned by enterprises due to the labor they have put in during the collection, processing, analysis, and other processes. However, this viewpoint ignores the protection of personal privacy.

In order to protect personal privacy, the country has successively introduced multiple data related regulations. Article 8 of the Data Security Law of the People's Republic of China states: "Data processing activities shall not endanger national security, public interests, or harm the legitimate rights and interests of individuals and organizations. Article 44 of the Personal Information Protection Law of the People's Republic of China states:" Individuals have the right to know and decide on the processing of their

personal information, and have the right to restrict or refuse others from processing their personal information. Currently, individuals and businesses protect their relevant rights and interests by signing a Privacy Agreement. Companies need to clearly inform individuals of the scope of their data usage, and individuals can license and authorize their personal information in conjunction with the services provided by the company. During this process, the enterprise obtained the right to process and use personal information related data.

The ownership of data is not an either or relationship between individuals and enterprises. Enterprises should classify and manage the data resources they hold. For the use of personal privacy data, it is necessary to accurately inform the relevant rights and interests parties and obtain their permission, in accordance with relevant laws such as the Personal Information Protection Law of the People's Republic of China, and use it in accordance with laws and regulations. Article 5 of the "Data Article 20" proposes that "market entities have the right to lawfully hold, use, and obtain profits from data collected and processed by various market entities in their production and operation activities that do not involve personal information and public interests." For data that does not involve personal information and public interests, enterprises have the ownership of the data and can authorize its use in combination with application scenarios.

3 Separation of Data Rights

The "Twenty Articles on Data" innovatively proposes to establish a property rights operation mechanism for the equal distribution of data resource ownership, data processing and use rights, and data product operation rights, and encourages enterprises to explore data authorization and use models. In the process of data rights verification and authorization, enterprises need to accurately distinguish between three types of rights and responsibilities.

3.1 Ownership of Data Resources

The ownership of data resources is the de facto control over the data, which can be conditionally shared with other entities within the scope permitted by the data owner. Without relevant equity restrictions, the holder of data resources can actually utilize the data resources, thereby losing control over the right to process and use data and the right to operate data products [6]. Therefore, it is necessary for data resource holders to strengthen the protection of data resource ownership rights and interests through technical and institutional management of data.

For enterprises, internal control over data resource ownership can be strengthened through institutional management and responsibility authorization. When granting external authorization, priority should be given to granting other rights and interests, and cautious authorization should be exercised.

Data resource holders should strengthen self-protection of data resources, adopt encryption, backup, recovery and other technical means and permission control to protect their control over data resources. For private data, it is necessary to increase the level of data security protection, and for the transfer of related rights and interests, authorization from individuals and organizations is required.

3.2 Right to Use Data Processing

The right to use data processing refers to the right to process and utilize data resources. Data processing users provide customers with higher quality services through the processing and utilization of data resources. The holder of data resources can transfer the right to use data processing, and can agree on the time limit, scope, and transfer conditions of the right to use data processing through relevant agreements. For private data, data processing users need to promptly notify individuals and organizations of the functions, scenarios, and other information related to the use of private data.

3.3 Right to Operate Data Products

The right to operate data products refers to the right to trade specific data products [7]. The scope of specific data products shall be agreed upon and confirmed by the data resource holder and the data product operator. It is not recommended to use raw data as a data product, otherwise the ownership of data resources may be easily compromised. Regarding private data, considering the difficulty of authorizing relevant rights through individuals or organizations, it is recommended to undergo desensitization and convert it into non private data before conducting transactions.

4 Exploration of the Separation of Three Rights in Railway Data

At present, the railway has been carrying out data aggregation and sharing for many years, and unclear data ownership is one of the important factors that hinder the smooth operation of the data chain. In 2023, the “Digital Railway Plan” will be released, emphasizing the exploration of the “separation of three rights” data property operation mechanism. Combined with the “Railway Data Management Measures (Draft for Comments)”, this article will explore the operation path of the “separation of three rights” of railway data, including the following three paths.

4.1 Divide and Conquer

Divide and conquer means that the rights to hold data resources, process and use data, and operate data products belong to different entities. The entity that holds the right to hold data resources actually holds the data resources, but cannot transfer the data resources to other entities without the permission of the authorized party. The common target of this entity is the data platform operator. The entity with the right to use data processing can conditionally obtain the data required for the production of data products (or services) and process and use the data. The common target of this entity is business system research and development units. Having the main body of data product operators, they can operate data products externally and provide feedback on new demands to data processors based on market demand. In this path, it is recommended that the ownership of data resources and the right to use data processing be undertaken by controllable entities of the enterprise to protect the controllability of data resources.

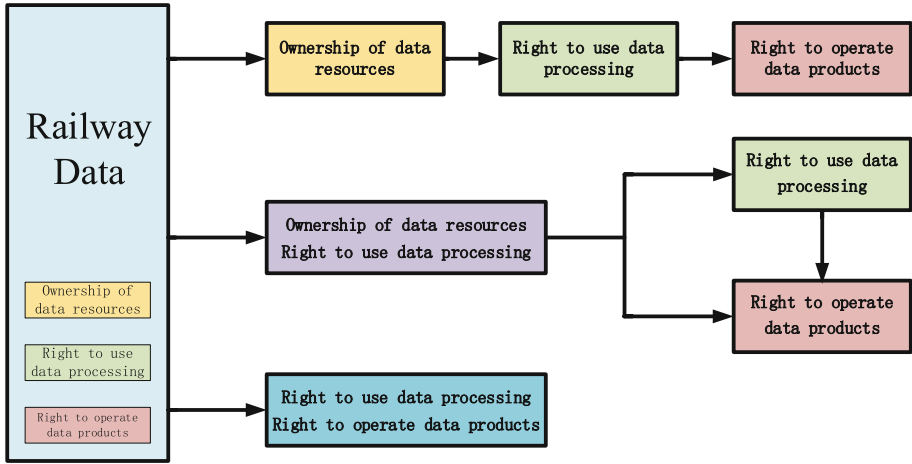


Fig. 1. Railway data three rights separation path

4.2 Having the Right to Hold Data Resources and the Right to Use Data Processing

The entity that holds the ownership and processing rights of data resources can process and extract the value of raw data resources. These data products can only improve the operational efficiency of enterprises and provide services to customers [8]. It is also possible to perform initial processing on the data by desensitizing and encrypting the data resources to ensure reliable data out of the domain, thereby providing reliable, stable, and controllable data resources to the outside world. But related data products cannot be traded with other parties.

4.3 Having the Right to Process and Use Data and Operate Data Products

The entity with the right to use data processing and the right to operate data products can process the obtained data resources into data products and conduct business transactions on the data products. When a company grants both of these rights to external entities, it is necessary to preprocess data resources, select reasonable technical solutions to provide data resources, and ensure the ownership rights of data resources.

5 Data Authorization Mechanism in Railway Data Sharing Scenarios

The railway data sharing scenario mainly relies on the railway data service platform for data sharing, but due to the business requirements of some specific scenarios, some data also supports direct sharing from the production system to the data demander. In the scenario of railway data sharing, there are mainly several roles, including business system research and development unit, data platform operator, railway bureau group company, and data applicant. As the main body of research and operation of the railway

production system, the business research and development unit actually stores and manages production data. China Railway can grant it the right to hold data resources and the right to process and use data. The data platform operator mainly provides services for railway data aggregation and sharing, and can be granted the right to hold data resources and the right to process data [9]. The platform operator can encrypt, desensitize, and process data according to the requirements of the data owner to ensure data security and compliant sharing. As a regional production entity, the Railway Bureau Group has complex usage scenarios for some data. China Railway may consider granting data resource ownership and data processing usage rights based on sharing conditions. For data applicants, generally only the right to use data processing is granted. The right to operate data products may be considered for establishing a dedicated data group in the future to provide unified data management services.

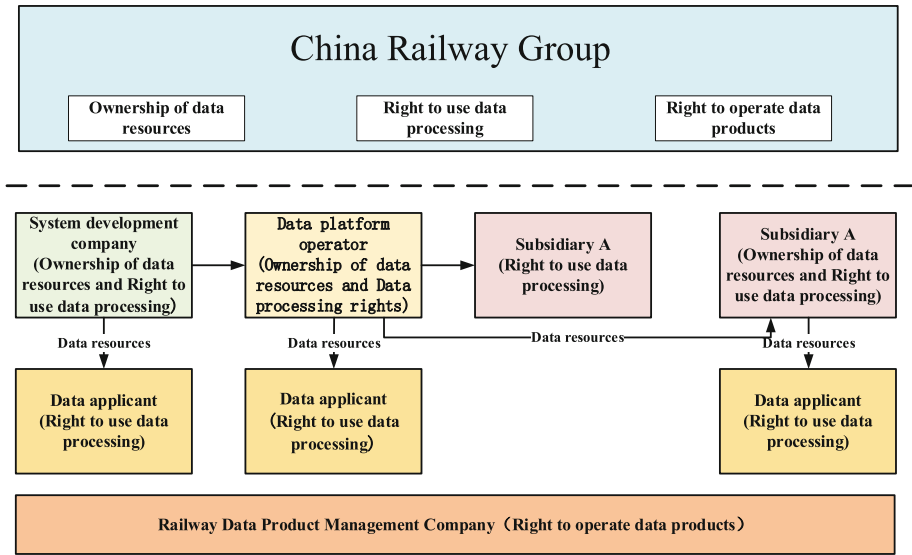


Fig. 2. Authorization for railway data sharing scenarios

6 Application of Privacy Computing Technology in Railway Data Authorization Scenarios

In the context of railway data authorization, the data authorizer is most concerned about the leakage of personal privacy data. When granting third-party data resource ownership and data processing usage rights, due to the replicability of data, it is highly likely to lead to uncontrollable personal privacy data. Privacy computing technology provides a feasible solution to this problem.

Privacy computing refers to a series of information technologies that analyze and calculate data while ensuring that the data provider does not disclose the original data

[10]. Privacy computing has characteristics such as “data is available but not visible, data remains unchanged and models move”, “data is available but not visible, data is controllable and measurable”, and “data value is shared instead of data sharing”. The main technologies for privacy computing currently include multi-party secure computing, federated learning, trusted execution environment, and multi-party intermediary computing [11].

The application of privacy computing has a transformative impact on the “three rights of data”. Firstly, privacy computing technology separates the ownership and usage rights of data. Under the technical framework of privacy computing, data holders can authorize the use of data to other parties for computation and analysis without disclosing the original data. This separation greatly ensures the stability of data holding rights. Secondly, with the support of privacy computing technology, data security is also fully protected. Due to strict protection of raw data, the risk of data leakage and abuse is avoided.

7 Conclusion

The data property rights operation mechanism of “separation of ownership of data resources, processing and use of data, and operation of data products” is conducive to promoting the structural separation and orderly circulation of enterprise data property rights. Enterprises can combine the data classification and grading protection system to promote data classification, grading, certification, authorization, use, and market-oriented circulation transactions. This article analyzes the relevant rights and interests of the three rights of data, explores the operation path of the separation of the three rights, proposes a data authorization mechanism based on the railway data sharing scenario, and explores the application of privacy computing technology in the railway authorization scenario, providing ideas for the confirmation and authorization of railway data rights and accelerating the smooth circulation of railway data.

Acknowledgments. This research was partially funded by the China Academy of Railway Sciences Corporation Limited Major subject (2023YJ151).

References

1. Shen, W.: Data property rights: from separation of two rights to separation of three rights *China Legal Rev.* (06), 125–137 (2023). (in Chinese)
2. Wang, N.: The legal function and normative construction of data resource ownership rights *J. Shanghai Univ. Polit. Sci. Law (Collect. Legal Stud.)* (03), 137–152 (2024). <https://doi.org/10.19916/j.cnki.cn31-2011/d.2024.03.007>. (in Chinese)
3. Fan, Z., Zhou, Z., Zhang, W.: Game analysis of enterprise data sharing from a supply chain perspective. *Heliyon* (4), e25678 (2024)
4. Cheng, J.: Research on data management for china railway based on computer science. *Appli. Mech. Mater.* (631–632), 1304–1307 (2014)
5. Wang, X.: Research on the definition and construction of data ownership. *Inter. J. Front. Sociol.* (4) (2024)

6. Zhang, S., Wang, N.: The proof and construction of the “dual order binary structure” of data property rights. *China Legal Rev.* (06), 138–156 (2023)
7. Huang, L., Du, W., Wu, Y.: Structural separation of data property rights based on the value chain of data element circulation. *Big Data* (02), 5–15 (2023)
8. Wang, K., Peng, B., Jiang, Z.: The value creation process of manufacturing service-oriented driven by new information technology: an exploratory case study based on XCMG. *Manag. Comment* (11), 275–285 (2021). <https://doi.org/10.14120/j.cnki.cn11-5057/f.2021.022>
9. Wu, Z.: Analysis of Multiple Rights and Responsibilities in Public Data Digital Economy, vol. (04), 12–22 (2024). <https://doi.org/10.19609/j.cnki.cn10-1255/f.2024.04.008>
10. Hu, F., Chen, T.: An efficiency optimization study of data governance legal issues in the framework of privacy computing. *Appli. Math. Nonlinear Sci.* (1) (2024)
11. Zhu, Z., Zhang, Z.: An analysis of privacy computing applications in the financial field. *Financial Developm. Res.* (03), 90–92 (2023). <https://doi.org/10.19647/j.cnki.37-1462/f.2023.013>



A Fair Traffic Management Approach for Electric and Conventional Vehicles in Urban Areas

Liang Zhang¹(✉), Huijuan Zhou², and Xiaoming Liu²

¹ Drexel University, Philadelphia, PA 19104, USA
lz465@drexel.edu

² North China University of Technology, Beijing 100144, China
{zhouhuijuan, liuxiaoming}@ncut.edu.cn

Abstract. With the increasing adoption of electric vehicles (EVs) on urban roads, traditional traffic management approaches that treat all vehicles equally are no longer sufficient to address the diverse impacts of EVs and conventional vehicles (CVs). This paper proposes a fair management approach addressing the distinct characteristics of EVs and CVs, focusing on optimizing traffic signal control to promote fair road usage and reduce delays. We introduce a fairness function that considers vehicle-specific attributes, including waiting times and vehicle credits, to allocate traffic signal resources in a way that supports both environmental goals and balanced access for all vehicle types. The effectiveness of this approach is demonstrated through two key analyses: delay heatmaps for multiple intersections before and after implementing the fair management strategy, and an evaluation of the impact of varying EV penetration rates on balanced and unbalanced traffic conditions. The results show that the proposed fairness function not only decreases delays but also improves overall traffic flow, particularly in scenarios with higher EV penetration. This study contributes to the development of sustainable urban traffic systems by promoting a fair transition towards greener and more efficient transportation systems.

Keywords: Electric Vehicles · Fairness Function · Traffic Signal Control

1 Introduction

In recent years, electric vehicles (EVs) have become an increasingly prominent presence on roads, driven by advancements in technology, government incentives, and a growing emphasis on reducing carbon emissions. While conventional vehicles (CVs) still dominate the market, EVs are expanding their share as countries worldwide aim to meet sustainability goals. Governments and manufacturers have invested heavily in EV infrastructure, such as charging stations, battery technology, and purchasing discounts, making EVs more accessible and practical for everyday use. As cities increasingly support EV adoption, ensuring fairness in transportation infrastructure and road usage is critical.

Ensuring fairness requires a balanced approach that fosters EV adoption while acknowledging that many people still depend on CVs. By designing policies and infrastructure that account for both vehicle types, cities can support a fair transition towards sustainable transportation. In traffic signal control systems, signal timings and road usages can be treated as resources to be allocated to different vehicle types and traffic conditions. How to address these issues has become the focus among researchers [1–3].

Traditionally, traffic signal control at intersections has utilized fixed or adaptive timing systems that are designed to be fair by treating all vehicles uniformly, without discrimination. This approach aligns with the notion of fairness when vehicles are treated anonymously and equally. However, if we consider fairness from the perspective of vehicle types, treating all vehicles the same may no longer be equitable. Vehicles vary in their levels, so a more tailored approach becomes essential. Two existing concepts offer relevant insights. The first involves queue management methods, such as those discussed in [4], where traffic control systems aim to balance queue lengths across multiple approaches to maintain fair waiting times. By using queue length as input, these systems strive to equalize congestion across all traffic streams. Reinforcement learning methods introduced in [5] and [6] expand on this concept by enhancing queue balance and generating optimal signal timing plans that adapt to changing traffic conditions.

A second related concept considers the heterogeneity among individual vehicles. This variability includes factors such as the Value of Time (VOT) and each vehicle's willingness to pay to reduce delay. For example, in connected vehicle networks, vehicles could “bid” for priority at intersections based on these attributes [7]. This approach highlights how accounting for individual differences among vehicles could create a more adaptable and responsive traffic management system. Building on this, traffic signal control strategies based on individual vehicle emissions and energy consumption have been further proposed [8, 9]. While these studies provide insights into the emissions profiles and energy consumption of different vehicle types, they do not directly address the unique characteristics of EVs and CVs. Therefore, our paper proposes an approach to enhance fairness in traffic signal control and road usage for two distinct vehicle types: electric vehicles (EVs) and conventional vehicles (CVs) within urban areas. Recognizing that EVs and CVs contribute differently based on their credits, we aim to address this disparity by developing a system that optimizes signal control to balance traffic flow while considering vehicle types. By incorporating waiting time and vehicle credits into traffic signal algorithms, our approach seeks to promote fair road access and usage, ensuring that urban traffic management systems support both environmental goals and fair treatment of all vehicle types.

2 Methodology

2.1 A Fairness Function

We first present the fairness function and then present a real-world scenario explaining how it applies to urban areas.

In a typical four-approach intersection (see Fig. 1), we assume that EVs and CVs arrive at specific rates. In this setup, EVs are marked in red, while CVs are marked in yellow. For simplicity, the intersection operates in four phases: the west and east

approaches are in red signal phase, while the north and south approaches are in green. Through advanced communication hardware, it is possible to gather comprehensive information about each vehicle, such as waiting time, speed, and location. This signalized intersection can be viewed as a system, with EVs and CVs as its components. Within this system, resources are allocated to these components, such as signal timing plans and the value of time (VOT). Each component, in turn, has unique properties that influence how these resources are distributed. These properties may include the priority assigned to each component type or distinct attributes that differentiate EVs from CVs.

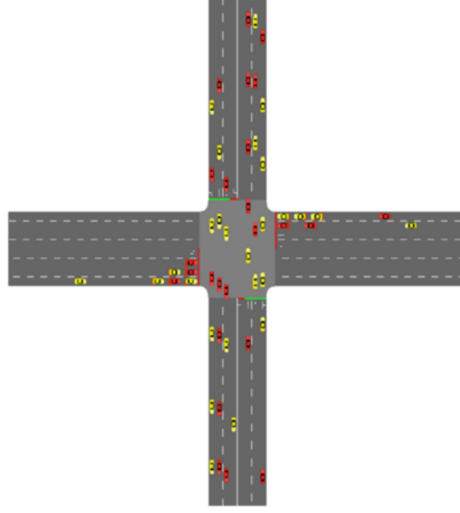


Fig. 1. A Typical 4-approach Intersection with EVs and CVs

We first present our fairness function as follows:

$$f(r, a) = \frac{\left(\frac{r_1}{a_1} + \frac{r_2}{a_2}\right)^2}{\left(\frac{r_1}{a_1}\right)^2 + \left(\frac{r_2}{a_2}\right)^2} - 1 \quad (1)$$

where r_1, r_2 represents the resources to be allocated to EVs and CVs, respectively. a_1, a_2 represents the attributes to be assigned to EVs and CVs, respectively. Note r_1, r_2 are normalized resources and $r_1 + r_2 = 1$. $f(r, a) \in [0, 1]$, where 0 indicates totally unfair and 1 indicates perfectly fair. One can find maximum fair values and decision variable values given r_1, r_2 or a_1, a_2 .

To illustrate the application of this approach in a traffic scenario, let's consider a numerical example. EVs are increasingly promoted in the market, often receiving incentives such as purchase discounts and reduced taxes. However, in urban traffic management—particularly in signal control or delay mitigation—EVs and CVs are generally treated equally, meaning drivers do not necessarily experience reduced travel times by choosing an EV. Establishing an incentive system for EVs within urban traffic flow could encourage more sustainable travel by offering favorable treatment for EVs. This fairness function makes such a system feasible.

As shown in Fig. 1, suppose we have three CVs and three EVs waiting on the west approach at an intersection. To favor EVs in this scenario, we could allocate certain attributes, such as priority credits, to each vehicle type, with a higher credit value assigned to EVs. This credit-based approach allows for a flexible, quantified means of prioritizing EVs, effectively creating a practical incentive within the urban traffic system. The waiting time for EVs and CVs are now treated as resources in order to assign credits accordingly. Now the equation becomes:

$$f(w, c) = \frac{(\frac{w_1}{c_1} + \frac{w_2}{c_2})^2}{(\frac{w_1}{c_1})^2 + (\frac{w_2}{c_2})^2} - 1 \quad (2)$$

where w_1, w_2 represents the average waiting time of EVs and CVs at the end of red time, respectively. c_1, c_2 represent the credits to be assigned to EVs and CVs, respectively. For example, suppose $w_1 = 30s, w_2 = 15s$, one can find the maximum f values = 1 when $c_1 = 2c_2$. One can disregard this assignment if $w_1 < w_2$, instead, equal credits are assigned to both EVs and CVs.

2.2 Traffic Signal Control

After we get the credits assigned to the EVs and CVs, the total credits of vehicles can be collected by approach and cycles. The fairness function can be further applied to traffic signal control. Here we adopted the measurement of credits and fairness functions proposed in our previous work [8]:

$$f(G_{max}) = \frac{(\frac{G}{c_g} + \frac{G_{max}}{c_r})^2}{(\frac{G}{c_g})^2 + (\frac{G_{max}}{c_r})^2} - 1 \quad (3)$$

One can find G_{max} by maximize f given other parameters, G_{max} is a constraint for the traffic signal plans for the incoming green phase in order to achieve the maximum fairness of the two approaches on a real-time basis. Then the algorithm is given in Table 1 below (Table 2).

3 Simulation Experiments

We collected the road configuration and traffic volumes at the intersections of the City of Tucson, Arizona. The existing traffic signal control systems in Tucson's intersections often rely on traditional models that require extensive calibration and fine-tuning to adapt to local conditions. The dataset of traffic volume is obtained from Pima Association of Government official website for the public access of transportation data [10]. The traffic signal timing data and road configuration is obtained by the author from field survey. Nine intersections are selected for model testing. The layout and the basic information of each intersection of the area is shown in Fig. 2.

The simulation was established in SUMO. We tested the average delays for each approach among the area and the delay fairness index [11] among different EV market

Table 1. Applied Traffic Signal Control Algorithm

Traffic Control Algorithm Considering EVs and CVs
1. Initialize time $t = 0$
2. While $t < T$ do:
3. Step 1: Determine the favored approach
4. Collect the credits for each approach
5. Identify the highest credits for each vehicle type
6. Identify the highest credits for each approach
7. Calculate the favored ratio $s = Cr/Cg$
8. If $s > \text{threshold 1}$:
9. Assign Cr as favored approach
10. Else if $s < \text{threshold 2}$:
11. Assign Cg as favored approach
12. Else:
13. Assign equal for Cr and Cg
14. End if
15. Step 2: Traffic Signal Control
16. If the favored approach phase is not green or $t > G_{max}$
17. Activate the yellow phase for Y seconds on the current green phase
18. Record the green time G for the current green phase
19. Switch the favored approach to the green phase for a minimum duration of G_{min}
20. Adjust G_{max} using fairness function f
21. Else:
22. Extend the current green phase by an additional 5 s
23. End if
24. Increment the time t for next cycle
25. End while

penetration levels with respect to different traffic volumes. The results are showing below (Fig. 3):

The figure presents a series of delay heatmaps across nine intersections, comparing the states before and after implementing fair management. Each row represents a unique intersection, with two heatmaps side-by-side for each intersection: the left showing the delay distribution before applying fair management and the right displaying the distribution after. In many intersections, areas with intense delays (highlighted in red) before fair management are either reduced or eliminated after the management. For instance, in Intersection 3, high delays in specific regions are diminished in the “After” heatmap, suggesting a more balanced flow. The heatmaps after fair management generally exhibit

Table 2. Fairness Measurement for Delay of EVs in Different Traffic Conditions

Penetration Rate	Balanced low	Balanced medium	Balanced high	Unbalanced low	Unbalanced medium	Unbalanced heavy
Limited (5%)	0.33	0.32	0.38	0.33	0.24	0.16
Low (20%)	0.54	0.37	0.56	0.46	0.25	0.22
Medium(50%)	0.59	0.57	0.58	0.44	0.36	0.23
High (80%)	0.58	0.56	0.66	0.47	0.44	0.42
Full (100%)	0.68	0.71	0.68	0.67	0.61	0.54

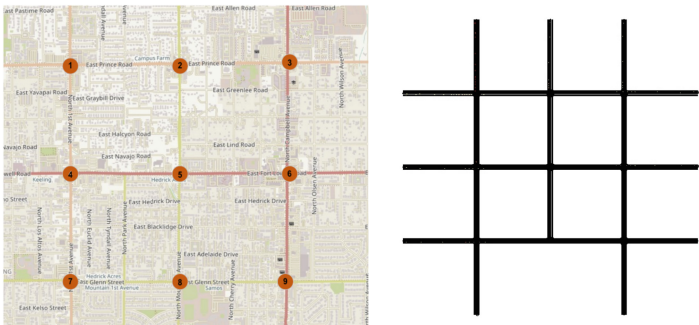


Fig. 2. Urban Area Road Configurations

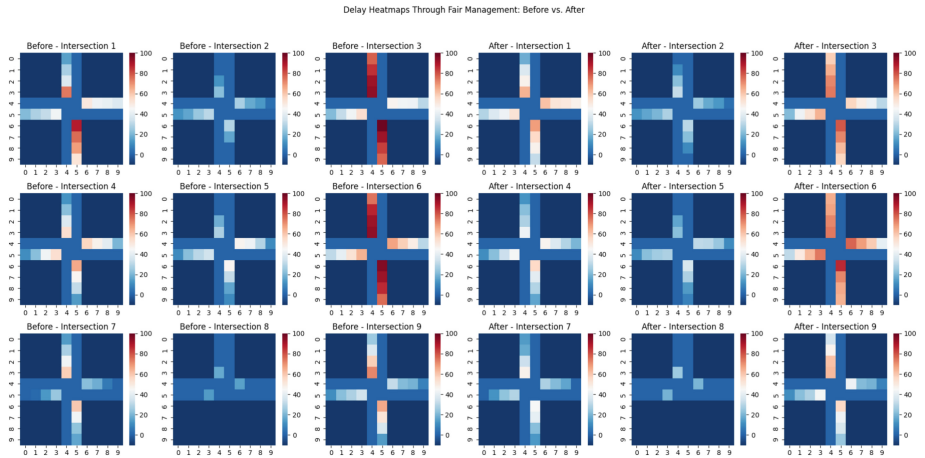


Fig. 3. Delay Heatmaps Through Fair Management: Before vs. After

a smoother, more evenly distributed pattern, indicating that delays have been mitigated and spread more uniformly. This suggests that fair management has helped balance delays across different approaches at each intersection.

The table suggests that increasing the penetration rate of EV type positively impacts traffic conditions, with the most substantial benefits occurring under balanced traffic conditions. High penetration rates lead to improved fairness, especially in scenarios with evenly distributed traffic. However, even in unbalanced conditions, a higher penetration rate contributes to performance gains, though the improvements are less pronounced than in balanced setups. This analysis supports the idea that achieving a high penetration rate is beneficial for overall traffic flow and fairness, particularly when traffic conditions are inherently balanced.

4 Conclusion

In conclusion, this paper proposes a fair management approach to address the distinct needs of electric vehicles (EVs) and conventional vehicles (CVs) within urban traffic systems. By introducing a fairness function that takes into account the differences in waiting times and credits assigned to each vehicle types between EVs and CVs, our approach aims to optimize traffic signal control to create a more equitable environment for both vehicle types. The effectiveness of this approach is demonstrated through two key figures: one showing delay heatmaps for multiple intersections before and after applying the fair management strategy, and the other highlighting the impact of different penetration rates on balanced and unbalanced traffic conditions.

References

1. Li, M., Wu, X., He, X., Yu, G., Wang, Y.: An eco-driving system for electric vehicles with signal control under V2X environment. *Trans. Res. Part C: Emerging Technol.* **93**, 335–350 (2018)
2. Majstorović, Ž, Tišljarić, L., Ivanjko, E., Carić, T.: Urban traffic signal control under mixed traffic flows: literature review. *Appl. Sci.* **13**(7), 4484 (2023)
3. Meng, X., Tang, S., Liu, X., Zhang, L.: Oversaturated traffic signal optimization based on active control. In: 2016 IEEE International Conference on Intelligent Transportation Engineering (ICITE), pp. 91–95. IEEE, Brazil (2016)
4. Lin, W.H., Lo, H.K., Xiao, L.: A quasi-dynamic robust control scheme for signalized intersections. *J. Intell. Trans. Syst.* **15**(4), 223–233 (2011)
5. Zhang, L., Lin, W.H.: Calibration-free traffic signal control method using machine learning approaches. In: 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), pp. 1–6. IEEE, Czech (2022)
6. Zhang, L.: Towards a self-organizing traffic control system at signalized intersections. Ph.D. dissertation, The University of Arizona (2023)
7. Carlino, D., Boyles, S.D., Stone, P.: Auction-based autonomous intersection management. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 529–534. IEEE, Netherlands (2013)
8. Zhang, L., Zhang, S., Tang, S., Liu, X., Zheng, G.: Towards sustainable urban intersections: An equitable vehicle emission-based traffic signal control system. In: 2023 China Automation Congress (CAC), pp. 6171–6176. IEEE, China (2023)

9. Zhang, L., Zhang, S., Liu, X., Chen, Z.: Sustainable traffic management: A framework utilizing energy consumption patterns. In: 2024 International Conference on Electrical. Computer and Energy Technologies (ICECET), pp. 1–6. IEEE, Australia (2024)
10. PAG Homepage. <https://pag.public.ms2soft.com/>
11. Jain, R., Duresi, A., Babic, G.: Throughput fairness index: Aa explanation. ATM Forum Contribution **99**(45) 1–13 (1999)



A Cooperative Control Model and Simulation Framework for Unmanned Aircraft Considering Multi-subject Communication

Zheng Zhang, Jingyuan Chen, Endian Wang, Ziqi Wang, Zihua Li, Gege Jiang,
and Yao Xiao^(✉) 

School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China
xiaoyao9@mail.sysu.edu.cn

Abstract. With the wide application of unmanned aerial vehicles (UAVs) in the fields of traffic monitoring and logistics distribution, the cooperative management of UAVs faces many challenges. In this study, a communication-based multi-object dynamics model for UAVs is proposed, and the existing RVO-3D model is improved to meet the cooperative obstacle avoidance requirements of multi unmanned aerial systems. By introducing the communication mechanism, the new model can effectively handle the dynamic interactions among multiple subjects, and improve the collision avoidance capability and cooperative efficiency of UAV clusters in complex 3D environments. In addition, a cross-platform simulation framework is developed to validate the performance of the proposed model and support the simulation and control of multi-UAS. The simulation results in simple and complex conflict scenarios show that the improved model exhibits excellent effectiveness and adaptability in multiple complex scenarios, which provides theoretical support and technical paths for the cooperative control of unmanned aerial systems.

Keyword: Multi-Agent Collaboration · Simulation Framework

1 Introduction

The rapid development of drone technology has led to an increasing number of applications in a variety of fields. Particularly in transportation and logistics, the ability of unmanned aerial vehicles (UAVs) to operate in groups is important for improving efficiency, reducing costs, and enhancing service accessibility. However, operation also brings a series of technical challenges, including coordinated control of swarm behavior, reliability of communication, and safety of operation [1].

Virtual simulation technology realizes the combination of reality and reality by creating a virtual model synchronized with the physical world, so that the UAV fleet can be accurately simulated and tested in a virtual environment. This approach not only improves the efficiency of UAV operation, but also predicts and evaluates potential risks and problems without affecting the actual operation, thus significantly improving the safety of the operation.

The core objective of this study is to develop a cross-platform integrated UAV virtual simulation framework and to explore multi-objective cooperative control strategies for effective management of UAVs in complex transportation systems. Our contribution lies in proposing an innovative multi-tool integration framework that enables efficient simulation of UAV swarm and provides a new cooperative control mechanism, in addition to conducting preliminary simulation tests on the framework to verify its effectiveness in different scenarios.

2 Status of Researches and Challenges

In recent years, the cooperative control of multi-UAV system has gradually become an important research direction in the field of automation and intelligent control. In a complex dynamic environment, it is important for the application of multi-UAV system to realize the efficient cooperation and obstacle avoidance ability of UAV group.

A mature model is Cellular Automaton (CA) [2]. CA is a grid dynamics model in which time, space and state are discrete, and spatial interaction and temporal causality are local. Duan et al. [3] focused on the concept of interaction between cellular neighborhoods of cellular automata, and proposed its combination with CASMA algorithm (concentration adaptive step size myxomycete algorithm). Fei et al. [4] reduced the complexity of cell states and the number of rules by using local configurations that can represent fluctuations between forward and backward movement of signals in the cell space. Although the CA model has significant advantages in simulating spatial patterns and temporal dynamics, its inherent rule definition limits its ability to fully express complex relational attributes. However, the interaction between cells is often not limited to simple spatial proximity. These attributes are essential for a comprehensive understanding and simulation of system behavior. Therefore, future research should focus on the further development and improvement of the CA model.

Taking into account the complex relationship between individuals, the social force model (SFM) has also been widely used. The social force model is another model that simulates crowd and traffic flow. Zhao et al. [5] first introduced the concept of the starting and ending speed ratio of the normalized expectation, and innovatively analyzed the influence of the continuously changing expected speed of the exercise individual on the exercise efficiency. Hou et al. [6] studied and established a micro-simulation model, which combined the social force model with the K-means clustering algorithm to improve the universality of this algorithm for various rules. Qi et al. [7] integrated the field model into the social force model, and improved the algorithm by introducing dynamic inertia weight and random reverse learning strategy, thus establishing a multi-layer model. Yang et al. [8] used the combination of Ant Colony Optimization (ACO) and social force model to navigate motion individuals with complete information. However, with the increase of environmental complexity, the interaction mode between individuals may be highly nonlinear and dynamic.

The velocity obstacle model is a good solution to the problem of dynamic complex environment. Fiorini et al. [9] proposed the VO algorithm. Its core idea is to define a set of velocity sets that will cause the robot to collide with obstacles, called Velocity Obstacle (VO). The advantage of the VO algorithm is that it can simultaneously handle

path planning and speed planning problems, and can consider the dynamic characteristics of the robot. Liu et al. [10] developed a dynamic perception algorithm, which uses VO to search the dynamic path of motion to verify the safety of motion, and enhances the smoothness and safety of the robot trajectory through nonlinear optimization. Gao et al. [11] solved the problem of path jitter caused by the speed obstacle method when the UAV flies from the starting point to the target point in the speed space by improving the reciprocal speed obstacle. Wakabayashi et al. [12] proposed a chance constraint based on the obstacle velocity (CCOV) method, which can be combined with the model predictive control (MPC) method to explain the uncertainty of position and velocity, thus effectively preventing collisions with high-speed obstacles. By solving the ORCA boundary, van den Berg et al. [13] transformed the conical region of the velocity obstacle into a half-plane, and optimized the calculation of the optimal velocity. The existing RVO2-3D model extends the speed domain to the three-dimensional space [14], which is of great significance to realize the efficient cooperation and obstacle avoidance ability of the UAV group.

With the development of UAV technology, especially the increasing demand for operation in three-dimensional space, the traditional two-dimensional model has been unable to meet the current research needs. The three-dimensional UAV simulation algorithm needs to consider more variables and complex spatial relationships, which makes the study of dynamic models more complex and challenging.

3 Collaborative Control Modeling and Simulation Implementation Architecture

3.1 Dynamical Model Improvement with Consideration of Communication

In UAV simulation research, dynamical models are key elements in simulating the flight behavior of UAVs, which directly affect the accuracy and reliability of simulation results. These models need to accurately capture the physical characteristics and flight dynamics of UAVs to ensure their performance in complex environments. In this study, we improve the existing RVO2-3D model and design a dynamics model that takes the communication mechanism into account.

Velocity obstacle (VO) is the core design of this model, assuming that there are two objects A and B in the scene (see Fig. 1(a)), with P_A as the mass point and $V_A - V_B$ as the velocity of A with respect to B. Among these velocities, the direction of the velocity that satisfies the intersection with the expanding B is the velocity obstacle VO_B^A of A with respect to B, i.e., the direction of the velocity in which A and B will collide. For A, the direction to avoid VO_B^A is the direction to avoid collision with B.

$$VO_B^A = \{V_A | \lambda(P_A, V_A - V_B) \cap B \oplus -A \neq \emptyset\} \quad (1)$$

However, the VO algorithm has the problem of jitter. The RVO algorithm improves the VO algorithm by assuming that the other party will also use the same strategy to avoid obstacles, and both parties will avoid obstacles together, which effectively avoids the jitter phenomenon. Under the RVO algorithm, the previous speed obstacles become

mutual speed obstacles, as shown in Eq. (2), and its improved RVO is the set of the average of the current speed and the speed belonging to the VO.

$$RVO_B^A(V_B, V_A) = \{V'_A \mid 2V'_A - V_A \in VO_B^A\} \quad (2)$$

On top of that, ORCA algorithm improves the RVO. The biggest difference between ORCA and RVO is the representation of VO. In the previous VO algorithm, VO is represented as two rays of the same origin in two dimensions, in the improvement of ORCA, the velocity is converted from the relative velocity to the same coordinate system, in order to compute the optimal velocity in the same coordinate system of the velocity, a time window τ is introduced to convert the displacement in τ to the velocity. The demonstration is shown in Fig. 2, and its mathematical interpretation is Eqs. (3)–(4), where $D(P_B - P_A, r_A + r_B)$ is the circle with $P_B - P_A$ as the center and $r_A + r_B$ as the radius. $VO_{A|B}^\tau$ means that there exists a time t , within the time period τ such that tv results in D , i.e., τ determines the intercept of the shading in Fig. 1(b). c is the set of collision avoidance considering the case of B motion.

$$VO_{A|B}^\tau = \{v \mid \exists t \in [0, \tau] : tv \in D(P_B - P_A, r_A + r_B)\} \quad (3)$$

$$CA_{A|B}^\tau(V_B) = \left\{ v \mid v \notin VO_{A|B}^\tau \oplus V_B \right\} \quad (4)$$

By solving the ORCA boundary to transform the conical region of the speed obstacle into a half-plane, the computation of solving the optimal speed is optimized at the same time, and becomes the problem of searching for the optimal direction after being divided by multiple half-planes in the space. Considering that there is a communication distance between UAVs, i.e., only UAVs within a certain horizon communicate with each other for obstacle avoidance, in the code implementation, in order to optimize the obstacle avoidance selection, the concept of time horizon T is proposed, which defines the time horizon of the agents that will be taken into account by the agents to consider the other agents in simulation, and it affects the forward-looking of the agents in predicting the future positions. That is, for the target obstacle B, it is necessary to calculate whether it is within the time horizon or not, as in Eqs. (5)–(6).

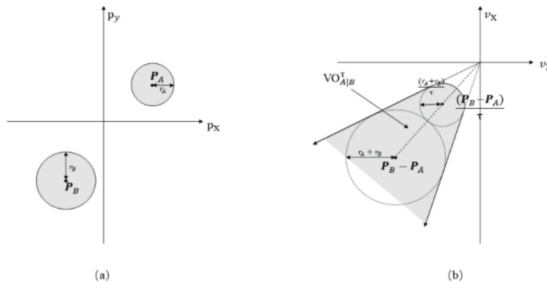


Fig. 1. (a) objects A and B in planar space; (b) the velocity that A cannot reach by avoiding collisions in velocity space during the time period τ

$$t_B = \frac{|\mathbf{P}_B - \mathbf{P}_A|}{V_A^{max}} \quad (5)$$

$$TVO_{A|B}^\tau = \begin{cases} \{v \exists t \in [0, \tau] :: tv \in D(\mathbf{P}_B - \mathbf{P}_A, r_A + r_B)\} & t_B \leq T \\ \emptyset & t_B > T \end{cases} \quad (6)$$

In this study, when applying the model, the temporal field of view is optimized so that the simulated UAV cluster can ignore its velocity obstacle plane when the obstacles are far away, increase the optional velocity direction of motion, and improve the motion flexibility. At the same time, the amount of computation in the simulation is reduced.

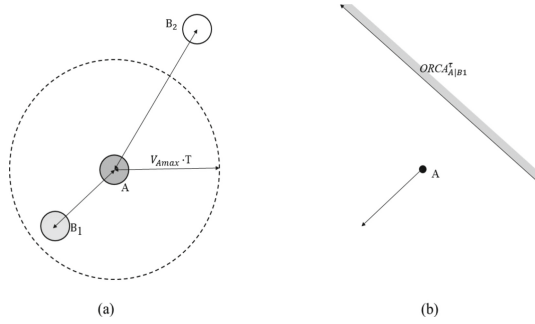


Fig. 2. (a) checking for obstructions in the time horizon; (b) calculating ORCA in the time horizon only

3.2 Framework Overview

In this study, an innovative multi-tool integrated UAV virtual simulation framework has designed, which aims to realize the high-precision simulation of the UAV group behaviors through the integration of the advanced simulation tools and flight control algorithms. The design of the framework takes into account the complex interactions among UAVs, especially in multi-UAV interaction scenarios, which is innovative and has potential for development.

The core of the framework design in this study is modularity and extensibility. The design interface organizes different simulation components as independent modules, and the modules transfer data to each other to realize the simulation together. This allows for easy upgrades and maintenance, as well as future expansion of new functionality and implementation of additional algorithms as needed. This modular design also facilitates the efficient integration of different tools and techniques, making the whole system more flexible and powerful. The specific architecture and communication flow is shown in Fig. 3.

In this framework, AirSim is chosen as the UAV simulation simulator after testing. As a UAV simulator based on real physics, AirSim provides us with a realistic flight

environment and supports a variety of UAV models and sensor data. It allows testing and verifying the flight behavior of UAVs in a virtual environment without the need of actual flight, and it can be quickly accessed to real UAVs for in-loop simulation after the simulation is completed [15]. In addition, AirSim provides a variety of interfaces for accessing a wide range of flight simulation data, allowing for improvements and the addition of rich modules on top of it. The AirSim simulator can be used as a plug-in for the Unreal engine to implement simulations, and can be placed into any Unreal project, so UE4 was chosen as the visual component of this framework. RVO2-3D has been widely used. However, it cannot communicate with AirSim directly, so this study adds an interface part with AirSim on the basis of the original RVO2-3D library.

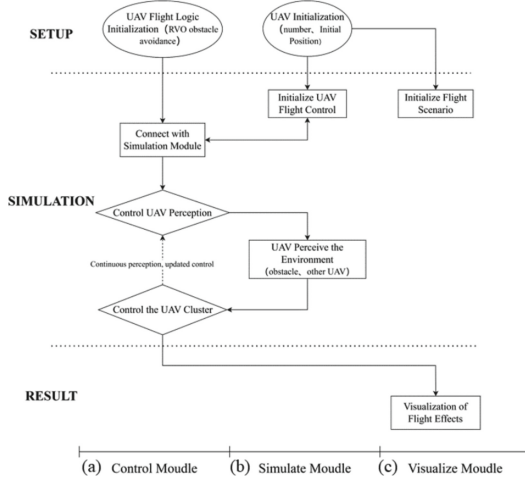


Fig. 3. Integrated tool framework: (a) Flight control module; (b) UAV model simulation module; (c) Environment simulation and visualization module.

4 Preliminary Simulation Results and System Evaluation

Two scenarios are set up to test the framework, in which the UAV models are from AirSim built-in models. In a simple conflict scenario and a complex conflict scenario, the initial position of the UAVs is relative to (0,0,0) in the NED coordinate system, and the coordinates of the origin indicate the player position in UE4. The unit of this coordinate system is m. The following simulation experiments have a time step of 150 steps, and the duration of each step is 0.1 s.

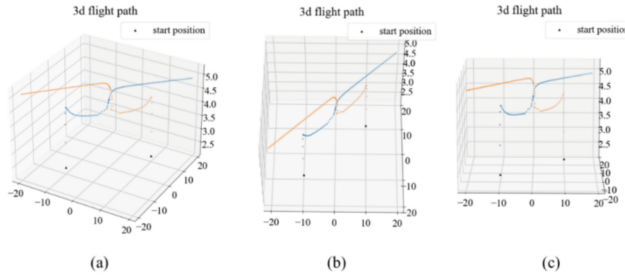
4.1 Simple Conflict Scenario Simulation

In this simulation experiment, two UAVs are set up, symmetrical about the origin of the NED coordinate system, and the direction of the setup is the direction pointing to the origin, and the queue arrangement of the UAVs is shown in Fig. 4.

Scenario1:

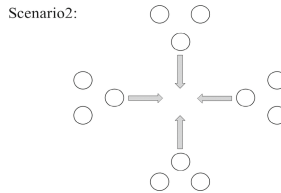
**Fig. 4.** Schematic diagram of a simple conflict scenario

The simulated three-dimensional trajectories of the UAVs are shown in Fig. 5, where the two UAVs take off to a predetermined altitude and travel in opposite directions, and continue to move in the initial direction after avoiding obstacles near the midpoint.

**Fig. 5.** (a) 3D simulation trajectory of a simple conflict scene; (b) top view of the simulation trajectory; (c) main view of the simulation trajectory

4.2 Simulation of Complex Conflict Scenarios

The simulation experiment set up 12 UAVs, 3 for a group, four groups of UAVs about the origin of the NED coordinate system symmetry, set the direction of the direction of the origin, the UAV queue arrangement is shown in Fig. 6, and Fig. 7 shows the 3D trajectory of UAV cluster movement under the simulation of complex conflict scenario.

**Fig. 6.** Schematic diagram of a complex conflict scenario

The simulation results in complex conflict scenarios show that the proposed simulation framework and multi-subject cooperative control strategy can effectively simulate the behavior of UAV swarms. The applicability of the system in large-scale simulation is preliminarily evaluated, showing good performance and scalability.

These preliminary results demonstrate the effectiveness of the framework in simulating UAV swarm behavior.

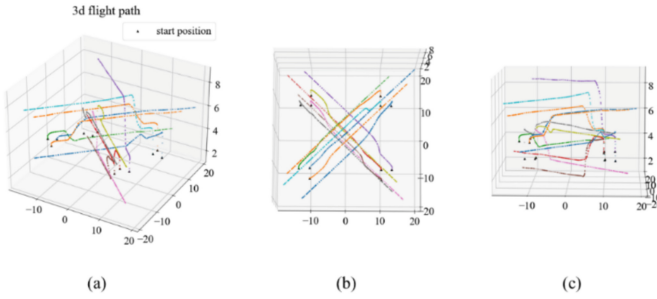


Fig. 7. (a) 3D simulation trajectory of complex conflict scene; (b) top view of simulation trajectory; (c) main view of simulation trajectory

5 Conclusions and Future Perspectives

5.1 Conclusions

In this study, a cross-platform integrated UAV virtual simulation framework is successfully implemented, and autonomous obstacle avoidance of UAV swarms is realized in two scenarios. The effectiveness of the framework is demonstrated. This cross-platform integration approach provides a flexible and scalable solution for UAV simulation.

Potential of multi-object cooperative control strategy: By applying the RVO2 model with improved time horizon, this framework can effectively simulate the complex interaction and cooperative behavior of UAV swarms. This provides a solid foundation for further research on cooperative control strategies for UAV swarms.

5.2 Future Perspectives

Although this research has achieved some results, there are still many areas that need further research and development. This study currently realizes a small-scale simulation, and it is hoped that the simulation can be performed based on a larger number of UAVs in the future research to test the simulation. It is hoped that the simulation can be based on a larger number of UAVs in future research to test the performance of the simulation.

Acknowledgments. The work is supported by the National Natural Science Foundation of China (Grant No. 72101276) and the National College Student Innovation and Entrepreneurship Training Program (Grant No. 20242273).

References

1. Skorobogatov, G., Barrado, C., Salami, E.: Multiple UAV systems: a survey. *Unmanned Syst.* **08**(02), 149–169 (2020). <https://doi.org/10.1142/s2301385020500090>
2. Blue V, Embrechts M, Adler J (1997) Cellular automata modeling of pedestrian movements. In: 1997 IEEE international conference on systems, man, and cybernetics. computational cybernetics and simulation. IEEE, pp 2320–2323. <https://doi.org/10.1109/ICSMC.1997.635272>
3. Duan, Y.P., et al.: Path planning optimization for swine manure-cleaning robots through enhanced slime mold algorithm with cellular automata. *Anim. Sci. J.* **95**(1), e13992 (2024). <https://doi.org/10.1111/asj.13992>
4. Fei, L.-J., Lee, J., Huang, X., Peper, F.: Effect of random fluctuations on minimizing the complexity of universal asynchronous cellular automata. *Physica D* **428**, 133052 (2021). <https://doi.org/10.1016/j.physd.2021.133052>
5. Zhao, M., Liang, C.: A novel extended social force model for studying the impact of the heterogeneity of pedestrian physical fitness on emergency evacuation efficiency. *Physica A* **653**, 130101 (2024). <https://doi.org/10.1016/j.physa.2024.130101>
6. Hou, X., Zhang, R., Yang, M., Cheng, S.: Modeling the lane-changing behavior of non-motorized vehicles on road segments via social force model. *Physica A* **633**, 129415 (2024). <https://doi.org/10.1016/j.physa.2023.129415>
7. Jun, H., Qi, Z., Ling, F., Yifan, L., Longcheng, Y., Juan, W.: Simulation of a multi-floor evacuation with the extended social force model in a multi-corridor environment. *Phys. Scr.* **98**(10), 105241 (2023). <https://doi.org/10.1088/1402-4896/acf802>
8. Yang, X., Yang, X., Li, Y., Zhang, J., Kang, Y.: Obstacle avoidance in the improved social force model based on ant colony optimization during pedestrian evacuation. *Physica A* **583**, 126256 (2021). <https://doi.org/10.1016/j.physa.2021.126256>
9. Fiorini, P., Shiller, Z.: Motion planning in dynamic environments using velocity obstacles. *Inter. J. Robot. Res.* **17**(7), 760–772 (1998). <https://doi.org/10.1177/027836499801700706>
10. Liu, W., Zhang, B., Liu, P., Pan, J., Chen, S.: Velocity obstacle guided motion planning method in dynamic environments. *J. King Saud Univ. Comput. Inform. Sci.* **36**(1), 101889 (2024). <https://doi.org/10.1016/j.jksuci.2023.101889>
11. Gao J, Zhang H, Tan L, Ren X (2022) UAV dynamic obstacle avoidance based on improved reciprocal velocity obstacle. In: *Journal of Physics: Conference Series*. IOP Publishing, p 012014. <https://doi.org/10.1088/1742-6596/2216/1/012014>
12. Wakabayashi, T., Yukimasa, S., Suzuki, S.: Dynamic obstacle avoidance for Multi-rotor UAV using chance-constraints based on obstacle velocity. *Robot. Auton. Syst.* **160**, 104320 (2023). <https://doi.org/10.1016/j.robot.2022.104320>
13. Van Den Berg J, Guy SJ, Lin M, Manocha D (2011) Reciprocal n-body collision avoidance. In: *Robotics Research: The 14th International Symposium ISRR*. Springer, pp 3–19
14. Snape J, Manocha D (2010) Navigating multiple simple-airplanes in 3D workspace. In: 2010 IEEE International Conference on Robotics and Automation. IEEE, pp 3974–3980
15. Shah S, Dey D, Lovett C, Kapoor A (2018) Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In: *Field and Service Robotics: Results of the 11th International Conference*. Springer, pp 621–635. https://doi.org/10.1007/978-3-319-67361-5_40



The Influence of Parameter Optimization of Rubber Nodes in High-Speed Train Turning Arms on Suppressing Hunting Motion

Tao Guo^{1,2}, Ya Li³, and Yongqiang Liu^{4,5}✉

¹ Zhan Tianyou College, Dalian Jiaotong University, Dalian 116028, China
sjc-guotao@tangche.com

² CRRC Tangshan Co., Ltd., Tangshan 063000, China

³ Shijiazhuang Enric Gas Equipment Co., Ltd., Shijiazhuang 051000, China
liya@enricgroup.com

⁴ State Key Laboratory of Mechanical Behavior and System Safety of Traffic Engineering Structures,
Shijiazhuang Tiedao University, Shijiazhuang 050043, China
liuyq@stdu.edu.cn

⁵ School of Mechanical Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China

Abstract. To solve the problems of frequent occurrence of lateral instability and abnormal wear of the tread surface during the service of high-speed trains, the stiffness characteristics and its parameters optimization of the rotating arm rubber node were analyzed in this paper. A high-speed train dynamic model was established using the dynamic analysis software Universal Mechanism with a type of 350 km/h high-speed train in China as the research object. Based on the mechanism of primary and secondary vehicle instability, the phenomena of abnormal hunting of the vehicle body and abnormal vibration of the bogie was reproduced through simulation, and the vehicle response and hunting characteristics were analyzed. The Poynting Thomson model, which can represent the stiffness frequency variation characteristic of rotating arm rubber node, was optimized by the genetic optimization algorithm. The research results indicate that the optimized arm node has a significant inhibitory effect on the instability of primary and secondary hunting movements.

Keyword: High-speed train · Hunting motion · Dynamical simulation · Rotating arm rubber note

1 Introduction

The mileage of high-speed railways in China has exceeded 45000 km, and the operating speed of high-speed trains can reach up to 350 km/h. With the increase of operating mileage and speed, the frequency of track disturbance continues to increase, the wheel rail interaction force increases, the vibration of various parts of the vehicle body increases,

and the wheel rail matching relationship deteriorates. The resulting wheel rail wear will seriously affect the stability of train operation, leading to a decrease in operational quality and an increase in maintenance costs for both the train and the track. Therefore, how to suppress the hunting of the vehicle body during small cone angles and the alarm problem of the frame during large cone angles is a difficult problem that troubles railway workers [1–3].

In terms of studying the causes and countermeasures of low-frequency train hunting, He conducted research and analysis on high-speed trains on a certain line [4]. Kaiser proposed an anti-yaw damping monitoring method for railway secondary suspension through a nonlinear constrained approach integrated with a randomly variable wheel-rail interaction [5]. Fu found that after the wheel tread was repaired and matched with the excessively polished steel rail, the wheel rail contact relationship was poor, and a too small equivalent conicity could easily cause hunting of the vehicle [6]; Zhao proposed a new semi-active control strategy on secondary lateral damper to avoid strong vibration of high-speed trains [7]. Chi established a vehicle system dynamics model, in which the frequency of the bogie's hunting motion increases with increasing speed, which can cause resonance and affect the stability of the train [8]; Wang summarized and explained several types of causes and prevention measures for car hunting in actual operation [9]; Xia explained the reason for vehicle body hunting at smaller equivalent conicity angles from the perspective of frequency capture, and proposed solutions [10].

These studies are all based on practical problems in train operation, analyzing the reasons for the occurrence of primary and secondary trains and proposing a series of solutions, which have good reference value. However, due to the contradictory analysis of the wheel rail contact state between the first and second hunting movements, there is a certain degree of mutual exclusion in the solution of simply suppressing vehicle instability and abnormal bogie vibration. How to consider the advantages of each solution and combine them is the key to improving the stability and comfort of the vehicle system as a whole.

2 The Dynamical Model of High-Speed Train

In order to study feasible solutions, a high-speed train of a certain type in China was taken as the research object, and a multi-body dynamics software UM was used to establish a high-speed train dynamics model. The bogie consists of two wheelsets, one frame, primary and secondary suspension systems. The dynamic model of the high-speed train consisting of one body and two bogies is shown in Fig. 1. The wheel tread adopts LMA type, and the steel rail adopts CHN60 type.

3 The Reproduction of Hunting Motion by Simulation

In practical engineering, high-speed trains often experience hunting and bogie alarms due to wheel rail wear. Therefore, this article uses the established dynamic model to simulate and reproduce these two types of hunting instability phenomena, in order to provide targeted solutions. At normal operating speeds, high-speed trains will experience two types of serpentine instability: one is low-frequency hunting motion characterized by

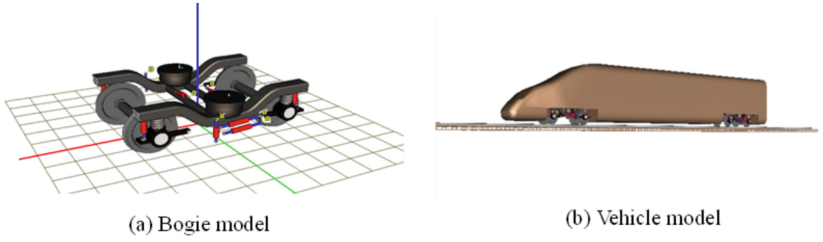


Fig. 1. The dynamic vehicle model of high-speed train

lateral hunting of the train body when the equivalent cone is small, with a lateral Sperling index exceeding 2.5 and a frequency of 1–2 Hz; Another type is high-frequency bogie acceleration alarm, which is prone to occur when the equivalent conicity is greater than 0.35. At this time, the hunting frequency is between 6–10 Hz. After filtering at 0.5–10 Hz, if the amplitude is greater than 0.8 g for more than 6 consecutive times, it is judged as unstable.

The factors considered in the simulation process include:

- (1) Track irregularities: German low interference spectrum, German high interference spectrum, and Chinese high-speed railway track irregularities;
- (2) Vehicle operating speed: from 280 to 350 km/h (with a step size of 5 km/h);
- (3) Line conditions: Straight track and horizontal curved track with a radius of 5500 ~ 12000 m on the main line (with a step size of 500 m for each change).

When the vehicle shakes, there is a significant harmonic vibration in the lateral acceleration of the vehicle body, as shown in Fig. 2, which is relatively larger than under normal operating conditions (the lateral acceleration of the vehicle body is generally 0.4 m/s^2); At this point, the main frequency is very prominent, around 1.7 Hz, and vibrations far from the main frequency are relatively small.

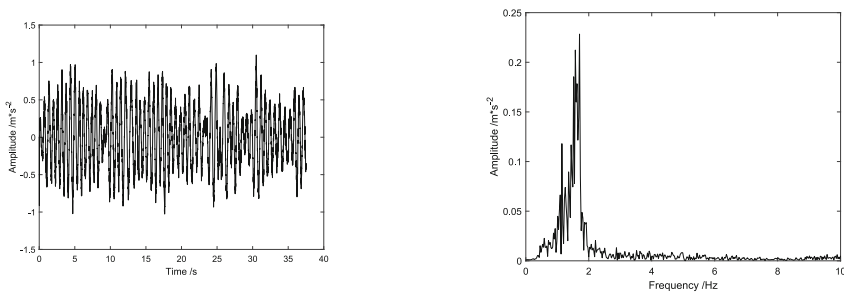


Fig. 2. Lateral acceleration of carbody in time and frequency domain

The acceleration alarm of the bogie generally occurs when the equivalent conicity is greater than 0.35. According to the route tracking situation, the alarm on a straight line is 26% higher than that on a curved condition. A multi rigid body model is used to simulate

the instability phenomenon of the bogie under large taper angle. The operating mileage of 200 km in the simulation is matched with worn steel rails, the vehicle running speed is 340 km/h, the wheel rail friction coefficient is 0.15, and the German high interference is used as the unevenness excitation. As shown in Fig. 3, there are significant harmonics after filtering the lateral acceleration at the end of the framework, with amplitudes exceeding 10 m/s^2 . The red dashed line represents the acceleration limit of 8 m/s^2 , with a main frequency of approximately 6 Hz.

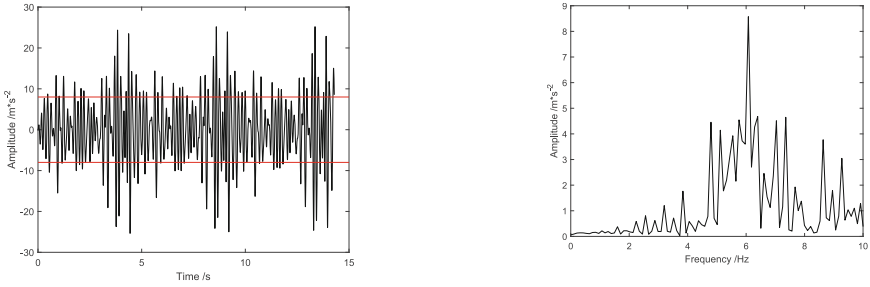


Fig. 3. Lateral acceleration of bogie frame in time and frequency domain

4 The Influence of Stiffness Frequency Variation Characteristics of Rubber Nodes in Rotating Arms

In order to analyze the influence of the longitudinal stiffness of the boom node on the primary hunting behavior, the relationship between the transverse Sperling index of the vehicle body and the longitudinal stiffness of the boom node was studied. The longitudinal stiffness of the turning arm node is designed to be 4 MN/m. The comparison between the lateral acceleration of the vehicle body and the lateral acceleration of the vehicle body under a single hunting motion is shown in Fig. 4.

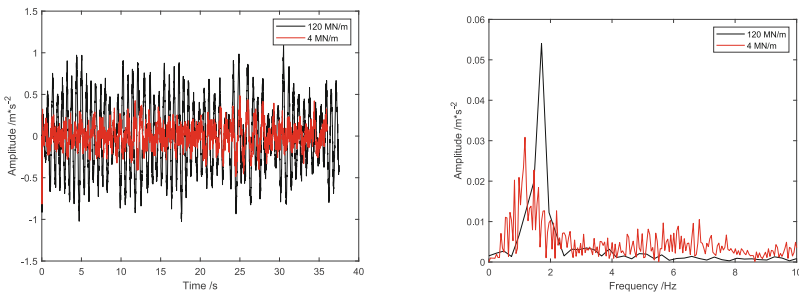


Fig. 4. Effect of optimal stiffness on lateral acceleration of carbody

The average lateral acceleration of the vehicle during a single hunting motion is 0.6 m/s^2 , while when the longitudinal stiffness of the pivot node is reduced to 4 MN/m , the average lateral acceleration of the vehicle is 0.2 m/s^2 , which is much lower than that during a single hunting motion. The longitudinal stiffness of the turning arm node is designed to be 266 MN/m . The comparison between the filtered lateral acceleration of the frame and the lateral acceleration of the frame under secondary hunting motion is shown in Fig. 5.

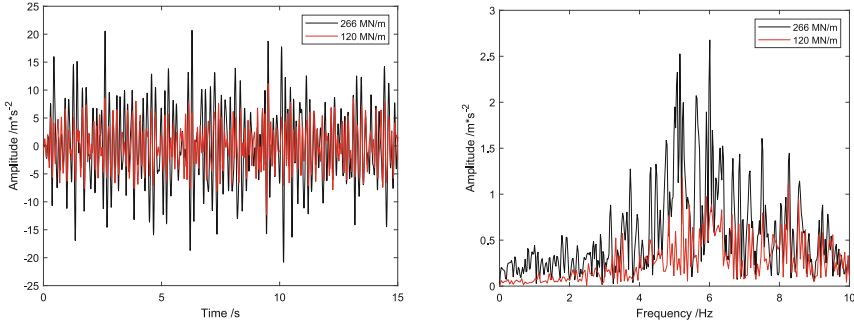


Fig. 5. Effect of optimal stiffness on lateral acceleration of bogie

From the above analysis results, it can be seen that reducing the longitudinal stiffness of the arm node to 4 MN/m can effectively suppress primary hunting instability, while increasing the longitudinal stiffness of the arm node to 266 MN/m can effectively suppress secondary hunting instability. However, due to the lack of consideration for these dynamic requirements in the design of the rubber joint of the rotating arm, it is difficult to meet the needs of both working conditions simultaneously. Therefore, there is a need for a rotating arm joint that can adjust its stiffness according to changes in motion frequency.

5 Multi-objective Optimization of the Mechanical Model of the Rotating Arm Rubber Node

The material of the arm positioning node is generally rubber, which has complex non-linear characteristics, specifically manifested as stiffness that can vary with amplitude, frequency, and temperature. This article is based on the variable stiffness arm positioning model, and the parameters are reset to adapt to the vehicle model, thereby improving the problems of hunting and steering frame alarms. The German company Freudenberg Schwab has developed a rotating arm positioning node with frequency dependent stiffness (FDS) characteristics.

In PT model, $x_0 = A_0 \cdot \sin(2\pi f \cdot t)$, is the amplitude, and is the frequency. The spring force in the series unit is equal to the damping force, that is.

Dynamic stiffness:

$$k_d = \frac{k_c^2 c^2 \omega^2}{k_c^2 + c^2 \omega^2} + k_s \quad (1)$$

Dynamic damping:

$$c_d = \frac{k_c^2 c}{k_c^2 + c^2 \omega^2} \quad (2)$$

The optimization design problem of the variable stiffness model parameters is described as the following multi-objective optimization problem: comprehensively consider selecting three parameters as the optimization variables for multi-objective optimization, with an optimization variable range of 0–300 MN, and two optimization objective functions, namely, the root mean square value of the lateral acceleration of the vehicle body under hunting problem and the root mean square value of the lateral acceleration of the frame under frame alarm problem. Their mathematical expressions can be written as:

$$\begin{aligned} & \min(C(x), B(x))^T \\ & s.t. = \begin{cases} 0 < x \leq 300 \\ \sqrt{\frac{\sum C(x)_i^2}{N}} \leq 0.416 \\ \sqrt{\frac{\sum B(x)_i^2}{N}} \leq 13.89 \end{cases} \end{aligned} \quad (3)$$

Write M files using UM software for the small taper operating conditions that are prone to causing hunting and the large taper operating conditions that are prone to causing frame alarms, and establish S functions in MATLAB SIMULINK. The output of the hunting problem is the lateral acceleration value of the vehicle body, and the output of the frame alarm problem is the lateral acceleration value of the frame.

Based on the joint simulation model, establish a fitness function and apply the genetic algorithm function GA in MATLAB to solve the multi-objective optimization problem of parameters. Partial parameter settings are as follows: the optimal population individual coefficient is 0.3, the population size is 100, and the maximum evolutionary generation is 200.

Further selection and optimization of a suitable set of non dominated solutions were carried out based on the two operating conditions of the train. At this point, the three parameters of the variable stiffness model were obtained, as shown in Table 1, and the dynamic stiffness with frequency variation is shown in Fig. 14. According to the variation of dynamic stiffness with frequency, it can be concluded that under the condition of small taper that is prone to hunting, the main frequency is 1.7 Hz, and the dynamic stiffness is around 48 MN. In the case of frame alarm, the main frequency is 6 Hz, and the dynamic stiffness is around 215 MN. The required dynamic stiffness values for both conditions are within our requirements. The variable stiffness model will be applied to the vehicle to test its dynamic performance from three aspects.

The comparison of lateral acceleration of the vehicle body under the action of fixed stiffness and FDS is shown in Fig. 6. The average lateral acceleration of the vehicle body during a hunting motion is 0.6 m/s². The lateral acceleration of the vehicle body

under the action of FDS is slightly lower than the lateral acceleration of the frame under constant stiffness, and the average lateral acceleration of the vehicle body is 0.4 m/s^2 , which is much lower than the lateral acceleration of the vehicle body during a hunting motion.

Table 1. Results of PT model

Variable	value
$k_s/\text{MN.m}$	2.89
$k_c/\text{MN.m}$	291.62
$c/\text{MN.m}$	79.83

The lateral acceleration of the frame under the action of FDS has decreased compared to the lateral acceleration of the frame under constant stiffness. During a period of time before and after 4 s, the decrease in acceleration is greater than that of other parts, with an average of 10%; When the lateral acceleration value of the framework is large, the suppression effect under FDS is more significant. The lateral acceleration PSD of the structure is shown in Fig. 6.

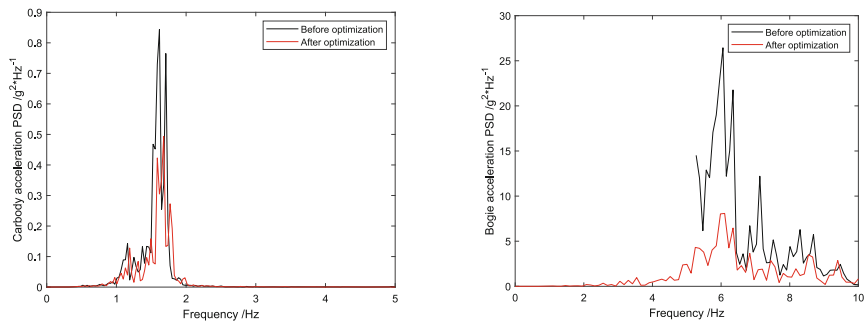


Fig. 6. PSD of the acceleration of the carbody and bogie

The main frequency is still 6 Hz, and the PSD decreases significantly within 4.8–6 Hz, with an average decrease of 26%; Within the frequency range of 0–2 Hz, the change in PSD is not very significant, and within 8–10 Hz, the PSD decreases again under the action of FDS, indicating that the FDS model has a certain effect on improving the structure alarm phenomenon.

6 Conclusion

A FDS node model was established, and its frequency and amplitude characteristics were analyzed. In order to better meet the basic dynamic performance of the vehicle model in the article and more efficiently solve the harsh hunting motion of the vehicle,

the parameters of the FDS model were optimized. Based on genetic algorithm, multi-objective optimization was performed on two types of harsh hunting motion conditions to obtain the optimal solutions for the three parameters of FDS.

The FDS model is applied to analyze the vehicle dynamics from two aspects. It analyzes the vehicle dynamics calculation when the vehicle has never experienced two types of severe hunting motion and uses an unworn tread and rail profile. This is the most basic dynamic performance. On the other hand, the dynamic calculation when the vehicle uses a worn tread and rail profile is analyzed, and the impact of using the FDS model on the vehicle during wear is analyzed; Test the effectiveness of two types of aggressive hunting movements, and achieve inhibitory effects on both types of hunting movements.

References

1. Sun, J., Chi, M., Jiao, W., et al.: Modal parameters-based hunting stability analysis of high-speed railway vehicles considering full range of equivalent conicity. *Proc. Institut. Mech. Eng. Part K: J. Multi-body Dyn.* **236**(3), 422–439 (2022)
2. Zhang, X., Liu, Y.Q., Yang, S.P., et al.: The first passage problem of a stochastic wheelset system. *Commun. Nonlinear Sci. Numer. Simul.* **128**, 107643 (2024)
3. Sun, J., Chi, M., Jin, X., et al.: Experimental and numerical study on carbody hunting of electric locomotive induced by low wheel-rail contact conicity. *Veh. Syst. Dyn.* **59**(2), 203–223 (2021)
4. He, X.S., Wu, H.C., Gao, F.L Experimental study on the mechanism of high-speed train hunting. *J. Dalian Jiaotong Univ.* **38**(001), 21–25 (2017). (in Chinese)
5. Kaiser, I., Strano, S., Terzo, M., et al.: Anti-yaw damping monitoring of railway secondary suspension through a nonlinear constrained approach integrated with a randomly variable wheel-rail interaction. *Mech. Syst. Signal Process.* **146**, 107040 (2021)
6. Fu, Z.B.: Research on abnormal hunting of the vehicle body of Hengliu Line EMU. *Railway Locomotive Rolling Stock* **39**(02), 46–49+54 (2019) (in Chinese)
7. Zhao, Y.W., Liu, Y.Q., Yang, s.p., et al.: a New Coordination Control Strategy on Secondary Lateral Damper for High-Speed Trains. *J. Vibrat. Eng. Technol.*, 1–14 (2021)
8. Chi, M.R., Zhang, W.H., Zeng, J., et al.: The influence of hunting motion on the stability of railway vehicles. *J. Vibrat. Eng.* **21**(06), 639–643 (2008). (in Chinese)
9. Wang, L.L.: Several common hunting and rectification methods in line maintenance. *Tech Wind* **2019**(20), 123–124 (2019). (in Chinese)
10. Xia, Z., Zhou, J., Liang, J., et al.: Online detection and control of car body low-frequency swaying in railway vehicles. *Veh. Syst. Dyn.* **59**(1), 70–100 (2021)



Multi-object Tracking Algorithm of Moving Scene Based on Bidirectional Buffering of Lost Trajectory

Chanzhen Xiong^(✉) and Yuanlong Wu

School of Electrical and Control Engineering, North China University of Technology,
Beijing 100044, China
xczkiong@ncut.edu.cn

Abstract. In the multi-object tracking (MOT) algorithm for sports scenes, the motion model based on Kalman filter cannot cope with the complex motion of objects and the frequent entry and exit of players in videos, which inevitably leads to identity switches. To solve this problem, we propose a bidirectional buffer trajectory management algorithm, based on the prior information that the number of players on the sports field will not change easily, we propose a bidirectional trajectory buffer method. When initializing a new trajectory, the new trajectory needs to match with the previously lost trajectory in advance, and try to recover the trajectory that once left the image, reducing the identity switch. We conducted comparative experiments on SportMOT and DanceTrack to prove the effectiveness of the proposed algorithm. Compared with the benchmark algorithm, the method improved HOTA by 1.1% on SportsMOT and reduced IDSw by 9.5%. On DanceTrack, HOTA was improved by 0.5%, and IDSw was reduced by 14%.

Keywords: Multi-Object Tracking · Kalman Filter · Trajectory Management

1 Introduction

Multi-Object tracking is an important fundamental topic in the field of computer vision, with the main objective of finding the position of each object in a video sequence and labeling it with the correct identity information. Multi-object tracking algorithms are divided into Joint Detection And Tracking (JDT) [1–4] and Tracking By Detection (TBD) [5–9] based on their implementation methods. Among them, joint detection and tracking train detection and tracking as a whole, which has good tracking performance for complex movements. However, it incurs higher design and training costs, as well as increased computational expenses. The detection and tracking algorithm divides detection and tracking into two main parts, which can be trained or adjusted separately according to different datasets. It has the characteristics of fast inference speed and flexible deployment.

The algorithm of the TBD framework demonstrates excellent tracking accuracy for Multi-Object tracking of densely packed pedestrians in monitoring scenarios, such as MOT17 [10] dataset and MOT20 [11] dataset, but performs poorly in datasets such as DanceTrack [12] and SportsMOT [8] where objects have long-term mutual occlusion, intense movement and deformation. SORT [13] algorithm uses Kalman filter to predict the position of the current frame of the object, calculates its IoU (Intersection over Union) loss matrix with the detection frame of the current frame picture, and uses the Hungarian matching algorithm to match, outputs the ID and detection frame of the successfully matched trajectory, and forms the basic framework of TBD. On the basis of SORT algorithm, researchers have proposed a series of improved algorithms, such as ByteTrack [5], OC-SORT [6], Hybrid-SORT [7]. OC-SORT proposes a method to improve the Kalman filter algorithm in dealing with complex motion and occlusion in Multi-Object tracking. When the trajectory is lost, it is not predicted, and when it is re matched, interpolation is used to restore the state value of the Kalman filter. The MixSort algorithm proposes an appearance feature assisted matching structure based on Mixformer, which achieves better tracking on SportsMOT by combining motion features. In addition to combining motion and appearance features for matching, the Hybrid-SORT algorithm also adds many weak clues, such as confidence changes, height states, etc., to further improve tracking accuracy.

The Kalman filter motion model, which is based on the assumption of linear motion and utilized within the TBD framework, struggles to effectively model objects that frequently exit the camera's field of view. To tackle these challenges, we introduce a Multi-Object tracking algorithm that leverages bidirectional buffering. Specifically, for objects that move out of the camera's range, we store their IDs and appearance features in a cache. When initializing a new object, we attempt to match it with the cached objects. If a successful match is found, we restore the trajectory and state of the matched object. This approach enhances the robustness and accuracy of the tracking algorithm in scenarios involving frequent object disappearance and reappearance.

2 Trajectory Management Strategy Based on Bidirectional Buffering

In order to reduce the generation of new IDs or ID switching caused by trajectory entering and exiting the camera, a trajectory management strategy based on bidirectional buffering was designed. Its logic is shown in Fig. 1. When the trajectory is lost on the left and right boundaries of the image, it is considered to be out of the camera range and its appearance features are stored. When a object is initialized from the left and right boundaries, its appearance features are matched with the previously stored appearance features. If the match is successful, the previous ID is used to reduce ID switching.

In the SORT algorithm, the initialization of trajectories can usually be expressed as the following formula.

$$tracklet_{untracked} = det_{unmatched} \text{ if } det.conf > init_thresh \quad (1)$$

Among them, *det* refers to all the detection boxes input to the tracker. *det_{unmatched}* refers to the detection boxes that still cannot match the current trajectory after participating in multiple rounds of matching with the tracker. The detection boxes contain position *det.dst* and confidence *det.conf*, *init_thresh* refers to the threshold for initializing the trajectory. It is usually set to 0.7, indicating that the trajectory will only be initialized when the confidence of the detection boxes is at least 0.7.

However, this cannot cope with the frequent entry and exit of objects caused by the camera swinging back and forth in sports scenes, which leads to ID switching. As shown in Fig. 1, when the athlete kicks the ball into the right half, the camera only captures the right half of the field, and the athletes on the left and middle of the field lose track. However, in reality, they are still on the field and the number of players has not changed. When the athlete launches a counterattack, the camera returns to the middle, and the athletes on the left and right sides of the field lost track.

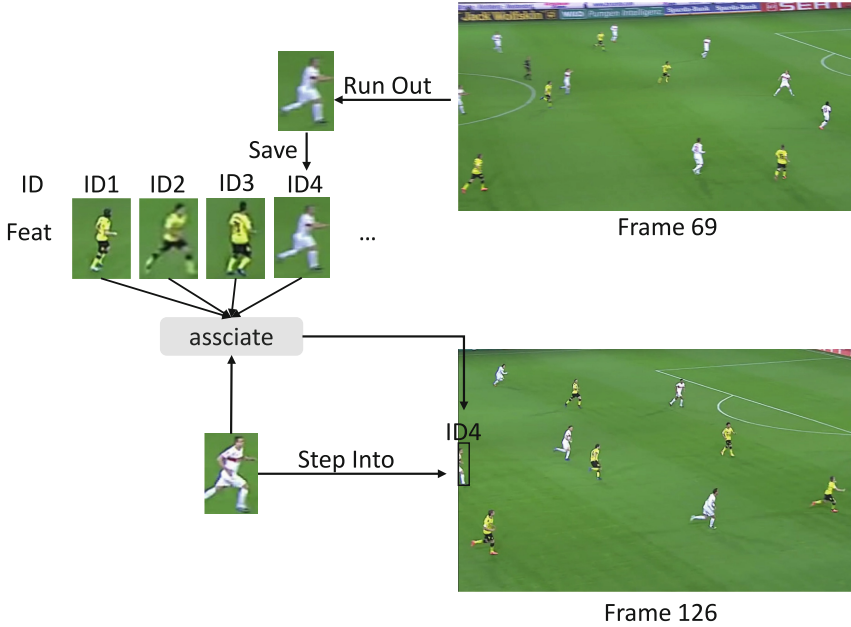


Fig. 1. Schematic diagram of bidirectional buffer trajectory management strategy

To solve this problem, we propose a bidirectional buffering trajectory management method. If the position of the object box where the trajectory is lost is located (x_1, y_1, x_2, y_2) , (x_1, y_1) is the upper left corner of the object box, (x_2, y_2) is the lower right corner of the object box, and its center coordinate is (x_c, y_c) , then the following formula is given.

$$(x_c, y_c) = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \quad (2)$$

For images, their width w is pixels and their height is w pixels. When the trajectory is lost and it satisfies $x_c < aw$, put its ID and corresponding appearance features into the buffer list on the left $buffer_{left}$. The a is proportional parameter for width is usually 0.16, or when the trajectory is lost, if it satisfies $x_c > (1 - a)w$, put its ID and corresponding appearance features into the buffer list on the right $buffer_{right}$, where a is the same as in the previous text. This creates $buffer_{left}$ and $buffer_{right}$, where the appearance features and corresponding IDs of all trajectories that meet the above conditions for disappearance are saved in order

When initializing, first determine the x-axis position of the initialization trajectory x_{init} whether meets the requirements $x_{init} < aw$ (or $x_{init} > (1 - a)w$). If satisfied, Further determine whether $buffer_{left}$ (or $buffer_{right}$) is empty, If not empty, use the loss matrix of REID to perform Hungarian matching between the initialized trajectory and the trajectory in the $buffer_{left}$ (or $buffer_{right}$). Otherwise, use the new ID to initialize the trajectory normally. The calculation method for the REID loss matrix of two trajectories with appearance feature vectors of and is shown in the following formula.

$$Cost_f = 1 - \frac{a \cdot b}{||a|| \times ||b||} \quad (3)$$

Among them, $Cost_f$ refers to the loss matrix by feature. a and b is the feature vectors of two trajectories.

3 Experimental Results and Analysis

3.1 Dataset and Evaluation Metrics

We conducted experiments on our algorithm on both the large-scale multi athlete dataset SportsMOT and the large-scale dancer Multi-Object tracking dataset DanceTrack. All experiments were conducted on an RTX3090 using Ubuntu 22.04, CUDA version 12.3, and PyTorch version 12.1.

SportsMOT contains a total of 240 videos and over 150K images with a resolution of 1280x720. There are 45 videos in the training set, 45 videos in the validation set, and 150 videos in the testing set. Each video contains three types of sports: football, basketball, and volleyball. The characteristic is that the motion is complex and fast, containing large and regular camera movements, and the object has similar appearance and motion patterns. At the same time, because it is a sports scene, the number of people on the field has hardly changed.

DanceTrack contains 100 videos, with most of the images having a resolution of 1920x1080. There are 40 videos in the training set, 25 videos in the validation set, and 35 videos in the testing set. It faces multi object tracking of characters in dance scenes, including indoor, outdoor, stage and other scenarios. There are a large number of dancers with highly nonlinear movements and similar appearance features, as well as frequent occlusions and intersections, which are of great significance for evaluating and improving tracking algorithms.

Since none of the above datasets have open-source annotations for their test sets, the experimental results were obtained by submitting them to the corresponding testing websites for evaluation.

We use HOTA [14], MOTA, IDF1, and IDSw as our main evaluation metrics to compare two tracking algorithms. HOTA refers to high-order tracking accuracy, which combines three indicators of detection, matching, and positioning accuracy for comprehensive evaluation, providing a more comprehensive and accurate evaluation method. MOTA stands for Multi-Object tracking accuracy, and compared to HOTA, it can better reflect the quality of the detection results when reflecting the tracking accuracy of the tracker. IDF1 can reflect the performance of the data association part of the tracker, while AssA can reflect the accuracy of the tracker matching. In addition, the experiment also referred to four indicators: DetA, and LocA. The values of the indicators HOTA, IDF1, MOTA, AssA, DetA, and LocA used in the experiment are percentages, and the unit of FPS is frames per second.

3.2 Detector

In order to fairly and intuitively compare the performance of our tracker with other trackers, we used the YOLOX-X [15] detector, which is used by most tracking algorithms that detect tracking frameworks first. When conducting experiments on the SportsMOT dataset, we used a detector provided by DeepEIoU, which was trained on the official weights of YOLOX-X and fine tuned for 80 rounds on the SportsMOT dataset. The input image size is 1440×800 , and data augmentation includes mosaic and mixup. The SGD optimizer with weight decay of 5×10 and momentum of 0.9 is used. The initial learning rate is 0.01, following the same training process as ByteTrack. On the DanceTrack dataset, we used a detector provided by OC-SORT, which was trained on the official weights of YOLOX-X and fine tuned for 20 rounds on the DanceTrack dataset.

3.3 Comparative Experiment

To verify the tracking performance of this algorithm, experiments were conducted on DanceTrack and SportsMOT test sets, and compared with ByteTrack, OC-SORT, and Hybrid SORT algorithms that detect first and match later. At the same time, it is also compared with joint detection and tracking algorithms such as FairMOT, MOTR, MOTRv2, and MOTIP. The detector used in the experiment is the YOLOX model provided by ByteTrack.

The experimental results on the DanceTrack dataset are shown in Table 1. After adding the bidirectional buffering trajectory management algorithm proposed in this paper to the Hybrid SORT algorithm, HOTA improved by 0.5%, IDF1 improved by 1.7%, and IDSw decreased from 1470 to 1268.

To further validate the results of this algorithm, comparative experiments were conducted on the SportsMOT test set. The experimental results on the SportsMOT test set are shown in Table 2. After adding the algorithm proposed in this paper, HOTA increased by 1.1%, Assa increased by 1.9%, and IDSw decreased from 2659 to 2409.

Table 1. The results of the DanceTrack test dataset

Tracker	HOTA	IDF1	MOTA	DetA	AssA	IDS _w
JDE						
FairMOT	39.7	40.8	82.2	66.7	23.8	\
MOTR	54.2	51.5	79.7	73.5	40.2	\
MOTRv2	69.9	71.7	91.9	83.0	59.0	\
MOTIP	67.5	72.2	90.3	79.4	57.6	\
TBD						
ByteTrack	47.7	53.9	89.6	71.0	32.1	3105
OC-SORT	61.3	61.5	92.3	82.2	45.8	1992
Hybrid-SORT	65.7	67.4	91.8	82.2	52.6	1470
Our	66.2	69.1	92.6	82.3	53.4	1268

Table 2. Comparison results on the SportsMOT test dataset

Algorithm	HOTA	IDF1	MOTA	AssA	DetA	LocA	IDS _w
FairMOT	49.3	53.5	86.4	34.7	70.2	83.9	21673
ByteTrack	64.1	71.4	95.9	52.3	78.5	85.7	3089
MixSort-Byte	65.7	74.1	96.2	54.8	78.8	85.7	2472
MixSort-OC	74.1	74.4	91.6	62.0	88.5	92.7	2781
Deep-ElIoU	77.2	79.8	96.3	67.6	88.2	92.4	2659
Our	78.3	83.3	96.7	69.5	88.3	92.4	2409

3.4 Visual Analysis

To further validate the effectiveness of the algorithm strategy proposed in this paper, the tracking results of SportsMOT and DnaceTrack were visualized. The first and second rows of Fig. 3 visualize the tracking results of the improved algorithm when the object in the DanceTrack test set is occluded. The third to fifth rows of Fig. 2 visualize the tracking performance of the improved algorithm on the SportsMOT dataset when the camera swings significantly left and right.



Fig. 2. The proposed algorithm achieves more stable tracking by handling ID switching caused by frequent camera movements in various sports scenes, using randomly sampled segments from DanceTrack (first and second rows) and SportsMOT (third to fifth rows).

4 Summary

In Multi-Object tracking algorithms, for lost trajectories, they are usually saved for prediction and update until they are re tracked or exceed 30 or 60 frames. However, this approach does not take into account the frequent entry and exit of athletes into the camera in sports scenes. We propose a trajectory management strategy based on bidirectional buffering, which saves the appearance features and IDs of trajectories lost on the image boundary. If a new trajectory is initialized on the image boundary, the appearance features of the new trajectory are matched with the appearance features in the cache. If the match is successful, the new trajectory is initialized with the cached ID, otherwise it is initialized with the new ID. Comparative and visualization experiments were conducted on the DanceTrack and SportsMOT datasets to demonstrate the effectiveness of our algorithm. Experiments have shown that compared to the baseline the addition of our algorithm significantly improves tracking accuracy.

Acknowledgments. It is sponsored by the State Key Lab of Intelligent Transportation System under Project No. 2024-A001.

References

1. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: FairMOT: on the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **129**, 3069–3087 (2021)
2. Zeng, F., Dong, b., zhang, y., wang, t., zhang, x., wei, y.: motr: End-to-End Multiple-Object Tracking with Transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., and Hassner, T. (eds.) *Computer Vision – ECCV 2022*, pp. 659–675. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19812-0_38
3. Zhang, Y., Wang, T., Zhang, X.: MOTRv2: bootstrapping end-to-end multi-object tracking by pretrained object detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22056–22065 (2023)
4. Gao, R., Zhang, Y., Wang, L.: Multiple Object Tracking as ID Prediction, <https://arxiv.org/abs/2403.16848v1>, Accessed 11 June 2024
5. Zhang, Y., et al.: ByteTrack: multi-object tracking by associating every detection box. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., and Hassner, T. (eds.) *Computer Vision – ECCV 2022*, pp. 1–21. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_1
6. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric SORT: rethinking SORT for robust multi-object tracking. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9686–9696. IEEE, Vancouver, BC, Canada (2023)
7. Yang, M., et al.: Hybrid-SORT: Weak Cues Matter for Online Multi-Object Tracking, <http://arxiv.org/abs/2308.00783> (2023)
8. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes. Presented at the *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023)
9. Huang, H.-W., Yet al.: Iterative scale-up expansioniou and deep features association for multi-object tracking in sports. In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 163–172. IEEE, Waikoloa, HI, USA (2024)
10. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: MOT16: A Benchmark for Multi-Object Tracking, <http://arxiv.org/abs/1603.00831> (2016)
11. Dendorfer, P., et al.: MOT20: a benchmark for multi object tracking in crowded scenes, <http://arxiv.org/abs/2003.09003> (2020)
12. Sun, P., et al.: DanceTrack: Multi-object tracking in uniform appearance and diverse motion. Presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022)
13. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468 (2016)
14. Luiten, J., et al.: HOTA: a higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* **129**, 548–578 (2021)
15. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO Series in 2021, <http://arxiv.org/abs/2107.08430> (2021)



A Cost-Sensitive Multi-scale Feature Multi-order Fusion Network for Bearing Fault Diagnosis Under Data Imbalance Conditions

Shuaiqing Deng^{1,2,3}, Zihao Lei^{1,2,3}(✉), Guangrui Wen^{1,2,3}, Yu Su^{1,2,3}, Zimin Liu^{1,2,3}, Zhangxuan Meng^{1,2,3}, and Zhifen Zhang^{1,2,3}

¹ National Key Laboratory of Aviation Power System and Plasma Technology, Xi'an Jiaotong University, Xi'an 710049, China

zihao_lei@163.com, {grwen, suyuchn, zzf919}@xjtu.edu.cn, liuzimin@stu.xjtu.edu.cn

² State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China

³ Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an Jiaotong University, Xi'an 710049, China

Abstract. Railroad transportation plays a vital role in the national economy, and fault diagnosis of its key components is of great significance in ensuring its safe operation. However, due to its long-term operation in a healthy state, it is tough to obtain fault signals, and this class imbalance data seriously degrades the diagnostic capability of intelligent diagnostic algorithms, leading to increased misclassification of minority classes. To address this problem, we propose a cost-sensitive multi-scale feature multi-order fusion network for bearing fault diagnosis under data imbalance conditions. Firstly, a lightweight multi-scale convolutional layer is employed to effectively extract the multi-scale features of the signal. Then the multi-order feature fusion is performed by a hornet to fully utilize the multi-scale information and enhance the ability to mine the limited diagnostic information of the minority class. And the excellent high-dimensional discriminative features are extracted through residual blocks. In addition, an improved cost-sensitive strategy is designed through the posterior probability distribution to rebalance the contribution of each class to the training process for improving the overall diagnostic performance of the model. Finally, the effectiveness and superiority of the proposed method are comprehensively verified by several sets of experiments with different imbalance rates.

Keywords: train transmission system · fault diagnosis · data imbalance · cost-sensitive learning · multi-scale feature extraction · multi-order feature fusion

1 Introduction

Railroad transportation is known for its high efficiency, low cost, and reduced emissions, making it a vital component of modern global transit systems. Recently, advancements in technology and increasing travel demands have led to significant enhancements in

train speed and frequency, thereby raising the safety and reliability standards for train operations [1, 2]. The train transmission system plays an essential role by transferring kinetic energy from the motor to the wheels, which is critical for normal train function [3]. However, the bearings within this system operate under high speeds and heavy loads for extended periods, increasing the likelihood of failure. Consequently, diagnosing faults in train bearings is essential for ensuring the safe and reliable operation of trains and enhancing economic performance [4].

In recent years, the rapid advancement of sensor and information technology has enabled the collection of extensive monitoring data from trains through various sensors. Intelligent diagnostic methods powered by deep learning can effectively extract valuable diagnostic insights from this vast amount of data, moving beyond the reliance on manual expertise. These approaches have gained considerable attention and application in the fault diagnosis field [5]. Peng et al. [6] accurately identified wheelset bearing faults in noisy environments using residual learning and a wide convolutional kernel. Jin et al. [7] diagnosed axlebox bearing faults through a deep belief network combined with a global optimization algorithm. Wang et al. [8] introduced a self-supervised representation learning-based framework to boost diagnostic performance, particularly with limited labeled data, demonstrating its effectiveness on train bearing datasets.

Although the application of intelligent diagnostic methods has greatly advanced the field of train bearing fault diagnosis, their training process relies on sufficient and balanced training data, which is hardly achievable in the monitoring system of trains. First of all, during train operation, due to the high priority given to safety, trains are usually operated in a normal condition with few faults occurring. Even if a few faults occur, the train is immediately shut down for maintenance, resulting in far fewer fault samples than normal samples that can be collected. Additionally, the occurrence probability of each fault type differs, leading to significant data imbalance across fault modes [9]. Training the model on such unbalanced data causes it to prioritize the diagnostic accuracy of the majority class, often resulting in higher misclassification rates for the minority class. Unfortunately, the fault class, which is a minority class, is more important in real-world operation, and its occurrence can seriously affect the safe operation of trains [10, 11].

To address this issue, numerous scholars have conducted extensive research and proposed targeted solutions, which can generally be categorized into data-level and algorithmic approaches [12]. Data-level methods mainly utilize under-sampling of minority classes [13], over-sampling of minority classes [14], hybrid sampling [15], and data generation methods [16] to construct a balanced dataset for the subsequent diagnosis process, thus reducing model training bias. Zhu et al. [17] developed an innovative over-sampling technique to effectively expand fault data. Li et al. [18] generated fault samples using a Wasserstein generative adversarial network to rebalance the data set. However, the effectiveness of these methods depends on the original data distribution, which may lead to erroneous noise samples when fault samples are scarce, and the additional samples can increase the model's training load. At the algorithmic level, researchers have sought to improve the model's attention to minority classes during training by assigning appropriate class weights to different health states. Geng et al. [19] introduced an unbalanced weighted cross-entropy based on class sample sizes to enhance fault diagnosis for train bogies. Zhao et al. [20] rebalanced training by assigning higher weights

to harder-to-train classes, while Deng et al. [21] incorporated evaluation metrics into the loss function to improve imbalanced fault diagnosis. Despite progress, these methods rely heavily on manual weighting based on the prior data distribution, which limits their generalizability across tasks. Furthermore, due to the scarcity of fault samples, models still face challenges in extracting distinct features from minority classes.

To tackle these challenges, this paper proposes a novel cost-sensitive multiscale feature multi-order fusion network (CSMFMOFN) for diagnosing train bearing faults under conditions of data imbalance. This network incorporates a lightweight, stable feature extraction module to thoroughly capture fault information in minority classes. Additionally, a novel cost-sensitive loss function, based on a posteriori prediction probability, is developed to enhance the model's diagnostic accuracy and generalization performance under data imbalance. The main contributions of this paper are as follows:

1. A lightweight and stable performance superior feature extraction and classification module (SFECM) is constructed. It extracts multiscale features with complementary performance in the signal through a lightweight multiscale convolutional layer, and realizes the higher-order fusion of multiscale features through the hornet, which effectively enhances the deep-mining capability of the limited diagnostic information of minority classes.
2. From the perspective of a posteriori prediction probability, a novel balanced softmax loss is constructed, which adaptively assigns appropriate weight penalties to each category. This rebalances the contribution of each category during model training, significantly improving diagnostic performance under data imbalance conditions.
3. Experiments across various imbalance ratios and data types validate the proposed method, demonstrating optimal diagnostic performance. This approach provides an effective end-to-end solution for fault diagnosis in trains under data imbalance scenarios.

2 Proposed Method

2.1 The Flowchart of the Proposed Method

This section details the structure and application flow of the proposed CSMFMOFN. The overall architecture, illustrated in Fig. 1, consists of four main steps:

1. Data acquisition and preprocessing. First, the raw vibration data of the equipment is collected from the train wheelset, and the data is min-max normalized. Then, the dataset is split into an unbalanced training set and a balanced test set.
2. Superior feature extraction and state recognition. An efficient, lightweight and stable SFECM is applied for multiscale feature extraction, fusion, and state recognition on the training data.
3. Cost-sensitive learning. Model parameters are optimized using a balanced softmax loss based on the posterior prediction probability, and the contribution of each category is adjusted adaptively during the model training process to achieve cost-sensitive adaptive weighting.
4. Model testing. The trained model is evaluated on a balanced test set to accomplish the task of fault diagnosis in data imbalance distribution.

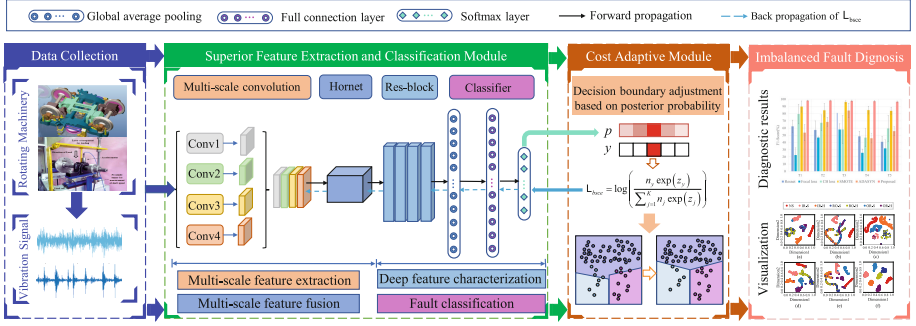


Fig. 1. The overall architecture of CSMFMOFN.

2.2 Superior Feature Extraction and Classification Module

Hornet. The base module of the hornet is the recursive gated convolution ($g^n\text{Conv}$) [22]. As shown in Fig. 2(a), $g^n\text{Conv}$ is constructed using linear projection, standard convolution, and elementwise multiplication, and implements an input adaptive spatial mixing function similar to self-attention through recursive operations to enable higher-order spatial interactions between different features. Assuming the input is $x_I \in \mathbb{R}^{WH \times C}$, we first employ ϕ_I to obtain a series of linearly projected features p_0 and $\{q_k\}_{k=0}^{n-1}$:

$$\left[p_0^{WH \times C_0}, q_0^{WH \times C_0}, \dots, q_{n-1}^{WH \times C_{n-1}} \right] = \phi_I(F_I) \in \mathbb{R}^{WH \times \left(C_0 + \sum_{0 \leq k \leq n-1} C_k \right)} \quad (1)$$

The resulting projection is then subjected to a recursive gated convolution operation, which can be expressed as:

$$p_{k+1} = f_k(q_k) \odot g_k(p_k) / \alpha \quad (2)$$

where $k = 0, 1, \dots, n-1$ denotes the interaction order, α scales the output to stabilize the training process, $\{f_k\}$ denotes a set of deep convolutional layers, and $\{g_k\}$ is used for dimensional matching between features of different orders and can be expressed as:

$$g_k = \begin{cases} \text{Identity}, & k = 0, \\ \text{Linear}(C_{k-1}, C_k), & 1 \leq k \leq n-1 \end{cases} \quad (3)$$

Finally, the output from the last gated convolution operation is fed into the projection layer ϕ_O to produce the final result. Each gated convolution increases the order of interaction between features, enabling flexible multi-order feature interaction and greatly enhancing the model's feature representation capacity. Additionally, a series of projected features can be convolved and linked together with just a single f , which can effectively reduce the model's computational burden, supporting a lightweight design. To further improve the model's representational learning ability, we adopt the structure from [23], shown in Fig. 2(b): a 2-layer MLP with GELU activation follows each layer, with LayerNorm (LN) applied before each layer and MLP. Residual connections

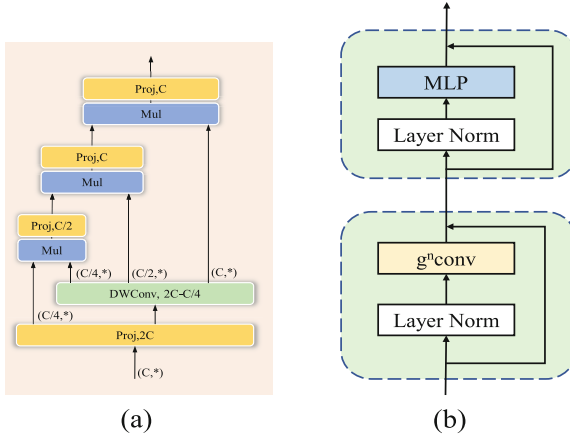


Fig. 2. Schematic diagram of the structure of hornet. (a) Structure of $g^n\text{Conv}$. (b) The overall framework of hornet.

help optimize gradient flow and prevent the degradation of feature map values across successive layers.

Overall Structure. Due to the complex and multifactorial coupling of the monitoring signals of mechanical systems, the features sensitive to mechanical equipment faults may be distributed in different scales. Thus, employing convolutional kernels with varying simultaneous frequency resolutions is crucial for extracting effective information from these signals. Therefore, in SFECM, multiple dilated convolutions with different simultaneous frequency resolutions are firstly connected in parallel as the first layer structure of the network, constituting a lightweight multiscale convolutional layer, which achieves efficient lightweight extraction of rich and complementary multiscale features and helps to alleviate the overfitting problem of minority classes. Subsequently, higher-order interactions and deep fusion of these multiscale features are then achieved through the hornet, enhancing the model's feature mining capability despite limited fault information. Finally, multiple residual blocks are stacked to further obtain deep discriminative feature representations, so as to obtain effective high-dimensional features for subsequent fault classification, and the structure of the residual blocks we use is shown in Fig. 3. The SFECM realizes multi-scale feature extraction and fusion of the original signals, and further extracts the deep discriminative features of the signals and identifies the health status of the equipment by means of the innovative network design. Its specific network parameters are shown in Table 1.

2.3 Cost Adaptive Module

To address the problem of majority class dominance and insufficient learning ability for minority class due to data imbalance during training, unlike the traditional reweighting method based on sample prior distribution, this section designs a novel cost-sensitive loss function based on the predicted probability obtained, from the perspective of posterior

**Fig. 3.** The structure of Res-block**Table 1.** The network parameters of the SFECM

Component	Operation	Parameters	Output shape
Multiscale Conv	Conv_1	4 @ 7×1 , $s = 1$	16×2048
	Conv_2	4 @ 7×1 , $s = 1$, $d = 2$	
	Conv_3	4 @ 7×1 , $s = 1$, $d = 3$	
	Conv_4	4 @ 7×1 , $s = 1$, $d = 4$	
BN + Relu	BN + ReLU	/	16×2048
Pooling	Maxpool	$16 @ 2 \times 1$	16×1024
Hornet	/	$k = 5$	16×1024
Resblock1	Conv + BN + ReLU	$16 @ 3 \times 1$, $s = 1$, $p = 1$	16×1024
	Conv + BN + ReLU	$16 @ 3 \times 1$, $s = 1$, $p = 1$	
Resblock2	Conv + BN + ReLU	$32 @ 3 \times 1$, $s = 2$, $p = 1$	32×512
	Conv + BN + ReLU	$32 @ 3 \times 1$, $s = 1$, $p = 1$	
Resblock3	Conv + BN + ReLU	$64 @ 3 \times 1$, $s = 2$, $p = 1$	64×256
	Conv + BN + ReLU	$64 @ 3 \times 1$, $s = 1$, $p = 1$	
Resblock4	Conv + BN + ReLU	$128 @ 3 \times 1$, $s = 2$, $p = 1$	128×128
	Conv + BN + ReLU	$128 @ 3 \times 1$, $s = 1$, $p = 1$	
Global Avepool layer	Global Avepool	/	128×1
Full connection layer	Linear	128, No. of classes	No. of classes
Softmax layer	Softmax	/	No. of classes

probability, which can effectively weight different classes thereby training the model without bias and mitigating the negative impacts brought by data imbalance.

The most commonly used loss function in deep learning is the cross-entropy(CE) loss, which is well suited to reduce the difference between predicted and true labels on balanced datasets. However, when applied to unbalanced datasets, it often leads to the majority class dominating the training process, resulting in increased misclassification of the minority class due to the same misclassification cost assigned to all classes. To address this issue, we introduce a novel balanced softmax loss [24] to optimize the model's training under conditions of data imbalance. First, we can understand the

softmax function as transforming the features output from the model's fully connected layer into predicted probabilities for each class, i.e.

$$p_k = P(y = k | x) = \frac{\exp(z^k)}{\sum_{i=1}^K \exp(z^i)} \quad (4)$$

where K represents the total number of health states contained in the dataset.

According to the Bayes theorem, it can be further obtained that

$$p_k = P(y = k | x) = \frac{P(x | y = k)P(y = k)}{P(x)} \quad (5)$$

In unbalanced fault diagnosis, it is generally assumed that $P(x | y = k)$ and $P(x)$ agree on the training and test sets, so we focus on the relationship between $P(y = k)$ in the training and test sets. For the balanced test set, the distribution probability of classes is denoted as:

$$\phi_{te} = p(y = k | x) = \frac{p(x | y = k)}{p(x)} = \frac{1}{M} \quad (6)$$

In the unbalanced training set, the distribution probabilities of the classes are

$$\phi_{tr} = P(y = k | x) = \frac{P(x | y = k)}{P(x)} = \frac{n_k}{\sum_{i=1}^K n_i} \quad (7)$$

where n_i represents the number of samples included in the i -th class of health states.

To effectively eliminate this variation in the labeling distribution arising from the data imbalance condition, the balanced softmax loss is defined as:

$$\mathcal{L}_{bsce} = -\log\left(\frac{n_y \exp(z_y)}{\sum_{i=1}^K n_i \exp(z_i)}\right) \quad (8)$$

3 Experimental Verification

To verify the superiority of the proposed method, multiple imbalance diagnosis tasks with different imbalance ratios are constructed on the bearing fault simulation dataset. Meanwhile, the proposed method can achieve state-of-the-art diagnosis performance when compared with multiple advanced methods.

3.1 Data Description

The bearing fault simulation dataset used [25] originates from the Sant Longowal Institute of Engineering and Technology. The structure of the test rig is illustrated in Fig. 4. This test rig allows for adjustments in load and rotational speed to simulate the operating

environment of train bearings. An accelerometer mounted on the test bearing collects three-channel vibration signals. In the experiment, three types of faults are artificially introduced: inner ring faults (IR), rolling element faults (RO), and outer ring faults (OR), with two fault severity levels for each type. Using this dataset, several representative diagnostic tasks with varying imbalance ratios and types are constructed. Details are provided in Table 2, where the imbalance factor β is defined as the ratio of the maximum to minimum sample sizes. Specifically, T1 and T4 are step imbalances, T3 is a nonlinear imbalance, while the data distribution of T2 and T5 is more random, which can simulate more realistic locomotive operation scenarios. A balanced test set is used to better demonstrate the comprehensive performance of the model, and each class in the test set contains 200 samples.

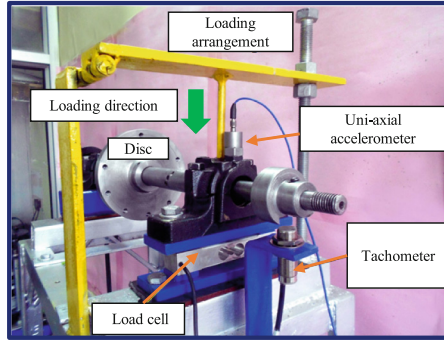


Fig. 4. Diagram of the bearing fault simulation test bench

Table 2. Design of the imbalance diagnostic task

Tasks	β	Train set						
		NS	IR-S	IR-H	RO-S	RO-H	OR-S	OR-H
T1	30:1	1200	40	40	40	40	40	40
T2	80:1	1200	300	60	200	30	15	15
T3	120:1	1200	400	220	280	80	120	10
T4	240:1	1200	80	80	30	30	5	5
T5	240:1	1200	200	20	200	20	10	5

3.2 Comparative Methods and Evaluation Metrics

To quantitatively assess the diagnostic performance of the proposed model, three evaluation metrics including accuracy, macro-F1 and macro-Gmean, are selected to evaluate the model. In addition, five benchmark methods are selected to validate the method's effectiveness. These five methods are Resnet, CB loss [26], Focal loss [20], SMOTE

[27], and ADASYN [28]. For the fairness of the comparison experiments, these five comparison methods adopt the same network structure as the proposed method, but the front-end multi-scale feature extraction and hornet are removed. During model training, all the methods use the stochastic gradient descent algorithm with restart (SGDR) optimizer [29] to optimize the parameters of the model with a training epoch of 400 times and a training batch size of 128.

3.3 Diagnostic Results

Analysis of Classification Results. To minimize the impact of random factors, each model is tested in ten replicate experiments, with results shown in Table 3 and Fig. 5. The table clearly indicates that the proposed method achieves optimal diagnostic results on all diagnostic tasks and has the best stability. Compared to the resnet model without any imbalanced diagnostic strategy, the Focal loss negatively impacts diagnostic performance, likely due to inappropriate weight settings that hinder the model’s representational learning. While CB loss and ADASYN improve performance to some extent, they do not yield consistently positive results across all tasks. Notably, the proposed method achieves a macro-F1 score exceeding 96% across all tasks, with a 12.38% improvement over the best alternative method in Task T4, demonstrating its clear superiority.

Table 3. Results of multiple methods on different diagnostic tasks

Tasks	Indicators	Method(%)					
		Resnet	Focal loss	CB loss	SMOTE	ADASYN	Proposed
T1	Accuracy	70.56	36.78	82.69	90.8	63.09	98.03
	Macro-F1	62.34	22.52	79.42	89.8	53.58	98.01
	Macro-Gmean	69.63	29.79	84.96	92.93	61.16	98.81
T2	Accuracy	64.66	57.01	71.41	86.21	75.02	98.2
	Macro-F1	57.29	46.86	67.32	84.69	68.49	98.19
	Macro-Gmean	63.95	53.23	74.1	89.66	74.42	98.91
T3	Accuracy	84.44	67.26	61.74	96.19	87.7	97.77
	Macro-F1	80.96	57.98	58.09	96.11	84.27	97.74
	Macro-Gmean	85.96	66.32	64.9	97.66	87.63	98.64
T4	Accuracy	58.71	38.49	54.66	85.39	55.86	97.27
	Macro-F1	48.56	25.69	46.68	84.86	45.81	97.24
	Macro-Gmean	57.13	32.48	55.42	90.41	52.81	98.34
T5	Accuracy	51.59	43.86	63.57	85.91	65.91	96.54
	Macro-F1	40.94	31.86	59.43	83.83	55.66	96.5
	Macro-Gmean	48.14	39.49	66.9	89.44	64.15	97.89

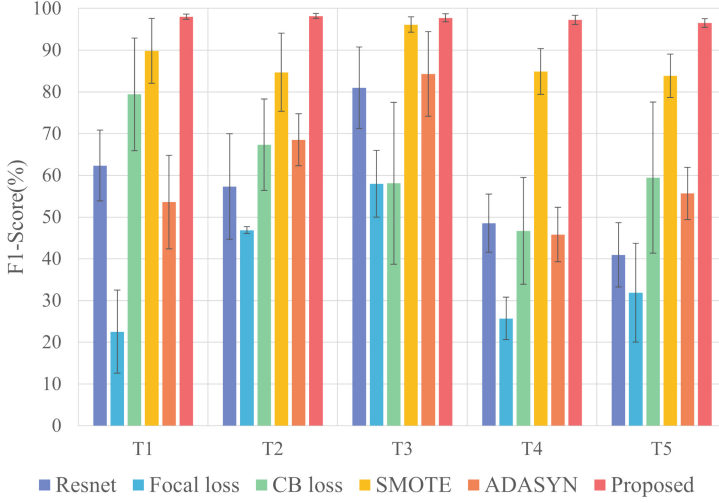


Fig. 5. F1-Score of multiple methods on different diagnostic tasks

Confusion Matrix Analysis. To further analyze the diagnostic capability for each class, the confusion matrix analysis is performed for the six methods under T4, as shown in Fig. 6. The results indicate that the proposed method demonstrates superior diagnostic performance across all fault classes, while the comparison methods have more confusions on minority classes. By implementing an improved cost-sensitive strategy, the proposed method significantly enhances the diagnostic performance for minority classes without compromising the accuracy for majority classes.

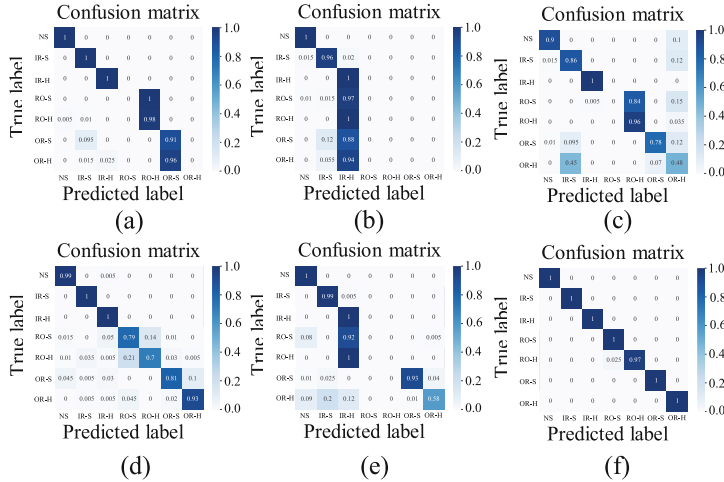


Fig. 6. Confusion matrix analysis under the T4 task. (a) Resnet. (b) Focal loss. (c) CB loss. (d) SMOTE. (e) ADASYN. (f) Proposed.

Feature Visualization Analysis. Under the same task, the model's deep features are visualized using t-distributed stochastic neighbor embedding (t-SNE) [30], with the results shown in Fig. 7. It is very apparent that the proposed method has clear classification boundaries and uniform feature distribution among various categories, while the comparison method has confusion problems among various categories. Through the extraction of multi-scale features and high-order fusion, the model can obtain effective discriminative features for various types of faults and improve separability between fault classes.

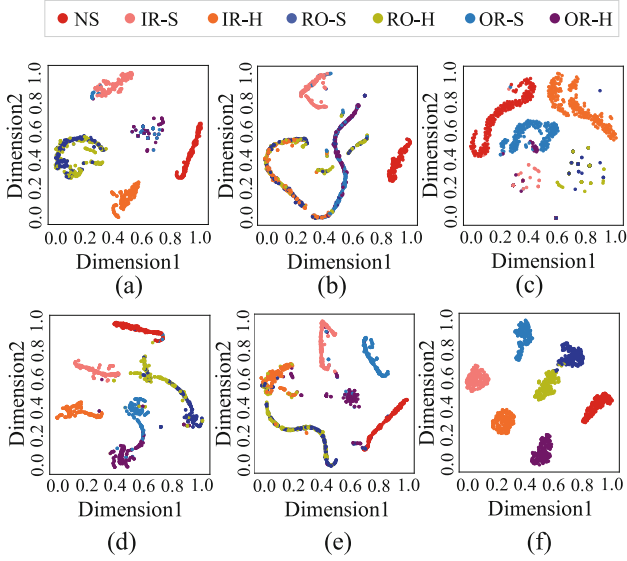


Fig. 7. Feature visualization analysis of different methods under the T4 task. (a) Resnet. (b) Focal loss. (c) CB loss. (d) SMOTE. (e) ADASYN. (f) Proposed.

4 Conclusion

To solve the problem that intelligent diagnosis model is not well-performed in data imbalance scenario, this paper proposes a novel data imbalance fault diagnosis method, CSMFMOFN, which can realize end-to-end fault diagnosis. Specifically, a lightweight and stable feature extraction module is constructed for deep mining fault feature information. Among them, a lightweight multiscale convolutional layer is employed to extract crucial multiscale features, the higher-order fusion of multiscale features is realized by the hornet, and high-performance deep discriminative features are further extracted by stacked multilayer res-block. A novel balanced softmax loss is constructed from the perspective of a posteriori prediction probability, which rebalances the training process by assigning appropriate penalty weights to each category during training. Finally, five sets

of diagnostic tasks containing diagnostic tasks with different imbalance ratios and imbalance types are constructed to systematically validate the proposed method. The results show that the proposed method can achieve satisfactory results on multiple diagnostic tasks, which is helpful for train bearing fault diagnosis in practical complex imbalanced scenarios.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (No. 2023YFB3207902).

References

1. Wei, Z., He, D., Jin, Z., et al.: Density-based affinity propagation tensor clustering for intelligent fault diagnosis of train bogie bearing. *IEEE Trans. Intell. Transp. Syst.* **24**(6), 6053–6064 (2023)
2. Zhong, H., He, D., Wei, Z., et al.: A novel meta-learning network with adversarial domain-adaptation and attention mechanism for cross-domain for train bearing fault diagnosis. *Meas. Sci. Technol.* **35**(12), 19 (2024)
3. Ding, A., Qin, Y., Wang, B., et al.: Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems. *Mech. Syst. Signal Process.* **210**, 15 (2024)
4. Zhang, X., Li, Y., Liu, Y., et al.: Adaptive fault diagnosis and decision-making method based on multi-spectrum evaluation and fusion for traction motor bearings. *IEEE Trans. Instrum. Meas.* **72**, 19 (2023)
5. Lei, Y., Yang, B., Jiang, X., et al.: Applications of machine learning to machine fault diagnosis: a review and roadmap. *Mech. Syst. Signal Process.* **138**, 106587 (2020)
6. Peng, D., Liu, Z., Wang, H., et al.: A Novel deeper one-dimensional CNN With residual learning for fault diagnosis of wheelset bearings in high-speed trains. *IEEE Access* **7**, 10278–10293 (2019)
7. Jin, Z., He, D., Wei, Z.: Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN. *Eng. Appl. Artif. Intell.* **110**, 11 (2022)
8. Wang, H., Liu, Z., Ge, Y., et al.: Self-supervised signal representation learning for machinery fault diagnosis under limited annotation data. *Knowl.-Based Syst.* **239**, 14 (2022)
9. Liu, Y., Li, X., Zhang, X., et al.: Imbalanced deep transfer network for fault diagnosis of high-speed train traction motor bearings. *Knowl.-Based Syst.* **293**, 17 (2024)
10. Chen, Z., Chen, J., Feng, Y., et al.: Imbalance fault diagnosis under long-tailed distribution: Challenges, solutions and prospects. *Knowl.-Based Syst.* **258**, 110008 (2022)
11. Ren, Z., Lin, T., Feng, K., et al.: A systematic review on imbalanced learning methods in intelligent fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 3508535 (2023)
12. Chang, S., Wang, L., Shi, M., et al.: Extended attention signal transformer with adaptive class imbalance loss for Long-tailed intelligent fault diagnosis of rotating machinery. *Adv. Eng. Inform.* **60**, 15 (2024)
13. Qian, M., Li, Y.: A novel adaptive undersampling framework for class-imbalance fault detection. *IEEE Trans. Reliab.* **72**(3), 1003–1017 (2022)
14. Yi, H., Jiang, Q., Yan, X., et al.: Imbalanced classification based on minority clustering synthetic minority oversampling technique with wind turbine fault detection application. *IEEE Trans. Industr. Inf.* **17**(9), 5867–5875 (2020)
15. Swana, E., Doorsamy, W., Bokoro, P.: Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. *Sensors* **22**(9), 3246 (2022)

16. Liu, S., Jiang, H., Wu, Z., et al.: Data synthesis using deep feature enhanced generative adversarial networks for rolling bearing imbalanced fault diagnosis. *Mech. Syst. Signal Process.* **163**, 108139 (2022)
17. Zhu, Q., Zhang, N., He, Y., et al.: Novel imbalanced fault diagnosis method based on CSMOTE integrated with LSDA and LightGBM for industrial process. In: 2022 8th International Conference on Control. Decision and Information Technologies (CoDIT), pp. 326–331. IEEE, Istanbul (2022)
18. Li, Y., Zou, W., Jiang, L.: Fault diagnosis of rotating machinery based on combination of Wasserstein generative adversarial networks and long short term memory fully convolutional network. *Measurement* **191**, 110826 (2022)
19. Geng, Y., Wang, Z., Jia, L., et al.: Bogie fault diagnosis under variable operating conditions based on fast kurtogram and deep residual learning towards imbalanced data. *Measurement* **166**, 108191 (2020)
20. Zhao, X., Yao, J., Deng, W., et al.: Normalized conditional variational auto-encoder with adaptive focal loss for imbalanced fault diagnosis of bearing-rotor system. *Mech. Syst. Signal Process.* **170**, 108826 (2022)
21. Deng, S., Lei, Z., Liu, J., et al.: A cost-sensitive dense network for fault diagnosis under data imbalance. In: 2022 International Conference on Sensing. Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD), pp. 1–6. IEEE, Harbin (2022)
22. Rao, Y., Zhao, W., Tang, Y., et al.: Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Adv. Neural. Inf. Process. Syst.* **35**, 10353–10366 (2022)
23. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30** (2017)
24. Ren, J., Yu, C., Ma, X., et al.: Balanced meta-softmax for long-tailed visual recognition. *Adv. Neural. Inf. Process. Syst.* **33**, 4175–4186 (2020)
25. Kumar, A., Kumar, R.: Vibration and acoustic data for defect cases of the cylindrical roller bearing (NBC: NU205E). IEEE Dataport (2022)
26. Hou, R., Chen, J., Feng, Y., et al.: Contrastive-weighted self-supervised model for long-tailed data classification with vision transformer augmented. *Mech. Syst. Signal Process.* **177**, 109174 (2022)
27. Chawla, N.V., Bowyer, K.W., Hall, L.O., et al.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intel. Res.* **16**, 321–357 (2002)
28. He, H., Bai, Y., Garcia, E.A., et al.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE (2008)
29. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint [arXiv:1608.03983](https://arxiv.org/abs/1608.03983)* (2016)
30. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. mach. learn. res.* **9**(11) (2008)



Research on Secure Storage Architecture of Railway Data Based on Data Lake

Jiang Wu^(✉), Dan Zou, Siqi Sun, Peiran Wang, and Xiaoning Ma

Institute of Computing Technologies, Chinese Academy of Railway Sciences, Haidian District,
Beijing 100081, China

15101100299@163.com, {zoudan, wangpeiran, maxiaoning}@rails.cn

Abstract. With the ongoing advancement of railway information technology, the volume of railway data is steadily increasing. This vast array of diverse data types presents significant challenges to railway data storage. In order to enhance the storage capacity of railway data, this paper proposes a secure storage architecture for railway data based on data lake. Firstly, a data classification standard is established based on business requirements and data types, and a cost-effective, high-efficiency storage scheme is proposed. Based on the scheme, a storage recommendation model suitable for railway application scenarios is constructed, and a data lake storage architecture including a source layer, an integration layer, and a market layer is designed. Then, based on business requirements and security standards, a three-tier data encryption acquisition scheme is proposed. Finally, a data access process based on data lake has been developed to ensure the secure storage of railway data.

Keywords: Data lake · Data storage · Data security

1 Introduction

With the continuous advancement of information technology, the total volume of railway data has gradually increased, reaching the PB level, which brings great challenges to data storage. At present, the primary storage architecture of railway systems is the data warehouse, which can store structured data while simultaneously enabling data processing and analysis. However, the traditional data storage architecture represented by data warehouses is unable to meet the new storage requirements due to the generation of a vast amount of unstructured and semi-structured data. Therefore, it has become an urgent priority to propose a data storage architecture that can meet the diverse requirements of various data storage and realize the needs of data mining and analysis.

In the realm of data storage, numerous scholars have explored the use of data lake architecture across various fields to solve the problems of large volumes of data, diverse data types, and multiple data requirements. Niu et al. proposed a standardized development method for real-time data services in intelligent oil fields based on a data lake architecture [1]. By establishing a multi-level data service management system that includes a sticking layer, dimension layer, and summary layer, they achieved comprehensive

standardization of data services. Ma et al. designed a data lake system for civil aircraft operational support [2], addressing the limitations of traditional databases and data warehouses in managing vast amounts of aviation data. This system meets the demands of exponential data growth, millisecond data processing, and multi-dimensional applications. Xiao et al. implemented a hierarchical design for the data resources of a hospital clinical big data platform [3], categorizing clinical data into four levels based on the data lake architecture: the original data lake, the comprehensive hospital research big data, the specialized disease database, and the research database. This approach facilitates hierarchical management and efficient utilization of data, enhancing both accessibility and operability. Xu et al. utilized data lake technology to establish a unified file directory and data security classification management system for meteorological information [4]. This initiative improves the security management and control of meteorological data, as well as the data mining and analysis capabilities, effectively addressing issues related to data dispersion and the challenges of mounted management.

In the realm of data extraction, privilege management utilizing encryption techniques has emerged as a crucial approach to ensuring data security. The encryption method known as Ciphertext-Policy Attribute-Based Encryption (CP-ABE) has garnered significant interest from numerous researchers. Zhang et al. proposed a hierarchical access control scheme based on an ordered binary decision diagram [5], integrating the ciphertext strategy of CP-ABE to mitigate the risk of data leakage during two-way interactions between users of power marketing big data and power enterprises. Zhao et al. introduced an encryption transmission scheme that combines CP-ABE with Advanced Encryption Standard (AES) [6], achieving efficient and low-consumption security measures by determining the length of the ciphertext and outsourcing decryption. This approach ensures the security of data transmission in industrial control systems. Yuan et al. developed a medical system identity authentication scheme based on CP-ABE and Schnorr signatures [7]. By integrating blockchain technology with the identity authentication mechanism, this scheme ensures that only authorized personnel can access and process sensitive information, thereby addressing the risks of privacy leakage and data security threats in medical data sharing.

To address the aforementioned challenges, this paper proposes a secure storage architecture and an encrypted access scheme for railway data based on a data lake framework. First, a data classification standard is established based on specific needs and data types, leading to the proposal of a cost-effective and highly efficient storage solution. Building on this foundation, a set of storage recommendation models is introduced, accompanied by a three-layer data lake storage architecture. Subsequently, an encryption acquisition solution is recommended in accordance with the established demand standard. Finally, a data access flow utilizing the data lake is proposed, ensuring the secure storage and retrieval of all types of data within the railway sector.

2 Data Storage

Data lake is a new generation of storage mode after data warehouse, which can store any type of data and ensure the integrity of the source data [8, 9]. In contrast to a data warehouse, data lake customizes the data processing and analysis steps according to

the stage of data utilization; specifically, data is processed and analyzed only when it is extracted, thereby enhancing the efficiency of data management. Currently, the commonly used storage methods of data lake can be categorized into two types: single-database storage and multi-database storage.

Single-database storage refers to the utilization of object and document databases for the storage and analysis of source data. This approach maximizes the use of available storage space and ensures the integrity of the source data, as it is stored and stored as a whole. However, research on the analysis of object and document data is still in its preliminary stage, and the research direction is mainly focused on database engine, and the dependence on existing NoSQL type database is large. The development of related functions is not perfect, and the corresponding analysis results have little reference significance and cannot meet business needs.

Multi-database storage refers to the replication and storage of the same database as the source data, enabling the processing and analysis of data while ensuring its integrity. However, due to the absence of relevant standard files, the database types of the source data vary. If the source data is stored according to the principle of using the same type of database, it will not only lead to a waste of storage space but also incur significant costs. At present, existing research on data lakes does not propose a standardized database construction scheme, under the premise of low cost and high efficiency, there is a lack of a database construction framework that meets the specific needs of railway operations. In addition, current research on data lakes primarily focuses on the source data for further processing. However, in the context of the railway industry, much of the data requires standardization or analysis. Utilizing the existing data lake architecture may result in redundant data standardization efforts, leading to significant time and labor costs.

Aiming at the current application problem of data lake in railway, this paper chooses the method of multi-database storage to study, constructs the data classification standard based on business and type, and proposes a low-cost and high-efficiency storage scheme on this basis, introduces the random forest algorithm to recommend the optimal storage path, and designs a three-layer data lake architecture including the paste source layer, the integration layer and the market layer. Finally, the railway data storage architecture based on data lake was formed.

2.1 Railway Data Classification Standards Based on Business and Type

At present, the most influential factors affecting data storage are business requirements and data types. Business requirements dictate the frequency of data access, while data types influence the selection range of the database.

In the context of business requirements, data is typically categorized into hot and cold data based on usage frequency. However, as application scenarios have expanded, hot data has evolved into two distinct categories: warm data and hot data. Cold data generally refers to information that is accessed infrequently, often utilized in data analysis scenarios. In contrast, warm and hot data are accessed more frequently and exhibit certain real-time performance characteristics. In real-time application scenarios, the timeliness requirements for data query operations are increasingly prioritized over those for data modification operations. Consequently, data that demands a higher level of timeliness and is commonly used in query scenarios is classified as hot data, while data with lower

timeliness requirements, typically used in modification scenarios, is classified as warm data.

In terms of data types, data is typically categorized into structured and unstructured data based on whether it can be organized in a tabular format. However, as the variety of data types continues to expand, unstructured data has evolved into two distinct categories: semi-structured data and unstructured data. Semi-structured data includes formats such as email and JSON, while unstructured data encompasses types such as video and images. There are significant differences between the two categories regarding data size.

2.2 Storage Scheme of Railway Data Lake Based on Classification Standard

Based on the classification criteria outlined above, railway data can be categorized into nine types. By considering the characteristics of various databases within the field, we propose a set of data lake storage solutions that fulfill the requirements of railway operations while ensuring low costs and high efficiency.

In terms of business requirements, existing databases can be categorized into two types: Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP). OLAP is primarily utilized for data analysis, while OLTP is designed for real-time applications. OLTP databases must satisfy varying requirements depending on the application scenario, selectively adhering to the Consistency, Partition Tolerance, and Availability (CPA) theory. In scenarios involving data modification, priority should be given to ensuring partition tolerance and consistency. This means that the data for each client must remain consistent before and after any modifications. Conversely, in data query scenarios, partition tolerance and availability take precedence, ensuring that query requests are answered under all conditions. Reason: Improved clarity, vocabulary, and technical accuracy while maintaining the original meaning.

In terms of data types, existing databases can be categorized into two main types: SQL and NoSQL. SQL databases are designed for structured data, while NoSQL databases accommodate semi-structured and unstructured data. To better suit specific use cases, NoSQL databases are optimized for unstructured data types, with examples including Redis and MongoDB. A statistical survey of the databases currently available in the market has been conducted, incorporating a comparison of cost, efficiency, application scenarios, and other relevant indicators. This analysis has led to the identification of a data storage solution that aligns with railway business requirements. The railway data lake storage scheme is shown in Tab 1.

Table 1. The Scheme of Railway Data Lake Storage

	Cold data	Warm data	Hot data
Structured	ClickHouse	PostgreSQL	PostgreSQL
Semi-structured	HDFS	Redis	ES
Unstructured	HDFS	Hbase	Cassandra

2.3 Recommendation Model Based on Random Forest Algorithm

Random forest model is a classical ensemble learning method based on Bagging strategy [10, 11]. It constructs multiple decision trees and combines their results through voting or averaging to obtain the final prediction. The algorithm can be divided into three steps, which are:

1. A portion of the dataset $D1$ is selected from a batch of manually labeled datasets M for training purposes. The data types and business requirements of the data in the training set are chosen as features to construct the corresponding decision tree A .
2. D is reintegrated into M , and then a consistent number of $D2$ samples is repeatedly selected for training. A corresponding decision tree, B , is constructed using the same features.
3. Step 2 is repeated until the number of decision trees reaches the desired forest size, N , at which point the model training is considered complete.

The process of random forest algorithm is shown in Fig. 1. Input the data into the model, select a significant number of judgment results through statistical analysis as the final classification for recommendations, and determine the optimal storage path for the data.

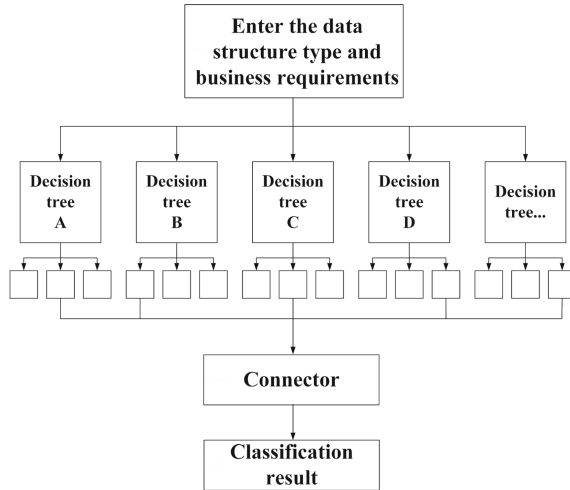


Fig. 1. The process of random forest algorithm

2.4 Hierarchical Architecture of Railway Data Lake

In various application scenarios, the requirements for the extent of data cleaning vary. These can be categorized into three stages: source data, initial data following cleaning and processing, and service data after encapsulation and integration. To address diverse business needs and minimize redundant data cleaning efforts, this paper proposes a three-layer data lake architecture. The Railway data lake architecture is shown in Fig. 2.

- (1) **Source layer.** Responsible for storing and receiving all types of source data. Its function aligns with that of a traditional data lake. When combined with the storage scheme proposed in this paper, the integrity of the source data can be ensured, thereby meeting the demands for high data integrity.
- (2) **Integration layer.** Responsible for mining and cleaning source data according to requirements. The integration layer is a fundamental component of the railway data lake architecture. It significantly reduces the costs associated with user data processing and accommodates the diverse needs of users through standardized processing. Given that the data within the data lake is stored in ciphertext form, operations in the integration layer are conducted using privacy-preserving computing techniques. This approach enables the initial integration and processing of data while ensuring the privacy and security of the information.
- (3) **Market layer.** Responsible for integrating and displaying data. The mart layer functions as an advanced representation of the source layer. When the integration layer performs standard business operations on the source data, it can generate a data collection within the mart layer. The market layer can provide more convenient and efficient data services for regular business, and reduce users' repeated operations.

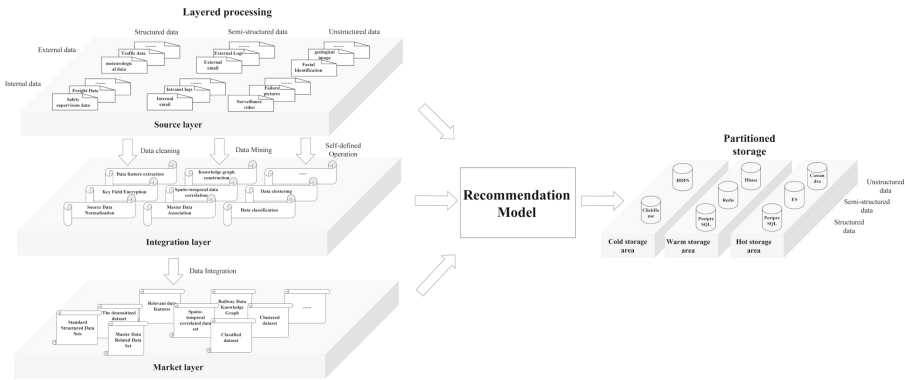


Fig. 2. Railway data lake architecture

3 Data Acquisition

At present, data lake architecture employs a role-based access control model to facilitate flexible access management. However, the data within the data lake originates from various data ownership parties. If the role model is utilized to regulate access to the data, multiple corresponding roles must be created, which escalates the management costs associated with the data lake. At the same time, the significance of various types of data varies. Currently, symmetric encryption methods are commonly employed for the storage and transmission of ciphertext in large data storage systems. However, these symmetric encryption methods necessitate a high level of security for the encryption key; if the key is compromised, the data may be at risk of theft [12]. This paper addresses

the issue of data access security by proposing a series of encryption access schemes tailored to various security levels, aligned with business requirements and data security classification standards. These schemes aim to reduce the costs and risks associated with data management while ensuring secure access to data.

According to the railway business level and the network security classification of the data source system, railway data is categorized into three risk levels: low risk, medium risk, and high risk. Distinct encryption methods are employed for each category of data based on its associated risk level. The hierarchical encryption scheme is shown in Fig. 3.

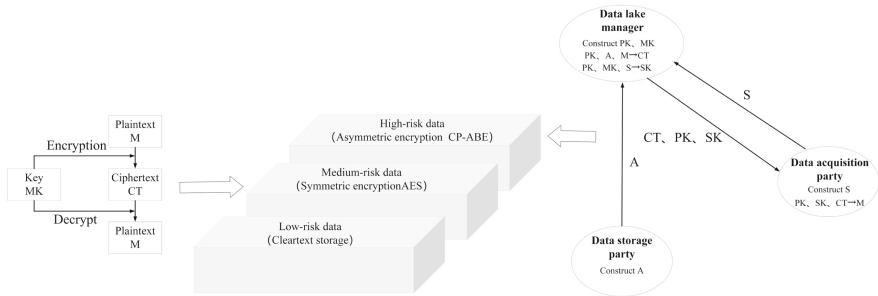


Fig. 3. Hierarchical encryption scheme

Low-risk data refers to information that has a minimal impact on daily business operations and information systems in the event of a breach. Examples include public train schedules and meteorological data. This type of data is typically accessible to the public and serves as reference information and standard indicators within daily business systems. Consequently, due to the ease of access to low-risk data and its limited potential for harm, it can be securely stored in plaintext format.

Medium-risk data refers to information that significantly impacts daily business operations and information systems in the event of a breach. Examples include public works maintenance data and locomotive basic ledger data. This type of data is typically utilized by business personnel and serves as either original or intermediate data within related business systems. Given the challenges associated with obtaining medium-risk data and its substantial influence, the AES symmetric encryption algorithm, known for its high encryption cost, is employed to store the data in ciphertext form. The AES symmetric encryption algorithm utilizes the same key for both encryption and decryption. It consists of three steps, which are:

1. The data storage party generates a key MK utilizing the AES algorithm, encrypts the original data with MK, and produces a ciphertext CT.
2. The data storage party transfers the ciphertext CT and the key MK to the data lake management entity, subsequently storing them within the data lake.
3. The data acquisition party retrieves CT and MK in accordance with its authority, subsequently decrypting the ciphertext to obtain the plaintext.

High-risk data refers to information that has a significant impact on daily business operations and information systems in the event of a breach. This category encompasses sensitive personal information pertaining to passengers or cargo owners, as well as critical details regarding railway infrastructure. Such data is typically targeted at specific business personnel and is deemed essential for particular business functions. Given the challenges associated with acquiring medium-risk data and the substantial consequences arising from its exposure, a dual-layer encryption approach is employed. Initially, AES algorithm serves as the first layer of encryption. Subsequently, CP-ABE algorithm—recognized for its high encryption cost—is applied to secure the ciphertext through double encryption. CP-ABE algorithm operates using distinct keys for both encryption and decryption, thereby enhancing the overall security of the data. It contains five steps, which are as follows:

1. The data lake manager generates the public parameters PK and the master key MK.
2. The data storage party constructs the access structure set A, and sends A to the data lake manager. The data lake manager combines the plaintext M (AES encrypted ciphertext), PK and A to encrypt it together to obtain the ciphertext CT, which is stored in the data lake.
3. The data acquisition party applies to extract the ciphertext CT, constructs its own attribute set S, and sends it to the data lake manager.
4. The data lake manager integrates PK, MK, and S to generate the private key SK for the data extraction party. Subsequently, it transmits CT, PK, and SK collectively to the data extraction party.
5. The data acquisition party decrypts the data by combining CT, PK and SK. If the attribute meets the requirements of the data storage party, the decryption succeeds to obtain the plaintext M, otherwise it fails.

4 Data Access Process

Based on the above data lake storage architecture and data encryption access scheme, this paper proposes a railway data access process based on data lake. The Railway data access process based on data lake is shown in Fig. 4.

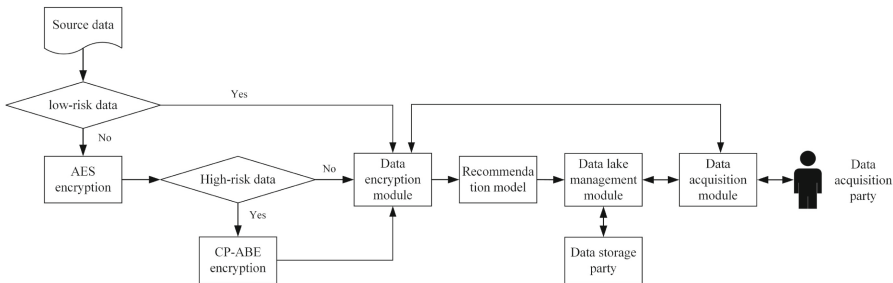


Fig. 4. The Railway data access process based on data lake

Data storage consists of five steps, which are:

1. The data storage party sends the source data to the data lake storage system.
2. The system determines whether the source data is low-risk data according to the data application information provided by the data storage party. If it is, it will directly send it to the data encryption module in the form of plaintext.
3. If not, it will directly send it to the data encryption module in the form of ciphertext. If not, it will perform CP-ABE secondary encryption, and then send it to the data encryption module in the form of secondary encryption.
4. The data encryption module sends the data to the recommendation model after encryption, and the optimal database type is selected by the recommendation model.
5. According to the results of the recommendation model, the data lake management module calls the data lake storage module to store the data into the data lake.

Data acquisition consists of four steps, which are:

1. The data acquisition party retrieves data information according to the data acquisition module, constructs its own attribute set and sends it to the data acquisition module together with the required data information.
2. The data acquisition module sends the required data information and the attribute set of the data extraction party to the data encryption module for judgment, and determines whether it has access permission according to the security level of the required data. If not, the information of extraction failure is fed back to the data extraction party, and the corresponding key is generated and fed back to the data acquisition module.
3. The data acquisition module sends the required data information to the data lake management module, which calls the data storage module to extract the required data, and feeds back to the data acquisition module.
4. The data acquisition module sends the required data and the corresponding key to the data acquisition party, and the source data can be obtained by decrypting the key.

5 Conclusion

This paper addresses the challenges posed by the diverse types and substantial volumes of data in railway storage by proposing a data lake-based architecture for railway data storage. First, a classification standard is developed based on business requirements and data types, accompanied by a cost-effective and efficient storage scheme designed to meet these needs. Building upon this framework, we introduce a storage recommendation model specifically tailored for railway business scenarios, alongside a comprehensive data lake storage architecture that includes source, integration, and market layers. Subsequently, in accordance with business requirements and established data security classification standards, we propose a set of encryption acquisition schemes catering to various security levels. Finally, we present a data access process grounded in the concept of the data lake to facilitate secure storage and retrieval of railway-related information. This study provides valuable insights for future efforts in railway management and holds significant implications for research pertaining to railway data storage.

Acknowledgments. This work is supported by Research on key technologies for the overall planning and implementation of digital railway (N2023S021).

References

1. Niu, Y., Yue, X., Chen, W., Wang, S.: Research on the standardization of real-time data services for smart oilfields using data lake platforms. *Netw. Secur. Data Gov.* **43**(09), 77–83 (2024). (In Chinese)
2. Ma, C.: Design and Implementation of Data Lake for Civil Aircraft Operation Support. *Comput. Meas. Control* **29**(07), 175–179 (2021). (In Chinese)
3. Xiao, N., Cao, L., Ye, Q., Feng, D.: Research on the layered design of data resources for hospital clinical research big data platform. *China Digital Med.* **7**(09), 90–94 (2022). (In Chinese)
4. Xu, J., Liu, X., Xi, X., Zhang, C.: Application of data lake in meteorological information system. *Modern Inf. Technol.* **6**(12), 127–129 (2022). (In Chinese)
5. Zhang, Y., Tian, X., Yan, Y., Lu, G.: A hierarchical strategy based access control scheme for power marketing big data. *J. East China Normal Univ. (Nat Sci Ed)* **05**, 169–184 (2021). (In Chinese)
6. Zhao, D., Shi, L., Xie, Y.: An encrypted transmission scheme for industrial control systems based on CP-ABE. *Comput. Technol. Dev.* **32**(10), 94–99 (2022). (In Chinese)
7. Yuan, Q., Liu, M., Shao, Y., Xia, Y., Wang, Y.: Medical system identity authentication scheme based on CP-ABE and schnorr zero knowledge proof. *J. Qiqihar Univ. (Nat. Sci. Edition)* **40**(05), 16–23 (2024). (In Chinese)
8. Sawadogo, P., Darmont, J.: On data lake architectures and metadata management. *J. Intell. Inf. Syst.* **56**(1), 97–120 (2021)
9. Nargesian, F., Pu, K., Ghadiri-Bashardoost, B., et al.: Data lake organization. *IEEE Trans. Knowl. Data Eng.* **35**(1), 237–250 (2022)
10. Adetunji, A.B., Akande, O.N., Ajala, F.A., et al.: House price prediction using random forest machine learning technique. *Procedia Comput. Sci.* **199**, 806–813 (2022)
11. Jibril, M.M., Bello, A., Aminu, I.I., et al.: An overview of streamflow prediction using random forest algorithm. *GSC Adv. Res. Rev.* **13**(1), 050–057 (2022)
12. Das, S., Namasudra, S.: Multiauthority CP-ABE-based access control model for IoT-enabled healthcare infrastructure. *IEEE Trans. Ind. Inf.* **19**(1), 821–829 (2022)



Research on Unplanned Skip-Stop Scheme for Interconnected Suburban Railway Based on Particle Swarm Multi-objective Optimization Algorithm

Haiwei Jin¹, Wenqian Huang¹, Changyu Chen¹, Ming Xu¹, and Rudong Yang²(✉)

¹ Shanghai Suburban Railway Operation Co., Ltd., Shanghai 201102, China
jinhaiwei@shmetro.com

² CASCO Signal Co., Ltd., Shanghai 200072, China
yangrudong@casco.com.cn

Abstract. The continuous expansion of urban spatial scale has made suburban railways play an increasingly important role in travel within metropolises and agglomerations. With the suburban railway interoperability and networked operation, organization becomes more complex, and the impact range of disturbances becomes larger. Simple solutions like adjusting section running time and compressing station stopping time are insufficient to quickly recover delays. This article aims to quickly derive an unplanned skip-stop scheme for suburban railway trains through the use of the Particle Swarm Optimization (PSO) algorithm. First, this article analyzes the demand for unplanned skip-stop scenarios in suburban railways and defines the model parameters, determining the calculation relationships between the parameters. Then, it constructs an integrated objective function and train operation constraints, considering passenger delay waiting time and delay impact duration. A PSO algorithm is designed to solve the model. The article conducts a case study using the peak passenger flow of the Shanghai Jiamin Line. The results show that the generated skip-stop scheme has high usability and fast iterative convergence, ensuring the quality of passenger travel services and providing a reference for the intelligent upgrade of suburban railway dispatch command systems.

Keywords: Interconnection and interoperability · Suburban railway · Multi-objective optimization · Unplanned skip-stop strategy

1 Introduction

In economically developed regions such as the Yangtze River Delta and Pearl River Delta, suburban railways that connect trunk railways and urban rail transit have ushered in a development boom, providing a key link for passengers within urban agglomerations to choose rail transit for their entire journey. At the same time, in order to improve the convenience of passenger travel in the region, the management department has put

forward new requirements for the interconnection and interoperability of multi-standard rail transit. Some lines in the planned railway network of Shanghai adopt the CTCS-2 + ATO signal system. However, in the scenario of interconnected operation, the operation organization becomes more complex. When disturbances occur, if not dealt with in a timely manner, the operational impact will quickly spread to the entire network. The C2 + ATO standard suburban railway adopts the CTC centralized dispatch system, so when disturbances occur, in order to eliminate the adverse effects of disturbances on train operation, the train dispatcher needs to issue adjustment plans in a timely manner. Dispatchers need to determine skip-stop plans as soon as possible and issue plans for execution in order to restore punctuality and reduce the impact of delays. This requires research on automatic generation methods for unplanned skip-stop plans.

Domestic and foreign scholars have conducted certain research on the problem of train skipping stops. Zhao [1] established a nonlinear integer programming model that minimizes travel time and operating costs by considering dynamic passenger flow demand. Fan [2] established a continuous approximation model for skip-stop operation and developed a solution algorithm using variational method and direct search method. Jihuai Chen [3] proposed an optimization model for the timetable of skip-stop rail transit trains under foreseeable high passenger flow, optimizing the total waiting time for passengers. Qiwei Jiang [4] considered the oversaturation and uneven distribution of passenger flow demand during peak hours of urban rail transit, and proposed optimization methods for skip station operation and multi-station passenger flow collaborative control. Chen [5] and Huang [11] established a nonlinear programming model, and further transformed it into a mixed integer quadratic programming model for solution. Hu [6] applied methods such as rushing, stopping, skipping, and dynamic stopping time to establish a model and design a nested genetic algorithm for solution. Taoyuan Yang [7] constructed a two-layer linear programming model based on skip station operation and multi-station passenger flow control to address the problem of train delay in urban rail transit emergencies.

In summary, there is relatively little research on suburban railways, especially for interconnected suburban railway networks, whose transportation organization, scheduling, and command modes are more complex than urban rail transit and require further research. This article combines the passenger demand, applies particle swarm optimization algorithm, and takes the objective function of minimizing the total additional waiting time and delay recovery time caused by the skip-stop scheme. Considering the constraints of operation, the train skip-stop scheme is solved.

2 Unplanned Skip-Stop of Suburban Railway Trains

Skip-stop [8] refers to the situation where a train does not stop at certain stations as planned during operation, and can be divided into two types: planned skip-stop and unplanned skip-stop. Planned skip-stop refers to an operational plan in which trains that have been determined before operation skip certain stations without stopping. Unplanned skip-stop refers to the decision made by a train during operation to temporarily skip certain stations without stopping due to unforeseen circumstances.

When a train is delayed due to an unexpected event and there is a significant difference from the original schedule, it can lead to uneven running intervals between the train and

the preceding and following trains, resulting in uneven full load rates between trains and passenger accumulation on the platform, leading to the following two demands: 1. Improve the punctuality rate of trains; 2. Reduce passenger waiting time [9, 10].

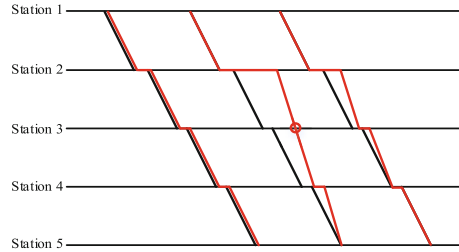


Fig. 1. Schematic diagram of the impact of train skipping on delay recovery

It's difficult to quickly restore the operation order by simply adjusting the interval operation time or shortening the stopping time, which may lead to the expansion of delay scope and the aggravation of propagation. Skip-stop, as a commonly used adjustment method in scheduling, can effectively save the time for train stopping, braking, and starting (as shown in Fig. 1), thereby better achieving delayed recovery. Therefore, this article studies the automatic generation of unplanned skip-stop schemes in the scenario of suburban railway delays.

3 Unplanned Skip-Stop Model for Suburban Railways

3.1 Definition and Calculation of Model Parameters

The definitions of model related parameters are shown in Table 1.

This article aims to study the impact of sudden delays in suburban railway trains on passenger waiting time and on-time recovery time, and establish a multi-objective composite optimization model. Among them, the impact on passenger waiting time is reflected in the fact that adopting different skip-stop combination strategies in delay scenarios can cause delays for passengers who originally got off at the corresponding skip-stop station. The impact on restoring punctuality is reflected in the combination of implementing skip-stop strategies and rush measures after a delay occurs, in order to ultimately recover the delay and restore the time required for the train to be on schedule.

Assuming the maximum passenger capacity C_{max} the train is large enough, without considering passenger retention; For station k , its passenger arrival rate can be expressed as:

$$\delta_k = \frac{E_k}{3600} \quad (1)$$

The number of people inside the train when it departs from each station is not only related to the number of people getting on and off at this station, but also to the number of people inside the train when it departs from the previous station, expressed as:

$$P_{i,j} = P_{i,j-1} + B_{i,j} - E_{i,j} \quad (2)$$

If train i skips to stop at station j , the actual number of people getting off at station $j - 1$ can be expressed as:

$$E_{i,j-1} = h_{i,j-1} + \lambda_1 P_{i,j-2} \quad (3)$$

If train i skips to stop at station j , the actual number of people getting off at station $j + 1$ can be expressed as:

$$E_{i,j+1} = h_{i,j+1} + \lambda_2 P_{i,j} \quad (4)$$

The relationship between the departure time, arrival time, and skip-stop plan of the train at the station can be expressed as:

$$d_{i,j} = a_{i,j} + w_{i,j}(1 - x_{i,j}) \quad (5)$$

3.2 Objective Function

The objective function T in this article consists of two parts: one is to minimize the total waiting time T_1 of passengers delayed due to unplanned skip-stop; the other is to minimize the total time T_2 required to return to the scheduled time under the unplanned skip-stop scheme.

The objective function T_1 is established by minimizing the total waiting time of passengers:

$$\min T_1 = \sum_{i=1}^N \sum_{j=1}^M (a_{i,j} - a_{i-1,j}) \delta_j I_0 x_{i,j} \quad (6)$$

Formula (6) indicates that the travel of some passengers is affected by the train skipping, and these passengers need to wait for the next train on top of their original ride. T_1 is the total extra waiting time caused by the train skipping.

The objective function T_2 is established by minimizing the time from the start of the delay to the restoration of the on-time schedule:

$$\min T_2 = t_e - t_s \quad (7)$$

Among them, the delay start time and the restoration time can be expressed as:

$$t_s = A_{i,j} + t \quad (8)$$

$$t_e = A_{i,j} + \sum_{j=r}^m (t_j^{sec} + w_{i,j} x_{i,j}), j \geq r, a_{i,j} \leq A_{i,j} \quad (9)$$

Normalize the two objective functions of Formula (6) and Formula (7) to obtain the objective function T :

$$\min T = \alpha T_1 + \beta T_2 \quad (10)$$

Table 1. Model parameter definition

Parameter	Definition
N	Delayed train assembly, $N = \{1, \dots, i, \dots, n\}$
M	Line station assembly, $M = \{1, \dots, j, \dots, m\}$
C_{max}	Maximum passenger capacity of the train
δ_j	Passenger arrival rate at station j
I_0	Train planning tracking interval
$P_{i,j}$	The number of people inside train i when it departs from station j
$h_{i,j}$	Normally, the number of people getting off the train i at station j
$w_{i,j}, w'_{i,j}$	The plan and minimum stopping time of train i at station j
$x_{i,j}$	Decision variable, $x_{i,j} = 1$ represents train i skipping stop at station j , $x = 0$ represents train i not skipping stop at station j
λ_1	The proportion choose to get off at the last stop before the skip-stop station
λ_2	The proportion choose to get off at the next stop after the skip-stop station
t_e	Punctuality restoration time
I_1	Minimum departure and arrival interval of train
I_2	Minimum tracking interval of train
I_3	Minimum turnaround time
$B_{i,j}$	Actual number of passengers getting on train i at station j
$E_{i,j}$	Actual number of passengers getting off train i at station j
B_k	Total number of passengers getting on train i at station k
E_k	Total number of passengers getting off train i at station k
$a_{i,j}$	Actual arrival time of train i at station j
$d_{i,j}$	Actual departure time of train i at station j
$A_{i,j}$	Planned arrival time of train i at station j
$D_{i,j}$	Planned departure time of train i at station j
ξ	Train types: 0 indicates national railway offline trains, 1 indicates suburban cross-line trains, 2 indicates urban express trains, 3 indicates urban slow trains
$t_j^{sec}, t_j^{sec'}$	Minimum and standard interval of trains from station j to station $j + 1$

3.3 Constraints

Based on the actual operation requirements of interconnected suburban railways, the generation of skip-stop plans must meet the following constraints:

Consider different types of train skipping constraints for interconnected operation modes:

$$x_{i,j}^{\xi} = \begin{cases} 0; & \xi = 0, 1 \\ 0, 1; & \xi = 2, 3 \end{cases} \quad (11)$$

One train cannot skip two stations consecutively:

$$x_{i,j}^{\xi} + x_{i,j+1}^{\xi} < 2 \quad (12)$$

Two trains cannot stop at the same station consecutively:

$$x_{i,j}^{\xi} + x_{i+1,j}^{\xi} < 2 \quad (13)$$

The first and last stations of the line cannot be skipped:

$$x_{i,1}^{\xi} = x_{i,m}^{\xi} = 0 \quad (14)$$

To ensure the lower limit of train service level, the total number of skip-stop stations of a single train shall not exceed W stations:

$$\sum_{j=1}^m x_{i,j}^{\xi} \leq W \quad (15)$$

The train needs to satisfy the departure and arrival interval constraints:

$$a_{i+1,j} - d_{i,j} \geq I_1 \quad (16)$$

The train arrival time must meet the minimum tracking interval constraint:

$$a_{i+1,j} - a_{i,j} \geq I_2 \quad (17)$$

The train departure time must meet the minimum tracking interval constraint:

$$d_{i+1,j} - d_{i,j} \geq I_2 \quad (18)$$

The train must meet the minimum turnaround time constraint at the final station:

$$a_{i,m} - D_{i,m} \leq d_{i+1,m} \quad (19)$$

4 Algorithm

4.1 Particle Swarm Optimization

This article adopts the particle swarm algorithm, Fig. 2 is the flow chart of the designed particle swarm algorithm.

4.2 Particle Encoding

The particles are encoded in binary format. The length of the particles is equal to the number of stations on the line. $x_{i,j}$ is a binary variable, indicating whether the train skips the station. The specific structure of the particles is shown in Fig. 3.

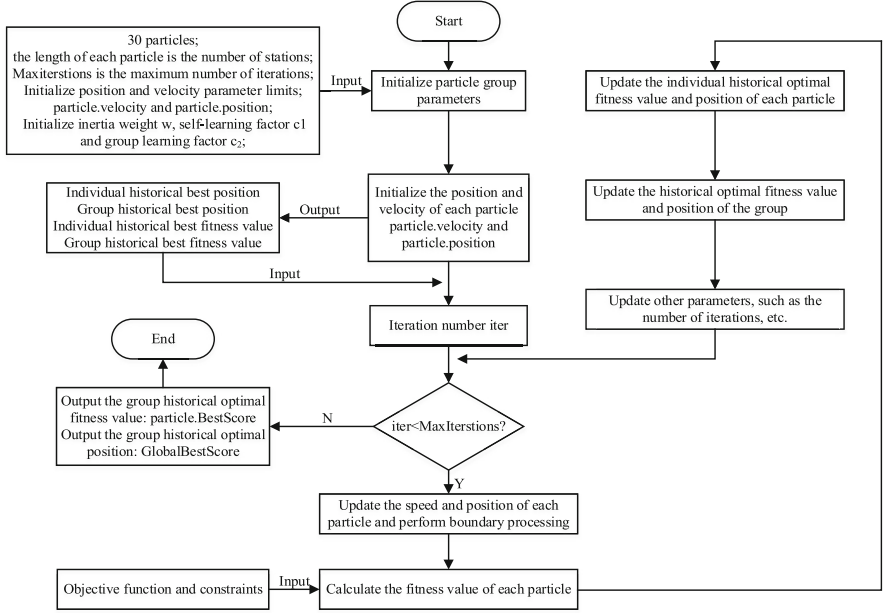


Fig. 2. The flow chart of Particle Swarm Algorithm

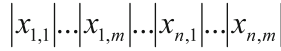


Fig. 3. Schematic diagram of particle structure

4.3 Fitness Function and Particle Swarm Evolution Strategy

The fitness function designed in this article is as follows:

$$Fit(Y) = C - T(X_i) - \sum_k^8 n_k(X_i, Y)q \quad (20)$$

In Formula (20), C is a sufficiently large integer, which represents the number of times the population Y generated based on particle X_i violates the k th constraint condition, and q is a positive penalty factor.

After the particle swarm is initialized, the interface function is used to obtain the input delay site, delay duration and other related parameters, and return them to the position update function. The position update of the particle swarm will check whether the constraints are met, screen out particles that do not meet the constraints and calculate the particle fitness function, and re-generate particles that do not meet the conditions randomly.

The objective function value is calculated based on the new particle position, and if the value is lower than the value of the previous position, it is replaced. The particle position is determined by the individual learning factor and the particle inertia weight, but the learning factor and inertia weight should not be too large, otherwise the particle

swarm will fall into a local optimum. When the particle swarm reaches the maximum number of iterations or meets the exit loop condition, the group is considered to be optimal.

4.4 Train Delay Recovery Strategy

According to the input delay scenario parameters, a combination of delay recovery strategies is performed;

1. On-time operation: by adjusting the interval operation time, stop time and return time, the train is constantly on time;
2. Skip-stop operation: by skipping stops, the number of stops is reduced;
3. Combination strategy: combining the operation on-time and skip-stop schemes to recover on time as soon as possible;

5 Case Analysis

5.1 Basic Data

This article takes the Jiamin Line of Shanghai's suburban railway as an example for analysis, and the data used is according to the preliminary design documents.

5.2 Simulation Verification

This article uses the morning peak OD data of Jiamin Line as input conditions to verify the proposed model and algorithm. Table 2 shows the passenger flow of the morning peak of Jiamin Line in the upward direction from 8:00 to 9:00. Combined with the actual operation of Shanghai Metropolitan Operation Company, the proportion of total delay time and total delay recovery time in the skip-stop recovery strategy is comprehensively weighed. The x_1 and x_2 in the objective function are taken as 0.7 and 0.3 respectively, the sub-initial population is 30, the maximum number of iterations is 100, the inertia weight is 0.5, the individual learning factor is 2, and the social learning factor is 2. The simulation environment is developed using VB.NET programming, and the model is solved to realize the automatic generation of the skip-stop plan according to the user-defined input of the delayed station and delay duration parameters. The generated skip-stop plan includes the stop plan, the recovery station on time, the delay duration and other parameters. Taking the Jiamin Line as an example, the delay scenario is set as a train arriving 8 min late at the 5th station (Nanxiang Station) on the upstream line. After entering the OD data, the program calculates the skipping plan as [0,0,0,0,0,1,0,0,1,0,0,0,0,0], that is, the skipping plan is executed at the 6th and 9th stations, and the train is restored to be on time at the 10th station. The optimal comprehensive objective function value is 33414 s (about 557 min). In this case scenario, the algorithm convergence process of the objective function is shown in Fig. 4. It can be seen that the algorithm converges very quickly and can quickly generate a skipping plan that meets the actual operation and production requirements. Through the train skipping and rushing operation strategy, the delayed train can be brought back to time in a shorter time to avoid further spread and expansion of the delay impact.

Table 2. Morning peak passenger flow OD on the Jiamin Line

Peak Hour	CBX	XCL	JJL	FML	NXZ	JY WL	JYL	TSL	HQ	YB SL	HXL	QB	QXL	XJL	YDL
CBX	0	201	130	103	220	205	408	40	231	13	12	16	111	4	9
XCL		0	90	264	793	252	289	45	212	8	8	18	12	3	7
JJL			0	230	964	470	463	60	385	23	17	32	23	7	11
FML				0	104	115	694	153	270	41	23	26	18	4	20
NXZ					0	959	1507	488	853	141	58	74	60	9	19
JYWL						0	173	102	484	76	48	89	49	13	15
JYL							0	290	733	135	52	248	96	19	77
TSL								0	835	19	34	98	62	15	60
HQ									0	665	142	930	778	99	310
YBSL										0	5	103	65	14	89
HXL											0	180	160	80	142
QB												0	511	490	889
QXL													0	179	1083
XJL														0	360
YDL															0

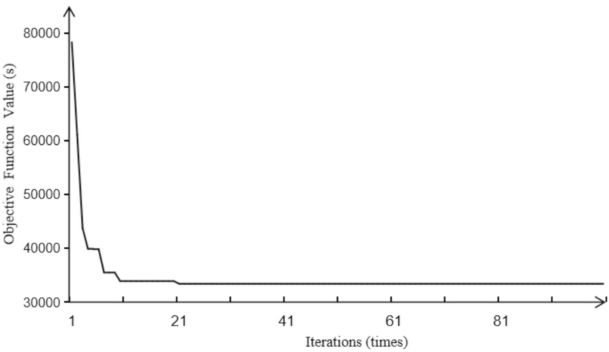


Fig. 4. Iterative graph of the objective function

6 Conclusion

Based on the interconnected operation demand of suburban railway and its adopted centralized dispatching system and national railway dispatching and command mode, this article designs an automatic generation model and algorithm for unplanned skip-stop schemes based on particle swarm multi-objective optimization for operational delays caused by normalized disturbances. The method considers the impact of the skip-stop scheme on passenger travel and the skip-stop constraints of various types of trains in the suburban rail transit, and constructs a comprehensive objective function that considers the

total waiting time for passenger delays and the total time required to return to punctuality. The model and algorithm are simulated and verified in combination with the preliminary design data of the Jiamin Line of the Shanghai Suburban Rail Transit. It can quickly converge to generate a skip-stop scheme with high availability, ensuring the travel service level of passengers when operational disturbances occur, and improving the emergency response efficiency of the suburban rail transit dispatching and command system.

References

1. Zhao, J.R., Ye, M., Yang, Z.Q., et al.: Operation optimizing for minimizing passenger travel time cost and operating cost with time-dependent demand and skip-stop patterns: nonlinear integer programming model with linear constraints. *Transport. Res. Interdisc. Perspect.* **9**(100309), 1–8 (2021)
2. Fan, W.B., Yu, R.: Planning skip-stop services with schedule coordination. *Transp. Res. Part E* **145**, 102–119 (2021)
3. Chen, J.H., Zhao, X.: Timetable optimization of skip-stop rail transit trains considering the predictable large passenger flow. *J. Guizhou Univ. (Nat. Sci. Ed.)* **38**(2), 111–117 (2021). (in Chinese)
4. Jiang, Q.W., Su, J.K., Chen, W.Y.: Collaborative control model of metro passenger flow considering skip-stop strategy. *J. Rail. Sci. Eng.* **18**(11), 2857–2864 (2021). (in Chinese)
5. Chen, Z.B., Li, S.K., D'Ariano, A., Yang, L.X.: Real-time optimization for train regulation and stop-skipping adjustment strategy of urban rail transit lines. *Omega* **110**(102631), 1–22 (2022)
6. Hu, Z.A., Xia, Y.M., Zhen, L.: Research on skip-stop scheme of urban rail transit under train delay. *J. Wuhan Univ. Technol. (Transport. Sci. Eng. Ed.)* **44**(05), 773–778 (2022). (in Chinese)
7. Yang, T.Y., Zhao, P., Yao, X.M., Zhang, P.: Coordinated optimization model of rail transit delayed train stop-skip pattern and passenger flow control. *J. Transport. Syst. Eng. Inf. Technol.* **21**(02), 105–110+132 (2021). (in Chinese)
8. Peftitsi, S., Jenelius, E., Cats, O.: Evaluating skip-stop policy in urban rail transit systems based on passenger cost. *J. Public Transport.* **25**, 100064 (2023)
9. Hu, Y., Li, S., Wang, Y., et al.: Robust metro train scheduling integrated with skip-stop pattern and passenger flow control strategy under uncertain passenger demands. *Comput. Oper. Res.* **115**, 106116 (2023)
10. Yuan, F., Sun, H., Kang, L., et al.: An integrated optimization approach for passenger flow control strategy and metro train scheduling considering skip-stop patterns in special situations. *Appl. Math. Model.* **118**, 412–436 (2023). <https://doi.org/10.1016/j.apm.2023.01.034>
11. Huang, J.Y., Zhang, Q.: Research on train delay skip-stop adjustment in urban rail transit based on virtual coupling technology. *Rail. Transp. Econ.*, 1–8 (2024). (in Chinese)



Adaptive Fault-Tolerant Control for Over-Actuated Ships Based on Safe Reinforcement Learning

Zongdi Tang^{1,2,4}, Yamin Huang^{1,2,3}, Yuanqiao Wen^{1,2,3}, Jie Liu⁵(✉),
and Linying Chen^{1,2,6}

¹ State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology (WUT), Wuhan, China

² National Engineering Research Centre for Water Transport Safety, WUT, Wuhan, China

³ Intelligent Transportation Systems Research Centre, WUT, Wuhan, China

⁴ School of Transportation and Logistics, WUT, Wuhan, China

⁵ Zhejiang Scientific Research Institute of Transport, Hangzhou, China
jie22073@gmail.com

⁶ School of Navigation, WUT, Wuhan, China

Abstract. This paper explores an adaptive fault-tolerant control method for unmanned surface vessels (USVs) equipped with multiple thrusters, which is based on safe reinforcement learning. The objective of the proposed method is to prevent the unexpected behavior of the USV and the subsequent accident due to the failure of one or more thrusters. Firstly, a framework of safe deep reinforcement learning for controlling USV is proposed; secondly, a disturbance network that impacts certain thrusters during specific scenarios is introduced to the environment module, which could simulate various failures on thrusters; thirdly, the control strategies of the USV are trained to track desired speed and course, and a safety constraint function is introduced to ensure the safety of USV during changes on speed and course. Lastly, the Line of Sight (LOS) guidance principle is applied to guide the USV to the designed trajectory. To validate the proposed method, the USV with four thrusters is employed in simulation, and scenarios with full-functional thrusters, one-thruster failure, and two-thruster failure are simulated. Compared with PID controller designed by pole placement method, the proposed method would help the USV track the trajectory when thrusters fail, ensuring both the safety and stability of the USV.

Keywords: Fault-tolerant control · Deep reinforcement learning · Safety constraints · Trajectory tracking

1 Introduction

With the development of artificial intelligence, unmanned vehicle technologies, etc., Unmanned Surface Vessels (USVs) have received much attention in recent years, which are expected to play an increasingly crucial role in maritime missions. For example, the

Korea Institute of Ocean Science and Technology utilizes USVs for marine pollution and ecological monitoring; the Zhejiang Scientific Research Institute of Transport utilizes USVs for channel inspections, identifying potential damage points in bridges and embankments. To extend the functions of USV in practice, the flexibility of the USV is the primary consideration. Figure 1 shows some over-actuated USVs with multiple thrusters in the practice.

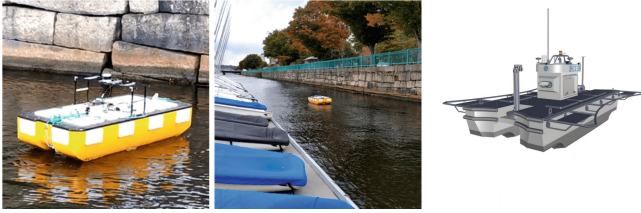


Fig. 1. Unmanned surface vessel with multiple thrusters (the first picture is from[1]).

Since the working environment of the USV is complex, the USV meets various challenges, and the failure of thrusters is one of the common incidents on USVs. For instance, the USV might lose its propellers when the ship grounds or collides with an obstacle, the propeller might be stuck by floating objects, the motor might shut down due to the overload protection, etc. In these cases, the behaviour of the USV might be unexpected and the subsequent accidents could be inevitable.

To fill the gaps, the fault-tolerant control for USVs is necessary. However, few USVs could handle the changes in the input degree of freedom, and fault-tolerant control for multi-output USVs remains limited. Hence, this paper explores multi-agent safe reinforcement learning for multi-output USVs, which contains PPO-Lagrangian training with LOS guidance for accelerated learning, Lagrange-constrained safety limitations, and periodic side thruster failure simulations during training.

Deep Reinforcement Learning (DRL), merging reinforcement learning with deep neural networks, achieved a breakthrough with DeepMind's 2013 Deep Q-Network (DQN) [2], which enabled strategy learning from high-dimensional inputs. DRL has since seen broad applications; for example, Wang et al. [3] proposed a distributed reinforcement learning-based leader-follower USV formation method, enhancing flexibility and scalability. Likewise, Du et al. [4] introduced the SLDDPG algorithm for safe target tracking under task-based safety constraints.

The rest of this paper is organized as follows. Section 2 introduces a fault-tolerant problem, relevant models, and its working environment. Section 3 outlines the steps of the PPO-Lagrangian algorithm, Sect. 4 shows cases, and the conclusion is presented in Sect. 5.

2 Problem Statement

Fault-tolerant control [5] for ship usually refers to the use of certain control strategies and methods to enable a ship to maintain or restore its normal operating state when faults or abnormal conditions occur in the ship system, thereby ensuring the safety and stability of the ship.

In this paper, the focus is on the fault-tolerant control in the case where the two front-side thrusters directly fail, considering three working conditions: no side thruster failure, one side thruster failure, and both side thrusters failure. The aim is to prevent abnormal control strategies of the USV under different working conditions, which could lead to accidents. Particularly for a USV with multiple thrusters, when one thruster fails while others operate normally, it should still be capable of completing the task. However, the key lies in how to execute the task safely. Therefore, fault-tolerant control for thruster failures is of crucial importance for the safe operation of the USV.

2.1 USV Model

The USV model presented in [6] is used, which contains four thrusters and is 0.75 m in length. Two thrusters at the stern offer the thrust at sway direction, and two thrusters at the bow offer surge forces. The layout of the thrusters and coordinates are presented in Fig. 2.

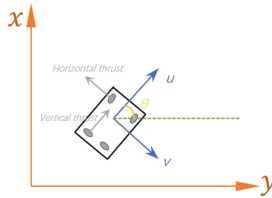


Fig. 2. Diagram of USV Coordinates and relevant forces

The kinematics of an unmanned ship with multiple thrusters is as follows:

$$\begin{cases} \dot{\eta} = J(\psi)v \\ Mv + D(v)\dot{v} = \tau + \tau_w \end{cases} \quad (1)$$

where, position (x, y) and yaw angle (ψ) of the USV in the coordinate system are described by η , the velocity vector (u, v) and yaw rate (r) are described by v :

$$\eta = [x, y, \psi]^T, v = [u, v, r]^T \quad (2)$$

The positive inertia matrix M and the damping matrix $D(v)$ are described as:

$$M = \begin{bmatrix} m_{11} & 0 & 0 \\ 0 & m_{22} & m_{23} \\ 0 & m_{32} & m_{33} \end{bmatrix}, D(v) = \begin{bmatrix} X_u & 0 & 0 \\ 0 & Y_v & Y_r \\ 0 & N_v & N_r \end{bmatrix} \quad (3)$$

In the matrix, m_{11} represents the equivalent mass along the longitudinal direction (surge), m_{22} and m_{33} are the equivalent masses for the transverse (sway) and yaw directions, respectively, m_{23} and m_{32} indicate the coupling effects between the sway and yaw. X_u is the damping coefficient along the longitudinal direction; Y_v and N_r are the damping coefficients for sway and yaw, respectively; Y_r and N_v represent the coupling damping between sway and yaw.

The $\tau = [\tau_x, \tau_y, \tau_r]^T$ denotes the control forces in surge and swag directions and moment, respectively. The $\tau_w = [\tau_{w1}, \tau_{w2}, \tau_{w3}]^T$ contains the disturbances from wind and wave forces in different directions. Thus, the resultant force can be expressed as:

$$\tau = T * n_Square + \tau_w \quad (4)$$

where T represents the thrust influence matrix, and n_Square denotes as a one-dimensional vector composed of the squares of the rotation speeds of each thruster.

2.2 Line of Sight (LOS) for Trajectory Tracking

The principle of navigation through the line of sight is reflected in the intuitive understanding it provides to the helmsman regarding steering and the USV's movement. This principle posits that by maintaining the controlled USV's heading aligned with the line of sight angle (LOS angle) [7], appropriate control can ensure that the vessel reaches the desired position, thereby achieving effective path tracking.

The formula for calculating the target heading angle is as follows:

$$\psi_{ref} = \arctan\left(\frac{y_t - y_c}{x_t - x_c}\right) \quad (5)$$

where (x_c, y_c) represents the current position of the USV, (x_t, y_t) represents the position of the next target point, and ψ_{ref} is the desired heading angle.

To control the moving body's trajectory, we define the lateral error as the perpendicular distance from its current position to the target path. A heading adjustment formula, using lateral error as feedback, can then gradually guide the moving body back to its desired path. The formula is as follows:

$$\psi_{cmd} = \psi_{ref} - \lambda \left(\frac{e_{cross}}{\Delta} \right) \quad (6)$$

where ψ_{cmd} is the adjusted actual heading angle, λ is the control gain used to regulate the magnitude of heading adjustment, and Δ is the look-ahead distance, typically set to twice the vessel's length. e_{cross} Denoted as the lateral error

2.3 PID Pole-Placement for Trajectory Tracking

The core idea of pole placement is to select controller parameters such that the closed-loop system's eigenvalues (poles) are positioned at desired locations on the complex plane, allowing precise control over the system's dynamic performance. Our designed K_p , K_i , and K_d are as follows:

$$K_p = wn * wn * M, K_i = 2 * \zeta * wn * M - D(v), K_d = \frac{1}{10} * wn * K_p \quad (7)$$

where ω_n represents the natural frequency, and ζ denotes the damping ratio. Thus, the control input is:

$$u = -K_p * e - K_d * e_v - K_i * e_{int} \quad (8)$$

where e represents the position error term, e_v the velocity error term, and e_{int} the accumulated error.

3 Deep Reinforcement Learning Control Methods

3.1 PPO-Lagrangian Algorithm

The traditional PPO (Proximal Policy Optimization) algorithm, introduced by OpenAI in 2017 [8], is a deep reinforcement learning algorithm that can be applied to reinforcement learning in high-dimensional continuous action spaces, like the DDPG (Deep Deterministic Policy Gradient) algorithm. Traditional reinforcement learning methods are generally divided into Q-value-based and policy-gradient-based approaches, with PPO being a variant of the latter. It is highly sensitive to parameters and operates as an on-policy strategy, meaning the agent used for training is the same as the one interacting with the environment.

Since PPO's high sensitivity to parameters, it can lead to poor policy stability, which poses significant safety concerns for USV. Therefore, a constrained Markov decision process (CMDP) is introduced based on the original Markov decision process (MDP) in PPO, and decision space is represented by a seven-tuple $(S, A, p, R, c, d, \gamma)$, consisting of the state space (S), action space (A), state transition probabilities (p), rewards (R), costs (c), cost threshold (d), and discount factor (γ).

(1) Constrained Markov decision process

The CMDP in PPO can be represented in Fig. 3. The agent selects an action a from the action space A and executes it according to the current environmental state. Upon executing this action, the agent transitions to the next state S_{t+1} in the environment. At each step, the agent receives both rewards and costs and seeks to maximize the reward while ensuring the cost does not exceed the constraint threshold. This kind of design turns the original problem into a constrained optimization problem.

In this process, the safety constraint function plays a crucial role, making its appropriate design one of the key factors affecting training efficiency.

A complete time step in the training process is illustrated in Fig. 3:

To ensure safety in this context, we establish safety constraint functions concerning changes in heading and variations in revolutions per minute (RPM) at each time step. The overall safety constraint function is represented as:

$$C_{total} = E_{(s_t, a_t) \sim \pi_\theta} \left[\gamma \sum_{k=0}^{\infty} c_{t+k} \right] \quad (9)$$

where, to satisfy the safety constraints, C_{total} must not exceed the cost threshold d . Under this condition, The reward function for the entire training process is:

$$\hat{R}(\lambda_i, s, a) = R(s, a) - \lambda_i C_{total} \quad (10)$$

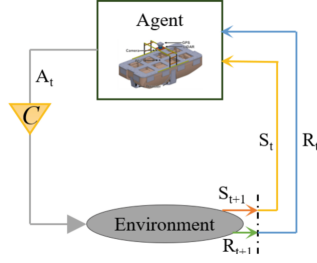


Fig. 3. A complete timestep

where, λ_i represents the Lagrange multiplier, initially set to 1, and $R(s, a)$ denotes the reward given the state S and action a .

(2) Proximal Policy Optimization

The proximal policy optimization (PPO) algorithm utilizes the gradient for optimizing the policy in the policy network under the constraints on the range of policy gradient. In the Trust Region Policy Optimization (TRPO) algorithm, a trust region could be introduced by KL divergence, which has been proven to enable stable training. Similarly, PPO's trust region can also be constrained by KL divergence, but a more popular approach is to control it through clipping, which prevents excessively large policy parameter updates.

Given an agent, the state observed in the current environment updates the known previous policy. The current policy θ is learned, and the ratio $r_t(\theta) = \frac{p_\theta(a_t|s_t)}{p_{\theta_{old}}(a_t|s_t)}$ between the new and old policies is considered. Therefore, the objective function of the proposed PPO can be expressed as follow [9]:

$$L(\theta) = E \left[\min(r_t(\theta)\hat{A}, \text{clip}(r_t(\theta), \frac{1}{1+\epsilon}, \frac{1}{1-\epsilon})\hat{A}) \right] \quad (11)$$

where, $L(\theta)$ is the objective function of PPO, where E can be regarded as the expected return value. $\hat{A}$ represents the advantage function, and ϵ is a small positive number used to define the size of the trust region.

To enhance sample efficiency, we apply importance sampling of policy gradients in this context. Suppose there is a function $f(x)$, and the data is sampled from the distribution p . If we need to sample from an arbitrary distribution q , Therefore, we need to transform the original expression:

$$E_{x \sim p}[f(x)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = E_{x \sim q(x)} \left[f(x)\frac{p(x)}{q(x)} \right] \quad (12)$$

Thus, by introducing importance sampling, the objective function needs to be modified as follows:

$$L^{\theta'}(\theta) = E_{(s_t, a_t) \sim \pi_{\theta'}} \left[\frac{p_\theta(s_t, a_t)}{p_{\theta'}(s_t, a_t)} A^{\theta'}(s_t, a_t) \right] \quad (13)$$

Simultaneously, we need to assess the efficacy of the current state-action pair. To determine whether to increase or decrease the probability of this action, we can calculate the advantage function, as detailed below.

Calculation of the State-Action Function:

$$Q^p(s_t, a_t) := E_{S_{t+1}:\infty, a_{t+1}:\infty} \left[\gamma \sum_{k=0}^{\infty} r_{t+k} \right] \quad (14)$$

The state-value function is:

$$V^p(s_t) := E_{S_{t+1}:\infty, a_{t+1}:\infty} \left[\gamma \sum_{k=0}^{\infty} r_{t+k} \right] \quad (15)$$

Thus, the advantage function is:

$$A^p(s_t, a_t) := Q^p(s_t, a_t) - V^p(s_t) \quad (16)$$

When $A > 0$, the action is considered beneficial, and the probability ratio is increased, though it does not exceed the clipping range. Conversely, when $A < 0$, the opposite is true.

Next, parameter updating is performed by taking the partial derivative of the objective function with respect to the parameter θ , followed by updating the parameters using gradient ascent:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} L(\theta) \quad (17)$$

Here, α represents the learning rate.

3.2 Lagrangian Constraints and Fault-Tolerant Control.

(1) Principle of PPO-Lag

To address this constrained optimization [10] problem, PPO-Lag incorporates the Lagrange multiplier method, integrating the constraints into the objective function. The central idea of the Lagrange multiplier method is to merge the constraints with the objective function, thus creating a new optimization problem. The objective function for policy optimization transforms into:

$$L^{\theta'}(\theta, C) = L^{\theta'}(\theta) - \sum_{i=0}^{\infty} \lambda_i (E_{\pi}[C_i] - d_i) \quad (18)$$

After updating the policy, the Lagrange multiplier λ_i is adjusted based on the extent of constraint violations. Specifically, if the i -th constraint is violated, the Lagrange multiplier λ_i increases, emphasizing this constraint more in the next optimization step. Conversely, if the constraint is satisfied, the Lagrange multiplier decreases.

The update rule for the Lagrange multipliers is as follows:

$$\lambda_i = \max(0, \lambda_i + \eta (E_{\pi}[C_i])) \quad (19)$$

Here, η is the learning rate, which controls the step size for updating the Lagrange multipliers. The Lagrange multiplier λ_i remains non-negative to ensure that the constraints are not excessively relaxed.

(2) PPO-Lag for USV

For multi-thruster USVs, thruster failure during a mission can challenge the controller's capacity to maintain safe vessel control. Issues such as equipment malfunction or motor overheating may also reduce or cut thruster power, endangering vessel safety. Thus, fault-tolerant control is crucial. Through a disturbance network in training, one or two thrusters face random disruptions in certain episodes, leading to a complete halt in rotational speed.

The disturbance network is a simple MLP using the four thrusters' rotational speeds as input. Hidden layers are fully connected with Tanh activation, and the output specifies disturbance on one thruster, two thrusters, or none. To support normal condition training, we keep disturbance frequency low, using a Poisson process with one disturbance every 20 episodes. A detailed schematic is provided below (Fig. 4):

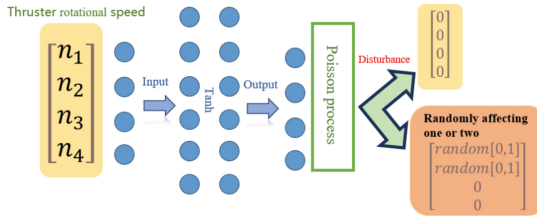


Fig. 4. Approximate design for network interference

Network interference affecting the thrusters significantly impacts the course and speed of the unmanned vessel. By implementing safety constraint functions, we aim for the agent to rapidly adapt its strategy, minimizing sudden changes in direction and speed. Even with only two or three thrusters active, the system should maintain a safety level and tracking precision comparable to four thrusters. However, due to model limitations, full control is lost if both main aft thrusters fail, so our discussion is limited to scenarios with the failure of the two forward thrusters.

The overall control flowchart is as follows:

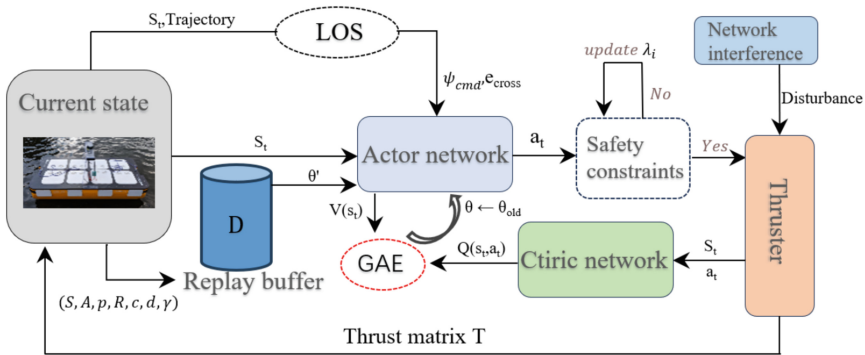


Fig. 5. Control Block Diagram

4 Experiment

4.1 Setup

The experimental setup utilizes a GEFORCE RTX 3070 Ti and an i7-12700H for reinforcement learning training. The scenario involves tracking a trajectory defined by the points (0,0), (15,15), (30,20), and (39,30) under slight wind and wave disturbances, see the dot-orange line in Fig. 5.

Scenario 1: all thrusters are fully functional

Scenario 2: one side thruster fails

Scenario 3: both side thrusters fail

The control group (CG) uses a pole-placement PID combined with LOS for trajectory tracking. The experimental group (EG) employs the following reinforcement learning parameters: a total of 20 million time steps, a discount rate of 0.99, an initial action distribution of 0.6, linearly decaying to 0.05, with a minimum action distribution of 0.2, and a clipping factor of 0.2. The learning rate for the actor-network is set at 0.0003, while for the critic network, it is 0.001.

4.2 Results

The trajectories of USV in different scenarios using the pole-placement PID (CGs) are shown in Fig. 6, and the results of using PPO-lag (EGs) in the same scenarios are shown in Fig. 7. Figure 8 shows the average reward and average cost of the proposed PPO-lag obtained from the safe reinforcement learning training. Figure 9 shows the lateral trajectory tracking errors over time in the boxplot.

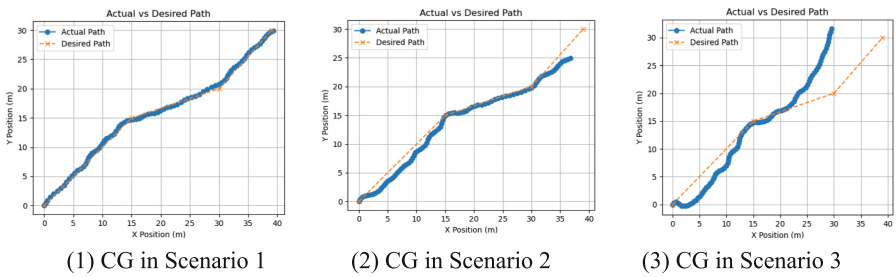


Fig. 6. Trajectory tracking result

The diagram below illustrates the lateral trajectory tracking error, with the absolute value of the error applied to demonstrate the tracking accuracy of the two trajectories.

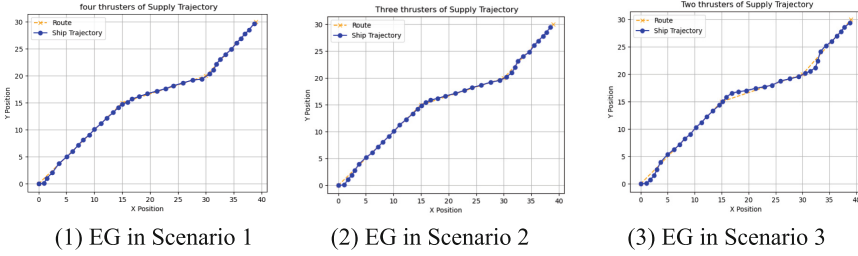


Fig. 7. The trajectory plots are as follows

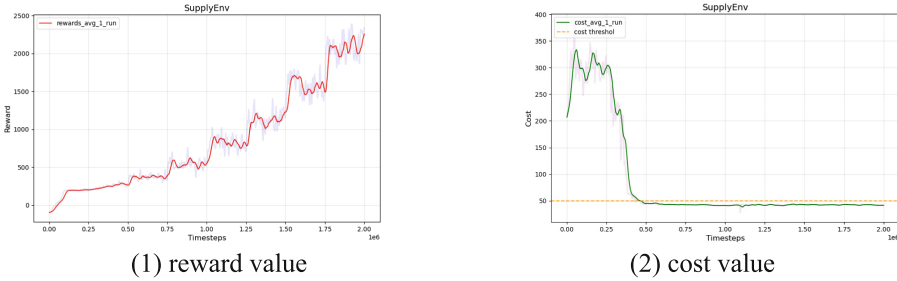


Fig. 8. Graphs of reward and cost values over time.

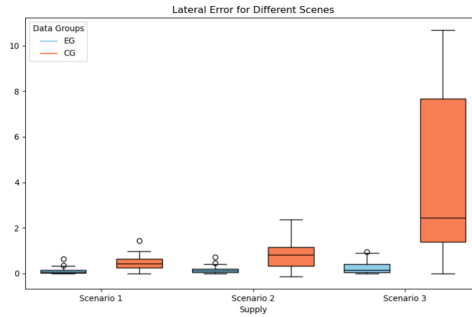


Fig. 9. Tracking errors under different conditions

4.3 Discussion

From Fig. 6, we can see that the designed PID controller could control the USV well when all thrusters are functional, and the average control errors are lower than 1 m. When one of the side thrusters is out of order, the controller xxx. However, when both side thrusters are out of order, the original PID controller might not control the ship to the desired path. Since pole-placement PID depends heavily on model accuracy, any model issues can lead to imprecise control, especially given the small scale of the unmanned vessel.

Figure 7 shows that with all thrusters functioning, the unmanned vessel achieves high trajectory accuracy. With one side thruster failing, errors increase slightly at turns, and with two thrusters down, errors become more pronounced in both turns and straight paths. Although reinforcement learning develops environment-suited strategies, safety constraints prevent the USV from performing maneuvers that might compromise safe operation.

In the PID scheme, side thrusters help reduce lateral errors, but if they fail, lateral control becomes challenging. In contrast, reinforcement learning does not require an accurate model and can develop effective control strategies through iterative trial and error with the main rear thruster.

5 Conclusion

Fault-tolerant control is crucial for the safety of USVs in complex working environments, and a reinforcement-learning method is introduced for controlling over-actuated ships when thrusters fail. Specifically, the PPO-lag method is developed and applied to the fault-tolerant control of the USV, enabling it to obtain appropriate control strategies under varying operating conditions.

To validate the proposed method, simulation with three scenarios, i.e., fully functional scenario, one-thruster-failure scenario, and two-thruster-failure scenario). The pole-placement PID controller is designed and employed in USVs, which is set as the control group; while the USV using the proposed method is set as the experiment group. The experiments show that the experiment group achieves higher tracking precision than the control group when one or two thrusters fail. The results imply that reinforcement learning exhibits a lower dependency on the model, developing effective control strategies through trial and error.

Future work will focus on multi-agent reinforcement learning for USV, aiming to expand action space and scale, enhancing robustness, and safety, and reducing reinforcement learning training time.

Acknowledgement. This work is supported by the Research Program of Sanya City through Grant No. 2022KJCX36, the National Science Foundation of China (NSFC) through Grant No. 52471384, and Research Program of Zhejiang Provincial Department of Transportation through Grant No. 2024006.

References

1. Wang, W., Shan, T., Leoni, P.: Roboat II: a novel autonomous surface vessel for urban environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1740–1747(2020)
2. Mnih, V., Kavukcuoglu, K., Silver, D., et al.: Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015)
3. Wang, S., Ma, F., Yan, X.: Adaptive and extendable control of unmanned surface vehicle formations using distributed deep reinforcement learning. *Appl. Ocean Res.* **110**, 102590 (2021)

4. Du, B., Lin, B., Zhang, C.: Safe deep reinforcement learning-based adaptive control for USV interception mission. *Ocean Eng.* **230**, 109–182 (2021)
5. Yu, W.Z., Xu, H.X., Han, X.: Fault-tolerant control for dynamic positioning vessel with thruster faults based on the neural modified extended state observer. *IEEE Trans. Syst. Man Cybern.-Syst.* **51**(9), 5905–5917 (2016)
6. Fossen, T.I.: Python Vehicle Simulator (2021). <https://www.fossen.biz/pythonVehicleSim>
7. Choi, J., Lee, W.-H., Lee, J.-H., Kim, S.-C.: LOS and NLOS identification in real indoor environment using deep learning approach. *Digital Commun. Netw.* **29**(6), 100913 (2023)
8. Schulman, J.: Proximal Policy Optimization Algorithms, vol. 2, no. 5, pp.99–110 (2017). [arXiv:1707.06347](https://arxiv.org/abs/1707.06347) [cs.LG]
9. Cheng, Y., Huang, L., Wang, X.: Authentic boundary proximal policy optimization. *IEEE Trans. Cybern.* **52**, 9428–9438 (2021)
10. Hu, X., Chen, P., Wen, Y., Tang, B., Chen, L.: Constraints driven safe reinforcement learning for autonomous driving decision-making. *IEEE Trans. Intell. Veh.* **9**(1), 593–612 (2024)



Parallel Path-Following Control of Fully-Actuated Autonomous Surface Vehicles Based on a High-Order Learning Extended State Observer

Jiaxue Xu^{1,2,3}, Haodong Liu^{1,2,3}, Nan Gu^{1,2,3}, Lu Liu^{1,2,3},
and Zhouhua Peng^{1,2,3} (✉) 

- ¹ School of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China
zhpeng@dlnu.edu.cn
- ² Dalian Key Laboratory of Swarm Control and Electrical Technology for Intelligent Ships,
Dalian 116026, China
- ³ State Key Laboratory of Maritime Technology and Safety, Dalian, China

Abstract. Unlike traditional path-following methods, this paper develops a parallel control framework to investigate the path-following behavior of autonomous surface ships (ASVs) in the presence of unknown disturbances and model uncertainties. Specifically, a high-order learning extended state observer is proposed to construct a digital-twin ASV system, enabling precise depiction of the motion characteristics of the actual ASV system. Based on the states and unknown total disturbances estimated by the digital-twin ASV system, a parallel path-following controller is developed to actively guide the ASV to achieve path-following task. Simulation results validate the efficacy of the proposed parallel path-following control scheme for the fully-actuated ASV.

Keywords: Autonomous surface vehicles · high-order learning extended state observer · parallel path-following control

1 Introduction

The ocean harbors abundant natural resources, and conducting efficient exploration, development, and protection of these resources is an essential guarantee for achieving sustainable development for mankind [1–3]. Due to its compact size and high flexibility, the autonomous surface vehicle (ASV) has become one of the important tools for exploring the ocean [4]. It is capable of performing various marine tasks, such as environmental monitoring and data collection, which are widely used in commercial and military fields [5, 6]. Reference [7–9] indicates that motion control under complex marine environmental disturbances faces significant challenges, and there is an urgent need to establish a new theoretical framework for motion control.

Path-following is a fundamental motion control scenario for ASV. In the literature, several motion control schemes are presented for marine vehicle to track the desired

path [10–12]. In [10], a secure path-following control scheme is developed for the ASV that enables it to track the given path under simultaneous cyber and physical attacks. In [11], static event-triggered mechanisms are designed to accomplish the path-following task with constrained communication resources. In [12], an aperiodic communication path-following problem is investigated to reduce the effects of sensor faults for under-actuated ASV. However, the above control strategies cannot characterize the motion states of ASV in virtual space. In contrast, parallel control technology, which is a powerful control paradigm proposed in references [13–15], has been applied to the marine ship motion control [16]. By integrating feedback information from both virtual and actual environments, parallel control technology can more effectively handle complex dynamic interactions, thus demonstrating unique advantages in vehicle motion control.

Inspired by the above research results, this paper introduces a parallel path-following control scheme for a fully-actuated ASV subject to model uncertainties and unknown uncertainties. At first, a high-order learning extended state observer (HL-ESO) is designed to construct a digital-twin ASV system, achieving characterization of the actual ASV system's motion characteristics in virtual space. In addition, a parallel path-following controller, which utilizes the system states and unknown total disturbances information estimated by the HL-ESO, is developed to actively guide the ASV in tracking the specified path.

2 Problem Formulation and Preliminaries

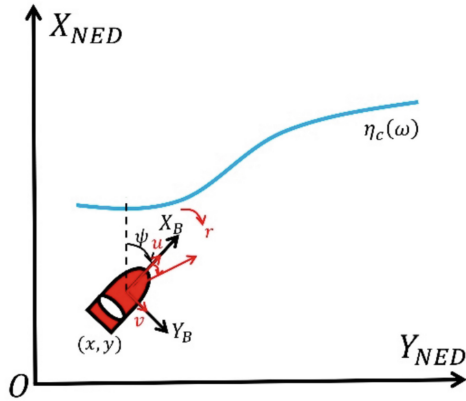


Fig. 1. A geometrical illustration of the path-following for the fully-actuated ASV.

Figure 1 present a geometric schematic illustrating the path-following problem of a fully-actuated ASV. The kinematics and kinetics of the ASV are expressed as follows:

$$\begin{cases} \dot{\eta} = R(\psi)v \\ M\dot{v} = g(v) + w(t) + \tau_v \end{cases} \quad (1)$$

where $\eta = [x, y, \psi]^T \in \mathbb{R}^3$ denotes the position and orientation in the earth-fixed reference frame $\{X_{\text{NED}} - Y_{\text{NED}}\}$, $v = [u, v, r]^T \in \mathbb{R}^3$ denotes the velocity vector in the body-fixed frame $\{X_{\text{B}} - Y_{\text{B}}\}$. $R(\psi) \in \mathbb{R}^{3 \times 3}$ and $M \in \mathbb{R}^{3 \times 3}$ are the rotation matrix and inertia mass matrix, respectively. The vectors $g(v) \in \mathbb{R}^3$, $w(t) \in \mathbb{R}^3$ and $\tau_v \in \mathbb{R}^3$ denote the internal model uncertainties, external environment disturbances, and control input of the system, respectively.

3 Parallel Path-Following Control

Unlike traditional path-following control methods, a parallel path-following control structure for a fully-actuated ASV is designed subject to external environment disturbances and internal model uncertainties, as illustrated in Fig. 2. Parallel control integrates advanced information technologies such as sensing, computing, communication, and control, enabling the mutual mapping and efficient collaboration between physical space and information space.

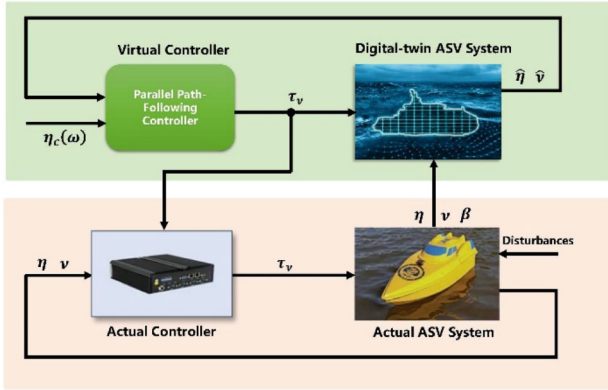


Fig. 2. The parallel path-following control architecture.

3.1 Digital-Twin ASV System

The kinetics of actual ASV system (1) can be rewritten as follows:

$$\dot{v} = f + \underline{M}^{-1} \tau_v \quad (2)$$

where

$$f = M^{-1}(g(v) + w(t) + \tau_v) - \underline{M}^{-1} \tau_v \quad (3)$$

denotes unknown total disturbances, and $\underline{M} = \text{diag}\{m_{11}, m_{22}, m_{33}\} \in \mathbb{R}^{3 \times 3}$ is the nominal diagonal inertial matrix.

Based on the Taylor series theorem and Fourier series theorem, a regressor vector $Y^* = [1, u, v, r, u^2, v^2, r^2, uv, ur, vr, uvr, u^3, v^3, r^3, u^2v, u^2r, v^2u, v^2r, r^2u, r^2v, \sin(\omega t), \cos(\omega t), \dots, \sin(n_1\omega t), \cos(n_1\omega t)] \in \mathbb{R}^{1 \times (20+2n_1)}$ is selected to precisely approximate the unknown total disturbances f . Here, n_1 and ω are the positive constants. Thus, the unknown total disturbances f in Eq. (3) can be transformed into

$$f = Y\vartheta + \varsigma \quad (4)$$

where $\vartheta \in \mathbb{R}^{60+6n_1}$ represents the bounded ideal weight, $Y = \text{diag}\{Y^*, Y^*, Y^*\} \in \mathbb{R}^{3 \times (60+6n_1)}$, and $\varsigma = [\varsigma_u, \varsigma_v, \varsigma_r]^T \in \mathbb{R}^3$ denotes the approximation error. Based on Eq. (2) and (4), it can be obtained as follows:

$$\dot{v} = Y\vartheta + \varsigma + \underline{M}^{-1}\tau_v \quad (5)$$

Assumption 1 ([16]). The acceleration β of actual ASV system, modeling errors $\hat{\varsigma}$ and variable θ are all bounded.

To characterize the kinematics and kinetics of the actual ASV system in the virtual space, a HL-ESO is designed to construct the digital-twin model as follows:

$$\begin{cases} \dot{\hat{\eta}} = R(\psi)\hat{v} - \lambda_1(\hat{\eta} - \eta) \\ \dot{\hat{v}} = Y\hat{\vartheta} + \hat{\varsigma} - \lambda_2(\hat{v} - v) + \underline{M}^{-1}\tau_v \\ \dot{\hat{\varsigma}} = -\lambda_3(\hat{v} - v) \\ \dot{\hat{\theta}} = \lambda_4 Y^T(\beta - \underline{M}^{-1}\tau_v - Y\hat{\theta}) \\ \dot{\hat{\vartheta}} = -\lambda_5(\hat{\vartheta} - \vartheta) \end{cases} \quad (6)$$

where β represents the acceleration of the actual ASV system; the vectors $\hat{\eta} = [\hat{x}, \hat{y}, \hat{\psi}]^T \in \mathbb{R}^3$, $\hat{v} = [\hat{u}, \hat{v}, \hat{r}]^T \in \mathbb{R}^3$, $\hat{\varsigma} = [\hat{\varsigma}_u, \hat{\varsigma}_v, \hat{\varsigma}_r]^T \in \mathbb{R}^3$ and $\hat{\vartheta} = [\hat{\vartheta}_u, \hat{\vartheta}_v, \hat{\vartheta}_r]^T \in \mathbb{R}^{60+6n_1}$ are the estimated states of the vectors η , v , ς , ϑ respectively.

The vector $\hat{\theta} \in \mathbb{R}^{60+6n_1}$ represents the estimated value of the intermediate variable θ . The observed gains $\lambda_1 \in \mathbb{R}^{3 \times 3}$, $\lambda_2 \in \mathbb{R}^{3 \times 3}$, $\lambda_3 \in \mathbb{R}^{3 \times 3}$ are positive definite coefficient matrices. $\lambda_4 \in \mathbb{R}$, $\lambda_5 \in \mathbb{R}$ are positive adaptive gains. Through the design of a digital-twin ASV system, we can not only accurately estimate the actual state of the system, but also effectively approximate the unknown total disturbances as $\hat{f} = Y\hat{\vartheta} + \hat{\varsigma}$. These results will be used in the design of the parallel path-following control law in the next subsection.

Letting $\tilde{\eta} = \hat{\eta} - \eta$, $\tilde{v} = \hat{v} - v$, $\tilde{\varsigma} = \hat{\varsigma} - \varsigma$, $\tilde{\theta} = \hat{\theta} - \theta$ and $\tilde{\vartheta} = \hat{\vartheta} - \vartheta$, it follows from (1) and (6) that:

$$\begin{cases} \dot{\tilde{\eta}} = R(\psi)\tilde{v} - \lambda_1\tilde{\eta} \\ \dot{\tilde{v}} = Y\tilde{\vartheta} + \tilde{\varsigma} - \lambda_2\tilde{v} \\ \dot{\tilde{\varsigma}} = -\lambda_3\tilde{v} - \tilde{\varsigma} \\ \dot{\tilde{\theta}} = \lambda_4 Y^T(\beta - Y\hat{\theta}) - \dot{\theta} \\ \dot{\tilde{\vartheta}} = -\lambda_5\tilde{\vartheta} + \lambda_5\hat{\theta} + \lambda_5(\vartheta - \theta) \end{cases} \quad (7)$$

3.2 Parallel Path-Following Control Law

Define a given path $\eta_c(\omega) = [x_c(\omega), y_c(\omega), \psi_c(\omega)]^T \in \mathbb{R}^3$ and a path-tangential angle $\psi_c(\omega) = \text{atan2}\left(\frac{\partial y_c(\omega)}{\partial \omega}, \frac{\partial x_c(\omega)}{\partial \omega}\right)$, where ω is a time-independent path parameter. The tracking error vector s_e is defined as

$$s_e = \hat{\eta} - \eta_c \quad (8)$$

Based on Eq. (6), the time derivative of s_e is expressed as follows:

$$\dot{s}_e = R(\psi)\dot{\hat{v}} - \lambda_1(\hat{\eta} - \eta) - \eta'_c \dot{\omega} \quad (9)$$

where $\eta'_c = \partial \eta_c / \partial \omega$ and $\dot{\omega} = v_0 - \omega_0$, with v_0 as a reference speed. The variable ω_0 , developed to achieve the path-following task, is defined as follows:

$$\dot{\omega}_0 = -\lambda_\omega(\omega_0 + k_0(\eta'_c)^T s_e) \quad (10)$$

where $\lambda_\omega \in \mathbb{R}$ and $k_0 \in \mathbb{R}$ are positive constants. To stabilize the error dynamic (9), the virtual guidance law vector $\alpha_v = [\alpha_u, \alpha_v, \alpha_r]^T$ is designed as follows:

$$\alpha_v = R^T(\psi) \left(-k_1 s_e / \left(\sqrt{\|s_e\|^2 + \Delta_1^2} \right) + \eta'_c v_0 \right) \quad (11)$$

where $\Delta_1 \in \mathbb{R}$ and $k_1 = \text{diag}\{k_{11}, k_{12}, k_{13}\} \in \mathbb{R}^{3 \times 3}$, with $k_{11} \in \mathbb{R}$, $k_{12} \in \mathbb{R}$, and $k_{13} \in \mathbb{R}$ as constants to be designed.

Letting $v_e = \hat{v} - \alpha_v$, the time derivative of v_e can be expressed as follows:

$$\dot{v}_e = Y\hat{\vartheta} + \hat{\zeta} - \lambda_2(\hat{v} - v) + \underline{M}^{-1}\tau_v - \dot{\alpha}_v \quad (12)$$

To stabilize the error dynamic (12), the parallel path-following controller is designed as follows:

$$\tau_v = \underline{M} \left(-k_2 v_e / \left(\sqrt{\|v_e\|^2 + \Delta_2^2} \right) - Y\hat{\vartheta} - \hat{\zeta} + \dot{\alpha}_v - R^T(\psi)s_e \right) \quad (13)$$

where $\Delta_2 \in \mathbb{R}$ and $k_2 = \text{diag}\{k_{21}, k_{22}, k_{23}\} \in \mathbb{R}^{3 \times 3}$, with $k_{21} \in \mathbb{R}$, $k_{22} \in \mathbb{R}$, and $k_{23} \in \mathbb{R}$ as constants to be designed.

Based on (9)–(13), the error dynamics are as follows that

$$\begin{cases} \dot{s}_e = -Q_1 s_e - \lambda_1 \tilde{\eta} + R(\psi)v_e + \eta'_c \\ \dot{v}_e = -Q_2 v_e - \lambda_2 \tilde{v} + R^T(\psi)s_e \\ \dot{\omega}_0 = -\lambda_\omega(\omega_0 + k_0(\eta'_c)^T s_e) \end{cases} \quad (14)$$

where $Q_1 = -k_1 s_e / \left(\sqrt{\|s_e\|^2 + \Delta_1^2} \right)$, $Q_2 = -k_2 v_e / \left(\sqrt{\|v_e\|^2 + \Delta_2^2} \right)$.

4 Stability Analysis

This section aims to analyze the stability of closed-loop systems.

Lemma 1. According to Assumption 1, the subsystem (7), regarded as a system with states being $\tilde{\eta}$, $\tilde{v}$, $\tilde{\zeta}$, $\tilde{\theta}$, $\tilde{\vartheta}$, and the inputs being $\dot{\zeta}$, $\dot{\theta}$ is ISS, and the error $\tilde{\eta}$, $\tilde{v}$ are bounded.

Lemma 2. The subsystem (14), regarded as a system with states being s_e , v_e , ω_0 , and the inputs being $\tilde{v}$, is ISS.

Proof. Construct a Lyapunov function as.

$$\dot{V}_1 = \frac{1}{2} \|s_e\|^2 + \frac{1}{2} \|v_e\|^2 + \frac{\omega_0^2}{2k_0} \quad (15)$$

Take the time derivative of V_1 , it can be obtained that

$$\begin{aligned} \dot{V}_1 &\leq -s_e^T Q_1 s_e - v_e^T Q_2 v_e - s_e^T \lambda_1 \tilde{\eta} - v_e^T \lambda_2 \tilde{v} - \frac{\lambda_\omega \omega_0^2}{k_0} \\ &\leq -\lambda_{\min}(Q_1) \|s_e\|^2 - \lambda_{\min}(Q_2) \frac{1}{2} \|v_e\|^2 + \lambda_1 \|\tilde{\eta}\| \|s_e\| + \lambda_2 \|\tilde{v}\| \|v_e\| - \frac{\lambda_\omega \omega_0^2}{k_0} \end{aligned} \quad (16)$$

Letting $\Xi_1 = [s_e^T, v_e^T, \omega_0]^T$, $S_1 = \text{diag}\left\{\lambda_{\min}(Q_1), \lambda_{\min}(Q_2), \frac{\lambda_\omega}{k_0}\right\}$ and $h_1 = \lambda_{\min}(\lambda_1) \|\tilde{\eta}\| + \lambda_{\min}(\lambda_2) \|\tilde{v}\|$. It follows that

$$\begin{aligned} \dot{V}_1 &\leq -\lambda_{\min}(S_1) \|\Xi_1\|^2 + h_1 \|\Xi_1\| \\ &\leq -\lambda_{\min}(S_1) (1 - \kappa_1) \|\Xi_1\|^2, \quad \forall \|\Xi_1\| \geq \frac{\|h_1\|}{\lambda_{\min}(S_1)\kappa_1} \end{aligned} \quad (17)$$

where $0 < \kappa_1 < 1$. Due to $\tilde{v}$ is bounded, then the subsystem (14) is input-to-state stability.

Based on the above discussion, it can be concluded that $\tilde{\eta}$ and s_e are bounded. Thus, the path tracking error of the actual ASV is derived as follows:

$$\|\eta - \eta_c\| \leq \|\eta - \hat{\eta}\| + \|\hat{\eta} - \eta_c\| \leq \|\tilde{\eta}\| + \|s_e\| \leq d_0 \quad (18)$$

where $d_0 \in \mathbb{R}^+$ is a bounded constant.

Theorem 1. Under Assumption 1, the system formed by the subsystem (7) and (14) is ISS, and all errors are uniformly ultimately bounded.

Proof. According to Lemma 1 and 2, the subsystem (7) and (14) are input-to-state stability. By Lemma 4.6 in [3], it follows that the cascade system formed by (7) and (14) with states being $\tilde{\eta}$, $\tilde{v}$, $\tilde{\zeta}$, $\tilde{\theta}$, $\tilde{\vartheta}$, s_e , v_e , ω_0 , and inputs being $\dot{\zeta}$, $\dot{\theta}$, is ISS. Due to the inputs $\dot{\zeta}$, $\dot{\theta}$ are bounded under assumption 1, then the error signals of the system are all bounded. Furthermore, the path tracking error of the actual ASV is also bounded.

5 Simulations

In this section, an example is provided to illustrate the effectiveness of the proposed parallel path-following control scheme for fully-actuated ASV. The model parameters of ASV are obtained based on [17], and the initial states of the ASV are chosen as $\eta_0 = [5 \text{ m}, 2 \text{ m}, 3 \pi/4 \text{ rad}]^T$ and $v_0 = [0 \text{ m/s}, 0 \text{ m/s}, 0 \text{ rad/s}]^T$. The observed gains of the digital-twin ASV system are chosen as $\lambda_1 = \lambda_2 = \text{diag}\{10, 10, 10\}$, $\lambda_3 = \text{diag}\{6, 6, 6\}$. The adaptive gains are set as $\lambda_4 = 0.5$, $\lambda_5 = 15$ and $k_1 = \text{diag}\{0.1, 0.1, 0.1\}$. The control parameter is selected as $k_2 = \text{diag}\{1, 1, 1\}$. The simulation results are presented in Figs. 3, 4 and 5. Figure 3 shows that the path-following performance is achieved, which can be further analyzed through the tracking error, clearly indicating that the system has a high tracking accuracy. Figures 4 and 5 show that the constructed digital-twin ASV system retains good estimation capability, demonstrating that it can not only depict the motion state of ASV in virtual space under complex marine environment conditions, but also effectively estimate the unknown total disturbance.

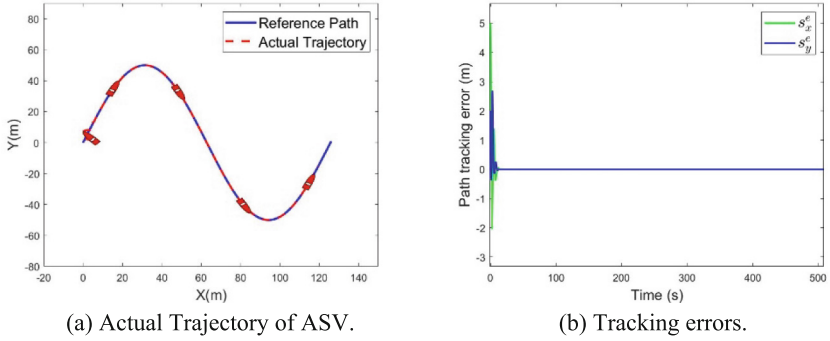


Fig. 3. The path-following performance of the proposed parallel control method.

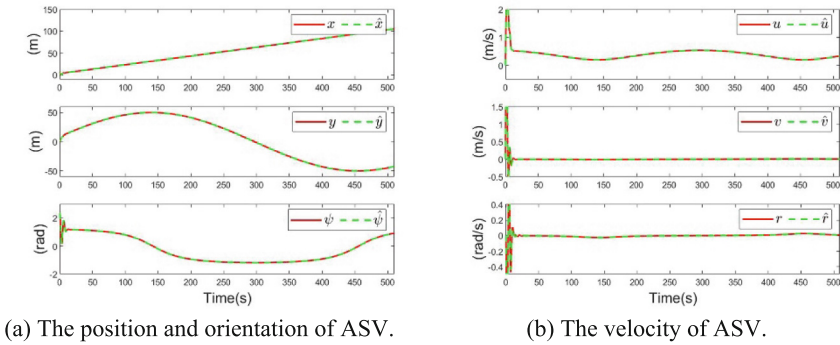


Fig. 4. The actual position, orientation and velocity estimated by digital-twin ASV system.

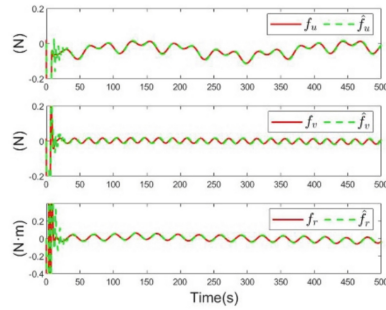


Fig. 5. The unknown total disturbances estimated by digital-twin ASV system.

6 Conclusion

In this paper, the parallel path-following of fully-actuated ASV is discussed. A parallel path-following control architecture is proposed informed by the actual-virtual interaction. In particular, the digital-twin ASV system is modeled by the designed HL-ESO, which can estimate the actual system state, unknown total disturbances, and modeling error. Next, based on the virtual guidance law, a parallel path-following controller is developed to actively drive the ASV, ensuring that it can achieve accurate path tracking behavior in real-time. Finally, by performing parallel execution, the digital-twin ASV system will be iteratively adjusted, progressively aligning with the actual ASV system. A simulation example confirms that the proposed method is effective.

Acknowledgments. This work was supported by the National Key R&D Program of China under Grant 2023YFB4302300, in part by the National Natural Science Foundation of China under Grants 52471372 and 52071044, in part by the Key Basic Research of Dalian under Grant 2023JJ11CG008, in part by the Doctoral Scientific Research Foundation of Liaoning Province under Grant 2024-BS-012, in part by the Fundamental Research Funds for the Central Universities 3132023508, and in part by the Cultivation Program for the Excellent Doctoral Dissertation of Dalian Maritime University 0034012406.

References

1. Peng, Z.H., Wang, J., Wang, D., Han, Q.-L.: An overview of recent advances in coordinated control of multiple autonomous surface vehicles. *IEEE Trans. Ind. Inf.* **17**(2), 732–745 (2021)
2. Wang, A., Liu, L., Qiu, J.B., Feng, G.: Leader-following consensus of multiple uncertain Euler-Lagrange systems via fully distributed event-triggered adaptive fuzzy control. *IEEE Trans. Cybern.* **54**(1), 76–86 (2024)
3. Wu, W.T., Peng, Z.H., Liu, L., Dang, D.: A general safety-certified cooperative control architecture for interconnected intelligent surface vehicles with applications to vessel train. *IEEE Trans. Intell. Veh.* **7**(3), 627–637 (2022)
4. Peng, Z.H., Wang, C., Yin, Y., Wang, J.: Safety-certified constrained control of maritime autonomous surface ships for automatic berthing. *IEEE Trans. Veh. Technol.* **72**(7), 8541–8552 (2023)

5. Bai, X.G., Li, B.H., Xu, X.F., Xiao, Y.J.: A review of current research and advances in unmanned surface vehicles. *J. Mar. Sci. Appl.* **21**(2), 47–58 (2022)
6. Dai, S.L., He, S.D., Lin, H., Wang, C.: Platoon formation control with prescribed performance guarantees for USVs. *IEEE Trans. Ind. Electron.* **65**(5), 4237–4246 (2017)
7. Gu, N., Wang, D., Peng, Z.H., Wang, J., Han, Q.-L.: Advances in line-of-sight guidance for path following of autonomous marine vehicles: An overview. *IEEE Trans. Syst. Man Cybern. Syst.* **53**(1), 12–28 (2022)
8. Liu, Z.X., Zhang, Y.M., Yu, X., Chi, Y.: Unmanned surface vehicles: An overview of developments and challenges. *Annu. Rev. Control.* **41**, 71–93 (2016)
9. Shi, Y., Shen, C., Fang, H.Z., Li, H.P.: Advanced control in marine mechatronic systems: a survey. *IEEE/ASME Trans. Mechatron.* **22**(3), 1121–1131 (2017)
10. Liu, Z.-Q., Ge, X.H., Han, Q.-L., Wang, Y.Q., Zhang, X.-M.: Secure cooperative path following of autonomous surface vehicles under cyber and physical attacks. *IEEE Trans. Intell. Veh.* **8**(6), 3680–3691 (2023)
11. Lv, M.G., Peng, Z.H., Wang, D., Han, Q.-L.: Event-triggered cooperative path following of autonomous surface vehicles over wireless network with experiment results. *IEEE Trans. Ind. Electron.* **69**(11), 11479–11489 (2021)
12. Zhu, K.J., Wang, Y.Q.: Event-triggered sensor fault estimation of unreliable networked unmanned surface vehicle system with correlated noises. *IEEE Trans. Veh. Technol.* **71**(3), 2527–2537 (2022)
13. Wang, F.-Y.: Parallel system methods for management and control of complex systems. *Control Decis.* **19**(2), 485–489 (2004)
14. Wang, F.-Y., Wang, X., Li, L., Li, L.: Steps toward parallel intelligence. *IEEE/CAA J. Automatica Sinica* **3**(4), 345–348 (2016)
15. Wei, Q.L., Wang, L.X., Lu, J.W., Wang, F.-Y.: Discrete-time self-learning parallel control. *IEEE Trans. Syst. Man Cybern. Syst.* **52**(1), 192–204 (2022)
16. Xu, J.X., Gu, N., Wang, D., Li, T.S., Han, B., Peng, Z.H.: Safety-critical parallel trajectory tracking control of maritime autonomous surface ships based on integral control barrier functions. *IEEE Trans. Intell. Veh.* **9**(5), 4979–4988 (2024)
17. Skjetne, R., Fossen, T.I., et al.: Adaptive maneuvering, with experiments, for a model ship in a marine control laboratory. *Automatica* **41**(2), 289–298 (2005)



The Development and Challenges of Intelligent Ship

Peng Zhang^{1,2,3,4}, Yong Ma^{1,2,3,4}(✉), Huihui Chen^{1,2,3,4}, Chengli Liu^{1,2,3,4},
and Zhongdai Wu^{4,5}

¹ School of Navigation, Wuhan University of Technology, Wuhan 430063, China
myongdl@whut.edu.cn

² Sanya Science and Education Innovation Park, Wuhan University of Technology,
Sanya 572000, China

³ State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology,
Wuhan 430063, China

⁴ Cosco Marine Technology Co Academician Workstation, Shanghai 200135, China

⁵ Cosco Marine Technology Co., Ltd., Shanghai 200135, China

Abstract. Intelligent ship, a pivotal focus for the future of the shipping industry, is revolutionizing traditional maritime operations. This article reviews the past decade development, current challenges, and future trends of intelligent ship, effectively outlining the development trajectory and charting a clear path forward. Over the past decade, global nations like Japan, South Korea, the US, and China have been proactively developing intelligent ship, integrating advanced technologies like modern IT and AI. Supported by policies and innovation, they've advanced ship autonomy, remote control, and energy efficiency, significantly propelling the shipping industry's transformation. At present, the growth of intelligent ship is robust, yet they confront multifaceted challenges—ranging from technological hurdles to regulatory frameworks, financial considerations, and the degree of market adoption—that these factors can impede their more rapid advancement. In future, the shipping sector must harness the potential of ongoing tech revolutions, refine policies, nurture skilled experts, and innovate business models to elevate the overall development level of intelligent ship.

Keywords: Intelligent ship · Development · Current Challenges · Future Trend

1 Introduction

It is widely recognized that intelligent ship play a crucial role in advancing the development of intelligent shipping. This encompasses various aspects such as technological innovation, green sustainability, quality and efficiency, industrial upgrading, and model innovation. Intelligent ship is gradually transforming the traditional maritime transportation model [1].

With the integration and application of advanced technologies, such as artificial intelligence, the Internet of Things, and big data analysis, intelligent ship has not only

enhanced the safety and efficiency of navigation but also significantly reduced operational costs and environmental impacts [2]. Globally, several countries and regions, including Japan, South Korea, Europe, and China, are actively promoting the research, development, and application of intelligent ship technology, striving to gain a competitive edge in the future shipping market [3].

Intelligent ship is poised to advance further in automation and smart capabilities, yet they will face technical, regulatory, and market challenges. With ongoing technological advancements, these vessels are on track for wider commercial deployment over the next few decades [4].

2 The Current Landscape of Intelligent Ship Development: Over the Last Decade

Intelligent ship represent an emerging trajectory for the maritime industry, reflecting a global upward trend in development. Countries such as China, Japan, South Korea, and various regions in Europe were at the forefront of researching and applying intelligent ship technologies over past decades. These nations, both at home and abroad, were making significant progress in standardization, information technology integration, and various exploratory initiatives, as depicted in Fig. 1.

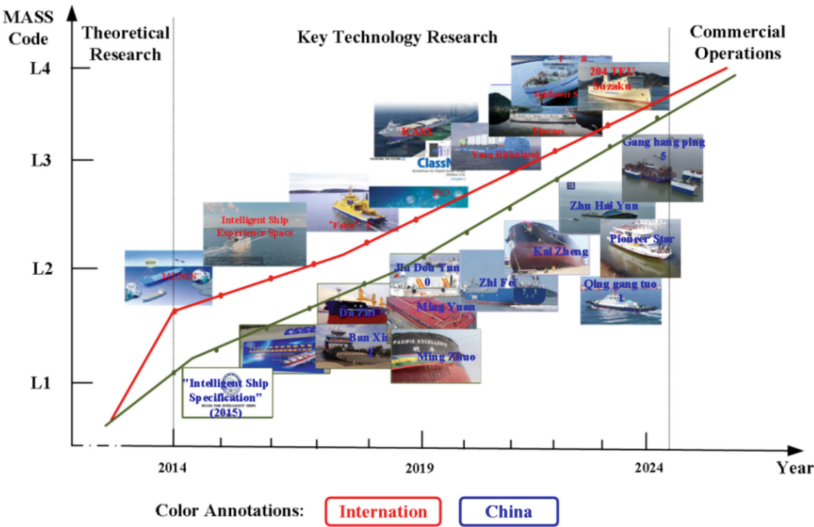


Fig. 1. The development of intelligent ship both at home and abroad

2.1 International Development

Over the past decade, the global intelligent ship sector has entered a pivotal phase of rapid growth.

In **2015**, Japan initiated a bottom-up approach to intelligent ship development, starting with the fundamentals of data communication standards. In **2016**, Europe advanced its exploration of intelligent ship technology through projects such as “Maritime Unmanned Navigation through Intelligence in Networks” (MUNIN) and “AUTO-SHIP.” By **2017**, prominent classification societies, including the American Bureau of Shipping (ABS) and Lloyd’s Register, issued specification documents on intelligent ship [5] to keep pace with technological advancements and market demands. Rolls-Royce launched an Intelligent Ship Experience Space, featuring a suite of digital products, and in **2018**, opened a cutting-edge research and development center for autonomous ships in Turku, Finland. In **2019**, Japan issued several key international standards, including ISO 16425, ISO 19847, and ISO 19848. In **2020**, South Korea pioneered advancements in intelligent navigation support systems with initiatives like the “Korean Autonomous Surface Ship” (KASS) project and the rollout of the “Hyundai Intelligent Navigation Assistance System” (HiNAS) [6]. The Mayflower Autonomous Ship, known as Mayflower400, was a groundbreaking unmanned maritime vessel developed by ProMare. By **2021**, South Korea aimed to lead in eco-friendly and autonomous ships, targeting a 75% market share by 2030. The Dutch Intelligent Shipping Platform had a roadmap for intelligent shipping technology across various vessels. The Yara Birkeland’s voyage in 2021 marked a significant milestone in intelligent ship technology. In 2022, Samsung Heavy Industries partnered with the American Bureau of Shipping (ABS) [7] to develop the Hull Stress Monitoring System (HSMS). In **2023**, European shipowners led the adoption of green fuels like methanol, ammonia, and hydrogen, collaborating with stakeholders to establish energy supply chains and infrastructure. Looking ahead to **2024**, South Korea’s new law [8], effective in 2025, aims to promote intelligent ship safety and growth by defining standards and trial areas. Japan’s Maritime Data Center has enhanced analytics and completed remote navigation tests, with autonomous ship trials like “MEGURI 2040.”

In summary, the development of intelligent ship is now in a phase of swift expansion, driven by technological innovation and market demand. The future outlook for intelligent ship is exceedingly promising, with further technological progress and supportive policies anticipated to broaden the spectrum of commercial applications in the near future.

2.2 Domestic Development

In the last decade, China has seen swift progress in the domain of Intelligent ship, with vigorous government and corporate efforts to advance research and application of associated technologies.

The year **2015** marked a pivotal moment for intelligent ship development, with a global surge in research and development initiatives. The China Classification Society (CCS) introduced the pioneering “Intelligent Ship Specification” (2015) [9], and by **2017**, the China Intelligent Ship Innovation Alliance was officially established in Shanghai [10]. Notably, the “Da Zhi” intelligent bulk carrier, developed by the China State Shipbuilding Corporation, became the world’s first intelligent ship to receive classification society certification. In **2018**, China launched the “Intelligent Ship Development Action Plan (2019–2021)” [11] to invigorate the shipbuilding industry, with the “Ban

Xin 2" ferry, featuring an intelligent navigation system, setting a new standard. The "Ming Yuan" achieved intelligent capabilities and secured global certifications from CCS and DNV GL. In **2019**, China's "Jing Dou Yun" completed its maiden autonomous cargo voyage, advancing intelligent ship technology toward commercial viability and navigation standardization. The "Ming Zhuo" a 400,000-ton VLOC, was equipped with smart systems for navigation, energy, and cargo management. The General Office of the Ministry of Industry and Information Technology issued the "Guidelines for the Construction of Intelligent Shipbuilding Standard System (2020 Edition)" [12] in **2020**, while the revised "Intelligent Ship Specification" (2020) [13] became effective, addressing the growing intelligent ship market. In **2021**, the "Zhi Fei" completed sea trials in Qingdao's Nu Island Port area, and by **2022**, Chinese ministries had issued policies promoting green, intelligent ship development. The "He Hua" an intelligent container ship, began operation with a secure network system, while the "Kai Zheng" an ultra-large smart oil tanker, received classification society notation, highlighting China's progress in smart vessel technology. The "White Paper on Intelligent Ship Development" (**2023**) detailed significant technological advancements, with the debut of the autonomous "Zhu Hai Yun" showcasing high domestic equipment rates and independent technology, underscoring China's maritime AI capabilities [14]. In **2024**, China's intelligent ship industry continued to advance with the "Gang Hang Ping 5" undergoing sea trials for intelligent navigation, and the launch of the green, smart inland vessel "Hua Nan 6" to support sustainability. The "Qing GangTuo 1" tugboat achieved zero emissions in electric mode, and the "Pioneer Star" an electric ship for cold regions, improved transportation efficiency in southern Chinese waterways.

In summary, the current wave of technological innovation and industrial transformation is creating substantial strategic opportunities for the growth of China's intelligent ship industry. Driven by technological advancements and rising market demands, intelligent shipping is experiencing rapid development, fostering new industries, models, and momentum that signal a promising future for the sector.

3 The Intelligent Ship Industry: Navigating the Challenges Ahead

Despite their burgeoning growth, intelligent ship confronts various challenges, such as spanning technology, regulatory frameworks, financial concerns, and market receptiveness, which may hinder their swifter progress [15].

3.1 Technical Challenges

While intelligent ship incorporates advanced technologies like autonomous navigation, intelligent perception, and remote control, these capabilities are still under refinement. Enabling self-directed voyages and onboard smart decision-making requires overcoming substantial technical obstacles. The maritime environment is inherently unpredictable, with constantly changing winds, waves, and currents posing major challenges. Therefore, intelligent ship must be equipped with sophisticated autonomous identification and collision-avoidance systems. The effectiveness of core sensors on intelligent ship, particularly their accuracy, adaptability, and reliability, is crucial. Current sensor performance

may not yet meet the demands of the next phase of development. Inaccurate data collection, processing, or synthesis can compromise high-precision situational awareness, directly impacting the autonomous system's decision-making and actions. To maintain stable connectivity, intelligent ship requires multi-layered network structures and communication channels. However, current communication technologies have yet to meet the high standards of connectivity and security necessary for intelligent ship operations.

3.2 Safety and Privacy Concerns

The operation of intelligent ship involves managing large volumes of sensitive data, which raises significant concerns about data security and privacy. Robust measures must be implemented when deploying AI technology on intelligent ship to prevent data theft or misuse.

3.3 Regulatory and Standardization Issues

Regulatory support and market demand are crucial for the adoption of intelligent ship, with technological maturity as a foundation. However, the development of regulations, standards, and testing frameworks for intelligent ship has lagged, highlighting the need for comprehensive regulatory structures and testing standards.

3.4 Cost Concerns

The high costs associated with developing and maintaining intelligent ship systems pose a challenge for many small and medium-sized enterprises, potentially limiting the widespread adoption of intelligent ship technologies.

3.5 Market Acceptance

As a relatively new concept, market acceptance of intelligent ship is still in its early stages. Demonstration projects and strategic marketing efforts are necessary to increase market recognition and acceptance of intelligent ship technology.

Overcoming these challenges requires collaboration among governments, industry bodies, businesses, and research institutions. Joint efforts in technological advancement, standardization, legislation, and market promotion are essential to drive the intelligent ship industry forward.

4 Trends in Intelligent Ship Development

A review of the past decade's progress and challenges in intelligent ship development reveals a clear path forward: driven by technological innovation, policy and standards establishment, talent cultivation, and evolving business models, the industry is well-positioned to advance intelligent ship development to new heights.

4.1 Driving Growth Through High-Tech Innovation

The concept of developing new quality productivity, as emphasized by General Secretary Xi Jinping, sets the course for efficient and green development in intelligent shipping. Intelligence, greening, localization, and safety are interconnected and mutually reinforcing dimensions of intelligent ship. Intelligence and greening represent the forefront of technological advancements in the industry, while localization strengthens international competitiveness and secures supply chains. Safety, in turn, provides the essential foundation and guarantee for all these advancements. With continuous technological progress and supportive policy measures, intelligent ship are positioned to make substantial breakthroughs across these areas, driving the high-quality growth of the entire shipbuilding industry.

4.2 Further Improvement Policies and Standards with Wider Applications

The CCS has launched the “Intelligent Ship Specifications 2024” supporting initiatives like “Maritime Power.” This guide introduces regulations for enhanced navigation systems, autonomous tugboat controls, and smart inland vessel navigation. It establishes benchmarks for electric propulsion and small diesel engine diagnostics, fostering smart ship technology. Amid growing global competition, entities such as IMO, IACS, and ISO are crafting international standards to regulate the development and operation of smart ship.

4.3 Developing Talent Cultivation Models

Cultivating talent for intelligent ship involves interdisciplinary education, blending ship engineering, IT, AI, and data skills. Practical training in ship operation and maintenance, plus simulation, is key. A global outlook with cooperation skills is vital for international collaboration. Understanding regulations and ethics in maritime and tech fields ensures responsible practice. These strategies develop talent to lead the maritime industry’s future.

4.4 Innovations in Intelligent Ship Operations

Intelligent ship development demands innovative, collaborative business models that tackle operational challenges and cater to shipowner and market needs. It moves from “product-focused” to “market-driven” fostering new services and models for shared prosperity and risk. Shipping firms globally are adopting digital and smart transitions, emphasizing ecosystem development, and leveraging tech like AI to enhance the shipbuilding industry and create an AI-driven intelligent ship ecosystem, offering both tech backing and market-aligned solutions [16].

Acknowledgment. This work was funded by the Innovation Research Team Project of the Hainan Natural Science Foundation, China (NO. 722CXTD518).

References

1. Demirci, S., Cicek, K.: Intelligent ship inspection analytics: ship deficiency data mining for port state control. *Ocean Eng.* **278**, 114232 (2023)
2. Zhu, Q., Xi, Y., Weng, J., et al.: The analysis of intelligent functions required for inland ship. *Ocean Eng.* **252**, 124148 (2024)
3. Ma, Y., Hu, Z., Wang, W., et al.: Development status and prospects of ship-borne intelligent equipment and systems. *J. Mech. Eng.* 1–12 (2024). (in Chinese)
4. Liu, C.: A Brief Introduction to the development and trends of intelligent ship technology. *China Water Transp.* **07**, 83–85 (2021). (in Chinese)
5. China Intelligent Ship Innovation Alliance. White Paper on the Development of intelligent ship - Ocean-going Ship Edition (2023). <https://www.54seaman.com/news/index/detail/id/140014.html>. Accessed 30 Sept 2024. (in Chinese)
6. Liu, J., Yang, F., Li, S., et al.: Testing and evaluation for intelligent navigation of ships: current status, possible solutions, and challenges. *Ocean Eng.* **295**, 116969 (2024)
7. Samsung Heavy Industries collaborates with ABS to develop intelligent ship hull monitoring technology Eworldship.com (2022). https://www.eworldship.com/html/2022/Shipyards_0728/184246.html. Accessed 30 Sept 2024. (in Chinese)
8. Effective next January! The South Korean National Assembly passes the Autonomous Shipping Act International Ship Network (2024). https://www.eworldship.com/html/2024/ShipbuildingAbroad_0114/199756.html. Accessed 30 Sept 2024. (in Chinese)
9. “Intelligent Ship Specification 2015”. China Classification Society (2016). <https://www.ccs.org.cn/ccswz/file/download?fileid=202312061011848316>. Accessed 30 Sept 2024. (in Chinese)
10. China Intelligent Ship Innovation Alliance holds its first member conference. E.Worldship.com (2017). https://www.eworldship.com/html/2017/Exhibition_1205/134407.html. Accessed 30 Sept 2024. (in Chinese)
11. Three departments jointly issue the “Intelligent Ship Development Action Plan (2019–2021)”. Ministry of Industry and Information Technology website (2018). https://www.gov.cn/xinwen/2018-12/30/content_5353550.htm. Accessed 30 Sept 2024. (in Chinese)
12. Notice of the General Office of the Ministry of Industry and Information Technology on printing and distributing the “Shipbuilding Intelligent Total Assembly Construction Standard System Construction Guide (2020 Edition)”. Ministry of Industry and Information Technology (2020). https://www.gov.cn/zhengce/zhengceku/2020-08/25/content_5537132.htm. Accessed 30 Sept 2024. (in Chinese)
13. “Intelligent Ship Specification 2020”. China Classification Society (2019). <https://www.ccs.org.cn/ccswz/file/download?fileid=202310120311311012>. Accessed 30 Sept 2024. (in Chinese)
14. AI Empowers the Ocean, the “2024 China Intelligent Shipping Industry Development Report” is Released. People’s Daily Online (2024). <http://sh.people.com.cn/BIG5/n2/2024/0706/c134768-40903678.html>. Accessed 30 Sept 2024. (in Chinese)
15. Ma, Y., Hu, Z., Wang, W., et al.: Development status and prospects of intelligent ship equipment and systems. *J. Mech. Eng.* 1–12 (2024). (in Chinese)
16. Zhang, P., Song, Z., Liu, Y., et al.: A study of engine room smoke detection based on proactive machine vision model for intelligent ship. *Expert Syst. Appl.* **241**, 122689 (2024)



Functional Safety Monitoring Strategy for Motor Control Unit of Autonomous Navigation Domain Controller in Unmanned Surface Vehicle

Ahui Xu^{1,4}, Yamin Huang^{1,2,3}, Yuanqiao Wen^{1,2,3}(✉), Jie Liu⁵, and Lihang Song^{1,3,6}

¹ State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology (WUT), Wuhan, China

wenyqwhut@foxmail.com

² Sanya Science and Education Innovation Park of Wuhan University of Technology, Wuhan, China

³ Intelligent Transportation Systems Research Centre, WUT, Wuhan, China

⁴ School of Navigation, WUT, Wuhan, China

⁵ Zhejiang Scientific Research Institute of Transport, Hangzhou, China

⁶ School of Transportation and Logistics, Wuhan University of Technology, Wuhan, China

Abstract. The motor control unit is the core device of the unmanned surface vehicle (USV) to realize the function of autonomous navigation, and ensuring its functional safety is the premise and foundation of system reliability. Firstly, Hazard and Operability (HAZOP) method [1] is used to identify the potential hazards, and fault tree analysis (FTA) is used to realize functional safety analysis for the possible failure of the motor control unit (MCU) of the Autonomous Navigation Domain Controller (ANDC) in USV. Secondly, a 3-layer monitoring strategy for MCU is proposed, and a fault diagnosis system was designed. Thirdly, to verify the proposed design, the quantitative risk analysis of the MCU with 3-layer monitoring strategy is introduced, which is based on FTA. The results show that the failure rate of the system reduced from 64.23×10^{-4} to 40.41×10^{-4} with the proposed 3-layer monitoring strategy in the case study. The result implies the potential of the proposed method for improving the safety of the USV.

Keywords: Monitoring Strategy · Domain Controller · USV

1 Introduction

As the intelligence and integration of Unmanned Surface Vehicle (USV) [2] increases, the complexity of the electronic and electrical architecture [3] systems of the ship is increasing. In order to realize the comprehensive monitoring and intelligent management of USV, China Classification Society (CCS) has stipulated the requirements for the intelligent integration platform [4] in the “Guidelines for Surveys of unmanned surface vehicle”. According to the requirements of the guidelines, the intelligent integration platform should be able to provide support for intelligent navigation and form a unified

integration platform for data acquisition, integration, interaction and control command transmission, so as to ensure the functional safety of USV during autonomous navigation.

Autonomous Navigation Domain Controller (ANDC) is an intelligent integration platform for USV, which is the core control unit of USV and is responsible for analyzing the navigation intention of USV and providing expected commands to the motor control unit (MCU). Motor control that is controlled by MCU is strongly related to the safety of USV. If the output motor control signal is abnormal during the navigation of USV, it may cause the USV out of control, or even accidents (e.g., collision, etc.). Thus, it has become a necessary and challenging issue to improve the functional safety of MCU.

To guarantee the functional safety of USV, many researchers have proposed various methods. Wen [5] proposed a 3-layer architecture for USV sailing state monitoring consisting of perception layer, transmission layer, and application layer. Hu et al. [6] established an IVHM system model based on intelligent body, which provides a reasoning basis for the comprehensive health analysis of USV. Cheng et al. [7] conducted research on the health management of USV's battery, integrated power system, battery management system and power supply system; Gu et al. [8] proposed the structure of USV's health management system. However, few researches studied on the functional safety strategy [9] of the motor control unit for USV, and cannot complete the monitoring of the navigation safety of USV based on the identification, transmission, and control of motor signals.

To fill the gap, this paper studied on the functional safety of the MCU of the ANDC of the USV, proposes a 3-layer monitoring strategy based on the domain controller based on the Infineon TC397 master controller chip that assigns the 6 CPUs to the 3-layer architecture. The designed 3-layer monitoring architecture realizes independent operation and redundancy coverage, and realizes the decoupling of the hardware and software, and the fault handling strategy is designed to enable the system to recover quickly or enter a safe state quickly.

The rest of this paper is organized as follows: the functional safety analysis of existing MCU and ANDC are presented in Sect. 2, followed by the designed 3-layer monitoring strategy in Sect. 3. The verification of the proposed method using FTA is presented in Sect. 4. Then, the main findings are collected in Sect. 5.

2 Functional Safety Analysis of MCU in ANDC

2.1 The Design of ANDC and MCU in USV

In order to redefine the design of the electronic and electrical architecture of the USV and lead the trend of centralizing the electronic and electrical architecture of the USV, a centralized platform for the USV ANDC is designed. The electrical and electronic system architecture of the USV is shown in Fig. 1, which contains different equipment modules, such as communication module, perception module, ANDC decision module. Among them, the ANDC decision module is composed of different functional units, including interface input/output unit, signal processing unit, data storage unit, power management unit, motor control unit and others, this paper mainly focuses on the motor control unit for safety monitoring [10] and fault analysis.

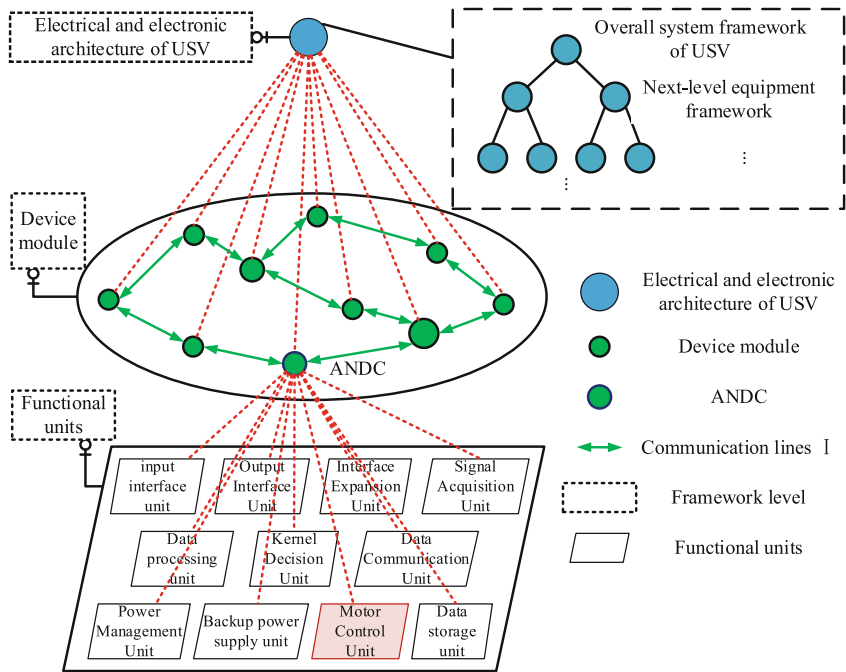


Fig. 1. Definition of relevant items of the motor control unit

In this paper, we mainly study the functional safety monitoring of motor control unit of ANDC, sort out the ways of input, transmission, decision-making, output and feedback signals of motor control unit, analyze the hazards and assess the risks of the relevant factors, determine the authenticity of motor data signals, the safety of communication and the reliability of function realization and other safety objectives, and finally formulate relevant monitoring strategies to ensure the safe and reliable realization of the functions of motor control unit of ANDC. Finally, relevant monitoring strategies are formulated to ensure the safety and reliability of the ANDC motor control unit functions.

2.2 Determination of the Objectives of Safety

In order to ensure the safety of USV navigation, to meet the “Guidelines for Surveys of unmanned surface vehicle”, and to improve the reliability of the USV control unit in the operation process, it is determined that the safety objective of this paper is “to avoid the abnormal working condition of ANDC motor control unit”.

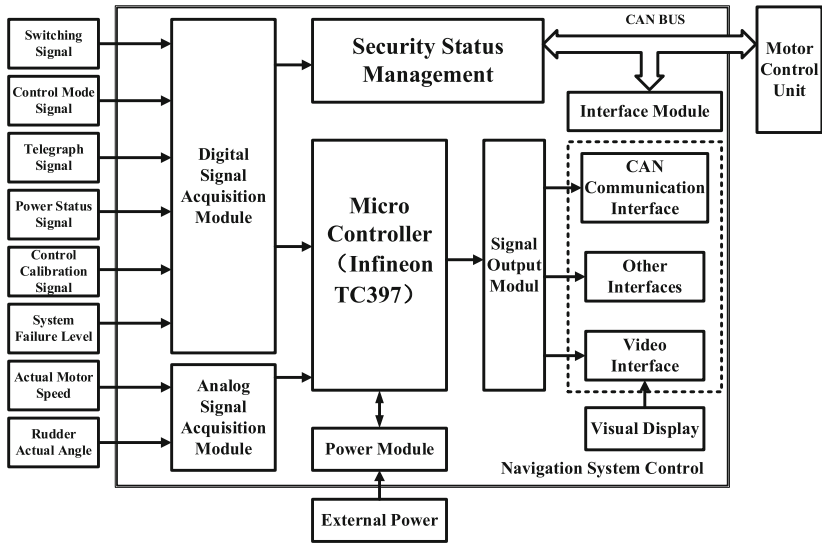


Fig. 2. Definition of relevant items of the motor control unit

2.3 Qualitative Analysis of the Original MCU

When sorting out the factors that contribute to the occurrence of abnormal events in the operating status of motor control units, it is necessary to consider hardware failure, software failure and abnormal external information interaction. In order to develop a set of scientific and effective functional safety requirements, and for the systematic analysis of the cause of the failure as well as the use of intuitive and easy-to-understand graphical analysis methods, so this paper adopts the Fault Tree Analysis (FTA) method of safety analysis, with the motor control unit is in abnormal working condition as the top event, to systematically analyze, looking for the basic fault events that lead to the signal output abnormality, as shown in Fig. 3 and details of the failure events are shown in Table 1.

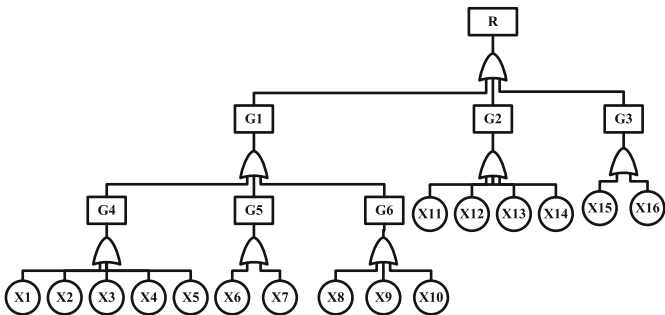


Fig. 3. FTA based avoidance of abnormal working conditions in motor control units (Not included in the monitoring architecture).

Table 1. Motor control unit work abnormal fault event information

Event number	Event description	Event type	Occurring Probability (1×10^{-4})/h
R	The motor control unit is in abnormal working condition	Top events	–
G1	Hardware failure	Intermediate events	–
G2	Software failure	Intermediate events	–
G3	Abnormal external information interaction	Intermediate events	–
G4	Energy supply anomaly	Intermediate events	–
G5	Abnormal platform function	Intermediate events	–
G6	Internal communication link anomaly	Intermediate events	–
X1	Battery loss	Basic events	0.381 ^①
X2	Overvoltage protection activation	Basic events	1.4 ^②
X3	Overload protection activation	Basic events	1.4 ^②
X4	Overcurrent protection activation	Basic events	1.4 ^②
X5	Static protection activation	Basic events	1.4 ^②
X6	Kernel anomaly	Basic events	11.4 ^②
X7	Memory anomaly	Basic events	11.4 ^②
X8	Signal interference	Basic events	8.7 ^②
X9	Signal short circuit	Basic events	7.4 ^②
X10	Signal disconnect	Basic events	7.4 ^②
X11	Addressing anomaly	Basic events	0.57 ^①
X12	Address conflict	Basic events	0.57 ^①
X13	Software driver abnormalities	Basic events	0.381 ^①
X14	Logic abnormal	Basic events	1.1 ^②
X15	Signal input abnormal	Basic events	5.7 ^①
X16	Abnormal feedback information	Basic events	3.81 ^①

* Data from expert experience, product technical manuals.

① product technical manuals ② expert experience.

0.381 = 1.14/3 (Three years warranty)

In this paper, we find the minimum cut set (MCS) of the fault tree of the motor control unit operating abnormally by the upstream method. Based on the FTA in Fig. 2, the structure function of the motor control unit is simplified to avoid abnormal operating conditions:

$$\begin{aligned}
 R &= G1 + G2 + G3 \\
 &= (G4 + G5 + G6) + (X11 + X12 + X13 + X14) + (X15 + X16) \\
 &= [(X1 + X2 + X3 + X4 + X5) + (X6 + X7) + (X8 + X9 + X10)] \\
 &\quad + (X11 + X12 + X13 + X14) + (X15 + X16)
 \end{aligned}$$

Thus, the MCS is obtained as {X1}, {X2}, {X3}, {X4}, {X5}, {X6}, {X7}, {X8}, {X9}, {X10}, {X11}, {X12}, {X13}, {X14}, {X15}, and {X16} for the abnormal fault tree of the motor control unit of the USV autonomous navigation domain controller.

The specific steps of the qualitative analysis algorithm are as follows:

- (1) Construct a two-dimensional array for storing cut sets;
- (2) Using loop nested statements by the top events, intermediate events and basic events are traversed, if the event and the previous level events are or gate relationship, then increase the number of cut set array; if the event and the previous level events are and gate relationship, then increase the capacity of the cut set array storage. Traverse to the top level event in turn to obtain the set of all cut set arrays of the fault tree and store them in a two-dimensional array;
- (3) Assign a prime number to the basic events of the fault tree, and the product of the prime numbers corresponding to each of the basic events of the tree that make up a cut set, in descending order;
- (4) Apply Boolean absorption operation to remove the corresponding cut sets, loop the operation from left to right, and finally the cut sets corresponding to the product of remaining prime numbers that are not integrable are the minimum cut sets of the top-level events;
- (5) Construct a six-tuple $\langle \text{FTAid}, \text{CSid}, \text{CSprob}, \text{CSimport}, \text{Facdesc} \rangle$. FTAid denotes the fault tree number; CSid is the cut-set number; CSprob is the cut-set probability; CSimport are the cut-set importance; Facdesc denotes the event description; CSid is a three-tuple $\langle \text{CSidnum}, \text{Facdesc}, \text{Factprimport} \rangle$, which denotes the number of the cut-set, and the underlying event of the smallest cut set, and Factprimport denotes the event probability importance.

2.4 Quantitative Analysis of the Original MCU

Calculate the top-level event failure probability by the tolerant repulsion theorem. Let the failure probability of the top-level event is not $P(T)$ for the sum of the probability of all the minimum cut set, C_i, C_j, C_k is the minimum cut set of the serial number i, j, k , the maximum cut set is n , in order to avoid the “combinatorial explosion”, remove the higher-order terms for the approximation of the calculation of the probability of failure is expressed as follows:

$$P_T = \sum_{i=1}^n P(C_i) - \sum_{i < j=2}^n P(C_i C_j) + \sum_{i < j < k=3}^n P(C_i C_j C_k) \quad (1)$$

The probabilistic importance is the contribution to the top-level event when the basic event occurs and reflects the size of the contribution to the change in the probability of system failure when the basic event failure changes. The minimum cut-set contribution of sequence number i is denoted as:

$$\lambda_i = P(C_i)/P_T \quad (2)$$

Based on the value of the probability of occurrence of the basic event given in Table 1, the top event failure probability is calculated as $64.22996001 \times 10^{-4}/h$ through Eq. (1), and the minimum cut-set failure probability and importance of the fault tree of the motor control unit of the USV autonomous navigation domain controller is calculated through Eq. (2) as shown in Table 2.

Table 2. MCS importance of fault tree for abnormal operation of motor control unit

Number	MCS	Event probability (1×10^{-4})/h	importance rating λ_i
1	{X1}	0.381	0.005931811
2	{X2}	1.4	0.021796682
3	{X3}	1.4	0.021796682
4	{X4}	1.4	0.021796682
5	{X5}	1.4	0.021796682
6	{X6}	11.4	0.177487266
7	{X7}	11.4	0.177487266
8	{X8}	8.7	0.135450808
9	{X9}	7.4	0.115211032
10	{X10}	7.4	0.115211032
11	{X11}	0.57	0.008874363
12	{X12}	0.57	0.008874363
13	{X13}	0.381	0.005931811
14	{X14}	1.1	0.017125964
15	{X15}	5.7	0.088743633
16	{X16}	3.81	0.059318113

3 Software Architecture Design for Functional Safety of Motor Control Units

In order to ensure the functional safety of the MCU of ANDC in the USV, a hierarchical, interactive and redundant monitoring software architecture is designed. The hierarchical design is to design the monitoring strategy for different functional layers of the

motor control unit, the interactive design is to transfer data signals and functional statuses between different functional layers, and the redundant monitoring is to design the backup for the 6 cores of the main control CPU. Therefore, this paper designs a 3-layer monitoring strategy for speed signal control of ANDC motor control unit. In order to fully utilize the ANDC and completely cover the 3-layer architecture, and to ensure the independence of each layer, this paper selects the UC EAORA07 as the basis, based on which the Infineon TC397 is used in the 3-layer architecture designed in this paper.

3.1 Hardware Platform

The autopilot processor inside EAORA07 is an Orin chip designed for embedded intelligent systems including autopilot systems. In view of the functional requirements of USV, EAORA07 is developed twice to design the USV autonomous navigation domain controller, which is internally modularized according to the functional units, including the power supply unit, the hardware interface unit, the data storage unit, the decision-making computation unit, the redundant communication unit, the performance expansion unit, and the safety detection unit, etc. The safety detection unit includes the motor control unit, which is used to monitor the functional safety of the motor control signals. The safety detection unit includes the motor control unit, which is used to monitor the functional safety of the motor control signals, and is realized on the EAORA07 platform with the Infineon TC397 microcontroller, which contains a 6-core TriCore™ architecture with a 300 MHz operating frequency, a 2.9 MB + 16 MB RAM with ECC (Error Correction Code) protection, and supports ASIL-D's highest level of safety. It supports the highest ASIL-D safety requirements. Together with other basic chips, it can be used to realize the design of hardware core safety architecture, i.e., functional safety monitoring of motor control signals.

3.2 Multi-layer Monitoring Strategy

Based on the TC397 platform with 6-core CPUs, a 3-layer monitoring strategy is designed for the functional safety of the motor control unit as well as an additional 3-layer redundant monitoring design, with the 3-layer architecture being the control layer (Level 1), the control monitoring layer (Level 2), and the controller monitoring layer (Level 3), and the additional 3-layer redundant monitoring architecture being the redundant control layer (Level 1–1), the redundant control monitoring layer (Level 2–1), and redundant controller monitoring layer (Level 3–1), as shown in Fig. 4.

The control layer (Level 1) mainly realizes the control function of the motor of the USV, analyzes the sailing information of the USV according to the control mode selection signal, the command signal of the remote control, the car clock signal of the shore end joystick, the sailing speed of the USV, the actual rotational speed fed back from the motor, etc., and at the same time, calculates the torque information of the USV, and finally outputs the control signal of the motor. In addition, it also contains corresponding diagnostic functions, including the reasonableness check of PWM signals, etc. Based on the diagnostic results, the control layer executes corresponding response commands.

The control and monitoring layer (Level 2) is the monitoring of the control layer (Level 1). Its main logic is to execute another algorithm model for calculating the torque

of the USV through another independent information acquisition, transmission and processing path to calculate the output motor rotation speed and compare it with the actual rotation speed fed back from the motor, if the difference is larger than the expected threshold value and the duration exceeds the set time, it means the action is abnormal, and the control and monitoring layer (Level 2) enters into the safety The control and monitoring layer (Level 2) enters the safe state and executes the corresponding action instructions until it gets rid of the warning.

The controller monitoring layer (Level 3) includes the hardware monitoring module and software monitoring program. In this paper, the TC397 is used as the motor control unit functional safety controller, and its hardware monitoring module is monitored and controlled by the power supply chip, which communicates with the TC397 through the SPI, and the TC397 needs to respond within a set period of time, and if the number of responses exceeds the set threshold within the set period of time, then the power supply chip will be triggered to perform the corresponding actions, including, but not limited to, the power restart, reset and other operations.

The redundant control layer (Level 1-1), redundant control monitoring layer (Level 2-1) and redundant controller monitoring layer (Level 3-1) in the redundant monitoring architecture are the functional backups of the other three monitoring architectures to ensure the safe operation of the motor control in the event of kernel lock-up, program runaway, etc., and thus ensure the safety and reliability of the USV autonomous navigation.

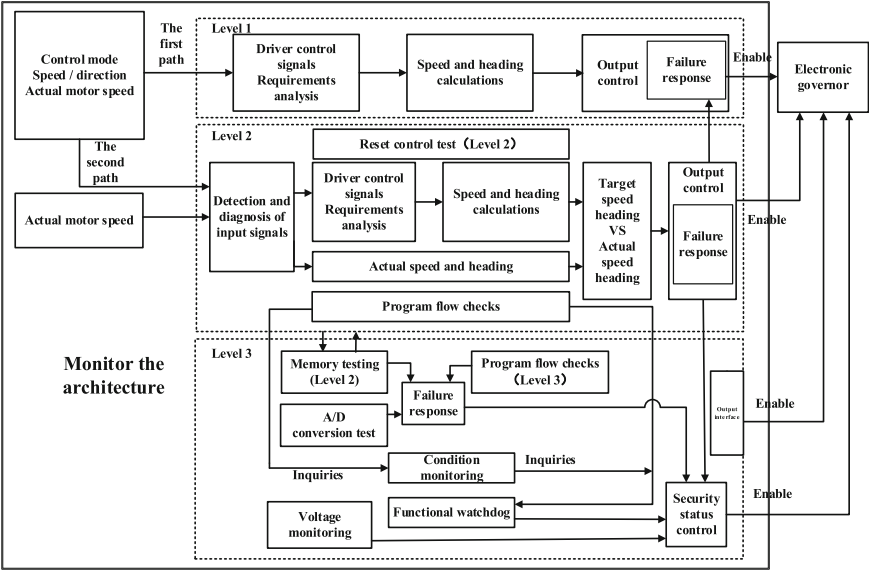


Fig. 4. 3-layer monitoring architecture for functional safety of motor control units

3.2.1 Design of Level 1

In order to realize the navigation control of USV, the control architecture designed in this paper contains two parts. One part is logic control, including input signals and control mode selection, the input signals include remote control PWM control signal; shore-end digital mapping communication data transmission, and then through the SPI for control signal input; autonomous navigation control instruction calculation, and then through the SPI communication for control signal input, the corresponding control modes are manual remote control, shore-end remote control, and autonomous navigation control, as shown in Table 3. The other part is motor control, which is to convert the received control instructions into PWM values and send them to the motor, and then control the direction of motor speed. The other part is motor control, the received control instructions are converted into PWM values and sent to the motor, which in turn controls the direction and size of the motor rotation speed, as shown in Table 4. The motor is then controlled by the control signal input via SPI communication.

Table 3. Motor control mode logic sorting

Control mode	Signal source	Instruction number	Instruction content	Signal input method
Manual remote control	Handheld remote	Mode 1	Manual remote control mode	PWM signal input
		Mode 2	Remote control mode	
		Mode 3	Emergency stop mode	
Shore-side remote control	Remote control panel	Mode 4	Remote driving Switching privilege mode	Ethernet/Digital Mapping/SPI

3.2.2 Design of Level 2

Level 2 is the control and monitoring layer, which is used to monitor the functional status of the control layer Level 1. Level 2 collects, transmits and calculates the direction and size of the motor rotation speed through a different data input path from Level 1, fits the torque output from the USV based on the motor rotation speed, and compares it with the torque obtained from the direction and size of the rotation speed fed back by the motor in real time. According to the different difference between the desired torque and the actual torque, the corresponding response action is executed, as shown in Table 5.

3.2.3 Design of Level 3

Level 3 Controller Monitoring Layer is a software architecture designed for the security status of the TC397 platform, which contains the following.

Table 4. Control mode content and communication

Control mode	Signal source	Instruction function	Signal transmission process
Manual remote control	Mode changeover paddles	Enables control mode switching	ADC signal acquisition
	Direct paddles	Keep the drone ship straight	↓ PWM input
	Steering paddles	Changing the course of an USV	↓ PWM output
Shore-side remote control	Telegraph equipment	Keeping the drone ship straight	Ethernet
	Control lever	Changing the course of an USV	↓ SPI ↓ ADC signal acquisition ↓ PWM output
Autonomous navigation	Domain controller	Calculation of the PWM value according to the corresponding decision algorithm	SPI ↓ ADC signal acquisition ↓ PWM output

Table 5. Motor desired speed and actual speed difference analysis fault response

Difference	Fault level	PWM actual (speed)	Failure response action
<10%	Trouble-free	0–2000 (0–10 kn)	Normal operation
10%–20%	Class 2 faults	500–1500 (0–5 kn)	Issuance of early warning, speed restrictions on USV, and reset of ANDC
≥20%	Class 1 faults	1000 (0 kn)	Issuing alarms and sending Emergency stop commands to USV

Remarks: PWM value 0–1000 is reverse, 1000 is stop, 1000–2000 is forward

TC397 status monitoring. The status monitoring of TC397 is realized through the watchdog function of Orin chip in ANDC, which is mainly composed of registers, counters and dog barking module. The upper Orin chip feeds the dog to TC397 through SPI communication at regular intervals, and if TC397 fails to reply within the specified time, it triggers the timer, and performs the reset operation when the count exceeds the preset value, which ensures the safe state of TC397, as shown in Fig. 5.

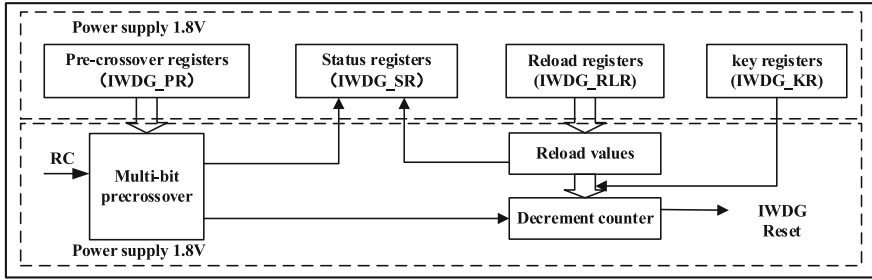


Fig. 5. Functional safety status monitoring trigger logic

Communication status monitoring. Ensure that the SPI communication between TC397 and Orin is normal and the CAN bus communication of control signal sent down by TC397 is normal. If the SPI communication abnormality will lead to the loss of upper level control commands and the CAN bus communication abnormality will lead to the loss of control of the motor of the USV, resulting in the danger of sailing the USV, at this time, the communication abnormality status will be warned to the shore end control platform and the reset operation will be executed to the ANDC.

Reset control test. Ensure that the controller can complete the reset operation when it detects an abnormal situation, conduct a test before the USV starts, only if the test is passed can the motor be issued control commands, if the reset control test is in a faulty state then the controller is always in a reset state until a manual reset can be lifted from its state.

The TC397 chip has 16MB of FLASH, 2.9MB of SRAM, and 129K of EEPROM. In order to ensure that its internal SRAM, EEPROM, and FLASH function properly, it is necessary to test its SRAM, EEPROM, and FLASH, and the specific tests include address buffer overflow test, addressing test, and bit flip test. Specific tests include address buffer overflow test, addressing test and bit flip test, etc.

A/D conversion test. Ensure that the A/D conversion function is normal, if there is an abnormality in the A/D conversion function, the received data will be parsed incorrectly, which will lead to incorrect information in the issued control command, causing the danger of USV navigation. At this time, the A/D conversion abnormality will be warned to the shore end control platform, and reset operation will be performed on TC397.

4 Fault Diagnosis System Design

4.1 Qualitative Analysis of the MCU with 3-Layer Design

Based on the 3-layer monitoring architecture designed above, the abnormal operation of the motor control unit is again analyzed qualitatively and calculated quantitatively, and compared with the probability of occurrence of the top event failure that has not been added to the monitoring architecture before, so as to verify the reliability and effectiveness of the monitoring architecture designed in this paper.

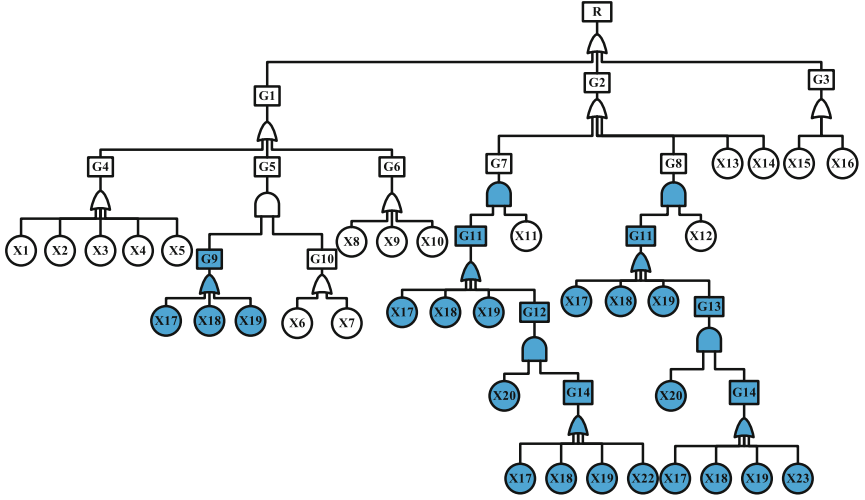


Fig. 6. FTA based avoidance of abnormal working conditions in motor control units (With the addition of a three-tier monitoring architecture)

After joining the 3-layer monitoring architecture, the fault tree of the motor control unit is analyzed again, as shown in Fig. 6. The highlighted part is the intermediate events and base events that appeared after joining the monitoring part, and the combed events are analyzed qualitatively and calculated quantitatively to get the new top event formula, the minimum cut set of the fault tree of the motor control unit working abnormally and the probability of failure.

$$\begin{aligned}
 R &= G1 + G2 + G3 \\
 &= (G4 + G5 + G6 + G7 + G8 + X13 + X14) + (X15 * X16) \\
 &= ((X1 + X2 + X3 + X4 + X5) + (G9 * G10) + (X8 + X9 + X10)) \\
 &\quad + ((G11 * X11) + (G11 * X12) + X13 + X14) + (X15 * X16) \\
 &= ((X1 + X2 + X3 + X4 + X5) + ((X17 + X18 + X19) * (X6 + X7)) + (X8 + X9 + X10)) \\
 &\quad + (((X17 + X18 + X19 + G12) * X11) + ((X17 + X18 + X19 + G13) * X12) + X13 + X14) + (X15 * X16) \\
 &= ((X1 + X2 + X3 + X4 + X5) + ((X17 + X18 + X19) * (X6 + X7)) + (X8 + X9 + X10)) \\
 &\quad + (((X17 + X18 + X19 + (X20 * G14)) * X11) + ((X17 + X18 + X19 + (X20 * G14)) * X12) + X13 + X14) + (X15 * X16) \\
 &= ((X1 + X2 + X3 + X4 + X5) + ((X17 + X18 + X19) * (X6 + X7)) + (X8 + X9 + X10)) \\
 &\quad + (((X17 + X18 + X19 + (X20 * (X17 + X18 + X19 + X22))) * X11) + ((X17 + X18 + X19 + (X20 * (X17 + X18 + \\
 &\quad X19 + X23))) * X12) + X13 + X14) + (X15 * X16)
 \end{aligned}$$

Thus, the MCS is obtained as {X1}, {X2}, {X3}, {X4}, {X5}, {X8}, {X9}, {X10}, {X13}, {X14}, {X15}, {X16}, {X17,X18,X19,X6,X7}, {X17,X18,X19,X20,X22}, {X17,X18,X19,X20,X23} for the abnormal fault tree of the motor control unit of the USV autonomous navigation domain controller.

Table 6. FTA of the MCU with the proposed strategy

Event number	Event description	Event type	Event occurring probability (1×10^{-4})/h
R	The motor control unit is in abnormal working condition	Top events	-
G7	Address monitoring	Intermediate events	-
G8	Address conflict monitoring	Intermediate events	-
G9	L3 Controller monitor layer failure	Intermediate events	-
G10	Platform function abnormal	Intermediate events	-
G11	L1 Control layer anomaly	Intermediate events	-
G12	Secondary addressing monitor	Intermediate events	-
G13	Address allocation monitor	Intermediate events	-
G14	L2 Control monitor layer exception	Intermediate events	-
X17	Abnormal power supply status	Basic events	0.381 ^①
X18	Reset function abnormal	Basic events	1.4 ^②
X19	Abnormal safety status	Basic events	0.381 ^①
X20	Auto addressing abnormal	Basic events	0.57 ^①
X21	Auto address abnormality	Basic events	0.57 ^①
X22	Secondary addressing abnormal	Basic events	0.57 ^①
X23	Secondary address abnormality	Basic events	0.57 ^①

*Data from expert experience, product technical manuals.

① product technical manuals ②expert experience.

4.2 Quantitative Analysis of the MCU with 3-Layer Design

Based on the value of the probability of occurrence of the bottom layer event given in Table 6, the top layer event failure probability is calculated as $40.40742498 \times 10^{-4}$ /h by Eq. (1), and the new minimum cut-set failure probability and importance of the fault tree of the motor control unit of the USV after adding the 3-layer monitoring architecture is calculated by Eq. (2), which is shown in Table 7.

Table 7. Importance of MCS when motor control unit works abnormally

Number	MCS	Event probability (1×10^{-4})/h	importance rating λ_i
1	{X1}	0.381	0.00942896
2	{X2}	1.4	0.034647098
3	{X3}	1.4	0.034647098
4	{X4}	1.4	0.034647098
5	{X5}	1.4	0.034647098
6	{X8}	8.7	0.215306964
7	{X9}	7.4	0.183134659
8	{X10}	7.4	0.183134659
9	{X13}	0.381	0.00942896
10	{X14}	1.1	0.02722272
11	{X15}	5.7	0.141063183
12	{X16}	3.81	0.094289602
13	{X17,X18,X19,X6,X7}	0.005	0.00012374
14	{X17,X18,X19,X20,X22}	0.0001	2.47479×10^{-6}
15	{X17,X18,X19,X20,X23}	0.0001	2.47479×10^{-6}

5 Conclusion

In this paper, to avoid abnormal working conditions of the motor control unit (MCU) in ANDC, a 3-layer monitoring strategy for the functional safety of the MCU is designed based on the 6-core CPU hardware platform of TC397. The potential hazards and initial event are identified by using HAZOP and Fault Tree Analysis (FTA). Then, the software architecture with 3-layer monitoring strategy is proposed, meeting the MCU functional safety requirements. The introduced 3 layers include control layer (Level 1), control monitoring layer (Level 2), and controller monitoring layer (Level 3), monitoring the state of the MCU and reset the MCU if necessary. To verify the proposed design, FTA is used to quantitatively calculate the failure rate of the new system. The result indicates that the failure rate time duration has been reduced from $64.23 \times 10^{-4}/h$ to $40.41 \times 10^{-4}/h$, showing the potential of the proposed strategy in the practice.

Acknowledgement. This work is supported by the Research Program of Sanya City through Grant No. 2022KJCX36, the National Science Foundation of China (NSFC) through Grant No. 52471384, and Research Program of Zhejiang Provincial Department of Transportation through Grant No. 2024006.

References

1. Durukan, O., Akyuz, E., Destanoğlu, O., et al.: Quantitive HAZOP and DS evidence theory-fault tree analysis approach to predict fire and explosion risk in inert gas system on-board tanker ship. *Ocean Eng.* **308**, 118274 (2024)
2. Congressional Research Service. Navy large unmanned surface and undersea vehicles: background and issues for congress (2020)
3. Waszecki, P., Mundhenk, P., Steinhorst, S., et al.: Automotive electrical and electronic architecture security via distributed in-vehicle traffic monitoring. *IEEE Trans. CAD Integr. Circuits Syst.* **36**(11), 1790–1803 (2017)
4. Chen, L., Xiaomei, Z., Min, W.: Research and practice of data acquisition technology for intelligent ship integration platform. *J. Shanghai Inst. Ship Transp. Sci.* **41**(02), 59–64 (2018). (in Chinese)
5. Xiaofei, W., Guangzhe, W.W., Feng, Z., et al.: Design of unmanned surface vessel navigation state monitoring system based on internet of things. *Shipbuilding Technol.* **52**(01), 19–27 (2024). (in Chinese)
6. Jian, H., Jianjun, Y., Wentao, C.: Analysis of integrated health management system for a certain type of USV. *China Ship Res.* **16**(01), 151–157 (2021). (in Chinese)
7. Yanming, C.: Research on Integrated Power System Health Assessment Model for Large Surface Unmanned Vessels. Wuhan University of Technology, Wuhan (2018). (in Chinese)
8. Bing, G., Qing, H., Houbao, S.: PHM technology for USV perception system. *Ship Electron. Countermeas.* **42**(3), 38–41 (2019). (in Chinese)
9. Yan, Z.: Research on the testing method of VCU control strategy for pure electric vehicles based on functional safety. Chongqing University of Technology (2024). (in Chinese)
10. Zhang, Y., Kuang, X., Xi, W.: Research on design of functional safety monitoring of intelligent driving vehicle software architecture In: International Association of Applied Science and Engineering. Conference proceedings of 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence (ACAi 2021), vol. 6. SAIC-GM-Wuling Automobile Co., Ltd.; CATARC (Tianjin) Automotive Engineering Research Institute Co., Ltd. (2021)



Research on Uplink Signal Processing of Balise Based on VMD Optimization

Qianshu Xue, Jing Xun^(✉), Minxue Fu, and Chenghao Cui

School of Automation and Intelligence, Beijing Jiaotong University, Beijing 100044, China
{24110007, jxun, 24120194, 24120193}@bjtu.edu.cn

Abstract. The noise, caused by electromagnetic interference during the transmission of uplink signals in the balise, could result in a high demodulation error rate at the receiver. A variational mode decomposition method based on optimized parameters is designed to reduce the noise. This method decomposes the signal into several intrinsic mode function components. The sum of the center frequency bandwidth of each intrinsic mode function component is the smallest, which can extract useful frequency components. The Archimedean optimization algorithm is used to optimize the penalty factor and the number of decomposed modes in the variational mode decomposition. Simulation experiments have shown that the method is capable of determining the optimal key parameters. This capability allows it to effectively separate the useful signal from the interference signal. Furthermore, by employing signal reconstruction, the method is able to filter out the interference signal. The experimental results prove that within the signal-to-noise ratio range of 1–10 dB, the bit error rate is reduced by an average of 43% by demodulating the signal reconstructed by variational mode decomposition, compared to the unprocessed noisy signal.

Keywords: Balise · Intrinsic mode function · Variational mode decomposition · Archimedes optimization algorithm

1 Introduction

Railway signal balises are important devices for vehicle-to-ground information transmission in train operation control systems. They transmit stored message information in the form of Binary Frequency-shift Keying (2FSK) signals to the Balise Transmission Module (BTM) on board the train. Due to the presence of radio frequency electromagnetic fields and other electromagnetic interference from train-borne equipment and railway line signal equipment, the demodulation end error rate increases, and the BTM cannot correctly receive the balise message [1].

In the study of balise uplink signal, Li et al. [2] analyzed the impact of different installation angles of the balise on the uplink transmission signal and the dynamic characteristics of the BTM. Geng et al. [3] utilized digital twin technology to analyze the energy and data transmission process between the on-board antenna unit and the balise, and conducted quantitative calculations and analysed on the impact of typical parameters

on the uplink signal. Wang et al. [4] established a balise transmission antenna model based on the principle of electromagnetic induction, analyzing the cause of the appearance of side lobes in the field strength. Moon et al. [5] used a discrete time differentiator to demodulate the signal, reducing the impact of frequency offset on the demodulator. Zhang et al. [6] and Zhang et al. [7] respectively used ensemble empirical mode decomposition and Duffing oscillators to noise reduction process for the uplink signal, but only considered the activation energy of the downlink and random noise, without testing for actual signal interference in the railway, and the mode components decomposed by ensemble empirical mode decomposition cannot clearly observe the effective frequency components. The Variational Mode Decomposition (VMD) method can decompose the signal into amplitude and frequency modulation components with different center frequencies and limited bandwidth, which is helpful for the extraction of useful frequencies and signal analysis and processing.

This paper constructs a 2FSK signal based on the characteristics of the balise uplink signal and the main interference sources in the field environment, uses the VMD algorithm to decompose and process the signal, and refers to the Archimedes Optimization Algorithm (AOA) in [8] to optimize the penalty factor and the number of decomposed modes in VMD, enhancing the correlation between the decomposition results and the original signal. Finally, the VMD decomposition results are analyzed, the uplink signal is reconstructed, and the noise reduction performance of the VMD processed signal is compared.

2 Related Work

2.1 Analysis of Uplink Signal of Balise

The balise transmission system consists of two parts: on-board equipment and ground equipment. The on-board equipment mainly includes the BTM, antenna unit, and on-board computer, while the ground equipment mainly includes the balise and track-side electronic unit [9], as shown in Fig. 1.

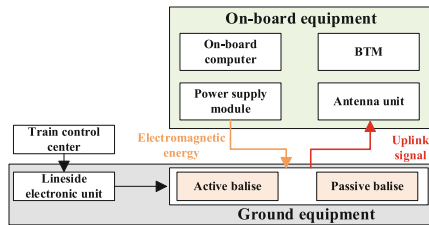


Fig. 1. Balise transmission system structure

When the train passes over the ground balise, the on-board antenna transmits 27.095 MHz electromagnetic energy to activate the balise. The message data stored in the balise, after being modulated by 2FSK, is sent as an uplink signal to the on-board equipment. The parameters of the uplink signal are shown in Table 1.

Table 1. Main parameters of the system

Parameter	Standard value /MHz	Allowable fluctuation range (%)
f_c	4.234	± 4.0
f_d	0.282	± 7.0
f_1	4.516	± 7.0
f_2	3.951	± 7.0

In Table 1, f_c represents the center frequency, f_d represents the frequency deviation, f_1 is the carrier frequency for transmitting '1', and f_2 is the carrier frequency for transmitting '0'. The transmission environment of the balise uplink signal is complex and variable. The main sources of interference include radio frequency electromagnetic field radiation, etc. [9], as shown in Table 2.

Table 2. Source of uplink signal interference for Balise

Interference Source	Frequency (Hz)
Radio frequency electromagnetic field radiation	80–1000 M
Radio telephone radio frequency radiation	800–960 M
Power frequency magnetic field	50
Electromagnetic activation energy magnetic field	27.095 M

2.2 Variational Mode Decomposition

The VMD algorithm redefines the Intrinsic Mode Function (IMF) as the amplitude modulation and frequency modulation component with limited bandwidth [10], and the expression is:

$$u_k(t) = A_k(t) \cos(\varphi_k(t)) \quad (1)$$

In Eq. (2), $A_k(t)$ represents the instantaneous amplitude, and $\varphi_k(t)$ represents the instantaneous phase. The essence of the VMD algorithm is to minimize the sum of the center frequency bandwidths of each IMF under a constrained variational model. Through the iterative solution of the algorithm, several narrowband IMF components are obtained. The constrained variational model expression for VMD is given by [11]:

$$\left\{ \begin{array}{l} \min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ \text{s. t. } \sum_K u_k = f \end{array} \right. \quad (2)$$

In Eq. (2), K represents the number of IMF components, $\{u_k\} = \{u_1, \dots, u_k\}$ represents the IMF components, and $\{\omega_k\} = \{\omega_1, \dots, \omega_k\}$ represents the center frequencies of each IMF component. To solve the above constrained variational problem, an augmented Lagrange function is introduced.

$$\begin{aligned}
 L(\{u_k\}, \{\omega_k\}, \lambda) &= \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\
 &+ \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_k u_k(t) \right\rangle
 \end{aligned} \quad (3)$$

In Eq. (3), α is the penalty factor, and λ are the Lagrange multipliers. The alternating direction multiplier method is used to solve for the optimal solution of the constrained variational model. The iteration formulas for the modal components u_k , center frequencies ω_k , and Lagrange multipliers λ are as follows:

$$\begin{cases}
 u_k^{n+1} = \arg \min_{u_k} L(\{u_{i \neq k}^{n+1}\}, \{\omega_k^n\}, \lambda^n) \\
 \omega_k^{n+1} = \arg \min_{\omega_k} L(\{u_k^n\}, \{\omega_{i \neq k}^{n+1}\}, \lambda^n) \\
 \lambda^{n+1} = \lambda^n + \rho \left(f(t) - \sum_k u_k^{n+1}(t) \right)
 \end{cases} \quad (4)$$

$$\sum_k \frac{\|u_k^{n+1} - u_k^n\|_2^2}{\|u_k^n\|_2^2} < \varepsilon, \quad \varepsilon > 0 \quad (5)$$

In Eq. (4), u_k , ω_k and λ are each calculated iteratively, and the updates are repeated until Eq. (5) is satisfied, where ε is the relative tolerance.

2.3 Based on AOA of Variational Mode Decomposition

Due to the complexity and variability of actual signals, it is challenging to determine the penalty factor α and the number of decomposition modes K in the VMD algorithm, which significantly affects whether the decomposition results meet expectations. To eliminate the influence of human subjective factors, the AOA is employed to find the optimal combination. Compared to other algorithms, the AOA has the advantages of strong local search capabilities and high optimization accuracy. Equation (6) is the calculation formula for envelope entropy.

$$\begin{cases}
 E_P = - \sum_{j=1}^N p_j \lg p_j \\
 p_j = a(j) / \sum_{j=1}^N a(j)
 \end{cases} \quad (6)$$

In Eq. (6), p_j is the normalized form of the $a(j)$, and $a(j)$ is the signal after Hilbert demodulation of the IMF component. This paper uses envelope entropy as the fitness function for the AOA to optimize α and K .

3 Experimental Validation and Results Comparison

3.1 Optimizing VMD Using AOA

Using AOA to Optimize VMD Algorithm, the implementation steps are as follows:

1. Based on actual measurements, the boundary range for α is determined to be [500, 3000], and K is [3, 7].
2. The maximum number of iterations, the number of optimizations, the number of individuals, and other parameters for the AOA are set.
3. $[\alpha, K]$ serve as the individual position coordinates for the optimization process. The iteration process is illustrated in Fig. 2.

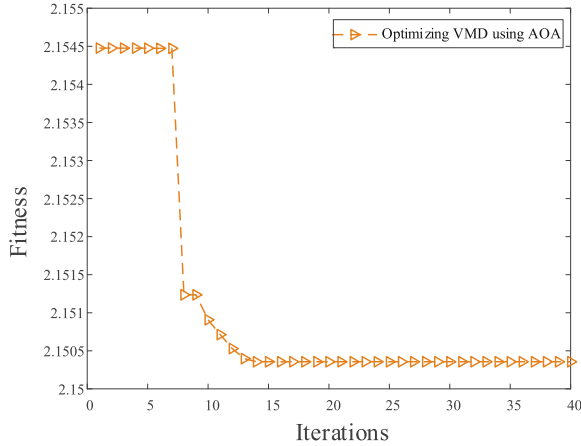


Fig. 2. AOA iteration process

As shown in Fig. 2, the AOA performs optimization searches from the 1st to the 40th iteration. After the 13th iteration, the envelope entropy value stabilizes, with the optimal fitness value being 2.1504. The final output result is [1398, 5], which means α is 1398 and K is 5.

3.2 VMD Decomposition of Uplink Signals

To verify the effectiveness of the VMD algorithm in processing uplink signals, the original uplink signal is generated according to the parameters in Table 1, and the signal expression is as shown in Eq. (7).

$$U_{\text{FSK}}(t) = A \cos(2\pi(f_c \cdot t + f_d \cdot \int s(t)dt) + \varphi) \quad (7)$$

where, A represents the amplitude of the signal, $s(t)$ represents the modulated message sequence, and φ represents the initial phase. The noise signals from the interference sources in Table 2 are added to simulate the on-site environment. Then, VMD was used to decompose the disturbed signal $U_{\text{FSK}}^{\text{NOISE}}(t)$, extract the useful frequency components, with the α set to 1398 and K set to 5, the decomposition result is shown in Fig. 3.

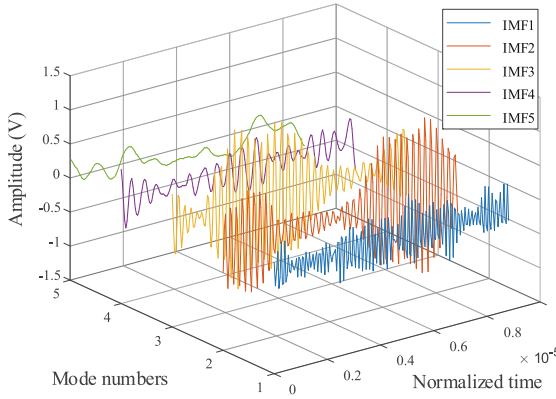


Fig. 3. VMD decomposition mode three-dimensional diagram

According to the principle of VMD decomposition, the frequencies of the decomposed modal signals are arranged from high to low. By analyzing the IMF components in Fig. 3, it is inferred from the changes in signal amplitude that IMF2 and IMF3 correspond to the 4.516 MHz and 3.951 MHz signal components in the original signal, while IMF1, IMF4, and IMF5 are noise signals. Hilbert spectral is performed on IMF2 and IMF3, as shown in Fig. 4.

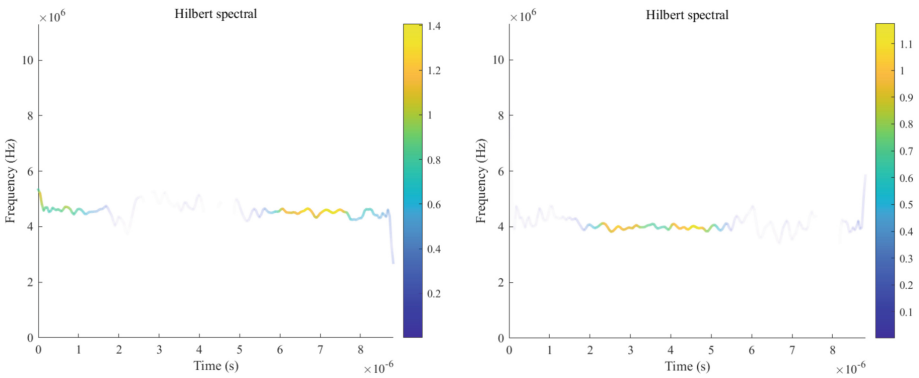


Fig. 4. Hilbert spectral analysis

Comparing Fig. 3 and Fig. 4, it can be observed that IMF2 and IMF3 correspond in terms of time-domain amplitude and energy, with higher energy where the amplitude

is larger. Additionally, at the weak energy regions of IMF2 (2 us-4.6us), there is a corresponding high energy in IMF3, and at the weak energy regions of IMF3 (0 us-1.3 us, 5.5 us-8 us), there is a corresponding high energy in IMF2.

In this experiment, IMF2 and IMF3 are identified as the useful frequency components, while the remaining IMF components are considered noise components. Selecting IMF2 and IMF3 to form the reconstructed signal $U_{\text{FSK}}^{\text{RES}}(t)$. To verify the effectiveness of the VMD algorithm in noise filtering, both $U_{\text{FSK}}^{\text{NOISE}}(t)$ and $U_{\text{FSK}}^{\text{RES}}(t)$ were subjected to demodulation processing, and the error rate was recorded. In this experiment, coherent demodulation was used for signal demodulation, and the demodulation Bits Error Rates (BER) within the Signal-to-Noise Ratio (SNR) range of 1 to 15 dB were recorded. Both the noisy signal and the reconstructed signal were tested 10 times at each SNR level, and the average error rate was taken as the final result. The variation of the error rate is shown in Fig. 5.

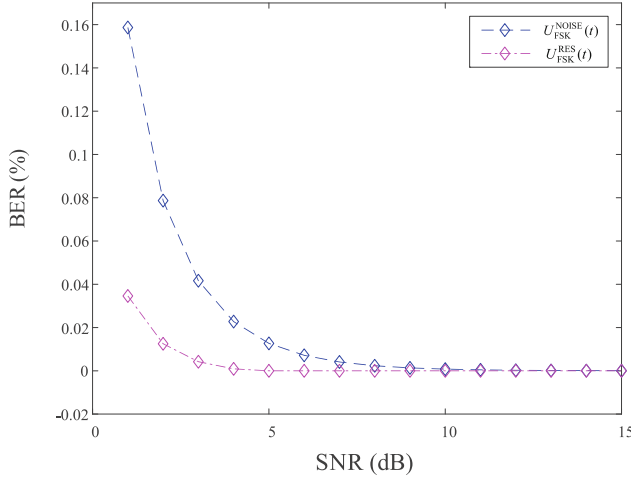


Fig. 5. Error rate comparison

The BER of the signal after coherent demodulation of VMD reconstructed signal in Fig. 5 is lower than that of the unprocessed noisy signal within the SNR range of 1 to 10 dB. After a signal-to-noise ratio of 5 dB, the BER of the noisy signal is 0.0072, and the BER of the reconstructed signal is 10^{-6} , which complies with the railway standards. The BER is reduced by an average of 43% by demodulating the signal reconstructed by VMD, compared to the unprocessed noisy signal.

4 Conclusion

The VMD algorithm was used for noise reduction processing of the uplink signal of the balise, and the AOA was employed to optimize the penalty factor and the number of decomposed modes in the VMD algorithm. The experiment demonstrated that after electromagnetic interference, the uplink signal of the balise could be effectively decomposed

using the VMD algorithm into a 4.516 MHz frequency signal, a 3.951 MHz frequency signal, and noise interference components. By selecting the appropriate IMFs for signal reconstruction, the BER of the demodulated signal was significantly reduced. This confirms the feasibility and effectiveness of the VMD algorithm in noise filtering, which has certain reference significance for on-site railway processing of noise interference in balise uplink signals.

Acknowledgments. This research is supported by National Key R&D Program of China (2024YFE0104400) and Beijing Natural Science Foundation (L231028).

References

1. Ye, K., Jiao, J., Xu, Q., Chen, F., Zhu, L.: Modeling and research on railway balise transmission system for underwater debris. *Appl. Sci.* **14**(16), 7306 (2024)
2. Li, J., Zhang, X., Zheng, Y.: Research on influence of installation angular deviation and train speed on the electromagnetic transmission performance of balise. *J. Electron. Meas. Instrum.* **37**(01), 212–221 (2023). (in Chinese)
3. Geng, Q., Wen, Y., Zhang, D., Xiao, J., Zhu, Y., Zhu, L.: Analysis of electromagnetic coupling characteristics of balise transmission system based on digital twin. *Appl. Sci.* **11**(13), 6002 (2021)
4. Wang, T., Zhao, L.: Cause analysis and application of balise system side-lope based on electromagnetic induction. *J. Chin. Railway Soc.* **41**(04), 111–117 (2019). (in Chinese)
5. Moon, S., Park, S., Hee, J.: Rapid balise telegram decoder with modified LFSR architecture for train protection systems. *IEEE Trans. Circuits Syst. II: Express Briefs* **66**(2), 272–276 (2019)
6. Zhang, Y., Liang, P.: Dispose of balise uplink signal based on ensemble empirical mode decomposition. *J. China Railway Soc.* **41**(03), 86–90 (2019). (in Chinese)
7. Zhang, H., Wang, R.: Application of Lyapunov Exponent Algorithm in balise signal chaotic oscillator detection. *J. Meas. Sci. Instrum.* **12**(03), 281–286 (2021)
8. Thikien, D., Shuchuan, C., Trongthe, N., et al.: An optimal WSN node coverage based on enhanced Archimedes Optimization Algorithm. *Entropy (Basel, Switzerland)* **24**(8), 1018–1037 (2022)
9. Xue, Q., Li, J., Zheng, Y.: Research on demodulation of balise uplink signal based on constant modulus algorithm. In: Ninth International Symposium on Sensors, Mechatronics, and Automation System 2023, ISSMAS, vol. 12981. Spie, Nanjing (2024)
10. Hu, X., Cao, Y., Tang, T., et al.: Data-driven technology of fault diagnosis in railway point machines: Review and challenges. *Transp. Saf. Environ.* **4**(4), tdac036 (2022)
11. Nikhil, K., Subhransu, R., Chandrashekhar, N.: VMD-enabled current-based fast fault detection scheme for DC microgrid. *IEEE Syst. J.* **16**(1), 933–944 (2022)



Real-Time Classification of Freeway Traffic Congestion Levels Based on Surveillance Video

Chuanyun Fu¹, Jinzhao Liu¹, Zhaoyou Lu¹, and Wei Bai^{2,3}(✉)

¹ School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China

² Department of Road Traffic Management, Sichuan Police College, 186 Longtouguan Road, Luzhou 646000, China

baiwei2019@scpolicec.edu.cn

³ Intelligent Policing Key Laboratory of Sichuan Province, Sichuan Police College, 186 Longtouguan Road, Luzhou 646000, China

Abstract. In response to the problem that the analysis of highway congestion in hourly units fails to accurately reflect real-time service levels, this study investigates a method for real-time grading of traffic congestion levels on highways based on video surveillance. The research employs YOLOv5 and the Deep SORT algorithm to extract vehicle trajectories and compute short-term traffic flow macro parameters (including flow rate, density, and speed) and micro parameters such as headway intervals on a one-minute basis. Using the entropy method, weights are assigned to indicators that characterize congestion states to construct a comprehensive congestion measurement index. This index is then used for real-time congestion level classification through Fuzzy C-Means (FCM) clustering, with validation performed using K-means algorithm. The results indicate that the proposed comprehensive congestion measurement index outperforms congestion indicators constructed solely from macro and micro parameters, providing a real-time assessment of traffic flow congestion status for different time periods.

Keywords: traffic congestion · congestion comprehensive index · traffic surveillance video · information entropy · K-means · FCM

1 Introduction

Highways are crucial for large-scale, long-distance transportation, and real-time congestion identification is essential for improving transportation efficiency. This has been a research hotspot in traffic control, with two main identification methods: traditional traffic flow models and machine learning algorithms. Traditional methods use macro models and mathematical equations to describe traffic flow, analyzing flow parameters to determine congestion. Li et al. [2] integrated traffic flow instability analysis for key urban interchanges, while Chen et al. [3] used probability density segmentation to classify vehicle speeds and establish an adaptive congestion index model. Yan et al. [4] employed big data to identify frequently congested areas and occasional events based

on travel time data. These methods are effective for single main roads and relatively uniform traffic flow networks but struggle with the complexity and uncertainty of urban traffic flow.

Machine learning offers advantages like handling complex data, real-time performance, and adaptability across environments and data sources, making it effective for traffic congestion identification. For instance, Ding et al. [5] used CNNs for image-based congestion detection, while Wikan et al. [6] employed MLP and LSTM networks to model traffic data. Muhammad et al. [7] introduced a deep learning model with ELMs for intersection congestion perception, and Imran et al. [8] integrated AI, image processing, and IoT for advanced traffic control. Xiao et al. [9] found that CNN-based feature fusion improves detection accuracy, and Luo et al. [10] combined deep learning and image processing for congestion identification. Zeng et al. [11] proposed a Directed Graph Convolutional Network for expressway congestion. Other methods include CART [12], Bayesian Networks [13, 14], and K-means Clustering [15]. Despite progress, challenges remain, such as model sensitivity to data quality and computational demands for real-time predictions. To effectively alleviate traffic congestion, congestion level classification is also necessary, primarily achieved through clustering methods. Li et al. [16] used system clustering to detect congestion in Beijing, establishing a three-level warning standard (red, orange, yellow).

To address research gaps, this study develops a comprehensive traffic congestion metric combining macro and micro traffic flow parameters from surveillance video for real-time freeway congestion level classification. It proposes a fuzzy C-means clustering method based on this metric. The study contributes by constructing a comprehensive metric for real-time congestion assessment and verifying its superiority over single-parameter metrics.

2 Video-Based Traffic Flow Information Extraction

This study uses computer vision (YOLOv5+Deep SORT) and deep convolutional networks to accurately identify vehicle positions and types in surveillance videos. YOLOv5 extracts features and detects objects, marking vehicle positions with bounding boxes. Deep SORT tracks vehicles by extracting feature vectors, matching them using a deep learning model and the Hungarian algorithm, and updating trajectories via a Kalman filter.

Detection boxes in red unrecognized areas, affecting travel time accuracy. To improve flow parameter accuracy, four lane markers and intervals were chosen, cropping the video frame to exclude the red area, forming a precise detection zone. This ensures accurate distances and travel times. Optimized process extracts precise vehicle trajectories (ID, lane, coordinates, etc.) and entry/exit times, enabling volume, density, speed, count, and headway calculations. For real-time analysis, traffic flow info is collected every minute.

3 Classification of Congestion Level

3.1 Congestion Composite Metric Index Construction

This study applies the entropy method to construct a congestion composite metric index (CCMI) by combining macro parameters such as traffic volume (Q), operating speed (V), and operating density (K) with the micro parameter of headway (h_t).

In information theory [19], entropy measures uncertainty of a stochastic process, i.e., unpredictability. Information amount relates to event probability; higher probability means larger information content. Information entropy represents expected content from all possible outcomes. Entropy values gauge parameter discreteness; smaller values indicate larger weights and greater discreteness. Higher discreteness means larger impact on congestion identification. The entropy method is used to determine parameter weights. Steps are as follows:

Step 1: Select parameters and data samples to construct the following initial parameter matrix:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} \end{bmatrix} \quad (1)$$

where x_{ij} represents the value of the j th parameter in the i th data sample, where $i=1,2,\dots,n$ and $j = 1,2,\dots,m$, with n being the total number of samples and m the number of parameters.

Step 2: To ensure the comparability of each parameter's contribution to the comprehensive index, all parameters need to be standardized so that they have the same dimension and range. In this paper, Min-Max normalization is applied, which is expressed as follows:

$$x_{ij}' = \frac{x_{ij} - x_{ij\min}}{x_{ij\max} - x_{ij\min}} \quad (2)$$

where x_{ij}' represents the value after Min-Max normalization, x_{ij} is the original data, and $x_{ij\min}$ and $x_{ij\max}$ are the minimum and maximum values of the original data, respectively.

Step 3: Calculate the proportion of the j th parameter in the i th data sample as follows:

$$p_{ij} = \frac{x_{ij}'}{\sum_{i=1}^n x_{ij}'} \quad (3)$$

where p_{ij} represents the proportion of the j th parameter in the i th data sample, and x_{ij}' is the normalized parameter value.

Step 4: Calculate the entropy value of the j th parameter as follows:

$$H_j = -\frac{1}{\ln n} \sum_{i=1}^n p_{ij} \ln p_{ij} \quad (4)$$

where H_j represents the entropy value of the j th parameter, and n is the total number of samples.

Step 5: Calculate the entropy weight. The entropy weight increases as the entropy value decreases, indicating that the parameter becomes more significant. The specific calculation is as follows:

$$w_j = \frac{g_j}{m - \sum_j^m H_j} \quad (5)$$

$$g_j = 1 - H_j \quad (6)$$

where g_j represents the entropy difference coefficient of the j th parameter, and w_j is the entropy weight of the j th parameter.

Based on the entropy weight, the comprehensive congestion measurement index T can be expressed as follows:

$$T = w_1 K + w_2 Q + w_3 \gamma_v + w_4 \gamma_{h_t} \quad (7)$$

where w_1 , w_2 , w_3 , and w_4 are the weight coefficients, and K , Q , γ_v , and γ_{h_t} are the four normalized parameters. Here, $\gamma_v = \frac{1}{v}$ and $\gamma_{h_t} = \frac{1}{h_t}$. The larger the value of the CCMI, the more severe the traffic congestion.

3.2 Congestion Level Classification

Traffic congestion levels are classified using clustering based on the congestion composite metric index. Clustering, an unsupervised ML algorithm, handles large datasets and uses preprocessing like association rules and classification for insights, enhancing feature extraction/classification. For this index, data may have noise or complex structures; Fuzzy C-Means (FCM), a soft clustering algorithm, is suitable. FCM allows data points in multiple clusters, using membership degrees to manage non-spherical clusters and offer flexible results. This paper employs FCM for clustering the CCMI, achieving congestion level classification.

To validate FCM clustering and the congestion composite metric index, a two-fold evaluation is done. Internal metrics like Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index assess clustering stability. Then, K-means clusters minute-level traffic density and speed data. The overlap between FCM and K-means results validates the index's practical significance and congestion level classification.

The Silhouette Coefficient evaluates clustering quality by measuring an object's relation to its cluster and nearest neighbor. Values range from -1 to 1; closer to 1 suggests better clustering. The formula is:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (8)$$

where $S(i)$ is the Silhouette Coefficient of point i ; $a(i)$ represents the average distance between point i and all other points within the same cluster; and $b(i)$ represents the average distance between point i and all points in the nearest neighboring cluster.

The Davies-Bouldin Index (DBI) determines the optimal cluster count by measuring overlap. A lower DBI suggests better clustering quality, expressed as:

$$DBI = \frac{1}{m} \sum_{i=1}^m \max_{j \neq i} \left(\frac{\sigma(C_i) + \sigma(C_j)}{d(C_i, C_j)} \right) \quad (9)$$

where m is the number of clusters; i and j represent the cluster numbers; $d(C_i, C_j)$ is the distance between clusters C_i and C_j ; $\sigma(C_i)$ is the average distance between points within cluster i ; and $\sigma(C_j)$ is the average distance between points within cluster j .

The Calinski-Harabasz Index evaluates clustering quality by assessing compactness and separation. A higher value implies better performance. The formula is:

$$CHI = \frac{TSS - WSS}{k(p - 1)} \quad (10)$$

where TSS is the total sum of squares (tracing the sample variance), WSS is the within-cluster sum of squares (intra-cluster variance), k is the number of clusters, and p is the number of features.

The K-means algorithm is simple, intuitive, fast, and widely applicable. It performs well on high-dimensional data with clear boundaries.

The overlap coefficient measures the overlap between two sets, useful when sizes differ. It's defined as the intersection size divided by the smaller set size:

$$\text{Overlap Coefficient} = \frac{|A \cap B|}{\min(|A|, |B|)} \quad (11)$$

where A is set I and B is set II.

4 Case Study

4.1 Case Data Description

In this study, surveillance video data from the Luzhou section of the Chengdu-Chongqing Ring Expressway in China was selected at one location during the peak hours (morning, noon, and evening) of the May Day holiday: 516 km + 80 m. 12 h of two-way traffic flow data was collected for experimental analysis. The selected road sections are basic expressway segments, where traffic flow conditions remain relatively stable, making them suitable for congestion analysis.

The extracted vehicle trajectory information includes each vehicle's ID, lane direction, time of entry into the detection zone, and time of exit from the detection zone.

4.2 Congestion Level Classification Using CCMI

Figure 1 shows the FCM clustering results of the congestion composite metric index for typical periods at the research locations. FCM clustering of the congestion composite metric index classifies traffic into three levels: Level 1 (smooth traffic, no congestion), Level 2 (slight congestion, minor impact), and Level 3 (moderate congestion, slightly reduced flow, still smooth). This detailed classification aids in accurate monitoring and analysis of traffic flow changes, providing valuable data support for traffic management authorities to implement optimization and management measures.

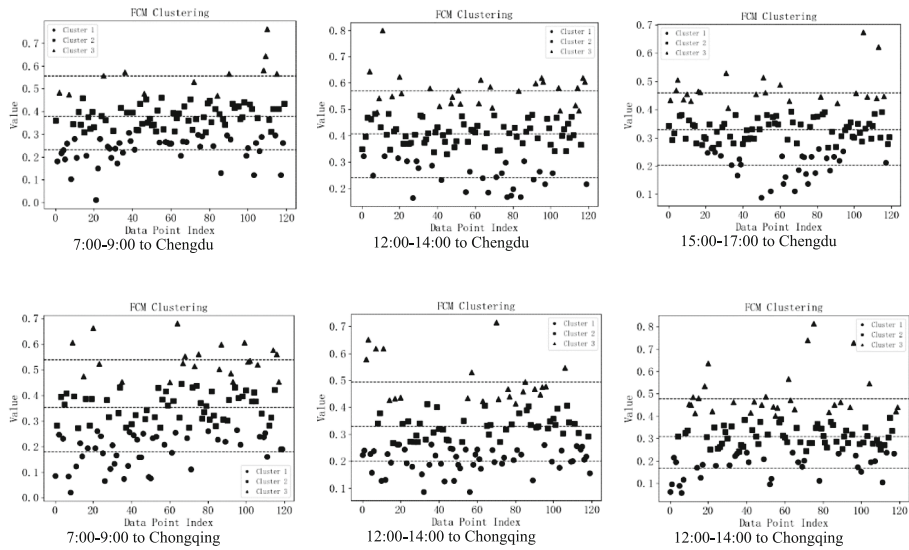


Fig. 1. FCM Clustering Classification Results

Validation of Congestion Clustering Classification. The evaluation metrics of the FCM clustering results are presented in Table 1. FCM clustering results were evaluated using metrics including the silhouette coefficient (0.5 to 0.58), Davies-Bouldin index (0.49 to 0.6), and Calinski-Harabasz index (189.48 to 300.52). These metrics indicate moderate cohesion and separation, good cluster separation and cohesion, and significant cluster separation and cohesion, respectively, demonstrating high quality and effectiveness of FCM clustering on the dataset.

Table 1. Evaluation Metrics of FCM Clustering Results.

Metrics	Range of results	Standard Scope
Silhouette Coefficient	0.5–0.58	–1–1
Davies-Bouldin index	0.49–0.6	$[0, \infty)$
Calinski-Harabasz index	189.48–300.52	$[0, \infty)$

To validate the effectiveness of FCM clustering results, K-means clustering was applied to density-speed data. Traffic flow states are categorized into three levels: low-density-low-speed, low-density-moderate-speed, and low-density-high-speed clusters. The low-density-low-speed cluster represents sparse traffic with lower speeds, typically indicating poor flow and mild congestion. The low-density-moderate-speed cluster refers to sparse traffic with moderate speeds, often seen when traffic flows smoothly but not at high speeds. The low-density-high-speed cluster indicates low traffic volume with high speeds, usually signifying good road conditions and unobstructed flow. These results

validate the practical significance of FCM clustering results for the proposed congestion composite index.

To further assess the superiority of the constructed congestion composite metric index, this study developed a metric T' composed of macroscopic parameters including traffic flow, traffic density, and traffic speed. The expression for this metric is:

$$T'=w_1/K + w_2/Q + w_3/\gamma_v \text{ (14)}$$

where w_1' , w_2' , and w_3' are the weight coefficients, K represents traffic density, Q denotes traffic flow, and γ_v is the speed congestion index.

To compare the effectiveness of the composite metric T , the macroscopic parameter metric T' , and the microscopic parameter index γ_{h_i} in classifying congestion levels, the overlap coefficient between FCM clustering and K-means clustering was computed, as shown in Table 1.

Table 2. Comparison Between FCM Clustering Results and K-means Clustering

Congestion Assessment Index	Maximum Values
T	62.5%
T'	54%
γ_{h_i}	50%

As shown in Table 2, the 62.5% overlap between the two clustering methods exceeds that of macroscopic parameter metrics and microscopic parameter indices, indicating the accuracy and reliability of the constructed congestion composite metric index. The 62.5% overlap in clustering results signifies significant consensus in traffic congestion state classification between FCM and K-means clustering, further validating that the CCMI accurately reflects highway section congestion conditions.

5 Conclusions

In this study, YOLOv5 and Deep SORT algorithms extracted vehicle trajectory data from highway surveillance videos, calculating macroscopic and microscopic traffic flow parameters. The entropy method determined parameter weights to construct a congestion composite metric index for accurately characterizing traffic congestion states. FCM clustering classified traffic congestion levels, and K-means clustered density-speed relationships, validating the practical significance of the composite metric index. Case analysis results show:

- (1) The overlap between FCM clustering results of the congestion composite metric index and K-means clustering results of traffic speed-density is 62.5%, indicating the index effectively evaluates traffic flow changes over time, enabling real-time congestion monitoring and providing valuable data for traffic management.
- (2) The overlap of FCM clustering results of the congestion composite metric index with K-means clustering is higher than that of indicators composed of macroscopic or microscopic parameters alone, suggesting the composite index more accurately reflects congestion conditions compared to single-type parameters.

Acknowledgments. The work was jointly supported by the National Natural Science Foundation of China (72371082), the Natural Science Foundation of Sichuan Province of China (2024NSFSC0184), the Opening Project of Intelligent Policing Key Laboratory of Sichuan Province (ZNJW2023KFZD001), and the Fundamental Research Funds for the Central Universities (FRFCUAUGA5710010222).

References

1. Jia, R., Dai, S., Huang, N., et al.: A review of traffic congestion discrimination methods. *J. South China Univ. Technol. (Nat. Sci. Ed.)* **49**(04), 124–139 (2021). (in Chinese)
2. Li, X., Liang, J., Zhang, W., et al.: A method for identifying interweaving zone congestion based on the construction of horizontal index from aerial video. *J. Railway Sci. Eng.* **20**(02), 494–505 (2023). (in Chinese)
3. Chen, D., Zhou, S., Chen, Y., et al.: Research on traffic operation state identification and application based on adaptive adjustment of congestion index. *J. Transp. Syst. Eng. Inf. Technol.* **22**(02), 137–144 (2022). (in Chinese)
4. Yan, X., Liu, X., Liu, Y., et al.: Urban traffic congestion identification and evaluation based on floating car big data and grid model. *J. Beijing Jiaotong Univ.* **43**(01), 104–113 (2019). (in Chinese)
5. Ding, J., Liu, J., Yang, Z., et al.: Traffic congestion detection based on deep learning. *J. Chongqing Univ.* **44**(04), 107–116 (2021). (in Chinese)
6. Wikan, D.S., Ahmad, S.M.S.: Traffic congestion prediction using multi-layer perceptrons and long short-term memory. In: 2020 10th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS). Malang, Indonesia, pp. 209–212. IEEE (2020)
7. Muhammad, H., Asma, K., Muath, J., et al.: Smart city intelligent traffic control for connected road junction congestion awareness with deep extreme learning machine. In: 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, pp. 1–4. IEEE (2022)
8. Uddin, M.I., Alamgir, M.S., Rahman, M.M., et al.: AI traffic control system based on deep-stream and IoT using NVIDIA Jetson Nano. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), DHAKA, Bangladesh, pp. 115–119. IEEE (2021)
9. Xiao, K., Shi, L., Guo, W., et al.: Multi-dimensional traffic congestion detection based on fusion of visual features and convolutional neural network. *IEEE Trans. Intell. Transp. Syst.* **20**(6), 2157–2170 (2019)
10. Luo, R., Yuan, H., Zhong, F., et al.: Research on road congestion identification based on convolutional neural networks. *J. Zhengzhou Univ. (Eng. Ed.)* **40**(02), 18–22 (2019). (in Chinese)
11. Zeng, Y., Shao, M., Sun, L., et al.: Traffic prediction and congestion control based on directed graph convolutional neural networks. *China J. Highway Transp.* **34**(12), 239–248 (2021). (in Chinese)
12. Sun, M., Wei, H., Li, X., et al.: Detecting traffic congestion points using CART decision tree classification. *Geomatics Inf. Sci. Wuhan Univ.* **47**(05), 683–692 (2022). (in Chinese)
13. Cao, Y., Wang, C., Yang, Y., et al.: Automatic real-time identification of multi-cause urban road traffic congestion based on Bayesian networks. *J. Highway Transp. Res. Dev.* **37**(11), 89–97 (2020). (in Chinese)
14. Cheng, X., Qu, X., Zhang, X., et al.: Analysis of chronic congestion propagation mechanisms based on dynamic Bayesian networks. *J. South China Univ. Technol. (Nat. Sci. Ed.)* **50**(11), 25–34 (2022). (in Chinese)

15. Mondal, M.A., Zeenat, R.: Identifying traffic congestion pattern using k-means clustering technique. In: 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India, pp. 1–5. IEEE (2019)
16. Li, Z., Di, X., Weng, J., et al.: Research on regional classification and evaluation methods based on spatiotemporal characteristics of congestion. *J. Highway Transp. Res. Dev.* **36**(02), 121–127 (2019). (in Chinese)
17. Fu, C., Lu, Z., Ding, N., Bai, W.: Distance headway-based safety evaluation of emerging mixed traffic flow under snowy weather. *Physica A* **642**, 0378–4371 (2024)



Numerical Analysis of the Influence of Free Surface on the Propulsive Efficiency of Biomimetic Fish Bodies

Chaoming Bao¹ , Kang Ren¹, Fei Guo², Daiyu Zhang¹ (✉), and Jiyang Li¹

¹ School of Naval Architecture and Ocean Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, Jiangsu, China

bao.chaoming@just.edu.cn, daiyu.zhang@outlook.com

² Shanghai Merchant Ship Design and Research Institute (SDARI), 2633, Zuchongzhi Road, Pudong, Shanghai 201203, China

Abstract. The study of the effect of free surface on the propulsion efficiency enhancement in fish body motion and its significance in biomimetic fish design is of great importance. This research establishes a numerical model of fish-shaped bodies and their oscillating tail models based on the NACA0012 airfoil. The motion of the fish body is controlled using field functions and dynamic mesh technology, and numerical analyses of the propulsion efficiency of the oscillating fish body under a free surface are conducted using the commercial software STAR-CCM. The study analyzes the impact of free surface distance on propulsion efficiency and reveals the structural distribution characteristics of the flow field under the influence of the free surface. The results indicate that oscillatory propulsion near the free surface can effectively reduce drag, with the drag decreasing as the free surface distance gets closer, although it is also influenced by the motion posture. When the fish body is very close to the free surface, the reactive effects of the free surface cause irregular variations in propulsion efficiency; however, the overall motion characteristics are still maintained.

Keywords: Biomimetic Fish · Free Surface · CFD Simulation

1 Introduction

Fish, with their unique body structure and flexible undulatory motion, demonstrate exceptional efficiency and maneuverability in underwater locomotion. Their undulatory propulsion mechanism is considered one of the most efficient forms of movement in nature, characterized by low energy consumption and high propulsion efficiency, which has garnered significant attention. This characteristic has driven research into biomimetic fish propulsion for applications in underwater robots and submersibles [1]. Particularly under varying environmental boundaries, the motion patterns and efficiency of fish bodies are notably affected, offering new avenues to explore the underlying mechanisms of efficient fish propulsion [2]. Studying the influence of environmental factors on fish body motion contributes to optimizing the performance of biomimetic fish propulsors and enhancing propulsion efficiency in specific environments, such as near the bottom or near the water surface.

The study of oscillatory propulsion in fish has a long history, particularly regarding the effects of factors such as body shape, oscillation patterns, and frequency on propulsion efficiency. Lighthill [3] established a theoretical framework for fish propulsion in his work on "Mathematical Models of Flow around Streamlined Fish Bodies," laying the theoretical foundation for this field. With the development of computational fluid dynamics (CFD) technology, numerical simulation has increasingly become an important tool in the research of biomimetic fish propulsion. CFD software can simulate the movement of fish bodies in fluid, providing detailed distributions of the flow field around the fish, thereby enabling precise calculations of propulsion efficiency and revealing the vortex characteristics generated by tail oscillation. Zhou [4] conducted studies on the autonomous swimming behavior of biomimetic fish under reverse flow conditions. Wolf et al. [5] pointed out that stiffness and frequency significantly affect thrust and exhibit complex interactions. Qin [6] analyzed the influence of movement frequency, oscillation amplitude, and incoming flow velocity on the swimming performance of koi as a biomimetic model.

The free surface effect is a critical factor in fish body motion. When a fish body moves near the water surface, gravitational and surface tension forces induce free surface oscillations. These oscillations not only alter the flow field distribution around the fish body but may also trigger instability in the wake vortices, thereby significantly impacting propulsion efficiency. This study employs STAR-CCM software to perform numerical simulations of fish body undulatory propulsion, aiming to analyze the mechanisms by which the free surface influences propulsion efficiency. By comparing flow field structures, vortex characteristics, and propulsion efficiency under varying water depths, the study examines the specific impacts of free surface oscillations on fish body undulatory propulsion.

2 Numerical Computation Model

To simplify the model, this study uses the NACA0012 airfoil as a substitute for a full two-dimensional fish model. As shown in Fig. 1, a rectangular computational domain is established, with the airfoil chord length L as the characteristic length. The computational domain spans 20 times the characteristic length in the flow direction and 10 times in the vertical direction. For boundary conditions, the right side of the domain is set as a pressure outlet, while all other boundaries are defined as velocity inlets.

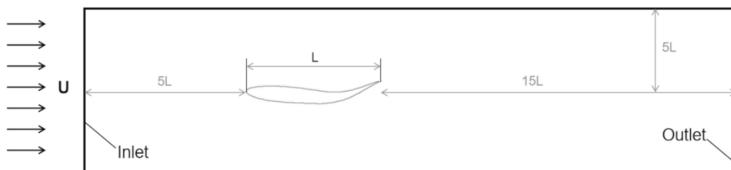


Fig. 1. Two-Dimensional Physical Model

Using the fluid density, airfoil chord length L , and inflow velocity U for nondimensionalization, the resulting dimensionless parameters include the Reynolds number Re .

Tuna species achieve locomotion by generating traveling waves along their bodies that move from head to tail, with the maximum amplitude occurring at the tail [7]. Through experimental observations, Videler and Hess [8] described this undulatory motion using Fourier series, indicating that the fin undergoes sinusoidal motion. They proposed a lateral wave equation to model the transverse oscillation of the fish body:

$$y(x, t) = a(x) \cos(kx - ct) \quad (1)$$

In this equation, x represents the position coordinate along the fish body, and $y(x, t)$ is the lateral position coordinate at time t . The wavenumber k is defined as $k = 2\pi/\lambda$, where λ is the wavelength; c denotes the wave speed; and $a(x)$ represents the waviness curve of fish-shaped bodies within the oscillation period:

$$a(x) = A + Bx + Cx^2 \quad (2)$$

In the equation, A , B , and C are coefficients used to control the oscillation amplitude of the fish-shaped body. Typically, A is set to 0.02 for optimal propulsion performance, and the relationship between B and C is given by $C = -2B$, with $B = -0.0825$ and $C = 0.1625$ [9].

The generation of thrust vortices or drag vortices by oscillating fins depends on whether the wave speed c is greater than or less than 1 [10]. This study selects a typical wave speed of $c = 1.5$ to investigate propulsion performance. Figure 2 illustrates the waviness curve of the fish body at a wave speed of 1.5, with the maximum tail oscillation amplitude a_m reaching 0.1.

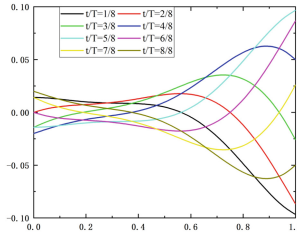


Fig. 2. Fish Body Wave Profile During Oscillation Cycle ($c = 1.5$)

When tunas exhibit two waveforms in their body, the forward speed reaches its peak [11]. In this study, λ is set to 0.5, indicating that the fish body oscillates with two waveforms.

The thrust coefficient and its components for the fish body under various motions are defined as:

$$C_D = \frac{F_D}{\frac{1}{2}\rho U^2 L}, C_L = \frac{F_L}{\frac{1}{2}\rho U^2 L} \quad (3)$$

In the equation, F_D represents the total force acting in the direction of fish movement. When F_D is negative, it indicates that the thrust generated by the fish body's oscillation is less than the drag it experiences. Conversely, when F_D is positive, it indicates that the thrust exceeds the drag. F_L denotes the lift force perpendicular to the direction of the incoming flow.

3 Numerical Method

Considering viscous and incompressible flow, the following continuity equation and motion equation can be established:

$$\frac{\partial u_i}{\partial x_i} = 0 \quad (4)$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{\partial P}{\partial x_i} + \frac{\partial}{\partial x_j} [(\gamma + \gamma_t) \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)] \quad (5)$$

In these equations, u_i ($i = 1, 2$) represents the fluid velocity components, x_i ($i = 1, 2$) denotes the spatial coordinates, P is the fluid pressure, γ is the kinematic viscosity, and $\gamma_t = c_\mu k^2 \varepsilon$ is the turbulent viscosity coefficient. Here, k represents the turbulent kinetic energy, ε is the turbulent dissipation rate, and c_μ is a constant.

In solving the governing equations, the computational domain is discretized using a mesh. This study employs the commercial CFD software STAR-CCM for numerical simulation, using the finite volume method for discretization. The equations are solved based on the Reynolds-Averaged Navier-Stokes (RANS) approach combined with the k - ω turbulence model. Temporal advancement is handled using the PISO algorithm, while the free surface is captured using the Volume of Fluid (VOF) model.

4 Convergence Analysis and Validation

In this study, the computational domain is discretized using a polygonal mesh. Given the significant impact of boundary layer flow characteristics on the propulsion performance of the airfoil, prism layer meshes are specifically arranged around the airfoil, and local mesh refinement is applied near the wall regions, as shown in Fig. 3. This approach aims to capture detailed flow variations within the boundary layer, thereby ensuring accurate resolution of flow features that critically affect the airfoil's propulsion performance.

To ensure the independence and reliability of the numerical simulation results, this study conducted multiple independent simulations based on the principle of control variables at a Reynolds number of 5×10^4 , a wave speed of 1.5, and a maximum amplitude of 0.1. This includes verification of mesh independence and time step independence, as shown in Table 1.

The computational results for various conditions in Fig. 4 show good agreement, indicating that mesh refinement and time step size have minimal impact on the results. Therefore, the Medium-scale mesh and time step will be selected for subsequent numerical calculations.

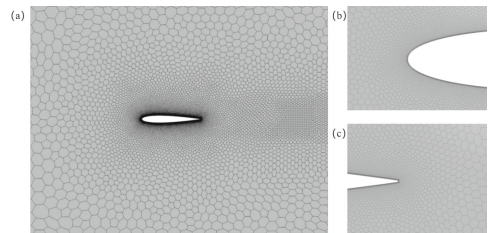


Fig. 3. Spatial Mesh Distribution: (a) Overall Mesh Distribution (b) Enlarged View of Head Mesh (c) Enlarged View of Tail Mesh

Table 1. Mesh and Time Step Convergence Verification

Mesh	Number of Mesh Elements	Time Step Size
Fine	47194	2.8×10^{-3}
Medium	31782	2.8×10^{-3}
Course	24343	2.8×10^{-3}
Mesh	Number of Mesh Elements	Time Step Size
Fine	31782	4.2×10^{-3}
Medium	31782	2.8×10^{-3}
Course	31782	1.4×10^{-3}

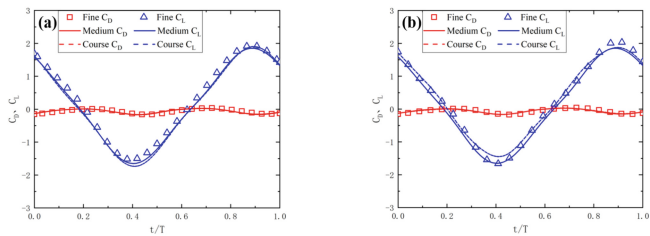


Fig. 4. (a) Mesh Quantity Verification (b) Time Step Verification

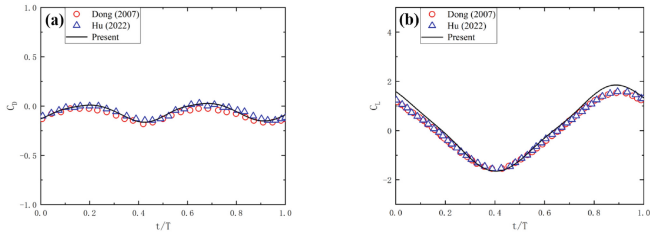


Fig. 5. (a) Drag Coefficient (b) Lift Coefficient

To validate our numerical approach, simulations were conducted following the models used by Dong [12] and Hu [13]. The computational model employs the same two-dimensional NACA0012 airfoil with a Reynolds number of 5×10^4 , and all other motion parameters were kept consistent with those in the references. The drag and lift coefficients of the fish body were calculated, with results shown in Fig. 5. As illustrated in Fig. 5, the results from this study align closely with those in the referenced literature, verifying the reliability of the numerical method employed in this study.

5 Results and Discussion

5.1 Impact of Free Surface on the Flow Field

Here, a fish body oscillating under the free surface is simulated. Each simulation completes at least 30 cycles, and the results of the numerical simulations indicate that the fish body reaches a steady state after 6 oscillation cycles. One of these cycles is selected for demonstration of the motion state.

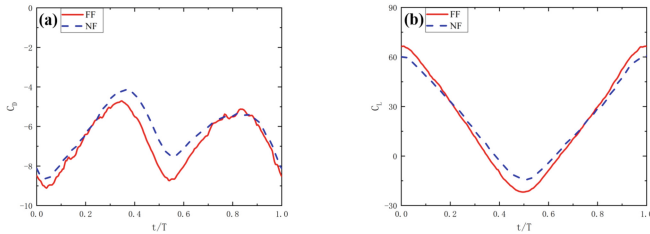


Fig. 6. (a) Drag Coefficient (b) Lift Coefficient

At a distance $d = 0.3L$ from the free surface (NF), the motion of a fish-shaped body with a period of 0.2 s was analyzed. The numerical results obtained were compared with those from a condition further away from the free surface (FF) to investigate the impact of the free surface on thrust. As shown in Fig. 6, the drag coefficient and lift coefficient when close to the free surface are significantly smaller than those observed when further away.

As shown in Fig. 7, the motion of the fish body with a period of 0.2s was analyzed at a distance $d = 0.3L$ from the free surface. The results indicate that at $t = 0.2T$, when the tail fin approaches the free surface and reaches its maximum oscillation amplitude, the drag coefficient attains its first peak value. By $t = 0.4T$, as the tail fin moves away from the free surface and returns to the center of the fish body, the drag coefficient drops to a low value, accompanied by the formation of a strong vortex at the tail. As the motion continues, at $t = 0.6T$, the tail fin is farthest from the free surface, and the oscillation amplitude peaks again, resulting in a second peak in the drag coefficient, although lower than the first peak. At this time, a strong and complete vortex forms at the tail, ready to detach. By $t = 0.8T$, the tail fin returns to the center, and the drag coefficient rises again to a value higher than the previous low, forming a weaker vortex. Ultimately, at $t = 1.0T$, the tail fin returns to its closest position to the free surface, completing one

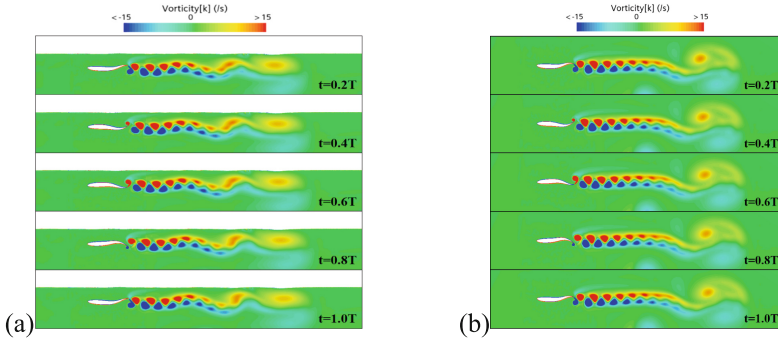


Fig. 7. Comparison of Vorticity Contours (a) with and (b) without Free Surface

cycle. Additionally, the free surface exerts an adhesive effect on the flow field structure; in the absence of a free surface, the flow field is uniform and stable, while near the free surface, vortices are distributed closer to the free surface.

5.2 Effect of Free Surface Distance on Propulsion Efficiency

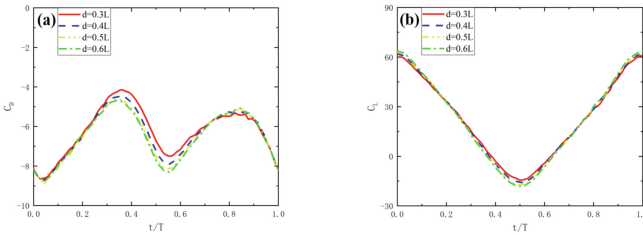


Fig. 8. (a) Drag Coefficient (b) Lift Coefficient

Figure 8 illustrates the variations in drag and lift coefficients during periodic motion. The asymmetric fluctuations of the drag coefficient around zero reflect the disruption of symmetry by the free flow. Both coefficients exhibit periodic oscillations, with the fluctuation frequency of the lift coefficient matching the frequency of the fish body movement, while the drag coefficient oscillates at double the frequency of the fish body. When the fish body moves away from both the free surface and the bottom, both coefficients exhibit sinusoidal fluctuations. Under the influence of the free surface, as the fish body approaches the free surface, the peak value of the lift coefficient increases with greater distance from the free surface, while the changes in the peak value of the drag coefficient are not significant. Conversely, when the fish body moves away from the free surface, both the peak values of the drag and lift coefficients increase with distance.

Figure 9(a) indicates that when the bionic fish is close to the free surface, a low-pressure region forms between the tail of the fish body and the free surface, while a high-pressure region is present on the opposite side. As the tail fin oscillates, two low-pressure centers are generated in the wake. These low-pressure centers are influenced

by the free surface and move in pairs towards the wake during one tail-beating cycle, gradually diffusing backward. As the distance d from the free surface increases, the low-pressure centers exhibit a uniform and orderly arrangement, with their intensity gradually diminishing.

Figure 9(b) shows that as the bionic fish moves within the flow field, high-speed regions form on both sides of the tail. As the distance d between the fish body and the free surface decreases, the high-speed region closer to the free surface gradually intensifies. The oscillation of the tail fin creates a high-speed strip in the wake, flanked by a series of low-speed regions. As the distance to the free surface decreases, the structure of the high-speed strip becomes increasingly complete.

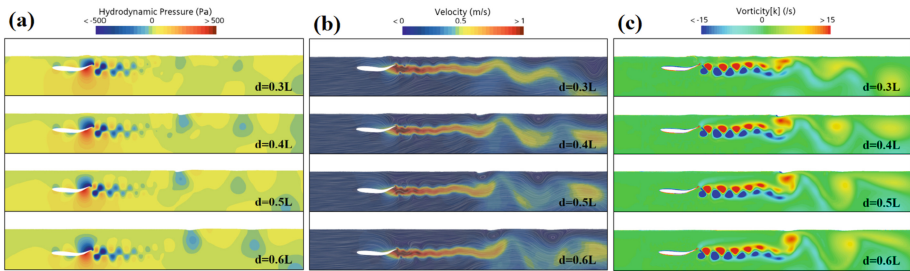


Fig. 9. Flow Field Structure at Different Distances from the Free Surface (a) Pressure (b) Velocity (c) Vorticity ($d = 0.3$ to 0.6)

Figure 9(c) illustrates that during one tail-beating cycle, the bionic fish releases two vortices from the tip of the tail fin: a clockwise vortex is generated when the tail fin moves away from the free surface, while a counterclockwise vortex is produced when it approaches the free surface. These two vortices couple to form a "vortex pair" that flows along the wake and gradually moves away under the influence of the free surface, with the angle between the free surface and the vortex pair decreasing as the distance d from the free surface increases. When d decreases, the free surface is affected by the motion of the bionic fish, resulting in fluctuations, and a portion of the tail end of the "vortex pair" is "adhered" to the free surface.

5.3 "Fragmentation" Phenomenon at Extremely Close Distances

We further analyzed the case with a minimal distance from the free surface, setting $d = 0.2 L$, and compared it with the motion state at $d = 0.4 L$. Figure 10 displays the variations in lift and drag coefficients over a complete cycle and their periodic changes after reaching a stable state. The results indicate that the drag coefficient exhibits irregular fluctuations: while the tail swings towards the free surface, the drag coefficient roughly maintains its original pattern. However, when the tail fin moves away from the free surface, the influence of fluctuations causes the drag coefficient curve to display a fragmented behavior. In contrast, the lift coefficient curve is minimally affected, demonstrating good stability.

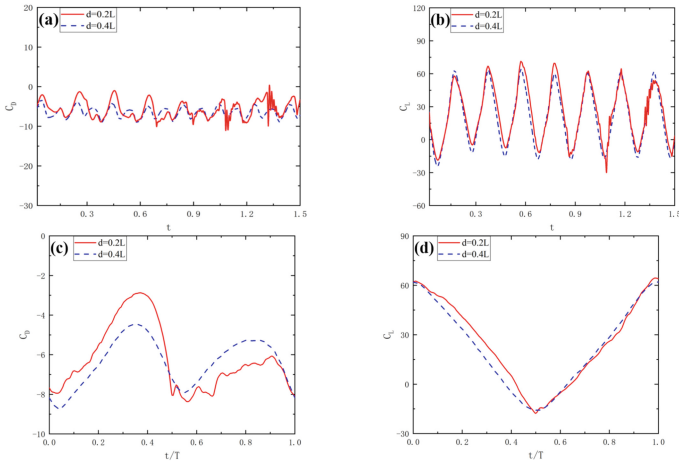


Fig. 10. “Fragmentation” Phenomenon (a) Overview of Drag Coefficient (b) Overview of Lift Coefficient (c) Drag Coefficient over One Cycle (d) Lift Coefficient over One Cycle

Figure 11 shows that when the fish body is very close to the free surface, the oscillation of the tail fin significantly disturbs the free surface, causing it to fluctuate with the movement of the tail fin, even resulting in splashing. At the same time, the fragmented free surface exerts a certain reaction on the movement of the fish body, but it does not completely interfere with its motion. During this process, the vortex pair dissipates more quickly, and the velocity and pressure fields decay more rapidly, affecting the lift and drag coefficients. However, the fundamental movement characteristics of the fish body are still maintained.

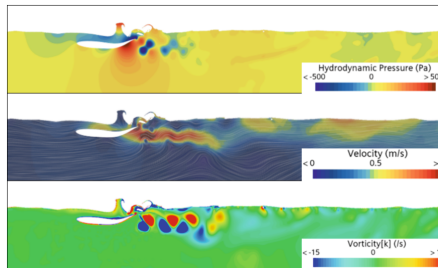


Fig. 11. Flow Field Structure Distribution Under the “Fragmentation” Phenomenon

6 Conclusion

This study investigates the motion of flexible fins beneath the free surface and the propulsion efficiency and flow field distribution at different distances from the free surface. The results are as follows:

1. The presence of the free surface has a positive effect on the flow field. Specifically, the drag and lift experienced by the fish body during motion are both lower when the free surface is present compared to the scenario without a free surface.
2. As the distance to the free surface decreases, the drag and lift coefficients of the fish body during oscillation gradually decrease. However, this is closely related to the motion posture: when the tail of the fish body reaches its maximum amplitude, the drag and lift coefficients are nearly consistent. Conversely, when the tail is positioned at the centerline, the closer the free surface, the lower the drag and lift coefficients.
3. In cases where the fish body is extremely close to the free surface, its oscillation can cause the free surface to break up, resulting in a reaction that complicates the variations in drag and lift. Nevertheless, the overall movement characteristics of the fish body remain well-preserved.

Funding. This research was supported by the National Natural Science Foundation of China under Grant No. 52201364 and 52372356. The Doctor of Innovation and Entrepreneurship Project of Jiangsu Province, China (approval no. JSSCBS20221267).

References

1. Triantafyllou, M.S., Triantafyllou, G.S., Yue, D.K.P.: Hydrodynamics of fishlike swimming. *Ann. Rev. Fluid Mech.* **32**(1), 33–53 (2000)
2. Li, G., et al.: Underwater undulating propulsion biomimetic robots: a review. *Biomimetics* **8**(3), 318 (2023)
3. Lighthill, S.J.: *Mathematical biofluidynamics*. Soc. Ind. Appl. Math. (1975)
4. Zhou, J., et al.: Study on the autonomous swimming behavior of bionic fish under countercurrent conditions. *Chinese J. Appl. Mech.* **39**(05), 908–914 (2022). (in Chinese)
5. Wolf, Z., et al.: Fish-like aquatic propulsion studied using a pneumatically-actuated soft-robotic model. *Bioinspir. Biomim.* **15**(4), 046008 (2020)
6. Qin, W., et al.: Analysis of oscillating propulsion characteristics based on koi BCF. *Sci. Technol. Eng.* **23**(08), 3200–3206 (2023). (in Chinese)
7. Domenici, P., Blake, R.W.: The kinematics and performance of fish fast-start swimming. *J. Exp. Biol.* **200**(8), 1165–1178 (1997)
8. Videler, J.J., Hess, F.: Fast continuous swimming of two pelagic predators, saithe (*Pollachius virens*) and mackerel (*Scomber scombrus*): a kinematic analysis. *J. Exp. Biol.* **109**(1), 209–228 (1984)
9. Videler, J.J.: Fish swimming. *Chapman & Hall google schola* **2**, 192–199 (1993)
10. Müller, U.K., Stamhuis, E.J., Videler, J.J.: Riding the waves: the role of the body wave in undulatory fish swimming. *Integr. Comp. Biol.* **42**(5), 981–987 (2002)
11. Curet, O.M., et al.: Mechanical properties of a bio-inspired robotic knife-fish with an undulatory propulsor. *Bioinspiration Biomimetics* **6**(2), 026004 (2011)
12. Dong, G.J., Lu, X.Y.: Characteristics of flow over traveling wavy foils in a side-by-side arrangement. *Phys. Fluids* **19**(5) (2007)
13. Hu, Q.-Q., Yu, Y.-L.: The hydrodynamic effects of undulating patterns on propulsion and braking performances of long-based fin. *AIP Adv.* **12**(3) (2022)



Real-Time LiDAR Point Cloud Semantic Segmentation for Unstructured Road Autonomous Driving

Huazhi Li^{1,2} , Guizhen Yu^{1,2}, Zhangyu Wang^{1,2} , Yang Chen^{1,2},
and Fei Zhao³

¹ School of Transportation Science and Engineering, Beihang University,
Beijing 100191, China

{lihuazhi, zywang}@buaa.edu.cn

² State Key Lab of Intelligent Transportation System, Beijing, China

³ School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China

Abstract. Lidar-based semantic segmentation is critical for autonomous driving, but existing methods struggle with real-time performance in unstructured environments like open-pit mining areas. This paper proposes a real-time semantic segmentation network leveraging an attention mechanism to improve accuracy and speed. Key innovations include a global contextual information extraction module, bilinear upsampling layers to reduce channel redundancy, and coordinate attention-based upsampling to alleviate structural redundancy. This paper tested the proposed method on an unstructured road open-pit mining dataset, achieving an average IoU of 85.6% and an inference speed of 5.90ms, which is 4.3% and 14.6% higher than the baseline (SalsaNext) in terms of accuracy and speed, respectively.

Keywords: Point cloud · semantic segmentation · unstructured road · attention mechanism

1 Introduction

The rapid development of artificial intelligence has brought advanced key technologies to the field of autonomous driving. Open-pit mining is emerging as a viable scenario for the implementation of autonomous vehicles due to its closed roads and relatively simple environment. Point cloud semantic segmentation plays a crucial role in autonomous driving by providing fine-grained environmental context for decision-making and planning, making it essential for unmanned mining vehicles. However, existing semantic segmentation networks face challenges in open-pit mining areas due to: (1) significant variations in obstacle sizes, which small convolutional kernels struggle to capture globally, leading to low accuracy in detecting different scales; (2) the introduction of channel redundancy with increasing feature dimensions and concatenation for upsampling, hindering accuracy improvement; (3) large model sizes and long inference times, limiting real-time deployment.

To address these challenges, this paper proposes a real-time LIDAR point cloud semantic segmentation network for autonomous driving on unstructured roads. The key contributions are:

- 1) Global Semantic Information Extraction Module: Multiple branches capture semantic spatial information at different scales, and 1×1 convolutions learn inter-channel relationships to improve segmentation accuracy.
- 2) Bilinear Upsampling Layer: This layer prevents channel redundancy by using interpolation for Range View (RV) images, and lightweight coordinate attention guides the network to learn both dimensional and spatial information, reducing model redundancy and improving inference speed.

The rest of the paper is organized as follows: Sect. 2 reviews related works; Sect. 3 details the methodology; Sect. 4 presents experimental results; and Sect. 5 concludes with future directions.

2 Related Work

In recent years, the mainstream point cloud semantic segmentation methods have proposed. These methods can be divided into four categories as point-based methods, voxel-based methods, projection-based methods, and fusion-based methods.

The point-based methods usually directly process the raw point cloud data without using any other preprocessing operations [1–3]. Reference [1] introduced Multilayer Perceptron (MLP) for point cloud feature learning, while [2] proposed Ball Query for local feature extraction. Methods like KPConv [3] improve accuracy by defining convolution kernels based on Euclidean distances. Point-based models preserve structure but struggle with large-scale scenes due to high parameter counts, often requiring combinations with other methods for better efficiency.

Voxel methods segment point clouds into uniform grids. FCPN [4] uses pyramid structures for multi-scale voxel features. However, voxel-based point clouds are sparse, and applying 3D dense convolution on sparse data can lead to unnecessary computation. To address this issue, OctNet [5] employs octree structures for efficient convolution. SSC [6] further optimizes OctNet by removing empty voxels to improve computation. However, voxel-based methods often lose structural features, impacting segmentation accuracy.

The projection-based semantic segmentation method maps the original point cloud onto a two-dimensional plane through spherical projection. Reference [7] used spherical mapping with CRFs, while others [8–13] introduced modules like Context Aggregation and Spatially Adaptive Convolutions to improve segmentation. Projection methods provide speed but sacrifice some accuracy, making them suitable for real-time applications. Recent innovations like FIDNet [14] and PolarNet [16] improve performance by balancing semantic details and computational efficiency.

In addition, some scholars have combined different methods to achieve a balance between accuracy and speed. Cylinder3D [17] addresses LiDAR sparsity, while AF2S3Net [18] balances accuracy and speed. These methods, though effective, are complex and challenging for real-time deployment on mobile units.

To address these limitations, this paper proposes a projection-based approach that strikes a balance between speed and accuracy for point cloud segmentation. Compared to point-based methods, it requires fewer parameters and offers faster inference, making suitable for real-time tasks in autonomous driving. Additionally, unlike grid-based methods, our approach is better suited to handle unstructured environments and improves perception accuracy, particularly in complex scenarios like open-pit mining.

3 Methodology

We propose a novel projection-based point cloud segmentation method, as shown in Fig. 1. The method including point cloud represent, backbone, and post-processing.

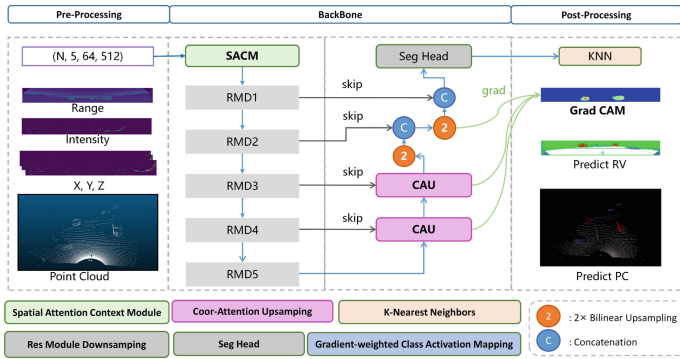


Fig. 1. The overall framework of the network

3.1 Point Cloud Represent

A projection-based semantic segmentation approach is utilized, where the original point cloud is mapped to a two-dimensional plane through spherical projection to obtain the input for the network. For a point cloud with 3D coordinates (x, y, z) , the formula for mapping to a pixel point (u, v) is:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}[1 - \arctan(y, x)\pi^{-1}]w \\ [1 - (\arcsin(z, r^{-1}) + f_{down})f^{-1}]h \end{pmatrix}, \quad (1)$$

where w and h are the width and height of the RV image, respectively, f is the vertical field of view of the lidar, f_{down} is the minimum vertical field of view that the lidar can scan, and r is the Euclidean distance from the origin to a point scanned by the lidar.

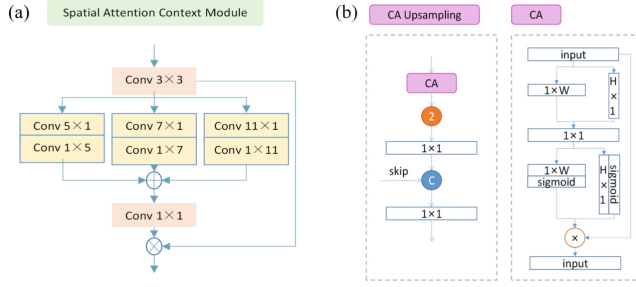


Fig. 2. The framework of Spatial Attention Context Module and Coordinated Attention. (a) Spatial Attention Context Module framework. (b) Coordinated Attention framework.

3.2 Backbone

The proposed Backbone architecture comprises three integral components: Spatial Attention Context Module, Bilinear Upsampling Layer, and Coordinated Attention. The framework of Spatial Attention Context Module and Coordinated Attention is shown in Fig. 2 (a) and Fig. 2 (b), respectively.

Spatial Attention Context Module. This module enhances the network’s ability to focus on relevant spatial features by employing an attention mechanism that emphasizes significant regions within the input data. The framework is as shown in Fig. 2 (a).

Small convolutions typically imply a small receptive field, making it challenging for the network to extract global semantic information. To further extract global contextual information, we use a convolutional attention [19] module before downsampling. The structure of the convolutional attention module includes a 3×3 depth convolution to aggregate local information, multi-scale convolutions to capture context information with different receptive fields and fuse them together. The fused feature map is then subjected to a 1×1 convolution to learn relationships between different channels. Finally, through a residual structure, the output of the initial 1×1 convolution is used as attention weights to guide the learning of spatial and channel features at different scales. To reduce the number of parameters, the multi-scale convolution part employs asymmetric convolutions [24] to extract features from both the horizontal and vertical directions of the feature map. Then this module can be expressed as follows:

$$\alpha = \text{Conv}_{1 \times 1}(\mathbf{F}_{in}), \quad (2)$$

$$\mathbf{Att} = \text{Conv}_{1 \times 1}\left(\sum_{i=0}^3 \text{Scale}_i(\mathbf{F}_{in})\right), \quad (3)$$

$$\mathbf{F}_{out} = \text{Conv}_{1 \times 1}(\mathbf{Att} \otimes \alpha), \quad (4)$$

where a represents the weight of channel features, F_{in} represents the input feature map, Att represents the weight of spatial features obtained through convolution with a 1×1 kernel, $\text{Conv}_{1 \times 1}$ represents the convolution operation using a 1×1 kernel, and Scale_i represents the usage of asymmetric convolutions at different scales. It is important to

note that we have replaced two layers of the original semantic extraction module with the convolutional attention module. This substitution has been found to maintain the accuracy of the semantic segmentation task while reducing the model's complexity.

Bilinear Upsampling Layer. This module facilitates the reconstruction of high-resolution outputs from low-resolution feature maps.

Existing networks often use Pixel-shuffle layer for upsampling, which directly converts the channel dimension into the spatial dimension, thereby resulting in significant channel redundancy. Therefore, the original Pixel-shuffle layer is replaced with a bilinear interpolation layer. This layer performs linear interpolation between four adjacent pixels to achieve feature map upsampling and effectively to improve the accuracy of the semantic segmentation task while reducing channel redundancy.

Coordinated Attention. Insufficient awareness of positional and dimensional information forces us to continually extract features using redundant ResBlocks, thereby exacerbating structural redundancy in the upsampling process. Therefore, the Coordinated Attention module [20] is introduced in this study, which embeds positional information into the channel dimension to coordinate the relationship between channels and spatial dimensions. The framework is as shown in Fig. 2 (b).

Specifically, we first perform two-dimensional global pooling on the input feature map to obtain feature vectors along different spatial directions. This helps capture the relative spatial relationships and position information. Moreover, we encode the feature vectors into an attention map using a sigmoid function. This attention map is sensitive to both direction and position, guiding the learning of objects in the input feature map by highlighting their importance based on spatial and positional cues.

3.3 Post-processing

Post-processing is introduced to address the issue of misclassification at the edges of obstacles when back-projecting RV images to 3D point clouds, where multiple lidar points are mapped to the same pixel. To tackle this problem, we employ the K-Nearest Neighbors (KNN) algorithm as [21]. This optimization step remaps the category of the current search origin to the category with the highest count among the N points, thereby improving the classification results.

4 Experimental Analysis

4.1 Setting and Datasets

Experiments were conducted using an 8GB GeForce RTX 3060Ti graphics card to train the model, considering factors such as optimization speed and the robustness of hyperparameters. The Gradient Descent with Momentum (GDM) algorithm was employed as the network optimizer to enhance convergence speed.

To evaluate the effectiveness of the proposed method, we collected data from typical unstructured road open-pit mining area and created a dataset in the Semantic KITTI format [22]. The data was acquired from coal and iron mines in Inner Mongolia, China,

resulting in 11,745 frames. The dataset consists of five categories: cars, mining trucks, pedestrians, roads, and other classes. Data was captured using an Ouster-64 LiDAR sensor mounted at the front of a truck, with a horizontal field of view of 180°. The LiDAR installation position is shown in Fig. 3.



Fig. 3. Data Collection Vehicle

4.2 Evaluation Metric

For the task of 3D semantic segmentation, mean accuracy (mAcc), Class Intersection over Union (IoU), and mean Intersection over Union (mIoU) are commonly used metrics to evaluate the accuracy of segmentation algorithms. The mining dataset used in this study include five categories: cars, mining trucks, pedestrians, roads, and others. True positives and true negatives are denoted as P_{ii} and P_{jj} , while false positives and false negatives are denoted as P_{ij} and P_{ji} , respectively.

- (1) mAcc: mAcc calculates the average accuracy over all classes by considering the ratio of correctly classified points to the total number of points in each class, which can be expressed as follows:

$$mAcc = \frac{1}{k+1} \sum_{i=0}^k \frac{P_{ii}}{\sum_{j=0}^k P_{ij}}, \quad (5)$$

where k is the number of categories.

- (2) IoU: IoU is one of the metrics used in semantic segmentation. It calculates the intersection over union ratio between the ground truth and predicted values for each class, can be represented as:

$$IoU = \frac{1}{n} \sum_{i=0}^n \frac{P_{ii}}{\sum_{j=0}^n P_{ij} + \sum_{i=0}^n P_{ji} - P_{ii}}, \quad (6)$$

where n represents the total number of points for a given class.

- (3) mIoU: mIoU is another metric used in semantic segmentation. It calculates the average intersection over union ratio across all classes, can be represented as:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{P_{ii}}{\sum_{j=1}^k P_{ij} + \sum_{i=0}^k P_{ji} - P_{ii}}, \quad (7)$$

where k represents the number of classes.

4.3 Accuracy and Efficiency

The segmentation results on the mine dataset are shown in Fig. 4. While all methods are able to segment pedestrians, our method demonstrates superior accuracy, achieving the highest performance among the compared approaches.

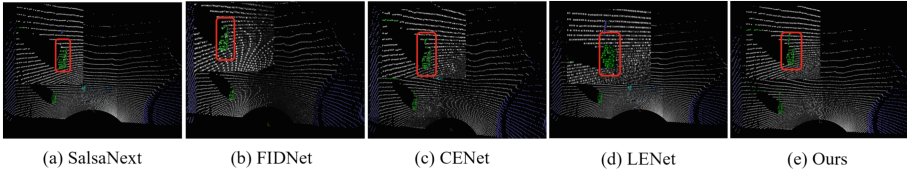


Fig. 4. Results on mine dataset. Our result exhibit more accurate than others.

Table 1 shows the results of recent semantic segmentation methods based on spherical mapping on our dataset. Our network achieves a balance between overall accuracy and inference speed. In terms of accuracy, mIoU performs best on the mining dataset, reaching 85.6%, which is a 4.2% improvement over our baseline and a 0.8% improvement over the previous best network as FIDNet [14]. In addition, we have also made significant improvements in speed, with a 14.6% improvement over our baseline, reducing the inference time from 6.91ms to 5.90ms.

4.4 Ablation Experiment

In this section, we conducted several ablation experiments on the validation set of the mining dataset to validate the improvement of each module in our proposed network. For each experimental setup, the size of the input RV image was 64×512 . Table 2 presents the performance of each module on the validation set.

Firstly, we replace the original pixel shuffle layer with a bilinear upsampling layer in SE [25]. The main purpose of this change is to address the negative impact of channel redundancy. As shown in the table, this results in a 0.1% increase in our mIoU while sacrificing only a small number of computational resources.

Secondly, we replace the context module in the baseline network with a convolutional attention module in CBAM [26]. We use larger convolutional kernels to expand the receptive field and fuse feature maps of different sizes to improve the detection accuracy of obstacles at different scales. Compared to the network with the bilinear upsampling

Table 1. Network accuracy and inference time test table based on spherical projection

Methods	IoU(%)					mIoU(%)	mAcc (%)	Infer Time (ms)
	C	T	P	R	O			
RangeNet ++ [21]	38.0	36.3	92.4	91.2	92.1	70.0	95.4	15.83
SqueezeSegV3[9]	57.7	56.8	94.8	93.5	94.2	79.4	96.7	50.00
SalsaNet [10]	43.2	47.5	92.1	91.7	92.3	73.4	95.6	7.83
SalsaNext [11]	63.7	65.9	94.7	92.5	93.6	82.1	96.3	6.91
FIDNet [14]	70.2	72.2	94.5	93.2	94.3	84.9	96.7	19.71
CENet [23]	69.5	70.9	94.8	92.7	93.8	84.3	96.7	14.3
LENet [15]	76.3	61.5	94.3	93.4	94.2	84.0	96.7	30.03
Ours	71.1	74.8	94.8	93.0	94.3	85.6	96.7	5.90

* The abbreviations used in this table are as follows: C = Car, T = Truck, P = Pedestrian, R = Road, and O = Others.

Table 2. Performance of Upsampling Modules with Different Attentions on Mining Dataset

Methods	IoU(%)					mIoU(%)	mAcc (%)	Infer Time (ms)
	C	T	P	R	O			
SalsaNext [11]	63.7	65.9	94.7	92.5	93.6	82.1	96.3	6.91
SE [25]	57.6	71.9	94.6	93.1	94.0	82.2	96.6	6.64
CBAM [26]	65.2	70.1	94.2	93.4	94.2	83.4	96.7	6.72
Ours	71.1	74.8	94.8	93.0	94. 3	85.6	96.7	5.90

* The abbreviations used in this table are as follows: C = Car, T = Truck, P = Pedestrian, R = Road, and O = Others.

module replaced, this leads to a 1.2% increase in accuracy, although there is some increase in inference speed, it remains within an acceptable range.

As shown in Table 3, we compared several attention algorithms used in recent years. By simply replacing them with the Coordinated Attention Module, we found that the Coordinated Attention Module provides the best guidance for both channel and spatial information. Compared to other algorithms, the Coordinated Attention Module has the shortest inference time and the highest accuracy.

Furthermore, we replace the redundant ResBlock with a coordinating attention module. This significantly reduces the inference time while focusing on areas of interest. Compared to the baseline network, this results in a 4.3% and 13.3% improvement in accuracy and speed, respectively, which is more suitable for real-time semantic segmentation in mining.

Table 3. The performance of different modules on the mine data set

Methods	IoU(%)					mIoU(%)	mAcc (%)	Infer Time (ms)
	C	T	P	R	O			
SalsaNext [11]	63.7	65.9	94.7	92.5	93.6	82.1	96.2	6.91
interpolate	63.9	67.8	94.8	91.6	92.9	82.2	96.3	7.14
Conv attention	65.9	67.2	94.9	93.0	93.9	83.0	96.5	6.47
Coor Attention	71.1	74.8	94.8	93.0	94.2	85.7	96.7	5.99

* The abbreviations used in this table are as follows: C = Car, T = Truck, P = Pedestrian, R = Road, and O = Others.

5 Conclusion

In summary, this paper proposes a lightweight segmentation method for non-structured roads and validates it on a mining dataset. The network reduces channel redundancy through bilinear upsampling, resulting in a 0.1% increase in overall accuracy compared to the baseline network. The use of convolutional attention modules to expand the receptive field and fuse feature maps from different receptive fields leads to a 2.3% increase in overall accuracy compared to the baseline network. The coordinating attention module focuses the network more on areas of interest during upsampling, simplifying the upsampling structure. Compared with the baseline (SalsaNext), the proposed method achieves 4.3% and 14.6% improvements in accuracy and speed, respectively.

Although we have achieved real-time semantic segmentation and deployed it in the field, our experiments have shown that point cloud mapping can lead to a loss of some three-dimensional structural features, resulting in decreased accuracy and the inability to guarantee segmentation accuracy. Combining point-based methods can effectively improve segmentation accuracy but may significantly reduce efficiency. Therefore, how to integrate various point cloud representation methods to balance segmentation accuracy and efficiency remains one of the key research directions in the field of autonomous driving point cloud segmentation.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant #52402488.

Reference:s

1. Charles, R.Q., Su, H., Kaichun, M., et al.: PointNet: deep learning on point sets for 3D classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 77–85. IEEE (2017)

2. Qi, C.R., Yi, L., Su, H., et al.: Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems, p. 30 (2017)

3. Thomas, H., Qi, C.R., Deschaud, J.E., et al.: KPConv: flexible and deformable convolution for point clouds. In: IEEE/CVF International Conference on Computer Vision, pp. 6411–6420. IEEE, Piscataway (2019)

4. Rethage, D., Wald, J., Sturm, J., et al.: Fully-convolutional point networks for large-scale point clouds. In: European Conference on Computer Vision (ECCV), pp. 596–611. IEEE, Piscataway (2018)
5. Riegler, G., Osman Ulusoy, A., Geiger, A.: OctNet: learning deep 3D representations at high resolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3577–3586 (2017)
6. Graham, B., Engelcke, M., Van, D.M.L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9224–9232. IEEE, Piscataway (2018)
7. Wu, B., Wan, A., Yue, X., et al.: Squeezeseg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D Lidar point cloud. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1887–1893. IEEE, Piscataway (2018)
8. Wu, B., Zhou, X., Zhao, S., et al.: Squeezesegv2: improved model structure and unsupervised domain adaptation for road-object segmentation from a Lidar point cloud. In: International Conference on Robotics and Automation (ICRA), pp. 4376–4382. IEEE, Piscataway (2019)
9. Xu, C., Wu, B., Wang, Z., et al.: Squeezesegv3: spatially-adaptive convolution for efficient point-cloud segmentation. In: European Conference on Computer Vision, pp. 1–19. Springer, Berlin (2020)
10. Aksoy, E.E., Baci, S., Cavdar, S.: SalsaNet: fast road and vehicle segmentation in Lidar point clouds for autonomous driving. In: IEEE Intelligent Vehicles Symposium (IV), pp. 926–932. IEEE, Piscataway (2020)
11. Cortinhal, T., Tzelepis, G., Erdal Aksoy, E.: SalsaNext: fast, uncertainty-aware semantic segmentation of LiDAR point clouds. In: International Symposium on Visual Computing, pp. 207–222. Springer, Berlin (2020)
12. Qiu, H., Yu, B., Tao, D.: GFNet: geometric flow network for 3D point cloud semantic segmentation. arXiv preprint [arXiv:2207.02605](https://arxiv.org/abs/2207.02605) (2022)
13. Graham, B., Engelcke, M., Van Der Maaten, L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9224–9232. IEEE, Piscataway (2018)
14. Zhao, Y., Bai, L., Huang, X.: FidNet: Lidar point cloud semantic segmentation with fully interpolation decoding. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4453–4458. IEEE (2021)
15. Ding, B.: LENet: lightweight and efficient LiDAR semantic segmentation using multi-scale convolution attention. arXiv preprint [arXiv:2301.04275](https://arxiv.org/abs/2301.04275) (2023)
16. Zhang, Y., Zhou, Z., David, P., et al.: PolarNet: an improved grid representation for online Lidar point clouds semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9601–9610 (2020)
17. Zhu X, Zhou H, Wang T, et al. Cylindrical and asymmetrical 3D convolution networks for Lidar segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9939–9948 (2021)
18. Cheng, R., Razani, R., Taghavi, E., et al.: 2-s3net: attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12547–12556. IEEE, Piscataway (2021)
19. Guo, M.H., Lu, C.Z., Hou, Q., et al.: Segnext: rethinking convolutional attention design for semantic segmentation. arXiv preprint [arXiv:2209.08575](https://arxiv.org/abs/2209.08575) (2022)
20. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)
21. Milioto, A., Vizzo, I., Behley, J., et al.: Rangenet++: fast and accurate Lidar semantic segmentation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4213–4220. IEEE, Piscataway (2019)

22. Behley, J., Garbade, M., Milioto, A., et al.: Semantickitti: a dataset for semantic scene understanding of Lidar sequences. In: IEEE/CVF International Conference on Computer Vision, pp. 9297–9307. IEEE, Piscataway (2019)
23. Cheng, H.X., Han, X.F., Xiao, G.Q.: CENET: toward concise and efficient Lidar semantic segmentation for autonomous driving. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 01–06. IEEE (2022)
24. Wang, Q., Wu, B., Zhu, P., et al.: ECA-Net: efficient channel attention for deep convolutional neural net-works. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)
25. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
26. Woo, S., Park, J., Lee, J.Y., et al.: CBAM: convolutional block attention module. In: European Conference on Computer Vision (ECCV), pp. 3–19 (2018)



A Review of the Following Behavior of Human-Driven Vehicles and Automated Vehicles in Adverse Weather Conditions

Qiong Lu^{1,2} , Haoda Fang¹, Xinzhaoh Jia¹, and Bing Wang^{1,2}

¹ School of Traffic and Transportation Engineering, Xinjiang University, 777 Huarui Street, Shuimogou District, Urumqi, Xinjiang, China

Qiong.lu@xju.edu.cn, 107552401777@stu.xju.edu.cn

² Xinjiang Key Laboratory of Green Construction and Smart Traffic Control of Transportation Infrastructure, Xinjiang University, 777 Huarui Street, Shuimogou District, Urumqi, Xinjiang, China

Abstract. This paper reviews the following behavior of human-driven vehicles (HDVs) and automated vehicles (AVs) under adverse weather conditions. As a basic framework for describing the following behavior, the car-following model is mainly used to simulate and predict the dynamic relationship between vehicles. Under adverse weather conditions (e.g., haze, snow, and rainfall), HDVs rely on the driver's experience and intuition, which involves greater uncertainty and risk. AVs can theoretically significantly improve driving safety and stability through a variety of sensors and intelligent algorithms, but under these conditions, the sensor performance may degrade, which in turn affects the driving safety of the vehicles. Studies have shown that adverse conditions significantly impact car-following behavior, and thus, the parameters of existing models need to be adapted to these environments. Future research will explore the relationship between HDVs, AVs, and fully-AVs to promote the development of smarter, safer, and more efficient intelligent transportation systems.

Keywords: Human-driven vehicles · Automated vehicles · Adverse weather · Car-following model

1 Introduction

Car-following models are mathematical frameworks in traffic flow theory used to describe and simulate the behavior of vehicles on roads. By quantifying the interactions between leading and following vehicles, these models study and predict car-following behavior to analyze the dynamic characteristics of traffic flow [1]. They aid in understanding traffic flow operations and uncover the underlying mechanisms of traffic phenomena.

Since the mid-20th century, significant advancements have been made in car-following models, encompassing disciplines such as traffic engineering, vehicle engineering, systems engineering, control science, psychology, and physics [2]. Recently,

the progress in automated driving technology has garnered widespread attention in the vehicle traffic field, attracting numerous experts and scholars to engage in research.

In recent years, the development of deep learning and other advanced technologies has invigorated the field of automated driving. Increasing numbers of automated vehicles (AVs) have transitioned from simulation labs to real-world roads, becoming integral parts of traffic flow and playing an increasingly significant role in smart urban transportation systems [3].

In adverse weather conditions, human-driven vehicles (HDVs) primarily rely on the driver's experience and intuition for car-following behavior. Drivers adjust following distance and speed based on visibility, road conditions, and the behavior of other vehicles to minimize accidents. However, this experience-based driving approach involves significant uncertainty and risk. In such conditions, drivers' reaction times and perception abilities decrease, leading to instability in car-following behavior and an increased accident rate. In contrast, AVs use sensors and intelligent algorithms to perceive environmental changes in real-time and make decisions and controls accordingly. This technology theoretically enhances driving safety and stability significantly under adverse weather conditions.

Traditional HDVs and emerging AVs face significant challenges in car-following behavior under adverse weather conditions. In situations like haze, snow, and rain, drivers' visibility and reaction times are affected, often leading to heightened stress and anxiety, impacting decision-making. Additionally, the performance of sensors in AVs may degrade under such conditions, affecting perception accuracy. These factors collectively have a substantial impact on car-following behavior. Therefore, studying and analyzing the car-following behavior of both HDVs and AVs under these adverse weather conditions holds essential theoretical and practical significance.

2 Car-Following Models for HDVs and AVs

Car-following models for HDVs primarily use dynamic methods to study the theoretical state of a following car trailing a lead car on a single lane without overtaking. In this context, various aspects of the car-following process can be represented through mathematical modeling. Originating in the 1950s, this theory examines how drivers control vehicle speed to maintain a desired relative distance from the vehicle ahead, aiming to enhance traffic flow management and optimization.

In 1950, Reuschel introduced the earliest car-following model [4]. In the early 1960s, Gazis, Herman, and Rothery at the General Motors Research Laboratories developed the GM (General Motors) series models, leading to various other car-following models. Table 1 shows the main car-following models for HDVs currently in use.

The models listed in the table include the Braking Distance Model, Optimal Velocity (OV) Model, Intelligent Driver Model (IDM), Full Velocity Difference (FVD) Model, and Path-Following Model. Despite their distinct forms and applications, they share the

common goal of simulating the dynamic behavior of vehicles in traffic flow, particularly car-following behavior. These models generally use variables such as vehicle speed, the speed of the lead car, gap distance, or desired gap distance to calculate vehicle acceleration, thereby describing dynamic changes. The IDM model, which considers

Table 1. Car-following models for HDVs.

Model Names	Authors	Model Equations	Model Content and Significance
Braking Distance Model	Gipps (1981) [5]	$a_n(t) = b \left(1 - \frac{v_n(t)}{v_{\max}} \right) a_n(t)$	Driver-based braking distance decision considering safe braking distance(s_{safe})
OV Model	Bando et al. (1995) [6]	$a_n(t) = \alpha (V(s_n(t)) - v_n(t))$	Acceleration $a_n(t)$ determined by optimal velocity $V(s_n(t))$ function for traffic flow simulation
IDM	Treiber, Hennecke, Helbing (2000) [7]	$\dot{v}_\alpha = a^{(\alpha)} \left[1 - \left(\frac{v_\alpha}{v_0(\alpha)} \right)^6 - \left(\frac{s^*(v_\alpha', \Delta v_\alpha)}{s_\alpha} \right)^2 \right]$	Considers multiple factors (e.g., $v_n(t)$, $s_n(t)$) suitable for diverse traffic scenarios, widely used in simulation and research
FVD Model	Jiang et al. (2001) [8]	$a_n(t) = \alpha (V(s_n(t)) - v_n(t)) + \beta (v_{n-1}(t) - v_n(t))$	Enhance the OV model by adding a speed difference term ($s_n(t) - v_n(t)$) to improve stability
Path-Following Model	Brackstone and McDonald (1999) [9]	$a_n(t) = k \cdot (v_{n-1}(t) - v_n(t)) + c \cdot (s_{n-1}(t) - s_{\text{desired}})$	Integrates speed $v_n(t)$ and distance $s_n(t)$ control for effective path following in real-world driving scenarios

quadratic terms of speed and gap distance, is suitable for various traffic conditions and widely used in simulations and research. The Path-Following Model combines speed and distance control, particularly applicable to path-following behavior in actual driving. The Braking Distance Model is based on the driver's braking decision and considers safe braking distances. The OV Model employs an optimal velocity function to determine acceleration, which is suitable for traffic flow simulation. The FVD Model enhances the OV Model by adding a speed difference term to improve model stability.

Although these models do not explicitly account for weather conditions, they significantly impact car-following behavior. For instance, in haze, snow, or rain, driver

reaction times increase, and desired following distances expand, potentially affecting model parameters. Braking distances lengthen on slippery or icy roads, necessitating adjustments to reflect longer stopping times. Additionally, optimal speed and speed difference terms might be modified in adverse weather to accommodate lower speed limits, ensuring safe vehicle spacing.

As the scale of mixed traffic flows composed of HDVs and AVs continues to expand, new challenges arise for car-following model theory. Traditional HDV car-following models typically rely on the driver's perception and reaction abilities, using vehicle distance, speed difference, and other factors for linear or nonlinear modeling. These models capture typical driver behaviors, such as maintaining safe distances and adjusting speed. However, individual driver differences and changing external environments often limit their accuracy and reliability. In contrast, AV car-following models leverage advanced sensors, computer vision, and machine learning technologies to perceive the surrounding environment in real-time, including vehicles, pedestrians, and road conditions. Through complex algorithms, AVs dynamically adjust speed and distance [10]. These models integrate current traffic information and can predict future traffic conditions, thereby optimizing car-following behaviors to enhance safety and efficiency. Mainstream AV car-following models are shown in Table 2.

In automated car-following research, various models have been developed to cater to traffic conditions and driving needs. Table 2 lists models such as the Gipps model, Cooperative Adaptive Cruise Control (CACC) model, Deep Q Networks (DQN) car-following model, Deep Reinforcement Learning car-following model, and the Autonomous Vehicle Integration Strategy (AVIS) model. Despite methodological differences, their common goal is to optimize car-following behavior to enhance traffic flow efficiency and safety. The CACC model uses vehicle-to-vehicle (V2V) communication to improve car-following performance and traffic efficiency by coordinating speed and distance, thus reducing congestion and enhancing road utilization. The Gipps model, a classic microscopic car-following model, considers acceleration and deceleration constraints and is widely used in traffic flow simulation and automated driving control. The Deep Reinforcement Learning and DQN car-following models leverage machine learning techniques to optimize car-following behavior by adapting strategies based on historical data and real-time environment information, effectively managing complex traffic scenarios. The AVIS model integrates sensor and communication data, offering real-time decision-making capabilities suitable for complex traffic environments and providing a comprehensive understanding of the surroundings for more accurate car-following decisions.

When considering the impact of weather on these models, it is evident that adverse conditions such as haze, snow, and rain can affect vehicle perception and road friction, thereby influencing car-following behavior. For instance, the CACC and AVIS models rely on sensor and communication technologies, which may experience interference under these weather conditions, affecting data accuracy and real-time performance.

Table 2. Car-following models for AVs.

Model Names	Authors	Model Equations	Model Content and Significance
Gipps Model	Gipps (1981) [11]	$v_n(t+T) = \min \left\{ v_n(t) + 2.5a \left(1 - \frac{v_n(t)}{v_0} \right) \sqrt{0.025 + \frac{v_n(t)}{v_0}}, \frac{v_n(t)^2}{2b(s_n - s_{\text{safe}})} \right\}$	The Gipps model, a classic microscopic car-following model, considers acceleration and deceleration constraints and is widely used in traffic flow simulation and automated driving control.
CACC Model	Milanes, et al. (2014) [11]	$a_n(t) = K_p(s_{n-1}(t) - s_n(t) - s_{\text{desired}}) + K_d(v_{n-1}(t) - v_n(t))$	Cooperative vehicle-to-vehicle communication technology enhances car-following performance and traffic efficiency
Deep Q-Network Car-Following Model	Zhu M et al. (2018) [12]	$Q^*(s, a) = \mathbb{E}_{s'} \left[r + \gamma \max_{a'} Q^*(s', a') \mid s, a \right]$	Deep Q-Networks optimize car-following strategies to adapt to complex dynamic traffic environments
Deep Reinforcement Learning Car-Following Model	B. R. Kiran et al. (2022) [13]	$Q(s_t, a_t; \theta) = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta')$	Deep reinforcement learning optimizes vehicle car-following behavior, enhancing traffic flow efficiency and safety with strong adaptability
AVIS Model	Lee et al. (2020) [14]	$a_n(t) = f(x_n(t), v_n(t), \text{sensor data})$	The AVIS model integrates sensor and communication data, offering real-time decision-making capabilities suitable for complex traffic environments and providing a comprehensive understanding of the surroundings for more accurate car-following decisions.

3 Changes in Car-Following Behavior Under the Influence of Adverse Weather Conditions

With the global increase in vehicles, traffic congestion and accidents have become frequent, with environmental factors being significant influences. In China, expansive geography results in traffic conditions impacted by specific weather phenomena like haze, snow, and rainfall, presenting challenges to traffic safety and capacity.

Haze-induced low visibility and slippery roads often lead to major traffic accidents, causing severe casualties and property damage. Fog increases driver stress, elevating driving risks and reducing vehicle speeds, thereby decreasing road capacity and burdening surrounding road networks, affecting the entire network [15, 16]. Ran Jinsong studied the effects of rain-fog conditions on highway car-following behavior, developing a corresponding model. Data analysis showed increased headway and spacing under rain-fog conditions, with the IDM model performing best using genetic algorithms for GHR, FVD, and IDM model calibration, supporting refined traffic management [17]. Wang Jiayu and colleagues analyzed driving behavior under foggy conditions using high-fidelity simulators to construct fog scenarios. Genetic algorithm-calibrated IDM models showed decreased safe distances, headway, and desired speeds, indicating increased traffic conflict risks, necessitating enhanced safety management [18]. Qin Yanyan used numerical simulations to examine highway tailgating risks under fog, proposing V2V communication-based improvements. Using TET (Time Exposed Time-To-Collision) and TIT (Time Integrated Time-To-Collision) as risk metrics, light, and dense fog conditions showed optimal risk reduction at a 60 km/h limit, with the highest risks at 80 km/h and 100 km/h. The method effectively reduced tailgating risks by 35.62% and 36.52% for TET and TIT, respectively, at a 95% confidence level [19]. Yu Lemei analyzed foggy driving psychological characteristics via surveys, incorporating acceleration probability into the NaSch model to develop a fog-influenced traffic flow model. Numerical simulations assessed traffic density, headway, speed, and queue length trends under varied visibility, using principal component analysis to build a traffic operation model, aiding decisions on speed limits and driving distances under different fog visibility levels [20].

Under icy and snowy conditions, impacts on car-following behavior are particularly significant. Firstly, ice and snow create a film on the road, reducing friction coefficients and significantly diminishing friction between vehicles and the road. This directly affects climbing ability, driving stability, and braking performance, greatly extending braking distances and potentially causing braking skids. The risk of skidding is higher during turns, increasing accident likelihood. Secondly, snow significantly reduces visibility, shortening drivers' sight distance. This can cause driver tension, reducing reaction speed and judgment and increasing the risk of accidents like scrapes. Zhang Chun studied driver car-following behavior under icy conditions using real vehicle data, genetic algorithms, and trajectory data to calibrate models under normal and icy conditions, constructing the corresponding car-following model. Subjective and objective evaluation methods were used to classify drivers, showing high consistency in results. Under icy conditions, drivers increased headway and operated more cautiously. He used the Task-Difficulty intelligent driver model (TDIDM) to reflect driver behavior better than traditional IDM models, providing theoretical support for traffic management and behavior norms under icy conditions [21]. Yang Longhai and colleagues used actual vehicle

data and MATLAB genetic algorithm toolbox to calibrate the TDIDM car-following model, analyzing driver behavior and constructing an improved intelligent driver model with human factors. They found drivers adjust headway under icy conditions as risk compensation, with the new model better describing this behavior, outperforming traditional models across scenarios and providing traffic management insights [22]. Tang Zhen analyzed the impact of snowy weather on highways in Xinjiang using field survey data and regression analysis, optimizing the Full Velocity Difference (FVD) model by adding headway acceleration influence and improving desired speed, determining minimum safe distance, optimized speed, sensitivity coefficients, and headway acceleration, constructing a minimum safe distance model, validated through simulations for stability and feasibility, supporting traffic management [23]. Wei Zhengchao analyzed winter icy road conditions, establishing an improved car-following model using Greenshields speed-density relationship and FVD model, proposing expected speed and minimum safe distance models for icy roads. Validation showed the new model excels in stability and acceleration asymmetry, aligning with actual traffic flow, aiding traffic management [24].

Under rainy conditions, Huang Zhaoguo et al. proposed a car-following risk determination model based on random forests, incorporating rainfall conditions and safe distances to refine traditional driving risk models. Using driving simulator data under rain conditions, the model was validated and trained with the random forest algorithm for risk feature extraction. Results showed the model's risk level error was less than one level, and it reduced average relative errors by 10.36% and 4.54% compared to artificial neural networks and support vector machines, respectively. This model is a practical tool for assessing car-following risk levels in rainy weather [25]. Yan Ran et al., by analyzing the impact of rain on traffic flow, quantified the effect of rainfall on car-following behavior using driver reaction delay time and road surface adhesion coefficients. They concluded that rain mainly affects optimization speed and safety distance, with minimal impact on vehicle speed differences [26]. Shan Huagang et al. collected traffic flow data under different rainfall intensities through field experiments and calibrated and cross-validated GHR, FVD, and IDM models using genetic algorithms. Results showed rainfall intensity significantly affects highway car-following behavior, with heavier rain leading drivers to maintain larger headways and lower following speeds [27].

Compared to traditional HDVs, AVs demonstrate better adaptability and interference resistance in complex weather conditions, offering higher stability. Thus, studying automated car-following models is crucial for ensuring driving safety and improving traffic efficiency in adverse weather. Traditional HDVs rely heavily on visual perception, making them susceptible to adverse weather like rain and snow. Wang Wenxuan's study on human driving behavior in adverse weather shows that adverse weather conditions mainly affect three aspects: speed changes, following distance changes, and headway time changes [28]. For speed changes, drivers reduce speed to ensure safety; for instance, the average speed drops from 17.38m/s in clear weather to 9.03m/s and 4.43m/s in rain and fog, respectively. For following distance changes, reduced visibility prompts shorter distances to gather information from the leading car. The average following distance decreases from 17.38 m in clear weather to 14.08m and 8.17m in rain and fog, respectively. For headway time changes, drivers adjust following distance

Table 3. Impact of Adverse Weather on Key Parameters of Human and Automated Car-Following Behaviors.

			Haze	Snow and ice	Rainfall
Human driving	Speed		Reduced by 7%-61% under varying visibility conditions [28]	Reduced by 30%-40% [30]	Reduced by 2%-10% [28]
	Traffic capacity		Reduced by 4%-91% under varying visibility conditions [28]	Reduced by 10%-30% [30]	Decreased by 13%-14% [29]
	Time Headway		Only 31.9% of vehicles fall within 2 s [19]	Increased by 2–4 s [30]	Increased by approximately 1 s [28]
	Driver behavior characteristics		Increased reaction time makes drivers more cautious [20]	Reaction times may increase, with significant differences in driving styles [30]	Reaction time increased, with physiological differences and psychological fluctuations in drivers [28]
Automated driving	Sensors	laser radar	Rarely causes detection errors [30]	Directly leads to false positives or detection failures [30]	Causes perceptual errors 30% of the time [30]
		Millimeter-wave radar	Negligible impact [30]	May cause minor errors in specific scenarios [30]	Rarely causes detection errors [30]
		Standard camera	Causes perceptual errors 30% of the time [30]	Causes perception errors 30% of the time [31]	Causes perceptual errors 30%-50% of the time [30]
	Average vehicle speed		Reduces by approximately 6% [17]	Reduces by 20%–35% [30]	Reduces by approximately 10% [28]

(continued)

Table 3. (continued)

		Haze	Snow and ice	Rainfall
	Headway time	Increases by approximately 17% [18]	Increases by 7%–15% [30]	Increases by approximately 8% [28]
	Braking distance	Increase is less than manual driving [20]	Approximately doubles [30]	Increase is less than that of manual driving [19]

and speed to maintain safety, with clear weather headway at 1.93 s, 1.79 s, and 1.99 s in rain and fog, respectively. He concludes that as visibility decreases, rain and snow significantly lower average speed and following distance. Under good visibility, drivers maintain speed as road conditions worsen but increase following distance. Adverse road conditions result in very low average speeds, with following distances increasing as precipitation intensifies. In AVs, road perception relies on multi-sensor systems, including LiDAR, millimeter-wave radar, and stereo vision cameras, compensating for the limitations of human vision and significantly enhancing reliability and safety [29].

However, icy weather affects AVs. Huang Jiaping optimized car-following data from the Waymo dataset using genetic algorithms, refining normal and icy weather parameters. His study showed that icy conditions significantly impact automated car-following behavior. The improved IDM model, validated with real data and genetic algorithms, proved effective. Simulations indicated that under severe icy conditions, following queues are less stable, acceleration-following is more sensitive to speed changes, and deceleration-following causes more severe traffic flow oscillations. A comprehensive risk evaluation model revealed that icy conditions pose the highest risks for AVs, with collision risks when perception errors exceed 3 s. This research is crucial for enhancing winter driving safety for AVs [30]. This study validated the improved model's effectiveness and provided essential references for advancing automated driving technology in adverse winter conditions, guiding further development and application in complex environments.

This section deeply analyzes the impacts of adverse weather conditions (such as haze, ice, and rain) on driving behavior, particularly comparing changes in key parameters between HDVs and AVs, as shown in Table 3.

Comprehensive data analysis indicates that under adverse weather conditions such as haze, ice, and rain, the performance of HDVs is significantly affected: speed typically decreases substantially to adapt to reduced visibility and lower road friction; reaction time increases as drivers need more time to identify and respond to potential hazards; and traffic capacity correspondingly decreases, especially in high traffic volumes. In contrast, AVs demonstrate higher stability and safety under these adverse conditions. Automated systems use sensors and algorithms to monitor environmental changes in real-time and respond quickly and accurately, maintaining more stable speeds and shorter reaction

times. Although traffic capacity is also impacted, the system's ability to better coordinate distances and speeds between vehicles makes overall traffic efficiency significantly superior to human driving. Therefore, future improvements in sensor technology are necessary to ensure reliability and accuracy under adverse weather conditions.

4 Trends in Car-Following Model Development

In the foreseeable future, automated traffic flows, composed of human and machine co-driving, may become a major component of traffic streams [31]. There will be a shifting trend between HDVs and AVs. As technology advances and proliferates, the proportion of AVs will gradually increase, while reliance on human driving will decrease. Simultaneously, automated technology will mature and be applied in specific environments and conditions [32].

This transitional period will pose new challenges and demands for studying car-following models. Future models should account for the interplay between HDVs, AVs, and fully-AVs, providing accurate and reliable predictions in complex and variable traffic environments, especially under adverse weather conditions. Specifically, research on car-following models may focus on the following aspects:

1. **Characteristics of Mixed Traffic Flows:** Investigate the interaction behaviors among different driving modes, including the mutual influence between HDVs and AVs, as well as the performance of AVs within mixed traffic streams [33]. These interactions can become more complex under adverse weather conditions, requiring particular attention.
2. **Dynamic Adaptability:** Develop car-following models capable of dynamically adapting to different driving modes and weather conditions to handle the complex variations in driving behaviors within traffic flows. This includes model optimization and parameter adjustments for driving scenarios and weather situations.
3. **Data-Driven Modeling Under Adverse Weather Conditions:** Utilize big data and machine learning techniques to extract driving behavior characteristics from massive micro-level vehicle trajectory data, enhancing the precision and adaptability of models. In particular, adverse weather conditions require large sample data for model training and validation to achieve high-accuracy predictions [34].
4. **Interdisciplinary Research:** Combine knowledge from intelligent perception, artificial intelligence, psychology, traffic engineering, and other fields to comprehensively analyze driving behavior and traffic flow characteristics, providing multi-dimensional perspectives and methods for innovation in car-following models. Driving behavior under adverse weather conditions necessitates deeper interdisciplinary research.

Future car-following models will not be limited to simulating and predicting a single type of driving behavior but will reflect various driving modes within complex traffic environments. The complexity and uncertainty of driving behavior under adverse weather conditions will increase, posing higher demands on car-following models. Research in this field will likely focus on the relationship between HDVs, AVs, and fully-AVs, becoming a central theme in traffic flow theory studies. This will provide robust technical support for achieving more intelligent, safe, and efficient traffic systems, advancing automated driving technology's gradual implementation and popularization.

Acknowledgements. This research is supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2018D01C054) and the Major Science and Technology Special Project of the Xinjiang Uygur Autonomous Region in 2020 “Research on Key Technologies for the Construction and Operation of Highway Tunnel Projects in Alpine and High-Altitude Areas of 20 km Class” Project 4 (Project No. 2020A03003/ 2020A03003-4).

References

1. Yan, L.: Study on car-following models. Dalian Jiaotong University (2006). (in Chinese)
2. Wang, D., Jin, D.: A review and prospect of car-following behavior modeling. *Chin. J. Highway Transp.* **25**(01), 115–127 (2012). (in Chinese)
3. Editorial Board of China Journal of Highway and Transport. Review of Academic Research in China's Automotive Engineering 2023. *China Journal of Highway and Transport* **36**(11), 1–192 (2023). (in Chinese)
4. Reuschel, A.: Vehicle movements in the column uniformly accelerated or delayed. *Oesterrich IngrArch*, (4): 193215 (1950)
5. Gipps, P.G.: A behavioural car-following model for computer simulation. *Trans. Res. Part B Methodol.* **15**(2), 105–111 (1981)
6. Bando, M., Hasebe, K., Nakayama, A., Shibata, A., Sugiyama, Y.: Dynamical model of traffic congestion and numerical simulation. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdisc. Top.* **51**(2), 103 (1995)
7. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **62**(2 Pt A), 1805–24 (2000)
8. Jiang, R., Wu, Q., Zhu, Z.: Full velocity difference model for a car-following theory. *Phys. Rev. E* **64**(1), 017101 (2001)
9. Brackstone, M., McDonald, M.: Car following: a historical review. *Trans. Res. Part F Traffic Psychol. Behav.* **2**(4), 181–196. (1999)
10. Wang, J., Huang, H., Zhi, P., et al.: Overview of automated driving development and key technologies. *Appl. Electron. Technique* **45**(06), 28–36 (2019). (in Chinese)
11. Milanés, V., Shladover, S.E., Spring, J., et al.: Cooperative adaptive cruise control in real traffic situations. *IEEE Trans. Intell. Trans. Syst.* **15**(1), 296–305 (2014)
12. Zhu, M., Wang, X., Wang, Y.: Human-like autonomous car-following model with deep reinforcement learning. *Trans. Res. Part C Emerg. Technol.* **97**, 348–368 (2018)
13. Kiran, B.R., et al.: Deep reinforcement learning for autonomous driving: a survey. In: *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926 (2022)
14. Mingmei, S.: A day-to-day dynamic model for mixed traffic flow of autonomous vehicles and inertial human-driven vehicles. *Transp. Res. Part E* **173** (2024)
15. Jiang, X., Huang, K.: Analysis of road traffic safety measures in foggy conditions and their effectiveness. *J. Harb. Inst. Technol.* **44**(6), 86–91 (2012). (in Chinese)
16. Gao, J.: Evaluation of the hazards of fog on highway operations. *J. Chongqing Jiaotong Univ. (Nat. Sci. Ed.)* **32**(2), 248–251 (2013). (in Chinese)
17. Ran, J.: Analysis and Calibration of Highway Car-Following Behavior in Rain and Fog. Chang'an University (2022). (in Chinese)
18. Wang, J., Li, Z., Zhang, J., et al.: IDM car-following model analysis on highway straight sections in fog and haze. *Transp. Res.* **1**(01), 86–91+98 (2015). (in Chinese)
19. Qin, Y., Xiao, T., Luo, Q.: Research on rear-end collision risk in foggy highways based on car-following model. *J. Wuhan Univ. Sci. Technol.* 1–8 (2024). (in Chinese)

20. Yu, L., Zhang, M., Wang, X.: Simulation study on safe speed limits in fog based on car-following model. *Sci. Technol. Eng.* **18**(33), 224–229 (2018). (in Chinese)
21. Zhang, C.: Study on Car-Following Model Considering Driver Behavior Under Ice and Snow Conditions. Harbin Institute of Technology (2019). (in Chinese)
22. Yang, L., Zhang, C., Qiu, X., et al.: Study on car-following behavior and model of chinese drivers under ice and snow conditions. *J. Transp. Syst. Eng. Inf. Technol.* **20**(06), 145–155 (2020). (in Chinese)
23. Tang, Z.: Research on Car-Following Model on Xinjiang Highways During Snowfall. Xinjiang Agricultural University (2016). (in Chinese)
24. Wei, Z.: Research on Car-Following Model on Urban Expressways Under Icy and Snowy Conditions. Harbin Institute of Technology (2012). (in Chinese)
25. Huang, Z., Guo, X., Jia, L.: Research on car-following risk behavior in rainy conditions using random forest. *Traffic Inf. Saf.* **38**(01), 27–34 (2020). (in Chinese)
26. Yan, R., Ye, X., Han, G.: Full velocity difference car-following model and calibration considering rain impact. *J. Wuhan Univ. Technol. (Transp. Sci. Eng. Ed.)* **40**(06), 1050–1053 (2016). (in Chinese)
27. Shan, H., Yu, C., Xu, J., et al.: Calibration of highway car-following models under different rain intensities. *J. Highway Transp. Res. Dev.* **41**(06), 190–198 (2024). (in Chinese)
28. Wang, W.: Study on Variable Speed Limit Control Strategy for Intelligent Connected Vehicles on Expressways in Adverse Weather. Tongji University (2022). (in Chinese)
29. He, J.: Speed Limit Methods for Highways Considering Automated Driving in Rainy Weather. Harbin Institute of Technology (2023). (in Chinese)
30. Huang, J.: Study on Car-Following Characteristics of Automated Vehicles Under Ice and Snow Conditions. Changsha University of Science & Technology (2022). (in Chinese)
31. Automated Driving Committee of China Highway & Transportation Society. Development Report on Cooperative Vehicle-Infrastructure Automated Driving (2019). (in Chinese)
32. Yang, L., Zhang, C., Qiu, X., et al.: Progress in car-following model research. *J. Traffic Transp. Eng.* **19**(05), 125–138 (2019). (in Chinese)
33. Gao, Q.: Study on car-following model and traffic flow characteristics of intelligent connected vehicles. *Highway* **66**(10), 229–235 (2021). (in Chinese)
34. Zhou, Y., Zhang, Y.: Car-following model in 5G environment. *Appl. Res. Comput.* **38**(03), 796–799 (2021). (in Chinese)



Optimization of Arrival and Departure Track Application Considering In-App Reconnection Decomposition

Mengjie Cui and Lei Nie^(✉)

Beijing Jiaotong University, Beijing 100044, China

Lnie@bjtu.edu.cn

Abstract. The concept of generalized train is proposed for the decomposition of reconnection in the station. According to the concept of reconnection decomposition train contained in the generalized train proposed in this paper, the reconnection decomposition is successfully integrated into the operation plan of the station-to-departure line. At the same time, a method for judging the route conflict of the reconnection decomposition train is proposed, So that the reconnection decomposition train fits the necessary constraints for the use of the departure line. Finally, taking the train schedule data of Beijing South Railway Station in a certain period of time as input, the model successfully obtains the application scheme of the arrival and departure track considering the decomposition of reconnection and the operation scheme of the arrival and departure track without considering the decomposition of reconnection., And calculate the utilization rate of Beijing South Railway Station to arrival and departure track under these two scenarios. The results show the effectiveness of this method.

Keywords: generalized train · reconnection decomposition train · application scheme of the arrival and departure track

1 Introduction

In order to solve the problem of insufficient train capacity during peak passenger flow hours, some foreign railway companies carry out the reconnection operation and decomposing operation of EMUs at railway stations. This method can better relieve the pressure of passenger flow during peak hours and reduce the waste of train capacity during off-peak hours. In recent years, domestic high-speed railways have also piloted the EMU in-app reconnection and decomposing operation mode. In this operation mode, there may be two trains occupying the same arrival and departure line at the same time, and the train in-app running process is more complicated, which puts forward new requirements for the arrival and departure track application model. In order to solve the construction problem of the arrival and departure track operation model under the reconnection decomposition operation mode, improve the operation efficiency of the arrival and departure track, and reasonably arrange the operation route and departure line of the reconnection decomposition train, this paper proposes an optimization model of the arrival and departure track operation plan considering the reconnection decomposition.

At present, there are few studies on the combination of arrival and departure track application and reconnection decomposition. In terms of arrival and departure track application research, some scholars have proposed different arrival and departure line application optimization models according to different optimization goals, and designed intelligent optimization algorithms to solve [1–6]. Some scholars have also proposed new solutions to solve the arrival and departure line application optimization problem [7–11].

2 Analysis of Decomposition Process of Train Reconnection in Station

The reconnection decomposition train is divided into two types: reconnection and decomposition. The schematic diagram of these two cases is shown in the figure below.

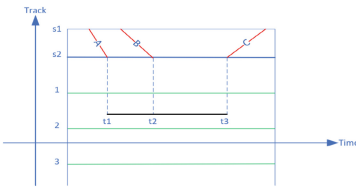


Fig. 1. Reconnection process

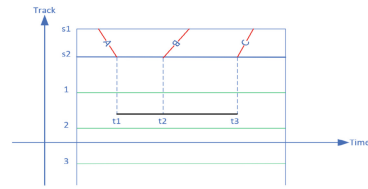


Fig. 2. Decomposition process

In Fig. 1, train A first arrives at the No. 2 track at station s2, and train B then enters the same track and reconnects with train A. After the reconnection, the bottom of train A and the bottom of train B are combined into a long train, which continues to serve as train C. In Fig. 2, after the long train A to be decomposed arrives at the No. 2 track at station s2, it is decomposed into short train A and train B. Train B departs first, and then train A departs.

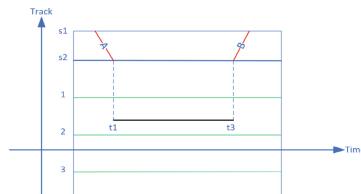


Fig. 3. Schematic diagram of the operation process of ordinary trains

Figure 3 shows the operation process diagram of the non-reconnection decomposition train. If the reconnection decomposition train and the non-reconnection decomposition train can be regarded as a class of trains, the difference between them can be eliminated, which is conducive to the construction of the arrival and departure line application model. Therefore, this paper proposes the concept of generalized train. The generalized train includes the reconnection decomposition train and the ordinary train. The reconnection

decomposition train refers to the whole composed of the three trains in Fig. 1 or Fig. 2. The ordinary train refers to the whole composed of the two trains in Fig. 3, which has one arrival time and arrival direction and one departure time and departure direction.

3 Optimization Model of Station-to-Departure Line Application Considering Reconnection Decomposition

Table 1. Symbol definition

variable	meaning
T	train assembly
L	Station to departure line collection
InD	Set of turnout sets in approach direction
$outD$	The set of turnout sets in the direction of operation
$tInD$	The assembly of the direction of the train entering the station
$tOutD$	Collection of train departure directions
TA	Arrival time of the train arriving at the station
TD	The train departs from the station at the departure time
t_i^z	The time when Train i uses the arrival and departure line
t_{safe}	Minimum safe time interval

Before building the model, it is necessary to generate the route conflict set P in advance. The (i., e, j, k) of each element in the set indicate that if i train is assigned to the e track line and j train is assigned to the k track, there will be an approach conflict. The reason for the approach conflict is that the paths passed by the two trains intersect in space, and the approaches of the two trains overlap in time (Table 1).

Determine whether (i., e, j, k) belongs to the P method:

For any (i., e, j, k), first determine whether there is time overlap between i train and j train in the four approach scenarios of pick-up and pick-up, pick-up and departure, departure and departure, and record the conflict situation.

If after the previous judgment, it is found that the approach times of i. and j overlap, then for the scene where the approach times overlap, continue to determine whether there is an approach conflict. For example, there is a time overlap between the pick-up route of the i train and the departure route of the j train, then continue to judge whether there is a space conflict between the pick-up route of the i train and the pick-up route of the j train. If there is a conflict, this element belongs to P, indicating (i., e, j, k) a conflicting combination.

The capital P consists of the following:

The P1 is the set of conflicts between the ordinary train i. and the ordinary train j. The construction process of the P1 is as follows:

If the i train that selects the e track and the j train that selects the k track are determined to conflict according to the above method, then the conflicting index (i, e, j, k) is added to P1.

The P2 is the set of conflicts between the ordinary train i and the reconnection decomposition train j . The construction process of the P2 is as follows:

The reconnection train j is regarded as two short marshalling trains $j1$ and $j2$, as shown in Fig. 1., where $j1$ has the same approach direction as the A train, the same arrival time as the A train, $t1$, the same exit direction as the reconnected train, and the same departure time as the reconnected train, which is $t3$. $j2$ has the same approach as the B train, and the arrival time is the same as the B train, which is $t2$. The exit direction is the same as the reconnected train, and the departure time is the same as the reconnected train. The decomposition process of $t3$ is similar to the reconnection process.

The ordinary train i . and $j1$ and $j2$ respectively follow the conflict judgment method to determine whether there is a conflict. If i conflicts with one of $j1$ and $j2$, then i occupies e and j occupies k , and adds the (i, e, j, k) of the indices of the ordinary train i ., the departure line e , the reconnection decomposition train j , and the departure line k to the P2.

The P3 is the set of conflicts between the reconnection decomposition train i and the reconnection decomposition train j , and the content of the P3 is as follows:

The reconnection decomposition train i is regarded as two short-marshalling trains $i1$ and $i2$, and the reconnection decomposition j is regarded as two short-marshalling trains $j1$ and $j2$. $i1, i2$ and $j1$ and $j2$ respectively are determined according to the above method to determine whether there is a conflict. If one of $i1$ and $i2$ conflicts with one of $j1$ and $j2$, then i occupies e and j occupies k . Then, add the conflicting reconnection decomposition train i , the arrival line e , the reconnection decomposition train j , and the index left parentheses i, e, j, k right parentheses of the departure line k to P4.

Objective function: the most balanced use of station to departure lines;

$$\min Z = \left(\sum_{e=1}^n \left(\sum_{i=1}^m t_i^z x_{ie} - \frac{\sum_{i=1}^m t_i^z}{n} \right)^2 \right)^{\frac{1}{2}} \quad (1)$$

Constraint 1: A train can only occupy one arrival and departure line.

$$\sum_{j \in L} x_{ij} = 1, i \in T \quad (2)$$

Constraint 2: The minimum safe interval between two trains occupying the same arrival and departure line must be satisfied.

$$t_j^a - t_i^d \geq t_{safe} - M(2 - x_{ie} - x_{je}), i, j \in T \quad (3)$$

Constraint 3: Two trains with overlapping operating periods on the arrival and departure line cannot be arranged on the same arrival and departure line.

$$x_{ie} + x_{je} \leq 1, \forall i, j \in T, i \neq j, t_i^d > t_j^a, t_j^d > t_i^a \quad (4)$$

Constraint 4: Trains with overlapping travel times cannot choose the hostile approach.

$$x_{ie} + x_{jk} \leq 1, \{i, e, j, k\} \in P \quad (5)$$

4 Example Verification

Taking the Beijing-Shanghai Expressway Station of Beijing South Railway Station as an example, the design scheme of the running diagram of a certain period of the Beijing-Shanghai high-speed rail is used as input, and the solver is used to solve the train arrival and departure arrangement scheme to verify the validity of the model (Fig. 4).

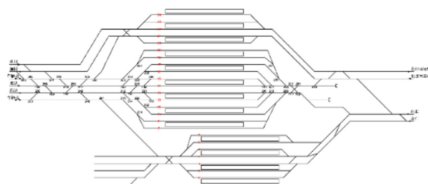


Fig. 4. Beijing South Railway Station Beijing-Shanghai Expressway Field

In this paper, the starting and ending operation time of the train on the arrival and departure line is set to 15 min, the reconnection operation time is set to 10 min, and the decomposition operation time is set to 5 min.

The train schedule for a certain period of the Beijing-Shanghai high-speed rail used in this paper is shown in the following table (Table 2):

After determining the operation chain corresponding to each arrival and departure line, for the convenience of calculation, it is assumed that the arrival and departure times of the trains are the same. The following table shows the arrival and departure times of each arrival and departure line in Beijing South Railway Station. According to the distribution range of the data in the table, the arrival and departure time of the train is determined as 4 min.

In this paper, the solution of the reconnection decomposition example is shown in the figure below. There are 5 final trains, 7 starting trains, 2 vertical folding trains and 2 reconnection decomposition trains arranged in the scheme (Fig. 5).

Among them, the A6 and A8 are reconnection trains, and the rest are ordinary trains (Fig. 6).

The calculation formula of the arrival and departure line utilization rate for the time period taken in the plan is as follows:

$$\eta = \frac{\sum t_i}{T} \quad (6)$$

In the solution, the capacity utilization rate of each incoming line is as follows (Table 3):

According to the arrangement scheme and utilization rate of arrivals and departures obtained by the solution, it can be seen that the model in this paper successfully realizes the arrangement of 16 generalized trains in the Beijing-Shanghai Expressway Field of Beijing South Railway Station. At the same time, the arrangement sequence in the scheme diagram and the utilization rate in the utilization rate table also reflect the equilibrium of the arrival and departure arrangement.

Table 2. quadrillion Shanghai High-speed Railway Beijing South Railway Station Beijing-Shanghai Expressway General Train Timetable

Generalized train number	Type of train	Arrival time	Departure time	pit stop direction	Outbound direction
A1	ordinary	9:59	10:14	Langfang	Depot
A2	ordinary	9:55	10:10	Langfang	Depot
A3	ordinary	9:51	10:06	Langfang	Depot
A4	ordinary	9:47	10:02	Langfang	Depot
A5	ordinary	9:43	9:58	Langfang	Depot
A6	reconnection decomposition	9:34\9:38	9:48\9:48	Langfang\Langfang	Langfang\Langfang
A7	ordinary	9:29	9:44	Langfang	Langfang
A8	reconnection decomposition	9:25\9:20	9:35\9:35	Langfang\Langfang	Langfang\Langfang
A9	ordinary	9:10	9:29	Langfang	Langfang
A10	ordinary	9:09	9:24	Depot	Langfang
A11	ordinary	9:05	9:20	Depot	Langfang
A12	ordinary	9:01	9:16	Depot	Langfang
A13	ordinary	8:57	9:12	Depot	Langfang
A14	ordinary	8:53	9:08	Depot	Langfang
A15	ordinary	8:49	9:04	Depot	Langfang
A16	ordinary	8:45	9:00	Depot	Langfang

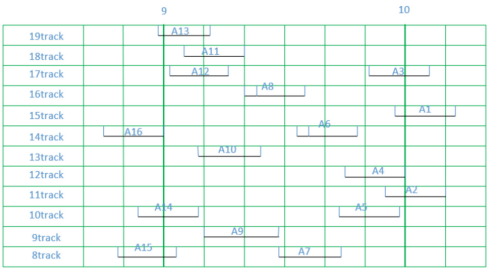


Fig. 5. Schematic diagram of the application plan of the incoming line considering reconnection decomposition

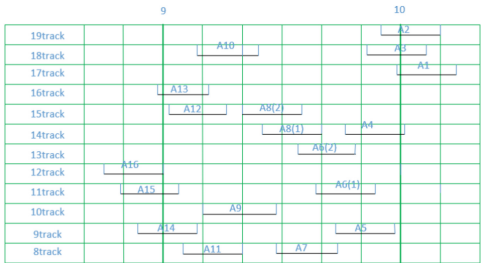


Fig. 6. Schematic diagram of the application plan of the incoming line without considering reconnection decomposition

Table 3. The statistics of the utilization rate of the AD track considering reconnection decomposition and without reconnection decomposition

AD track	Time of Occupying AD Track	Number of trains	Utilization Rate of ADTrack	Surplus Capacity of AD Tracks
19	22.9\22.9	1\1	25.4%\25.4%	74.6%\74.6%
18	22.9\45.8	1\2	25.4%\50.9%	74.6%\49.1%
17	45\22.5	2\1	50%\25%	50%\75%
16	22.5\22.5	1\1	25%\25%	75%\75%
15	21.9\43.8	1\2	24.3%\48.7%	75.7%\51.3%
14	42\43	2\2	46.7%\47.8%	53.3%\52.2%
13	21.5\21.5	1\1	23.9%\23.9%	76.1%\76.1%
12	21.9\21.9	1\1	24.3%\24.3%	75.7%\75.7%
11	22.5\45	1\2	25%\50%	75%\50%
10	45\26.5	2\1	50%\29.4%	50%\70.6%
9	26.9\45.8	1\2	29.9%\50.9%	70.1%\49.1%
8	45.8\45.8	2\2	50.9%\50.9%	49.1%\49.1%

5 Conclusion

In this paper, for the special treatment of trains that are reconnected and decomposed on the arrival and departure lines, the three trains participating in the reconnection decomposition operation are regarded as a whole to meet the relevant constraints of arranging on the same arrival and departure line; when dealing with the time conflict and the approach conflict of the train, the reconnection decomposition train that is regarded as a whole is split into two trains to judge the conflict or not. This treatment method reasonably meets the requirements of the hard constraints in the model. At the same time, the operation example of Beijing South Railway Station shows that the model in this paper can handle the station-to-departure arrangement with reconnection decomposition trains.

References

1. Billionnet, A.: Using integer programming to solve the T-platforming problem. *Transp. Sci.* **37**(2), 213–222 (2003)
2. Corman, F., D'Ariano, A., Pacciarelli, D., et al.: A tabu search algorithm for rerouting trains during rail operations. *Transp. Res. Part B; Methodol.* **44**(1), 175–192 (2010)
3. Wang, B., Hou, L., Liu, H.: Application optimization method of passenger dedicated line station to departure line. *Transp. Syst. Eng. Inf.* **12**(02), 105–110 (2012). (in Chinese)
4. Lian, Z., Shuai, B., Xu, M.: Method for adjusting the operation plan of high-speed railway station to departure line after train delay. *Railway, Transp. Econ.* **44**(02), 1–7 (2022). (in Chinese)
5. Lang, Y.: Research on comprehensive optimization model and algorithm of throat area and arrival and departure line of high-speed railway station. *China Academy of Railway Sciences* (2021). (in Chinese)
6. Zwaneveld, P.J., Kroon, L.G., Hoesel, S.P.M.V.: Routing Ts through a railway station based on a node packing model. *Eur. J. Oper. Res.* **128**(1), 14–33 (2001)
7. Flamini, M., Pacciarelli, D.: Real time management of a metro rail terminus. *Eur. J. Oper. Res.* **189**(3), 746–761 (2008)
8. Peng, Q., Zhang, Y., Lu, G., Li, W., Shi, T.: Application adjustment of railway passenger station to departure line based on time and space resources. *J. Tongji Univ. (Natural Science Edition)* **47**(07), 1011–1021 (2019). (in Chinese)
9. Xu, J., Du, W., Chang, J., et al.: Optimal arrangement of section station-to-departure line application based on genetic algo. *China Railway Sci.* **02**, 112–117 (2003). (in Chinese)
10. Ren, Y., Zhang, Q., Yuan, Z., Sun, Y.: Research on the optimization of the operation of the arrival and departure lines of high-speed rail stations under safety constraints. *J. China Saf. Sci.* **31**(01), 179–185 (2021). (in Chinese)
11. Lu, G., Peng, H., He, B., Zhang, B.: Railway station approach time conflict degree and its approximate calculation method. *China Railway Sci.* **42**(06), 181–193 (2021). (in Chinese)



Optimization of Vehicle Body Modal Recognition Measuring Point Based on Particle Swarm Optimization

Xiao Sun and Gong Dao (✉)

College of Transportation Engineering, Tongji University, Shanghai 201804, China
{1953640, gongdao}@tongji.edu.cn

Abstract. Vibration is an important index to evaluate the operating safety, comfort and performance of rail transit vehicles. In order to further obtain the vibration characteristics and characteristics of structures, modal analysis is generally used as an effective way. If the structure has a natural mode in a certain frequency range, its characteristics can be estimated and determined by modal analysis method, and the actual structural vibration response of the structure to external or internal excitation source excitation and transmission under the natural frequency range corresponding to the mode can be calculated on this basis. This paper takes rail vehicle as the research object, aiming at the test modal analysis, the layout of modal test sensor is further studied. By optimizing the layout position of modal test sensor of rail vehicle, the results of modal identification based on sensor test data are more accurate, convenient and effective. Based on particle swarm optimization algorithm, this paper proposes an improved scheme for the sensor configuration optimization existing in the current rail vehicle body modal analysis test test test. Based on engineering practice experience, combined with the actual test scenario of rail vehicle body modal test, and according to the characteristics of the three-dimensional sensor used in the actual test, an intelligent particle swarm algorithm based on modal confidence method is proposed. For the actual test of the rail vehicle body mode, the position of the measuring point obtained by the particle swarm optimization algorithm is adjusted according to the test ability of the laboratory and the consideration of the integrity of the vibration mode, so as to maximize the test ability.

Keywords: Rail Vehicle · Measurement Point Optimization · Particle Swarm Optimization · Modal Test

1 Introduction

Vibration is an important index to evaluate the operating safety, comfort and performance of rail transit vehicles. In order to further obtain the vibration characteristics and characteristics of the structure, the generally used effective way is modal analysis. If the structure has an inherent mode in a certain frequency range, its characteristics can be estimated and determined by modal analysis method, and the actual structural vibration

response of the structure to external or internal excitation source excitation and transmission can be calculated according to the corresponding natural frequency range of the mode [1].

For simple structures, the location of measuring points can be determined by visual means of observation of vibration patterns. Generally speaking, this method of measuring points layout has a good effect. For the modal analysis test of a large and complex mechanical mechanism, it is necessary to first select the test degree of freedom, and then conduct modeling and computational simulation of the model through finite element software. Finally, the obtained measurement points are used as the target measurement points of the test for modal test experiments [2].

Before carrying out the structural modal test experiment, it is necessary to select and optimize the measuring points to determine the best number of measuring points and the optimal spatial coordinate position. Therefore, under the condition of limited measuring points, the research on the optimization and arrangement of measuring points has important engineering practical value and scientific research significance for the modal identification of vehicle body.

In modal test, the sensor layout directly affects the quality of frequency response function and the effect of modal recognition. The key point of sensor layout is the position and number of sensors used in the test. In general, during the test, the sensor usually adopts the traditional uniform distribution measuring point method, that is, the section is divided along the length direction of the car body, and the measuring point is evenly arranged on the section [3]. However, the uniform distribution measuring point method is completely based on the experience of engineers. Some measuring points may be at the position of modal nodes, which makes little contribution to the identification of modal information.

Gong et al. [4–6] proposed a method for determining car body and bogie based on modal assurance criteria, modal similarity and phase difference for current high-speed EMU, studied the influence of multi-modal vibration on the riding quality of high-speed EMU trains, and determined the target mode of vibration control. Juang et al. [7] put forward a feature system implementation algorithm and a time-domain analysis method based on the system minimum realization theory, which is suitable for the structure of multi-input multi-output system. The impulse response was used to construct the Hankel matrix and perform singular value decomposition on it. The system matrix was constructed with the decomposition results to realize modal identification and parameter acquisition. Guyan et al. [8] proposed Guyana Reduction method, which divided the model degrees of freedom into principal degrees of freedom and secondary degrees of freedom, then established constraint equations to optimize the model, and applied kinetic energy equation or strain energy equation to iteratively reduce the model, so as to obtain the exact information of modal principal degrees of freedom. By placing the measuring points at the position of the principal degrees of freedom, the modal information of the system can be better identified [9, 10]. Came et al. [11] defined the Modal Assurance Criterion (MAC) as the evaluation index to measure the spatial Angle of the modal vector. The index takes the minimum non-diagonal element value as the core idea when optimizing the sensor layout. The lower the value of non-diagonal element of the index

of modal assurance criterion, the greater the Angle of modal vector in space, the lower the similarity between modal vectors, and the better the mode recognition effect.

2 Theoretical Algorithm

2.1 Modal Confidence Method Theory

Modal confidence method, also known as minimum modal confidence method, is a method of base MAC modal confidence index. The smaller the inner product cosine value of the modal vector, the higher the mutual independence between the modal vectors, the stronger the anti-noise interference ability, and the better the mode recognition effect. According to the theory of structural dynamics, the modal vectors of each order on the structural node are orthogonal to each other, but because the degree of freedom of the measurement point is much smaller than the degree of freedom of the actual rail vehicle body structure during the actual modal test, it is difficult to ensure that the modal vectors obtained by analyzing and processing the actual test data are orthogonal to each other. If the Angle of the modal vector space is found to be too small, it indicates that the sensor group arranged by the method is unreasonable, which makes the recognition of part of the modal information blocked. Therefore, in order to evaluate the spatial intersection Angle of modal vectors, an evaluation scheme for sensor arrangement, namely modal confidence matrix (MAC matrix), was defined.

MAC matrix reflects the degree of independence of the modes by comparing the correlation coefficient between the modes of each order. If s sensors are arranged in an n -degree-of-freedom system, the modal order can be identified as m . The modal formation matrix of S -degree-of-freedom identified by the test scheme is, and since the degree of freedom of the test is less than the degree of freedom of the system n , therefore. $\Phi_S \in R^{s \times m}$. Modal confidence matrix $MAC = (mac_{ij})$, the expression can be written as:

$$mac_{ij} = \frac{(\varphi_{si}^T \varphi_{sj})^2}{(\varphi_{si}^T \varphi_{si})(\varphi_{sj}^T \varphi_{sj})} \quad (1)$$

The above formula is used to represent the Angle between the i mode vector and the j mode vector. Where, $\Phi_S = (\varphi_{s1} \varphi_{s2} \dots \varphi_{sm})$, φ_{si} and φ_{sj} represent column i and column j vectors of the mode modes of the test degrees of freedom, respectively. Therefore, it can be deduced that when the value of the non-diagonal element mac_{ij} ($i \neq j$) of the MAC matrix is lower and closer to 0, it can be proved that the orthogonality of the two modal modal vectors is better, the linear independence of the two modal modal modes is higher, the rationality of the sensor arrangement is higher, and the recognition and analysis effect of each order modal modal is better. Conversely, when the value of the non-diagonal element mac_{ij} ($i \neq j$) of MAC matrix is closer to 1, it is proved that the orthogonality of the two mode mode vectors is worse, the mode mode recognition effect is poor, and there may be modal coupling or failure to identify the concerned target modes.

Modal confidence method has defects, mainly including low modal strain energy in the determined sensor position and poor identification effect on noise interference, resulting in easy loss of information. More importantly, the above effective independent

method can only be used for measurement point optimization in single-dimensional coordinate system, and is highly applicable to some structures, such as narrow and long structures such as Bridges, whose length is much larger than their width and height. Its modes can only be considered vertical. To be extended to the modal recognition of rail vehicle body, the algorithm needs to be further optimized.

2.2 Particle Swarm Optimization Algorithm Based on Modal Confidence Method

The mathematical expression of particle swarm algorithm is as follows: a particle swarm composed of q particles flies and searches at a certain speed in D -dimensional space, and each particle has three attributes: current spatial position, historical optimal spatial position and velocity. Assuming that the current is the s iteration, then the i -th particle in the D -dimensional space can be represented as:

$$\begin{cases} X_i^S = (x_{i1}^S, x_{i2}^S, x_{i3}^S, \dots, x_{iD}^S) \\ V_i^S = (v_{i1}^S, v_{i2}^S, v_{i3}^S, \dots, v_{iD}^S) \\ P_i^S = (p_{i1}^S, p_{i2}^S, p_{i3}^S, \dots, p_{iD}^S) \end{cases} \quad (2)$$

where X_i^S is the current position of the particle in space, V_i^S is the velocity of the particle, and P_i^S is the current individual optimal position of the particle. X_i^S, V_i^S, P_i^S The optimal position of the current group can be represented by the following formula

$$P_g = (p_{g1}, p_{g2}, p_{g3}, \dots, p_{gD}) \quad (3)$$

At this time, the fitness of each particle is calculated, and it can be determined that in the s generation, the current optimal position of each particle and the optimal position of the group, and the particle updates its own speed and position according to the three parameters of the current speed, the optimal position of the individual and the optimal position of the group, which can be expressed in the following formula

$$v_{in}^{s+1} = \omega v_{in}^s + c_1 r_1 (p_{in}^s - x_{in}^s) + c_2 r_2 (p_{gn}^s - x_{in}^s) \quad (4)$$

$$x_{in}^{s+1} = x_{in}^s + v_{in}^{s+1} \quad (5)$$

2.3 Finite Element Model and Modal Analysis Results

The car body of the analysis and calculation object in this paper is an aluminum alloy structure, and the parts of the car body are divided into four parts according to the location, including side wall, floor, roof and end wall. According to the distribution quality of wiring and cables in the car, it is applied through the distributed mass point according to the specific distribution area. For the auxiliary equipment with mass concentration, such as traction inverter, high pressure box, air spring attached air cylinder, auxiliary brake module, etc., a node is created in the corresponding center of gravity position of the equipment and connected through the elastic unit. The main material of the car body is aluminum alloy, and the material parameters are shown in Table 1.

Table 1. Material parameters of car body.

Parts	Materials	Modulus of elasticity	Poisson's ratio	Density
Car body	6005A-T6	69Gpa	0.33	2.70e-9 t/mm3
Traction beam under car	6082-T6	70Gpa	0.30	2.70e-9 t/mm3

The modal distribution of the reassembled body is shown in Table 2. It can be seen that the first typical mode of the car body, diamond mode, appears at 7.15Hz. At 14.56Hz, a second-order rhomboid mode appears, and at 17.55Hz, a third-order rhomboid mode appears, after which no obvious car body mode appears. Generally, when considering the vehicle body modes, the third order rhomboid modes are concerned at most, that is, the modes within 20Hz are considered, and the first four order elastomer modes are mainly concerned, as shown in Fig. 1. The typical modes of the vehicle body are basically guaranteed to cover, which can effectively guarantee the optimization results of the vehicle body measurement points.

Table 2. Mode distribution table of the repaired body.

Mode of vibration	Mode frequency
First-order diamond	7.15
First order droop	8.41
First order twist	9.55
First order horizontal bend	14.04

3 Optimization of Modal Analysis Measuring Points Based on Particle Swarm Optimization

3.1 Optimization of Measuring Point Set

For the modal testing personnel, the modal measuring points are usually evenly arranged according to the structural characteristics, so as to ensure that the identified modes have a good consistency, but when using the uniform distribution measuring point method to arrange, it may cause errors due to the symmetry of the structure itself. Because the symmetrical structure has symmetrical modes, and it is divided into the same direction symmetry and reverse symmetry, if the measurement point is arranged in a symmetrical position, there will be a risk of partial modal loss. In order to select the optimal configuration of the measuring points for modal testing, a certain amount of candidate measuring points should be selected to form an initial measuring point set, and the optimization algorithm of measuring point layout should be used to optimize the configuration until the required number of test sensors is retained. Therefore, 160 feasible measuring points are selected as candidate measuring points set among the points where sensors can be arranged on the vehicle body, as shown in Fig. 2.

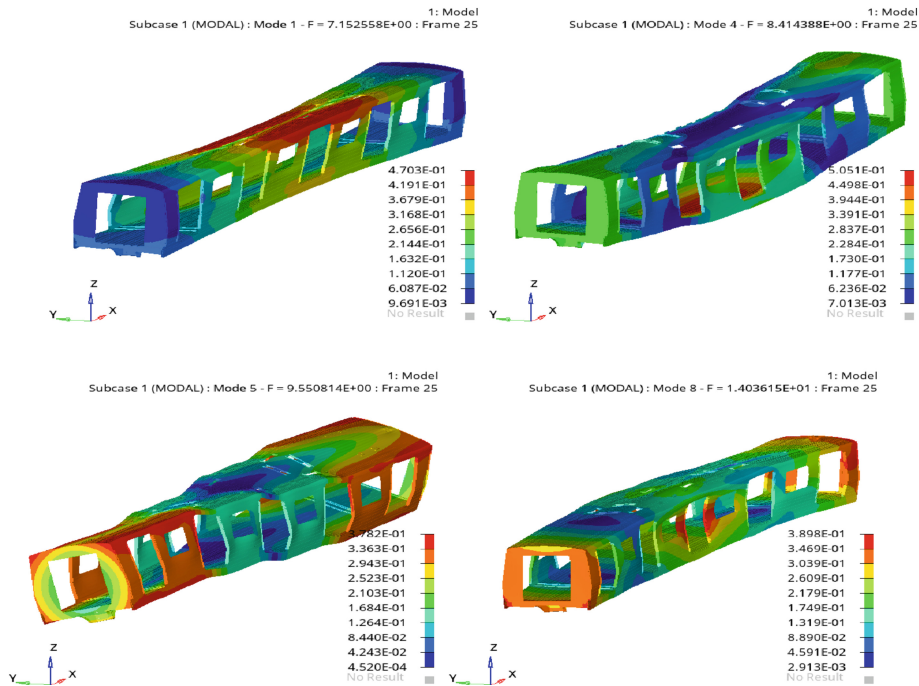


Fig. 1. Reconditioning the body mode.

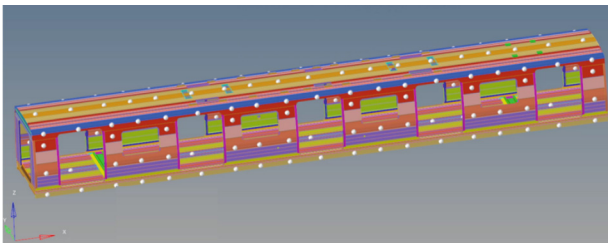


Fig. 2. Location diagram of candidate measuring point set.

The modal confidence matrix is used as the criterion to measure the quality of the measuring point arrangement, and the modal modal matrix of the candidate measuring point is extracted by finite element simulation to calculate the MAC matrix. By comparing the maximum value of the non-diagonal elements of the MAC matrix, the MAC matrix of the candidate measuring point set is shown in Fig. 3. The maximum value of the non-diagonal elements of the MAC matrix of the candidate measuring points set ensures that the maximum value of the non-diagonal elements is less than 0.25. However, subject to the actual test conditions, 160 sensors cannot be arranged during the modal test test. Therefore, it is necessary to optimize the configuration of measuring points. Under the premise of ensuring that the maximum value of non-diagonal elements of

the MAC matrix is not higher than 0.25 and the number of sensors is higher than the concerned target modal order, the number of sensors required should be reduced as much as possible to make full use of test resources.

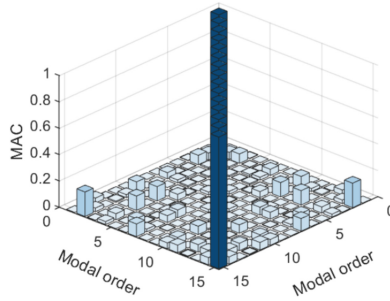


Fig. 3. MAC matrix diagram of candidate test point set.

The vertical and vertical (XYZ direction) mode mode information of the candidate measurement point set was extracted and optimized by particle swarm optimization algorithm. MAC matrix was calculated according to the mode mode mode matrix extracted by finite element simulation. The maximum minimization of non-diagonal elements of MAC matrix was selected as the objective function, and the initial population was selected to be 80, 40 and 20, respectively, for optimization calculation.

After optimization, three groups of measurement points were generated. Node numbers were used in the finite element model to confirm the positions and mark them. The optimized measurement points were selected and simulated in the finite element software to extract the vibration information obtained from the simulation calculation for modal calculation. It can be determined from Fig. 4 and Table 3 that the maximum value of non-diagonal elements of MAC matrix gradually decreases with the decrease of the number of measuring points, and gradually increases when the number of measuring points is reduced to a certain extent. If we focus on the first four order elastomer modes of the car body, the minimum number of measuring points is 40.

The mode mode mode information of the candidate measuring points set in XYZ direction was extracted, optimized by particle swarm optimization algorithm. MAC matrix was calculated according to the mode mode mode matrix extracted by finite element simulation. The maximum minimization of non-diagonal elements of MAC matrix was selected as the objective function, and 40 measuring points were selected for optimization calculation.

After optimization, 40 measurement points are generated. Node numbers are used in the finite element model to confirm the positions and mark them. The location diagram of the optimized measurement points is shown in Fig. 5. The optimized measuring points were selected for polycondensation processing in the finite element software, and the modal shape matrix obtained by simulation calculation was extracted for MAC matrix calculation. The MAC matrix obtained was shown in Fig. 6, and the maximum value of non-diagonal elements was 0.13.

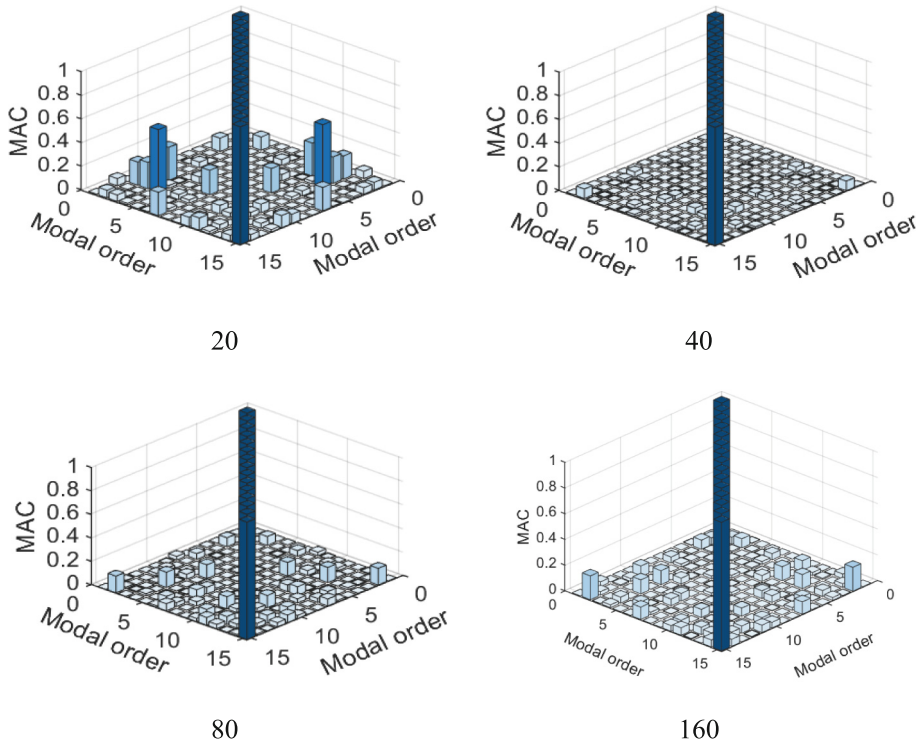


Fig. 4. Comparison of the number optimization effect.

Table 3. Modal frequencies of different test schemes.

Number of test points	160	80	40	20
Vibration form	Frequency (Hz)	Frequency (Hz)	Frequency (Hz)	Frequency (Hz)
First order diamond	7.11	7.15	7.15	9
First order droop	8.4	8.41	8.41	9.21
First order twist	9.53	9.5	9.55	10.4
First order horizontal bend	13.79	13.95	14.04	14.98

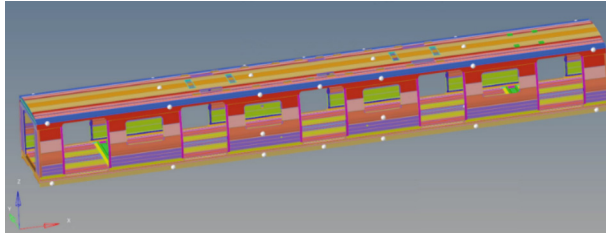


Fig. 5. Location diagram of measuring points optimized by particle swarm optimization algorithm.

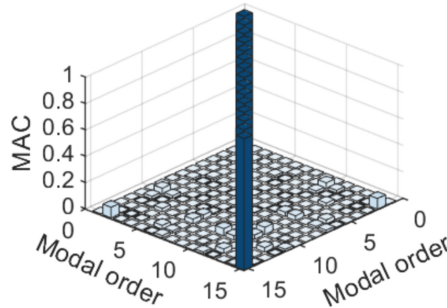


Fig. 6. MAC matrix of particle swarm optimization of measuring points.

4 Conclusion

When optimizing the measuring point, the multiple degrees of freedom of each measuring point are unified into the whole consideration. The mode mode of XYZ direction can be considered comprehensively by choosing the uniform measuring point method, but the mode mode coupling is serious, which is not easy to separate the natural mode and forced excitation. The particle swarm optimization method comprehensively considers the contribution degree of measuring points to the XYZ direction mode, and the distribution is more uniform, which has guiding significance for vehicle modal test sensors.

Acknowledgements. The authors gratefully acknowledge the financial support from National Natural Science Foundation of China (No. 52375115), National Natural Science Foundation of China (No. 52475122) and Supported by the Fundamental Research Funds for the Central Universities (No. 22120240338).

References

1. Wang, Z., Gong, D.: A survey on countermeasures of railway vehicle stability decrease caused by the evolution of hollow-worn wheels. *J. Vibr. Control* (2024)
2. Liu, G., Gong, D., Zhou, J.: Research on resonance mechanism of articulated train car body based on modal vibration extraction method. *J. Vibr. Control* (2024)

3. TB/T 3502–2018. Modal Test method and Evaluation of railway passenger cars and EMUs. null, 2018–04–12. (in Chinese)
4. Dao, G., et al.: Frequency veering between car body and under-chassis equipment of railway vehicles in vertical bending mode. *Mech. Syst. Sig. Process.* 109768 (2023)
5. Gong, D., Duan, Y., Wang, K., Zhou, J.: Modelling rubber dynamic stiffness for numerical predictions of the effects of temperature and speed on the vibration of a railway vehicle car body. *J. Sound Vib.* **449**, 121–139 (2019)
6. Gong, D., Zhou, J., Sun, W., Sun, Y., Xia, Z.: Method of multi-mode vibration control for the carbody of high-speed electric multiple unit trains. *J. Sound Vib.* **409**, 94–111 (2017)
7. Juang, J.-N., Pappa, R.S.: An eigensystem realization algorithm for modal parameter identification and model reduction. *J. Guidance, Control, Dyn.* **8**(5), 620–627 (1985)
8. Robert, J.G.: Reduction of stiffness and mass matrices. *AIAA J.* **3**(2), 380–380 (1965)
9. Zhou, K., You, T., Gong, D., Zhou, J.: Effects of uncertain suspension parameters on dynamic responses of the railway vehicle system. *Probab. Eng. Mech.* **71**, 103405 (2023). <https://doi.org/10.1016/j.probengmech.2022.103405>
10. Zhang, Y., Zhou, K., Zhou, J., Chang, C., Gong, D., Sun, W.: Frictional vibration analysis of train braking system considering wheel-rail attachment and multi-body friction. *Int. J. Non-linear Mech.* **162**, 104715 (2024). <https://doi.org/10.1016/j.ijnonlinmec.2024.104715>
11. Came, T.G., Dohrman, C.R.: A modal test design strategy for modal correlation. In: *Proceedings of the 13th International Modal Analysis Conference*, SPIE The International Society for Optical Engineering, Bethel (1995)



High-Precision Belt Material Flow Tracking Research in Bulk Cargo Ports

Changan Li¹(✉), Baodong Sun¹, Huiqiang Li¹, Baoxin Qin², and Xiaoshu Qiao²

¹ National Energy Economic and Technological Research Institute Co., Ltd.,
Beijing 102299, China
11330437@ceic.com

² CHNENERGY (Tianjin) Port Co., Ltd., Binhai New District, TianJin 300270, China

Abstract. With the development of the global economy, bulk cargo ports, as important logistics hubs, are increasingly valued for their operational efficiency and safety. Belt conveyors, as the core transportation equipment in bulk cargo ports, play a significant role in improving work efficiency, reducing costs, and ensuring safety through precise tracking of material flow. This paper proposes a high-precision belt material flow tracking method that segments the belt and uses array elements to track the weight and position information of the material flow on each segment, combining the belt's operating speed and detection pulse information to achieve precise tracking of the belt material flow. This method can effectively prevent belt idling and material overload, reduce equipment wear and energy waste, and enhance the level of intelligent operation in the port.

Keywords: Bulk cargo ports · Belt material flow tracking · High precision · Intelligent control

1 Introduction

Bulk cargo ports, as key hubs in global trade, undertake a large amount of loading and unloading tasks for bulk goods such as coal and ore. Belt conveyors have become indispensable transportation tools within ports due to their efficient transportation capabilities. However, traditional belt control systems often lack precise tracking of material flow, leading to issues such as equipment idling and material overload, which increase operational costs and safety risks. Therefore, developing a high-precision belt material flow tracking technology is of great practical significance for improving port operation efficiency, reducing energy consumption, and ensuring work safety.

2 Technical Background

In bulk cargo ports, belt conveyors are crucial for connecting docks, stockyards, and ship loading/unloading areas, impacting operational efficiency and costs. Material flow tracking technology has advanced from simple, low-precision methods to more complex, high-precision systems.

2.1 Traditional Belt Material Flow Tracking Technology

Early detection methods relied on mechanical switches and contact sensors, which were limited in providing weight information and precise position tracking of material flow. They were also sensitive to wear and environmental factors, leading to high maintenance costs [1–3]. Photoelectric detection technology, while offering faster response times, still lacked weight information and was sensitive to ambient light. Ultrasonic detection could determine material flow position but not weight and was limited by material surface characteristics. Radar technology, though non-contact and high-precision, was costly and sensitive to environmental interference [4, 5].

2.2 Development Needs of Material Flow Tracking Technology

As port operations become more automated and intelligent, the demand for material flow tracking technology increases. High-precision tracking is essential for optimizing work processes, reducing waste, and improving safety. Real-time monitoring is necessary to adjust work parameters promptly, and the technology must be stable and reliable in varying conditions. Cost-effectiveness is critical for widespread adoption, and integration with other port systems is needed to enhance overall intelligence.

2.3 Limitations of Existing Technology

Existing technologies are limited in their ability to provide detailed material flow information, are sensitive to environmental conditions, and are often costly, which hinders their application in port operations. They also struggle to integrate with other information systems, limiting the improvement of port intelligence [6].

2.4 Directions for Technological Innovation

To overcome these limitations, future developments should focus on multi-sensor fusion for comprehensive monitoring, intelligent algorithms for data analysis, environmental adaptability for stable operation, cost-reducing innovations, and system integration for collaborative operations [7–9]. This approach will meet the precision, real-time performance, stability, and cost-effectiveness requirements of modern port operations, as addressed by the high-precision belt material flow tracking method proposed in this study.

3 Key Technical Implementation

The belt material flow tracking system proposed in this study aims to achieve high-precision tracking of material flow on belt conveyors in bulk cargo ports. The system uses a variety of sensors, advanced algorithms, and computing technologies to achieve real-time monitoring of material flow weight, position, and speed. The following are key technical details of the system implementation.

3.1 Acquisition of Material Flow Weight Information

Material flow weight information is the foundation of material flow tracking and the key to achieving material flow control and optimization. The system acquires material flow weight information in the following two ways:

3.1.1 Integration of Belt Scales

For belt conveyors equipped with belt scales, the system directly obtains the instantaneous flow rate of the material flow from the belt scale. Belt scales can provide continuous and real-time material flow weight data, which is an important means of achieving high-precision material flow tracking [10, 11]. The system connects to the interface of the belt scale and reads the instantaneous flow rate data output by the belt scale in real-time [12].

3.1.2 Determination of Periodic Material Feeding Quantity

For belt conveyors not equipped with belt scales, the system obtains material flow weight information by determining the periodic material feeding quantity of the belt. The periodic material feeding quantity refers to the amount of material that falls onto the belt from the upstream conveying equipment within a unit scanning period. The system determines the periodic material feeding quantity through the following steps:

- (1) Upstream Equipment Monitoring: The system monitors the operating status of the upstream conveying equipment, including operating speed and material falling points.
- (2) Material Falling Quantity Calculation: Based on the operating parameters and material characteristics of the upstream conveying equipment, calculate the material falling quantity per unit time.
- (3) Periodic Material Feeding Quantity Determination: Multiply the calculated material falling quantity by the scanning period time to obtain the periodic material feeding quantity.

3.2 Calculation of Belt Periodic Operating Distance

The belt periodic operating distance refers to the distance the belt travels within a unit scanning period and is a key parameter for calculating the position of the material flow. The system calculates the belt periodic operating distance through the following steps:

3.2.1 Monitoring of Belt Speed

The system uses speed sensors to monitor the operating speed of the belt in real-time. Speed sensors can be contact types, such as rotary encoders, or non-contact types, such as laser speedometers.

3.2.2 Determination of Scanning Period Time

The system determines the scanning period time based on the configuration and operating parameters of the scanning equipment. The scanning period time refers to the time required for the scanning equipment to complete one full scan.

3.2.3 Calculation of Periodic Operating Distance

The system uses the following formula to calculate the belt periodic operating distance:

$$S = v \times T \quad (1)$$

where S represents the belt periodic operating distance, v represents the belt speed, and T represents the time length of a unit scanning period.

3.3 Material Flow Tracking Algorithm

The material flow tracking algorithm is the core of the system, responsible for dynamically updating each array element based on the positions pointed to by the head pointer and tail pointer, the belt periodic operating distance, and the weight information of the belt material flow to achieve material flow tracking. The following are the key steps of the material flow tracking algorithm:

3.3.1 Definition of Array Elements

The system divides each belt into multiple segments according to a preset fixed scale, and each segment of the belt corresponds to multiple attribute array elements. Array elements include weight information and position information of the material flow on each segment of the belt. Array elements with the same attributes are arranged in order according to the corresponding belt segment to form a material flow array.

3.3.2 Setting of Head Pointer and Tail Pointer

The system sets a head pointer and a tail pointer for each belt. The head pointer is used to store the storage address of the material flow array corresponding to the material feeding port, and the tail pointer is used to store the storage address of the material flow array corresponding to the belt discharge port.

3.3.3 Update of Array Elements

The system updates the array elements according to the following steps:

- (1) Update of the material feeding port array elements: The system updates the array elements at the material feeding port based on the position pointed to by the head pointer, the belt periodic operating distance, and the belt's periodic material feeding quantity.
- (2) Update of the discharge port array elements: The system updates the array elements at the discharge port based on the position pointed to by the tail pointer and the belt periodic operating distance.
- (3) Accumulated weight calculation: The system accumulates the weight information of the material flow on each segment of the current belt to obtain the accumulated weight of the material flow between the current belt and the upstream conveying equipment.

3.3.4 Implementation of Material Flow Tracking

The system uses the following formula to calculate the periodic material feeding quantity $Z(i)$ of the belt during the i -th sampling:

$$Z(i) = F(i) \times T(i) \quad (2)$$

where $F(i)$ represents the instantaneous flow rate of the belt scale, and $T(i)$ represents the unit scanning period time.

If the belt periodic operating distance during the n -th sampling does not exceed the preset fixed scale length, it is determined that during the n -th sampling, the weight information of the array elements $Ad[t]$ and the adjacent array elements $Ad[t + 1]$ corresponding to the same attribute at the position pointed to by the head pointer of the belt to be monitored is:

$$Ad[t] = Z(1) + Z(2) + \dots + Z(n) \quad (3)$$

$$Ad[t + 1] = Ad[t + 2] = \dots = Ad[t + m - 1] = 0 \quad (4)$$

where m represents the number of array elements with the same attributes as the belt to be monitored.

If the belt periodic operating distance during the n -th sampling exceeds the preset fixed scale length, it is determined that during the n -th sampling, the weight information of the array elements $Ad[t]$ and $Ad[t + 1]$ corresponding to the segment of the belt material flow is:

$$Ad[t] = Z(1) + Z(2) + \dots + Z(n) - \frac{(L(1) + L(2) + \dots + L(n)) - K}{L(n)} \times Z(n) \quad (5)$$

$$Ad[t + 1] = \frac{(L(1) + L(2) + \dots + L(n)) - K}{L(n)} \times Z(n) \quad (6)$$

where K is the length of the preset fixed scale, and $L(n)$ is the belt periodic operating distance during the n -th sampling.

3.4 System Module Implementation

The system consists of the following three main modules, which work together to achieve high-precision tracking of material flow.

3.4.1 Information Acquisition Module

The information acquisition module is responsible for real-time acquisition of weight information of belt material flow, belt operating speed, and belt detection pulse information. This module includes:

- (1) Weight information acquisition unit: Connects to the belt scale interface to read material flow weight data in real-time.
- (2) Speed information acquisition unit: Monitors the belt operating speed through speed sensors.
- (3) Pulse information acquisition unit: Obtains belt detection pulse information through pulse sensors.

3.4.2 Belt Periodic Operating Distance Acquisition Module

The belt periodic operating distance acquisition module calculates the belt periodic operating distance based on the belt operating speed and belt detection pulse information. This module includes:

- (1) Speed processing unit: Processes the data from speed sensors to calculate the real-time speed of the belt.
- (2) Pulse processing unit: Processes the data from pulse sensors to determine the average number of pulses within a scanning period.
- (3) Distance calculation unit: Calculates the belt periodic operating distance based on speed and pulse data.

3.4.3 Material Flow Tracking Module

The material flow tracking module updates the array elements of each belt to be monitored based on the positions pointed to by the head pointer and tail pointer, the belt periodic operating distance, and the weight information of the belt material flow, to achieve material flow tracking. This module includes:

- (1) Array management unit: Manages the material flow array and maintains the order and attributes of the array elements.
- (2) Pointer management unit: Manages the head pointer and tail pointer, updating the positions pointed to by the pointers.
- (3) Tracking algorithm execution unit: Executes the material flow tracking algorithm, updates the array elements, and achieves material flow tracking.

3.5 Technical Advantages and Innovation Points

The key technical implementation of the system has the following advantages and innovation points:

- (1) High-precision tracking: The system achieves high-precision tracking of material flow weight and position by integrating various sensors and advanced algorithms.
- (2) Real-time monitoring: The system can monitor the state of material flow in real-time, respond promptly to changes in material flow, and optimize work processes.
- (3) Strong environmental adaptability: The sensors and algorithms used by the system have good environmental adaptability and can work stably in harsh port environments.
- (4) High cost-effectiveness: The system reduces costs through technological innovation while ensuring high-precision tracking, improving cost-effectiveness.
- (5) High system integration: The system can integrate with other information systems in the port, achieving data sharing and collaborative operations, and enhancing the overall level of port intelligence.

In summary, the key technical implementation of the system provides strong technical support for high-precision tracking of belt material flow in bulk cargo ports and lays a solid foundation for the development of port intelligence.

4 Experiments and Results

To verify the effectiveness of the proposed high-precision belt material flow tracking method, a series of experiments were conducted in an actual bulk cargo port environment. This section will detail the experimental setup, process, results, and analysis.

4.1 Experimental Setup

The experiment was conducted in a typical bulk cargo port, involving several belt conveyors of different lengths and widths. The purpose of the experiment was to verify the accuracy, stability, and practicality of the proposed tracking method in an actual environment.

4.1.1 Experimental Environment

The experimental environment includes the following parts:

- (1) Belt conveyors: Five belt conveyors with lengths ranging from 100 m to 500 m were selected as the experimental subjects.
- (2) Sensor layout: Speed sensors, pulse sensors, and belt scales were installed on each belt conveyor.
- (3) Data acquisition system: A high-sampling-rate data acquisition system was used to ensure that the detailed information of belt operation and material flow changes could be captured.
- (4) Environmental monitoring equipment: Environmental monitoring equipment was installed to record the environmental conditions during the experiment, such as temperature, humidity, and wind speed.

4.1.2 Experimental Materials

The following materials were used in the experiment:

- (1) Materials: Coal and ore were used as materials in the experiment to simulate the typical operating conditions of bulk cargo ports.
- (2) Sensors and instruments: High-precision speed sensors, pulse sensors, and belt scales were used.

4.1.3 Experimental Parameters

The key parameters in the experiment include:

- (1) Belt speed: Different belt speeds were set to simulate different operating conditions.
- (2) Scanning period: Different scanning periods were set to test the performance of the system under different scanning frequencies.
- (3) Material flow weight: Different material flow weight conditions were simulated by changing the upstream material falling quantity.

4.2 Experimental Process

The experiment was divided into the following steps:

4.2.1 System Calibration

Before the start of the experiment, all sensors and instruments were calibrated to ensure the accuracy of the data.

4.2.2 Data Collection

During the experiment, real-time data such as belt speed, pulse signals, and material flow weight were collected.

4.2.3 Material Flow Tracking

The proposed tracking method was used to track the weight and position of the material flow based on the collected data in real-time.

4.2.4 Data Analysis

The collected data were analyzed to verify the accuracy and stability of the tracking method.

4.3 Experimental Results

The experimental results show that the proposed high-precision belt material flow tracking method can accurately track the weight and position information of the belt material flow, with the error controlled within $\pm 0.5\%$, meeting the needs of high-precision tracking. At the same time, the method shows good stability and reliability under different environmental conditions.

4.3.1 Tracking Accuracy

By comparing with the actual material flow weight, the accuracy of the tracking method was verified. The results show that the tracking method can accurately track the material flow weight, and the error is within an acceptable range.

4.3.2 Tracking Stability

Under different belt speeds and material flow weight conditions, the tracking method can work stably, without obvious fluctuations or increased errors.

4.3.3 Environmental Adaptability

Under different environmental conditions, such as different temperatures and humidities, the tracking method can maintain stable performance, showing good environmental adaptability.

4.4 Result Analysis

4.4.1 Tracking Accuracy Analysis

Tracking accuracy is a key indicator for evaluating the performance of the tracking method. By comparing the tracking results with the actual values, we can draw the following conclusions:

- (1) The tracking method can accurately track the material flow weight, with the error controlled within $\pm 0.5\%$, which is sufficiently accurate for practical applications.
- (2) Even when the material flow weight changes significantly, the tracking method can still maintain high precision, indicating that the tracking method has good dynamic response capabilities.

4.4.2 Tracking Stability Analysis

Tracking stability is an important indicator for evaluating the reliability of the tracking method. By analyzing the experimental data, we can draw the following conclusions:

- (1) During the experiment, the tracking method showed good stability and could maintain stable tracking performance even when the material flow conditions changed.
- (2) The stability of the tracking method is not affected by belt speed and material flow weight, indicating that the tracking method has good robustness.

4.4.3 Environmental Adaptability Analysis

Environmental adaptability is an important indicator for evaluating the practicality of the tracking method. By conducting experiments under different environmental conditions, we can draw the following conclusions:

- (1) The tracking method can maintain stable performance under different temperature and humidity conditions, showing good environmental adaptability.
- (2) The tracking method is not affected by external electromagnetic interference, which is particularly important in the complex electromagnetic environment of ports.

4.5 Experimental Conclusion

The experimental results fully prove the effectiveness of the proposed high-precision belt material flow tracking method. The method can not only accurately track the weight and position of the material flow but also maintain stable performance under different operating and environmental conditions. These characteristics make the method valuable for actual operations in bulk cargo ports.

5 Conclusion

The high-precision belt material flow tracking method proposed in this paper has been fully proven to be effective, accurate, and stable in the tracking of material flow on belt conveyors in bulk cargo ports after detailed theoretical analysis, experimental verification, and result analysis. The following is a summary of this study and suggestions for future work.

5.1 Method Effectiveness

The experimental results show that the belt material flow tracking method proposed in this study can effectively track the weight and position information of the material flow on the belt conveyor. By accurately measuring the belt periodic operating distance, obtaining real-time material flow weight information, and combining an advanced tracking algorithm, the method can ensure high precision and stability of the tracking results. The error is controlled within $\pm 0.5\%$, which is significant in practical industrial applications and can greatly improve the efficiency and safety of port operations.

5.2 Technical Advantages

The technical advantages of this study are mainly reflected in the following aspects:

- (1) High-precision tracking: By precise measurement and real-time data processing, high-precision tracking of material flow weight and position is achieved.
- (2) Real-time monitoring capability: The system can respond to changes in material flow in real-time, providing immediate data support for port operations.
- (3) Environmental adaptability: Under different environmental conditions, including temperature, humidity, and electromagnetic interference, the system can maintain stable performance.
- (4) Cost-effectiveness: Compared to traditional high-end tracking technologies, this method significantly reduces costs while ensuring high precision.
- (5) System integration: This method can be seamlessly integrated with existing port information systems, achieving data sharing and collaborative operations.

5.3 Practical Application Value

The belt material flow tracking method proposed in this study has significant value in practical applications:

- (1) Improve work efficiency: By accurately tracking material flow and optimizing work processes, reduce material waste and equipment idling, and improve work efficiency.
- (2) Reduce operational costs: Reduce equipment wear and energy consumption, lower maintenance costs, and achieve cost savings.
- (3) Enhance work safety: Avoid material overload and equipment idling, reduce the risk of accidents, and ensure the safety of workers and equipment.
- (4) Support intelligent development: Provide technical support for port intelligence and automation, promoting the development of port logistics to a higher level.

5.4 Contribution to Port Intelligence

The contribution of the belt material flow tracking method proposed in this study to port intelligence is mainly reflected in:

- (1) Data support: Provide accurate material flow data for port intelligent systems, supporting intelligent decision-making.
- (2) Process optimization: Optimize the material flow transportation process through real-time monitoring and data analysis, improving overall work efficiency.

- (3) Forecasting and early warning: Use tracking data for trend analysis and forecasting, achieving early warning of material flow abnormalities.
- (4) System integration: Promote the integration of the material flow tracking system with other port information systems, achieving information sharing and collaborative operations.

5.5 Suggestions for Future Work

Although this study has achieved significant results, there are still some areas that require further research and development:

- (1) Algorithm optimization: With the development of artificial intelligence technology, more advanced algorithms, such as deep learning, can be considered for material flow tracking to further improve tracking accuracy and intelligent level.
- (2) System scalability: Research how to extend the material flow tracking system to more port equipment and more complex operating environments to achieve broader applications.
- (3) Data analysis and mining: In-depth study of material flow data, mining the potential value behind the data, and providing more decision support for port operations.
- (4) System integration testing: Conduct system integration testing in actual port environments to verify the system's performance and stability under different conditions.
- (5) User interaction and interface design: Develop more user-friendly user interaction interfaces, making it easier for operators to use the system and improving system usability.

5.6 Conclusion

In summary, the high-precision belt material flow tracking method proposed in this study has shown excellent performance in both theory and practice. The method can not only improve the work efficiency and safety of bulk cargo ports but also promote the development of port intelligence. With the continuous advancement of technology and the deepening of applications, the method is expected to play a greater role in future port logistics.

References

1. Sabih, M., Farid, M.S., Ejaz, M., et al.: Raw material flow rate measurement on belt conveyor system using visual data. *Appl. Syst. Innovation* **6**(5), 88 (2023). <https://doi.org/10.3390/asi6050088>
2. Zhang, X., Yang, Z., Zhang, M., et al.: Quantitative monitoring method for conveyor belt deviation status based on attention guidance. *Appl. Sci.* **14**(16), 6916 (2024). <https://doi.org/10.3390/app14166916>
3. Mahmood, M.S., Shareef, I.R.: Applications of artificial intelligence for smart conveyor belt monitoring systems: a comprehensive review. *J. Européen des Systèmes Automatisés*. 1195–1206 (2024). <https://doi.org/10.18280/jesa.570426>

4. Ji, J., Miao, C., Li, X., et al.: Speed regulation strategy and algorithm for the variable-belt-speed energy-saving control of a belt conveyor based on the material flow rate. *PLOS ONE* (2021). <https://doi.org/10.1371/journal.pone.0247279>
5. Sun, G.: Application of intelligent main coal flow transportation system in mines. *Jiangxi Coal Sci. Technol.* **03**, 267–269 (2024). (in Chinese)
6. Yang, Y.P., Shi, Y., Li, L., et al.: Development of integrated control system based on coal flow scene monitoring. *Shaanxi Coal* **43**(07), 166–170+176 (2024). <https://doi.org/10.20120/j.cnki.issn.1671-749x.2024.0732>. (in Chinese)
7. Duan, X.X., Miao, D., Wang, Y.M., et al.: A laser detection method for belt conveyor coal flow based on b-spline curve fitting. *J. Tianjin Univ. Technol. Educ.* **34**(02), 56–61 (2024). <https://doi.org/10.19573/j.issn2095-0926.202402009>. (in Chinese)
8. Lian, X.: Design and application analysis of a collaborative control system for belt conveyor coal flow. *Coal* **33**(07), 103–105 (2024). (in Chinese)
9. Wang, F.: Research on energy-saving control system of belt conveyor based on coal flow visual detection. *Min. Equip.* **06**, 135–137 (2024). (in Chinese)
10. Ji, X.L., Zhang, W.J., Wang, Y.Q., et al.: Research on high-precision coal flow detection of belt conveyor based on machine vision. *Mine Autom.* **50**(05), 75–83 (2024). <https://doi.org/10.13272/j.issn.1671-251x.2024030028>. (in Chinese)
11. Yong, J.W., Qiao, T.Z., Ji, J., et al.: A coal flow monitoring method for conveyor belts based on binocular real-time matching. *Coal Technol.* **43**(06), 220–223 (2024). <https://doi.org/10.13301/j.cnki.ct.2024.06.046>. (in Chinese)
12. Li, Y.D.: Energy-saving operation control of belt conveyor based on coal flow detection. *Mech. Manag. Dev.* **39**(03), 189–191 (2024). <https://doi.org/10.16525/j.cnki.cn14-1134/th.2024.03.071>. (in Chinese)



Enhancing Online HD Map Construction with Trajectory Guidance in Challenging Conditions

Ziying Yao, Zhongxia Xiong, and Xinkai Wu(✉)

Beihang University, Beijing, China
xinkaiwu@buaa.edu.cn

Abstract. High-definition maps (HD maps) are crucial for precise navigation of autonomous vehicles. However, creating and maintaining HD maps involves substantial costs and challenges. Hence, the online generation of HD maps using on-board sensors has gained considerable attention. Existing methods often encounter data visibility difficulties due to factors such as occlusion, long distance, etc., which are common challenges in autonomous driving. In this paper, we propose the framework to improve the accuracy of online HD map generation by introducing implicit clues from the trajectories of traffic participants. We specifically leverage the trajectory data from the driving scene and encode trajectories as an additional branch. Although trajectory data may contain tracking errors or noises, we adopt an attention-based architecture to adaptively focus on relevant trajectory information, thus significantly improving the performance. We use nuScenes dataset to evaluate the proposed method, and the experiments demonstrate that trajectory information can enhance the performance of online map construction in rasterized formats.

Keywords: Autonomous driving · Online HD map · Trajectory

1 Introduction

High-definition (HD) maps are an indispensable task for autonomous driving, offering critical elements like road boundaries, road dividers, and crosswalks [1, 2]. This detailed information is vital for downstream tasks such as trajectory prediction [3, 4] and path planning [5]. With HD maps, autonomous vehicles can effectively optimize the decision-making process with lane following, lane changing, and obstacle avoidance. Traditional methods for creating HD maps are built using SLAM-based approaches, which are time-consuming and difficult to implement, and they cannot be considered a one-time effort since they must be frequently updated and expanded to cope with changes in urban environments and road conditions. Consequently, there is a growing recognition of the importance of reducing reliance on HD maps. To tackle the challenges, researchers are placing increasing emphasis on online HD map construction leveraging onboard sensors to create maps in real time. Recently, several researches [1, 6] have employed multi-modal online sensor inputs (such as LiDAR data and surrounding multi-view RGB camera images) to generate HD maps. While these methods have demonstrated promising

results, they still encounter challenges in certain areas in challenging visibility conditions, which can be due to factors such as faded lane markings, occlusions from vehicles or pedestrians, and adverse weather conditions that obscure visibility and hinder the accurate detection of map elements in autonomous driving scenarios.

Hence, integrating additional information, such as prior knowledge about road conditions and vehicle behavior, becomes essential. This additional information can help mitigate the limitations of online perception, improving the overall accuracy and reliability of autonomous navigation in challenging environments. Recent researches have attempted to use prior maps to supplement perception data. For example, Xiong *et al.* [7] design memory-based priors without utilizing existing offline maps, leading to updates solely through model training, which results in limited transferability to new scenes. Similarly, Sun *et al.* [8] focuses on existing HD maps by simulating past maps through the introduction of offsets and erasure of map elements. However, this method struggles to accurately represent the differences between past and present HD maps. Jiang *et al.* [9] embeds OSM map priors into the model in an attention-based manner by converting vector maps into raster representations. These outdated maps, due to infrequent updates, often do not accurately reflect real-time changes in road structures, making them inadequate for current navigation needs.

In this paper, we have introduced a framework, which utilizes trajectory information to guide the construction of high-definition maps, because trajectory data helps to supplement and enhance the mapping process by providing insights into real-time vehicle movements and road usage. This information can compensate for areas where lane markings may be ambiguous or obscured, allowing us to identify critical features based on actual driving behavior. By leveraging trajectory data, we can improve the accuracy and relevance of HD maps, ensuring they better reflect real-world conditions and support effective autonomous tasks. Experiments on the nuScenes dataset show that the proposed method improves performance under various challenging visibility conditions. In conclusion, we present the following contributions:

- (1) We propose a framework that uses trajectory priors and encodes weakly aligned trajectory features to assist online HD map construction implicitly.
- (2) We achieve improved results in HD map generation on public benchmarks in diverse conditions.

2 Related Work

2.1 Online HD Map Construction

For the advancement of autonomous driving technologies, online HD map construction is an essential task [5, 10]. Traditionally, HD maps have been created through manual annotation and offline processes, which involve the integration of point cloud data via SLAM algorithms [11]. While this method provides high accuracy, it is resource-intensive and lacks the capability for real-time updates.

Recent research has shifted focus toward the use of onboard sensors, enabling a more efficient and cost-effective approach to generating online HD maps [6, 12]. For example, Li *et al.* [1] utilizes pixel-wise labeling along with heuristic post-processing techniques, measuring performance through metrics such as Intersection over Union

(IoU) and Average Precision (AP). Emerging techniques [2, 13, 14] have further evolved this field by employing end-to-end vectorized HD map generation methods that leverage advanced Transformer architectures [15]. However, these approaches predominantly depend on onboard sensors, which limits their effectiveness in challenging conditions, such as occlusions and adverse weather, where accurate map generation becomes more complex.

2.2 Priors for HD Map Construction

Prior information plays a crucial role in providing effective initialization for map construction, thereby simplifying the model learning process. Liu *et al.* [16] introduced the concept of Mask-Active Instance, concentrating on learning map instance segmentation outcomes and providing semantic information for instance queries. Li *et al.* [17] employs 2D lane detection outcomes as priors to initialize the querying process. Both Luo *et al.* [18] and Jiang *et al.* [9] encode standard clarity maps (SD maps) into additional conditional branches and utilize mask autoencoders to capture the prior distribution associated with HD maps. Xiong *et al.* [7] introduced a globally self-updating neural map prior, significantly improving the performance of local map inference. Sun *et al.* [8] categorizes existing maps into three distinct types and enhances the matching algorithm of query-based map estimation models.

Despite these advancements, numerous studies have overlooked the potential of utilizing trajectory information to enhance online historical predictive maps as a valuable source of prior information. Trajectory data provides real-time insights into vehicle behavior and road usage patterns, making it a critical resource for improving map accuracy. Our work aims to fill this gap by harnessing the native representations of maps and incorporating trajectory data to improve the current perception analysis.

3 Methodology

3.1 Overview

The pipeline of the proposed method is shown as Fig. 1. Using N surrounding views images $\{I_i\}_{i=1}^N$ as input, the backbone EfficientNet [19] is followed to generate multi-scale image features F_{img} . In this study, we adopt a rasterized representation for online HD map generation, as it is more suitable for encoding inputs and outputs [9]. The output is M that presents the predicted label of each grid (divider, boundary, pedestrian crossing) in the Bird's Eye View (BEV) space. The pipeline can be formulated as:

$$M = f_{seg}(D_{cross}(f_{bev}(F_{img}), T)) \quad (1)$$

where T is the trajectory prior of traffic participants (eg. Car, pedestrian, bicycle, truck) around the ego vehicle obtained by the real-time detection and tracking algorithm. f_{bev} is a feature transformation module that convert surrounding image features into BEV features. D_{cross} Means the cross-attention layer to introduce prior features into BEV features. f_{seg} is the segmentation head.

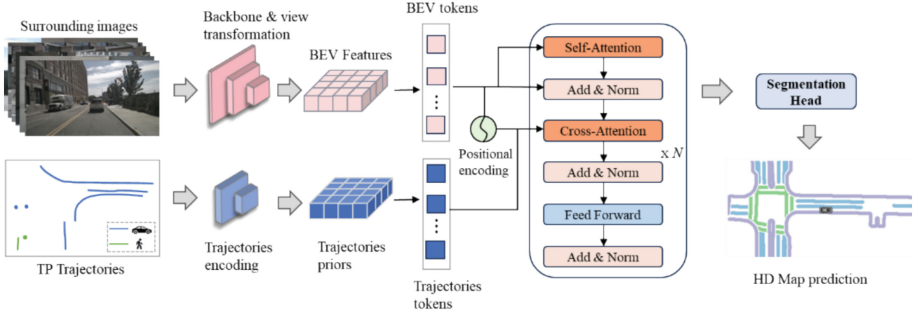


Fig. 1. The framework of the proposed method.

3.2 Trajectory Prior Generation

The detailed overview of the implementation of trajectory prior module. Considering the inherent challenges of onboard sensing, such as road marking invisibility due to far distance, unclear ground lane markings and so on, combining trajectory priors can provide implicit guidance and consistent information.

Firstly, we use existing detection method CenterPoint [20] and multiple objects tracking algorithms Poly-MOT [21] to process the data and save it as local files in advance. We specifically employ the nuScenes [22] dataset for our research. In order to simulate real-time tracking data, we classify these tracking trajectories by scene, arrange them in chronological order by timestamp. Then we transform these trajectory data into the current time's ego-vehicle coordinate system. Although we have obtained different types of trajectory priors, there may be errors in the trajectory due to possible detection and tracking errors. Therefore, we first limit the area and set the time window according to the current time to remove the out-of-area and fragmented trajectories to reduce possible trajectory errors, and select relevant trajectories according (2).

$$T_{selected} = \{T_i | Area(T_i, T_{ego}) \wedge (t_{current} - \Delta t \leq t_i \leq t_{current}) \wedge |T_i| > n\} \quad (2)$$

where $T_{selected}$ is the set of selected trajectories, each composed of multiple trajectory points, and T_i is the i -th trajectory, $|T_i|$ denotes the number of points in trajectory, ensuring that only trajectories with more than n points are included. $Area(T_i, T_{ego})$ Determines whether the trajectory T_i is within the effective range of the ego vehicle according the x and y axis in BEV space. $t_{current}$ Represents the current timestamp of ego vehicle, and Δt is the time window interval, specifying a duration within which trajectories are deemed relevant to the current time period.

3.3 Trajectory Features Embedded

After extracting trajectory information and rasterizing it, the rasterized trajectory priors inherently represent implicit road information, which may inevitably lead to spatial and quantitative mismatches with the actual map. To achieve a deeper and more effective fusion of BEV features with trajectory features, we employ the multi-head cross-attention module, which allows the model to utilize cross-attention to determine optimal alignment

positions, making effective use of trajectory priors to implicitly guide the generation of more effective BEV features.

Specifically, it begins by employing a convolutional network to reduce the resolution of the BEV features. Allowing to capture higher-level semantic information while reducing memory consumption. These features are then combined with sine position encodings and compressed into a linear format, producing the BEV query. The rasterized trajectories are processed through a convolutional network in conjunction with sine position embeddings, generating the trajectory prior tokens. Next, we utilize the multi-head cross-attention module to improve the BEV query features through incorporating information from the trajectory prior, thereby guiding the map generation process with relevant trajectory data. The process can be formulated as:

$$Q' = \text{Cat}(H_1(Q_{bev}, F_{traj}), \dots, H_k(Q_{bev}, F_{traj})) \quad (3)$$

$$F'_{BEV} = Q_{bev} + \text{Linear}(Q') \quad (4)$$

where the H denotes a single cross-attention head, and k indicates the number of heads. Linear acts as a projection layer, and F'_{BEV} features are updated in the multi-head cross-attention module that integrates the embedded trajectory prior. Following this, these enhanced BEV features are passed through a typical segmentation head to generate predictions for HD map elements.

4 Experiment

4.1 Dataset and Evaluation Protocol

In this paper, we evaluated the method using the widely used datasets, nuScenes [22]. The perception range is defined by the vehicle's moving direction following [9], measuring two ranges: 60×30 m, 120×60 m with a map resolution of 0.3 m. We employed Intersection over Union (IoU) and mean IOU (mIOU) as the evaluation metric for segmentation results. In line with [9], we set the IoU threshold at 0.2, and focused on three map elements including lane dividers, road boundaries and pedestrian crossings.

4.2 Experimental Parameters

We trained the method was trained on 4 NVIDIA GeForce RTX 3090 GPUs in batchsize 16. The learning rate is starting at 5×10^{-4} with a StepLR learning rate schedules, and the Adam optimizer is applied. For image features extraction, we utilized EfficientNet-B0 pretrained on ImageNet as the backbone. These were subsequently converted into BEV features through a LSS view transformation module.

4.3 Experimental Results

We present a comparative analysis of the performance metrics for two methods, baseline and the proposed method, across two perception ranges (60×30 m, 120×60 m) for

three elements: lane dividers, pedestrian crossings, and road boundaries. Our method consistently outperforms the baseline method HDMapNet [1] in all metrics, indicating its superior effectiveness in identifying and segmenting map elements. For lane dividers, our method shows a 5.5% improvement in the 60×30 m range and a 6.1% increase in the 120×60 m range. Notably, pedestrian crossings see a substantial enhancement of 3.8% in the smaller range 60×30 m and a remarkable 10.5% in the larger range 120×60 m. Road boundaries also benefit from the proposed method, with a 5.0% improvement in the smaller range and a 4.4% increase in the larger range. The mean Intersection over Union (mIoU) values indicate the overall improvements, with increases of 5.2%, 6.5% for the two ranges, respectively. Overall, these results highlight that we significantly improve the accuracy and coherence of segmented map elements compared to the baseline method (Table 1).

Table 1. Performance Evaluation on the nuScenes *val* set.

Range	Method	Divider	Ped Crossing	Boundary	mIoU
60 x 30 m	Baseline	39.8	18.6	40.4	32.9
	Ours	42.0 (+5.5%)	19.3 (+3.8%)	42.4 (+5.0%)	34.6 (+5.2%)
120 x 60 m	Baseline	41.1	23.7	40.9	35.2
	Ours	43.6 (+6.1%)	26.2 (+10.5%)	42.7 (+4.4%)	(+6.5%)

4.4 Visualization Results

We show the visualization results in Fig. 2 with two different scenes. This figure compares the proposed method and the baseline method against the ground truth for HD map elements, specifically lane dividers (green), road boundaries (blue), pedestrian crossing (orange). To demonstrate the impact of trajectories, we represent trajectories as black lines and overlay it on the ground truth map. On the left are images from six panoramic on-board cameras, and on the right is the map predicted from the BEV perspective. On the camera images, the red dotted boxes highlight areas in the vehicle camera images where visibility is compromised due to factors such as unclear road markings or far distance from the vehicle. On the right side, the red dotted boxes in the BEV perspective maps show the corresponding predictions.

The proposed method (top-right) predicts lane dividers and boundaries more accurately than baseline method, maintaining well-defined lines along the curve. It uses the consistency between vehicle trajectories and roads to provide useful guidance about map structures. Our approach demonstrates improved performance in these challenging conditions compared to the baseline, successfully capturing map details despite visibility issues. This highlights how implicit information in trajectories enhances the prediction performance of map construction.

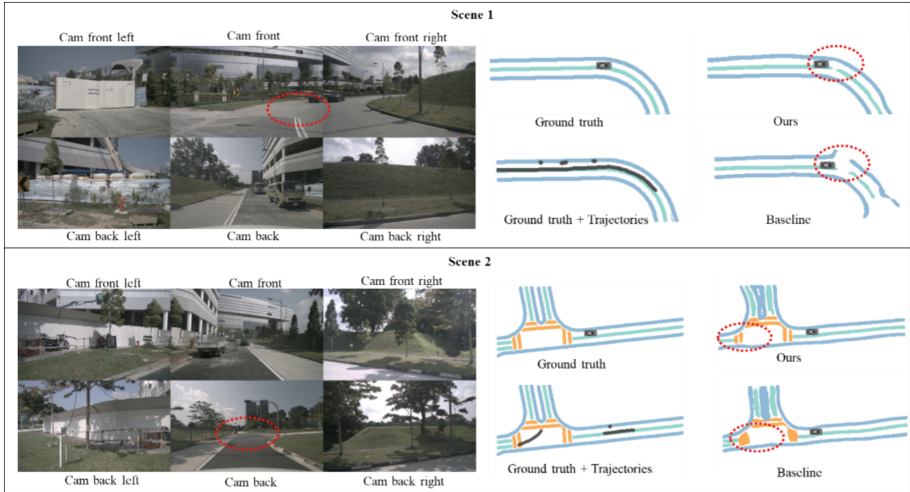


Fig. 2. Visualization Analysis. Blue lines indicate road boundaries, green lines are lane dividers, orange lines are pedestrian crossings, and black lines represent the trajectories.

5 Conclusion

In this paper, we improve the construction of online HD maps. The core of the proposed method is the encoding of trajectory data from traffic participants within the scene, utilizing the attention module to fuse trajectory features with map BEV features. This fusion allows us to obtain implicit information representing map constructs when the camera view is obstructed. The experiments indicate that our method can enhance the performance of online HD map construction, thereby ensuring the latest road information for automatic navigation.

Acknowledgments. This work was partially supported by the National Key R&D Program of China (2023YFC3805400), the National Science Foundation of China (52172376), and the Fundamental Research Funds for the Central Universities (KG16312001).

References

1. Li, Q., Wang, Y., Wang, Y., Zhao, H.: HDMaNet: an online HD map construction and evaluation framework. In: International Conference on Robotics and Automation, pp. 4628–4634. IEEE, USA (2022)
2. Liao, B., et al.: MapTR: structured modeling and learning for online vectorized HD map construction. In: International Conference on Learning Representations (2022)
3. Liang, M., et al.: Learning lane graph representations for motion forecasting. In: 16th European Conference on Computer Vision, pp. 541–556. Springer, Glasgow, UK (2020)
4. Zhou, Z., Ye, L., Wang, J., Wu, K., Lu, K.: HiVT: hierarchical vector transformer for multi-agent motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8823–8833. IEEE, USA (2022)

5. Hu, Y., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17853–17862. IEEE, Canada (2023)
6. Liu, Y., Wang, Y., Wang, Y., Zhao, H.: VectorMapNet: End-to-end vectorized HD map learning. arXiv preprint [arXiv:2206.08920](https://arxiv.org/abs/2206.08920) (2022)
7. Xiong, X., Liu, Y., Yuan, T., Wang, Y., Wang, Y.s, Zhao, H.: Neural map prior for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17535–17544. IEEE, Canada (2023)
8. Sun, R., Yang, L., Lingrand, D., Precioso, F.: Mind the map! Accounting for existing map information when estimating online HDMaps from sensor data. arXiv preprint [arXiv:2311.10517](https://arxiv.org/abs/2311.10517) (2023)
9. Jiang, Z., et al.: P-MapNet: Far-seeing Map Generator Enhanced by both SDMap and HDMap Priors. arXiv preprint [arXiv:2403.10521](https://arxiv.org/abs/2403.10521) (2024)
10. Jin, B., et al.: ADAPT: action-aware driving caption transformer. In: IEEE International Conference on Robotics and Automation, pp. 7554–7561. IEEE, UK (2023)
11. Houston, J., et al.: One thousand and one hours: self-driving motion prediction dataset. In: Conference on Robot Learning, pp. 409–418. USA (2023)
12. Liao, B., et al.: Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction. arXiv preprint [arXiv:2303.08815](https://arxiv.org/abs/2303.08815). (2023)
13. Yuan, T., Liu, Y., Wang, Y., Wang, Y., Zhao, H.: StreamMapMet: Streaming mapping network for vectorized online HD map construction. arXiv preprint [arXiv:2308.12570](https://arxiv.org/abs/2308.12570) (2023)
14. Zhang, G., et al.: Online map vectorization for autonomous driving: A rasterization perspective. arXiv preprint [arXiv:2306.10502](https://arxiv.org/abs/2306.10502). (2023)
15. Vaswani, A., et al.: Attention is all you need, Advances in neural information processing systems, 30, USA (2017)
16. Liu, X., Wang, S., Li, W., Yang, R., Chen, J., Zhu, J.: MGMap: mask-guided learning for online vectorized HD map construction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14812–1482. IEEE, USA (2024)
17. Li, H., Huang, Z., Wang, Z., Rong, W., Wang, N., Liu, S.: Enhancing 3D Lane Detection and Topology Reasoning with 2D Lane Priors. arXiv preprint [arXiv:2406.03105](https://arxiv.org/abs/2406.03105) (2024)
18. Luo, K., et al.: Augmenting Lane Perception and Topology Understanding with Standard Definition Navigation Maps. arXiv preprint [arXiv:2311.04079](https://arxiv.org/abs/2311.04079) (2023)
19. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. USA (2019)
20. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3D object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11784–11793. IEEE, USA (2021)
21. Li, X., et al.: Poly-MOT: a polyhedral framework for 3D multi-object tracking. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 9391–9398. IEEE, USA (2023)
22. Caesar, H., et al.: nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11621–11631. IEEE (2020)



Traffic Signal Timing Optimization Based on D3QLight

Haoyue Wang^(✉), Zundong Zhang, and Shilin Zhang

School of Electrical and Control Engineering, North China University of Technology,
Beijing 100144, China

22108030123@mail.ncut.edu.cn, {zdzhang, zhangshilin}@ncut.edu.cn

Abstract. Addressing the optimization of coordinated control of traffic signals at intersections within road traffic networks, this paper proposes a deep reinforcement learning algorithm based on D3QN to enhance the efficiency of intersection passage and alleviate traffic congestion. By delving into the underlying implicit features contained within traffic state information [1], the algorithm empowers intelligent intersection signal control units with the capability to autonomously learn based on traffic conditions. It utilizes a deep neural network to learn traffic state information and predict the impact of different control strategies on future traffic flows, thereby selecting the optimal signal control strategy. The D3QN algorithm estimates the value function, taking the current state as input and providing expected returns for each action. It selects actions to maximize long-term rewards, enhancing intersection signal efficiency and network optimization. Experiments show it can quickly stabilize and optimize intersection signal control, with good results in traffic networks.

Keywords: D3QLight Algorithm · Deep Reinforcement Learning · Urban Traffic Simulation · Traffic Signal Control

1 Introduction

With the surge in global population and the increasing demand for mobility, reducing traffic congestion through road expansion is a challenging task due to the limitations of urban space. Moreover, urban traffic congestion not only leads to low travel efficiency but also increases environmental pollution and energy consumption, and to some extent, increases the likelihood of traffic accidents. Deep Reinforcement Learning (DRL) is a rising research area in traffic signal control, optimizing signal timings by learning from dynamic traffic conditions to boost efficiency and responsiveness. In large networks, a multi-agent DRL system enhances robustness with distributed decision-making, enabling local traffic signal control. Agents observe, analyze, and adjust signals based on traffic data and feedback, offering real-time analysis and dynamic adjustments beyond traditional methods, and building an intelligent system for improved network coordination and efficiency.

In this paper, the feasibility and effectiveness of deep reinforcement learning in formulating traffic signal timing plans are studied. Li Li et al. [2] established a deep

neural network for Q-function learning in reinforcement learning without accounting for adjacent intersection signal timing coordination. Aslani et al. [3] used the Actor-Critic algorithm to develop an adaptive traffic signal control method, integrating policy and value functions to optimize decision-making and dynamically adjust traffic light timings to alleviate congestion based on real-time conditions. Xu et al. [4] proposed a DRQN-based deep reinforcement learning approach for traffic signal control at key road network nodes, leveraging LSTM to approximate value functions and capture traffic's temporal dynamics, selecting optimal actions greedily Zhang et al. [5] proposed a multi-agent reinforcement learning method for adaptive coordinated traffic signal control, utilizing shared parameter training and lead-follow phase sequence design to enhance main road signal coordination.

This paper tackles optimal signal timing by introducing a D3Qlight-based method for single intersection traffic control. It outlines mainstream deep learning methods in Sect. 2, details the D3Qlight algorithm in Sect. 3, and presents the signal timing plan design with experimental setup and simulation results in Sect. 4, demonstrating the method's effectiveness. Section 5 provides a summary.

2 Deep Reinforcement Learning Algorithms

A) Dueling DQN Algorithm

Reinforcement learning is a field of machine learning that emphasizes taking actions based on observations of the environment. Combining traffic signal control with reinforcement learning can effectively improve the efficiency of intersections 6. Dueling DQN is a variant of DQN that separates the value and advantage functions within a neural network to enhance the estimation of state-action values. It features two parallel streams: one for state values and another for action advantages, which share parameters but have separate output layers. At the same time, Dueling DQN divides the entire model structure into two parts, one being the state value function, which is only related to the state and independent of the specific actions to be taken, denoted as, The other part is the advantage function, which is related to both the state and the action, denoted as, The value function can be re-expressed as:

$$Q(s, a, w, \alpha, \beta) = V(s, w, \alpha) + A(s, a, w, \beta) \quad (1)$$

where w represents the network parameters that are common to both parts, α represents the network parameters unique to the value function, and β represents the network parameters unique to the advantage function.

The state value function estimates the value of a state, that is, the expected cumulative reward given a state:

$$V(s; \theta, \beta) = \sum_a \pi(a|s)(Q(s, a; \theta, \alpha) - A(s, a; \theta, \alpha)) \quad (2)$$

where $\pi(a|s)$ is the probability of choosing action a in state s , $Q(s, a; \theta, \alpha)$ is the action value output by the advantage network, and $A(s, a; \theta, \alpha)$ is the advantage of action a in state s .

The advantage function estimates the advantage of each action relative to the average action:

$$A(s, a; \theta, \alpha) = \text{ReLU}(Z(s; \theta, \alpha)) - \frac{1}{|A|} \sum_{a'} \text{ReLU}(Z(s; \theta, \alpha)) \quad (3)$$

where $\text{ReLU}(Z(s; \theta, \alpha))$ is the value output by the advantage network's output layer.

The combination formula for the value function gives us the action value, but this formula cannot discern the individual contributions of $V(s, w, \alpha)$ and $A(s, a, w, \beta)$ in the final output. To make this discernibility possible, the actual combination formula used is as follows:

$$Q(s, a, w, \alpha, \beta) = V(s, w, \alpha) + A(s, a, w, \beta) - \frac{1}{|A|} \sum_{a'} A(s, a'; \theta, \alpha) \quad (4)$$

Figure 1's design enhances Dueling DQN's ability to capture state values and action advantages in complex settings, improving learning efficiency.

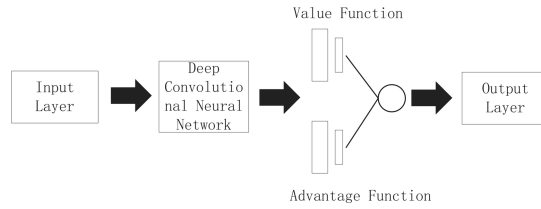


Fig. 1. Flowchart of the Dueling DQN Algorithm

B) DQN algorithm

The DQN algorithm uses experience replay technology to train a neural network to approximate the action-value function⁷. The DQN algorithm, depicted in Fig. 2, inputs the state and outputs Q-values for all actions, choosing the action with the highest Q-value. It uses a deep convolutional neural network to approximate the value function and experience replay for training. The core of DQN is to use deep neural networks' non-linear fitting to manage high-dimensional state spaces.

- 1) The action-value function of DQN is approximated using a neural network, which is a form of non-linear approximation. The network structure used by DQN consists of three convolutional layers followed by two fully connected layers. In formulaic terms, the value function is denoted as $Q(s, a; \theta)$. At this point, updating the network is essentially updating the parameters θ . Once θ is determined, the network parameters are set.
- 2) The main feature of DQN is the introduction of experience replay, which involves adding a quintuple $(s_t, a_t, r_t, s_{t+1}, done)$ to an experience pool. These quintuples are later used to update the Q-network parameters. In this context, both s_t and s_{t+1} are in vector form, the action and reward are scalars, $done = False$.

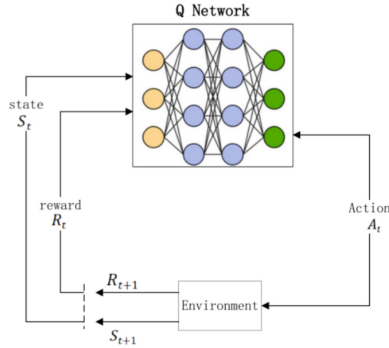


Fig. 2. Flowchart of the DQN Algorithm

3 D3Qlight

The D3Qlight algorithm combines the concepts of Dueling DQN and Double DQN, aiming to enhance performance and stability. Dueling DQN separates the state value function and the advantage function to better estimate the value of state-action pairs, while Double DQN uses two networks to mitigate the issue of overestimating Q-values. The D3Qlight framework integrates these two approaches to improve the performance and stability of Q-learning algorithms. The reinforcement learning control framework based on the D3Qlight algorithm is shown in Fig. 3.

D3Qlight Algorithm Framework:

- 1) Network Architecture: Like Dueling DQN, D3Qlight has a state value network and an advantage network, sharing parameters but with separate output layers. The state value network estimates state value, while the advantage network estimates the relative advantage of each action.
- 2) Target Value Update Strategy: D3Qlight uses a Double DQN-like strategy, selecting actions from the main network and calculating target Q-values from the target network to reduce Q-value overestimation.
- 3) Training Process: The D3Qlight training process involves experience replay for storing and sampling agent-environment interactions and target network updates to stabilize learning.

In D3Qlight, the target value update formula combines elements of Dueling DQN and Double DQN, using one network for action selection and another for estimating target Q-values. It follows Dueling DQN's approach for value and advantage functions, while adopting Double DQN's strategy of selecting actions from the main network and calculating target Q-values from the target network, ensuring Q-network training stability. Overall, D3Qlight combines the architecture of Dueling DQN with the target value update strategy of Double DQN, separating the state value function and advantage function, and using two networks to mitigate the overestimation of Q-values, thereby enhancing the performance and stability of the Q-learning algorithm. The network architecture diagram is the same as Dueling DQN, and the algorithm flowchart is the same as DDQN.

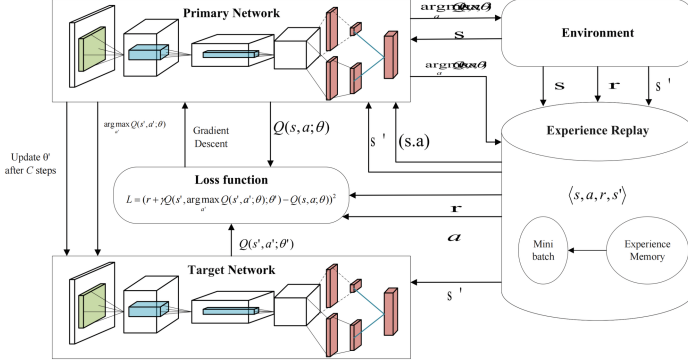


Fig. 3. Schematic Diagram of the Principle of D3QLight-based Reinforcement Learning Algorithm

D3QLight's execution process can be summarized as follows:

D3QLight algorithm

1. Define configurable parameters: total training session N , maximum number of training steps per session n ;
2. Initializing neural network parameters: Evaluation network `eval_net` parameters θ ;
3. Target network `target_net` parameters $\theta' \leftarrow \theta$; Setting the greediness coefficient ϵ ;
4. Setting the learning rate α .
5. for episodes from 0 to N do
6. Initialize the road network environment and load the traffic flow data;
7. for t from 1 to n do
8. Sampling states s_t from the microscopic traffic simulation platform SUMO and using `eval_net` to map states to Q-value estimates for all actions;
9. Based on the greedy policy ϵ and the currently selected action a_t
10. Execute the action a_t and obtain the reward r_t and the next state s_{t+1}
11. Store the experience $\{s_t, a_t, r_t, s_{t+1}\}$ into the replay buffer
12. Randomly sample B $\{s_t, a_t, r_t, s_{t+1}\}$ from the replay buffer
13. The neural network was trained using the mini batch mode
14. Calculate the loss function L and use gradient descent to update the parameters θ of the `eval_net`
15. If the number of experiences in the replay buffer $\%T = 0$, then
16. update the `target_net` parameters .
17. end for
18. end for

4 Experiments and Results Analysis

We designed 5 different traffic phases, each with different green light duration configurations. The experiment is divided into a training phase and a testing phase. During the training phase, the agent learns the optimal strategy by interacting with the environment; during the testing phase, the performance of the learned strategy is evaluated.

A. SUMO Design

This study created a 12-lane, four-way intersection simulation in SUMO, as shown in Fig. 4. Each approach has 3 lanes (left-turn, through, right-turn). The intersection operates in 4 phases (E-W through, E-W left, S-N through, S-N left), with right turns unregulated. Phases switch with a 3-s yellow light.

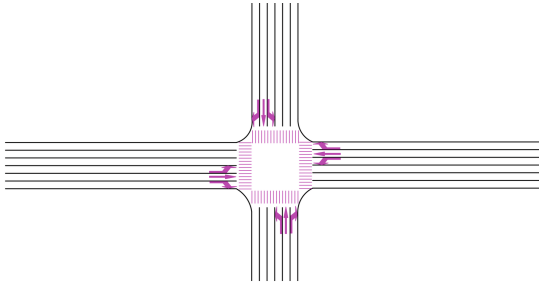


Fig. 4. Intersection Schematic Diagram

B. Basis for Signal Timing Design

This study used SUMO simulation and optimization models to test signal timing plans, focusing on vehicle queue length and delay time as efficiency indicators. Real-time data on these metrics from SUMO informed the optimization of signal timing.

C. Agent Design

• State Representation Design

In this study, the state representation design of the agent is aimed at capturing the key characteristics of the traffic signal control task to enable the agent to make accurate decisions. The state representation integrates multiple dimensions of traffic parameters. The agent, positioned at the central intersection, has a state defined by “occupancy rate at the approaches of the central intersection,” “vehicle speed at the approaches of the central intersection,” “traffic flow at the central intersection,” and “congestion length at the central intersection.” This comprehensive data allows the agent to make accurate decisions, monitor traffic, predict congestion, and optimize signal timings to minimize delays and enhance efficiency in real-time.

• Action Selection Design

The definition and selection of actions directly affect the capacity of the intersection. The action selection design evaluates strategy robustness and effectiveness through simulation and testing. It dynamically adjusts green light durations in various traffic conditions, considering the optimal range for traffic phases to avoid unnecessary oscillations or control latency. The agent uses an epsilon-greedy strategy to balance exploration and exploitation for effective traffic signal control decision-making. When initializing the D3QLight network, the agent’s predicted actions are random, a choice strategy known as exploration. As the

convolutional neural network converges, more accurate Q-values for each action can be obtained, and the amount of exploration should gradually be reduced, prioritizing the selection of actions with larger Q-values, this choice strategy is known as exploitation⁹. This strategy balances exploring optimal and new actions to maximize rewards and discover better signal timing plans. The design includes north-south phases for straight and left turns, and east-west phases for four scenarios (east-west straight, east-west left, east left, and west left turns), resulting in three distinct schemes.

- **Reward Function Design**

In this study, we adopted a reward function based on time loss to evaluate the agent's performance in the traffic signal control task. Time loss refers to the time lost by vehicles at intersections due to waiting for traffic lights, which is a key indicator of traffic flow efficiency. Our goal is to design a reward mechanism that encourages the agent to reduce time loss for vehicles, thereby improving the efficiency of intersection passage. Specifically, we defined the reward function R to calculate the agent's performance at each time step, as shown below:

$$R = \begin{cases} 0, & \text{if TimeLoss} = -1 \text{ or TimeLoss} = 0 \\ \frac{1}{\text{TimeLoss}}, & \text{otherwise} \end{cases} \quad (5)$$

In this study, TimeLoss is measured by the E3 detector to assess intersection delays, which are integrated into the reward mechanism. The reward function, using parameter MID_e3_det, calculates rewards based on the inverse of time loss, incentivizing smoother traffic flow. The agent learns to optimize traffic by adjusting signal timings to maximize rewards through continuous learning and error correction. Real-time time loss data is obtained from the E3 detector via SUMO's API for experimental analysis.

D. Results Analysis

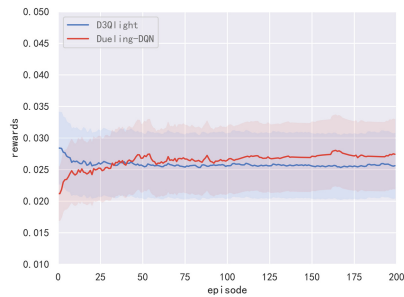


Fig. 5. Reward Comparison Diagram

Figure 5 shows D3QLight outperforms others in cumulative rewards, adapting quickly to traffic for more rewards. It leverages structural advantages by separating state and

advantage functions and using Double DQN's update strategy for better learning efficiency and stability. Fig. 5 highlights the D3Qlight algorithm's superior training efficiency, showing quick convergence to optimal rewards and sustained stability. It also shows robust performance in traffic signal timing, with rapid adaptation to traffic changes and steady reward growth, underscoring its effectiveness for real-time, adaptive traffic management in complex scenarios.

5 Conclusion

This study introducing a novel D3Qlight-based deep reinforcement learning algorithm for optimizing coordinated intersection signal control. The algorithm endows intelligent intersection signal control units with the ability to learn autonomously and make dynamic adjustments based on real-time traffic conditions. Through deep neural networks, the intelligent unit learns and understands complex traffic state information, predicts the potential impact of various control strategies on future traffic flows, and selects the optimal signal control strategy accordingly. Experiments indicate it effectively addresses intersection signal control optimization and yields positive outcomes in traffic networks.

Acknowledgments. This work was supported by the Yuxiu Innovation Project of NCUT under grant No.P2024NCUTYXCX108.

References

1. Baolin, Y., Dong, C., Chunyuan, L., et al.: Traffic signal control method based on dueling double DQN. *Comput. Measur. Control* **32**(07), 154–161 (2024). <https://doi.org/10.16526/j.cnki.11-4762/tp.2024.07.023>. (in Chinese)
2. Li, L., Lv, Y., Wang, F.: Traffic signal timing via deep reinforcement learning. *IEEE/CAA J. Automatica Sinica* **3**(03), 247–254 (2016)
3. Aslani, M., Mesgari, M.S., Wiering, M.: Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events. *Transp. Res. Part C: Emerg. Technol.* **85**, 732–752 (2017)
4. Xu, M., Wu, J., Huang, L., Zhou, R., Wang, T., Hu, D.: Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning. *J. Intell. Transp. Syst.* **24**(1), 1–10 (2020)
5. Zhang, W., Yan, C., Li, X., et al.: Distributed signal control of arterial corridors using multi-agent deep reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **24**(1), 178–190 (2022)
6. Zhimin, L., Baolin, Y., Yaodong, Z., et al.: Traffic signal control method based on deep reinforcement learning. *J. Zhejiang Univ. (Engineering Science)* **56**(6), 1249–1256 (2022). (in Chinese)
7. Zhai, P., Zhang, Y., Shaobo, W.: Intelligent ship collision avoidance algorithm based on DDQN with prioritized experience replay under COLREGs. *J. Mar. Sci. Eng.* **10**(5), 585 (2022)
8. Xiqun, C., Yizhang, Z., Chaofeng, L.: Intersection signal phase and timing optimization method based on mixed proximal policy optimization. *Syst. Eng. Inf. Sci.* **23**(1), 106–113 (2023). (in Chinese)
9. Jianhui, L.: Traffic signal control strategy based on D3QN. *Comput. Sci.* **46**(S2), 117–121 (2019). (in Chinese)

10. Kewen, X.: Research on the optimization design of signal timing at road horizontal intersections. *Eng. Constr.* **36**(03), 644–646+649 (2022). (in Chinese)
11. Guojia, Z., Shifeng, S.: Optimization of traffic light control based on D3QN. *Comput. Modernization* (07), 30–35+47 (2023). (in Chinese)



Research on the Transportation Organization Mode of Low-Speed Maglev System

Zinuo Li¹ , Jun Liu^{1,2} , Wu Bai¹ , Yu Liu^{3,4}, and Qingying Lai¹

¹ School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
laqingying@bjtu.edu.cn

² North China University of Technology, Beijing 100144, China

³ CRRC TANGSHAN Co., Ltd., Tangshan 063035, China

⁴ CRRC Technology Innovation (Beijing) Co., Ltd., 15Th Research Institute, Beijing 10000, China

Abstract. Due to the increasing demand for passenger and freight transportation at the end of backbone lines, existing transportation modes cannot fully alleviate these transportation pressures. The low-speed maglev system can provide both passenger and freight services simultaneously. The system adopts unmanned driving and automatic loading and unloading technology, and can serve as an innovative, intelligent, and autonomous rail transit mode connected to the end of the backbone line. This article takes typical cities in China as an example to predict and analyze the demand for passenger and freight transportation at the end of their backbone lines. Next, define the concept of transportation organization mode for low-speed maglev system and divide it into four categories: single passenger transportation, single freight transportation, mixed passenger and freight line, and mixed passenger and freight formation. Finally, a comprehensive evaluation method for the transportation organization mode of low-speed maglev system based on FCE-AHP model was constructed. After calculation, the comprehensive scoring results of the transportation organization mode of LSM system were obtained as follows: mixed passenger and freight line, mixed passenger and freight formation, single passenger transportation, and single freight transportation. The research results of this article can provide reference for the future design, planning, and operation of low-speed maglev system.

Keywords: Backbone Line End Access · Passenger and Freight Demand · Low-speed Maglev System · Intelligent Autonomy · Prediction and Analysis · Transportation Organization Mode · Comprehensive Evaluation Model

1 Introduction

With the acceleration of urbanization, road traffic congestion is becoming increasingly severe, and the development of urban rail transit systems has effectively alleviated this problem. In August 2022, the Ministry of Housing and Urban Rural Development and the National Development and Reform Commission issued the “14th Five Year Plan for National Urban Infrastructure Construction”, which clearly proposed to promote

the construction of urban rail transit by classification, and to promote the planning and construction of low volume rail transit systems in eligible Type II large cities based on urban transportation needs [1]. Meanwhile, due to the structural transformation and adjustment of the Chinese economy, the express logistics industry has developed rapidly. In 2023, the volume of China's express delivery business reached 132.07 billion pieces, and the revenue from express delivery business reached 1207.40 billion yuan [2].

In this context, LSM(LSM) system can be designed as a new type of rail transit system that provides both passenger and freight services, demonstrating potential application value when connected at the end of backbone lines. The LSM system is shown in Fig. 1. The system adopts the form of a "rail holding car" embedded in the track beam of the vehicle running gear, and adopts unmanned driving and automatic loading and unloading technology. It has significant advantages such as small footprint, small turning radius, strong climbing ability, intelligent autonomy, and wide applicability scenarios [3]. In addition, the system can achieve zero power suspension, low-power operation, without generating exhaust emissions and electromagnetic radiation, and has low construction and operation costs [4].

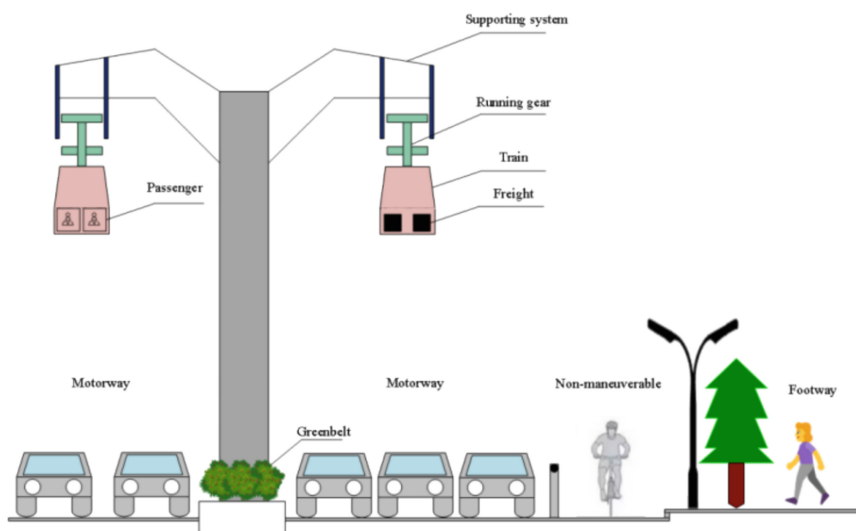


Fig. 1. Schematic diagram of LSM system.

With the continuous development of maglev technology, domestic and foreign experts' research on maglev mainly focuses on key technologies, operation control, and optimization of operation diagrams. Lai et al. [7] established an integrated energy-saving optimization model for the control of medium speed maglev trains by integrating the power supply system and train control system. Lai et al. [8] simultaneously considered traction energy consumption and suspension energy consumption, and established an optimization model for the energy-saving timetable of medium speed maglev. At present, research on the organization of passenger and freight transportation of wheel

rail systems mainly focuses on high-speed railways and subways. The freight transportation modes of China's high-speed rail include four modes: dynamic inspection confirmation car, freight special train, passenger multiple unit train carrying, and passenger multiple unit reserved carriage [9]. Li et al. [10] proposed an operational strategy for using subways for freight transportation and constructed an optimization model for the design of combined train services to maximize the revenue generated by freight services and the profit generated by balancing costs.

This article takes typical cities in China as an example to predict the demand for passenger and freight transportation connected at the end of backbone lines, and analyze their development trends. Then define and propose the concept and types of transportation organization modes for LSM systems. Finally, a comprehensive evaluation method for the transportation organization mode of LSM system based on FCE-AHP model is constructed.

2 Prediction and Analysis of Passenger and Freight Demand for LSM System

This article takes typical cities in China as an example and uses trend extrapolation to predict and analyze the demand for passenger and freight transportation at the end of backbone lines. The specific prediction function is shown in Table 1.

Table 1. Passenger and freight demand prediction function.

Prediction type	Prediction object	Prediction function
Passenger transportation (Taking the daily average passenger volume of the subway as an example)	Shenzhen	$y = 19.15x^2 - 77.372x + 78.133835$
	Chengdu	$y = 6.78x^2 - 27.356x + 27.591093$
	Wuhan	$y = 13.58x^2 - 54.883x + 55.439853$
	Chongqing	$y = 6.5x^2 - 26.258x + 26.510201$
Freight transport (Taking Jinhua City as an example)	Annual freight volume of railway	$y = 27662.07x^2 - 111.202710x + 111.759502646$
	Annual air cargo volume	$y = 103.68x^2 - 416.801x + 418887.294$
	Annual business volume of express delivery	$y = 0.95x^2 - 3.806x + 3.824702$
	Number of express delivery service vehicles	$y = 1615.66 \ln(x - 0.3) + 3690.59$

The passenger and freight volume connected at the end of the backbone line and its predicted trend are shown in Fig. 2. From Fig. 2 (a), it can be seen that the subway passenger volume in these cities has experienced varying degrees of growth in the past few years and is expected to continue to increase in the coming years. From Fig. 2 (b), it can be seen that the freight volume of railways and aviation in Jinhua City has been increasing year by year.

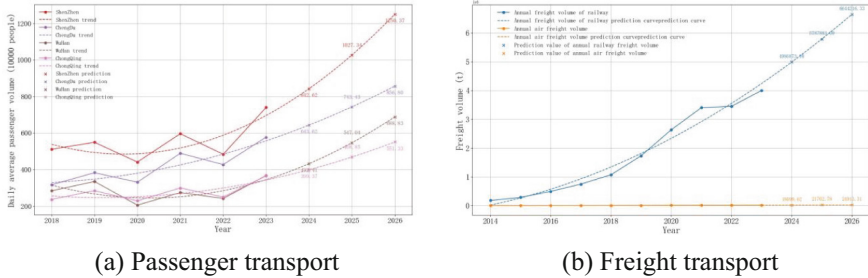


Fig. 2. Passenger and freight volumes connected to the end of backbone lines and their predicted trends.

3 Comprehensive Evaluation of Transportation Organization Mode for LSM System

Based on the current situation and predictive analysis of passenger and freight transportation demand at the end of the backbone line, this section combines the characteristics of the LSM system in terms of “passenger and cargo capacity, flexible grouping, unmanned driving, and automatic loading and unloading”. Firstly, the transportation organization mode of the system is proposed, and then a comprehensive evaluation method for the transportation organization mode of the LSM system based on the FCE-AHP model is constructed.

3.1 Definition of Transportation Organization Mode for LSM System

This article defines the transportation organization mode of the LSM system as: based on the characteristics of the LSM system in terms of lines, stations, vehicles, traction power supply, operation control, transportation organization, etc., the fixed facilities and mobile equipment of the system are reasonably arranged in space to meet the passenger and freight transportation needs of the bone trunk line end access.

Based on the characteristics of the LSM system in terms of “passenger and cargo capacity, flexible grouping, unmanned driving, and automatic loading and unloading”, four transportation organization modes suitable for the system are proposed: single passenger transportation, single freight transportation, mixed passenger and freight line, and mixed passenger and freight formation.

3.2 Evaluation of Transportation Organization Mode for LSM System

This article combines the development trend of passenger and freight transportation demand at the end of backbone lines and the characteristics of LSM systems to establish an evaluation model for the transportation organization mode of the system.

Determine Evaluation Indicators

This article divides the evaluation indicators for the transportation organization mode of low-cost small capacity maglev systems into six categories: economic cost, market capacity, system capacity, operational organization, service level, and environmental impact. It includes 6 primary indicators and 12 secondary indicators, as shown in Table 2.

Table 2. Evaluation index system for transportation organization mode of LSM system.

Evaluate the overall objective	Symbol	First level indicators	Symbol	Second level indicators	Symbol
Transportation organization mode comprehensive evaluation index system	A	Economic cost	B_1	Construction cost	B_{11}
				Operating cost	B_{12}
		Market traffic volume	B_2	Passenger volume	B_{21}
				Freight volume	B_{22}
		System capacity	B_3	Passenger turnover	B_{31}
				Freight turnover	B_{32}
		Operating Organization	B_4	Operability	B_{41}
				Flexibility	B_{42}
		Service Level	B_5	Ticket price	B_{51}
				Full load rate	B_{52}
		Environmental impact	B_6	Energy consumption	B_{61}
				Noise	B_{62}

Building an Evaluation Model

This article uses the Fuzzy Comprehensive Evaluation (FCE) method to evaluate the transportation organization mode of LSM systems, and uses the Analytic Hierarchy Process (AHP) to determine the weight of evaluation indicators in the FCE method. The specific steps of the FCE-AHP model are shown in Fig. 3.

Weight Calculation of Evaluation Indicators

Establish a weight judgment matrix for primary and secondary indicators in sequence,

Table 4. Weight and compatibility analysis results of secondary indicators.

B_{ij}	B_{i1}	B_{i2}	Eigenvector	Root Squared	Weight	Characteristic value	$\lambda_{\max}$	CI	CR
B_{1j}	0.65	0.69	1.90	0.63	1.94	3.07	3.04	0.02	0.04
	0.22	0.23	0.78	0.26	0.79	3.03			
B_{2j}	0.75	0.75	1.50	0.75	1.50	1.99	2.00	0.00	0.00
	0.25	0.25	0.50	0.25	0.50	1.99			
B_{3j}	0.75	0.75	1.50	0.75	1.50	1.99	2.00	0.00	0.00
	0.25	0.25	0.50	0.25	0.50	1.99			
B_{4j}	0.80	0.69	2.33	0.78	2.48	3.19	3.08	0.04	0.08
	0.09	0.08	0.21	0.07	0.21	3.01			
B_{5j}	0.56	0.65	2.67	0.53	2.87	5.38	5.15	0.04	0.03
	0.11	0.13	0.87	0.17	0.88	5.08			
B_{6j}	0.55	0.60	1.65	0.55	1.66	3.03	3.02	0.01	0.02
	0.18	0.20	0.63	0.21	0.63	3.01			

Evaluation and Scoring Results

The comprehensive weights of the primary and secondary indicators calculated from Tables 3 and 4, using the expert scoring method, are shown in Table 5.

Evaluate Results

The evaluation level of the transportation organization mode of the LSM system is assigned using a percentage system, with excellent being 100 points, good being 80 points, medium being 60 points, general being 40 points, and difference being 20 points, which is $V = (100, 80, 60, 40, 20)^T$. Based on the calculation method in Fig. 3 and the expert scoring results in Table 5, the comprehensive scores for single passenger transportation, single freight transportation, mixed passenger and freight line, and mixed passenger and freight formation are calculated as $D_1 = 71.8$, $D_2 = 65.6$, $D_3 = 78.4$, $D_4 = 72.8$, respectively.

Table 5. Calculation result of comprehensive weight of evaluation indicators.

B_i (w_i)	B_{ij} (w_{ij})	Expert scoring					W_{ij}
		Excellent	Good	Middle	Generally	Difference	
B_1 (0.44)	B_{11} (0.63)	10/10/10/10	0/0/0/0	0/0/0/0	0/0/0/0	0/0/0/0	0.282
	B_{12} (0.26)	10/10/10/10	0/0/0/0	0/0/0/0	0/0/0/0	0/0/0/0	0.116
B_2 (0.20)	B_{21} (0.75)	10/0/8/7	0/0/2/3	0/0/0/0	0/0/0/0	0/10/0/0	0.149
	B_{22} (0.25)	0/10/7/6	0/0/2/2	0/0/1/1	0/0/0/1	10/0/0/0	0.050
B_3 (0.18)	B_{31} (0.75)	10/0/8/7	0/0/2/2	0/0/0/1	0/0/0/0	0/10/0/0	0.133
	B_{32} (0.25)	0/10/7/6	0/0/2/2	0/0/1/1	0/0/0/1	10/0/0/0	0.044
B_4 (0.09)	B_{41} (0.78)	6/8/2/2	2/1/2/3	1/1/6/5	0/0/0/1	1/0/0/0	0.071
	B_{42} (0.07)	9/3/6/8	1/3/2/2	0/2/2/0	0/1/0/0	0/1/0/0	0.006
B_5 (0.05)	B_{51} (0.53)	8/8/6/5	2/2/2/2	0/0/2/2	0/0/0/1	0/0/0/0	0.027
	B_{52} (0.17)	6/8/7/7	2/1/2/1	1/1/1/1	1/0/0/1	0/0/0/0	0.009
B_6 (0.04)	B_{61} (0.55)	4/6/6/4	5/2/2/3	0/1/2/2	0/1/0/1	1/0/0/0	0.020
	B_{62} (0.21)	3/4/5/5	3/3/3/3	2/2/2/2	1/1/0/0	1/0/0/0	0.008

Note: “* /* /* /*” refers to the expert’s separate ratings for the transportation organization mode of LSM systems, including “single passenger transportation/single freight transportation/mixed passenger and freight line/mixed passenger and freight formation”

4 Conclusion

This article first predicts and analyzes the demand for passenger and freight transportation of LSM systems connected at the end of backbone lines and their future development trends. Then, the concept and types of transportation organization modes for LSM systems are defined and proposed. Next, an evaluation method for the transportation organization mode of LSM systems based on the FCE-AHP model is constructed. Finally, the comprehensive evaluation results of the transportation organization mode of LSM systems considering different aspects are as follows: mixed passenger and freight line, mixed passenger and freight formation, single passenger transportation, and single freight transportation mode. In the future, we will collect more expert ratings on the transportation organization mode of LSM systems, analyze the impact of different indicator weight changes on the evaluation results of transportation organization modes, and further improve the accuracy of the evaluation model for the transportation organization mode of LSM systems.

Acknowledgements. The authors gratefully acknowledge the financial support from the National Key R&D Program of China (Grant No. 2023YFB4302104-1).

References

1. National Development and Reform Commission. “14th Five-Year Plan” for National Urban Infrastructure Construction. (2022-07-07) [2024-10-16]. https://www.gov.cn/zhengce/zhengceku/2022-07/31/content_5703690.htm. (in Chinese)
2. State Post Bureau of the People’s Republic of China. 2023 Statistical Bulletin on the Development of the Postal Industry. (2024-05-10) [2024-10-16]. <https://www.spb.gov.cn/gjyzj/c100276/202405/ff1ab12da9d74425b7ddef9e38de8916.shtml>. (in Chinese)
3. Hu, B., Pan, Y., Lu, Z., et al.: Construction practice of china’s first suspended monorail test line. *Urban Rapid Rail Transit* **32**(01), 38–42 (2019). (in Chinese)
4. Li, X., Li, J.: Overall design of Xingguo test line for suspended permanent magnet levitation monorail transit system. *Railway Invest. Surveying* **49**(05), 114–118 (2023). (in Chinese)
5. Li, Y., Yan, Y., Cheng, H.: Design and analysis of levitation magnets for hybrid EMS Maglev trains. *J. Nat. Univ. Defense Technol.* **05**, 94–98 (2006). (in Chinese)
6. Gao, Y., Gu, B.: Statistical analysis of power consumption of urban rail transit networks in china - the third research report based on data analysis of china association of metro. *Urban Mass Transit* **22**(08), 1–7 (2019). (in Chinese)
7. Lai, Q., Liu, J., Wang, Y., et al.: Energy-efficient operation of medium-speed maglev through integrated traction and train control. *IET Intel. Transport Syst.* **18**(2), 409–431 (2024)
8. Lai, Q., Liu, J., Guo, S., et al.: Energy-efficient train timetabling for a medium-speed maglev line considering propulsion and suspension energy consumption. *IET Intel. Transport Syst.* **17**(9), 1860–1878 (2023)
9. Zhen, L., Gao, J., Zhang, N., et al.: A decision model for high-speed railway express systems. *Comput. Ind. Eng.* **177**, 1–14 (2023)
10. Li, Z., Shalaby, A., Roorda, M.J., et al.: Urban rail service design for collaborative passenger and freight transport. *Transp. Res. Part E: Logistics Transp. Rev.* **147**, 1–23 (2021)



Downstream Intersection Entrance Flow Arrival Rate Estimation Using Sampled Trajectory Data

Yingjie Bai¹, Shuyan Jiang¹, Ke Lin³, Zhijie Zhou¹, and Kai Lu^{1,2}(✉)

¹ School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510640, China

kailu@scut.edu.cn

² State Key Laboratory of Subtropical Building and Urban Science, Guangzhou 510640, China

³ Guangdong Zhenye UCTRL Technology CORP., Zhongshan 528405, China

Abstract. This paper classified and defined traffic flows from various directions to achieve a refined estimation of vehicle inflow rates at the downstream intersections. Using sampled trajectory data from these flows, the paper analyzed vehicle arrivals, selected an exponential distribution as the prior distribution, and constructed the likelihood function for vehicle arrivals. The Maximum A Posteriori (MAP) estimation method was then applied to estimate the traffic arrival rate. Results have indicated that the Mean Absolute Percentage Error (MAPE) for arrival rate estimations from the proposed model has fluctuated around 13% under different penetration rates of connected vehicles, which is significantly lower than the average arrival rate estimation error. This has demonstrated the model's strong estimation performance. The proposed model has effectively adapted to varying penetration rate scenarios, achieving precise arrival rates and traffic volume estimates for traffic flows from different directions.

Keywords: Traffic Flow Analysis · Trajectory Data · Arrival Rate · Maximum A Posteriori · Likelihood Function · Penetration Rate

1 Introduction

In urban traffic research, traffic volume is commonly used as an input-output variable to study traffic control, traffic conditions, and other areas. However, issues like missing or biased traffic data often occur in practical applications. Therefore, estimating vehicle arrival rates with sampled data has drawn significant attention from scholars.

Traditional methods for obtaining traffic volume primarily involve traffic surveys [1] and detector measurements [2–4]. For instance, Li et al. [5] utilized the ST-DBSCAN method to process trajectory feature data obtained from surveys, analyzing the spatiotemporal distribution of traffic volume. Wu [6] integrated audio and video data based on the characteristic sequence correlation between them, enhancing the accuracy of traffic volume detection. An et al. [7] estimated the cyclic traffic arrival rate at signalized intersections by acquiring license plate data through recognition techniques. However, current traffic surveys require substantial human resources and time, and the accuracy of

the data cannot be effectively guaranteed. Additionally, intelligent detectors' installation and maintenance costs are high, and they are susceptible to environmental influences, often leading to issues such as data loss or bias. Data completion is still necessary.

In response to this, some scholars have integrated mathematical models [8, 9] and modern information technology [10, 11] to conduct data mining and short-term vehicle arrival prediction, thereby mitigating the impact of missing data. Sun [12] has constructed an arrival rate estimation model based on Gaussian processes and two-dimensional Gaussian mixture models, taking into account the relationship between shock wave models and vehicle trajectories. Kwon et al. [13] have employed ASTGCN (Attention-Based Spatial-Temporal Graph Convolutional Networks) to integrate the spatiotemporal characteristics of detection data for traffic volume completion. Chen et al. [14] have improved the Back Propagation Neural Network by combining the position update methods of SSA (Sparrow Search Algorithm) and CSO (Chicken Swarm Optimization), achieving a fitting prediction of traffic volume. However, much of the current research focuses on vehicle arrivals on a single lane, often neglecting the arrival conditions of traffic flows from different upstream directions. Moreover, modern information technology has prediction errors, which often leads to discontinuous vehicle trajectories within the road network, resulting in significant deviations from reality.

With the improvement of vehicle networking degree, the acquisition of vehicle trajectory data becomes more accurate and convenient. How to use vehicle trajectory data to estimate vehicle arrival rates has become a research hotspot for scholars. Therefore, this study will take sampled vehicle trajectory data as the basis. Through Bayesian inference and using the Maximum A Posteriori estimation method, it will refine the division of traffic flow and finally achieve the estimation of the arrival rate of inbound traffic flow at the downstream intersection.

2 Definition of Traffic Flow in Different Directions

To achieve more accurate traffic volume estimation for a specific flow direction leaving the downstream intersection I_j , the incoming traffic flows at the entrance of this direction are classified and defined. The incoming traffic flows at the intersection I_j are defined as $f_{S(i \rightarrow j)}$, with their arrival rates denoted as $\lambda_{S(i \rightarrow j)}$, where S takes values of 1, 2, 3, HN, representing left-turning traffic $f_{1(i \rightarrow j)}$, through traffic $f_{2(i \rightarrow j)}$, right-turning traffic $f_{3(i \rightarrow j)}$, and merging traffic $f_{HN(i \rightarrow j)}$ respectively. Explanations are as follows:

Based on the traffic flow entering a specific flow direction at the entrance of the downstream intersection I_j from different incoming directions (left turn, through, right turn) at the upstream intersection I_i , the traffic flows $f_{1(i \rightarrow j)}$, $f_{2(i \rightarrow j)}$ and $f_{3(i \rightarrow j)}$ are determined. Additionally, considering the signal phase settings within the cycle at the intersection, the situation can be divided into two cases: whether there is a mixture of traffic flows from different directions entering intersection I_j . As shown in Fig. 1, if there is no mixture, the vehicle arrival conditions for the traffic flow $f_{1(i \rightarrow j)}$, $f_{2(i \rightarrow j)}$ and $f_{3(i \rightarrow j)}$ are estimated separately. Otherwise, if two or more of the traffic flows $f_{1(i \rightarrow j)}$, $f_{2(i \rightarrow j)}$, or $f_{3(i \rightarrow j)}$ are released during the same time period, they are combined into a merged traffic flow $f_{HN(i \rightarrow j)}$, where N takes values of 1, 2, ..., q , and the value of q depends on the number of time periods with mixed entry.

Taking the southbound entry merging into the traffic flow as an example

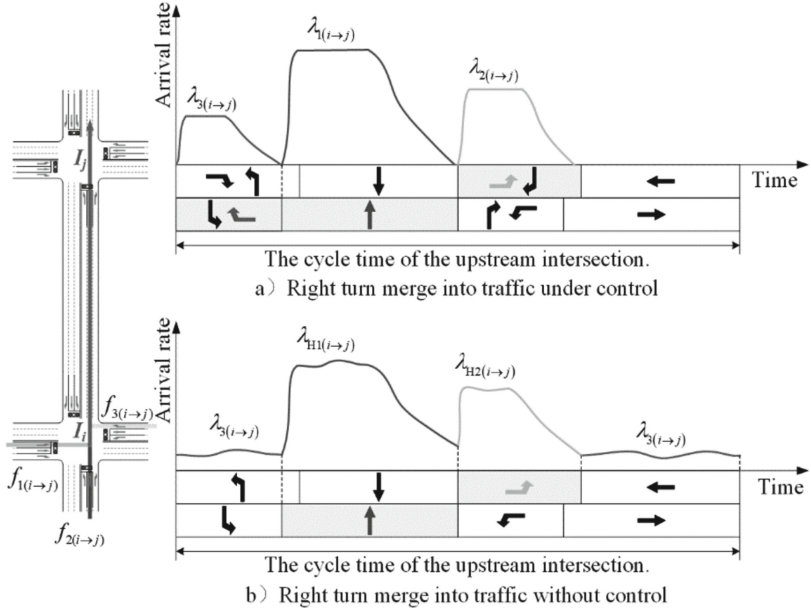


Fig. 1. Signal Cycle Arrival Rate Diagram

3 Vehicle Passing Time Calculation

To accurately estimate the arrival conditions of traffic flows in different directions, we calculate vehicles' actual and ideal passing moments. If vehicle n has a stopping state at the intersection, $S_{(n,j)}$ is set to 1; otherwise, it is set to 0. If $S_{(n,j)} = 0$, the actual passing time $T_{P(n,j)}$ through the stop line at the entrance of the intersection I_j is calculated using interpolation based on the detection time of the vehicle's trajectory points before and after the stop line and its distance from the stop line. If $S_{(n,j)} = 1$, the stopping time $T_{S(n,j)}$ of the first stop at the entrance of the intersection I_j is detected, and the distance $D_{(n,j)}$ between the vehicle and the stop line at that time is recorded. Combining this with the average ideal driving speed $\bar{V}$ of the traffic flow on the current road section, the ideal passing time $T_{E(n,j)}$ through the stop line at the entrance of the intersection I_j is calculated, as shown in Eq. (1).

$$T_{E(n,j)} = \begin{cases} T_{P(n,j)}, & S_{(n,j)} = 0 \\ T_{S(n,j)} + \frac{D_{(n,j)}}{\bar{V}}, & S_{(n,j)} = 1 \end{cases} \quad (1)$$

Based on the actual passing times of vehicles at intersection I_i , the trajectory data converging at intersection I_j is separated and incorporated into the data sets of traffic flows in different directions as defined. To facilitate the subsequent establishment of an arrival rate estimation model, the continuous arrival times within a signal cycle are discretized at intervals of k seconds. The number of vehicles ideally passing through the

stop line at the intersection I_j during time period is counted, and the vehicle arrival ratio $P(t_m)$ is calculated as shown in Eq. (2).

$$P(t_m) = \frac{1}{N} \sum_{n=1}^N F\{T_{E(n,j)} = t_m\} \quad (2)$$

In the equation, N represents the total number of vehicles for which trajectories can be collected; m is the index of the time interval after discretization; $F\{T_{E(n,j)} = t_m\}$ is an indicator variable that determines whether the ideal passing time of any collected vehicle n falls within the period t_m , taking the value of 1 if true and 0 otherwise.

4 Traffic Flow Arrival Rate Estimation Model

4.1 Classification of Traffic Flow Arrival Conditions

The sampled vehicle trajectory data is categorized based on their arrival conditions. We traverse the sampled trajectory data of the S direction traffic flow, excluding trajectories that do not provide valid information for estimating the arrival rate. The vehicle trajectories are sequentially incorporated into data sets for different conditions. The specific conditions within the same signal cycle are outlined as follows:

1) The first sampled vehicle in all S direction traffic flows has a stopping trajectory.

Trajectories satisfying this condition are denoted as A. Based on any sampled vehicle n_A in A, the number $N_{S1(n_A, i \rightarrow j)}$ of arriving vehicles at the intersection for corresponding traffic flow within time period $[T_{R(n_A, j)}, T_{E(n_A, j)}]$ is estimated, according to Eq. (3).

$$N_{S1(n_A, i \rightarrow j)} = \left\lceil \frac{G(T_{P(n_A, j)} - T_{G(n_A, j)})}{h_s} \right\rceil \quad (3)$$

$$N_{S1(n_A, i \rightarrow j)} = N(T_{R(n_A, j)}, T_{E(n_A, j)}) \sim \text{Poisson}(\lambda_{S(i \rightarrow j)} T_{S1(n_A, i \rightarrow j)}) \quad (4)$$

$$T_{S1(n_A, i \rightarrow j)} = \int_{T_{R(n_A, j)}}^{T_{E(n_A, j)}} P(t_m) dt_m \quad (5)$$

where $\lceil \cdot \rceil$ represents the integer operation; $G(T_{P(n_A, j)} - T_{G(n_A, j)})$ is the green light duration from the start of the green phase for the corresponding direction at intersection I_j to the actual passing time of vehicle n_A ; $T_{R(n_A, j)}$ and $T_{G(n_A, j)}$ are the red and green light start times for the corresponding phase when vehicle n_A passes through intersection I_j ; h_s is the saturation headway; Poisson indicates that vehicle arrivals follow a Poisson distribution; $T_{S1(n_A, i \rightarrow j)}$ is the time interval corresponding to the number of arriving vehicles $N_{S1(n_A, i \rightarrow j)}$, as shown in Fig. 2.

2) In the S direction traffic flow, the trajectories of two consecutive sampled vehicles, vehicle n_B^* (the leading vehicle) and vehicle n_B (the trailing vehicle), both are stopping trajectories.

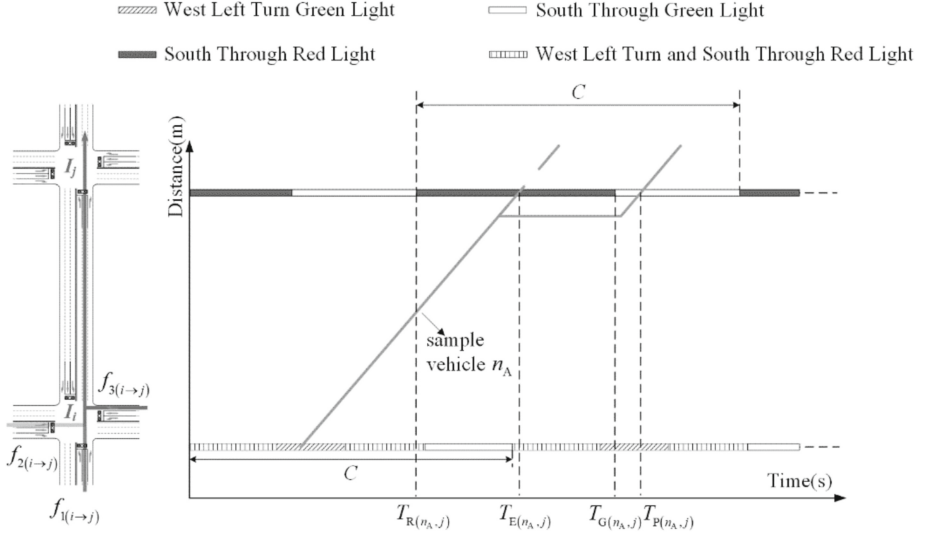


Fig. 2. Diagram of Sampled Vehicles in Set A

Trajectories meeting this condition are denoted as set B. Based on adjacent sampled vehicles n_B^* and n_B in B, as shown in Fig. 3, the number $N_{S2(n_B, i \rightarrow j)}$ of arriving vehicles in time interval $[T_{E(n_B^*, j)}, T_{E(n_B, j)}]$ is estimated according to Eq. (6).

$$N_{S2(n_B, i \rightarrow j)} = \left\lceil \frac{G(T_{P(n_B, j)} - T_{P(n_B^*, j)})}{h_s} \right\rceil \quad (6)$$

$$N_{S2(n_B, i \rightarrow j)} = N(T_{E(n_B^*, j)}, T_{E(n_B, j)}) \sim \text{Poisson}(\lambda_{S(i \rightarrow j)} T_{S2(n_B, i \rightarrow j)}) \quad (7)$$

$$T_{S2(n_B, i \rightarrow j)} = \int_{T_{E(n_B^*, j)}}^{T_{E(n_B, j)}} P(t_m) dt_m \quad (8)$$

3) For the S direction traffic flow, the two adjacent sampled vehicles, vehicle n_C^* (the leading vehicle) and vehicle n_C (the trailing vehicle), have travel trajectories that are respectively a stopping trajectory and a non-stopping trajectory.

Trajectories satisfying this condition are denoted as set C. Based on adjacent sampled vehicles n_C^* and n_C in C, as shown in Fig. 4, the number $N_{S3(n_C, i \rightarrow j)}$ of arriving vehicles during time interval $[T_{E(n_C^*, j)}, T_{P(n_C, j)}]$ is estimated according to Eq. (9).

$$N_{S3(n_C, i \rightarrow j)} = \left\lceil \frac{G(T_{P(n_C, j)} - T_{P(n_C^*, j)})}{h_s} \right\rceil \quad (9)$$

$$N_{S3(n_C, i \rightarrow j)} \geq N(T_{E(n_C^*, j)}, T_{P(n_C, j)}) \sim \text{Poisson}(\lambda_{S(i \rightarrow j)} T_{S3(n_C, i \rightarrow j)}) \quad (10)$$

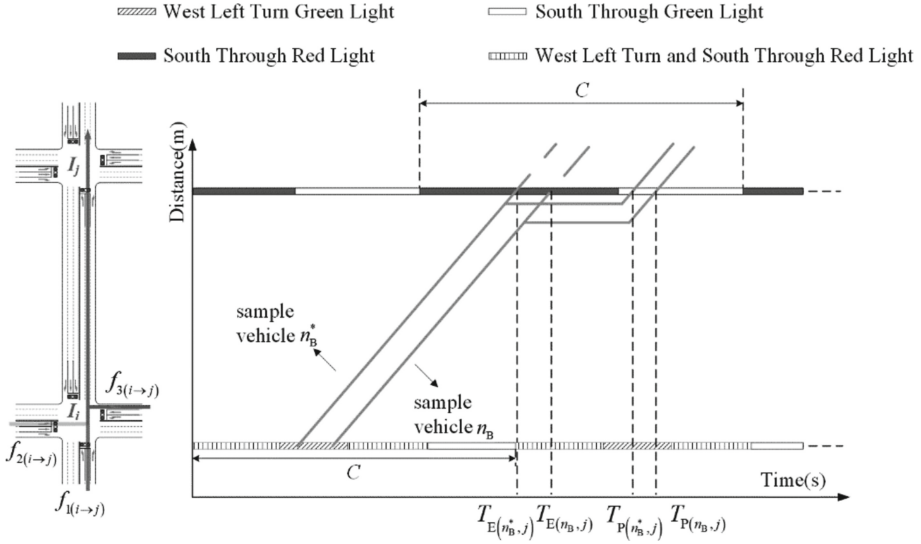


Fig. 3. Diagram of Adjacent Sampled Vehicles in Set B

$$T_{S3}(n_C, i \rightarrow j) = \int_{T_{E(n_C^*, j)}}^{T_{P(n_C, j)}} P(t_m) dt_m \quad (11)$$

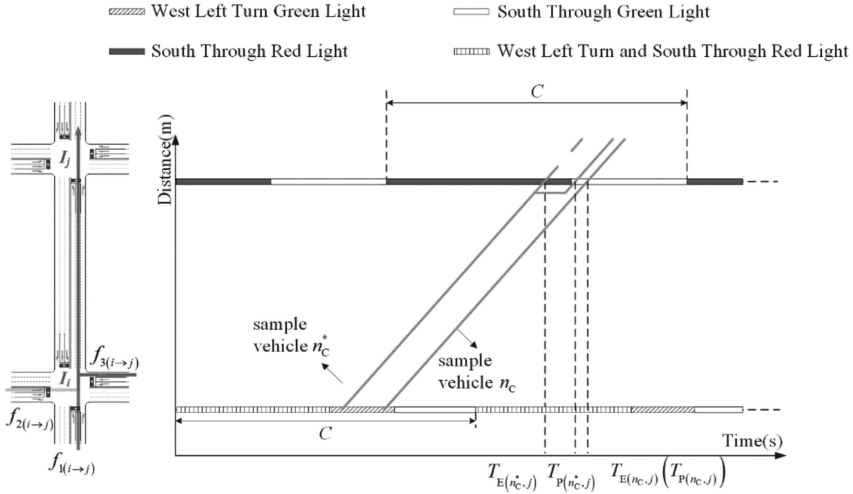


Fig. 4. Diagram of Adjacent Sampled Vehicles in Set C

4.2 Construction of Likelihood Function

Based on the three aforementioned arrival scenarios, the likelihood function for the S direction traffic flow is constructed as shown in Eq. (12), where h represents the possible number of vehicles between any two adjacent sampled vehicles n_C^* and n_C in Set C, with a maximum value of $N_{S3(n_C, i \rightarrow j)}$. In Set A, a total of a sampled vehicle trajectories were collected, in Set B, b pairs of adjacent sampled vehicle trajectories were collected, and in Set C, c pairs of adjacent sampled vehicle trajectories were collected.

$$L(A, B, C | \lambda_{S(i \rightarrow j)}) = \prod_{n_A=1}^a \left\{ \frac{(\lambda_{S(i \rightarrow j)} T_{S1(n_A, i \rightarrow j)})^{N_{S1(n_A, i \rightarrow j)}} e^{-\lambda_{S(i \rightarrow j)} T_{S1(n_A, i \rightarrow j)}}}{N_{S1(n_A, i \rightarrow j)}!} \right\} \cdot \prod_{n_B=1}^b \left\{ \frac{(\lambda_{S(i \rightarrow j)} T_{S2(n_B, i \rightarrow j)})^{N_{S2(n_B, i \rightarrow j)}} e^{-\lambda_{S(i \rightarrow j)} T_{S2(n_B, i \rightarrow j)}}}{N_{S2(n_B, i \rightarrow j)}!} \right\} \cdot \prod_{n_C=1}^c \left\{ \sum_{h=0}^{N_{S3(n_C, i \rightarrow j)}} \frac{(\lambda_{S(i \rightarrow j)} T_{S3(n_C, i \rightarrow j)})^h e^{-\lambda_{S(i \rightarrow j)} T_{S3(n_C, i \rightarrow j)}}}{h!} \right\} \quad (12)$$

By Bayesian inference and assuming an exponential prior distribution (with parameter $\beta_{S(i \rightarrow j)}$ as a known empirical value and $\lambda_{S(i \rightarrow j)}$ positive), the posterior probability function of the arrival rate is constructed. The prior distribution function and posterior probability expression are in Eqs. (13) and (14).

$$P(\lambda_{S(i \rightarrow j)}) = \lambda_{S(i \rightarrow j)} e^{-\beta_{S(i \rightarrow j)} \lambda_{S(i \rightarrow j)}} \quad (13)$$

$$P(\lambda_{S(i \rightarrow j)} | D) = \frac{P(\lambda_{S(i \rightarrow j)}) \cdot L(A, B, C | \lambda_{S(i \rightarrow j)})}{P(D)} \quad (14)$$

4.3 Arrival Rate Estimation

The parameter estimation is conducted using the Maximum A Posteriori estimation method. Since the denominator $P(D)$ of the posterior probability is a constant and does not influence the location of the maximum of the posterior probability function, it can be omitted. The numerator part of the posterior probability function is taken logarithmically and simplified to obtain Eq. (15).

$$\begin{aligned} LP(\lambda_{S(i \rightarrow j)} | D) &= \ln [P(\lambda_{S(i \rightarrow j)}) L(A, B, C | \lambda_{S(i \rightarrow j)})] \\ &= \ln \lambda_{S(i \rightarrow j)} - \beta_{S(i \rightarrow j)} \lambda_{S(i \rightarrow j)} \\ &\quad + \sum_{n_A=1}^a \left\{ N_{S1(n_A, i \rightarrow j)} \cdot (\ln \lambda_{S(i \rightarrow j)} + \ln T_{S1(n_A, i \rightarrow j)}) - \lambda_{S(i \rightarrow j)} T_{S1(n_A, i \rightarrow j)} - \ln(N_{S1(n_A, i \rightarrow j)}!) \right\} \\ &\quad + \sum_{n_B=1}^b \left\{ N_{S2(n_B, i \rightarrow j)} \cdot (\ln \lambda_{S(i \rightarrow j)} + \ln T_{S2(n_B, i \rightarrow j)}) - \lambda_{S(i \rightarrow j)} T_{S2(n_B, i \rightarrow j)} - \ln(N_{S2(n_B, i \rightarrow j)}!) \right\} \end{aligned}$$

$$+ \sum_{n_C=1}^c \left\{ \hat{N}_{S3(n_C, i \rightarrow j)} \cdot (\ln \lambda_{S(i \rightarrow j)} + \ln T_{S3(n_C, i \rightarrow j)}) - \lambda_{S(i \rightarrow j)} T_{S3(n_C, i \rightarrow j)} - \ln(\hat{N}_{S3(n_C, i \rightarrow j)}) \right\} \quad (15)$$

where $\hat{N}_{S3(n_C, i \rightarrow j)}$ is the estimated value of $N_{S3(n_C, i \rightarrow j)}$ as per Eq. (16); l is the loop iteration count; g is the possible vehicle count between adjacent sampled vehicles n_C^* and n_C in Set C, having a maximum value of $N_{S3(n_C, i \rightarrow j)}$.

$$\hat{N}_{S3(n_C, i \rightarrow j)} | N_{S3(n_C, i \rightarrow j)}, \lambda_{S(i \rightarrow j)}^{(l)} = \sum_{h=1}^{N_{S3(n_C, i \rightarrow j)}} h \cdot \frac{\left(\frac{\lambda_{S(i \rightarrow j)}^{(l)} T_{S3(n_C, i \rightarrow j)}}{h!} \right)^h}{\sum_{g=0}^{N_{S3(n_C, i \rightarrow j)}} \frac{\left(\frac{\lambda_{S(i \rightarrow j)}^{(l)} T_{S3(n_C, i \rightarrow j)}}{g!} \right)^g}{N_{S3(n_C, i \rightarrow j)}}} \quad (16)$$

The average value of arrival rate $\lambda_{S(i \rightarrow j)}$ for the S direction traffic flow is denoted as average arrival rate $\bar{\lambda}_{S(i \rightarrow j)}$. By iteratively calculating using the MAP estimation method combined with Eq. (16) until convergence, the result satisfies Eq. (17).

$$\bar{\lambda}_{S(i \rightarrow j)}^{(l+1)} = \frac{1 + \sum_{n_A=1}^a N_{S1(n_A, i \rightarrow j)} + \sum_{n_B=1}^b N_{S2(n_B, i \rightarrow j)} + \sum_{n_C=1}^c \hat{N}_{S3(n_C, i \rightarrow j)}}{\beta_{S(i \rightarrow j)} + \sum_{n_A=1}^a T_{S1(n_A, i \rightarrow j)} + \sum_{n_B=1}^b T_{S2(n_B, i \rightarrow j)} + \sum_{n_C=1}^c T_{S3(n_C, i \rightarrow j)}} \quad (17)$$

Furthermore, multiplying $\bar{\lambda}_{S(i \rightarrow j)}^{(l+1)}$ by the vehicle arrival ratio the arrival rates $\lambda_{S(i \rightarrow j)}$ of traffic flows in different time periods, which satisfy Eq. (18).

$$\lambda_{S(i \rightarrow j)} = \frac{\bar{\lambda}_{S(i \rightarrow j)}^{(l+1)} \int_{T_{R(n, i \rightarrow j)}}^{T_{E(n, i \rightarrow j)}} P(t_m) dt_m}{k} \quad (18)$$

5 Case Study

5.1 Case Scenario

A case study was conducted at the actual intersection of Tongji Road in Chancheng District, Foshan City. The intersection of Fenjiang Middle Road-Tongji Road was the upstream intersection I_1 , while the intersection of Tiyu Road-Tongji Road was the downstream intersection I_2 . The intersection structure and signal control scheme are shown in Fig. 5. This paper selected the traffic flow entering the west through direction of the intersection of Tiyu Road-Tongji Road as the object for estimating the traffic arrival flow rate. The traffic flow arrival rate in this direction was estimated based on the proposed model. VISSIM-based simulation experiments were conducted under real road conditions to obtain vehicle trajectory data for analysis. Each run lasted 4200 s, with the first 600 s as the warm-up period, and no data collected.

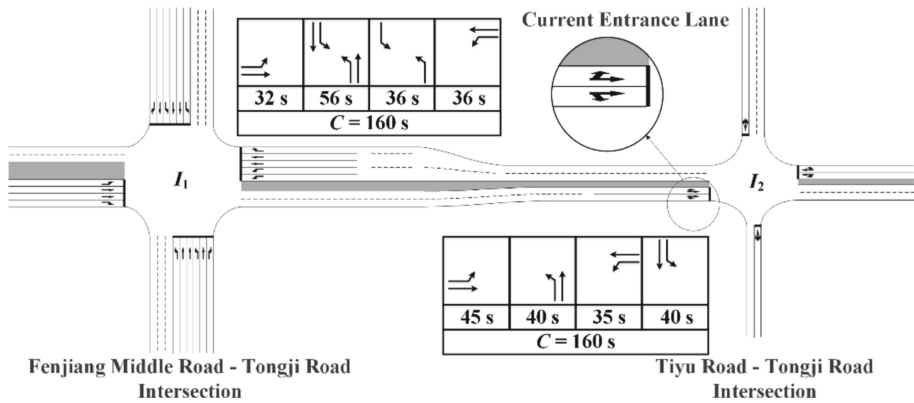


Fig. 5. Intersection Structure and Signal Control Scheme

5.2 Results Analysis

To evaluate the effectiveness of the proposed model, we focused on the through merging traffic flow at the west entrance of the Fenjiang Middle Road-Tongji Road intersection and the left-turn merging traffic flow at the north entrance. The vehicle arrival ratio of historical trajectory data within the signal cycle was calculated. To assess the accuracy of the model's estimation, we utilized the Mean Absolute Percentage Error (MAPE) as the evaluation metric for the through merging traffic flow. However, considering the potential exaggeration of MAPE in smaller sample sizes with lower actual vehicle counts, such as left-turn merging traffic flow, we used the Mean Absolute Error (MAE) as the evaluation metric for such samples. Additionally, we derived the average arrival rate of the respective traffic flows from the full sample data. Using this as a reference, we calculated the MAPE and MAE for both the proposed model and the average arrival rate to enable a comparative evaluation of the model's performance.

Lane 1 represents the through-right lane at the west entrance of the Tiyu Road-Tongji Road intersection, while Lane 2 represents the through-left lane. The overall traffic flow results were derived by calculating the weighted average of the evaluation metrics for these two lanes. Figure 6 shows the curves of the actual arrival rate, the estimated arrival rate by the model proposed, and the average arrival rate for the through merging traffic flow in Lane 2 at a 15% penetration rate. Due to the overestimation of the number of arriving vehicles in the third arrival scenario, the estimated arrival rate of the proposed model in the final calculation results also shows a slight overestimation compared to the actual arrival rate. At a penetration rate of 15% for the through merging traffic flow, the discrepancy between the estimated and actual values for each lane and the aggregate traffic flow is 13.69%, 12.87%, and 13.27% respectively, which is substantially lower than the error range of 38.76% to 61.37% for the average arrival rate. This demonstrates that the model proposed in this paper exhibits superior estimation performance, with the estimated arrival rates based on this model showing a small deviation from the actual values. For the left-turn merging traffic flow, in comparison to the MAE values of 0.96 to 1.61 obtained from the average rate, the MAE values calculated by the proposed model range from 0.40 to 0.61, indicating a smaller error margin. This suggests that

the proposed model maintains good estimation accuracy even when the actual observed values of the samples are relatively small, resulting in reduced errors.

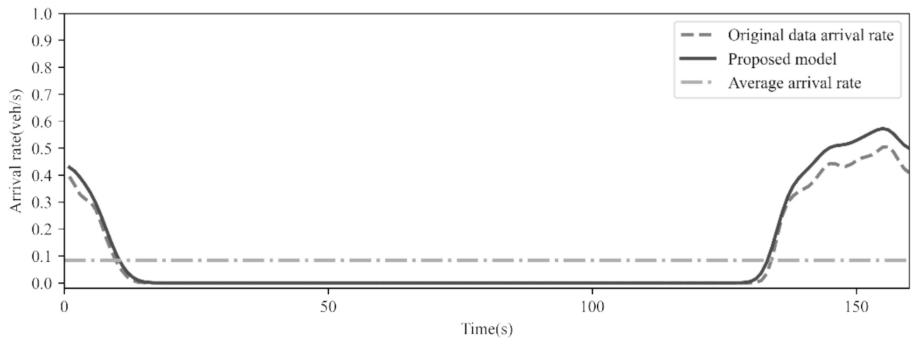


Fig. 6. Actual Arrival Rate, Estimated Arrival Rate by the Proposed Model, and Average Arrival Rate for the Through Merging Traffic Flow in Lane 2 at 15% Penetration Rate

5.3 Accuracy Analysis

To analyze the accuracy of the estimated arrival rate under different penetration rates, we used different random sampling rates to represent various connected vehicle penetration rates. We calculated the MAPE values for both the proposed model and the average arrival rate. Sampling analysis was conducted at every 5% penetration rate increment up to 20%, and then at every 10% increment thereafter. We set different random seeds and conducted three simulations. The weighted average of MAPE is calculated according to the amount of effective trajectory data obtained from each simulation. The results are shown in Table 1. At a penetration rate of 15%, the MAPE for the through merging traffic first dropped below 15%, with an error of 13.27%, while the average arrival rate error was 49.90%. As the penetration rate increased, the estimated arrival rate error of the proposed model fluctuated around 13%, which was much smaller than the average arrival rate error. When the penetration rate was low, the estimation accuracy decreased slightly due to the small sample size. But compared to the average arrival rate, the proposed model still had better estimation accuracy, indicating it can adapt well to arrival rate estimation under different penetration rate scenarios.

Table 1. MAPE of Through Merging Traffic Flow

	Lane 1		Lane 2		Through Merging Traffic Flow	
Penetration Rate	Average Arrival Rate %	Proposed Model %	Average Arrival Rate %	Proposed Model %	Average Arrival Rate %	Proposed Model %
5%	130.51	15.32	65.63	19.77	97.58	17.58
10%	63.20	22.59	59.00	16.68	61.07	19.59
15%	61.37	13.69	38.76	12.87	49.90	13.27
20%	61.85	11.53	32.69	17.12	47.06	14.37
30%	25.14	12.57	24.81	13.98	24.97	13.28
40%	22.96	12.44	24.68	17.13	23.83	14.82
50%	29.10	9.22	31.53	16.65	30.33	12.99
60%	34.24	9.33	32.77	18.00	33.49	13.73
70%	34.98	9.23	28.30	17.02	31.59	13.18
80%	45.94	11.34	28.63	16.38	37.16	13.90
90%	41.42	10.89	29.34	16.99	35.29	13.98
100%	34.17	11.08	27.52	17.41	30.79	14.30

6 Conclusion

This paper categorizes and defines each traffic flow direction, constructs a vehicle arrival likelihood function based on the sampled trajectory data for each direction, and employs the maximum a posteriori (MAP) estimation method to estimate the arrival rate at the downstream intersection entrance. The case study results show that when the penetration rate reaches 15%, the model’s estimation error drops below 15% for the first time, demonstrating good estimation performance with a MAPE of 13.27%, much lower than the average arrival rate error of 49.90%. As the penetration rate of connected vehicles varies, the model’s arrival rate estimation error remains around 13%, much smaller than the average arrival rate, indicating a good estimation effect. The model can effectively adapt to different penetration rate scenarios and provide accurate estimates of arrival rate and traffic volume for each traffic flow direction.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (52172326, 61773168), the Open Foundation of the State Key Laboratory of Subtropical Building and Urban Science (2023KA04), and the Guangdong Basic and Applied Basic Research Foundation (2023A1515012815).

References

1. MiSeon, K., PyongKun, K., KilTaek, L., et al.: Method for obtaining better traffic survey data. *Electronics* **10**(7), 833 (2021)
2. Long, X., Wang, H., Wang, R.: Taxi trajectory characteristics analysis based on frequent sequence mining. *J. S. China Univ. Technol. (Natural Science Edition)* **52**(06), 24–33 (2024). (in Chinese)
3. Tan, C., Liu, L., Wu, H., et al.: Fuzing license plate recognition data and vehicle trajectory data for lane-based queue length estimation at signalized intersections. *J. Intell. Transp. Syst.* **24**(5), 449–466 (2020)
4. Shan, X., Yu, W., Li, Z., et al.: Vehicle trajectory-based traffic volume prediction on urban roads with fast-communication license plate recognition data. *IEEE Trans. Intell. Transp. Syst.* **25**(3), 2768–2778 (2024)
5. Li, J., Li, W., Lian, G.: Urban resident travel survey method based on cellular signaling data. *ISPRS Int. J. Geo Inf.* **12**(8), 304 (2023)
6. Wu, Y.: Traffic volume detection algorithm based on audio visual fusion. Chongqing Jiaotong University (2024). (in Chinese)
7. An, C., He, Y., Lu, J., et al.: One-stage estimation of cyclic arrival rates using license plate recognition data. *J. Intell. Transp. Syst.* 1–14 (2024)
8. Lu, K., Chen, H., Jiang, S., et al.: CPS-based rolling optimization control model for single intersection signal. *J. SE Univ. (Philosophy and Social Science)* **54**(02), 379–388 (2024). (in Chinese)
9. Wang, S.L., Patwary, A.U.Z., Huang, W., et al.: A general framework for combining traffic flow models and Bayesian network for traffic parameters estimation. *Transp. Res. Part C: Emerg. Technol.* **139**, 103664 (2022)
10. Dodia, A., Kumar, S.: A comparison of YOLO based vehicle detection algorithms. In: International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), pp. 1–6 (2023)
11. Shao, Q., Piao, X., Yao, X., et al.: An adaptive composite time series forecasting model for short-term traffic flow. *J. Big Data* **11**(1), 102 (2024)
12. Sun, L.: Arterial signal optimization based on group vehicle trajectory reconstruction. Southeast University (2023). (in Chinese)
13. Kwon, D., Lee, C., Kang, H., et al.: Large-scale network imputation and prediction of traffic volume based on multi-source data collection system. *Transp. Res. Rec.* **2677**(9), 30–42 (2023)
14. Chen, L., Hao, W., Li, Q., et al.: Traffic volume forecast model based on BP neural network optimized by improved sparrow search algorithm. *J. Harbin Inst. Technol.* **56**(07), 94–101 (2024). (in Chinese)



Multi-spatial Scales Delay Time Index Calculation Method Based on Digital Road Networks

Kai Lu^{1,2,4}, Kun Zeng¹, Yaoting Wu³, Jingfeng Wei¹, and Xingdong Deng^{3,4}(✉)

¹ School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510640, China

² State Key Laboratory of Subtropical Building and Urban Science, Guangzhou 510640, China

³ Guangzhou Urban Planning and Design Survey Research Institute, Guangzhou 510050, China
dengxingdong@gzpi.com.cn

⁴ Collaborative Innovation Center for Natural Resources Planning and Marine Technology of Guangzhou, Guangzhou 510060, China

Abstract. In the context of digital road networks, lane-level data support and dynamic traffic perception enable the precise calculation of operational parameters for different lanes during various time periods. Delay metrics play a crucial role in evaluating traffic states. The delay time index defined in this paper intuitively reflects traffic delay states and eliminates the influence of network scale on delay assessment, making it an appropriate evaluation metric. By establishing relationships among the indices at different levels of the road network through weighting, and considering factors such as traffic flow and free-flow travel time, this paper presents definitions and calculation methods for the weight coefficients of various evaluation objects within the network. Based on the constructed weight system, a multi-scale delay time index calculation method based on weight coefficients is derived. Case results indicate that the traffic states evaluation metrics obtained through weight coefficient estimation are consistent with those derived from simulation and defined calculations. Additionally, when the data acquisition rate for the road network is 10%, the relative accuracy of the estimates can reach 96.0% at an 80% cumulative probability. Utilizing foundational operational data provided by digital road networks, not only enables the step-by-step calculation of the delay time index across multi-spatial scales of the road network, but the overall traffic states of the network can also be estimated based on partial traffic operation data, effectively supporting the assessment needs of the overall traffic states of the network in scenarios where the digital road network construction is incomplete or coverage is insufficient.

Keywords: Digital Road Networks · Traffic Delay · Weight Coefficient · Multi-Spatial Scales · Step-By-Step Calculation

1 Introduction

The digital road network, as a new generation of information technology that integrates transportation infrastructure with digital technologies, digitizes traffic elements and road conditions within urban road networks. This provides critical data support for the perception of road traffic states, situational assessment, and traffic management. The primary task for the effective application of digital road networks is to conduct an objective and comprehensive evaluation of traffic states, which is also a fundamental basis for subsequent traffic situation analysis and management decision-making. The digital road network can provide dynamic, high-resolution lane-level trajectory data, offering new methods and ideas for the refined evaluation of traffic states.

In recent years, numerous studies have been conducted both domestically and internationally in the field of urban traffic states evaluation. Early in the United States, evaluation methods based on indicators such as Level of Service (LoS) and Road Congestion Index (RCI) were proposed to measure changes in traffic states [1]. As the problem of traffic congestion has become increasingly severe, relevant standards and norms have gradually been established in China. Among these, the Traffic Performance Index (TPI), also known as the congestion index, has become an important tool for quantifying the smoothness or congestion of urban road networks.

Existing research primarily focuses on the calculation of traffic state indicators and the classification and identification of traffic flow states. Most studies measure road traffic states through the computation of traffic performance indicators, with representative indicators including segment speed [2], road traffic density [3], and traffic volume [4]. Pattara-Atikom et al. [5] determined road congestion status by calculating vehicle speed variations on segments. Lu Kai et al. [6] constructed a traffic congestion index using vehicle travel time and free-flow travel time to assess the status at various levels of the road network. As traffic demand and network complexity increase, delays have gradually become a critical indicator for evaluating traffic states. Delay is an important metric for assessing intersection performance and measuring service levels, playing a significant role in evaluating road traffic states [7].

Research on traffic delays has mainly concentrated on the establishment of delay indicators and their applications. Zhang et al. [8] developed a highway accident delay prediction model using simulation methods, considering the impact of different vehicle types. The results indicated that this method outperforms traditional approaches in terms of accuracy and practicality. Zhang et al. [9] analyzed actual traffic data from Beijing and Shenzhen, proposing that delay indicators can indicate the reliability of traffic conditions. Song et al. [10] utilized urban bus delay indicators to identify segments and causes of traffic delays, further revealing the application of delay indicators in traffic state identification. Some studies have considered the allocation of traffic weights, which involves evaluating the relative importance of different levels within the network. Wong [11] accounted for variations in traffic cycles affecting delays in the model, proposing weight coefficients for delays and stop counts in intersection delay evaluations. Su et al. [12] adopted a multi-indicator system to integrate delay states across various components of the road network, establishing a weighted index system to enhance the analytical accuracy of urban traffic states. Innovative methods for calculating traffic delay indicators have improved the accuracy of delay estimation and computation with the introduction

of new technologies in recent years. Liu et al. [13] proposed a real-time delay estimation model based on cooperative vehicle-road systems at signalized intersections, applicable to dynamic traffic environments. Rani and Shaw [14] developed a segment delay estimation method using neural networks, improving estimation accuracy by considering multiple factors. However, the existing research on calculating and assessing traffic states has mainly concentrated on designated spatial scales. Moreover, although current studies propose the weighted calculation of traffic state indicators, there is still a lack of detailed analysis regarding the definition and computation methods of weight coefficients. The physical significance of the calculated results remains unclear, and in-depth research on multi-scale calculations of urban road traffic states is still lacking.

To address the aforementioned issues, this paper first defines the Delay Time Index (DTI). To further analyze the relationships among various levels of the DTI, a weighting approach is employed to construct the relationships within the road network. Utilizing the lane-level detailed data provided by digital road networks, factors such as traffic flow and free-flow travel time are considered to define the weighting coefficients. Under the constructed traffic weighting system, the DTI is derived and calculated for sub-segments, segments, and even the entire network. This reveals the interrelations of the DTI across different layers and establishes a method for tiered calculation of the DTI at multiple spatial scales. Consequently, this approach enables a comprehensive and refined assessment of traffic delays within the digital road network.

2 Definition of Delay Time Index at Different Spatial Scales

The delay index, as a core parameter for measuring the service level and traffic efficiency, serves as a crucial reference for assessing road congestion conditions and optimizing traffic resource allocation. In transportation system evaluations, ensuring the comparability and applicability of evaluation metrics is essential. The traditional delay indicator (delay time) struggles to accurately reflect the true congestion differences across road networks of varying scales. For instance, while two networks exhibit the same delay time, a larger-scale network experiences relatively lighter actual congestion due to its broader coverage area. This indicates that relying solely on delay time is insufficient for objectively assessing the traffic states of different networks.

To address this issue, this study defines the Delay Time Index (DTI) as the ratio of delay time to free-flow travel time. Through this normalization process, the impact of road network scale on delay assessment is eliminated. The application of the DTI enhances the comparability of traffic state evaluations across networks of different scales, providing a more scientific foundation for analyzing and comparing the traffic states of various networks. Depending on the subject of evaluation, the DTI is defined and calculated for lanes, sub-segments, segments, intersections, roads, sub-areas, and entire networks.

Definition 1 - Lane Delay Time Index: The lane delay time index is defined as the ratio of the total delay time of all vehicles on a lane to the total free-flow travel time.

The calculation formula is as follows:

$$DI_l^L = \frac{\sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \quad (1)$$

where DI_l^L represents the delay time index for lane L_l , $S_{L,l}^V$ is the set of vehicles that traveled on lane L_l during the evaluation period, $d_{(v,l)}^V$ is the delay time for vehicle V_v traveling through lane L_l , and $t_{f(v,l)}^V$ is the free-flow travel time for vehicle V_v traveling through lane L_l .

Definition 2 - Sub-segment Delay Time Index: The sub-segment delay time index is defined as the ratio of the total delay time of all vehicles on a sub-segment to the total free-flow travel time. The calculation formula is as follows:

$$DI_u^U = \frac{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \quad (2)$$

where DI_u^U represents the delay time index for sub-segment U_u , and $S_{U,u}^L$ is the set of lanes included within sub-segment U_u .

Definition 3 - Segment Delay Time Index The segment delay time index is defined as the ratio of the total delay time of all vehicles on a segment to the total free-flow travel time. The calculation formula is as follows:

$$DI_s^S = \frac{\sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \quad (3)$$

where DI_s^S represents the delay time index for segment S_s , and $S_{S,s}^U$ is the set of sub-segments included within segment S_s .

In practice, the definitions and calculations of delay time indices for intersections, roads, sub-zones, and networks are similar to those for segments; therefore, they will not be elaborated here. The formulas are provided as follows:

$$DI_i^I = \frac{\sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \quad (4)$$

$$DI_r^R = \frac{\sum_{s \in S_{R,r}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V + \sum_{i \in S_{R,r}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{s \in S_{R,r}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_{R,r}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \quad (5)$$

$$DI_z^Z = \frac{\sum_{s \in S_{Z,z}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V + \sum_{i \in S_{Z,z}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{s \in S_{Z,z}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_{Z,z}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \quad (6)$$

$$DI^A = \frac{\sum_{s \in S_A^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V + \sum_{i \in S_A^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{s \in S_A^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_A^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \quad (7)$$

In Eqs. (4) to (7), DI_i^I , DI_r^R , DI_z^Z , and DI^A represent the delay time indices for intersection I_i , road R_r , sub-zone Z_z , and network A, respectively. $S_{L,i}^L$ denotes the set of lanes contained within intersection I_i , $S_{R,r}^S$ is the set of segments contained within road R_r , $S_{R,r}^I$ is the set of intersections contained within road R_r , $S_{Z,z}^S$ is the set of segments contained within sub-zone Z_z , $S_{Z,z}^I$ is the set of intersections contained within sub-zone Z_z , S_A^S is the set of segments contained within network A, and S_A^I is the set of intersections contained within network A.

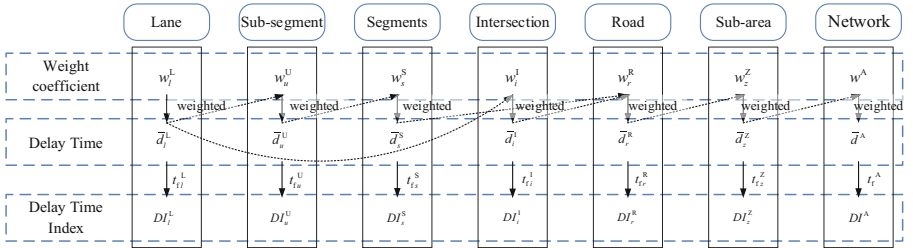


Fig. 1. Schematic of Traffic Delay Calculation Index Relationships at Multiple Spatial Scales in the Road Network

In practice, the DTI across different levels of the network are interrelated. However, how can this relationship be reasonably established? From a spatial composition perspective, there exists a containment relationship among the evaluation objects at various levels; thus, the relationships among the indices can be constructed through weighting. By calculating the weighting coefficients of evaluation objects at different spatial scales and utilizing the weighting system to establish the relationships among the DTIs for lanes, sub-segments, segments, intersections, roads, sub-areas, and the entire network, a method for calculating the delay time index across multiple spatial scales within the network can be established. This enables multi-spatial-scale calculations of DTIs in digital road networks and quantifies the impact of the traffic delay status of different

components on the overall traffic delay status within the network. The diagram of the index relationships is shown in Fig. 1.

3 Calculation Methods for Multi-Spatial-Scale Delay Time Index in Road Networks

3.1 Weight Coefficients

To reasonably quantify the relationship between DTIs at different levels, this study proposes constructing these relationships through weight coefficients. Given the differences in traffic states among evaluation objects at various levels within the road network, the weight coefficients are defined by considering both the traffic flow and free-flow travel time of each evaluation object. The weight coefficient is calculated as the ratio of the total free-flow travel time of all vehicles passing through a given evaluation object over a specified period to the total free-flow travel time of all vehicles in the entire road network. The traffic weight coefficients for all components within the same spatial scale belonging to that evaluation object are summed to obtain the traffic weight coefficient for this evaluation object, as shown in the following formula:

$$w_{(v,l)}^V = \frac{t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} \quad (8)$$

$$w_l^L = \frac{\sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{v \in S_{L,l}^V} \frac{t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{v \in S_{L,l}^V} w_{(v,l)}^V \quad (9)$$

$$w_u^U = \frac{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{l \in S_{U,u}^L} \frac{\sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{l \in S_{U,u}^L} w_l^L \quad (10)$$

$$w_s^S = \frac{\sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{u \in S_{S,s}^U} \frac{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{l \in S_{S,s}^L} w_l^L \quad (11)$$

$$w_i^I = \frac{\sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{l \in S_{L,i}^L} \frac{\sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{NV} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} = \sum_{l \in S_{L,i}^L} w_l^L \quad (12)$$

$$\begin{aligned}
w_r^R &= \frac{\sum_{s \in S_{R,r}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_{R,r}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{N^V} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} \\
&= \sum_{s \in S_{R,r}^S} w_s^S + \sum_{i \in S_{R,r}^I} w_i^I = \sum_{l \in S_{R,r}^L} w_l^L \quad (13)
\end{aligned}$$

$$\begin{aligned}
w_z^Z &= \frac{\sum_{s \in S_{Z,z}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_{Z,z}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V}{\sum_{v=1}^{N^V} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} \\
&= \sum_{s \in S_{Z,z}^S} w_s^S + \sum_{i \in S_{Z,z}^I} w_i^I = \sum_{l \in S_{Z,z}^L} w_l^L \quad (14)
\end{aligned}$$

where $w_{(v,l)}^V$, w_l^L , w_u^U , w_s^S , w_i^I , w_r^R and w_z^Z represent the traffic weight coefficients for vehicle V_v , lane L_l , sub-segment U_u , segment S_s , intersection I_i , road R_r , sub-area Z_z , respectively.

Based on the definition of traffic weight coefficients within the road network, the traffic weight coefficients of all lanes, sub-segments, segments, and intersections in the network are summed separately, with the total equal to 1, as represented by the following formula:

$$w^A = \sum_{s=1}^{N^S} w_s^S + \sum_{i=1}^{N^I} w_i^I = \sum_{u=1}^{N^U} w_u^U + \sum_{i=1}^{N^I} w_i^I = \sum_{l=1}^{N^L} w_l^L = \sum_{l \in S_A^L} \sum_{v \in S_{L,l}^V} w_{(v,l)}^V = 1 \quad (15)$$

where w^A represents the total traffic weight coefficient of the road network, N^S denotes the number of segments in the network, N^I is the number of intersections in the network, N^U is the number of sub-segments in the network, N^L is the number of lanes in the network, and S_A^L denotes the collection of lanes contained within road A.

3.2 Multi-scale Traffic Delay Time Index Calculation

Under the weight coefficient system, the delay time index for each lane, sub-segment, segment, intersection, road, sub-area, and the entire network is calculated using both delay time and free-flow travel time, while revealing the relationships between the various indices. Based on Eqs. (1) to (7) and incorporating the weight coefficient calculation formulas, the construction of the delay time index formulas under the weight system is

derived, as shown in Eqs. (16) to (22).

$$DI_l^L = \frac{\sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{v=1}^{N^V} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V} \times \frac{\sum_{v=1}^{N^V} \sum_{l \in S_{V,v}^L} t_{f(v,l)}^V}{\sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{v \in S_{L,l}^V} (DI_{(v,l)}^V \times w_{(v,l)}^V)}{\sum_{v \in S_{L,l}^V} w_{(v,l)}^V} \quad (16)$$

$$DI_u^U = \frac{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} (DI_{(v,l)}^V \times w_{(v,l)}^V)}{\sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} w_{(v,l)}^V} = \frac{\sum_{l \in S_{U,u}^L} (DI_l^L \times w_l^L)}{\sum_{l \in S_{U,u}^L} w_l^L} \quad (17)$$

$$DI_s^S = \frac{\sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{l \in S_{S,s}^L} \sum_{v \in S_{L,l}^V} (DI_{(v,l)}^V \times w_{(v,l)}^V)}{\sum_{l \in S_{S,s}^L} \sum_{v \in S_{L,l}^V} w_{(v,l)}^V} = \frac{\sum_{u \in S_{S,s}^U} (DI_u^U \times w_u^U)}{\sum_{u \in S_{S,s}^U} w_u^U} \quad (18)$$

$$DI_i^I = \frac{\sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} (DI_{(v,l)}^V \times w_{(v,l)}^V)}{\sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} w_{(v,l)}^V} = \frac{\sum_{l \in S_{L,i}^L} (DI_l^L \times w_l^L)}{\sum_{l \in S_{L,i}^L} w_l^L} \quad (19)$$

$$DI_r^R = \frac{\sum_{s \in S_{R,r}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V + \sum_{i \in S_{R,r}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{s \in S_{R,r}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_{R,r}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{l \in S_{R,r}^L} (DI_l^L \times w_l^L)}{\sum_{l \in S_{R,r}^L} w_l^L} \quad (20)$$

$$DI_z^Z = \frac{\sum_{s \in S_{Z,z}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V + \sum_{i \in S_{Z,z}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{s \in S_{Z,z}^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_{Z,z}^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{l \in S_{Z,z}^L} (DI_l^L \times w_l^L)}{\sum_{l \in S_{Z,z}^L} w_l^L} \quad (21)$$

$$DI^A = \frac{\sum_{s \in S_A^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V + \sum_{i \in S_A^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{s \in S_A^S} \sum_{u \in S_{S,s}^U} \sum_{l \in S_{U,u}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V + \sum_{i \in S_A^I} \sum_{l \in S_{L,i}^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} = \frac{\sum_{l \in S_A^L} \sum_{v \in S_{L,l}^V} d_{(v,l)}^V}{\sum_{l \in S_A^L} \sum_{v \in S_{L,l}^V} t_{f(v,l)}^V} \\ = \sum_{l \in S_A^L} (DI_l^L \times w_l^L) \quad (22)$$

In fact, the Delay Time Index at any level can be aggregated layer by layer using the Delay Time Indices of its subordinate levels and the corresponding weight coefficients.

For example, in Eq. (22), DI^A can also be expressed in terms of sub-areas and roads, with the formulas represented as $\sum_{z \in S_A^Z} (DI_z^Z \times w_z^Z)$ and $\sum_{r \in S_A^R} (DI_r^R \times w_r^R)$, respectively.

Here, S_A^Z denotes the set of sub-areas contained within the road network A, while S_A^R represents the set of roads included within the road network A.

4 Case Analysis

4.1 Basic Experimental Conditions

The central urban area of Chancheng District in Foshan City, Guangdong Province, is taken as a research case. Based on the actual foundational data of the road network, a simulation network is constructed using VISSIM software to analyze and validate the methods proposed in this paper. Due to the current incompleteness of digital road network construction and the lack of comprehensive real-world data, simulation technology effectively mimics the digital road network and validates the effectiveness of the proposed multi-scale Delay Time Index calculation method. The structure of the case road network, the OD traffic demand of sub-areas, and the signal control schemes at each intersection are illustrated in Fig. 2. The road network comprises 13 main roads, 20 signal-controlled intersections, and 19 primary traffic flow input/output nodes. The simulation runs for 28,800 s, with the first 1,800 s serving as a warm-up period, and the vehicle operation data collected after 1,800 s is used for analysis and calculation.

4.2 Results Analysis

4.2.1 Delay Time Index at Different Levels

Taking the road network sub-area division scheme shown in Fig. 2 as an example, starting from the 1800 s of the simulation experiment, the simulation time from the 1800 s to the 3600 s is selected as the evaluation period. Based on the multi-spatial-scale delay time index calculation method proposed in this paper, the delay time index results for each lane are computed, and subsequently, the delay time index results for each sub-area are derived.

Using the vehicle operation data, the travel time for each vehicle traversing each lane in the network can be calculated, which in turn allows for the computation of the delay time index for each lane. Based on the lane results and their corresponding weight coefficients, the delay time index at various levels can be further calculated. The calculation results of traffic state indicators for each lane and road are shown in Tables 1 and 2.

Taking sub-area Z_1 as an example, based on the lane composition of sub-area Z_1 , the traffic weight coefficients w_1^Z and w^A for sub-area Z_1 and network A are derived using vehicle operation data.

$$w_1^Z = \sum_{s \in S_{Z,1}^S} w_s^S + \sum_{i \in S_{Z,1}^I} w_i^I = \sum_{l \in S_{Z,1}^L} \sum_{v \in S_{L,l}^V} w_{(v,l)}^V = 0.4508 \quad (23)$$

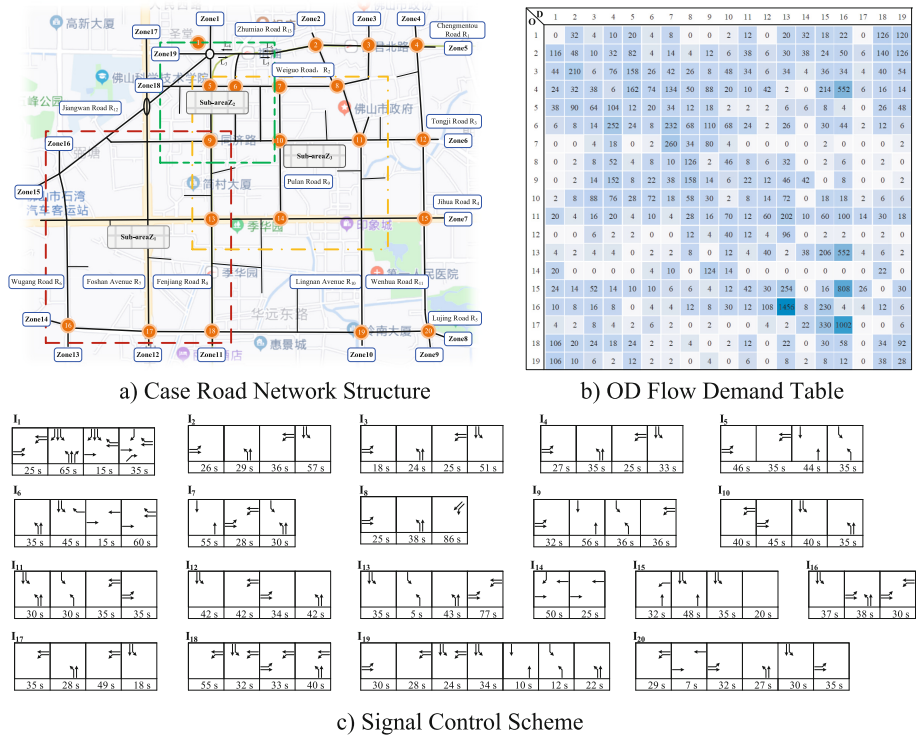


Fig. 2. Experimental Case Road Network Conditions

Table 1. Calculation of Traffic Operation Status Indicators for Lanes

Lane Number	$q_l^L(\text{veh/h})$	$\bar{d}_l^L(\text{s})$	$t_{II}^L(\text{s})$	w_l^L	DI_l^L
L ₁	888	15.70	12.46	0.0055	1.26
L ₂	672	13.19	16.14	0.0054	0.82
L ₃	1026	25.97	14.84	0.0076	1.75
...	...	...	...	...	...
Total	23480	116.04	85.04	1	1.36

$$w^A = \sum_{l=1}^{N^L} w_l^L = \sum_{l \in S_A^L} \sum_{v \in S_{L,l}^V} w_{(v,l)}^V = \sum_{v=1}^{N^V} w_v^V = 1 \quad (24)$$

Furthermore, by utilizing the traffic weight coefficient w_l^L and its corresponding delay time index DI_l^L , the delay time index for sub-area Z_1 and network A can be calculated

Table 2. Calculation of Traffic Operation Status Indicators for Roads

Road Number	Segment Number	$q_l^L/(\text{veh/h})$	$\bar{d}_l^L(\text{s})$	$t_{ql}^L(\text{s})$	w_l^L	w_r^R	DI_r^R
R ₁	S _W	1648	2.52	24.27	0.020	0.047	0.21
	S _E	3278	4.67	16.50	0.027		
R ₂	S _W	3076	93.91	18.37	0.028	0.066	3.48
	S _E	3100	54.25	24.16	0.038		
R ₃	S _W	1920	74.38	42.81	0.041	0.098	1.48
	S _E	3626	40.53	31.20	0.057		
...	...	...	...	...	...	...	...
Total		23480	116.04	85.04	1	1	1.36

as follows:

$$DI_1^Z = \sum_{l \in S_{Z,1}^L} (DI_l^L \times w_l^L) = 0.65 \quad (25)$$

$$DI^A = \sum_{l \in S_A^L} (DI_l^L \times w_l^L) = 1.36 \quad (26)$$

4.2.2 Validity and Sensitivity Analysis

1) Validity Verification

By obtaining all vehicle operation data for each lane, sub-section, road segment, roadway, sub-area, and the entire network through simulation experiments, the traffic delay time index for the road, sub-area, and network was calculated based on the definition of the traffic delay time index, as shown in Table 3. A comparison of Tables 1, 2 and 3 indicates that the derived results are completely consistent with the outputs from the simulation software.

It can be observed that the calculation method for traffic state evaluation indicators based on road lane weight coefficients not only accurately reflects the actual traffic states of the network but also enables a refined description of the traffic states across multiple spatial levels of the network.

2) Sensitivity Analysis

Although the traffic states of lanes typically exhibit complex variability, the free-flow travel time is fixed, and traffic flow remains relatively stable over a period. Consequently, the weight coefficients also demonstrate a degree of stability during that time. By utilizing historical traffic flow data, lane weight coefficients can be predetermined. Combining these with real-time vehicle operation data allows for real-time estimation of the network delay time index using the aforementioned calculation method.

Table 3. Evaluation Results of Traffic Operation Status from Simulation Output

Evaluation Object	Road						Sub-area			Network
	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	Z ₁	Z ₂	Z ₃	
Number of Vehicles	2222	2829	2691	2001	2307	4661	13484	5960	8606	23480
Delay Time Index	0.21	3.48	1.48	1.48	0.35	1.46	0.65	2.74	1.89	1.36

Taking the entire network A and sub-area Z₁ as examples, vehicle operation data from the simulation experiment was used to randomly extract portions of the data at various collection rates (5%, 10%, 15%, 20%, 25%, 30%). This enables the estimation of lane delay time indices based on the weight coefficients. For each data extraction ratio, 100 random experiments were conducted, and the results were statistically compared with the network delay time index obtained from the full sample calculation, yielding the relative accuracy of the approximate estimations, as shown in Fig. 3. The relative accuracy reflects the degree of closeness between the approximate estimation based on partial sampled data and the precise calculation based on all collected data. The calculation method is as follows:

$$R_A = \frac{R_C - |d|}{R_C} \times 100\% \tag{27}$$

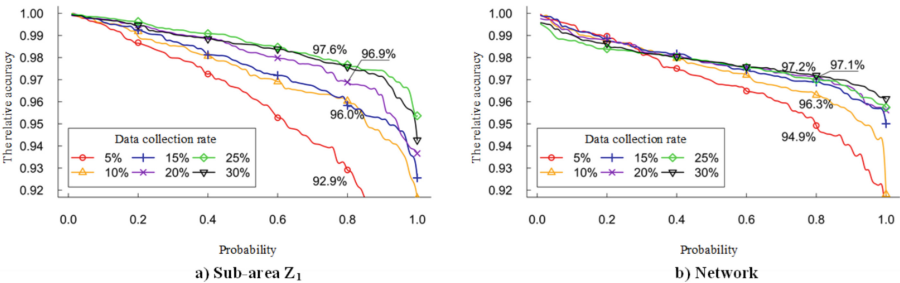


Fig. 3. Relative Accuracy of Approximate Estimation of Network Delay Time Index at Different Data Acquisition Rates

In the equation, R_A represents the relative accuracy, d denotes the difference between the approximate estimation based on partial sampled data and the precise calculation result based on all collected data, R_C is the precise calculation result based on all collected data.

As shown in Fig. 3, when the data collection rate is 5%, the relative accuracy of the estimation can reach 92.9% at an 80% cumulative probability; similarly, at a data collection rate of 10%, the relative accuracy can reach 96.0% at the same cumulative probability. This indicates that as the data collection rate increases, the

relative accuracy of the estimation also rises. Therefore, the multi-scale traffic state index calculation model proposed in this study can accurately estimate the traffic states of the entire network by randomly obtaining traffic operation data from a subset of vehicles. This is particularly important in contexts where the current digital road network construction is incomplete or insufficiently covered.

5 Conclusion

- 1) The multi-scale delay time index calculation method proposed in this paper, based on a digital road network environment, achieves a composite calculation of traffic delay time indices for sub-segments, segments, roads, sub-areas, and the entire network by defining lane weight coefficients. This method not only provides a scientific tool for the refined evaluation of traffic states but also contributes theoretical support for the construction of traffic evaluation systems and decision support systems under digital road networks.
- 2) The minimum level of this study is the lane level. Defining its weight coefficients allows for a refined description of the traffic states in different characteristic segments of the road, thereby providing a realistic foundation for the precise management and governance of urban roads.
- 3) Utilizing high-resolution lane-level data provided by the digital road network and partially acquired real-time traffic operational data, the weighted calculation method proposed in this paper can approximate the overall network's delay time index even when data acquisition is incomplete. This indicates that the proposed method retains high practicality and application value, even in scenarios where the construction of digital road networks is not yet complete.
- 4) Building upon the definitions of weight coefficients and traffic states evaluation methods provided in this paper, future research could integrate the driving characteristics and management needs of different turning types of traffic flows. This would further enrich and expand the application of weight coefficients in digital road network environments and enhance the accuracy and adaptability of traffic states assessments by incorporating additional traffic factors (such as real-time events and emergencies). This will be an important direction for subsequent research, promoting the in-depth application of digital road networks within complex traffic systems.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (52172326), the Open Foundation of the State Key Laboratory of Subtropical Building and Urban Science (2023KA04), and the Guangdong Basic and Applied Basic Research Foundation (2023A1515012815).

References

1. American Transportation Research Board. Road Capacity Manual. Trans. by Ren, F., Liu, X., et al. People's Transportation Publishing House, Beijing (2007). (in Chinese)
2. Wang, J., Wang, L.: Congestion analysis of traffic networks with direction-dependent heterogeneity. *Physica A* **392**(2), 392–399 (2013)
3. Guan, J.: Urban traffic composite index, travel index, and mathematical model. *J. Transp. Syst. Eng. Inf. Technol.* **2004**(01), 49–53 (2004). (in Chinese)
4. Du, M., Jiang, X., Anthony, C.: Identifying critical links using network capacity-based indicator in multi-modal transportation networks. *Transportmetrica B: Transp. Dyn.* **10**, 1126–1150 (2022)
5. Pattara-Atikom, W., Pongpaibool, P., Thajchayapong, S.: Estimating road traffic congestion using vehicle velocity. In: *Proceedings of the 6th International Conference on ITS Telecommunications*, pp. 1001–1004 (2006)
6. Lu, K., Wu, Y., Deng, X., et al.: Step-by-step calculation model of road traffic operation index suitable for multi-spatial scales. *J. Southeast Univ. (Nat. Sci. Edn.)* **52**(06), 1178–1188 (2022). (in Chinese)
7. Sun, C.: Research on urban traffic network state evaluation and analysis. Ph.D. thesis, South China University of Technology (2010). (in Chinese)
8. Zhang, H., Ni, F., Yang, S.: Prediction of freeway traffic incident delay based on simulation. In: *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pp. 167–171. IEEE (2012)
9. Zhang, L., Zeng, G., Guo, S., et al.: Comparison of traffic reliability index with real traffic data. *EPJ Data Sci.* **6**(1), 1–15 (2017)
10. Song, G., Kim, S., Jung, U., et al.: An experimental study on the modeling system of the public bus traffic delay index. In: *Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City*, pp. 514–520 (2019)
11. Wong, S.: Derivatives of the performance index for the traffic model from TRANSYT. *Transp. Res. Part B Methodol.* **29**(5), 303–327 (1995)
12. Su, F., Dong, H., Jia, L., et al.: On urban road traffic state evaluation index system and method. *Mod. Phys. Lett. B* **31**(01), 1650428 (2017)
13. Liu, H., Rai, L., Wang, J., et al.: A new approach for real-time traffic delay estimation based on cooperative vehicle-infrastructure systems at the signal intersection. *Arab. J. Sci. Eng.* **44**, 2613–2625 (2019)
14. Rani, P., Shaw, D.: A neural network approach for link delay prediction. In: *2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, pp. 1–4. IEEE (2017)



Behavioral Experimental Study on Multi-mode Combined Travel Route Choice Considering Information Acceptance

Xiaohui Sun¹ , Jiaxin Zhang², and Min Chen¹

¹ School of Traffic and Transportation Engineering, Xinjiang University, Urumqi 830017, China
xhsun347@xju.edu.cn

² College of Civil Engineering and Architecture, Xinjiang University, Urumqi 830017, China

Abstract. Based on the multi-mode transport network environment, an experimental platform for multi-mode combined route choice behavior was built to explore the multi-mode combined route choice laws of subjects with different information acceptance willingness. Firstly, based on information acceptance willingness data, the subjects were divided into three classes with respectively low, medium and high willingness to accept information by potential profile analysis. Secondly, z-Tree software was used to set up the transport network with corresponding route cost function, and 200 rounds of experiments about multi-mode combined route choice for subjects of different classes were then conducted. Finally, the experimental data was processed and analyzed and the following results were obtained. The subjects were almost unwilling to change strategies after 40 rounds of experiments, and the experimental transport network reached a balanced state. There are differences in the multi-mode combined route choice behavior among subjects with different information acceptance willingness: the average travel cost shows low > medium > high; The decision switching frequency shows medium > high > low; Compared with subjects with low information acceptance willingness, subjects with medium information acceptance willingness are more inclined to direct response pattern, while subjects with high information acceptance willingness are more inclined to opposite response pattern.

Keywords: Travel Route Choice Law · Behavioral Experiment · Multi-Mode Combined Route · Information Acceptance Willingness · Potential Profile Analysis

1 Introduction

In recent years, with the continuous expansion of urban areas, the emergence of new modes of travel, and the improvement of travel connection facilities, multi-mode combined travel has gradually become the main trend of residents' daily travel under the background of urbanization. Compared with single-mode travel, multi-mode combined travel involves the connection and transfer of multiple travel modes, and its route decision-making behavior is more complicated. However, the current research mainly focuses on single-mode travel route choice, and the route decision-making behavior law under multi-mode combined travel is still unclear.

At the same time, the current research on travel route choice behavior mainly collects data through questionnaire, subsequently constructs models to analyze the law of travel route choice behavior. For example, Zhou et al. built route choice prediction model based on the route choice data obtained by questionnaire survey to study the law of travelers' route choice [1]. Questionnaire survey provides an effective means to obtain abundant single-mode route choice intention data, but it is difficult to deal with the complexity of multi-mode combined travel. On the one hand, the combined of various travel modes makes it necessary to set a large number of travel route choice scenarios. On the other hand, as a new travel trend, residents have not yet formed stable travel habits with multi-mode combined travel. The exploration of the law of multi-mode combined travel route choice behavior needs data that can reflect the evolution process of behaviors, which further aggravates the complexity of the scenario setting of travel route choice. In view of this, it is necessary to design more effective data collection methods for multi-mode combined travel route decision behavior research. The travel behavioral experiment carried out by Mahmasani et al. can flexibly set a variety of complex travel scenarios, and automatically form new travel scenarios by processing choice information of the subjects [2]. Therefore, compared with questionnaire survey, behavioral experiment is more suitable for studying the behavior law of multi-mode combined travel route choice including new travel modes.

Another feature of current travel route choice behavior research is that with the development of information and communication technology, more and more studies focus on the impact of travel information on travel behavior. For example, Xu found that travel information has greater impact on route choice during peak periods [3]. Liu found that excessive travel information would greatly increase the route switching behavior [4]. Liu found that travel information has a significant impact on route choice at the initial stage, and with the formation of travel habits, the influence of travel information gradually weakened [5]. With the deepening of research, some scholars have further explored the mechanism of travelers' information acceptance willingness in travel route choice. For example, Hu et al. established route choice model and found that travelers' actual travel decisions are jointly influenced by travel information acceptance willingness and decision-making habits [6, 7]. In the existing studies on route choice considering information acceptance willingness, travelers are usually considered to be homogeneous, and lack of categorical discussion about the influence of different information acceptance willingness on travel route choice behavior.

Based on this, this study takes multi-mode transport network as the research object, built experimental platform for multi-mode combined travel route choice behavior considering travel information acceptance willingness, and excavates the behavioral laws of travelers in travel route decision making based on the analysis of subjects' information acceptance characteristics, so as to provide reference for traffic control under multi-mode combined travel.

2 Information Acceptance Willingness Analysis

The Travel Information Acceptance Willingness questionnaire designed includes personal socioeconomic attributes, travel attributes and psychological latent variable attributes. According to the relevant research on information acceptance, as well as the

characteristics of multi-mode combined travel behavior, eight psychological latent variables, subjective norm, perceived usefulness, perceived ease of use, information acceptance intention and information acceptance willingness were selected under the framework of technology receptive model [8]. Information cognition was selected under the framework of micro-blog user information acceptance model [9]. Highly self-monitoring was selected to explore subjects’ strategic regulation behaviors under different travel information [10]. Time sensitivity was selected to explore the sensitivity of the subjects to time cost.

From December 22, 2023 to January 8, 2024, data on travel information acceptance willingness was collected through online questionnaire survey. A total of 340 questionnaires were distributed and recovered, and 317 valid questionnaires were obtained after excluding invalid ones, with an effective rate of 93.24%.

The classification of travelers according to their information acceptance willingness was conducted by potential profile analysis. In potential profile analysis, it is generally believed that smaller AIC, BIC, and aBIC values are better, Entropy values > 0.8 indicate classification accuracy of more than 90%, and significant LMR and BLRT values indicate that a model with K classes is superior to a model with K-1 classes. As can be seen from Table 1, AIC, BIC and aBIC indexes are all decreasing with the increase of the number of classifications, indicating that the model is constantly improving. Entropy is always > 0.8, indicating that the model has a high classification accuracy; LMR and BLRT values are not significant after the third classes, indicating that three classes is superior to four classes. Therefore, this study divided travelers into three classes according to information acceptance willingness.

Table 1. Potential profile analysis results.

Class	k	AIC	BIC	aBIC	Entropy	LMR	BLRT	Class probability
1	48	18461	18642	18490				
2	73	17788	18062	17831	0.874	0	0	0.306/0.694
3	98	17609	17977	17666	0.84	0.025	0.024	0.297/0.473/0.230
4	123	17469	17931	17541	0.871	0.494	0.497	0.151/0.480/0.158/0.211
5	148	17386	17942	17473	0.892	0.768	0.769	0.032/0.174/0.126/0.476/0.192

Class 1 has the lowest score on each latent variable of travel information acceptance willingness, class 2 has the second score, and class 3 has the highest score. Therefore, defined classes 1, 2 and 3 as low, medium and high information acceptance respectively.

3 Design of Behavioral Experimental

In order to observe the route choice laws of subjects in multi-mode combined travel route choice, this study relied on z-Tree software [11] and set up 200 rounds of travel route choice behavior experiments. Taxi, bus, subway and shared electric bicycle were

provided in the experiment. At the beginning of each month, the subjects select three travel modes to form a travel scenario and then conduct 20 rounds travel route decisions during the month. Each round of experiment simulates an actual travel route choice process. After each round of choice, subjects can get their travel cost, account balance, account ranking, as well as the travel cost of other routes included in this scenario. Subjects can refer to the information for the next round of travel route choice.

Based on the relevant research results of multi-mode combined travel route choice [12] and combined with the experimental purpose, the experimental basic transport network was set up, as shown in Fig. 1.

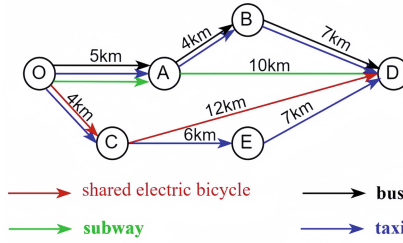


Fig. 1. Experimental basic transport network.

According to the difference of route segment's travel mode, the transport network forms a total of 13 combined travel routes. The travel route cost consists of free flow cost and congestion cost. The free flow cost refers to the passage time of the route in the case of free flow, while the congestion cost is used to describe the growth of the passage time with each additional subject choice of the route. The setting of route cost function in this study is shown in formula (1):

$$T_N = \sum_{i=1}^j n_i \times s_j + \eta_{NE} a_N \pm 2\sigma \quad (1)$$

where, T_N is the actual cost time of route N; n_i is the time required per unit kilometer of each travel mode (minutes); s_j is the distance of each route segment (km); η_{NE} is the selection coefficient of route N; a_N is the number of people who choose route N; σ is the standard deviation of the cost distribution.

In the behavioral experiment, when all subjects find the optimal decision, the transport network reaches stability, and the decision set of the subjects is called the Nash equilibrium solution. The Nash equilibrium cost solution is shown in formula (2):

$$T_{NE}(a_{NE}, \eta_{NE}) \leq \min\{T_N(a_N + 1, \eta_N)\} \quad (2)$$

where, T_{NE} is the transport network equilibrium cost under the Nash equilibrium; a_{NE} is the number of people who choose route N under the Nash equilibrium.

4 Behavior Law Analysis

4.1 Experimental Equilibrium Analysis

Firstly, in order to observe the process of system stabilization, used Sorenson Similarity Index (SSI) to analyze the dynamic evolution of transport network equilibrium. The value range is $[0, 1]$, and the closer to 1, the more the system tends to be equilibrium. The calculation method is shown in formula (3):

$$SSI_r = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N \frac{2 \min(a_{NE}, a_N)}{a_{NE} + a_N} \quad (3)$$

where, r is the number of experiment rounds; M is the starting point, in this study $M = 1$; N is the number of routes. SSI of 200 rounds of experiments calculated from the above equation is shown in Fig. 2. From the overall trend, SSI values fluctuated around 0.5 in the first 40 rounds of experiments, and reached 0.7 in the 40th round and then fluctuated around 0.7. It can be considered that subjects were in the stage of exploring road network and comparing travel route cost in the first 40 rounds, and formed their own stable decisions after the 40th round, then the system reached a stable state.

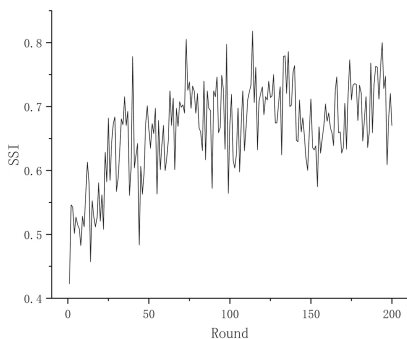


Fig. 2. Dynamic evolution of SSI.

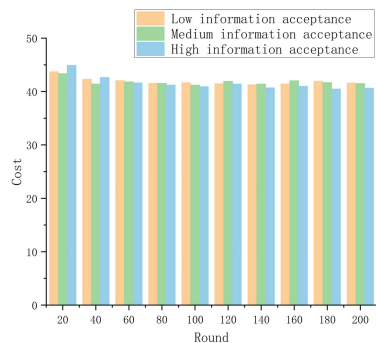


Fig. 3. Three classes subjects travel cost.

Secondly, in order to observe the specific decision-making performance of the three classes subjects in the process of equilibrium evolution, the three classes' travel cost of 200 rounds are shown in Fig. 3. The total travel costs of low, medium and high information acceptance are 117600 min, 117292 min and 116312 min respectively. From the overall trend analysis, in the first 40 rounds of experiments, the average travel cost of the three classes subjects is high, and shows a trend of gradual decline. After 40 rounds, the travel cost leveled off. Through statistical calculation, the average cost of the first 40 rounds is 45.11 min, while the average cost of the 41 to 200 rounds is 41.50 min. Compared with the first 40 rounds, the cost of the experiment after 40 rounds is closer to 39.86 min of the Nash equilibrium solution.

Based on the analysis of the above two points, it is considered that the system has reached the equilibrium state, and 40th rounds as the critical point of the transport network reaching the equilibrium.

4.2 Route Choice Law Analysis

Decision Switching Frequency Analysis

In order to observe the subjects' switching frequency of decision in the process of multiple experiments, statistics the number of three classes subjects who switched decision in each round of experiments, as shown in Fig. 4.

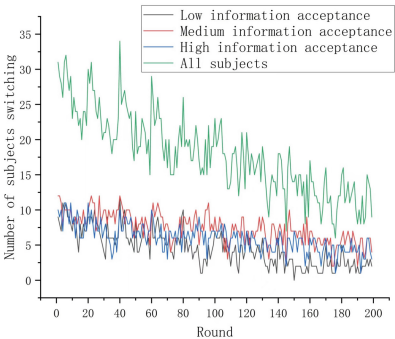


Fig. 4. Frequency of decision switching.

On the whole, subjects' decision switching frequency showed a trend of decreasing in the fluctuation, indicating that subjects gradually formed a stable strategy with the progress of the experiment. From Fig. 4, the trend of decision switching frequency of three subjects is same as the overall. Meanwhile, the decision switching frequency of the three classes subjects presents medium > high > low information acceptance.

Low information acceptance has the lowest information acceptance willingness and are not sensitive to information. After each round of the experiment, low information acceptance switch strategy lest and are most likely to maintain the original decision. Medium and high information acceptance show higher information acceptance willingness than low information acceptance, and are more inclined to switch strategies according to the travel information, showing a higher frequency of decision switch. In addition, the information acceptance willingness is not only reflected in the willingness to acceptance information feedback, but also includes the ability to process information. When a route's cost in the last round is too high, the medium information acceptance tends to change the strategy, while the high information acceptance may consider the majority subjects will switch the strategy in the next round, and the cost of the route is greatly reduced, so maintain the strategy. This kind of information processing ability and the existence of risk game psychology make the strategy switch frequency of high information acceptance lower than the medium information acceptance.

Decision Response Pattern Analysis

In order to verify the existence of game psychology and analyze its influence on route choice, this study introduced Q coefficient (Yule's Q) to identify the decision response mode of the subjects, and then explored the influence of the response mode of the subjects on route decision making.

Q coefficient is an index proposed by Yule to measure the correlation degree of two attributes, and the value range is $[-1, 1]$. In the decision-making response mode of subjects, the closer to 1 indicates that subjects tend to direct response pattern, that is, there is no game psychology; the closer to -1 indicates that subjects tend to opposite response pattern, that is, there is game psychology. The calculation formula is as follows (4):

$$Q = \frac{c_- \times s_+ - c_+ \times s_-}{c_- \times s_+ + c_+ \times s_-} \quad (4)$$

where, c_+ and c_- respectively represent the statistical number of times that the traveler's earning is higher (lower) than the equilibrium earning in the last round of decision-making and changes the strategy in the next round; s_+ and s_- respectively represent the number of statistical times that the traveler's earning is higher (lower) than the equilibrium earning in the last round of decision-making and the original strategy remains unchanged in the next round. In this study, when the system reaches equilibrium, the route equilibrium cost is 41.77, that is, the equilibrium earning is -41.77 .

By extracting the decision sequence and cost sequence of each subject, Q coefficients of the three classes of subjects were calculated, as shown in Fig. 5.

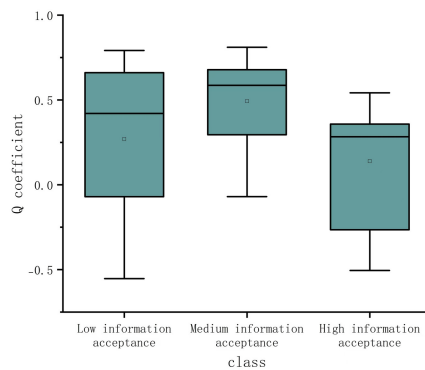


Fig. 5. Q coefficient distribution of three classes subjects.

From Fig. 5, in the multiple routing experiments, some subjects have opposite response pattern, and the direct response patterns are more than the opposite response patterns in the three classes subjects. Compared with low information acceptance, medium information acceptance is more flexible in decision-making adjustment, and are more inclined to direct response pattern, which can timely adjust strategies according to the cost of the last round of travel. High information acceptance is more inclined to the opposite response pattern, such subjects are more complex in information processing, and have game psychology, rather than immediately switch the decision. Based on the travel costs of the three classes of subjects, it can be seen that appropriate decision switching frequency and game psychology can help subjects reduce travel costs. In addition, information acceptance willingness is not only manifested by holding a positive attitude

towards information and adjusting strategies in time, but also by processing information and being able to predict risks and play games. If the travel choice behavior of travelers is predicted according to the completely rational hypothesis, the result will be inconsistent with the actual situation.

5 Conclusion

This study selected different classes of subjects to enter the multi-mode combined travel route choice behavioral experiment. Based on the experiment data deeply analyzed the route choice behavior law, the following main conclusions are drawn:

- (1) Travelers have different information acceptance willingness. According to information acceptance willingness, the subjects can be divided into three classes: low, medium and high information acceptance, and the higher the travelers' information acceptance willingness, the lower the route travel cost.
- (2) Information acceptance willingness is not only manifested by holding a positive attitude towards information, but also manifested by the ability to process information. Appropriate decision switching frequency and game psychology can help travelers reduce travel route costs.

References

1. Zhou, Y., Wang, P., Zheng, S., et al.: Modeling dynamic travel mode choices using cumulative prospect theory. *Transp. Res. Part A Policy Pract.* **179**, 103938 (2024)
2. Mahmassani, H.S., Chang, G., Herman, R.: Individual decisions and collective effects in a simulated traffic system. *Transp. Sci.* **20**(4), 258–271 (1986)
3. Xu, X., Xie, L.: Research on route choice behavior of urban rail transit passengers under induced Information. *Transp. Syst. Eng. Inf.* **22**(03), 63–73 (2022). (in Chinese)
4. Liu, S., Wang, S., Huang, G., et al.: An experimental study on the influence of charging and information on routing behavior under uncertain environment. *J. Tongji Univ. (Nat. Sci.)* **52**(02), 241–251 (2019). (in Chinese)
5. Liu, K.: Analysis of the conflict between car commuter's route choice habitual behavior and traffic information search behavior. *Sensors* **22**(12), 4382 (2022)
6. Hu, X., Qin, H., Guo, J., et al.: Study on the route-selection behavior of bus passengers based on an evolutionary game under the condition of information guidance. *Appl. Sci.* **12**(13), 6703 (2022)
7. Agrawal, S., Peeta, S.: Hybrid route choice model incorporating latent cognitive effects of real-time travel information using physiological data. *Transport. Res. F Traffic Psychol. Behav.* **81**, 223–239 (2021)
8. Ajzen, I.: The theory of planned behaviour: reactions and reflections. *Psychol. Health* **26**(9), 1113–1127 (2011)
9. Chen, J., Yu, H., Zhang, C.: Traffic Information acceptance behavior in social network environment. *J. Chongqing Jiaotong Univ. (Nat. Sci. Edn.)* **39**(12), 13–19 (2019)
10. Qu, Y., Gao, Z., Li, X.: Modeling and simulating herding behavior and information spreading process in pedestrian flow. *Transp. Syst. Eng. Inf.* **14**(05), 188–193 (2014)

11. Fischbacher, U.: Z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ. J. Econ. Sci. Assoc.* **10**(2), 171–178 (2007)
12. Li, D., Jin, C., Yang, M., et al.: Incorporating multi-level taste heterogeneity in route choice modeling: from disaggregated behavior analysis to aggregated network loading. *Travel Behav. Soc.* **19**, 36–44 (2020)



Distributed Adaptive Coordination Control for Autonomous Road Traffic at Multiple Intersections

Honghai Ji¹, Qirui Shi¹, Shida Liu¹(✉), Lingling Fan², and Li wang¹

¹ School of Electrical and Control Engineering, North China University of Technology,
Beijing 100144, China
lsdshiwo@hotmail.com

² School of Automation, Beijing Information Science and Technology University,
Beijing 100192, China

Abstract. As societal demands on traffic control increase, traditional model-based traffic control methods are showing their limitations. This paper introduces a Distributed Adaptive Coordination Control (DACC) algorithm, which integrates data-driven control and multi-agent systems, aimed at optimizing traffic signal control across multiple intersections to effectively balance queuing times. Through joint simulation experiments with SUMO and Python, and comparison with the Proportional-Integral-Derivative Coordination Control (PIDCC) algorithm, the DACC algorithm has demonstrated advantages in reducing average queuing times and improving traffic flow efficiency. The results indicate that the DACC algorithm excels in managing traffic flow variations.

Keywords: Data-Driven Control · Multi-Agent Coordination Control · Traffic Signal Optimization

1 Introduction

In the era of big data, Intelligent Transportation Systems (ITS) are undergoing a revolutionary transformation, with previously difficult-to-access vehicle, road, and environmental data now easily obtainable. The variety, scope, and accuracy of sensor data have significantly improved, and societal advancements have placed higher demands on traffic control. Traditional Model-Based Traffic Control (MBTC) [1] theories have proven to be insufficient, making Data-Driven Intelligent Transportation Systems (DDITS) [2] a focal point in this field. In recent years, data-driven control methods such as model-free adaptive control [4], iterative learning control [5], and lazy learning control [6] have received increasing attention.

Multi-agent coordination control [7] involves multiple agents working collaboratively to complete specific tasks. The study [8] presents a multi-agent coordination control algorithm that performs well at individual intersections. Another study [9] introduces a collaborative multi-agent reinforcement learning method that enables the system to overcome instability issues and reduce both the average travel time and its variability.

The paper [10] introduces a value-decomposition-based spatiotemporal graph attention multi-agent deep reinforcement learning model, significantly reducing average vehicle delays and the number of stops within controlled road networks. Additionally, the study [11] proposes a regional boundary intersections collaborative control method, leveraging the model-free data-driven advantages of multi-agent deep reinforcement learning, allowing for detailed regulation of the network's traffic flow distribution from a macro to micro perspective, thus enhancing operational efficiency. Lastly, the paper [11] introduces a multi-objective reinforcement learning model for network-wide coordinated signal control, addressing the critical limitations of data transmission in field implementations and analyzing the traffic state across the entire network rather than just examining each isolated intersection.

This paper combines data-driven control with multi-agent coordination control to propose a Distributed Adaptive Coordination Control algorithm for autonomous traffic signal control, aimed at achieving queue equilibrium across multiple intersections and directions. The main contributions of this paper are as follows:

- 1) The DACC algorithm is introduced, which effectively balances queue times under varying traffic flow conditions and can impartially handle both undersaturated and oversaturated traffic flow conditions at multiple intersections.
- 2) Joint simulations using SUMO and Python were conducted, validating the effectiveness of the algorithm in comparison with the PIDCC algorithm.

This paper is structured as follows. Section 2 introduces the multi-intersection model design. Section 3 completes the design of the controllers. In Sect. 4, we conducted joint simulations using SUMO and Python to validate the effectiveness of the algorithm. Section 5 summarizes the work of this paper and provides an outlook for future research.

2 Problem Formulation

The phase sequence at the intersection is as shown in Fig. 1: the first phase allows north-south through and right-turn traffic, the second phase allows north-south left-turn traffic, the third phase allows east-west through and right-turn traffic, and the fourth phase allows east-west left-turn traffic.

The multi-intersection setup, consisting of two standard four-phase signal-controlled intersections connected together, is shown in Fig. 2.

The model for balancing queuing at multiple intersections is as follows:

$$x_{i,m}(k+1) = x_{i,m}(k) + T_i(k)q_{i,m}(k) - s_{i,m}g_{i,m}(k) \quad (1)$$

where $x_{i,m}(k)$ represents the number of queued vehicles at intersection i in the m direction during the k cycle, $T_i(k)$ represents the cycle duration, and $q_{i,m}(k)$ is the vehicle arrival rate. $s_{i,m}(k)$ is the variable saturation flow rate affected by uncontrollable factors such as vehicles making U-turns and violations, and $g_{i,m}(k)$ represents the green phase duration.

The green times and signal cycles for all intersections must satisfy the following constraints:

$$T_i(k) = \sum_{m=1}^4 g_{i,m}(k) + t_{i,L} \quad (2)$$

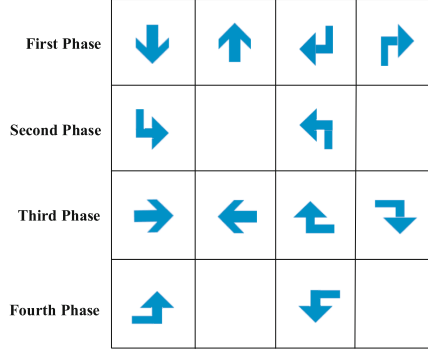


Fig. 1. Phase Sequence Diagram at the Intersection.

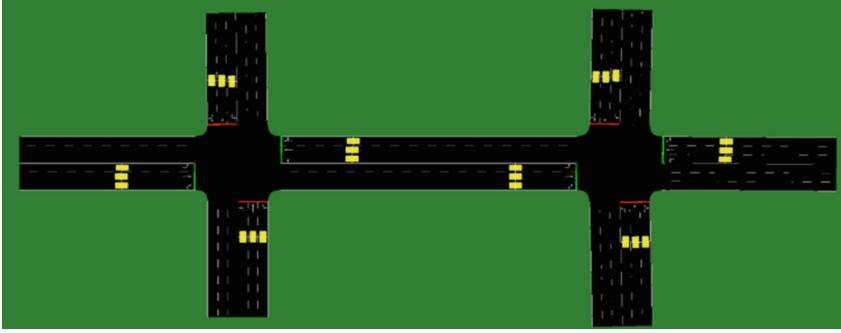


Fig. 2. Schematic Diagram of the Multi-Intersection Layout.

where $t_{i,L}$ represents the total loss time, including the yellow light duration.

The saturation level of the approach lanes at intersections is defined as follows:

$$d_{i,m}(k) = \frac{T_i(k)q_{i,m}(k)}{s_{i,m}g_{i,m}(k)} \quad (3)$$

According to the definition of saturation, under the dynamics of the intersection traffic flow system, the state is undersaturated when $d_{i,m}(k) < 1$, critically saturated when $d_{i,m}(k) = 1$, and oversaturated when $d_{i,m}(k) > 1$.

To avoid wasting green light time, it is necessary to introduce time-varying control signal cycles $T_i(k)$ and green light time constraints to ensure that the number of queued vehicles is physically meaningful and meets $x_{i,m}(k) \geq 0$. Thus, under these constraints, the green light duration becomes:

$$\bar{g}_{i,m}(k) \triangleq \text{sat}(g_{i,m}(k)) = \begin{cases} \bar{g} & \text{if } g_{i,m}(k) > \bar{g} \\ g_{i,m}(k) & \text{if } \underline{g} \leq g_{i,m}(k) \leq \bar{g} \\ \underline{g} & \text{if } g_{i,m}(k) < \underline{g} \end{cases} \quad (4)$$

where $\bar{g} = \min\{[x_{i,m} + T_i q_{i,m}]/s_{i,m}, g_{i,m,\max}\}$ and $\underline{g} = g_{i,m,\min}$.

Treat the queue length of each segment at the two intersections as an agent, with each agent containing traffic information about adjacent intersections. The topology of the multi-intersection system is shown in Fig. 3.



Fig. 3. Topology Structure of the Multi-Intersection System.

3 Control Scheme

Each segment's queue at the intersection is treated as an agent that contains traffic information about adjacent intersections. The distributed consistent coordination error for queuing in direction m at intersection i is constructed as follows:

$$e_{i,m} = \sum_{n \in N_i^m} a_{mn}^i (x_{i,n} - x_{i,m}) + \sum_{n \in N_j^m} b_{mn}^i (x_{j,n} - x_{i,m}) + c_{i,m} (x_r - x_{i,m}) \quad (5)$$

where $A = [a_{mn}]$, $m, n \in [1, M]$ represents the internal coordination adjacency matrix within the distributed signal controller network topology, and $B = [b_{mn}]$, $m, n \in [1, M]$ represents the external coordination adjacency matrix. $c_{i,m}$ denotes the connection coefficient between the queuing in direction m and the expected queue. N_i^m represents the m -th internal coordination direction at intersection i . N_j^m represents the m -th external coordination direction at intersection i , belonging to intersection j .

The design of the DACC strategy for multiple intersections is as follows:

$$\begin{aligned} \bar{u}_{i,m}(k) = \frac{p_{i,m}}{p_{i,m} + \mu} & \left\{ u_{i,m}(k-1) - c \left[\sum_{n \in N_i^i} a_{mn}^i (e_{i,m} - e_{i,n}) \right. \right. \\ & \left. \left. + \sum_{n \in N_j^j} b_{mn}^i (e_{i,m} - e_{j,n}) + c_{i,m} e_{i,m} \right] - \bar{\varphi} \hat{\theta}_{i,m}(k) e_{i,m}(k) \right\} \end{aligned} \quad (6)$$

The distributed form of the parameter learning law is as follows:

$$\hat{\theta}_{i,m}(k) = \sum_{i=1}^N \frac{1}{1 + \kappa \alpha_{i,m}} \left\{ \hat{\theta}_{i,m}(k-1) + \bar{\varphi} p_{i,m} \alpha_{i,m} e_{i,m} \left[\sum_{n \in N_m^i} a_{mn}^i (e_{i,m} - e_{i,n}) + \sum_{n \in N_m^j} b_{mn}^i (e_{i,m} - e_{j,n}) + c_{i,m} e_{i,m} \right] \right\} \quad (7)$$

4 Simulation Result

In the simulation experiment, 3000 random vehicles were modeled by modifying the arrival rate parameter p in SUMO to dynamically control the traffic flow. The arrival rate parameters were set to 1.0, 0.6, 1.4, 1.0 and the algorithm parameters for DACC were set as: $k = 0.01$, $p_{i,m} = 0.5$, $\bar{\varphi} = 0.001$, $\alpha_{i,m} = 0.1$, $c = 0.001$, $c_{i,m} = 0.025$, $\mu = 0.001$. This was compared PIDCC algorithm. The PIDCC algorithm formula is as follows:

$$u_i(k) = u_i(k-1) + K_i e_{i,m}(k) + K_p (e_{i,m}(k) - e_{i,m}(k-1)) + K_d (e_{i,m}(k) - 2e_{i,m}(k-1) + e_{i,m}(k-2)) \quad (8)$$

where K_p is the proportional gain, K_i is the integral gain, K_d is the differential gain, and $e_{i,m}(k)$ is the coordination error for cycle k . The simulation parameters for PIDCC are set as $K_p = 0.5$, $K_i = 0.2$, $K_d = 0.1$.

The performance of DACC at intersection i is shown in Fig. 4. In the evaluation of multi-intersection traffic network operation, as shown in Fig. 5, the average queuing time per 300 vehicles is calculated, with a total of 10 counts, showing that DACC consistently outperforms the PIDCC algorithm.

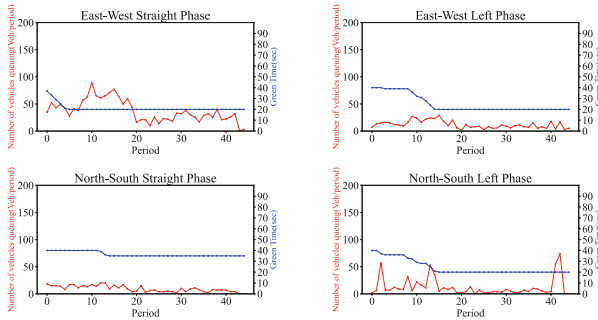


Fig. 4. Changes in Queued Vehicles and Green Light Duration at Intersection i .

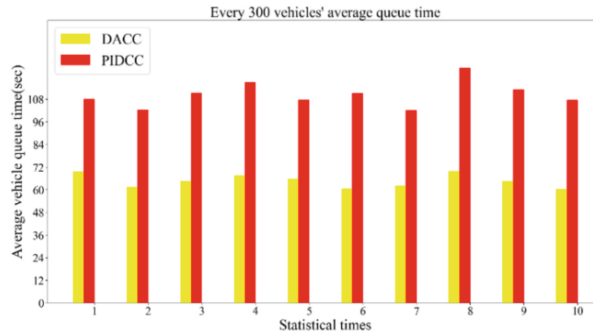


Fig. 5. Bar Chart Comparing Average Queuing Times of Vehicles.

5 Conclusions

This paper addresses the issues of traffic overflow and green time loss in complex traffic networks within autonomous traffic control systems. By integrating the concept of queue balancing with multi-agent collaboration theory, a systematic model for multiple intersections has been developed. The DACC algorithm is proposed, which exhibits effective control under varying traffic conditions, achieving adaptive traffic signal control in an autonomous traffic environment.

Acknowledgments. This work is supported by Beijing Natural Science Foundation under Grant L241054, National Key Research and Development Plan under Grant 2023YFB4704404, R&D Program of Beijing Municipal Education Commission (KM202310009010).

References

1. Papageorgiou, M., Diakaki, C., Dinopoulou, V., et al.: Review of road traffic control strategies. *Proc. IEEE* **91**(12), 2043–2067 (2003)
2. Zhang, J., Wang, F.Y., Wang, K., et al.: Data-driven intelligent transportation systems: a survey. *IEEE Trans. Intell. Transp. Syst.* **12**(4), 1624–1639 (2011)
3. Hou, Z., Jin, S.: *Model Free Adaptive Control*, 2nd edn. CRC Press, Boca Raton (2013)
4. Hou, Z., Huang, W.: The model-free learning adaptive control of a class of SISO nonlinear systems. In: *Proceedings of the 1997 American Control Conference* (Cat. No. 97CH36041), vol. 1, pp. 343–344. IEEE (1997)
5. Xu, J.X., Tan, Y., Lee, T.H.: Iterative learning control design based on composite energy function with input saturation. *Automatica* **40**(8), 1371–1377 (2004)
6. Hou, Z., Liu, S., Tian, T.: Lazy-learning-based data-driven model-free adaptive predictive control for a class of discrete-time nonlinear systems. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(8), 1914–1928 (2016)
7. Lewis, F.L., Zhang, H., Hengster-Movric, K., et al.: *Cooperative Control of Multi-Agent Systems: Optimal and Adaptive Design Approaches*. Springer, London (2013). <https://doi.org/10.1007/978-1-4471-5574-4>
8. Zhang, H., Liu, X., Ji, H., et al.: Multi-agent-based data-driven distributed adaptive cooperative control in urban traffic signal timing. *Energies* **12**(7), 1402 (2019)

9. Liu, J., Qin, S., Su, M., et al.: Multiple intersections traffic signal control based on cooperative multi-agent reinforcement learning. *Inf. Sci.* **647**, 119484 (2023)
10. Bie, Y., Ji, Y., Ma, D.: Multi-agent Deep Reinforcement Learning collaborative Traffic Signal Control method considering intersection heterogeneity. *Transp. Res. Part C Emerg. Technol.* **164**, 104663 (2024)
11. Kang, L., Huang, H., Lu, W., et al.: Optimizing gate control coordination signal for urban traffic network boundaries using multi-agent deep reinforcement learning. *Expert Syst. Appl.* **255**, 124627 (2024)



Optimization Method Based on Improved PointNet++ Network Parameters

Lei Kou¹, Hongzheng Zhao¹, Ruiyu Tang¹, Junsheng Su², Xuqiang Zhao², Wei Yan¹,
and Feng Guo¹(✉)

¹ School of Qilu Transportation, Shandong University, Jinan 250061, China
fengg@sdu.edu.cn

² China Railway Sixth Bureau Group Traffic Engineering Branch, Beijing 100070, China

Abstract. PointNet ++ suffers from the problems that the network learning rate cannot be dynamically adjusted, poor robustness to noise and outliers, insufficient local feature extraction and lack of capture of global contextual feature information in the feature learning process, and insufficient efficiency of feature transfer. This paper aims to improve the network for the above problems. The main improvement includes three major parts: training strategy, feature learning, and feature propagation, and then the experimental environment and evaluation indexes carry out separate ablation experiments on these three parts to illustrate the enhancement effect on the network, the enhancement effect on the network segmentation effect after the combination of all the modules, and finally carry out the comparison between before and after the improvement of the network. The results show the improvement effect is obvious.

Keywords: Transmission line · galloping · adaptive · vibration characteristics

1 Introduction

The tunnel is the “throat” of the transport network and its operation and management are complicated. As lifespan increases, some tunnels exhibit varying degrees of disease. An efficient method of detecting cracks in tunnel lining can provide strong technical support to the maintenance department, improve the quality of tunnel maintenance, and effectively extend the service life of tunnels [1]. A 3D point cloud is a collection of data points represented graphically in 3D Euclidean space. In recent years, almost every application area based on remote sensing data processing has started to use deep learning as a key data processing method. The current research focus has shifted from scene-wide classification of 3D LiDAR point clouds to semantic classification or segmentation [2]. The application of deep learning methods for point cloud semantic segmentation has yielded significant results. These deep learning techniques can be categorized into two main groups based on the input data. The goal of point-based semantic segmentation methods is to group points in a 3D point cloud based on their semantics. Semantic segmentation methods can be categorized into indirect and direct methods (i.e., point-based methods). Indirect point-based methods can be further categorized into projection and voxel-based techniques.

Projection-based approaches focus on converting a 3D point cloud into a 2D image using existing techniques that involve projecting the point cloud onto a plane, and then processing the projected image using existing 2D CNNs to determine category labels. The main mainstream projection techniques include multi-view projection [3], bird's eye view projection [4] and spherical projection [5].

Lawin et al. [6] proposed an alternative method for semantic segmentation of 3D point clouds, which projects a point cloud onto a collection of resulting 2D images. The corresponding images are then fed into a multi-stream 2D-CNN architecture for semantic segmentation, overcoming the limitations of 3D-CNN. Boulch et al. [7] developed a model based on generating a 2D view of a 3D scene called the SnapNet model. Bieder et al. [8] proposed a semantic mesh map estimation method based on sparse input data, and built method for generating a dense semantic base instance. Liong et al. [9] proposed a multi-view fusion network (AMVNet) with two decoupled networks so that complexity points can be learned and have more flexibility. In order to provide a complete backprojection of the spatial structure, the multiview projection technique must finely capture the entire point cloud. However multi-view images lose important spatial information from real-world scenes [10]. In order to capture the scene more comprehensively, it is necessary to consider factors such as the number of views, viewing distance, and virtual camera angle when capturing the image [11]. Aksoy et al. [12] proposed a semantic segmentation network for point clouds called SalsaNet, which employs an encoder-decoder architecture with a ResNet module for segmenting the point cloud data projected onto a bird's eye view. Wu et al. [13] used spherical projection to convert 3D point clouds to 2D images and proposed an end-to-end channel called SqueezeSeg which utilizes a convolutional neural network to output labeled point-by-point data.

A fully 3D convolutional neural network based on a 3D point cloud labeling system was developed in which no a priori scene knowledge for segmentation is required [14]. Dai et al. [15] introduced ScanComplete, a data-driven approach that predicts a comprehensive 3D model that can increase with the scene size to manages the cubic growth of data while dealing with large scenes with various spatial scales. Rethage et al. [16] developed a fully convolutional network design for large voxel-based point clouds where both structured and unstructured data are applicable. Han et al. [17] proposed a full-cycle deformation point cloud processing method for mountain tunnels ICP algorithm, which cuts the tunnel point cloud into a single circular point cloud utilizing the ICP algorithm to accurately align the two-phase point cloud, and after alignment, by calculating the deformation at the highest point, it can obtain the Deformation and settlement at the top of the tunnel.

At present, there are fewer studies on semantic segmentation of point clouds in the field of construction tunnels, and similar studies are mostly focused on subway tunnels, where the main reason is that it is difficult to obtain a large amount of point cloud data of construction tunnels, which leads to the inability to construct relevant datasets that can be used for network learning. On the other hand, the pitch-arch step is a presence that cannot be ignored in tunnel construction safety; however, the existing detection methods are fewer and less accurate. Therefore, this paper carries out a study to improve the key structure of the correlation network to segment the point cloud at construction.

2 Experimental Settings and Assessment Indicators

2.1 Experimental Settings

The experiments in this paper run on a virtual environment containing Python, CUDA, Pytorch, and other tools built based on Anaconda on a Windows system. The hardware configuration is as follows: CPU is Intel Core i5-12490F 3.0 Ghz, 6 cores, and 12 threads, GPU is NVIDIA GeForce RTX 3060 with 12 GB of video memory and 32 GB of RAM. The software configuration is as follows: the experimental environment is the virtual environment built by Anaconda on Windows, Python version is 3.10, and the Pytorch version is 2.0. 3.10, Pytorch version 2.0, CUDA version 12.2.

2.2 Assessment Indicators

The network performance evaluation metrics common in the field of semantic segmentation of 2D images are also applicable to semantic segmentation tasks in the point cloud domain, such as Overall Accuracy (OA) and Intersection over Union (IoU).

The overall accuracy OA represents the ratio of the number of correctly classified points to the total number of points in the point cloud data, and its mathematical expression is:

$$OA = \frac{\sum_i^k P_{ii}}{\sum_{i=0}^k \sum_{j=0}^k P_{ij}} \quad (1)$$

Although the OA value can characterize the overall segmentation accuracy better, for point cloud data with an imbalanced number of categories, the OA value is more affected by the category with a high number of points and cannot characterize the classification accuracy of each category well.

The intersection-union ratio IoU can well characterize the classification accuracy of each category, Intersection represents the intersection of the point set predicted to be a category and the real point set of the category, Union represents the concatenation of the two-point sets, and IoU is the ratio of the number of points in the Intersection to the number of points in the Union, i.e. = Intersection/Union.

2.3 Improvement Methods and Improvement Effects

The network structure diagram of the improved point cloud semantic segmentation neural network based on PointNet++ proposed in this paper is shown in Fig. 1, which consists of three main parts: improved training strategy, improved feature learning and improved feature propagation. Among them, the improved training strategy includes more advanced scheduler, optimizer, and Loss function, and this paper analyzes in detail how the hyperparameters affect the network segmentation in Chapter 4, and the feature learning part of the network includes three parts, namely, independent feature learning module, spatial attention module, and InvResMLP module. For each point cloud input into the network, its original features include D-dimensional spatial coordinates ($D = 3$) and C-dimensional additional features consisting of spatial normals and coordinates (C

spatial attention mechanism can gradually extract higher-level and more comprehensive feature representations, so as to better express the structural and semantic information of the data, and the attention mechanism between layers can enhance the expressive ability of the network, so as to make it better adapted to the complex tunnel point cloud data. The stacking of multiple SA_Layer layers enhances the extraction of global contextual information and also enhances the capture of local feature information. In the SA_Layer layer, the input feature is x_i , which contains N elements, and the following three representations are obtained for each element by a 1×1 convolution operation, Query(q_i), Key(k_i), Value(v_i). These representations can be obtained by the following equations.

$$q_i = W_q \cdot x_i \quad (3)$$

$$k_i = W_k \cdot x_i \quad (4)$$

$$v_i = W_v \cdot x_i \quad (5)$$

Here: W_q, W_k, W_v is the weight matrix, next, the similarity between Query and Key is computed to obtain the attention weight a_{ij} , which indicates the degree of association between elements x_i and x_j .

$$a_{ij} = q_i \cdot k_j / \sqrt{d} \quad (6)$$

Here: $\sqrt{d}$ is the dimension of the Query or Key (whose square root is usually taken as the scaling factor). The similarity is normalized by the Softmax function to obtain the attention weight e_{ij} .

$$e_{ij} = \text{Softmax}(a_{ij}) = \frac{\exp(a_{ij})}{\sum_{j=1}^N \exp(a_{ij})} \quad (7)$$

Finally, the value of each element is weighted and summed using the attention weights to obtain the output representation of the attention mechanism y_i .

$$y_i = \sum_{j=1}^N e_{ij} \cdot v_j \quad (8)$$

The whole process can be described as obtaining the attention weights by calculating the similarity between Query and Key, and then applying these weights to Value and weighting and summing them to obtain the output representation. The spatial attention module in the network of this paper is shown in Fig. 2.

An Inverted Residuals MLP (InvResMLP) layer is inserted after the last three SetAbstractions to further capture feature information and reduce the loss of feature information. This block is used to enhance the performance and representation of the multilayer perceptron. The InvResMLP module reverses this order by first lifting the dimensionality, then applying a Depthwise Separable Convolution, and finally lifting the dimensionality back to the original one. The process consists of several key layers: the expansion layer, the Depthwise Separable Convolution layer, and the projection layer. In the expansion

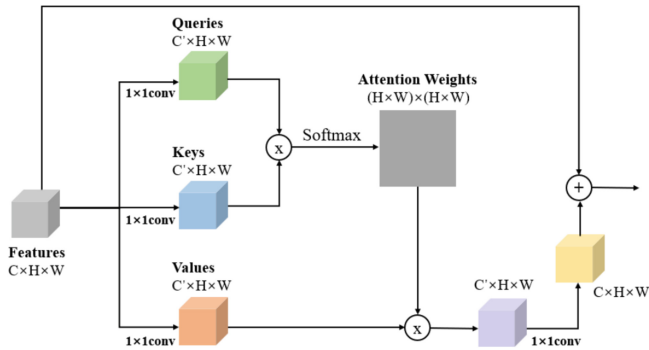


Fig. 2. Diagram of spatial attention module principle.

layer, 1×1 convolution is used to expand the number of input channels, increase the feature dimension, and improve the expressive ability of the network; in the depth convolution layer, depthwise separable convolution is used to realize the convolution operation by separating spatial convolution and channel-by-channel convolution to reduce the number of parameters and computation; finally, the number of channels is restored to the original number of output channels by 1×1 convolution in the projection layer.

The depth separable convolution layer has an important role, which contains two steps: depth convolution and point-by-point convolution. Deep convolution performs a separate convolution operation for each channel of the input, but instead of using a separate convolution kernel for each channel as in a traditional convolution operation, a single convolution kernel is applied to each channel of the input, and this convolution kernel is slid over the channel dimensions to process the features of each channel separately. Deep convolution produces multiple output channels, each corresponding to the convolution result of one input channel. Point-by-point convolution is a convolution operation using a 1×1 convolution kernel that integrates and combines multiple channel feature maps produced by deep convolution, a step that linearly combines features between channels through a series of 1×1 convolutions. The point-by-point convolution operation is typically used to reduce the number of channels and reduce the dimensionality of the features to reduce the computational effort. The InvResMLP module and the depth-separable convolutional layer in it are shown in Fig. 3.

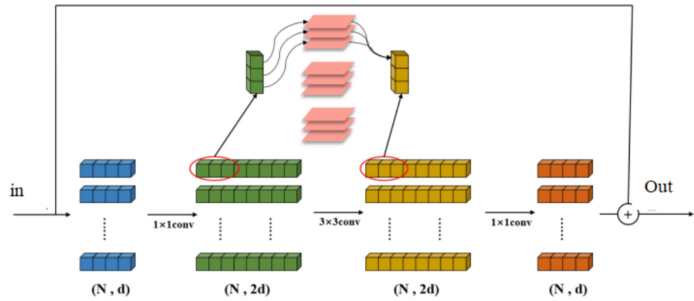


Fig. 3. D InvResMLP module structure diagram.

3 Experimental Design

3.1 Acquisition of Data Sets

As In a high-speed project in Shandong, a construction tunnel to carry out a number of field tests. The equipment used in the field test is the scanning device Hovermap.

In order to ensure that the spatial coordinates of the data obtained from the handheld laser scanning can be compatible with the geodetic coordinates of the real space, it is necessary to carry out control point measurements in the tunnel before data collection to ensure the subsequent alignment of the point cloud data. The following two methods are used to select the control points for on-site data collection: the prism target and the cross mark, the prism target is set on both sides of the secondary lining of the newest section of the tunnel, and the cross mark is set at the mouth of the tunnel, and the geodetic coordinates of the control points required for data collection are obtained from the existing control points at the construction site by pointing measurements.

The handheld device needs to move inside the tunnel for acquisition, by setting up the acquisition origin at the mouth of the tunnel going around the tunnel along the edge of the tunnel for a week, and then returning to the acquisition origin to realize the trajectory closure, and in the process of moving to realize the scanning of the sheltered part, and the comprehensive results and analysis of the above results in the tunnel scenario, the data obtained by the handheld scanning device has a more complete structure. Therefore, to guarantee the comprehensiveness and completeness of each structure in the construction of the data set, this paper adopts the handheld laser scanning equipment to collect the field data of the tunnel, and the field data collection is shown in Fig. 4.

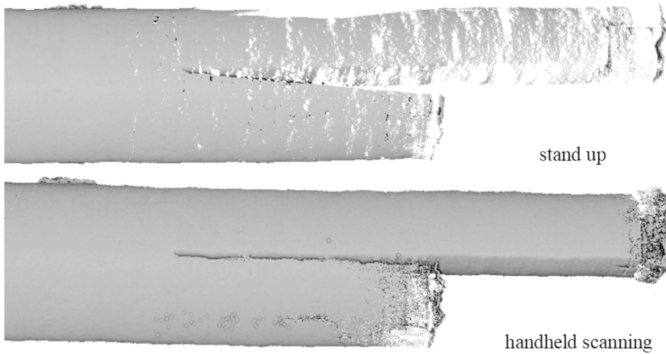


Fig. 4. Comparison of results between and handheld scanning.

3.2 Comparison Networks

In order to illustrate the best segmentation effect of the improved network, this paper sets up algorithm comparison experiments, and the experimental settings of npoints, batchsize, and Epoch are used with the same parameters of 5000, 12, and 50, respectively,

Table 1. Algorithm Comparison.

Methods	Back-arch	palm face	Back-arch trestle	Mid	Trolley	Initial Support	Pipeline	Ground	mIoU	OA
PointNet++	91.9	73.6	72.3	64.9	75.1	95.4	79.2	88.4	80.10	90.08
Ours	97.0	89.9	94.9	91.2	93.6	99.0	94.6	95	94.38	98.23

and the segmentation performance of the network before and after the improvement is shown in Table 1.

Figures 5 and 6 show the segmentation results of PointNet++, the network in this paper, and the IoU of each category, respectively. The networks before and after the improvement are more accurate in recognizing the first branch, and the IoU of the first branch category reaches 99% after the improvement. For the two categories of elevated arch trestle bridge and middle and next door, the segmentation effect is most improved, and the IoU is improved by 22.6% and 26.3% respectively, and the IoU is improved by more than 20%. The improvement of segmentation effect for the elevated arch trestle and the middle and next door indicates that the proportion of segmentation error of neighboring structures decreases, so the IoU of the categories of elevated arch, palm face and cart are also improved, the recognition effect of the elevated arch category before the algorithm is improved is better, and after the improvement, the IoU reaches 97%, the IoU of the palm face and the cart is also improved more greatly, respectively, by 16.3% and 18.5%. The segmentation results of the two categories of pipeline and ground also reach a high level of 94.6% and 95.0%, respectively, which are 15.4% and 6.6% higher than that before improvement. The above analysis shows that the segmentation results of the improved algorithm in all categories have been greatly improved.

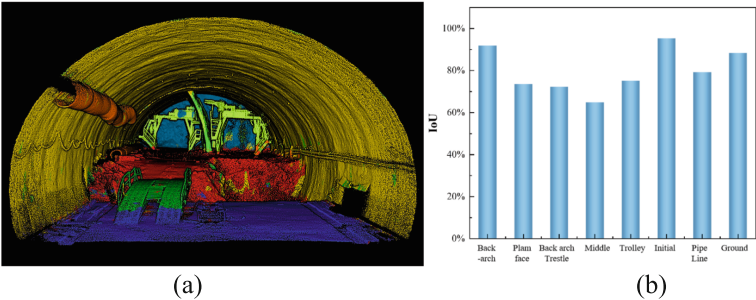


Fig. 5. PointNet++ segmentation results (a) visualization effect (b) IoU of each category.

By comparing and analyzing the results of the network segmentation before and after the improvement, the improved network has a substantial increase in the IoU and final mIoU for each category, and the improved network has an increase in the mIoU and OA of 14.28% and 8.15%, respectively, compared to the original network.

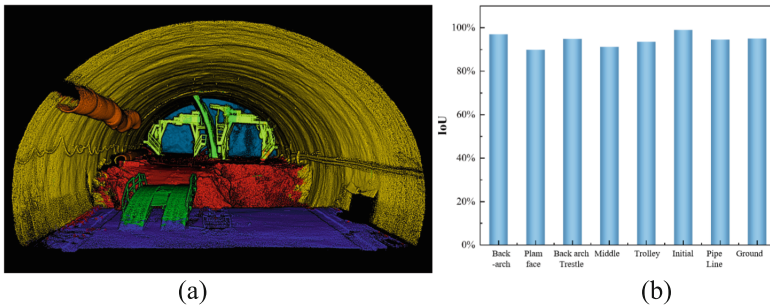


Fig. 6. Ours network segmentation results (a) Visualization effect (b) IoU of each category.

4 Conclusions

In this paper, we propose a deep learning neural network architecture with an attention mechanism and an inverted residual MLP, which employs AdanW, Cosine Decay, and a loss function with label smoothing to enable the network to adaptively reduce the learning rate as well as to mitigate the model's sensitivity to the label noise in the training data; the MLP is enhanced by integrating the attention mechanism with the InvResMLP into the network loop, which increases the capture of inter-neighborhood feature information and long-distance semantic information thereby enhancing the segmentation performance of the network; using Kd-Tree and Gaussian kernel functions for point cloud up-sampling, the use of the channel attention mechanism during feature fusion enhances the model's attention to the input features. The most significant contributors to improving network performance were identified through ablation experiments to be the spatial attention layer and the inverted residual MLP layer, with the network mIoU and OA improved by 14.28% and 8.15%, respectively, when this module was used alone.

Subsequent improvement work can propose a new way of acquiring samples, e.g., the number of points per sample remains the same but there is no restriction on the spatial size of the acquired samples, or a module to reduce the impact of duplicate points on feature learning, e.g., labeling duplicate-point related features to reduce the weight of such features during feature learning.

References

1. Hou, L.L., Zhang, Q., Du, Y.L.: Width estimation of hidden cracks in tunnel lining based on time-frequency analysis of GPR data and back propagation neural network optimized by genetic algorithm. *Automat. Constr.* **162**, 105394 (2024). <https://doi.org/10.1016/j.autcon.2024.105394>
2. Diab, A., Kashef, R., Shaker, A.: Deep learning for LiDAR point cloud classification in remote sensing. *Sensors* **22**(20), 7868 (2022)
3. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 945–953 (2015)

4. Beltrán, J., Guindel, C., Moreno, F.M., Cruzado, D., García, F., De La Escalera, A.: BirdNet: a 3D object detection framework from LiDAR information. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2018, pp. 3517–3523 (2018)
5. Cho, S., Kim, C., Park, J., Sunwoo, M., Jo, K.: Semantic point cloud mapping of LiDAR based on probabilistic uncertainty modeling for autonomous driving. *Sensors* **20**(20), 5900 (2020)
6. Lawin, F.J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F.S., Felsberg, M.: Deep projective 3D semantic segmentation. In: Felsberg, M., Heyden, A., Krüger, N. (eds.) CAIP 2017. LNCS, vol. 10424, pp. 95–107. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64689-3_8
7. Boulch, A., Guerry, J., Le Saux, B., Audebert, N.: SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Comput. Graph.* **71**, 189–198 (2018)
8. Bieder, F., Wirges, S., Janosovits, J., Richter, S., Wang, Z., Stiller, C.: Exploiting multi-layer grid maps for surround-view semantic segmentation of sparse LiDAR Data. In: IEEE Intelligent Vehicles Symposium (IV), pp. 1892–1898 (2020)
9. Liong, V.E., Nguyen, T.N.T., Widjaja, S., Sharma, D., Chong, Z.J.: AMVNet: assertion-based multi-view fusion network for Lidar semantic segmentation. *Computer Vision*, [ArXiv:2012.04934](https://arxiv.org/abs/2012.04934) (2020)
10. He, Y., et al.: Deep Learning Based 3D Segmentation: A Survey (2021)
11. Robert, D., Vallet, B., Landrieu, L.: Learning multi-view aggregation in the wild for large-scale 3D semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5565–5574 (2022)
12. Aksoy, E.E., Baci, S., Cavdar, S.: SalsaNet: fast road and vehicle segmentation in LiDAR point clouds for autonomous driving. In: IEEE Intelligent Vehicles Symposium, pp. 926–932 (2020)
13. Wu, B., Wan, A., Yue, X., Keutzer, K.: SqueezeSeg: convolutional neural nets with recurrent crf for real-time road-object segmentation from 3D LiDAR point cloud. In: IEEE International Conference on Robotics and Automation, pp. 1887–1893 (2018)
14. Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S.: SEG cloud: semantic segmentation of 3D point clouds. In: 2017 International Conference on 3D Vision, pp. 537–547 (2017)
15. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Niebner, M.: Scan complete: large-scale scene completion and semantic segmentation for 3D scans. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4578–4587 (2018)
16. Rethage, D., Wald, J., Sturm, J., Navab, N., Tombari, F.: Fully-convolutional point networks for large-scale point clouds. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 625–640. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_37
17. Han, Z., Wang, Y., Xie, X., Li, P., Tang, G.: A 3D laser scanning mountain tunnel processing method based on ICP algorithm. In: Fourth International Conference on Geology Mapping, and Remote Sensing (2023)



Research on Coverage Path Planning of Unmanned Surface Vehicle Swarm for Maritime Search and Rescue

Lootha Loserian Muriatoy¹, Gaihui Wang², Yixuan Dong¹, Dingxiang Sun³,
Zhiqiu Huang⁴, and Haoliang Wang¹ (✉)

¹ Marine Engineering College, Dalian Maritime University, Dalian, China
{yif, haoliang.wang12}@dlmu.edu.cn

² Changjing Survey Planning, Design and Research Co., Ltd., Wuhan, China
wanggaihui@cjwsjy.com.cn

³ Zhejiang Dahua Technology Co., Ltd., Hangzhou, China

⁴ Shenzhen Maritime Safety Administration, Shenzhen, China

Abstract. This paper studies a path planning problem of unmanned surface vehicle (USV) swarm for maritime search and rescue. Firstly, an environmental model suitable for the path planning of USV swarm is established by using a grid method. Secondly, A task area is divided by using a K-means algorithm to convert the coverage path planning problem into a traveling salesman problem, which is then solved using an artificial bee colony algorithm. A coverage path planning method for maritime search and rescue is designed based on the K-means algorithm, artificial bee colony algorithm, and the traveling salesman problem. Finally, simulation results demonstrate the effectiveness of the proposed coverage path planning algorithm.

Keywords: Maritime Search and Rescue · Unmanned Surface Vehicle · Coverage Path Planning · Artificial Bee Colony Algorithm

1 Introduction

With the development and utilization of marine resources and the rapid development of maritime transportation, maritime operations are becoming much busier, and the probability of maritime traffic safety accidents is also increasing. Maritime search and rescue refers to the activities of organizing and implementing a series of rescue actions in the maritime environment when accidents, disasters, or emergencies occur. In recent years, unmanned surface vehicles (USVs) are gradually applied to maritime search and rescue [1–3].

According to whether the target point position is clear, it can be divided into known target path planning and unknown target path planning. For the known target path planning, it is usually necessary to plan an optimal safe path from the starting point to the target point based on optimization algorithms. In order to improve the path planning capabilities of for single USV and unmanned underwater vehicle, deep reinforcement learning algorithms, improved rapid expansion random tree algorithms, and artificial

potential field algorithms are proposed in literature [3–8]. Literature [9–11] studies the path planning problem of USV swarm, reasonable optimization objectives have been set to ensure the maximization of task efficiency. For unknown target path planning, in order to search for unknown targets in the specified area as much as possible, a coverage path planning can be used. While there is little research referring to USV coverage path planning for maritime search and rescue.

To address this issue, this paper proposes a coverage path planning algorithm for maritime search and rescue based on the artificial bee colony algorithm and the multiple traveling salesman problem, aiming to achieve effective path planning for the search and rescue area.

2 Coverage Path Planning Algorithm

This section determines the research objectives of USV coverage path planning, and proposes an overall idea by using the K-Means algorithm for task allocation, converting the coverage path planning into a traveling salesman problem, and solving it with the artificial bee colony algorithm.

2.1 Grid Method

For the research on the coverage path planning of USV swarms, factors such as the number of USVs, the initial position of USVs, and whether there are obstacles in the sea area need to be considered. The grid method is a commonly used environmental modeling method, which is particularly suitable for the coverage path planning. This method divides the environment into regular grids, each grid cell represents an area in the environment. Each cell can be represented as a passable area or an impassable area (including obstacle areas, controlled sea areas, and marine protected areas), thereby simplifying the representation of obstacles and path finding process.

To better express the environmental model based on the grid method, the following five definitions are needed:

1. $K = \{(x, y) | x, y \text{ is a point in the search and rescue area}\}$. 2. The reachable area $S_\phi = \{(x, y) | ocu(x, y) = 1\} \subseteq K$, where $ocu(x, y)$ is the grid value, it has:

$$ocu(x, y) = \begin{cases} 1, \text{ Passable Area} \\ 0, \text{ Impassable Area} \end{cases} \quad (1)$$

3. The unreachable area, $S_\Delta = \{(x, y) | ocu(x, y) = 0\} \subseteq K$, obviously, $S_\phi \cap S_\Delta = \emptyset$ and $S_\phi \cup S_\Delta = K$. 4. After the USV swarm starts the search and rescue mission, the grids passed are marked as $ocu(x, y) = 2$. 5. The maximum length of the environmental area established is L_{Bor} , the maximum width is W_{Bor} , and the size of each grid (length and width) is D_{Squ} , thus the size of the environmental map is $(L_{Bor}/D_{Squ}) \times (W_{Bor}/D_{Squ})$.

2.2 Task Allocation Based on K-Means Algorithm

When multiple USVs carry out coverage path search, they may repeat the search of the same area due to path design flaws, which greatly reduces efficiency and may even cause

collisions among USVs. Therefore, it is necessary to divide the search and rescue area and assign tasks to each USV.

The K-Means algorithm is an iterative clustering analysis algorithm, which has the advantages of good clustering effect and fast convergence, and can perform unsupervised learning without prior data marking. In the K-Means algorithm, the number of clusters K is a hyperparameter that needs to be input manually. The core task of K-Means is to find the K optimal centroids based on the set K , and assigns the data closest to these centroids to the clusters represented by these centroids.

The K-Means algorithm can be designed as follows: 1. First, randomly select K points from the samples as cluster centers; 2. Calculate the distance between the other samples in the sample and these K cluster centers, and assign these samples to the category of the nearest cluster center; 3. Calculate the average value for each category to find the new cluster centroid; 4. Compare the K cluster centroids calculated this time with the previous calculation, if the cluster centroids change, go to step 2, otherwise go to step 5; 5. When the centroids do not change, stop and show the clustering results.

2.3 Artificial Bee Colony Algorithm

In the standard artificial bee colony algorithm, initialization is required at first. A food source position represents a feasible solution to the problem, and an employed bee corresponds to the food of a food source. Therefore, the number of employed bees (i.e., population size: SN) should be equal to the number of food sources around the hive. The population is randomly initialized as follows:

$$x_{ij} = x_{\min} + \text{rand}(0, 1)(x_{\max} - x_{\min}) \quad (2)$$

where $i = 1, 2, 3 \dots n$, represents the i th solution, $j = 1, 2, 3 \dots n$, represents the j th dimension, and x_{ij} is the current position. When the three types of bees start working and passing on information, as the swarm begins to search for and develop new food sources, new positions will emerge, and with the transmission of information, the quality of the food sources will become better and better, achieving maximum benefit. At this time, the convergence of the result position becomes more and more convergent.

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \quad (3)$$

where $k \neq i$.

During the search process, employed bees will compare the newly discovered food sources with known food sources, and choose the food sources with better quality or closer to the hive as the ideal food sources. After the search task, the information is passed to the onlooker bees. The onlooker bees also compare the information passed by the employed bees with the known information in the same way, and finally find the best food source.

The probability of onlooker bees choosing the best food source is given by:

$$p_i = f_i / \sum_{i=1}^n f_i \quad (4)$$

where i indicates the selection of the i th food source.

2.4 Traveling Salesman Problem

The traveling salesman problem (TSP) can be described as follows: A salesman needs to complete sales tasks in several cities, and it is needed to choose the shortest route to pass through all other cities and return to the starting point. From the perspective of graph theory, the problem is to calculate a Hamiltonian circuit with the smallest weight from all circuits of a weighted undirected graph. The TSP can be divided into symmetric and asymmetric problems. This paper studies the symmetric TSP, which needs to meet three requirements: the traveler is unique; the round-trip distance between any two adjacent cities is the same; the traveler starts from the starting point and visits each city with the shortest path.

Let $c_{ij}(i, j \in V)$ represent the distance between cities i and j , and $c_{ij}(i, j \in V) = c_{ji}(i, j \in V)$, which means the distance is symmetric. Given a weighted graph $G = (V, E)$, where $V = \{0, 1, 2, \dots, n\}$ represents the set of vertices and E represents the set of edges. The objective function is defined as follows:

$$\min f = \sum_{i=1}^n \sum_{j=1}^n c_{ij} x_{ij} \quad (5)$$

$$x_{ij} = \begin{cases} 1, & i, j \in V \\ 0, & i, j \notin V \end{cases} \quad (6)$$

$$\sum_{i=1}^n x_{ij} = 1, i = 1, 2, \dots, n \quad (7)$$

$$\sum_{j=1}^n x_{ij} = 1, j = 1, 2, \dots, n \quad (8)$$

$$\sum_{i \in S} \sum_{j \in S} x_{ij} \leq |Q| - 1, 2 \leq |Q| \leq n - 2, Q \subset \{1, 2, \dots, n\} \quad (9)$$

In the above mathematical model, Eq. (5) is the objective function, and it can be used to calculate the minimum distance the traveler passes through all cities; Eq. (6) depends on whether the traveler has visited the city, taking the value of 1 if visited, and 0 otherwise. Equations (7) and (8) stipulate that each city will only be passed once, with values set to make this determination; Eq. (9) is used to prevent the formation of cycles in the true subset of traveler cities. Ultimately, the minimum Hamiltonian value that meets all prerequisites is taken as the optimal solution.

2.5 Full Coverage Path Planning Design

Through the above description, the artificial bee colony algorithm can be used to solve the TSP: 1. Number the target cities and calculate the path distances between all target points; 2. Set parameters such as the distance matrix, the number of swarms, and the maximum number of iterations to obtain the initial solution of the artificial algorithm;

3. Employed bees carry out search work; 4. Record the data of the target points; 5. Calculate the probability of the target points being selected using Eq. (4); 6. Scout bees search for the selected target points; 7. Compare the fitness values of the two points and choose the better new target point; 8. Generate the optimal target point through Eq. (3); 9. Determine whether the maximum number of iterations has been reached, if not, start the planning again from step 3, if it is reached, show the corresponding solution.

The above method can transform the coverage path planning problem into a multi-traveling salesman problem for solution. The target area is gridded, the sea area is marked as “target points”, and the USV swarm is used as the “bee swarm”. The target sea area is allocated to specific USVs using the K-Means algorithm, and the coverage path is planned using the artificial bee colony algorithm. The general path planning for the TSP is used to simulate the path planning for the USV for maritime search and rescue. The artificial bee colony algorithm has the advantages of easy implementation and simple comparison of advantages and disadvantages, to reasonably plan the best area coverage path.

2.6 Algorithm Design for Coverage Path Planning

There are several steps to achieve the algorithm design for coverage path planning. 1. Assign a number to each city and calculate the distance between all cities; 2. According to the total number of cities, provide the required distance matrix; 3. Initialize the parameters needed during the algorithm execution, including the number of swarms, the maximum number of iterations, the dimension, and initialize the swarm to generate the initial solution; 4. The leading bee completes the local target point neighborhood search task; 5. Solve and record the fitness value corresponding to each local target point; 6. Calculate the probability of the local target point being selected according to Eq. 4; 7. The following bee searches near the selected local target point; 8. Compare the fitness values of the new and old local target points to obtain the optimal target point; 9. Determine whether the number of searches has reached the specified value. If it has reached while has not been updated, the leading bee is turned into a scout bee; otherwise, proceed to Step10; 10. The scout bee generates a new optimal target point according to Eq. 3; 11. If the maximum number of iterations has been reached, show output the solution; otherwise, continue the search task from Step 4.

3 Simulation Results Analysis

In actual maritime search and rescue missions, electronic chart information can be extracted, or modern tools such as radar can be used to detect the target sea area to determine whether there are any obstacles. The coverage path planning for the USV swarm can be planned based on the presence or absence of obstacles. This paper uses a grid map to display the task area, which can more directly and clearly show the coverage of the planned path by the program.

3.1 Coverage Path Planning Simulation Analysis Without Obstacles

A. Single USV coverage path in obstacle-free area

When the search and rescue area is small, or the number of USVs is insufficient, it is necessary to use a small number of USVs to carry out the search and rescue in the target sea area.

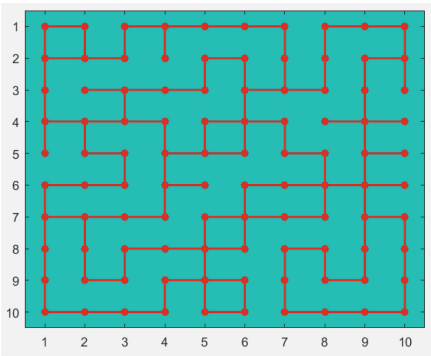


Fig. 1. Coverage path of a single USV in an obstacle-free sea area

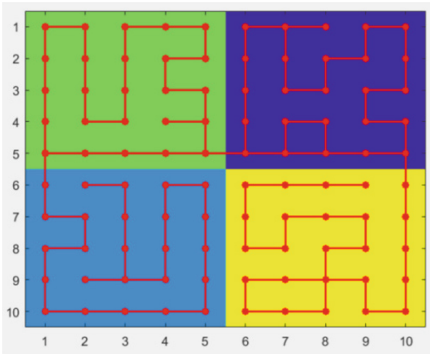


Fig. 2. Coverage paths of four USVs in an obstacle-free sea area

Figure 1 shows the grid map of the coverage path of a single USV in an obstacle-free sea area. All coordinate points of the grid map are represented by red dots, the red solid line represents the path between two coordinate points, and the green background is the task area of this USV. Simulation results show that when a single USV carries out maritime search and rescue in a sea area without static obstacles, the proposed path planning method can successfully plan a path that covers the search and rescue area.

B. Multiple USVs coverage path in obstacle-free sea area

When the search and rescue area is large, or there are more USVs available in the USV swarm that can carry out search and rescue task, multiple USVs can be used to carry out search and rescue in the target sea area.

Figure 2 shows the grid map of the coverage path of four USVs in an obstacle-free sea area. The green, purple, blue, and yellow backgrounds represent the search and rescue areas of the four USVs, respectively. It can be seen that the task area is divided relatively evenly, and the red marks on the grid map are also connected by red solid lines. It can be seen that the proposed path planning method can reasonably plan the coverage path for multiple USVs.

3.2 Coverage Path Planning Simulation Analysis with Static Obstacles

In actual maritime areas, static obstacles may exist, which can not only cause dangerous situations such as grounding for the USVs involved in the mission but also affect the coverage path planning for the USV swarm. To address the challenges that USV swarms may encounter when carrying out search and rescue missions in areas with static obstacles, thus the proposed path planning method is tested to plan the coverage paths for the USV swarm in areas with static obstacles.

A. Single USV Coverage Path in Sea Area with Static Obstacles

Figure 3 shows the grid map of the coverage path of a single USV in a sea area with static obstacles. The yellow part is the open sea area, and the blue rectangular parts are the static obstacle. The yellow part is divided, the red dots are the grid map coordinates of the open sea area, and the red solid line is the path. Simulation results show that in the task area with static obstacles, the path planning method proposed in this paper can plan the coverage path for a single USV and can fully cover the search and rescue area.

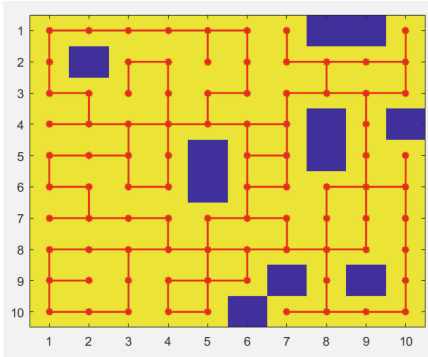


Fig. 3. Coverage path of a single USV in sea area with static obstacles

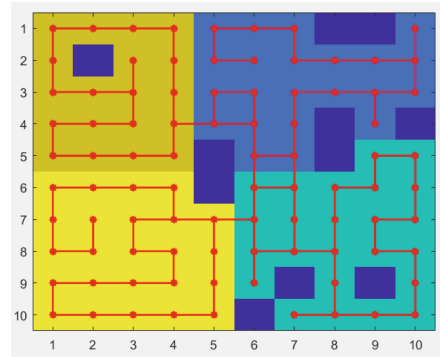


Fig. 4. Coverage paths of four USVs in sea area with static obstacles

B. Multiple USVs Coverage Path in Sea Area with Static Obstacles

Figure 4 shows the grid map of the coverage path of four USVs in a sea area with static obstacles. The brown background area, yellow background area, blue background area, and cyan background area are the task areas allocated by the program for the four USVs, and the dark blue rectangles are the static obstacle. As shown in Fig. 4, due to the irregular position of the static obstacles, the task areas allocated by the program are determined by the static obstacles. Simulation results show that the proposed path planning method can reasonably allocate task areas for the four USVs according to the position of the static obstacles, and complete the coverage path planning for the search and rescue area of the USV swarm without omitting any area.

4 Conclusion

This paper conducts research from two aspects: the establishment of the environmental model and the coverage path planning of USVs. It first describes the grid method used to establish the environmental model. Then, it proposes the research objectives of coverage path planning and transforms the coverage path planning problem into a traveling salesman problem for study, using the K-Means algorithm to allocate the task area and the artificial bee colony algorithm to plan the coverage path. Finally, simulation results demonstrate the effectiveness of the proposed coverage path planning method for different numbers of USVs.

Acknowledgments. This work was supported by the Doctoral Scientific Research Foundation of Liaoning Province (2023-BS-077), the Postdoctoral Research Foundation of China (2024M751980), the Open Project of State Key Laboratory of Maritime Technology and Safety (SKLMTA-DMU2024Y3), and Bolian Research Funds of Dalian Maritime University (3132023616).

References

1. Zhang, W., Shan, L., Chang, L., et al.: SVF-RRT*: a stream-based VF-RRT* for USVs path planning considering ocean current. *IEEE Robot. Autom. Lett.* **8**(4), 2413–2420 (2016)
2. Zhao, J., Zhang, L., Zheng, L., Li, N.: The design and research of intelligent search and rescue device based on sonar detection and marine battery. In: 2017 International Conference on Computer Network, Electronic and Automation, ICCNEA, pp. 383–387 (2017)
3. Wang, Y., Liu, W., Liu, J., Sun, C.: Cooperative USVCUAV marine search and rescue with visual navigation and reinforcement learning-based control. *ISA Trans.* **137**, 222–235 (2023)
4. Akbar, A., Fatichah, C., Dikairono, R.: Autonomous surface vehicle in search and rescue process of marine casualty using computer vision-based victims' detection. In: 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia, CENIM, pp. 1–6 (2022)
5. Chu, Z., Wang, F., Lei, T., et al.: Path planning based on deep reinforcement learning for autonomous underwater vehicles under ocean current disturbance. *IEEE Trans. Intell. Veh.* **8**(1), 108–120 (2023)
6. Fan, X., Guo, Y., Liu, H., et al.: Improved artificial potential field method applied for AUV path planning. *Math. Probl. Eng.* **2020**(1), 1–21 (2020)
7. Hadi, B., Khosravi, A., Sarhadi, P.: Deep reinforcement learning for adaptive path planning and control of an autonomous underwater vehicle. *Appl. Ocean Res.* **129**(103326), 1–14 (2022)
8. Yu, F., Shang, H., Zhu, Q., et al.: An efficient RRT-based motion planning algorithm for autonomous underwater vehicles under cylindrical sampling constraints. *Auton. Robots* **47**(3), 281–297 (2023)
9. Tan, G., Zhuang, J., Zou, J., et al.: Adaptive adjustable fast marching square method-based path planning for the swarm of heterogeneous unmanned surface vehicles (USVs). *Ocean Eng.* **268**(113432), 1–15 (2023)
10. Mahmoud, Z.S., Abbasi, A., Yazdani, A., et al.: Uninterrupted path planning system for multi-USV sampling mission in a cluttered ocean environment. *Ocean Eng.* **254**(111328), 1–18 (2022)
11. Jin, X., Er, M.: Cooperative path planning with priority target assignment and collision avoidance guidance for rescue unmanned surface vehicles in a complex ocean environment. *Adv. Eng. Inform.* **52**(101517), 1–13 (2022)



Research on the Method of Mooring Position Selection for Ships Riding to Single Anchor Based on Nautical Practice

Xiao Liu^{1,2} , Hua Deng¹, Bing Wang², Ke Zhang³ , Yong Wu¹,
and Xingya Zhao²  

¹ School of Navigation Technology, Jiangsu Maritime Institute, Nanjing 21100, China

² School of Navigation, Wuhan University of Technology, Wuhan 430063, China
zhaoxingya@whut.edu.cn

³ China Waterborne Transport Research Institute, Beijing 100088, China

Abstract. This study presents an advanced method for selecting anchoring positions for ships riding to a single anchor, with a focus on safety, feasibility, and optimization in the context of autonomous ship navigation. Grounded in nautical practice, the research defines the criteria for identifying safe anchorage waters and determining the appropriate anchor chain length to ensure secure mooring. The proposed method employs a gridding technique to systematically analyze the anchorage area, enabling a comprehensive search for feasible anchoring positions while accounting for environmental conditions and vessel dynamics. An optimal selection algorithm is integrated into the process to enhance decision-making in autonomous anchoring. Simulation experiments validate the effectiveness of the approach across various scenarios, demonstrating its robustness and practical applicability. The findings contribute to improved safety and efficiency in autonomous anchoring operations. The paper concludes with a critical discussion of the results, implications for autonomous maritime operations, and suggestions for future research.

Keywords: Intelligent Mooring · Autonomous Navigation · Automatic Mooring Position Selection

1 Introduction

1.1 Background

With the rapid development of the global shipping industry and continuous advancements in automation technology, autonomous ship navigation has emerged as a key research focus in the maritime field. Autonomous ships not only enhance navigation efficiency and reduce costs but also significantly improve safety and environmental performance. However, they face numerous challenges in practical operations, one of the most critical being autonomous anchoring. Anchoring is an essential operation when a vessel is stationary at a port, shelter, or other designated waters, and its safety directly

impacts the safety of both vessels and personnel [1]. Therefore, developing an efficient and reliable autonomous anchoring system is crucial for achieving fully autonomous navigation.

Traditional anchoring operations rely on the experience and judgment of the crew, but this approach is characterized by high subjectivity and potential for significant error. In complex and dynamic marine environments, it is challenging to ensure the precision and safety of each anchoring operation solely through human effort. In recent years, advancements in sensor technology, computer vision, and machine learning have opened new opportunities for autonomous anchoring research.

Autonomous anchoring refers to the process by which a vessel utilizes onboard sensors to perceive the maritime traffic environment, assesses its own condition and the hydrological environment, identifies a safe anchoring position [2], and completes path planning and decision-making based on the vessel's maneuverability and constraints posed by static and dynamic obstacles. The vessel then navigates to the anchoring position under the guidance of a tracking control system to complete the anchoring task [3]. The selection of the anchoring position is a fundamental prerequisite, and its outcome is critical to making sound anchoring decisions.

1.2 Literature Review

When ocean-going vessels arrive at their destination ports, they often face situations where they need to anchor at the anchorage area to wait for available berths due to berth unavailability. With a high density of anchored vessels in the anchorage area, selecting a safe anchoring position is crucial for the vessel's safety. Currently, domestic scholars have conducted relevant research on ship anchoring, achieving significant results in anchorage capacity, anchoring safety, anchorage layout and optimization, as well as anchoring maneuver simulations.

Makino [4] analyzed the motion trajectories of anchored vessels in Osaka Bay under severe weather conditions by examining the anchorage environment, establishing a prediction model to enhance the safety of anchored vessels. Gunnu [5] focused on anchoring safety based on the stability of vessels engaged in anchoring operations, proposing standards for the critical static heel angle and critical roll angle by considering vessel heading and mooring line direction, validated through anchoring accident cases. Cao [6] analyzed the movement patterns of anchored vessels, establishing a radius model and a safe distance model for anchoring positions based on vessel type, draft, and other factors. Monte Carlo simulation was employed to measure the anchoring area and determine the anchoring point.

Usui [7] applied an environmental stress model to anchorage layout and ship navigation scenarios, analyzing the relationship between anchorage layout and ship maneuvering difficulty, calculating the safe waters required from the perspective of vessels navigating through anchorage waters. Zou [8] studied single-anchor anchoring for Capesize bulk carriers, analyzing factors such as anchoring depth, vessel draft differences, anchor chain equipment, and anchoring motion, proposing a numerical simulation model for describing anchoring behavior to provide a theoretical approach for safe anchoring. Sasa [9] focused on the anchoring motion of a 28,000-ton bulk carrier, reproducing dragging

anchor motion under severe weather conditions and assessing the safety of the anchoring process considering wave conditions.

In the field of anchoring decision-making, scholars at home and abroad have explored anchoring safety through mooring forces, anchoring methods [10], anchoring radius, and anchorage layout, while also analyzing and simulating ship anchoring maneuvers. However, research on intelligent single-anchor anchoring position selection methods based on nautical practice remains relatively scarce.

1.3 Contribution

The selection of an intelligent anchoring position is central to intelligent anchoring decision-making and is one of the critical functions of autonomous navigation systems. Research in this area is essential for the advancement of intelligent ships. In nautical practice, ocean-going merchant vessels primarily employ single-anchor anchoring, making this study focused on the method for selecting anchoring positions for intelligent single-anchor anchoring based on practical navigation scenarios. This research addresses the following key issues:

- 1) Based on nautical practice, the study determines the safe anchor chain length and establishes the corresponding safe anchorage area for vessels.
- 2) The proposed method utilizes gridding theory to systematically analyze the anchorage area, conducting a comprehensive search for feasible anchoring positions while considering environmental conditions and vessel dynamics. Additionally, the method integrates an optimal selection algorithm to enhance the safety of ship anchoring.

1.4 Outline

To thoroughly explain the proposed method and the associated analysis, the content of this paper is structured as follows: Sect. 2 presents a study on the method for calculating safe waters for ships riding to a single anchor based on different environmental chain lengths. Section 3 develops an optimal anchoring position selection method for intelligent ships' autonomous anchoring, utilizing gridding theory and the defined safe anchorage waters. To validate the effectiveness of the proposed method, Sect. 4 conducts a case study, followed by an analysis and discussion of the results. Section 5 concludes the study and offers recommendations for future research.

2 Safe Waters for Ships Riding to Single Anchor

2.1 Safe Anchor Chain Length for Anchoring

Single-anchor mooring is a commonly used mooring method for ocean-going vessels, where the ship is secured at sea using an anchor chain and anchor fluke deployed from either the port or starboard bow, maintaining a stable mooring state. Compared to other mooring methods, single-anchor anchoring does not require multiple anchoring points, thus saving operational time and costs, making it widely used in maritime transport and mooring operations.

To maintain a stable anchoring state, it is essential for anchored vessels to ensure sufficient anchor chain length to provide adequate holding force. Under the influence of gravity, the released anchor chain length is divided into two parts: the suspended chain length in the water and the chain length lying flat on the seabed, as illustrated in Fig. 1.

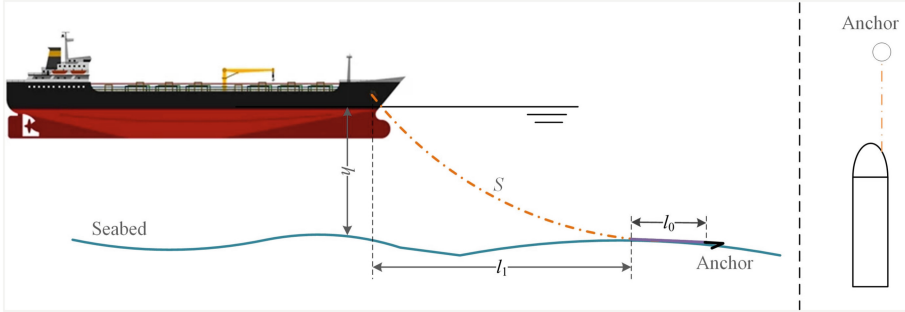


Fig. 1. Schematic Diagram of Single-Anchoring for Ships, Side View (Left) and Top View (Right).

In the left-side diagram, the blue line represents the seabed, the orange dashed section indicates the anchor chain suspended in the water, and the purple section depicts the chain lying on the seabed. The right-side diagram provides a top view of the vessel in a single-anchor anchoring configuration.

The force that secures the vessel at sea is also known as the anchoring mooring force, primarily provided by the combined effects of the anchor holding power and the frictional force generated between the chain lying on the seabed and the seafloor, as shown in Eq. (1).

$$P_0 = P_a + P_c = \lambda_a \cdot W_a + \lambda_c \cdot w_c \cdot l_0 \quad (1)$$

In Eq. (1), P_0 represents the anchoring mooring force, P_a denotes the anchor's holding power, P_c is the holding power of the chain lying on the seabed. λ_a is the anchor's holding coefficient, which ranges from 4 to 6 depending on the anchor type, W_a is the anchor's weight in air (kg). λ_c is the holding coefficient of the anchor chain (0.6 for mud and 0.75 for sand), w_c is the weight per unit length of the anchor chain in air (kg), l_0 represents the length of the chain lying on the seabed (m).

The length of the anchor chain lying on the seabed can be calculated using Eq. (2).

$$l_0 = S_0 - s = S_0 - \sqrt{h(h + \frac{2F}{w_c})} \quad (2)$$

where, S_0 is total length of the released anchor chain (m), s is length of the suspended anchor chain(m), h represents vertical distance from the hawsepipe to the seabed (m), F is net external force acting on the vessel.

From the perspective of static equilibrium, the fundamental requirement for safe anchoring is to ensure sufficient anchoring force, which should be capable of withstanding the net external force acting on the anchored vessel, that is, $P_0 \geq F$.

The net external force acting on a single-anchor vessel in this area is typically approximated by the wind force exerted on the vessel's frontal surface, as shown in Eq. (4).

$$F = 0.5 \cdot \rho_a \cdot V_a \cdot C_a \cdot A_T \quad (3)$$

In the equation, F represents the wind pressure in newtons (N), ρ_a is the air density, taken as 1.226 kg/m^3 , V_a denotes the relative wind speed in meters per second (m/s), C_a is the wind pressure coefficient, and A_T is the projected area of the vessel above the waterline in square meters (m^2).

Therefore, by substituting Eq. (2) into Eq. (1), the safe length of the anchor chain for a single-anchor anchoring state can be determined, as shown in Eq. (4).

$$S_0 = s + l = \sqrt{h(h + 2F/w_c)} + (F - \lambda_a \cdot W_a)/\lambda_c \quad (4)$$

2.2 Safe Anchorage Waters

For Ocean-going vessels using single-anchor mooring, factors such as wind, waves, anchoring position, and vessel mass cause the vessel to pivot around the anchor fluke, resulting in a circular or irregular elliptical or wave-like trajectory. The actual movement trajectory based on AIS data, as shown in Fig. 2, illustrates the anchoring path of the bulk carrier JINFUHAN 69 at Changjiang Kou Anchorage No. 1.

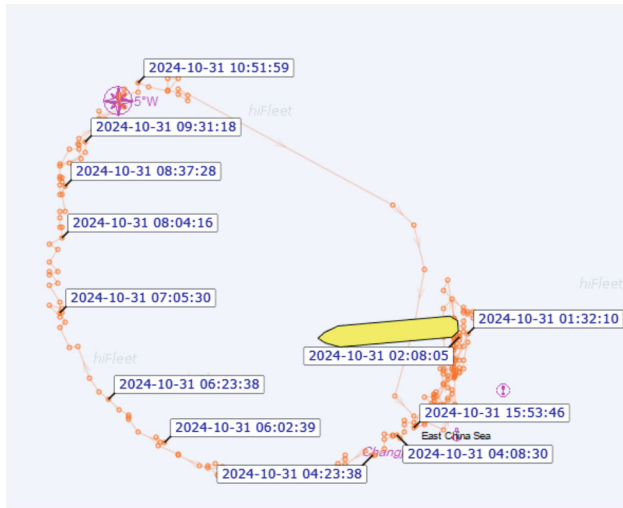


Fig. 2. Ship Anchoring Trajectory Based on Real AIS Data.

From a Vessel Safety Perspective, the Anchoring Waters for a Single-Anchor Vessel Are Taken as a Circular Area with the Radius of Maximum Swinging Area, as Shown in Fig. 3.

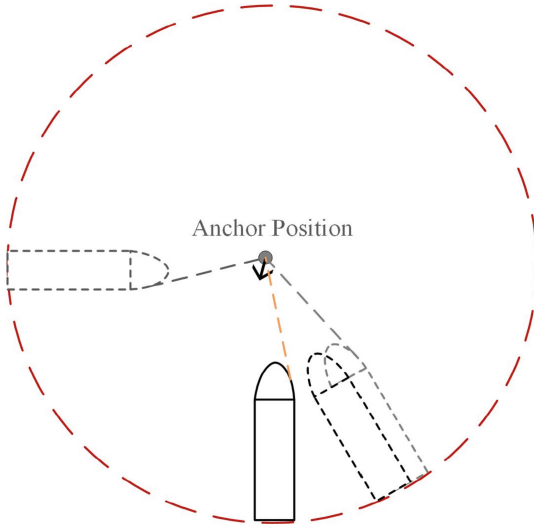


Fig. 3. Schematic Diagram of Anchoring Waters for a Single-Anchor Vessel.

The radius R of the anchoring waters depends on the length of the anchor chain deployed and the vessel's length overall (LOA) at the time of anchoring, as shown in Eq. (5).

$$R = L + l_0 + l_1 \quad (5)$$

In nautical practice, due to the numerous and complex external forces acting on the vessel, the length of the anchor chain for single-anchor anchoring is often estimated using empirical formulas. The UK estimates this length using Eq. (6), while domestic estimates based on weather conditions are made using Eq. (7).

$$R = L + 39\sqrt{h} \quad (6)$$

$$\begin{cases} R = L + 3h + 90 \\ R = L + 4h + 145 \end{cases} \quad (7)$$

3 Anchoring Position Selection Method

As a specialized area for vessel mooring, emergency maneuvers, and sheltering from storms, the layout of an anchorage is determined by factors such as port size, types and density of arriving vessels, and the hydrological environment. Prior to arrival, the vessel's autonomous anchoring decision-making module should comprehensively account for external disturbances, such as wind and currents, while considering anchorage layout, water depth, seabed composition, and the vessel's load condition to determine the appropriate anchor chain length and calculate the required safe anchoring area. Anchorage gridding is then used to identify feasible anchoring positions, and an optimal anchoring position is selected from this set.

3.1 Anchorage Gridding

In the practice of harbor anchorage layout, the anchorage area is often shaped as a regular quadrilateral or polygon. As shown in Fig. 4, this type of anchorage layout is the most prevalent in practical applications.

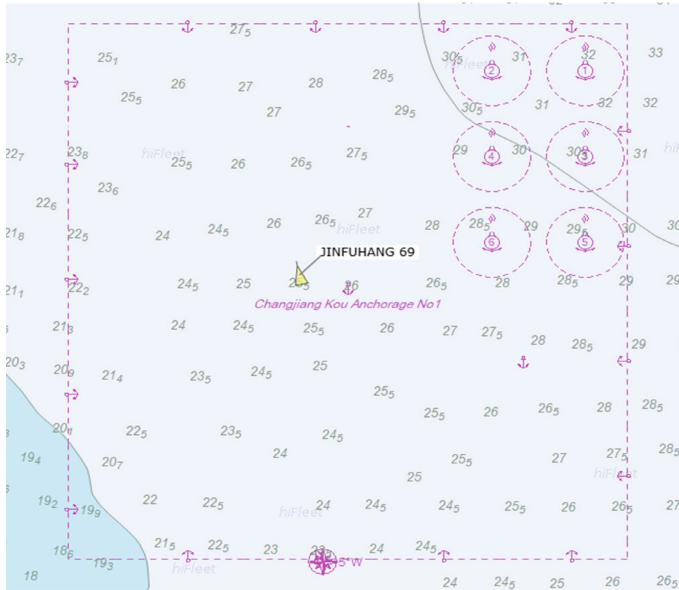


Fig. 4. Schematic Diagram of Anchorage Layout Plan.

To facilitate computer-based perception and decision-making within the anchorage environment, this study grids the anchorage waters by dividing the designated anchorage area at the destination port into numerous fixed-size cells, based on gridding theory. After gridding, each grid cell's center point is designated as a preliminary anchoring position, evaluated individually using an enumeration method. If the safety area for anchoring falls within the boundaries of the anchorage waters and does not overlap with the safety zones of other anchored vessels, the grid center is marked as a feasible anchoring position.

To correspond with grid cells in subsequent research, each cell in the grid map is encoded. This study applies a sequential numbering method for grid identification, setting the bottom-left cell as the origin and numbering from left to right and bottom to top, beginning with the natural number 1. Principle Diagram of Anchorage Gridding, as Shown in Fig. 5.

3.2 Feasible Anchoring Position Search Method

Before arriving at the anchorage, the vessel uses information on water depth and the positions of already anchored vessels to determine the necessary safe anchoring area

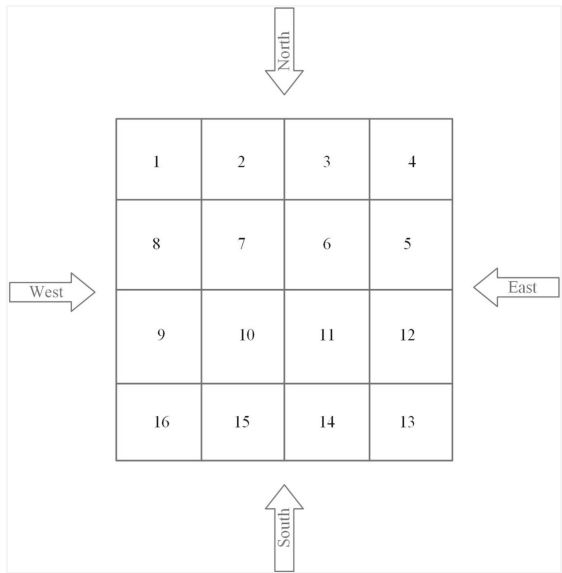


Fig. 5. Schematic of gridding of anchorage waters.

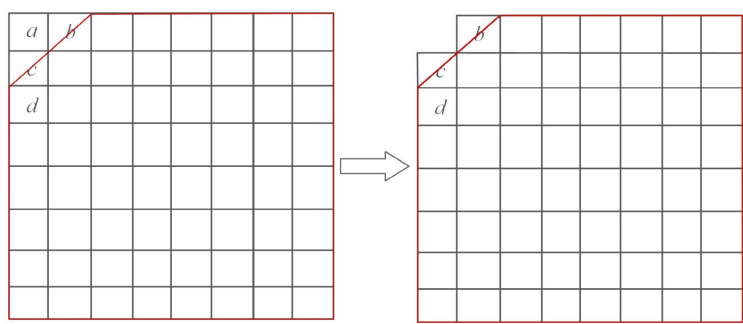


Fig. 6. Schematic Diagram of Removing Grids with Centers Outside the Rasterized Waters.

for itself, as well as the safe zones for other anchored vessels. Based on the gridded anchorage area, feasible anchoring positions are then selected.

After performing large-scale gridding, some grid cells may have only a small portion or none of their area within the designated gridding waters. The next step is to eliminate these cells by determining whether the center point of each grid cell lies within the boundaries of the anchorage waters. If the center point is within the area, the grid cell is retained; if it is outside, the cell is removed. As shown in Fig. 6, the red area represents the desired gridding range, initially processed through large-scale gridding. For the four grid cells labeled *a*, *b*, *c*, and *d* in the figure, the center point of grid A lies outside the gridded water area, so it is removed, while the center points of the remaining grids lie within the area and are thus retained.

By establishing a planar Cartesian coordinate system for the anchorage area, two-dimensional coordinates are assigned to existing vessels or other objects within the anchorage that may obstruct anchoring operations. Using the method for calculating safe anchoring areas, a safe zone is defined for each anchored vessel. Anchoring positions that satisfy the required safe distance between vessels are then selected in combination with the safe area for the current vessel, as shown in Eq. (8).

$$\begin{cases} d_n = \sqrt{(x_o - x_n)^2 + (y_o - y_n)^2} \\ \text{Min}(d_n) \geq D_s \end{cases} \quad (8)$$

where, the point (x_n, y_n) represents the position of existing vessels or other objects that may obstruct anchoring operations within the planar Cartesian coordinate system of the anchorage area, while the point (x_o, y_o) denotes an anchoring position of own ship that satisfies the safe distance D_s for anchoring.

Assume the anchorage area is divided into m grid cells, with (x_o^k, y_o^k) representing the center coordinates of each grid cell ($k = 1, 2, \dots, m$). If a grid center satisfies the above equation, it is considered a feasible anchoring position for the vessel. The search then continues to the next grid cell until all cells have been evaluated.

Based on the above method, all feasible anchoring positions within the anchorage area can be determined, forming the set $M = [P_1, P_2 \dots P_j]$, where j represents the number of feasible anchoring positions within the anchorage.

3.3 Optimal Anchoring Position Selection Method

Anchoring position allocation refers to the process of assigning anchoring positions when a vessel arrives at the anchorage. In practice, when the anchorage is designated for non-single-position anchoring, a random allocation mechanism is often used, where anchoring positions are assigned in the order of vessel arrival. The primary allocation methods are fixed-point anchoring and free anchoring. Vessels employing fixed-point anchoring proceed to a specific anchoring position assigned by the Vessel Traffic Service (VTS) center.

A feasible anchoring position is one that meets the basic requirements for safe vessel anchoring. In navigational practice, vessels using free anchoring rely on the pilot's professional judgment, taking into account factors such as anchorage depth, available water space, and ease of access to the anchoring position, to select the safest and most advantageous position. The center of the largest safe area radius is designated as the optimal anchoring position.

In the set M of feasible anchoring positions obtained in Sect. 3.2, there are m feasible positions. For each feasible anchoring position (x_o^p, y_o^p) ($p = 1, 2, \dots, m$), the distance $\text{Dis}^{(m)}$ between this position and each anchored vessel is calculated. The safest anchoring point P^m is then selected as the optimal anchoring position.

4 Simulation Experiment

In this section, the study area used in this paper is introduced first. Moreover, the analysis of the anchoring position selection is carried out to presented method.

4.1 Study Area

An anchorage of Huizhou port of China is selected, with abundant port resources and dense traffic flow. With the development of ports, the number of ships is constantly increasing, and the demand for anchorage is also increasing day by day. The limitation of anchorage is shown in Table 1 and Fig. 7.

Table 1. The limitation of anchorage

Position	Latitude/N	Longitude/E
1	22°27.1′	114°41.496′
2	22°27.1′	114°43.489′
3	22°29.077′	114°43.504′
4	22°29.077′	114°41.496′

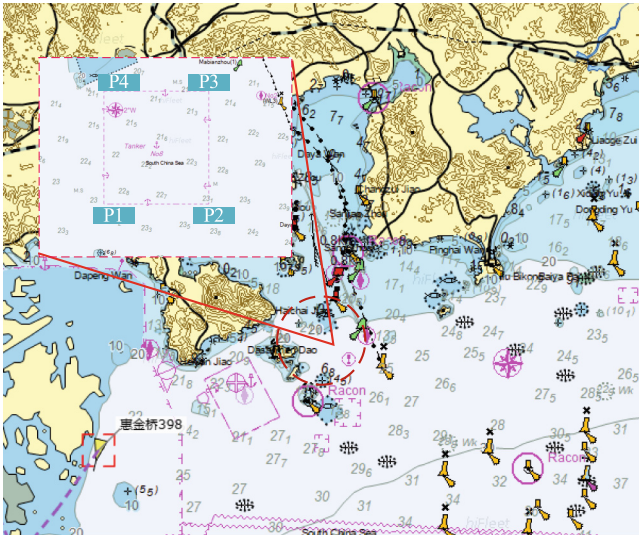


Fig. 7. The study area of the case study.

4.2 Case Study

This case study selects the situation of the anchorage at a certain moment in November 2024 based on real data and scenario calculations. Among them, there are three ships anchored, and the ship parameters are shown in Table 2. This ship is an oil tanker with a length of 225 m and needs to anchor here. According to the algorithm proposed in this paper, the results are shown in the Fig. 8.

As can be seen from Fig. 8(a), there exists a safety extreme low value in the vicinity of the anchored ship, which is in accordance with nautical practice. With the increasing distance, the safety value of the ship increases, and the optimal anchor position is at the highest point in the figure, i.e., the lower left corner of the present anchorage. That is, as shown in Fig. 8(b), the location where the ship anchors is the optimal anchor position. The anchor position at this location is furthest away from the other anchor positions of the vessel.

Table 2. The anchored ships’ particular

Position	Latitude/N	Longitude/E
BEI HAI ZHAN WANG	22.473763	114.697866
LIAN XING HU	22.459483	114.713831
BEI HAI FEN JIN	22.478070	114.715805

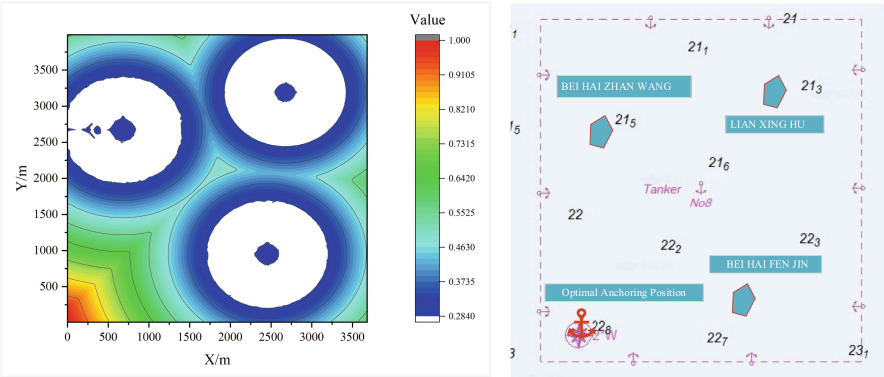


Fig. 8. The feasible anchoring position and optimal anchoring position.

5 Summary and Discussion

This study addresses the issue of autonomous single-anchor position selection for future intelligent vessels. It quantifies safe anchoring areas by integrating vessel dimensions, anchorage hydrological characteristics, and vessel anchoring orientation. A feasible anchoring position search method is developed based on gridding theory and defined safe anchorage zones. Finally, an optimal anchoring position selection method is proposed using an anchoring position allocation mechanism, and digital simulation experiments are conducted to validate the effectiveness of the method. This research provides theoretical support for intelligent anchoring decision-making for vessels.

Acknowledgment. The work was supported by The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 24KJA580002) and The National Natural Science Foundation of China (Grant NO. 52071249, 52271367).

References

1. China Classification Society. Rules for Intelligent Ships (2024)
2. Mehmet, K.: Risk assessment of ship anchorage handling operations using the fuzzy bow-tie method. *Ocean Eng.* **236**, 109500 (2021)
3. Zhu, F., Hu, Q., Xu, H., et al.: Study on the drift caused by ship leaving out of control with dragging anchor in Zhejiang Coastal. In: Xian: 2018 4th International Conference on Environmental Science and Material Application (ESMA), 15–16 December (2018)
4. Gao, X., Makino, H.: Analysis of anchoring ships around coastal industrial complex in a natural disaster. *J. Loss Prev. Process Ind.* **50**, 355–363 (2017)
5. Gunnu, G.R., Moan, T.: Stability assessment of anchor handling vessels during operations. *J. Marine Sci. Technol.* **23**, 201–227 (2018)
6. Cao, L., Wang, X., Zhang, W., et al.: Research on intelligent detection algorithm of the single anchored mooring area for maritime autonomous surface ships. *Appl. Sci.* **12**(12), 6009 (2022)
7. Usui, H.: Navigating between anchored ships and manoeuvring difficulty. *J. Navig.* **55**(2), 277–291 (2002)
8. Zou, Y., Shen, C., Xi, X.: Numerical simulations on the motions of anchored capesize ships. *J. Navig.* **65**(1), 145–158 (2012)
9. Sasa, K., Incecik, A.: Numerical simulation of anchored ship motions due to wave and wind forces for enhanced safety in offshore harbor refuge. *Ocean Eng.* **44**, 68–78 (2012)
10. Lee, S.W., Sasa, K., Chen, C., et al.: Novel safety evaluation technique for ships in offshore anchorage under rough seas conditions for optimal ship routing. *Ocean Eng.* **253**, 111323 (2022)



Study on the Scale of Urban Rail Station Connection Facilities Based on the Concept of Integration

Xiaohui Sun¹ , Jingqi Zheng², and Yuling Fu²

¹ School of Traffic and Transportation Engineering, Xinjiang University, Urumqi 830017, China
xhsun347@xju.edu.cn

² College of Civil Engineering and Architecture, Xinjiang University, Urumqi 830017, China
fuyuling@stu.xju.edu.cn

Abstract. The planning of urban rail station connection facilities is an important influencing factor to improve the efficiency of urban rail connection, enhance the service capacity of urban rail transportation and expand the scope of passenger attraction. As an important part of the planning of station connection facilities, the refined estimation of the scale of connection facilities is of guiding significance for optimizing the scale of the existing configuration of connection facilities and predicting the scale of future station connection facilities. The connection demand prediction model is constructed to obtain the sharing rate of each connection mode, and the refined land use information of each entrance and exit of the station is integrated with the sharing rate of each connection mode, which is substituted into the integrated scale measurement model of the connection facilities, and the Gao Shengqiao Metro Station in Chengdu is used as an example to quantitatively study the scale of the configuration of the connection facilities of the urban rail station by combining the functional classification of the station's land use and the scale measurement model.

Keywords: Transport Integration Concept · Hybrid Choice Model · Connection Facility Sizing · Connection Mode Choice

1 Introduction

Urban rail transit has become the preferred mode of transportation for residents of large and medium-sized cities because of its economic, fast, green and safe characteristics, but urban rail transit has the problems of low density of the line network and low coverage of stations, which need to realize the integrated transfer with ground transportation, and the formation of the integrated transfer system cannot be separated from the support of the perfect bargaining system around the rail stations. Therefore, it is necessary to refine the prediction of the passenger flow of the urban rail transit stations, and on this basis, to estimate the scale of the connection facilities around the urban rail stations, to better construct the integrated connection system.

Established studies on the measurement of urban rail transit connection sizes are mainly based on connection demand forecast data, and the size of non-motorized, motorized, or multi-modal transportation connection facilities are measured. Persons Brinckerhoff et al. modeled connection demand forecasting by comparing the time it takes travelers to reach the transfer station and the time it takes to reach their destination [1]. The size measurement of non-motorized connection facilities at urban rail stations mainly focuses on bicycle parking space size measurement. Wang et al. studied the scale of bicycle parking at metro stations [2]. Studies on the measurement of the size of motor vehicle connection facilities are dominated by studies on park-and-ride facilities, and Cao measured their size by combining the average utilization rate of parking lots [3]. Scholars have studied the integration of connection facilities and suggested spatial layout and scale for different stations. Xiao et al. optimized the formula and parameters for measuring the scale of connection facilities [4].

In summary, it is essential to predict the demand for urban rail transit connection facilities to optimize their configuration. However, existing research has not sufficiently refined demand predictions for connections at each entrance and exit of stations. Furthermore, current literature has yet to incorporate new transportation modes, such as car-sharing, motorcycles, and electric vehicles, into the connection facility configurations at urban rail transit stations. Therefore, the scale of connection facilities is refined from the concept of transportation integration, and a model is constructed to measure the scale of the configuration of connection facilities.

2 Calculation Process for the Scale of Urban Rail Station Connection Facilities

The scale of the connection facilities is based on the prediction of the total urban rail passenger flow, and the number of connecting passengers is determined by the total passenger flow, the diverted passenger flow at the entrances and exits, and the sharing rate. Configuration of connection facilities needs to clarify the passenger flow demand at each entrance and exit to realize rapid distribution, and the overall calculation is divided into four parts.

Firstly, the total passenger flow forecast is the basis for the configuration of the site connection facilities, and the total passenger flow of the site can be derived from the site construction feasibility report.

Secondly, urban rail transit stations are predicted in terms of passenger flow in different directions. Existing studies have shown that the land use and land development intensity around the station have a positive correlation with its passenger flow [5]. Urban land is divided into eight categories according to the land use category in which it is expressed. Building density was adopted as a quantitative indicator. Assuming that the passenger flow at each entrance and exit is proportional to the land use and development intensity, the process of predicting the passenger flow in each direction is as follows:

- (1) Scope division. Based on the distance between the urban rail station and its neighboring stations, the passenger flow attraction area is divided, and the passenger flow attraction range of each entrance and exit of the station is delineated.

- (2) Data collection. The passenger flow attraction range of each entrance and exit is divided into several traffic districts according to the nature of the land, and the nature of the land, land area and building density of each traffic district are determined.
- (3) Proportion calculation. Calculate the ratio according to the floor area of the traffic district and other indicators, combined with the attraction weight of the land use nature.

Thirdly, based on China's support for green transportation, the principle of ensuring the proportion of various connection methods and increasing the sharing rate of public transportation and non-motorized vehicles should be followed in predicting the passenger flow sharing rate.

Finally, using the above data, the size of the connection facility configuration was measured.

3 Model Construction for Measuring the Scale of Integration of Connection Facilities at Urban Rail Stations

Connection demand forecasting is the prediction of the probability of choice of different connection modes for travelers when traveling on urban rail. Influencing factors include individual socioeconomic attributes, travel attributes and psychological attributes. Structural equation modeling is used to measure the psychological attributes and substitute them into the Mixed Logit model to predict the connection mode share rate. The Mixed Logit model overcomes the irrelevant scheme independence of the traditional Logit model and follows the utility maximization theory [6]. After measuring the psychological latent variables through AMOS software, they were substituted into the Mixed Logit model and the model was solved using STATA software.

3.1 Walking Connection Facility Scale Measurement Modeling

Pedestrian connection facilities include sidewalks, street crossing facilities (underpasses or flyovers) and distribution plazas, etc. [7]. For urban rail stations, there are passengers walking in and out of the station for connection, so take the sum of peak hourly inbound and outbound passenger flow as the basis for calculating the width of the walking channel [8].

$$W_P = (Q_w/Q_W) \times W \quad (1)$$

where, Q_w is the number of pedestrians in the peak hour; Q_W is the design capacity of a single pedestrian access belt; W is the width of a single pedestrian access belt.

3.2 Model for Sizing Non-motorized Connection Facilities

Non-motorized connection facilities include bicycle connection facilities and electric vehicle connection facilities, and bicycles are divided into shared bicycles and private bicycles. For urban rail stations, the value of the connection passenger flow is taken as the basis for calculation, and parking lot turnover coefficients are used as the basis for calculating the size of the non-motorized connection facilities [9].

(1) Model for sizing bicycle connection facilities:

$$S_b = \frac{Q_b \times s_b}{\varphi_b \times \alpha_b} \quad (2)$$

where, S_b is the area of bicycle parking area, m^2 ; Q_b is the number of bicycle connections in the peak hour; s_b is the area of single bicycle parking, taken as $1 m^2$; φ_b the number of passengers per vehicle, taken as 1 person/vehicle; and α_b is the turnover rate of bicycle parking, taken as 1 time/vehicle-hours for private bicycles, and 4 times/vehicle-hours for shared bicycles.

(2) Model for sizing electric vehicle connection facilities:

$$S_{eb} = \frac{Q_{meb} \times \beta_{eb} \times s_{eb}}{\varphi_{eb} \times \alpha_{eb}} \quad (3)$$

where, S_{eb} for the electric car parking area, m^2 ; Q_{meb} for the number of full-day connection; for the motorcycle car connection in the electric car connection ratio, take 0.8; φ_{eb} for a single electric car parking area, take $1.5 m^2$; s_{eb} the number of passengers per vehicle, take 1 person / vehicle; α_{eb} for the electric car parking turnover rate, take 2.0.

3.3 Model for Sizing P&R Connection Facilities

P&R parking lots are divided into two categories: parallel and integrated. Parallel parking is independent of the rail transit, travelers need to park and then transfer; integrated parking and transportation station integration, travelers can directly transfer. In practice, the layout is mostly parallel. Connection facilities include private car and motorcycle parking and car-sharing drop-off points, and their scale measurement model is like that of bicycle connection facilities, with scale related to park-and-ride passenger flow, turnover rate, average number of passengers, utilization rate and parking ratio [10].

(1) Model for sizing private car connection facilities:

$$S_{car} = N_{car} \times s_{car} = \frac{Q_{car} \times \beta_{car} \times s_{car}}{\varphi_c \times \gamma \times \alpha_{car}} \quad (4)$$

where, S_{car} is the area of the private car parking lot, m^2 ; N_{car} is the number of parking spaces in the private car parking lot, parking spaces; s_{car} is the floor space of a single car, taking $30 m^2$ /space; Q_{car} is the full-day private car connecting passenger flow; β is the parking ratio, taking 0.7; φ_c is the number of passengers per car, taking 2; γ is the parking lot utilization, taking 1.0; α_{car} is the parking lot turnover rate, taking 1.2.

(2) Model for sizing motorcycle and car-sharing connection facilities:

$$S_m = N_m \times s_m = \frac{Q_{meb} \times \beta_m \times s_m}{\varphi_m \times \alpha_m} \quad (5)$$

$$S_{cs} = N_{cs} \times s_{cs} = \frac{Q_{cs} \times s_{cs}}{\varphi_{cs} \times \phi_{cs} \times \alpha_{cs}} \quad (6)$$

where, S_{cs} is the area of the private car parking lot, m^2 ; N_{cs} is the number of parking spaces in the private car parking lot, spaces; s_{cs} is the footprint of a single car, taken as $25 \text{ m}^2/\text{space}$; Q_{cs} is the full-day private car connection passenger flow; φ_s is the number of passengers per car, taken as 2; φ_{cs} is the ratio of the number of parking spaces to the fleet size, taken as 2.5, α_{car} is the turnover rate of the parking lot, taken as 1.2, and α_{cs} is the parking lot turnover rate, taken as 1.2.

3.4 K&R Connection Facility Sizing Modeling

The K&R stopping facilities include rental vehicle stops and pick-up and drop-off mini-van stops, and the size of the stops is related to the number of berths for temporary parking connections, the footprint of a single connecting vehicle, the peak hourly connection traffic, and the number of passengers carried in a single vehicle [5].

(1) Measurement of the size of the parking facilities for cabs or online vehicles:

$$S_{taxi} = N_{taxi} \times s_{taxi} = \frac{Q_{taxi} \times s_{taxi}}{\varphi_{taxi} \times \lambda} \quad (7)$$

where, S_{taxi} is the area of the K&R parking facility, m^2 ; N_{taxi} is the number of parking spaces in the P&R parking lot, spaces; s_{taxi} is the footprint of a single small car, which is taken to be $15 \text{ m}^2/\text{space}$; Q_{taxi} is the peak hour cab connection demand; φ_t is the number of passengers per vehicle, which is taken to be 1.3; and λ is the peak hour coefficient, which is taken to be 1.0.

(2) Measurement of the size of the parking facilities for pick-up and drop-off of small cars:

$$S'_{car} = N'_{car} \times s_{car} = \frac{Q_{car} \times \beta' \times s'_{car}}{\varphi_{c1} \times \lambda \times T_M} \quad (8)$$

where, S'_{car} is the area of P&R parking lot, m^2 ; N_{car} is the number of parking spaces, parking spaces; s'_{car} is the floor space of a single car, taking $30 \text{ m}^2/\text{space}$; Q_{car} is the peak hour private car connection passenger flow; β' is the proportion of parking, taking 0.3; φ_{c1} is the number of passengers per car, taking 1.3; λ is the peak hour coefficient, taking 1.0; T_M is the peak hour frequency of the urban rail train, calculated according to the peak hour departure interval, taking 12 when the departure interval is 5 min. Interval calculation, when the departure interval is 5 min, take 12.

3.5 Conventional Transit Scale Measurement Model

The connection between buses and subways is achieved through bus stops, which are calculated based on parameters such as maximum hourly passenger volume and interval time between buses [11].

Conventional buses and the subway are connected by bus stops, which can be divided into straight-line stops and harbor-type stops and are calculated by parameters such as the maximum number of passengers served per hour at the stops and the interval time between buses.

The number of bus parking spaces N_{bus} is equal to the ratio of the number of regular bus service passengers Q during peak hours to the maximum number of passengers q served per hour at a single station.

$$N_{bus} = \frac{Q(t_b N + T_b)}{3600 N \omega} \quad (9)$$

where, N is the number of passengers boarding and alighting on road buses, taken as 15 people; ω is the offset stopping time and arrival time fluctuation discount factor, taken as 0.833; t_b is the average boarding and alighting time of each passenger, taken as 3 s; and T_b is the time interval between arrivals of road bus vehicles, taken as 240 s.

4 Example Studies

Chengdu Metro Gao Shengqiao Station is taken as an example to apply the forecast data of connection mode sharing rate obtained from the hybrid selection model and the station's connection facility configuration model for example analysis. Gao Shengqiao Metro Station is the interchange station of Chengdu Metro Line 3 and Line 5, the station halls and platforms of Line 3 and Line 5 are independent, and the stations of the two lines are connected through the interchange channel, with seven entrances and exits in total, and five entrances and exits, namely Exit A, Exit B, Exit E, Exit F, and Exit G, are laid out on the west side of the First Ring Road, and two entrances and exits, Exit C and Exit D, are laid out on the east side.

4.1 Forecast of Connection Demand and Scale of Connection Facilities at Gao Shengqiao Station

To better confirm the scale of Gao Shengqiao Station's connection facilities, it is necessary to determine the connection demand at each entrance and exit of the station under the situation of all-day total passenger flow and peak hourly passenger flow.

(1) Total passenger flow forecast

Chengdu Metro Line 5 Phase II project as a background, to get the long-term forecast of the metro station average daily inbound passenger Q1 flow up to 50,000, outbound passenger Q2 up to 60,000, transfer passenger Q3 up to 25,000, morning peak hourly inbound passenger Q4 for 0.45 million, outbound passenger Q5 for 0.55 million, evening peak hourly inbound passenger Q6 for 0.55 million, outbound passenger Q7 for 0.50 million. 0.50 million.

(2) Directional passenger flow prediction

Based on the distance between Gao Shengqiao Station and adjacent stations and existing studies, the passenger attraction area is determined. 0.6 km are selected as the passenger attraction area of Gao Shengqiao Station, and the area is divided into

seven parts using ArcGIS 10.8 based on the locations of the seven entrances and offline surveys. Through the formula of passenger flow sharing ratio, combined with the floor area, building density and land attraction weight, the passenger flow sharing ratio of Gao Shengqiao entrances and exits is calculated, as shown in Table 1.

Table 1. Passenger flow sharing ratio at entrances and exits.

A	B	C	D	E	F	G
7.62%	6.96%	33.70%	19.52%	9.66%	6.74%	15.80%

(3) Calculation of passenger flow of each connection at Gao Shengqiao station

The use of Chengdu rail transit survey data to construct a hybrid selection model, get the departure end and arrival end of the connection mode prediction ratio data, based on the principle of passenger flow sharing rate prediction, the need for green connection mode proportion of moderate amplification, the use of STATA to get to the consideration of the choice of the probability of each connection mode as shown in Table 2.

Table 2. Selection probability of each connection method.

Connection Method	Walking	Shared Bicycle	Bus	Car Sharing	Taxi	Private Car	Private Bicycle	Motorcycle
Origin	39.5%	14.93%	21.44%	2.35%	12.93%	4.79%	0.31%	3.76%
Destination	37.12%	22.34%	25.04%	1.72%	13.31%	0.00%	0.47%	0.00%

(4) Calculation of the scale of site connection facilities

① Scale of pedestrian connection facilities at entrances and exits.

Based on the service area of each entrance and exit, the morning peak or evening peak passenger flow is selected as the calculation basis, and the entrance and exit walkway widths are calculated using the pedestrian connection facility scale measurement model, as shown in Table 3.

② Non-motorized vehicle connection facilities scale measurement.

According to the service area of each entrance and exit to select the morning and evening peak hour inbound or outbound passenger flow as the basic data, the use of passenger flow sharing ratio and the bicycle connection facilities established in the previous section of the size of the measurement model for the size of each entrance and exit of the bicycle connection facilities to configure as known in Table 4.

Table 3. Walkway width setting.

Entrance and Exit	Description of service area	Passenger Flow Selection	Passenger Flow Fetch	Pedestrian traffic(person)	Width Setting (<i>m</i>)
A	Commercial areas, residential areas, attractions	Morning Peak Flow	$Q_4 + Q_5$	301	3
B	Comprehensive area	Morning and Evening Peak Passenger Traffic	$Q_6 + Q_7$	271	3
C	Residential area, schools	Morning Peak Flow	$Q_4 + Q_5$	1331	4
D	Government units, school	Morning Peak Flow	$Q_4 + Q_5$	771	4
E	Comprehensive area	Morning and Evening Peak Passenger Traffic	$Q_6 + Q_7$	377	3
F	Living area	Morning Peak Flow	$Q_4 + Q_5$	266	3
G	Comprehensive area	Morning and Evening Peak Passenger Traffic	$Q_6 + Q_7$	616	4

③ P + R parking lot size estimation

The Gao Shengqiao station uses a full-day inbound passenger flow and connection facility sizing model to calculate the size of the P + R parking lot. The calculation of motorcycle connection facilities considers that only 22% of users park their vehicles in off-station connection facilities, the same as for electric vehicles, as shown in Table 5.

④ K + R Stopping Facilities

Based on the K + R connection facility configuration principle, taxi temporary parking facilities will be set up on both sides of the road, and temporary parking spaces at both entrances and exits will be uniformly configured according to the number of passengers entering during peak hours, as shown in Table 6.

⑤ Conventional bus stopping facilities

According to the location of the site’s entrance/exit space, traffic network and existing bus stops, the conventional bus connection facility scale model is used to calculate the conventional bus stopping berths at Gao Shengqiao Station as known in Table 7.

Table 4. Bicycle connection facility scale settings.

Entrance and Exit	Passenger Flow Selection	Passenger Flow Fetch	Private Bicycle Passenger (person)	Bicycle Connection Facility (m^2)
A	Morning peak inbound traffic	Q4	86	22
B	Morning peak outbound passenger flow	Q5	92	23
C	Morning peak inbound traffic	Q4	379	95
D	Morning peak outbound passenger flow	Q5	258	65
E	Morning peak outbound passenger flow	Q5	128	32
F	Evening peak outbound passenger flow	Q7	81	20
G	Morning peak outbound passenger flow	Q5	209	52

Table 5. P+R parking lot size facilities.

Connection Method	Passenger Flow Selection	Passenger Flow Fetch	Entrance and Exit	Passenger flow (person)	Number of parking spaces (pcs)	P+R Connection Facilities (m^2)
Private Car	All-day inbound passenger flow	Q1	A/B/E/F/G	784	229	8187
			C/D	892	261	9314
Motorcycle	All-day inbound passenger flow	Q1	A/B/E/F/G	176	18	39
			C/D	200	20	44
Car Sharing	All-day inbound passenger flow	Q1	A/B/E/F/G	550	92	2339
			C/D	625	104	2661

Table 6. K&R stopping point scale settings.

Passenger Flow Selection	Traffic Fetch	Entrance and Exit	Taxi Passenger flow (person)	Pick-up and drop-off passenger cars (person)	Number of parking spaces (pcs)	K+R Connection Facilities (m^2)
Morning peak inbound traffic	Q4	A/E	84	9	2	30
		B/F/G	144	15	3	30
		C	164	17	3	45
		D	95	10	2	30

Table 7. Scale settings for conventional transit connection facilities.

Connection Method	Traffic Selection	Passenger Flow Fetch	Entrance and Exit	Passenger flow(person)	Number of stops (number)	Bus Connection Facilities (m^2)
Bus	Morning peak inbound traffic	Q4	A/E	167	1	39
			B/F/G	285	2	78
			C	325	2	78
			D	188	1	39

5 Conclusion

Based on the total amount of passenger flow, firstly, the land attraction weight was applied to get the passenger flow sharing ratio of each entrance and exit, secondly, the hybrid selection model was applied to get the probability of connection selection, and finally, the scale prediction model of the connection facilities was established to get the refined scale of each connection mode, and the example analysis was carried out with the Chengdu Railway Transportation travelers, and the main research conclusions are as follows:

- (1) According to the land use situation within 600m of urban rail transit stations, the proportion of passengers entering and exiting each entrance and exit was obtained, and the passenger flow sharing rate model for entrance and exit connections was established based on the three indexes of building category, building density and floor area, and the passenger flow shared by each entrance and exit connection was calculated accordingly in the peak hour.
- (2) On the basis of the passenger flow shared by each entrance/exit of the urban rail transit station, the refined prediction model of the scale of the connection facilities was obtained by combining the probability of the connection mode selection and the predictability of the scale of the connection facilities. And take Chengdu Gao Shengqiao Station as an example to estimate the scale of bicycle parking lot, regular

bus stop, cab stop, P+R parking lot, K+R parking lot and bicycle parking lot around the urban rail transit station.

References

1. Persons, B., Qyade, P., et al.: Park and Ride Demand Estimate. Submitted MOTRO. Seattle. WA. 7578 (1995)
2. Wang, Q., Du, G., Yang, H.: Xi'an bike and subway traffic cohesion planning. *Technol. Econ. Areas Commun.* **11**(03), 4–6+10 (2009). (in Chinese)
3. Cao, G., Yang, Z., Zuo, Z.: The planning method of regional parking based on railway transit park-and-ride. *J. Transp. Syst. Eng. Inf. Technol.* **16**(04), 153–158 (2016). (in Chinese)
4. Xiao, S., Ling, X., Wu, D.: Calculation method of the scale of urban rail transit station integrated transfer facilities. *Urban Mass. Transit.* **25**(04), 82–86 (2022). (in Chinese)
5. Qing, K., Lei, W., Yong, L., et al.: Urban land development intensity: new evidence behind economic transition in the Yangtze River Delta. *J. Geogr. Sci.* **32**(12), 2453–2474 (2022)
6. Yin, Y., Wang, H., Xiong, J., et al.: Estimation of optimum supply of shared cars based on personal travel behaviors in condition of minimum energy consumption. *Environ. Dev. Sustain.* **23**, 13324–13339 (2021)
7. Aromal, V., Naseer, M.: Decision-making framework for prioritizing the improvement of pedestrian facilities in urban areas using integrated Delphi, AHP, and TOPSIS approach. *Transp. Res. Rec.* **2677**(1) (2023)
8. Wu, P., Duan, Y.: Discussion on sidewalk width based on street integration. *J. Municipal Technol.* **40**(06), 52–54+79 (2022). (in Chinese)
9. Pei, Y., Pan, Y.: A forecasting method for scale of interchange facilities at urban rail transit stations. *J. Transp. Inf. Safety* **36**(04), 106–112 (2018). (in Chinese)
10. Liang, Z.: Urban park-and-ride planning study. *Highway* **04**, 181–186 (2009). (in Chinese)
11. Liu, Y.: Research of bus stop berth number based on M/G/1 queuing model. *J. Transp. Eng. Inf.* **15**(01), 131–136 (2017). (in Chinese)



Optimization Method of the Turnover Plan of Rolling Stocks of Urban Rail Transit Considering Service Life and Maintenance

Xianghao Kong^(✉), Dan Wei, Jinduo Chen, Wanting Jia, and Hui Li

School of Automobile and Traffic Engineering, Liaoning University of Technology,
Jinzhou 121001, Liaoning, China
sgqhmy02012@163.com, qcdanwei@lnut.edu.cn

Abstract. With the continuous development of urban rail transit, the usage of rolling stock in urban rail transit continues to increase, accelerating the consumption of rolling stocks. How to extend the service life of rolling stocks is of great significance for reducing the operating costs of urban rail transit and improving passenger service levels. In the long-term use of rolling stocks, when the maintenance standards are met, rolling stocks needs to be repaired in a timely manner. In this context, how to ensure the balance of rolling stock usage while completing transportation tasks is of utmost importance. Starting from this practical problem, an optimization method is proposed to design an efficient genetic algorithm for problem solving. Finally, taking subway line 3 in a certain city as an example, the calculation analysis shows that the method proposed in the article can effectively ensure the balance of rolling stock usage, thereby extending the service life of rolling stocks.

Keywords: Urban Rail Transit · Operation Plan of Rolling Stocks · Maintenance Plan of Rolling Stocks · Service Life of Rolling Stocks · Genetic Algorithm

1 Introduction

In recent years, urban rail transit has developed rapidly in China due to its efficient characteristics. This article refers to a train set composed of multiple motor cars and trailers as the rolling stock [1]. The full day train task undertaken by the rolling stock is called the transportation task. Under the long-term rolling stock operation plan, the maintenance plan of the rolling stock needs to be considered [2]. During a train inspection cycle, due to the different operating mileage of the train, the degree of wear and tear is not the same, which will affect the use of maintenance regulations on the rolling stock, leading to the phenomenon of early maintenance and delayed maintenance, ultimately shortening the service life of the rolling stock [3]. In this context, it is very important to ensure the balance of rolling stock usage while completing transportation tasks.

Zhang Bo [4] adopted a segmented periodic cycle solving strategy, and designed corresponding algorithms based on whether maintenance constraints were considered

to solve single day and multi day rolling stock utilization plans separately. Nielsen et al. [5] designed a real-time adjustment method based on rolling time windows and demonstrated the feasibility of real-time adjustment of high-speed train operation plans through practical cases. Wang et al. [6] explored the issues of energy-efficient timetabling and rolling stock circulation planning based on automatic train operation levels for metro lines. By transforming the model, Cplex and Gurobi were used to solve it. Peng Qiyuan et al. [7] designed reasonable neighborhood search operators for transport and inspection tasks, and solved the problem with the help of large-scale neighborhood algorithm. Yuan et al. [8] designed a hybrid algorithm to obtain high-quality solutions by combining genetic algorithms and general solvers. Zhou Wenliang et al. [9] take the energy saving of train operation and the minimization of the total cost of rolling stocks utilization time as the optimization objective, and design an efficient particle swarm algorithm to solve the model. Yuhua et al. [10] proposed a mixed integer linear programming (MILP) model that integrates train schedules, vehicle scheduling, and passenger allocation. The model is solved using an iterative programming heuristic method to meet passenger demand and reduce operating costs. Xu Peijuan et al. [11] established a linear optimization mathematical model to achieve short-term and efficient solution of the optimal value of the model, verifying the efficiency of the synchronous real-time adjustment method.

In this paper, the running map data is used as the initial solution input to realize the optimization of rolling stocks utilization preparation for multi-day urban rail transit, balancing the utilization efficiency of each rolling stock so as to prolong the service life of rolling stocks.

2 Model Construction

2.1 Construct a Multi-Day Rolling Stocks Operation Optimization Programming Problem Network

This paper constructs a long-term rolling stock utilization planning model. The following parameters are defined: the depot is the collection $S = \{s_m | m = 1, 2, 3, \dots, M\}$, s_m is the depot, and M is the total number of the depot; Organize the train number information with the same service number on the train diagram into a transportation task, The set of transportation tasks is $T = \{t_w | w = 1, 2, 3, \dots, W\}$, t_w is the transportation task, its starting station is the starting depot t_{ws} of the first shift, the final station is the finishing depot t_{we} of the last shift, the mileage t_{wl} is recorded, the travel time t_{wt} is recorded, and W is the total number of tasks. Compilation date set $D = \{d_x | x = 1, 2, 3, \dots, X\}$, d_x is compilation date, X is the total number of days of compilation date. According to X , randomly select a suitable transportation task to continue it, and a transportation route tr_i can be obtained. After all transportation tasks are completed, a collection of transportation routes can be obtained $TR = \{tr_i | i = 1, 2, 3, \dots, I\}$, I is the total number of transport routes.

2.2 Optimization Model of Urban Rail Transit Rolling Stocks Planning

Define the following parameters: N_{s_m} is the total number of vehicles stored in the s_m depot, $Tis(t_w, d_x)$ and $Tie(t_w, d_x)$ respectively represent the starting station and the final

station of task t_w on the d_x day of the establishment date. For each transport line tr_i , The 0–1 parameter $tr_i(t_w, d_x)$ indicates that the task t_w is in the transport line tr_i on the d_x day of the compilation date, with 1 indicating that it is in and 0 indicating that it is not. C represents a fixed cost for the use of a rolling stocks, then the transport route cost Z is:

$$Z = C + H \quad (1)$$

$$H = (|\bar{t} - t_i|)\delta + \left(\left|\bar{l} - l_i\right|\right)(1 - \delta) \quad (2)$$

where H is the equilibrium index function; $\bar{t}$ and $\bar{l}$ are the average driving time and average mileage respectively; t_i , l_i are the traveling time and mileage of the transport line tr_i respectively; δ is the correction factor required for time to be converted into cost; $1 - \delta$ is the correction factor for mileage to be converted into cost.

Establish an optimization model for urban rail transit vehicle operation planning with the lowest total operating cost as the objective function.

$$\min \sum_{i=1}^I Z \quad (3)$$

$$\sum_{i=1}^D tr_i \leq N_{s_m} \quad \forall s_m \in R \quad (4)$$

$$\sum_{x=1}^X \sum_{w=1}^W tr_i(t_w, d_x) = 1 \quad tr_i \in TR \quad (5)$$

$$t_{w1e}(t_{w1}, d_x) = t_{w2s}(t_{w2}, d_{x+1}) \quad w1, w2 \in TR, x < X \quad (6)$$

$$\sum_{x=1}^X tr_i(t_w, d_x) * t_{wl} \leq L * 1.1 \quad \forall i \leq I, t_w \in TR \quad (7)$$

3 An Algorithm for Solving an Optimization Model of the Rolling Stocks Utilization Planning Problem in Urban Rail Transit

3.1 Algorithm Analysis

In the urban rail transit rolling stocks planning model established in this paper, the number of transportation tasks is relatively large, so genetic algorithm is used to solve it (Table 1).

Generate the Initial Solution. The encoding form adopted in this paper is the real matrix encoding mode.

When making initial calculations, $C\{tri\} = \emptyset$, $H\{(0)\} = \emptyset$. Each time a new rolling stocks is generated, the first transportation task with the earliest departure time is selected as the starting task from T , and the accumulated mileage l_i and accumulated driving time t_i are recorded at the same time. When the maintenance value is reached, the connection is stopped and maintenance is carried out, and the maintenance task (tri, dx) is scheduled

Table 1. Parameter description

Symbol	Implication
$(t_{wn,m}, d_n)$	The rolling stock m on the n day of the transport task
$C\{tr_i\}$	The transport task set has been completed
$H\{(tr_i, d_x)\}$	Maintenance task set
TR	Transportation route pooling
p_s	Replication probability
p_c	Crossover probability
p_m	Variation probability

for the current rolling stocks. After completing the overhaul task, it can continue until the next overhaul moment occurs, or the entire establishment cycle is completed.

Construct Fitness Function. The fitness function is generally constructed based on the objective function. Therefore, we will utilize the equilibrium function to construct the fitness function i.e.:

$$f_n = \left[\frac{t_i}{T} * \partial + \frac{l_i}{L} * (1 - \partial) \right] * 0.01 \quad (8)$$

After consulting a lot of data and the opinions of relevant experts, the weight of running time and running mileage is set as $\alpha = 0.1$, $1 - \alpha = 0.9$ respectively. Where T is the total running time of all individuals and L is the total running mileage of all individuals.

Genetic Operator.

1) selection

In this paper, the roulette method is used for the selection operation of chromosomes.

Suppose the total population is N and the fitness value of individual n is f_n , then the probability of individual being selected is:

$$P_n = \frac{f_n}{\sum_{n=1}^N f_n} \quad (9)$$

Among them, individuals with very low fitness function value are screened and not genetically manipulated to ensure the excellence of offspring individuals.

2) Crossover

Under a certain probability, two individuals that can be crossed are selected, and the interchangeable interval between them is added to the offspring. It is also necessary to consider the operation state of the rolling stocks and maintenance constraints.

3) Mutation

In this paper, a probabilistic mutation operator based on location is used to mutate some individuals according to a certain probability.

Termination Condition Judgment. Set the number of iterations as the termination condition. When the loop reaches a certain number of iterations, the algorithm ends and outputs the current optimal solution.

3.2 Algorithm Step

Based on the principles and rules of the genetic algorithm mentioned above, the algorithm flow chart shown in Fig. 1 is obtained.

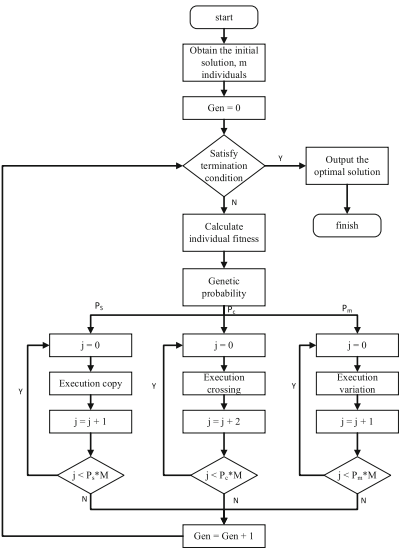


Fig. 1. Algorithm flow chart

4 Case Study

Relying on the actual data of Metro Line 3 in City C on weekdays and days off for analysis. The line has an operating mileage of 67.09 km, with 45 stations and 3 depots.

4.1 Transportation Task Data Collation

The planning period is 14 days. Consider the daily journey tasks completed each rolling stock as a transportation task. 0104 indicates the fourth transport task on day 1 (Table 2).

4.2 Parameter Setting

In this paper, 14 days is a planning cycle, and the maintenance level is biweekly inspection. The reference values are as follows (Table 3):

Table 2. Partial transportation task data

Serial number	Departure depot	Return to the depot	Departure time	Return time	mileage(km)
0101	A	A	5:17:00	10:35:15	156.4
0102	A	A	5:20:00	20:26:35	402.7
0103	A	A	5:23:00	0:04:20	563.1
0104	A	A	5:26:00	22:31:25	498.1
0105	A	A	5:29:00	0:28:15	563.1
0106	A	A	5:32:00	9:28:15	100.1

Table 3. Maintenance specification parameters

category	Overhaul cycle indicator		Maintenance time
	Travel distance(km)	interval	
Biweekly inspection	50000	0.5 month	0.5 day

Considering the global convergence and solving speed of the genetic algorithm, the main control parameter values are set as follows:

Ps: Genetic algorithm replication probability, the value is 0.09;

Pc: Genetic algorithm cross probability, the value is 0.9;

Pm: Genetic algorithm variation probability, the value is 0.01;

Number of iterations: 1000 times.

4.3 Analysis of Results

By using the above algorithm to solve the optimization model, the convergence graph shown in Fig. 2 is obtained, indicating that the method proposed in this paper can quickly achieve the solution of the model. After 246 iterations, the algorithm achieved convergence with a final objective function value of 1686406. Finally, the optimized compilation diagram of rolling stock application shown in Fig. 3 was obtained. In Fig. 3, the horizontal axis represents the number of days, and the vertical axis represents the rolling stocks; The yellow legend, green legend, and blue legend represent transportation tasks with A, B, and C as the starting and ending points, respectively. The black legend represents bi weekly inspection.

By analyzing the mileage graphs of the three garages using rolling stocks (Fig. 4), as shown in Table 4, it can be seen that the mileage of the rolling stocks used at each of the garages is relatively balanced, and overall, the mileage extremes are within 1% of the average mileage used. During the bi-weekly inspection cycle, the utilization rate of the rolling stock is above 90%, which largely reduces the overuse of the rolling stocks as a way to extend the service life of the rolling stock.

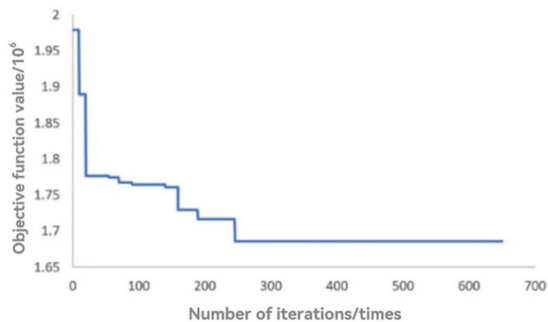


Fig. 2. Algorithm convergence diagram

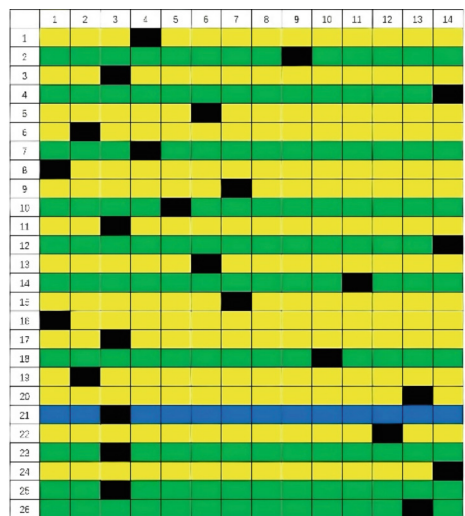


Fig. 3. Partial undercarriage utilization preparation chart

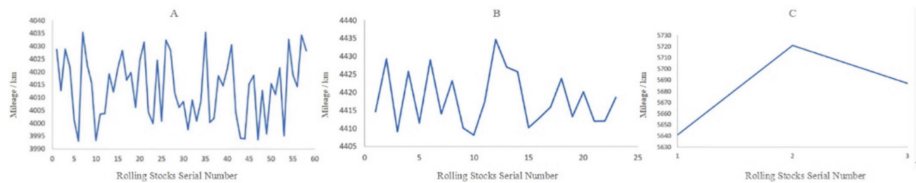


Fig. 4. Mileage statistics of rolling stock depots 1, 2, and 3 respectively

From the optimization results, it can be seen that even in the multi day planning of urban rail transit, the method proposed in this paper can still meet the planning requirements in terms of balancing the service life of rolling stocks, the utilization efficiency of rolling stocks, and the solving efficiency.

Table 4. Rolling stocks mileage of each depot

Depot	Rolling stocks count	Mileage(km)		Average mileage	Very poor mileage
		Maximum	Minimum		
A	56	4035.4	3992.9	4013.7	42.5
B	23	4434.6	4408.2	4418.2	26.4
C	3	5640.9	5721.1	5681	80.2

5 Summarize

The purpose of this article is to consider extending the service life of the rolling stock under the multi day planning conditions of urban rail transit. Based on the C City Subway Line 3, using the proposed optimization method, design an optimization model. The objective function of the model consists of two parts: the number of rolling stocks used to complete all transportation tasks and the balance index function. The results show that this optimization method can not only ensure that each rolling stock is in a balanced state of use while completing transportation tasks within the preparation cycle, effectively improving the service life of the rolling stock, but also provide convenience for the preparation of rolling stock maintenance plans and balancing the daily number of rolling stock maintenance tasks in each depot.

References

1. Wei, G., Mao, B., Wang, Y.: Research on the preparation method of urban rail transit rolling stock utilization plan. *Urban Rail Transp. Res.* **16**(06), 97–100 (2013). (in Chinese)
2. Yang, R.: Research on Optimization Problem of Train Undercarriage Utilization in Urban Rail Transit. Beijing Jiaotong University (2015). (in Chinese)
3. Long, X.: Research on Optimization of Urban Rail Transit Vehicle Maintenance and Utilization Plan. Beijing Jiaotong University (2018). (in Chinese)
4. Zhang, B.: Research and System Development on the Optimization of Underframe Utilization Plan Preparation for Urban Rail Transit Considering Underframe Maintenance Plan. Beijing Jiaotong University (2019). (in Chinese)
5. Nielsen, L.K., Kroon, L., Maróti, G.: A rolling horizon approach for disruption management of railway rolling stock. *Eur. J. Oper. Res.* **220**(2), 496–509 (2019)
6. Wang, Y., Zhu, S., D'Ariano, A., et al.: Energy-efficient timetabling and rolling stock circulation planning based on automatic train operation levels for metro lines. *Transp. Res. Part C: Emerg. Technol.* **129**, 103209 (2021)
7. Peng, Q.-Y., Liao, P.-R., Biu, S., et al.: Synergistic optimization of underfloor operation and maintenance of urban rail transit under all-weather operation conditions. *China Railway Sci.* **43**(06), 194–206 (2022). (in Chinese)
8. Yuan, J., Gao, Y., Li, S., et al.: Integrated optimization of train timetable, rolling stock assignment and short-turning strategy for a metro line. *Eur. J. Oper. Res.* **301**(3), 855–874 (2022)
9. Zhou, W., Huang, Y., Deng, L.: Optimization of urban rail schedules considering operational energy saving and underframe utilization. *J. Railway Sci. Eng.* **20**(02), 473–482 (2023). (in Chinese)

10. Yuhua, Y., Marcella, S., Dario, P., et al.: Train timetabling with passenger data and heterogeneous rolling stocks circulation on urban rail transit line. *Soft. Comput.* **27**(18), 12959–12977 (2022)
11. Xu, P., Niu, Y., Qin, Y., et al.: Synchronized real-time adjustment method of subway train operation and underframe turnover plan. *J. Railway Sci. Eng.* 1–12 (2023). (in Chinese)



Short-Term Passenger Flow Prediction of Urban Rail Transit Based on Station Clustering and ConvLSTM

Zhengda Tian^(✉), Dan Wei, Lei Huang, Yunzhou Tan, and Lvqi Yao

School of Automobile and Traffic Engineering, Liaoning University of Technology,
Jinzhou 121001, China

tianzhengda1@163.com, qcdanwei@lnut.edu.cn

Abstract. As the scale of urban rail transit systems in many of our cities gradually expands, the accuracy of short-term passenger flow forecasting, which serves as a basis for train operation planning, has become increasingly important with the surge in passenger traffic. To improve the short-term passenger flow forecasting accuracy for subway stations, an optimization model, ConvLSTM is proposed. The model clusters similar stations and dates by K-Means clustering algorithm, and based on the clustering results, feature extraction and prediction are performed using one-dimensional convolutional neural network and Long Short-Term Memory (LSTM). Taking the AFC swipe data of Wuhan for three weeks in 2022 as an example, the 15min granularity inbound passenger flow is predicted using the ConvLSTM model, and the ARIMA, SVM, LSTM, and ConvLSTM models are evaluated by two evaluation metrics (MAPE, RMSE). The results show that the stations of the whole network can be categorized into: Hybrid station with a commercial bias, Residential Station, Hybrid station, Hub Station, Hybrid station with a residential bias; and the days can be categorized into: weekdays, Saturdays and Sundays. The ConvLSTM model has the highest prediction accuracy on hybrid stations, with Mean Absolute Percentage Error of 6.886%, 10.546%, and 11.793% on weekdays, Saturdays, and Sundays, respectively, as well as outperforming the control model on a sample of six types of stations. It shows that ConvLSTM has higher prediction accuracy in short-term prediction than single model.

Keywords: subway · short-term passenger flow prediction · station clustering · CNN · LSTM

1 Introduction

By summarizing and analyzing the characteristics of passenger flow and extracting the key factors that affect the changes in passenger flow, combined with these characteristics for scientific prediction of subway passenger flow, it can better assist subway operators in making decisions [1]. Before conducting passenger flow forecasting for rail transit, it is common to analyze the historical passenger flow characteristics from different perspectives, including temporal and spatial angles [2].

The short-term passenger flow forecasting of urban rail transit is mainly based on mathematical and statistical models at the initial stage [3], such as historical average model, exponential smoothing model, ARMA model, etc. However, due to the high requirement of timeliness for short-term passenger flow prediction, these single models cannot meet this requirement, and many scholars combine multiple models for prediction [4]. Javad Abdi [5] and others proposed a combined prediction model that combines a fuzzy neural network model with a time-varying learning model. As variants of RNN models, LSTM and GRU models and related variants have also been applied by some scholars in the field of passenger flow prediction [6–9]. Long Short-Term Memory models are a special kind of Recurrent Neural Networks [10, 11].

In this paper, a combined model based on K-clustering algorithm and ConvLSTM is proposed to classify the passenger flow in urban railway stations into station attributes and date attributes, extract the similar station attribute passenger flow and date attribute passenger flow in two steps by K-Means clustering algorithm, and input the similar feature passenger flow into ConvLSTM prediction model.

2 Model Building

2.1 Station Clustering Analysis

Clustering models can be divided into hierarchical clustering, K-Means clustering, EM algorithm and DBSCAN algorithm. In this paper, K-Means clustering is selected, Elbow method and contour coefficient method are used to determine the number of clusters k , and Z-SCORE method is selected to standardize the clustering variables in order to optimize the K-Means clustering algorithm. The station attribute characteristics are determined according to the station passenger flow distribution characteristics; the intra-week passenger flow time characteristics are determined according to the weekly passenger flow distribution. The division of each time period is shown in Table 1:

Table 1. Segmentation of passenger flow periods

designation	work shift
morning peak	7:00–9:00
evening peak	10:00–16:00
Mt Pingping	17:00–19:00

The clustering factors are defined as shown in Table 2:

Table 2. Cluster factor definition

Clustering factor number	Clustering factor definition	data type
A1	Average weekday morning peak passenger flow/total inbound passenger flow	numerical value
A2	Average weekday evening peak passenger flow/total inbound passenger flow	numerical value
A3	Average Weekday Peak Passenger Flow/Total Inbound Passenger Flow	numerical value
A4	Average Saturday morning peak passenger flow/total inbound passenger flow	numerical value
A5	Average Saturday Peak Inbound Traffic/total inbound passenger flow	numerical value
A6	Average Saturday Night Peak Passenger flow/total inbound passenger flow	numerical value
A7	Average Sunday Morning Passenger Flow/Total Inbound Passenger Flow	numerical value
A8	Sunday Flat Peak Traffic/Total Inbound Traffic	numerical value
A9	Sunday evening peak passenger flow/total inbound passenger flow	numerical value
A10	Average of peak weekday traffic	numerical value
A11	Peak passenger flow on Saturday	numerical value
A12	Peak passenger flow on Sunday	numerical value

The combination of A1–A12 clustering factors is used as a vector factor for the first step of clustering, denoted as $A = [A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12]$. Next, B1–B17 are defined as the hourly passenger flow at the station from 6:00–23:00 as the vector factor for the third clustering step, denoted as $B = [B1, B2, B3, \dots, B16, B17]$.

2.2 Predictive Modelling

In this section, based on the above clustering results, a combined model of clustering algorithm and ConvLSTM is constructed to predict station inbound passenger flow. The clustering algorithm is used to obtain station categories. The model is trained on three consecutive weeks of data using station samples from different categories. Then the next week's data is predicted and the prediction effect of subcategories is calculated.

3 Example Analysis

3.1 Datasets

Wuhan subway swipe card data in March 2022 was selected to test the prediction model. Swipe data contains a time stamp, station name, station number, entry/exit type, and card number. Match and sort the card data by certain rules to get the 15-min out passenger flow at each station. The data structure is shown in Table 3:

Table 3. Data structure of 15min passenger flow at the station

Field	Typology
Date	Date
Station Number	Varchar
Inbound traffic	Integer
Outbound traffic	Integer
Starting time	timestamp
Ending time	timestamp

Four weeks of complete station passenger flow data are selected, with the first three weeks as the training set and the last week as the test set.

3.2 Clustering Result Recognition

From Fig. 1, it is clear that the first step clustering is at a low level of SSE and a high level of profile coefficient at a classification number of 6.

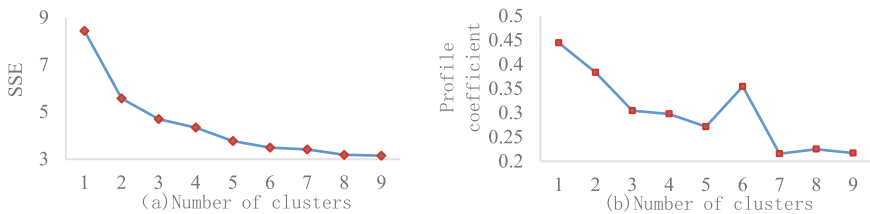


Fig. 1. Cluster k-value assessment

The K-Means algorithm $k = 6$ was selected and some of the clustering results are shown in Table 4.

The changes in inbound and outbound passenger flows at the six types of stations are shown in Fig. 2.

Table 4. Cluster results of station parts

station name	First step clustering categories
Chong ren Road	I
Crab Point	I
National Expo Centre South	II
Hanyang Railway Station	II
Long yang village	III
Luo Jia Zhuang	III
Dajun Mountains	IV
Wuchang railway station	IV
Che Cheng Dong Road	V
Wuhan Business District Station	V
Military Games Village	VI
Zhanjiang Road	VI

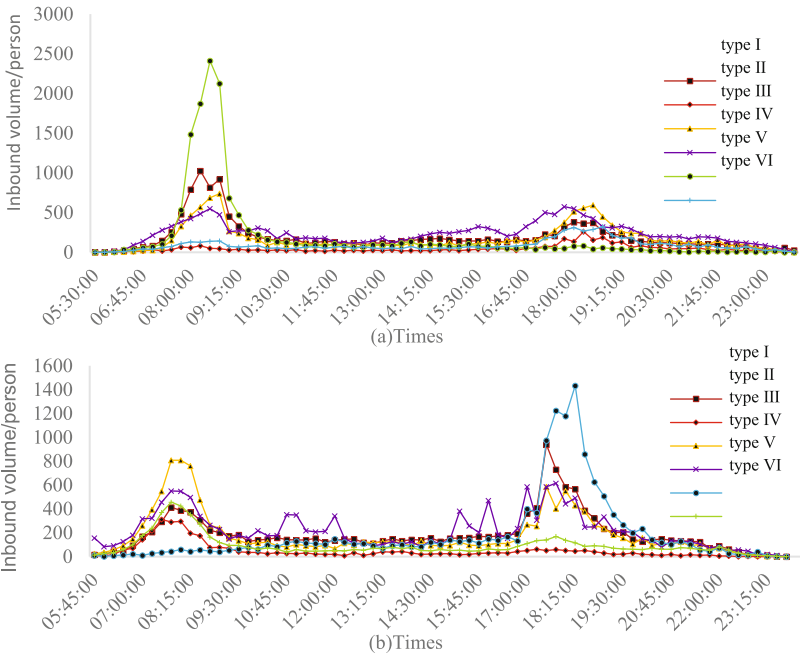


Fig. 2. Passenger flow entering and exiting the station by category

Type I stations, which are characterized by low morning peaks and high evening peaks of inbound traffic, are defined as mixed-occupation, residential-occupational stations. Similarly, Type II stations are residential stations. Type III stations are mixed type stations, Type IV stations are hub stations, Type V stations are vocational stations, and Type VI stations are mixed-occupancy biased residential stations. The intra-week category division is formed based on station characteristics. Category I is Monday to Friday, Category II is Saturday, and Category III is Sunday.

The weekdays, Saturdays and Sundays were trained using the ConvLSTM model respectively. Six typical stations were selected as training and validation samples for the prediction model, Crab Point Station, National Expo Centre South, Long yang Village Station, Wuchang Railway Station, Wuhan Business District Station, and Zhanjiang Road Station.

3.3 Analysis of Projected Results

The ConvLSTM model is used to predict six types of station samples. After repeated testing, this chapter uses a 15-min time interval, a convolutional kernel with a model size of 2*2, maximum pooling, 64 neurons in the CNN layer, 100 hidden neurons in the LSTM layer, a time window of 36 iterations until convergence, an activation function of ReLU, and an optimization algorithm of Adam, and the ARMIA model is used in the prediction performance comparison. In the prediction performance comparison, the ARMIA model is used, whose model order is determined by the BIC criterion; the single LSTM model is used, whose hyperparameter settings are the same as those of the LSTM layer in the ConvLSTM model; and the SVM model is used, whose kernel function is the radial basis kernel function. The (MAPE) and (RMSE) statistical metrics are used for comparison, as shown in Table 5.

Table 5. Evaluation of model prediction results

Assessment of indicators	Station type	weekly category	ARIMA	SVM	LSTM	ConvLSTM
MAPE	I	Weekday	22.900	21.892	18.144	15.120
		Saturday	20.345	20.223	26.522	15.855
		Sunday	27.398	27.276	24.011	16.904
	II	Weekday	20.894	20.456	17.908	12.181
		Saturday	22.457	22.097	19.026	16.433
		Sunday	29.126	28.896	26.685	19.211
	III	Weekday	13.088	12.748	9.991	6.886
		Saturday	16.565	16.416	14.791	10.546
		Sunday	21.245	20.885	15.211	11.793

(continued)

Table 5. (continued)

Assessment of indicators	Station type	weekly category	ARIMA	SVM	LSTM	ConvLSTM
	IV	Weekday	31.958	30.424	27.625	23.021
		Saturday	35.057	33.690	27.592	22.993
		Sunday	32.896	31.251	27.157	22.631
	V	Weekday	18.389	17.653	14.617	11.676
		Saturday	24.172	23.374	20.920	13.768
		Sunday	23.337	23.097	20.523	14.923
	VI	Weekday	24.690	24.937	11.054	17.545
		Saturday	21.112	20.119	17.041	18.201
		Sunday	26.852	26.801	15.054	17.103
RMSE	I	Weekday	36.299	34.629	31.687	26.406
		Saturday	33.098	31.609	37.057	28.134
		Sunday	45.907	45.448	38.656	22.645
	II	Weekday	37.033	37.366	28.188	23.526
		Saturday	40.114	39.953	33.760	28.159
		Sunday	41.217	40.499	37.174	27.907
	III	Weekday	22.117	21.874	22.405	16.643
		Saturday	38.588	37.855	34.789	20.621
		Sunday	33.348	42.555	29.883	17.686
	IV	Weekday	52.462	51.727	46.813	39.011
		Saturday	45.782	43.905	37.407	31.173
		Sunday	40.748	39.852	35.986	29.989
	V	Weekday	32.653	32.751	28.231	32.230
		Saturday	41.570	41.820	33.790	22.548
		Sunday	44.678	4.308	32.670	23.490
	VI	Weekday	37.740	36.381	21.488	27.225
		Saturday	49.613	48.422	40.142	33.452
		Sunday	49.027	46.919	43.025	35.854

From Table 5, it can be seen that the prediction accuracy of ConvLSTM for weekdays, Saturdays and Sundays is better than other models. The prediction accuracies for the six types of stations are $III > VI > II > I > V > IV$, as shown in Fig. 3.

The correlation coefficients were tallied and described using the Pearson correlation coefficient. The average correlation coefficients for specific categories are shown in Table 6.

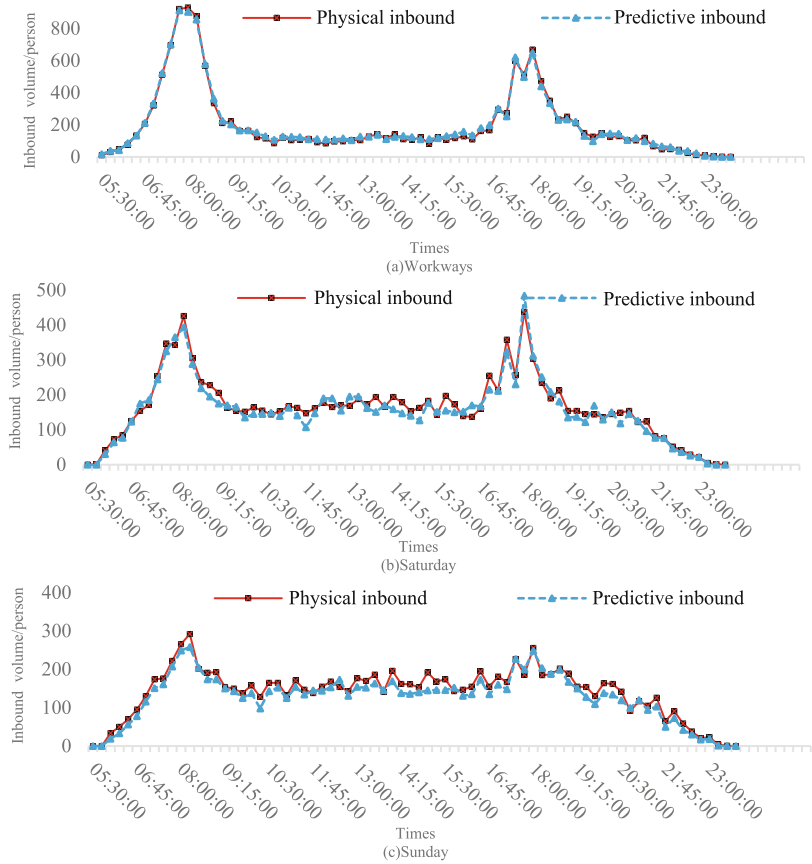


Fig. 3. Forecast results of inbound passenger flow for each week category

Table 6. Average correlation coefficient of each category

form	Average correlation coefficient for the working day	Average Saturday correlation coefficient	Average weekday correlation coefficient
I	0.81	0.63	0.61
II	0.85	0.66	0.63
III	0.92	0.72	0.72
IV	0.73	0.52	0.52
V	0.90	0.67	0.69
VI	0.83	0.58	0.55

4 Conclusion

Using Wuhan metro swipe card data for validation, the ConvLSTM model shows an average decrease of 3.82%, 6.34%, and 6.01% in the absolute percentage errors on weekdays, Saturdays, and Sundays, and an average decrease of 3.96, 8.81, and 9.97 in the root-mean-square errors, compared to the LSTM model, suggesting that ConvLSTM is better than a single-model network in short-term prediction.

References

1. Cao, Y., Sun, Y., Lin, L., et al.: Research on short-time inbound passenger flow prediction of urban rail transit based on CNN-LSTM. *Traffic Transp.* **40**(02), 94–99 (2024). (in Chinese)
2. Gao, D., Xu, Q., Chen, P., et al.: Spatial characteristics of urban rail transit passenger flows and fine-scale built environment. *J. Transp. Syst. Eng. Inf. Technol.* **21**(06), 25–32 (2021). (in Chinese)
3. Zeng, C., Wu, J., Luo, X.: Literature review of short-term passenger flow forecast for urban rail transit. *Railway Transp. Econ.* **43**(8), 105–111 (2021). (in Chinese)
4. Ni, J., Yu, L., Jing, X.: Forecast and research of urban rail transit passenger flow based on ARIMA model. *Intell. Comput. Appl.* **11**(04), 135–138 (2021). (in Chinese)
5. Abdi, J., Moshiri, B., Abdulhai, B., et al.: Short-term traffic flow forecasting: parametric and nonparametric approaches via emotional temporal difference learning. *Neural Comput. Appl.* **23**, 141–159 (2013)
6. Ma, X., Tao, Z., Wang, Y., et al.: Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transp. Res. Part C: Emerg. Technol.* **54**, 187–197 (2015)
7. Zhao, Z., Chen, W., Wu, X., et al.: LSTM network: a deep learning approach for short-term traffic forecast. *IET Intell. Transp. Syst.* **11**(2), 68–75 (2017)
8. Tang, Q., Yang, M., Yang, Y.: ST-LSTM: a deep learning approach combined spatio-temporal features for short-term forecast in rail transit. *J. Adv. Transp.* 1–8 (2019)
9. Yu, W., Xia, S., Lan, Jie., et al.: A comparative study of passenger flow prediction methods in Subway Stations. *Intell. Rail Transp.* **61**(04), 98–103 (2024). (in Chinese)
10. Qian, X.: Incorporating CNN-LSTM and SVM with wavelet transform methods for tourist passenger flow prediction. *Soft. Comput.* **28**(3), 2719–2736 (2024)
11. Bao, J., Yu, H., Wu, J.: Short-term FFBS demand prediction with multi-source data in a hybrid deep learning framework. *ET Intell. Trans. Syst.* **13**(9), 1340–1347 (2024)



Study and Implementation of File Transparency Encryption Technology for Intelligent Connected Vehicles Based on SM4

Qi Wang¹, Yongzhong Li¹(✉) , and Yuheng Li²

¹ School of Automotive Engineering, Nantong Institute of Technology, Nantong 226010, China
wqi.003@126.com

² Internet of Things and Big Data Major of Macao Polytechnic University, Macao, China

Abstract. With the developing of vehicle networking technology and artificial intelligence technology, intelligent connected vehicles are facing more and more serious cybersecurity threats and challenges. The privacy data in intelligent connected vehicles is becoming increasingly urgent. For privacy information protection problem of Android platform, we proposed a new transparent encryption scheme based on SM4 algorithm for intelligent connected vehicles. A transparent encryption method for Android platform is designed and implemented. The secret-key is generated by the Phone MAC address that fundamentally guarantees the security of user information. The system design adopts the MVP design pattern and completes the file transparent encryption system. Finally the system test results proven that the transparent encryption system can effectively encrypt and decrypt the specified files to protect the privacy data for Android smart terminal of intelligent connected vehicles.

Keywords: Intelligent Connected Vehicles · File Transparency Encryption · SM4 · Android

1 Introduction Preface

With the rapidly develop of Internet of Vehicles and intelligence technology in recent years. The concepts such as V2V, V2I and even V2X has revealed a new mode of future traffics. However, while Internet of Vehicles brings convenience to people, the security problem of the Internet of Vehicles is gradually emerging [1]. Especially the attack of the attacker on the vehicle-mounted network is likely to cause serious consequences such as the loss of control and privacy leakage [2]. At present, many car manufacturers use mobile terminal APPs to remotely control the air conditioner, open and close car doors, and start the car [3]. Connecting the APP to the car's local area network may pose a risk of being implanted with malicious code. In addition, mobile terminals have the risk of privacy data leakage after sensitive driver data is attacked, especially when obtaining account passwords and other private data, which directly affects the driving safety of the car.

At present, with the privacy information protection technology's developing, file transparent encryption is becoming mature [4–6]. However, file transparent encryption method is originally used in Windows system, and the application software for file encryption of Android platform is not file transparent encryption and the methods are uneven. Users of Android platform must enter the passwords to verify encrypt and decrypt files every time. This inconvenience greatly reduces efficiency of encryption and the user operation [4]. The file transparent encryption technology can be classified to two classes: the hook function transparent encryption and the file filter driven encryption [4–6]. In this article, we proved a transparent encryption system based on Android file filter drive.

2 Developing Architecture of Android and Algorithm SM4

2.1 Developing Architecture of Android System

The developing architecture of Android platform is based on Linux system and it is a bottom-up developing structure. It is consisting of four layers, that is Application layer, Application framework, Android library and Linux Kernel layer, as shown in Fig. 1 [17].

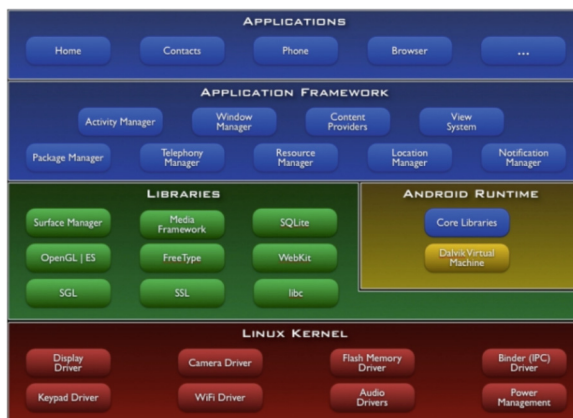


Fig. 1. Android developing system architecture

2.2 Encryption Algorithms SM4 and AES

Algorithm SM4. Algorithm SM4 [8] is also a block cipher method. Algorithm SM4 is similar to the algorithm DES and AES, but the architecture of SM4 is different from DES and AES. SM4 is designed four branch generalized unbalanced Feistel structure. For SM4 algorithm, the encryption process structure and the decryption structure are same procedures, except the round keys are used opposite from encryption, that is the order of decryption keys is opposite to the order of encryption keys. The algorithm structure is shown in Fig. 2.

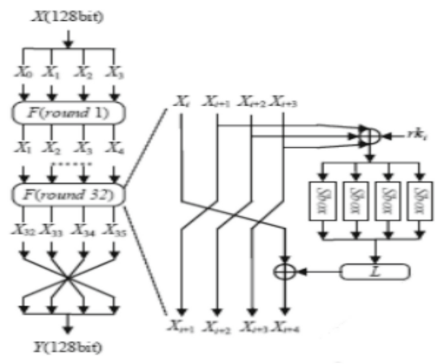


Fig. 2. Structure of SM4 for Encryption and decryption

Algorithm AES. AES is a block advanced encryption standard developed by NIST of United States. The plaintext of AES algorithm is a 128-bit or 16-Bytelength and the byte length of secret key of AES can be selected from 128-bits, 192-bits, or 256-bits [8–10]. The AES Algorithm are proved four operations functions to encrypt plaintext. Encrypt functions are AddRoundKey(), SubBytes(), ShiftRows(), and MixColumns(). The decryption functions are AddRoundKey(), InvSubBytes(), InvShiftRows(), and InvMixColumns(). The function AddRoundKey() is self-invertible [10].

3 Transparent Encryption Method

File transparent encryption is referred to the procedures of encrypting and decrypting files that is transparent and imperceptible to users. It does not affect the user’s operation. It is does not need to input encryption or decryption keys for user [5–11]. Transparent encryption method generally can be divided into two type’s encryption technologies, that is hook function interrupted transparent encryption and file filter-driven transparent encryption. According to encryption process and the efficiency, file filter-driven transparency technology is encrypting and decrypting the file in the driver layer, it has a high efficiency in driver layer. So we adopt file filter-driven transparent encryption technology in this paper (Fig. 3).

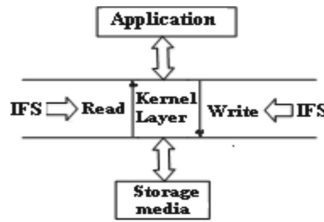


Fig. 3. Transparent Encryption of File Drives

4 Design and Implementation of File Transparent Encryption System for Android

4.1 Overall Designs

The file transparent encryption system is designed for Android platform. It is designed for the specified files encryption and decryption. The design frame of file transparent encryption system is shown in Fig. 4.

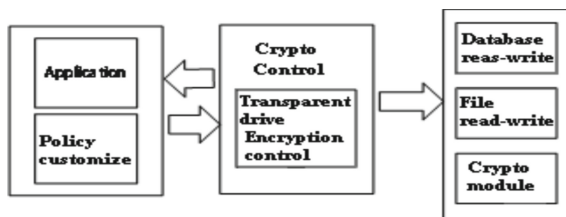


Fig. 4. Design Framework of Transparent Encryption System

The transparent encryption system design is using MVP (Model, View, and Presenter) framework. MVP is evolved from MVC framework [16]. For encryption algorithm, the SM4 and AES encryption algorithm is adapted in our file transparent encryption system. SM4 is developed by China, and AES is mostly used encryption algorithm. The system authentication of the user is used by screen lock when the Android phone is to open and starts.

4.2 Encryption and Decryption Modules Design and Implementation

The encryption module's performance directly affect the security and effectiveness of whole systems [17–20]. So in the encryption algorithm module, we have designed two optional algorithms: SM4 algorithm and AES algorithm. Figure 5 shows the encryption modules diagram. Whether using algorithm of SM4 or AES, the functions to be implemented by the encryption module are consistent.



Fig. 5. Encryption modules diagram

Transparent encryption technology does not require users to input a key. The method of generating keys for encrypt and decrypt are involved to read the Android terminal's MAC address, it is generally placed under `/sys/class/net/plan0`. For MAC address performing some rounds of substitution and other operations then form a key, which is only

used as the current device’s key and transmitted it to the algorithm of encryption and decryption in the system kernel module for use. Because of each Android device has its only one key. So this feature plays an important role to protect the privacy data of Android files.

For the decryption algorithm module, the process is similar to the encryption process, the only different is using the algorithm. The diagram of decryption module is similar to Fig. 5.

4.3 The System Design and Implementation

The overall design of the file transparent encryption system is involved to the design of application layer module and kernel module. The function of the application layer module design is to complete the selection and customization of encryption strategies (SM4 or AES algorithm), as well as the functions of interacting with users and authentication of user. The kernel module completes system monitoring, file transparent encryption and decryption functions, and file data reading and writing functions based on the selection of algorithm and to set the module of application. The design diagram of the system is shown in Fig. 6.

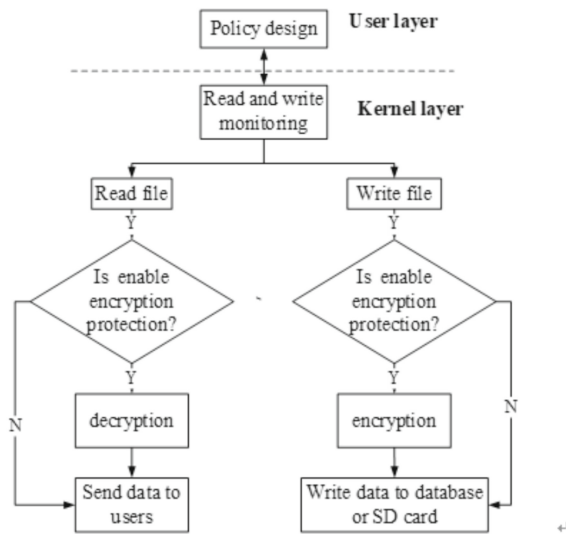


Fig. 6. The diagram of overall design

In the Android, most kernel level core services require functional support from the Linux system kernel layer. Because of the transparent encryption technology of file filtering driver is used in this system, so it is used of the file read and write monitoring module requires the Linux system kernel functions to achieve the system file read and write monitoring. In the file system of Linux, an index node is used to associate with the file system, which includes information about files, file directories, and with hardware

devices. The file index node link one-to-one with each file in the Linux system. So it is easily to control the all files, directories and hardware devices with the nodes on Linux kernel. The Linux kernel files read-write hierarchy is shown in Fig. 7.

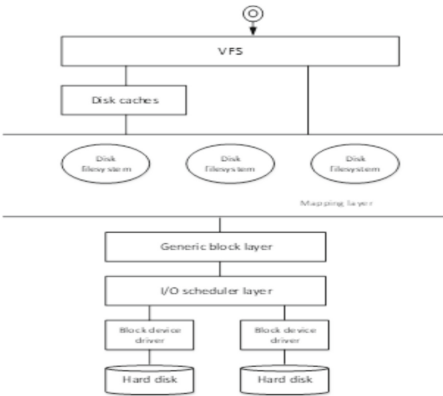


Fig. 7. The Linux kernel files read-write hierarchy

4.4 Linux Kernel Module Design

Linux system manages files through index nodes, while Linux manages files through system call functions. The system call function of Linux is implemented through soft interrupts. In Linux system, it has provided a program to load the kernel modules, called LKM (Loadable Kernel Module), LKM is used to dynamically add and extend the function of the Linux kernel layer [11]. The kernel module diagram is shown in Fig. 8.

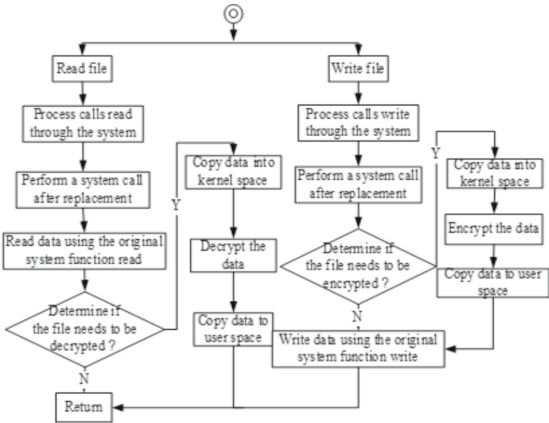
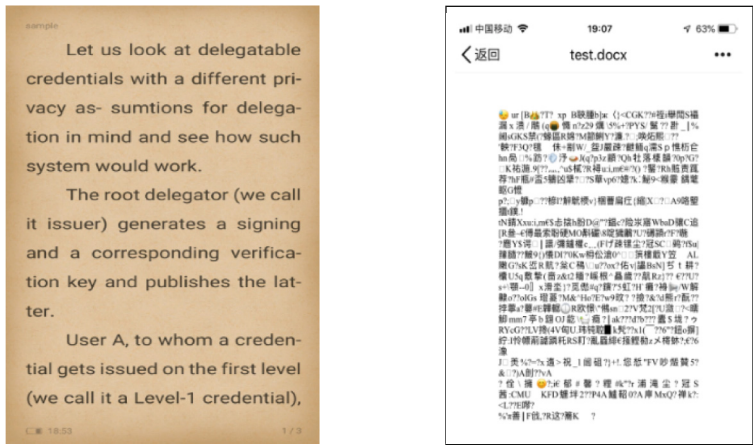


Fig. 8. The Diagram of Kernel Module

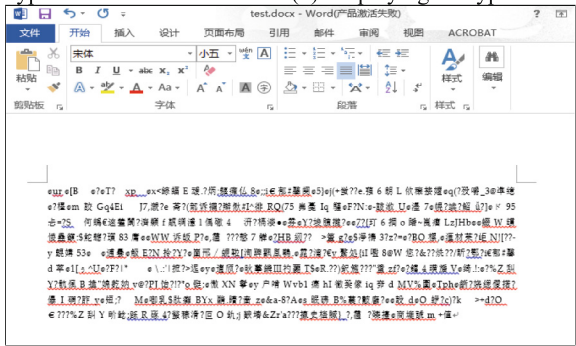
When user is needed to read a file, the system performs the process of reading the file, which is the decryption process. When performing the decryption process, the interception system will call the read program, and the kernel module monitoring system will obtain the file name and file structure. Verifying the file name and file structure in encryption policy that selected from the application layer, if the file name and structure matched to the encryption policy, the file is copied into the kernel layer and decrypt it. The key is the MAC address transformation key obtained from application layer. When user is needed to write a file, the system performs the process of writing the file, which is the encryption process.

5 System Testing

The file transparent encryption system was packaged into a mobile terminal APP and installed on API18 emulator and Android system smartphones for system function testing. The transparent encryption system for txt text file and doc file formats on Android



(a) Displaying encrypted DOC on local Phone (b) Displaying encrypted DOC on other Phone

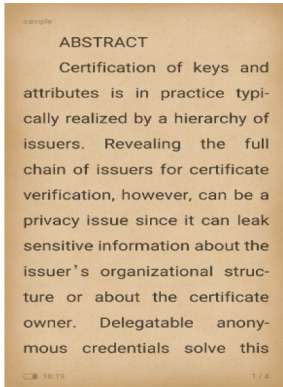


(C) Displaying encrypted DOC File on PC

Fig. 9. Results of encrypted DOC File

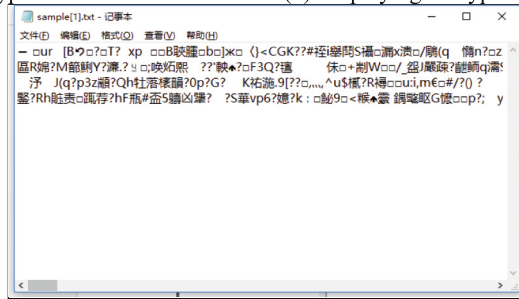
terminal SD cards was successfully implemented. After successful encryption, the file is displayed normally on the local Android phone. Changing the terminal or displayed it on a PC or another phone terminal is garbled, thus achieving privacy transparent encryption protection for Android terminal files. As shown in Fig. 9, these are the results of the DOC file protection test.

As shown in Fig. 10, these are the results of the TXT file protection test.



(a) Displaying encrypted TXT on local Phone

(b) Displaying encrypted TXT on other Phone



(C) Displaying encrypted TXT File on PC

Fig. 10. Results of encrypted TXT File

6 Conclusions

In this paper, we provide a file transparent encryption system based on the Android system. This method employs file filtering driven transparent encryption technology to intercept system calls and obtain keys through host MAC address translation. The algorithms of encryption and decryption were written into the system kernel layer to complete transparent encryption and decryption. Finally, the system examine results show that transparent file encryption system can effectively protect the privacy data.

Acknowledgements. This project supported by Doctor Research Initiation Fund Project of Nantong Institute of Technology (2023KK (B) 02), 2024 China Private Education Association Planning Project (CANFZG24287), 2023 Nantong University of Technology Teaching and Research Reform Project (2023JJG003).

References

1. Yan, W.: Car hackers are busy - how to secure the internet of vehicles (EB/OL). 01 May 2015. <https://club.autohome.com.cn/bbs/threadowner/08de703dc7dd8079/40936819-1.htm>. (1 Ch)
2. Shukla, A.S., et al.: Entropy-based anomaly detection in a network. *Wirel. Pers. Commun. Int. J.* **99**(4), 1487–1501 (2018)
3. Nan, Y.: Kang rongbao analysis of security threats and protection strategies for connected vehicles. *Commun. Technol.* **12**, 1421–1426 (2015)
4. Shen, C.: Research and Design of File Transparent Encryption Technology Based on Android Platform, M.S. dissertation, Jiangsu University of Science and Technology, China (2018)
5. Wang, Y.-M., Li, Y.-Z., Lu, S.-W.: Research and implementation of file transparent encryption technology based on android. *Comput. Technol. Dev.* **09**, 137–140 (2014)
6. Mei, K.-Z., Li, Y.-Z.: Transparent file safety encryption method of enterprise LAN based on filter driver. *Comput. Technol. Dev.* **04**, 238–241 (2012)
7. Guo, L.: *The First Line of Code-Android*, Second Edition. The People's Posts and Telecommunications Press, Beijing (2016)
8. Office of State Commercial Cipher Administration. Block Cipher for WLAN Products-SMS4 [EB/OL]. <http://www.oscca.gov.cn/UpFile/2006021016423197990.pdf>
9. Zhu, S.-B.: Design and Implementation of AES Algorithm Based on FPGA, M.S. dissertation, West Normal University, China (2016)
10. Luo, Z.-Y.: Improvement and Application of the AES Algorithm Based on Multi-core Android Platform, M.S. dissertation, Shanghai Normal University, China (2016)
11. Yang, D., Ye, P., Zhen, L.: The application of transparent decryption in trusted storage of electronic documents. *Electron. Sci. Technol.* **04**(04), 147–150 (2017)
12. Qian, J.: Design and Implementation of Phone Privacy Protection System Based on Android", M.S. dissertation, University of Electronic Science and Technology, China, (2014)
13. Wang, Q., Zhou, Q., Liu, Y., et al.: Research on file system transparent encryption techniques. *Comput. Technol. Dev.* **03**, 147–150 (2010)
14. Zhou, S.: *The Reveal and Prevention in Android Security Technology*. The People's Posts and Telecommunications Press (2015)
15. Sun, X., Yang, T., Hu, X.: A file system protection scheme based on file filter driver and true crypt. *Modern Comput. (professional)* **03**, 77–80 (2016)
16. Zeng, L.: Application research of MVP for android. *Comput. Eng. Softw.* **06**, 75–78 (2016)
17. Fu, C.-J.: Design of transparent encryption system for documents based on windows kernel. *J. Chongqing Univ. Educ.* **28**(03), 171–173 (2015)
18. Alshambri, H.A., Alassery, F.: Securing fog computing for e-learning system using integration of two encryption algorithms. *J. Cyber Secur.* **3**(3), 149–166 (2021)
19. Kumar, T.M., Karthigaikumar, P.: Implementation of a high-speed and high-throughput advanced encryption standard. *Intell. Autom. Soft Comput.* **31**(2), 1025–1036 (2022)
20. Xu, C., Mei, L., Cheng, J., Zhao, Y., Zuo, C.: IoT services: realizing private real-time detection via authenticated 'conjunctive searchable encryption. *J. Cyber Secur.* **3**(1), 55–67 (2021)



Cognition and Modeling of Urban Low-Altitude Airspace

Sidao Chen^{1,2}, Xuejun Zhang^{1,2(✉)}, and Weidong Zhang³

¹ School of Electronic Information Engineering, Beihang University, Beijing 100191, China
{chensd, zhxj}@buaa.edu.cn

² State Key Laboratory of CNS/ATM, Beihang University, Beijing 100191, China

³ Beijing Key Laboratory for Network-Based Cooperative Air Traffic Management,
Beihang University, Beijing 100191, China
vidon@buaa.edu.cn

Abstract. The rapid advancement of UAV technology has highlighted the importance of managing low-altitude airspace and planning efficient, safe flight paths in complex urban environments, where refined airspace modeling and connectivity analysis are essential for effective UAV operations. In this study, we propose a low-altitude airspace modeling and visibility analysis method based on the Alpha-Shape algorithm and space syntax theory to address the challenge of UAV airspace cognition in complex urban environments. Utilizing the Alpha-Shape algorithm, we construct the reachable urban airspace that encompasses buildings and obstacles, adapting to various flight requirements by adjusting algorithm parameters. We then generate visibility graphs through isovist analysis to quantitatively evaluate the spatial connectivity of different regions within the airspace. Case studies of complex urban areas demonstrate that densely built environments significantly obstruct UAV flights, while zones with open views exhibit superior connectivity, making them ideal for UAV path planning. This method is shown to be feasible to UAV path planning and can be further extended to related fields such as low-altitude route design, critical hub node identification and takeoff and landing site selection.

Keywords: UAV · Airspace · Space Syntax

1 Introduction

In recent years, the rapid development of unmanned aerial vehicle (UAV) technology has made low-altitude airspace management and path planning critical topics in UAV applications. Particularly in urban environments, the widespread use of UAVs necessitates refined airspace cognition and the design of efficient flight routes. Buildings and obstacles within low-altitude airspace create a complex three-dimensional flight airspace. Defining reachable airspace in such airspace and effectively assessing its connectivity are fundamental to ensuring the safe and efficient operation of UAVs. Therefore, precise modeling and analysis of low-altitude airspace in complex urban environments are of significant importance for the extensive application of UAVs.

Existing research primarily focuses on UAV path planning and dynamic airspace management [1, 2], however, systematic and quantitative analyses of low-altitude airspace structure and accessibility are still lacking. Complex buildings and obstacles in urban environments significantly impact airspace reachability and connectivity. Traditional path planning methods often rely on simplified two-dimensional or three-dimensional models, neglecting these critical factors. This limitation reduces planning precision and increases the risk of UAVs encountering unforeseen obstacles during actual flights, thereby affecting flight safety and efficiency. Moreover, urban low-altitude airspace is constrained by multiple spatial factors; building density and height variations severely limit available flight space for UAVs. In high-density building clusters and narrow airspaces, UAV flight connectivity and visibility are particularly restricted. Therefore, precise modeling of complex three-dimensional urban airspace and systematic quantitative analysis of its reachability and connectivity are imperative.

To address these issues, we propose a low-altitude airspace cognition and modeling method based on the Alpha-Shape algorithm and space syntax theory. By following the approach in [3], we utilize the Alpha-Shape algorithm to construct the reachable airspace for UAVs in urban environments as the first step. This algorithm accurately captures the influence of buildings and other three-dimensional obstacles. By adjusting algorithm parameters, the Alpha-Shape algorithm adapts to airspace morphological characteristics in different urban settings and further constructs the reachable airspace for low-altitude UAV flights. Then, based on the constructed reachable airspace, we use space syntax to quantify the field of view and connectivity represented as isovist values at different locations to generate the visibility graph. By evaluating the isovist values, we can identify areas with minimal obstructions and strong connectivity, providing quantitative data for UAV takeoff and landing site selection and path planning, and laying a data foundation for low-altitude route design. Further, conducting in-depth analysis of the visibility graphs can enhance understanding of UAV operational characteristics in complex urban environments.

The primary contribution of this study lies in presenting a low-altitude airspace modeling and analysis method that integrates the Alpha-Shape algorithm with space syntax theory. Unlike previous research that predominantly focuses on path planning, this work emphasizes airspace reachability and connectivity issues, offering new perspectives and analytical tools for UAV operations in complex urban settings. By systematically constructing the reachable airspace and quantifying its visibility, this research establishes a solid theoretical foundation for UAV path planning, takeoff and landing site selection, and low-altitude route design.

2 Related Work

2.1 Airspace Delineation

Airspace delineation is a fundamental prerequisite for UAV flight path planning. Traditional methods for delineating urban low-altitude airspace primarily fall into three categories: airspace planning based on urban road networks [4]; airspace delineation incorporating urban infrastructure [5, 6]; and airspace partitioning using available urban

airspace grids [7]. The first two approaches heavily rely on ground transportation networks, resulting in airspace layouts that closely mirror ground-level configurations. This reliance overlooks the impact of high-rise buildings and complex terrains on three-dimensional space, particularly in densely built-up areas. Consequently, the available airspace is significantly reduced, failing to provide sufficient flight space for UAVs.

The third method typically employs fixed grid scales for partitioning the airspace, maintaining uniform grid sizes and shapes regardless of whether the area is densely built or open. This lack of flexibility impedes precise segmentation of complex urban spaces, especially in high-density building areas, leading to inaccurate delineation of available airspace.

2.2 Airspace Structural Analysis

In traditional air traffic management, airspace is treated as a fully continuous space, and analyses of airspace structures predominantly employ graph theory to examine aviation networks. However, this approach provides limited insight into the structures of low-altitude airspace. Upon reviewing the literature [8, 9], we found that space syntax theory, originally applied to the relationship between urban blocks and human activities, is inherently well-suited for analyzing low-altitude airspace structures. Utilizing this theory allows for flexible and quantitative analysis of reachable airspace. Specifically, the concept of isovists [10] enables the calculation of the visible range in all directions from any given point, converting it into quantitative metrics of visibility and connectivity. This gives space syntax theory a natural advantage in complex three-dimensional airspaces, as it effectively captures the impact of buildings and obstacles on lines of sight and sensor occlusions.

3 Method

In this study, we employ the Alpha-Shape algorithm to extract the outer contours of urban buildings, enabling the calculation of UAV reachable airspace under multiple constraints. By adjusting the key parameter α , the algorithm processes building point cloud data to generate a minimal enveloping structure that precisely captures the reachable boundaries of complex airspaces for both individual and multiple buildings. After determining the reachable airspace, we discretize it into uniform grids, converting the continuous space into discrete units to facilitate subsequent analysis. We then apply space syntax theory to quantitatively evaluate the visibility of each grid unit. This analysis delineates the spatial structure of urban low-altitude reachable airspace.

According to [11], many countries require urban UAVs to maintain a certain distance from man-made structures during operation. Consequently, we define the reachable airspace for UAVs as the complement of the zones obtained by outwardly expanding buildings by the mandated distance within the entire managed airspace. Therefore, efficiently and accurately identifying and expanding the airspace occupied by multiple buildings is crucial for constructing the reachable airspace. To address this, we employ the Alpha-Shape algorithm. The core concept of this algorithm is to control the structure of a point cloud using the parameter α , which determines the connection distances between points.

The alpha-shape algorithm, as shown in Fig. 1, identifies the shape of a scattered point set by creating boundaries based on a parameter α . Circles of radius α are drawn between pairs of points (yellow dots); if a circle can connect two points without enclosing any other points, they are linked with an edge. This process builds a boundary (dark lines) that captures both the outer and inner contours of the point set, allowing concavities to be included. Smaller values yield more detailed, concave boundaries, while larger values result in more convex shapes, eventually approximating the convex hull as α increases. The flexibility of the alpha-shape algorithm makes it ideal for accurately defining complex point clusters and identifying cavities within them.

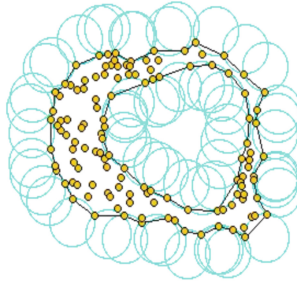


Fig. 1. Schematic Diagram of the Alpha-Shape Algorithm [12]

Building on the reachable airspace model, we conduct a quantitative spatial analysis to assess visibility and connectivity, aiming to identify key regions for UAV flight. Using isovist analysis from space syntax theory, we can quantify visibility, construct a visibility graph for the entire airspace, and evaluate connectivity. Isovist is the volume of space visible from a given point in space, together with a specification of the location of that point. The steps for constructing the visibility graph are as follows:

1. Sampling Point Generation: Discretize the reachable airspace $R \in \mathbb{R}^3$ by gridding it to generate a set of discrete sampling points $P = \{p_1, p_2 \dots p_n\}$, where each point $p_i \in R$ and the set P provides complete coverage of R .
2. Isovist Calculation: For each sampling point $p_i \in R$, compute its corresponding isovist area $A(p_i)$, which reflects the visibility at point p_i . The isovist area is defined as Eq. (1).

$$A(p_i) = \int_{\Omega} f(p_i, \theta) d\theta \quad (1)$$

where Ω represents all possible line-of-sight angles from p_i , and $f(p_i, \theta)$ denotes the maximum visible distance from p_i in direction θ .

3. Visibility Graph Construction: Based on the isovist areas $A(p_i)$, we can construct the visibility graph $G(V, E)$ of the reachable airspace, where:
 - $V = \{p_1, p_2 \dots p_n\}$ is the set of all sampling points.

- $E = \{e_{ij} | i \neq j, \text{ and there are no obstacles blocking between } p_i \text{ and } p_j\}$ is the set of edges representing connectivity between points, defined as Eq. (2).

$$e_{ij} = \begin{cases} 1, & \text{if } p_i \text{ and } p_j \text{ are visibly connected} \\ 0, & \text{if } p_i \text{ and } p_j \text{ are obstructed by an obstacle} \end{cases} \quad (2)$$

4. **Connectivity Analysis:** After constructing the visibility graph $G(V, E)$, we further analyze the spatial connectivity T using the following formula Eq. (3).

$$T = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i \neq j} d_{ij} \quad (3)$$

where d_{ij} is the Manhattan distance between points p_i and p_j . A higher T indicates better local connectivity, making these regions preferred choices for UAV flight paths and takeoff or landing sites.

4 Numerical Experiment

4.1 Experiment Set Up

In this study, we select a typical mixed-use urban area for analysis. The region measures 5 km in length and 4 km in width, encompassing a total area of 20 square kilometers. It includes several parks, a central business district, residential neighborhoods, stadiums, major transportation hubs, and narrow streets and alleys, representing a prototypical urban development structure. Figure 2 presents a satellite image of the selected area. Within this region, we construct a reachable airspace model and perform visibility graph computations and further analyses based on the obtained graph. By modeling this experimental area, we aim to validate the applicability and effectiveness of the proposed method in real-world scenarios.



Fig. 2. Experimental Area

4.2 Construction of the Reachable Airspace Map

We use the Alpha-Shape algorithm to generate the reachable airspace for UAVs. As depicted in Fig. 3, the boundaries of the reachable airspace are significantly influenced by the density and shape of buildings. Figure 3 (a) illustrates the reachable airspace at $h = 30$ m, $\alpha = 30$ m, clearly showing how obstacles between buildings disrupt the continuity of low-altitude airspace. Particularly in densely built regions, the Alpha-Shape algorithm constricts the airspace boundaries, resulting in a notable reduction in the UAVs' reachable range. This observation confirms that the distribution of buildings in urban environments directly affects the availability of airspace for UAV operations.

By adjusting the alpha parameter, as shown in Fig. 3 (b), the reachable airspace at the same altitude with $\alpha = 100$ m is visualized, revealing substantial blockage between buildings and a consequent decrease in available airspace. Comparative analysis of reachable airspace under different alpha values demonstrates that smaller alpha parameters can capture the subtle gaps between buildings, making the reachable airspace more precise and suitable for rotary UAVs with small turning radii that require meticulous obstacle avoidance. Conversely, larger alpha values generate more expansive boundaries, accommodating fixed-wing UAVs with larger turning radius which is more suitable for long-distance flights. The experimental results indicate that the Alpha-Shape algorithm effectively constructs low-altitude reachable airspace for UAVs and allows for the adjustment of modeling precision to meet diverse flight requirements.

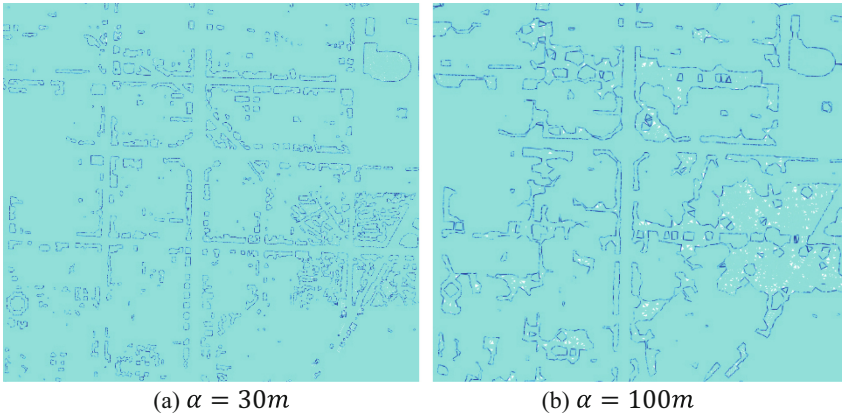


Fig. 3. Reachable Airspace Map.

4.3 Visibility Graph Analysis

Building upon the reachable airspace model, we further generate a visibility graph for the region and conducts isovist analysis on the sampling points within the experimental area to obtain a detailed distribution of visibility. Figure 4(a) presents a heatmap of

the visibility graph, where areas with high isovist values indicate wide-open views and strong spatial connectivity, whereas regions with low isovist values suggest significant line-of-sight obstructions, resulting in limited visibility.

The analysis reveals that sampling points near densely built areas exhibit noticeably lower isovist values (colder color), highlighting the substantial impact of building obstructions on visibility within these regions. Consequently, UAVs operating in these areas experience restricted visual ranges. In contrast, regions with fewer buildings or more open spaces show relatively higher isovist values (warmer color), indicating better spatial connectivity and making them more suitable as candidates for UAV flight paths and landing or takeoff points.

Figure 4(b) illustrates a heatmap of UAV visitation frequencies, generated by uniformly releasing UAVs within the reachable airspace and allowing them to fly randomly. The results demonstrate a high degree of consistency between frequently visited areas and the high isovist value regions depicted in Fig. 4(a). This correlation underscores the effectiveness of isovist analysis in identifying areas with strong spatial connectivity, validating its potential application in UAV path planning by accurately predicting UAV flight trajectories in real-world scenarios.

Furthermore, isovist analysis reveals the significant effects of building height and shape on airspace visibility. Taller and more complex buildings lead to lower isovist values, reducing visible range and spatial connectivity. This limitation constrains UAV flight space, underscoring the need for precise airspace management in densely built urban areas.

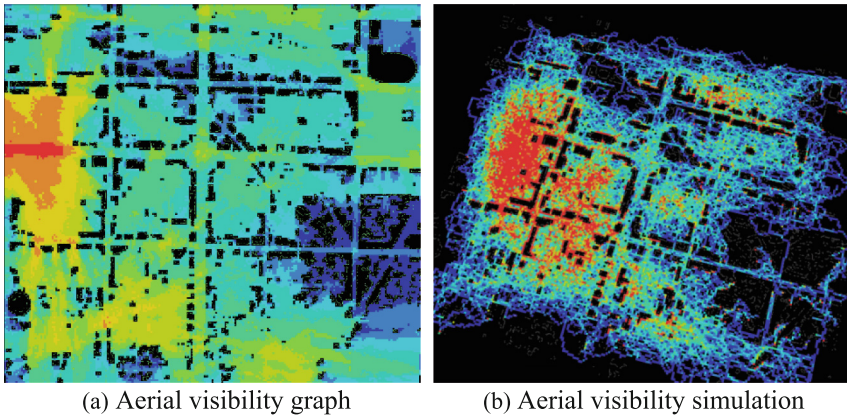


Fig. 4. Visibility Graph Analysis.

5 Result and Discussion

The experimental result shows the Alpha-Shape algorithm can effectively constructs reachable airspace in urban environments and can be adjusted to accommodate various flight requirements by modifying its parameters. Simultaneously, isovist analysis

provides a quantitative foundation for generating visibility graphs, offering an intuitive depiction of visibility and connectivity within low-altitude urban airspace. Through the analysis of these visibility graphs, we further validated that obstacles in densely built areas significantly impede UAV flight, whereas regions with open views exhibit stronger spatial connectivity, thereby enhancing UAV path planning and optimization.

Notably, although the Alpha-Shape algorithm accurately delineates airspace boundaries, temporary changes in buildings or obstacles—such as construction activities or moving barriers—in complex dynamic environments may compromise the effectiveness of the reachable airspace. Future research should incorporate dynamic obstacle detection and integrate real-time data to update the reachable airspace model, thereby improving its applicability and robustness.

In summary, the proposed method for constructing reachable airspace and generating visibility graphs provides an effective quantitative tool for managing urban low-altitude airspace. The experimental results confirm the feasibility and effectiveness of this approach. Future studies will expand its applications by integrating space syntax techniques, such as integration and connectivity analysis, particularly in areas like low-altitude route design, critical hub node identification, and dynamic airspace safety management.

6 Conclusion

This study introduces a low-altitude airspace modeling and visibility analysis method based on the Alpha-Shape algorithm and space syntax theory, applied to complex urban environments. By effectively constructing reachable airspace by Alpha-Shape algorithm and further generating visibility graphs through isovist analysis, this research provides quantitative tools and theoretical support for UAV operations within urban low-altitude airspace.

The primary contribution of this study lies in the precise definition of low-altitude airspace boundaries using the Alpha-Shape algorithm, coupled with the quantification of visibility distribution across different locations within the airspace through space syntax theory. This approach offers new perspectives for UAV path planning and low-altitude route design, particularly in complex urban airspace. Future research will integrate additional space syntax analysis tools to further enhance the precision and efficiency of low-altitude airspace management, thereby advancing the widespread application of UAV technology in urban environments.

Acknowledgments. This study is supported by the National Natural Science Foundation of China (Grants No. U2333214).

References

1. Huang, T., Fan, K., Sun, W.: Density gradient-RRT: An improved rapidly exploring random tree algorithm for UAV path planning. *Expert Syst. Appl.* **252**, 124121 (2024)
2. Huo, F., Zhu, S., Dong, H., Ren, W.: A new approach to smooth path planning of Ackerman mobile robot based on improved ACO algorithm and B-spline curve. *Robot. Auton. Syst.* **175**, 104655 (2024)

3. Cho, J., Yoon, Y.: Extracting the topology of urban airspace through graph abstraction. *Transportation Research Part C: Emerging Technologies*. **127**, 103116 (2021)
4. Mohamed Salleh, M.F.B., Low, K.H.: Concept of Operations (ConOps) for Traffic Management of Unmanned Aircraft Systems (TM-UAS) in Urban Environment. In: *AIAA Information Systems-AIAA Infotech @ Aerospace*. American Institute of Aeronautics and Astronautics, Grapevine, Texas (2017)
5. Doole, M., Ellerbroek, J., Hoekstra, J.: Estimation of traffic density from drone-based delivery in very low level urban airspace. *J. Air Transp. Manag.* **88**, 101862 (2020)
6. Bauranov, A., Rakas, J.: Designing airspace for urban air mobility: A review of concepts and approaches. *Prog. Aerosp. Sci.* **125**, 100726 (2021)
7. Mohamed Salleh, M.F.B., Wanchao, C., Wang, Z., Huang, S., Tan, D.Y., Huang, T., Low, K.H.: Preliminary Concept of Adaptive Urban Airspace Management for Unmanned Aircraft Operations. In: *2018 AIAA Information Systems-AIAA Infotech @ Aerospace*. American Institute of Aeronautics and Astronautics, Kissimmee, Florida (2018)
8. Hegazi, Y.S., Tahoon, D., Abdel-Fattah, N.A., El-Alfi, M.F.: Socio-spatial vulnerability assessment of heritage buildings through using space syntax. *Heliyon*. **8**, e09133 (2022)
9. Morrow, E., Mackenzie, I., Nema, G., Park, D.: Evaluating Three Dimensional Vision Fields in Pedestrian Micro-simulations. *Transportation Research Procedia*. **2**, 436–441 (2014)
10. Yu, R., Ostwald, M.J.: Spatio-visual experience of movement through the Yuyuan Garden: A computational analysis based on isovists and visibility graphs. *Frontiers of Architectural Research*. **7**, 497–509 (2018)
11. Cho, J., Yoon, Y.: How to assess the capacity of urban airspace: A topological approach using keep-in and keep-out geofence. *Transportation Research Part C: Emerging Technologies*. **92**, 137–149 (2018)
12. Sun, C., Wang, L., Tan, J., Zhao, B., Zhou, T., Kuang, Y.: A high-accuracy roundness measurement for cylindrical components by a morphological filter considering eccentricity, probe offset, tip head radius and tilt error. *Meas. Sci. Technol.* **27**, 085008 (2016)



Simulation Analysis of Mixed Traffic Flow Safety in Risk Scenarios on Highways Under Foggy Conditions

Ying Zhao¹, Haijun Li^{1,2(✉)}, Jianqiang Wang^{1,2}, and Yan Huang^{1,2}

¹ School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China
13240012@stu.lzjtu.edu.cn, {lihaijun, yanhuang}@mail.lzjtu.cn

² Key Laboratory of Railway Industry on Plateau Railway Transportation Intelligent Management and Control, Lanzhou Jiaotong University, Lanzhou 730070, China

Abstract. To clarify the safety of mixed traffic flow on highways under foggy conditions, this study conducted a systematic simulation analysis of mixed traffic flow in risk scenarios. By examining the information perception patterns of human-driven vehicles (HDVs), automated vehicles (AVs), and intelligent connected vehicles (ICVs), this study developed corresponding car-following models. Simulation scenarios with varying risk levels were created, and the simulation data were analyzed using indicators such as the cumulative reciprocal of time to collision (TTC), average speed, and speed standard deviation. The results indicate that, compared to HDVs, non-HDV vehicles (AVs and ICVs) significantly enhance the safety and stability of mixed traffic flow. In high-risk scenarios (deceleration of 6 m/s²), the cumulative reciprocal of TTC for ICVs is 12.8% lower than that for AVs. Furthermore, ICVs better maintain traffic flow stability than AVs, with the speed standard deviation in high-risk scenarios being reduced by an extra 13.1% for ICVs. Notably, ICVs have even shorter perception reaction times and display smaller speed adjustment oscillations. The findings of this study offer valuable insights for managing mixed traffic flow on highways during foggy conditions in the future.

Keywords: mixed traffic flow · car-following model · foggy condition · AVs · ICVs

1 Introduction

Automated Vehicles (AVs) and Intelligent Connected Vehicles (ICVs) are capable of effectively perceiving road environments and transmitting traffic information through their equipped sensors, cameras, and advanced wireless communication technologies, thereby improving road safety [1], and reducing the severity and frequency of traffic accidents [2]. In the existing road traffic network, the majority of vehicles are still traditional Human Driven Vehicles (HDVs). When promoting AVs and ICVs, the penetration rate of these vehicles is expected to gradually increase. Therefore, before the large-scale deployment of these two types of vehicles, it is important to conduct simulation

tests under various adverse driving scenarios at different penetration rates to ensure that the mixed traffic flow with a gradually increasing penetration rate can still effectively improve traffic safety.

Foggy conditions, as a typical adverse driving scenario, reduce a driver's ability to perceive the surrounding environment, resulting in poorer vehicle control and closer following distances [3], which greatly increases the likelihood of rear-end collisions on highways. To improve traffic safety in foggy conditions, researchers often use the information prompt functions of Connected Vehicles (CVs) to provide drivers with beyond-line-of-sight information [4, 5]. However, CVs are still fundamentally human-driven vehicles, requiring drivers to interpret the prompting information and make decisions accordingly. Hu et al. found that the distractions and increased cognitive load caused by these information prompt systems could negatively impact drivers' ability to drive safely in foggy conditions [6]. Therefore, CVs' prompting information cannot fundamentally eliminate the adverse effects of fog on drivers, the widespread deployment of AVs or ICVs, which do not require human driving, maybe a more effective approach to improving traffic safety in foggy conditions. However, no studies have yet systematically tested the safety of mixed traffic flows under foggy conditions.

Given this, the study aims to analyze the composition of mixed traffic flow, select or develop an appropriate car-following model, and systematically examine the safety of mixed traffic flow on foggy highways under different penetration rates and risk scenarios.

2 Method

2.1 Information Perception Mode for Different Vehicle Types

In this study, the vehicle types include HDVs, AVs, and ICVs. Among them, HDVs primarily perceive traffic information during the car-following process through the driver's visual and auditory. AVs do not require driver involvement during car-following process. These vehicles use high-precision radar sensors to obtain information about the leading vehicle and rely on the control algorithms to perform car-following. ICVs, based on AVs, use wireless communication technology to obtain information from multiple vehicles ahead, allowing for more reasonable adjustment of the vehicle's longitudinal motion state.

2.2 Car-Following Model for HDVs, AVs and ICVS

Huang et al. developed the Fog-related Intelligent Driver Model (FIDM), which effectively replicates the heterogeneity of the driver and the characteristics of the car-following distance of HDVs on highways in foggy conditions [7]. Therefore, this study used the FIDM to simulate the car-following behavior of HDVs in foggy conditions, and its mathematical formula is as follows:

$$a_n(t) = \begin{cases} a_{\max} \left[1 - \left(\frac{v_n(t)}{V_{\max}} \right)^\delta - \left(\frac{d_n^*(t)}{d_n(t)} \right)^2 \right] & \text{if } r_1 < (1 - p_1 f_m v_k) \\ ca_{\max} & \text{otherwise} \end{cases} \quad (1)$$

$$d_n^*(t) = S_0 + v_n(t)T_n(t) + \frac{v_n(t)\Delta v_n(t)}{2\sqrt{a_{\max}b}} \quad (2)$$

$$c = \begin{cases} 1 & \text{if } r_2 < 0.5 \\ -1 & \text{otherwise} \end{cases} \quad (3)$$

$$T_n(t) = \begin{cases} T_1 + (T_2 - T_1)r_4 & \text{if } r_3 > (1 - p_2 f_m V_k) \\ T_n(t - \Delta t) & \text{otherwise} \end{cases} \quad (4)$$

The definitions of the relevant parameters can be found in the previous study [7].

Zhou et al. used a linear controller to simulate the longitudinal car-following behavior of AVs [8]. The main features of this controller are its simplicity and stability. Based on the model, this study adds acceleration change constraints to improve passenger comfort. The specific control model is as follows:

$$s_n^{des}(t) = \delta_1 + \tau_1 \cdot v_{n-1}(t) \quad (5)$$

$$\Delta s_n(t) = s_n(t) - s_n^{des}(t) \quad (6)$$

$$a_n^*(t) = q \cdot \Delta s_n(t) + v_{n-1}(t) - v_n(t) \quad (7)$$

$$a_n(t) = \begin{cases} a_n^*(t) & \text{if } a_{dec} \leq a_n^*(t) \leq a_{acc} \\ a_{acc} & \text{if } a_{acc} < a_n^*(t) \\ a_{dec} & \text{otherwise} \end{cases} \quad (8)$$

$$v_n(t) = a_n(t)\Delta t + v_n(t-1) \quad (9)$$

Among them, a_{acc} represents the vehicle's maximum acceleration; a_{dec} represents the vehicle's maximum deceleration. The definitions of the other parameters can be found in the previous study [8].

ICVs integrate the speed and position information of multiple vehicles ahead into the control of the longitudinal car-following behavior of the ICVs. Based on the linear controller for AVs, this study establishes the control model of ICVs:

$$s_n^{des}(t) = \delta_1 + \tau_1 \cdot v_{n-1}(t) \quad (10)$$

$$\Delta s_n(t) = \sum_{i=1}^m \gamma_m (s_{n-m+1}(t) - s_{n-m+1}^{des}(t)) \quad (11)$$

$$\sum_{i=1}^m \gamma_m = 1 \quad (12)$$

$$a_n^*(t) = q \cdot \Delta s_n(t) + \sum_{i=1}^m \gamma_m v_{n-m}(t) - v_n(t) \quad (13)$$

$$a_n(t) = \begin{cases} a_n^*(t) & \text{if } a_{dec} \leq a_n^*(t) \leq a_{acc} \\ a_{acc} & \text{if } a_{acc} < a_n^*(t) \\ a_{dec} & \text{otherwise} \end{cases} \quad (14)$$

$$v_n(t) = a_n(t)\Delta t + v_n(t-1) \quad (15)$$

Among them, m represents the number of preceding vehicles that are capable of wireless communication; γ represents the weight factor assigned to the different preceding vehicles, with the sum of all weight factors equal to 1.

3 Simulation Experiment Design and Evaluation Indexes

3.1 Experiment Design and Parameters

The numerical simulation scenario in this study is a bidirectional, four-lane highway with heavy fog. According to the regulations of road traffic safety in China, the speed limit on the highway in heavy fog conditions was set at 40 km/h. The duration of each numerical simulation was set to 2,000 s, with a time step interval of 0.1 s. The number of vehicles in the traffic flow was set to 10 (excluding the leading vehicle).

In this study, the traffic flow was divided into three types: the traffic flow composed of HDVs as the control group, the mixed traffic flow composed of HDVs and AVs, and the mixed traffic flow composed of HDVs and ICVs. The penetration rates of AVs and ICVs were set at 20%, 40%, 60%, 80%, and 100%. For each simulation condition, 20 simulations were conducted. Besides, this study sets different deceleration rates (2 m/s^2 , 4 m/s^2 , and 6 m/s^2) for the leading vehicle to create various levels of potential rear-end collision risks. The leading vehicle performs a deceleration behavior with a duration of 1 s at 1,010 s of simulation time, then maintains the reduced speed for 39 s, after which it accelerates for 1 s to restore the speed limit. Furthermore, the data collection period for the evaluation indicators is from 1,010 s to 1,050 s.

In this study, the model parameters used were sourced from previous research [7, 8]. $a_{\max}$ was set to 0.73 m/s^2 ; δ was set to 4; b was set to 1.67 m/s^2 ; S_0 was set to 2 m; p_1 was set to 0.001 s^{-1} ; p_2 was set to 0.5 s^{-1} ; T_1 was set to 0.5 s; T_2 was set to 2.0 s; $V_{\max}$ was set to 50 km/h; f_m and V_k were set to 1; δ_1 was set to 7.28 m; τ_1 was set to 0.79 s; k was set to 0.12; a_{acc} was set to 0.73 m/s^2 ; a_{dec} was set to 1.67 m/s^2 ; m was set to 3; The proportions of γ were set to 0.5, 0.3, and 0.2, respectively.

3.2 Evaluation Indexes

TTC, as the most commonly used evaluation indicator in traffic safety analysis. Some researchers use the reciprocal of TTC instead of TTC, which effectively avoids situations where TTC changes too drastically or becomes undefined [9]. The specific expression is as follows:

$$C_{1/ttc} = \sum_{n=1}^N \sum_{j=1}^{T_3} \left[\frac{\Delta v_n(j)}{s_n(j)} \right] \quad (16)$$

$$\Delta v_n(j) = \begin{cases} v_n(j) - v_{n+1}(j), & v_n(j) \geq v_{n+1}(j) \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

The definitions of the relevant parameters can be found in the previous study [9].

Previous studies have used average speed to assess the potential accident risk when vehicles driving in foggy conditions [10]. It is generally believed that, under the same foggy conditions, the lower the vehicle speed, the lower the accident risk. In addition, the speed standard deviation was used to measure the stability and driving comfort of drivers during the speed adjustment process.

4 Results

4.1 Different Penetration Rates

This study employed multivariate analysis of variance (MANOVA) to test the correlation between influencing factors and variables. All hypothesis tests were conducted at a significance level of 0.05.

The results of MANOVA for the penetration rate and the comparison results of the cumulative reciprocal of TTC, average speed, and speed standard deviation under traffic flow types are shown in Fig. 1. The results indicate that the penetration rate of non-HDVs (AVs or ICVs) has a significant effect on the cumulative reciprocal of TTC, average speed, and speed standard deviation. As the penetration rate of non-HDVs increases, the cumulative reciprocal of TTC shows a clear downward trend. When the penetration rate of non-HDVs reaches 100%, their cumulative reciprocal of TTC is only 27.7% of that for HDVs. Under the same penetration rate, the cumulative reciprocal of TTC for ICVs is significantly lower than that of AVs, indicating that ICVs further reduce the risk of accidents during the car-following process. When the penetration rate of non-HDVs is 100%, their average speed and speed standard deviation are reduced by 8.9% and 33.5%, respectively, compared to HDVs. This suggests that non-HDVs can quickly adjust their speeds to meet speed limits in various risk scenarios, reducing the risk of accidents. Furthermore, during the speed adjustment process, the speed standard deviation of ICVs is significantly smaller than that of AVs, indicating that the speed adjustment process of ICVs is more stable and smoother compared to that of AVs.

Figure 2 shows the speed profiles of HDVs, AVs, and ICVs in a high-risk scenario. ICVs have a notably faster perception and reaction time than AVs and HDVs. Compared to HDVs, both AVs and ICVs quickly and stably adjust their speed to the corresponding speed limit. Furthermore, the speed fluctuation range of ICVs is smaller.

4.2 Different Risk Levels

Figure 3 presents the results of the MANOVA for the deceleration of the leading vehicle and a comparison of the cumulative reciprocal of TTC, average speed, and speed standard deviation under different traffic flow types. As the deceleration increases, the cumulative reciprocal of TTC increases significantly, indicating that the potential accident risk in foggy conditions increases. In the same risk level, the cumulative reciprocal of TTC for ICVs is significantly lower than that for AVs, suggesting that ICVs can further reduce the potential accident risk. In addition, the speed standard deviation also increases as the risk level rises, indicating that in high-risk scenarios, the range of vehicle speed fluctuations

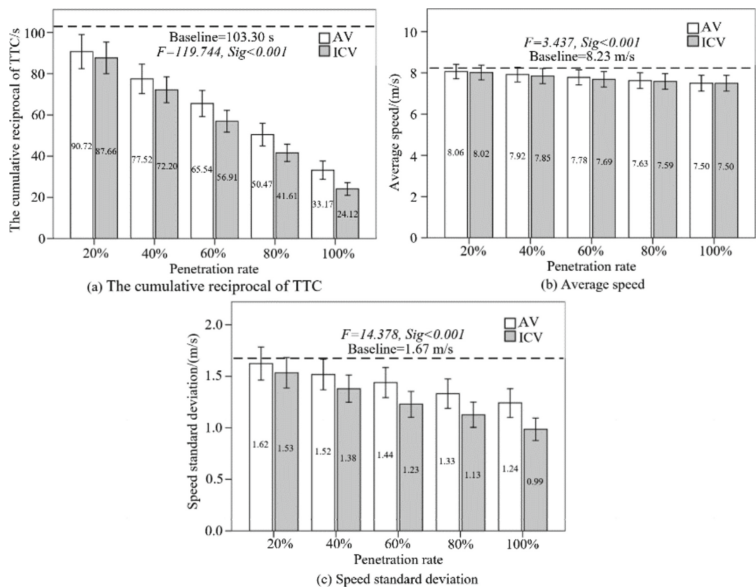


Fig. 1. Relationship between traffic flow types and variables under different penetration rates

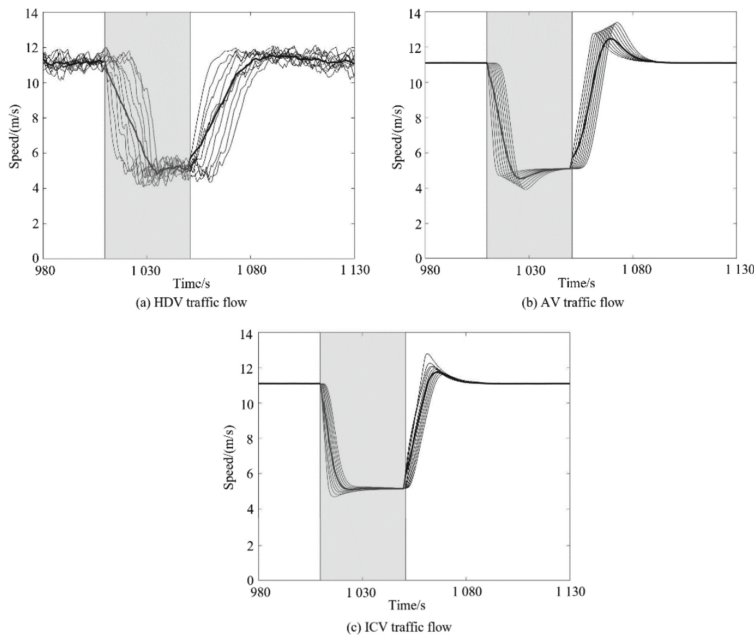


Fig. 2. The speed profiles of different traffic flows

increases, which leads to a decrease in traffic flow stability. The speed standard deviation of ICVs is significantly lower than that of AVs, with a reduction of approximately 12.5%, indicating that, ICVs are better able to control vehicle speed fluctuations and maintain better stability in the traffic flow compared to AVs.

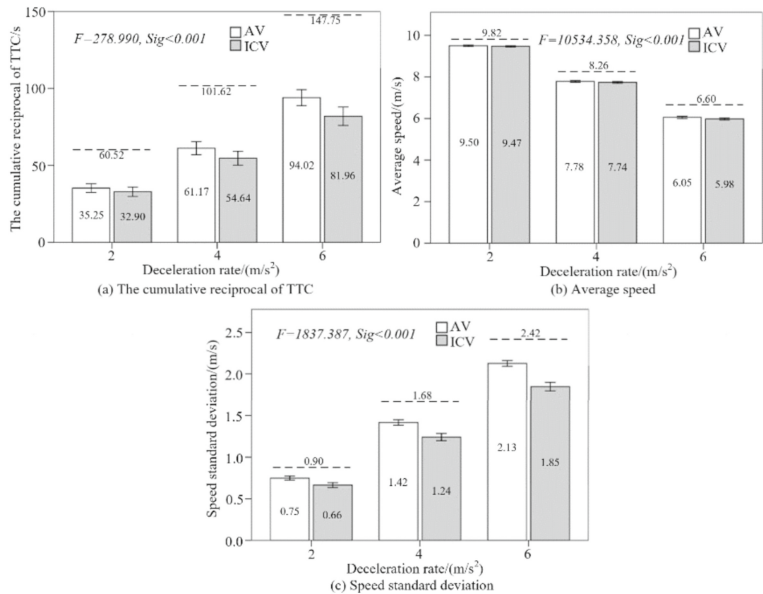


Fig. 3. Relationship between traffic flow types and variables under different risk levels

5 Conclusion

This study analyzes the information perception modes of ICVs, AVs, and HDVs in foggy conditions, selecting or constructing corresponding car-following models. Numerical simulation experiments were conducted under various penetration rates and risk levels. The results show that, in different risk levels and penetration rate scenarios, both AVs and ICVs can improve traffic safety better than traditional HDVs in foggy conditions. As the penetration rate increases, the ability of AVs and ICVs to reduce potential accident risks and stabilize traffic flow is further enhanced, with these improvements being more significant in high-risk scenarios. Moreover, under the same risk level or penetration rate conditions, ICVs show a significantly stronger effect on improving the safety and stability of traffic flow compared to AVs.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (52462045, 72401119, 52462047) and the Joint Innovation Fund Project of Lanzhou Jiaotong University and Beijing Jiaotong University (LH2024016).

References

1. Sultana, T., Hassan, H.M.: Does recognizability of connected and automated vehicles (CAVs) platoons affect drivers' behavior and safety? *Transport. Res. F: Traffic Psychol. Behav.* **103**, 368–386 (2024)
2. Boggs, A.M., Arvin, R., Khattak, A.J.: Exploring the who, what, when, where, and why of automated vehicle disengagements. *Accid. Anal. Prev.* **136**, 105406 (2020)
3. Huang, Y., Yan, X., Li, X., et al.: Using a multi-user driving simulator system to explore the patterns of vehicle fleet rear-end collisions occurrence under different foggy conditions and speed limits. *Transport. Res. F: Traffic Psychol. Behav.* **74**, 161–172 (2020)
4. Zhang, Y., Yan, X., Li, X.: Effect of warning message on driver's stop/go decision and red-light-running behaviors under fog condition. *Accid. Anal. Prev.* **150**, 105906 (2021)
5. Wu, Y., Abdel-aty, M., Park, J., et al.: Effects of crash warning systems on rear-end crash avoidance behavior under fog conditions. *Transp. Res. Part C: Emerg. Technol.* **95**, 481–492 (2018)
6. Hu, D., Feng, X., Zhao, X., et al.: Impact of HMI on driver's distraction on a freeway under heavy foggy condition based on visual characteristics. *J. Transp. Saf. Secur.* **14**(6), 905–928 (2022)
7. Huang, Y., Yan, X., Li, X., et al.: Improving car-following model to capture unobserved driver heterogeneity and following distance features in fog condition. *Transp. A: Transp. Sci.* **20**(1), 2048917 (2024)
8. Zhou, H., Zhou, A., Li, T., et al.: Significance of low-level control to string stability under adaptive cruise control: algorithms, theory and experiments. *Transp. Res. Part C: Emerg. Technol.* **140**, 103697 (2022)
9. Li, C., Zhao, X., Xie, D., et al.: Exploring the impact mechanism of communicating information on the perturbation propagation. *Transp. A: Transp. Sci.* **20**(1), 2035013 (2024)
10. Li, X., Yan, X., Wong, S.C.: Effects of fog, driver experience and gender on driving behavior on s-curved road segments. *Accid. Anal. Prev.* **77**, 91–104 (2015)



Research on the Innovation and Environmental Sustainability Development of Coal Storage Facilities in Chinese Ports

Yuhua Wei^(✉)

CHN Energy Technology and Economics Research Institute, Beijing 102209, China
20027019@ceic.com

Abstract. With the increasing global emphasis on environmental protection and sustainable development, port coal storage facilities, as a key link in the coal supply chain, have a direct impact on the operational efficiency of the coal supply chain. The stability of their operation is of great significance to energy security. The innovation and sustainable development of coal storage facilities in ports are particularly important. This article provides an in-depth analysis of the advantages and disadvantages of different coal storage facilities, innovative technologies and management strategies for port coal storage facilities, as well as the current status of environmental sustainability, and explores the challenges faced at present. Through the case study and analysis of the innovation of coal storage facilities in Huanghua Port, this article proposes innovative management strategies and suggestions aimed at promoting the environmental sustainability of coal storage facilities in the port.

Keywords: Coal Storage · Innovation · Environmental Sustainability

1 Introduction

Against the backdrop of global energy structure transformation and increased awareness of environmental protection [1, 2], China, as the world's largest producer and consumer of coal, plays a crucial role in optimizing its energy structure and improving environmental quality to achieve sustainable development goals. Port coal storage facilities, as key nodes in the coal supply chain [3], are not only related to the stability of energy supply, but also an important factor affecting regional environmental quality. With the advancement of technology and the strengthening of environmental regulations, traditional coal storage methods are facing many challenges, and there is an urgent need to improve environmental friendliness and operational efficiency through innovative management and technological innovation.

This study focuses on the current situation and challenges of innovative technology application, management strategy optimization, and environmental sustainability development of coal storage facilities in Chinese ports. Through comprehensive analysis of the development history, technological innovation, and management practices

of coal storage facilities in ports, a combination of quantitative and qualitative methods is adopted. Through in-depth analysis of the experience of Huanghua Port silo case in technology innovation and management optimization [4], as well as its contribution to environmental sustainability development, targeted innovative management strategies and suggestions for the sustainable development of coal storage facilities in ports are proposed to help achieve efficient, environmentally friendly, and sustainable development of the coal supply chain. The aim is to provide reference for relevant policy formulation and practical operation, assist China's port coal storage facilities in achieving green transformation, and jointly promote the optimization of global energy structure and environmental protection awareness. Upgrade.

2 Types and Characteristics of Coal Storage Facilities

2.1 Types of Coal Storage Facilities

According to the structural characteristics of coal storage facilities, they can be divided into cylindrical coal bunker type, yard type, and dome shaped coal bunker type.

A silo type coal storage facility is a closed upright container that can be in the form of a single coal bunker, a row of coal bunkers, or a group of bunkers. Although spherical coal bunkers, trough shaped coal bunkers, and square coal bunkers differ structurally from circular silos, they all belong to fully enclosed containers and can be classified as cylindrical coal bunkers.

A yard type coal storage facility usually refers to an open-air, semi enclosed, or fully enclosed coal storage yard, which is a shed type structure used to store coal with its surroundings and roof not enclosed, semi enclosed, or fully enclosed [5]. The yard type coal storage facilities can be divided into strip coal storage yards and circular coal storage yards according to the shape of the coal yard [6, 7]. According to the different roof structures, yard type coal storage facilities can be divided into steel grid structure coal storage yard, portal steel frame structure coal storage yard, pipe truss structure coal storage yard, spherical grid structure coal storage yard, and inflatable membrane coal storage yard [8]. The yard type coal storage facilities are divided into stacker reclaimer coal storage yard, coal drop tower coal storage yard, and wharf coal storage yard according to the process form. The dome shaped warehouse has a dome shaped or polygonal curved top cover, and the structure covering the entire dome is a PVC air cushion [9].

The Classification of Coal Storage Facilities is shown in Table 1. Schematic diagram of Coal Storage Facilities is show in Fig. 1.

2.2 Analysis of the Advantages and Disadvantages of Different Coal Storage Facilities

The cylindrical coal bunker has the advantages of small footprint, high site utilization rate, good environmental protection effect, high operational reliability, and easy automation operation. However, it has disadvantages such as small storage capacity, high construction cost, and high risk of spontaneous combustion and blockage [10, 11].

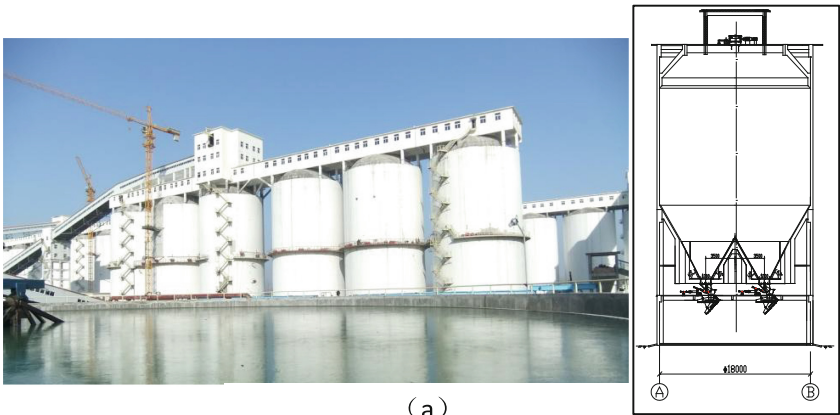
Table 1. Classification of Coal Storage Facilities

level 1	level 2	level 3
Silo type		Cylindrical coal bunker
		Ball coal bunker
		Slot coal bunker
		Square coal bunker
Yard type	Degree of closure	Open pit coal storage yard
		Semi enclosed coal storage yard
		Fully enclosed coal storage yard
	Yard shape	Strip coal storage yard
		Circular coal storage yard
	roof structure	Steel grid structure coal storage yard
		Portal steel frame coal storage yard
		Steel truss structure coal storage yard
		Spherical grid structure coal storage yard
		Inflatable membrane structure coal storage yard
	process form	Stacker reclaimer coal storage yard
		Coal drop tower coal storage yard
		Zhanqiao Coal Storage Yard
Dome shaped warehouse		

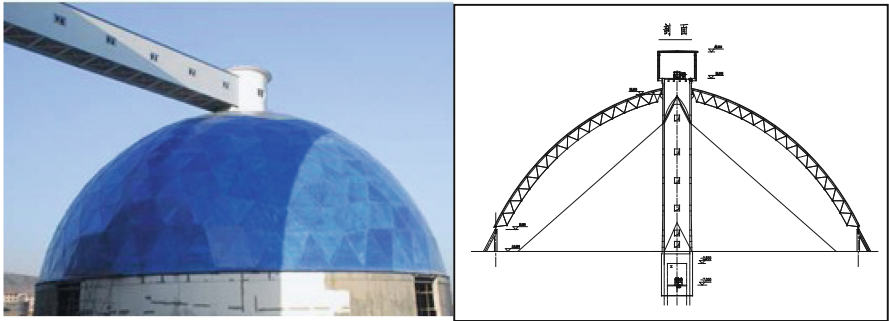
Open air coal storage yards and semi enclosed coal storage yards have poor environmental protection effects and do not meet environmental requirements.

Closed coal storage yards are developed on the basis of open-pit and semi closed coal storage yards. Compared with open-pit coal storage yards, closed coal storage yards have the advantages of improving the surrounding environment and shorter construction period, but they require higher investment. Compared with cylindrical coal bunkers, enclosed coal storage yards have the advantages of lower foundation treatment costs and shorter construction periods, but they occupy a larger area and have lower loading efficiency. The strip coal storage yard has the advantages of strong adaptability and is conducive to long-term expansion, but it has disadvantages such as occupying a large area and low efficiency of bulldozer operations. Compared with strip coal storage yards, circular coal storage yards have the advantages of small footprint and high degree of automation, but require higher investment. The advantages and disadvantages of coal storage yards with different maintenance structures are the same as those of circular coal storage yards.

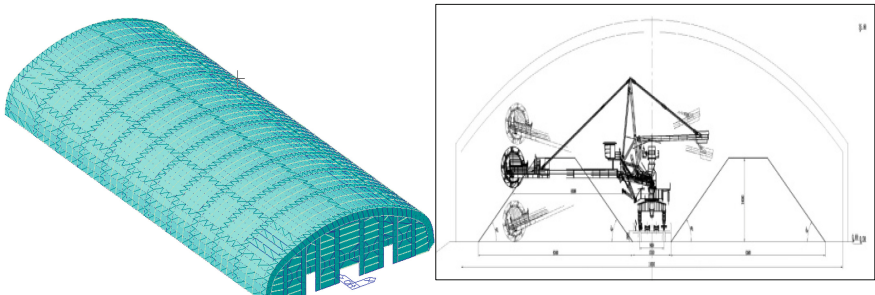
The dome shaped warehouse has the advantages of short construction period, low investment, and large coal storage capacity per unit area, but it has disadvantages such as high construction technology requirements and high maintenance costs [12].



(a)
Cylindrical coal bunker



(b)
Cylindrical coal drop tower coal storage yard



(c)
Strip stacker reclaimer type coal storage yard

Fig. 1. Schematic diagram of Coal Storage Facilities.

The advantages and disadvantages of different coal storage facilities are shown in Table 2.

Table 2. The advantages and disadvantages of different coal storage facilities

level 3	Advantages	Disadvantages
Cylindrical coal bunker	1.Easy to expand; 2.Good sealing performance, excellent environmental protection, high level of automation, small footprint, and low maintenance costs; 3.Capable of storing multiple coal types and has the function of blending coal; 4.The structure has strong seismic resistance, wind resistance, and corrosion resistance	1.The lifting height for coal is large; 2.High construction technical requirements and long construction period are needed; 3.There is a high level of investment required
Ball coal bunker	1.Small footprint; 2.Low operating costs	1.High construction technical requirements and long construction periods; 2.High investment and maintenance costs
Slot coal bunker	1.Large coal storage capacity; 2.Low operating costs;	1.High requirements for safety ventilation; 2.Large engineering volume, high construction difficulty, and substantial investment; 3.Inconvenient handling of coal spontaneous combustion
Square coal bunker	1.Small land occupation; 2.Short construction period	1.Small coal storage capacity; 2.Poor stability
Open pit coal storage yard	1.Low investment required; 2.Short construction period	Poor environmental protection effects
Semi enclosed coal storage yard	1.Low investment required; 2.Short construction period	Poor environmental protection effects
Fully enclosed coal storage yard	Good environmental protection effects	High investment is required
Strip coal storage yard	1.Convenient handling of coal spontaneous combustion; 2.Short construction period	1.Occupies a large area; 2.Poor auxiliary working environment

(continued)

Table 2. (continued)

level 3	Advantages	Disadvantages
Circular coal storage yard	1.The area occupied is moderate; 2.Good load-bearing performance	1.The coal pile has poor formability, a low fill factor, and loose stacking; 2.High construction technical requirements and long construction periods are needed; 3.The cost level is high
Steel grid structure coal storage yard	High safety and reliability; widely applied	Construction period is long
Portal steel frame coal storage yard	1.Low investment; 2.Short construction period	Not suitable for spans greater than 120 m
Steel truss structure coal storage yard	Easy to install, with a short construction period	Not suitable for spans greater than 160 m
Spherical grid structure coal storage yard	The structural overall stability is good	1.The structural span is limited; 2.High construction technical requirements, long construction periods, and high maintenance costs; 3.Poor corrosion resistance
Inflatable membrane structure coal storage yard	1.Short construction period; 2.Strong adaptability	1.Short lifespan; 2.High operating costs
Stacker reclaimer coal storage yard	1.High level of automation; 2.Flexible and reliable operation	1.Poor working environment; 2.High investment and significant energy consumption
Coal drop tower coal storage yard	Low investment is required	1.The working environment is poor; 2.The level of automation is low
Zhanqiao Coal Storage Yard	1.Simple structure with large storage capacity; 2.Low investment is required	1.Poor working conditions; 2.Low level of automation
Dome shaped warehouse	1.Strong earthquake and wind resistance capabilities; 2.Short construction period	High investment and maintenance costs

3 Research on the Innovation Path of Coal Storage Facilities in Huanghua Port

Huanghua Port, as an important port for the transportation of coal from the west to the east and from the north to the south, this article takes Huanghua Port's innovative measures as an example to analyze and study the innovative achievements of coal storage facilities in terms of technology and management.

3.1 Application of Technological Innovation in Huanghua Port Coal Storage Facilities

At present, the technological innovation of Huanghua Port's coal storage facilities is mainly reflected in the upgrading and transformation of intelligent systems, the application of silo groups, the use of energy-saving facilities, and the application of green development technologies.

Intelligent upgrade and transformation: Bringing together innovative technologies such as artificial intelligence, cloud computing, big data, 5G, and the Internet of Things, it achieves the integration of "flipping, stacking, retrieving, and loading" and internal processes, improving the quality and efficiency of ship loading; The intelligent upgrade and transformation of the yard have been completed. By collecting and modeling the operating data of the stacker and reclaimer, the coal quantity control and edge detection of stacking and reclaimer have been carried out, effectively providing the efficiency and stability of stacking and reclaimer, realizing unmanned operation of stacking equipment, and greatly reducing the labor intensity and labor cost of frontline workers.

Application of silo group: The Huanghua Port Phase III project has built 48 cylindrical warehouses with a single warehouse capacity of 30000 tons, forming a storage method of yard + silo. The storage types are reasonably allocated according to the length of the storage cycle, and the advantages of organization and coordination are obvious.

Energy saving facilities and green development technologies: changing the driving mode of the belt conveyor to variable frequency drive not only saves energy, but also improves the durability of the equipment; By implementing measures such as coal dust cake making and windproof greening systems, the dust concentration has been effectively reduced, achieving ultra-low emissions of coal dust; By adopting various technologies such as infiltration, retention, storage, purification, utilization, and discharge, it helps to prevent and control water pollution, protect water ecology, and utilize water resources, promoting the green and sustainable development of ports.

3.2 Application of Management Innovation in Huanghua Port Coal Storage Facilities

The innovation in management of Huanghua Port's coal storage facilities is mainly reflected in intelligent, ecological, green, and refined management measures.

Intelligent management: Huanghua Port adopts "intelligent" full process control, builds an intelligent operation scheduling system, and creates a digital operation management platform that integrates production scheduling, operation monitoring, safety warning, emergency command, and video conferencing.

Ecological and Green Management: Through innovative measures such as an ecological environment intelligent control platform, a multi-dimensional cloud monitoring system for port dust has been established to monitor the real-time distribution of dust concentration and intelligently determine the location of dust generation, achieving the scheduling of dust suppression equipment, realizing the full chain intelligent management and control of coal dust, and achieving near zero emissions. It has become the first dry bulk port in China to win the "China Five Star Green Port" award.

Refined management: Applying the concept of "sponge city" to port construction, promoting modern management methods such as refined management, standardized operations, and information technology construction, establishing a sound management system and process system, improving work efficiency and operation quality, and further enhancing the company's brand image and market competitiveness [13].

4 Environmental Sustainable Development Strategy

This article combines the analysis and research of the innovative path achievements of Huanghua Port's coal storage facilities, and proposes environmental sustainability development strategies for the port's coal storage facilities, including green ecological top-level design, environmental pollution prevention and control, water resource management, digital transformation development, and coordinated development of clean and low-carbon.

4.1 Green Ecological Development Top-Level Design

Based on the mechanism of dust generation and suppression in ports, the traceability of dust pollutants in port areas and surrounding areas, and the design of environmental management and decision-making platforms, we aim to enhance the level of informatization and intelligence in monitoring, controlling, and managing air pollutants in ports.

4.2 Environmental Pollution Prevention and Control and Water Resource Management

Promote the application of coal storage facilities such as cylindrical coal bunkers and enclosed storage yards to realize the non open storage of coal, reduce the exposure time of coal during storage, reduce the opportunities for coal dust diffusion, and reduce the pollution of coal dust on the environment.

Adopting precise watering, intelligent belt cleaning and other technologies to prevent coal dust from flying during operation, effectively controlling dust pollution.

Promote the concept of "sponge port", collect and treat spray water and other materials for reuse, and achieve the reuse of water resources.

Establish an ecological intelligent control system, predict and formulate the optimal water spraying and dust prevention plan in advance through big data analysis, and automatically control the dust spraying equipment based on the plan to achieve precise dust prevention.

4.3 Digital Transformation Development

In the operation and management of coal storage facilities in ports, advanced information technologies such as the Internet of Things, network communication technology, big data analysis and prediction, and energy management systems [14] are adopted to achieve real-time monitoring and optimized management of energy consumption. Digital technology is used to improve coal storage efficiency, reduce energy consumption and emissions.

4.4 Clean and Low-Carbon Coordinated Development

By promoting green and low-carbon construction in ports through technological innovation and demonstration applications, photovoltaic power generation systems are installed on the upper part of enclosed coal sheds to achieve self use, reduce plant electricity consumption, and decrease dependence on external electricity. At the same time, surplus electricity generated after local consumption is connected to the grid, generating certain economic benefits.

5 Conclusion

The innovation and environmental sustainability research of coal storage facilities in ports is a multidimensional and interdisciplinary field. By adopting clean energy and pollution reduction technologies, the impact of port operations on the environment can be reduced; Through technological innovation and management optimization, the efficiency of coal storage in ports can be effectively improved, achieving green, intelligent, and efficient development of ports. Future research should further explore more innovative technologies and management strategies, including but not limited to new energy storage technologies, intelligent monitoring systems, environmental impact assessment tools, and methods for effectively integrating these technologies and management strategies to adapt to constantly changing environmental policies and market demands, and achieve environmental sustainability of port coal storage facilities.

References

1. D. G, R.C, S. R.A review on environmental impacts of renewable energy for sustainable development. *Int. J. Environ. Sci. Technol.* **21**(5), 5285–5310 (2023)
2. Li, J., Sun, H., Cheng, L., et al.: The path driving china's energy structure transformation from the perspective of policy tools. *Sustainability* **16**(19), 8703–8703 (2024)
3. Yinglin, S., Sun, L., Shuyun, Z.: Exploration and application of automated control system for port coal storage and distribution bases. *Port Eng. Technol.* **59**(02), 6–9 (2022)
4. Huanbo, N.: Research and application on management strategies for large coal storage silo groups. *Transp. Manage. World* **06**, 124–128 (2020)
5. China Nonferrous Metal Industry Association. Technical Specification for Steel Silos: GB 50884-2013. China Planning Press, Beijing (2013)
6. Peng, W.: Analysis of coal storage methods in thermal power plants. *Global Market* **33**, 1 (2016)

7. Jianguo, L., Zihao, Z., Longzhe, J., et al.: Effect of coal dust content on the low-temperature oxidation of silo coal. *ACS Omega* **7**(42), 37442–37451 (2022)
8. Zhenggang, C.: Design and Construction of Large-Span Prestressed Steel Structure Dry Coal Sheds. China Architecture & Building Press, Beijing (2019)
9. Aimin, Z.: Prospect of application of the air-film domed concrete silo for coal storage in power plant. *Coal Preparation Technol.* **47**(3), 86–90 (2019)
10. Sun, J.: Research and practical application of safety technology for coal storage silos. *Digit. Commun. World* **10**, 169+269 (2017)
11. Rotter, J.: Guide for the Economic Design of Circular Metal Silos. Taylor and Francis, 01 November 2002
12. Chang, X., Li, X., Guo, Y.: Preliminary exploration of the scheme for using air-membrane reinforced concrete dome storage for raw coal bins at the Kekegai Coal preparation plant. *Huadian Technol.* **39**(11), 33–35+38+78 (2017)
13. Zhang, J.: Post-evaluation Study on the Coal Storage Silo Project of Huanghua Port. Yanshan University (2020)
14. Jiongyao, Z.: Design of intelligent energy consumption management system. *Comput. Program. Skills Maintenance* **07**, 100–102 (2019)



A Smart Tourism Recommendation Algorithm Based on Profiling Technology and Bipartite Graph

Zexian Guo^(✉), Yuemin Lin, and Zhouyang Luo

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
{22251037, 22251012, 22251046}@bjtu.edu.cn

Abstract. With the development of travel industry, there is an increasing demand for high-quality tourism services. However, there are problems such as the lack of individuation in tourism recommendation services and generalization in tour route planning. In order to solve these problems, this paper first collects data and proposes the concept of tourism personality and its classification based on tourist portraits. Then, it tests the “tourism personality” for users. Next, it constructs a tourism resource database based on tourism resource portrait and big data analysis. Finally, it plans tour routes and time arrangements rationally based on the bipartite graph by using the depth-first search (DFS) and breadth-first search (BFS) algorithms. Combining the four aspects of meals, living, traveling, shopping and accommodation, it generates an integrated and personalized intelligent tourism recommendation algorithm and implements the algorithm through an APP called “Zhi Lv Hui You”. This algorithm providing high-quality and comprehensive customized tourism services for tourists and promoting the process of intelligent tourism and smart city construction.

Keywords: Wisdom Tourism · Scheme Generation · Tourist Portraits · Tourist Resources · Bipartite Graph · Tourism Personality

1 Introduction

As the people’s demand for a better life continues to rise, China’s tourism industry is developing rapidly. According to the statistics of the domestic tourism sample survey, in the first quarter of 2024, the number of domestic tourists was 1.419 billion, 203 million more than the same period last year, an increase of 16.7%. However, the current personalized tourism services cannot meet the needs of tourists. In September 2023, the notice issued by the State Council Office, “Several Measures to Release the Potential of Tourism Consumption and Promote High-quality Development of Tourism”, proposing to “increase the supply of high-quality tourism products and services” [1]. At present, the research results in the field of tourism program recommendation can be mainly divided into two kinds of tourism route recommendation methods based on tourism resource portrait and tourist portraits.

The tourism recommendation method based on the tourism resource portrait mainly includes analyzing tourists' visit information to scenic spots, providing tourism recommendation schemes based on the geographic information of target tourist attraction [2], and defining tourist spots' labels according to their own characteristics. Such as the use of user reviews, scenic spot ratings [3], regional popularity and other dimensions of the scenic spot label features [4] to provide personalized recommendation services for tourists.

In the travel recommendation methods based on user portraits, the portraits constructed mainly include user portraits based on ontology [5], subject model [6], behavior and interest [7], while the personalized proposal recommendation methods include content-based recommendation [8], association rule-based recommendation [9], collaborative filtering recommendation [10] and mixed recommendation [11].

At present, there are few studies on tourism route recommendation based on the combination of tourism resource portrait and user portrait, which lacks the synthesis of multiple data. This study takes Beijing as an example, combines tourist portraits and tourism resource portraits, and proposes a travel recommendation algorithm and APP based on bipartite graph to ensure the rationality of route planning, aiming to provide tourists with intelligent, personalized and integrated travel plans and provide new plans and ideas for optimizing the supply of tourism services.

2 Data Collection

We takes Beijing as an example to collect tourist data and tourism resources data. First of all, we collect tourist data through questionnaires. The contents of the questionnaire include basic information of tourists (the number of days of the tour and personal travel habits), specific preferences in catering and accommodation (consumption level, location, taste preferences, etc.), as well as travel interest preferences (natural scenic spots, commercial streets or historical and cultural attractions.), whether they have shopping needs, etc. Finally, 5,000 tourist data were collected through online questionnaires to build a tourist database.

The tourism resource data is collected by using python crawler technology through several major web pages or software platforms with a large number of users, and then sorted according to different dimensions. The classification and sources of tourism resource data are shown in Table 1. A total of more than 90 restaurants, 90 hotels, 80 famous scenic spots, and nearly 60 large shopping malls are collected. At the same time, the longitude and latitude data of the location are identified, and the tourism resource database is constructed.

3 Portrait and Travel Program Generation

3.1 Tourist Portraits and Tourism Resource Portrait

Tourists' basic information and travel preferences of meals, living, traveling, shopping and accommodation were collected through questionnaires, and the tourist portraits were carried out. Tourism resource portraits should be paired with each other from four aspects:

Table 1. Tourism resource data

Category of tourism resources	Data classification	Data source
Catering resources	Price, taste	Meituan
Accommodation resources	Price, location	Xiecheng
Scenic spot resources	History, art, nature, business, entertainment	Beijing travel network
Shopping resources	-	Dianping

meals, living, traveling, shopping and accommodation. The two portraits were matched with each other, and the overall portrait model was shown in Fig. 1. Finally, according to the user’s travel preference, the concept of “tourism personality” is put forward, which means travel preference and travel personality, which are classified into four categories: Culture type, Relaxation type, Experience type and Comprehensive Exploration type. The tourism personality is tested for users, as shown in Table 2.

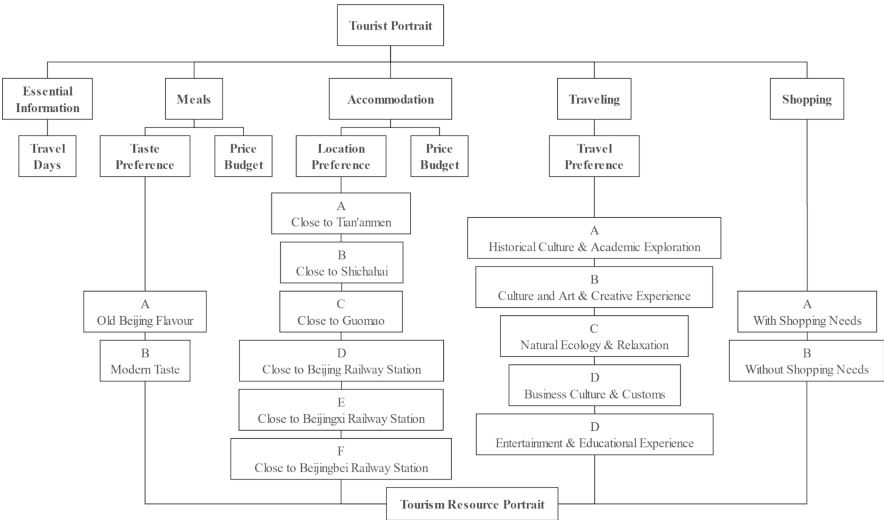


Fig. 1. Portrait of tourists

3.2 Tourist Portraits and Tourism Resource Portrait

Considering the time factor, add the scenic spot label 1/0.5, respectively, which means that the recommended travel time is one day/half a day, and recommend the corresponding plan according to the expected travel days of tourists; Considering the spatial factors, the accommodation places of tourists are first determined, and the places of traveling, meals, shopping and other places are generated based on the preferences of tourists and

Table 2. Classification of tourist cells

Travel lattice division	Paraphrase
Culture	This type of tourist is interested in cultural sites and enjoys learning about the local history and culture
Relaxation	This type of tourist tends to avoid the hustle and bustle, opting for natural scenery and resorts to relax and recharge
Experience	This type of tourist values a diverse range of experiences, including adventure activities, theme parks, or unique local experiences
Comprehensive Exploration	This type of tourist are interested in a wide range of travel experiences, have a comprehensive spirit of exploration and an open attitude

bipartite graph to ensure that the generated tourism plans and routes are reasonable and feasible.

Bipartite graph is a special graph in which the nodes can be divided into two disjoint sets, and each edge connects the nodes in the two sets, which has the characteristics of no circles (bipartite graph does not contain the cycle of odd edges), which meets the needs of tourists for staggered catering and sightseeing attractions. Therefore, we take the bipartite graph as the basis of matching algorithm.

Aiming at the longitude and latitude coordinates in the tourism resource data, in order to optimize the travel itinerary recommendation mechanism, this study adopts the method of constructing bipartite graph to make more appropriate arrangements for tourists' travel arrangements. The dichotomous graph is represented by an adjacency list (using a list to store the nodes connected by each node), and the polar coordinates of each node relative to the current node and the number of that point are the contents of the adjacency list.

In order to generate a bipartite graph, the formula for calculating relative polar coordinates using latitude and longitude is first listed, and the spherical distance between two points is calculated using the Haversine formula:

$$a = \sin^2\left(\frac{\Delta\Phi}{2}\right) + \cos(\Phi_1) \cdot \cos(\Phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right) \quad (1)$$

$$d = R \cdot 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right) \quad (2)$$

Calculate the azimuth:

$$\theta = \text{atan2}(\sin(\Delta\lambda) \cdot \cos(\Phi_2), \cos(\Phi_1) \cdot \sin(\Phi_2) - \sin(\Phi_1) \cdot \cos(\Phi_2) \cdot \cos(\Delta\lambda)) \quad (3)$$

In the above formula Φ_1 , Φ_2 is the latitude (radian) of the two points, $\Delta\lambda = \lambda_1 - \lambda_2$ is the longitude difference (radian) of the two points, d is the distance between the two points, θ is the azimuth Angle. For each new point in the bipartite graph, it is first judged which set of the bipartite graph it belongs to (take set A and B as an example). Assuming

that the point belongs to set A, the relative polar coordinates are calculated with the points in set B in turn, the points that meet the distance limit are taken (to avoid too long distance), and the new nodes are updated to the adjacency list.

When the path is recommended, the tourist accommodation point is used as the starting point to form a bipartite map, and the depth-first algorithm (DFS) is used to find the path. By limiting the polar coordinate angles of each three sides of the path, the planned route return trip is too long. If tourists have shopping needs, the selected shopping point will be added to the binary graph as the end point, and the breadth-first algorithm (BFS) will be used to find the appropriate path.

4 Software Implementation

By analyzing the user’s choices in questions 1 to 6, the user’s travel needs are determined, the user’s choices in question 7 determined the user’s tourism personality type, as shown in Fig. 2. The final travel plan is generated by determining whether the user has consumption needs and giving corresponding recommendations, as shown in Fig. 3. The accommodation part is based on the region, hotel type and budget range selected by the user. If the hotel is not found in the initial screening, continue to adjust the screening conditions in the next step. The recommendation content of the travel and dining part changes according to the user’s choice, and will avoid repeated recommendation according to the result, which is similar to the recommendation of the accommodation part. The shopping section will recommend different shopping areas according to the type of tourists if the user has relevant needs.

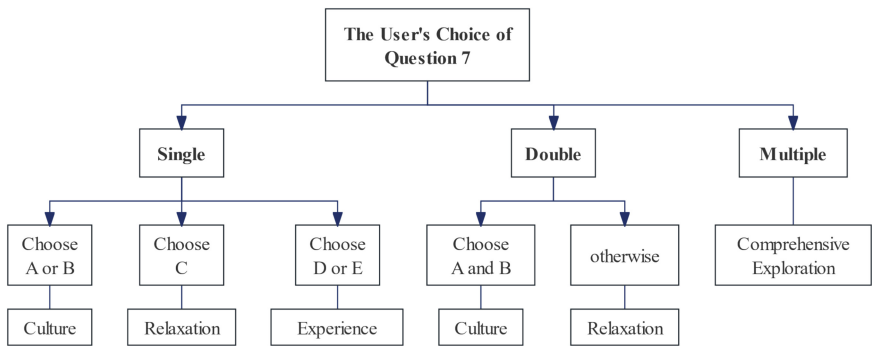


Fig. 2. Determination of tourism personality

The software part is written by Android Studio and the overall framework is shown in Fig. 4. *WelcomeActivity* is the initial interface, displaying welcome information and product introduction to guide users. *MainActivity* is the main interface, displaying survey content and storing user input results. Use *TravelPlanCalculator* to generate the final travel plan and pass it to *ResultActivity* to display the result. *TravelQuestion* is the data model and initializes the list of questions and options and passes it to *MainActivity* for display. *TravelPlanCalculator* is the core logic. By collecting user choices, analyzing

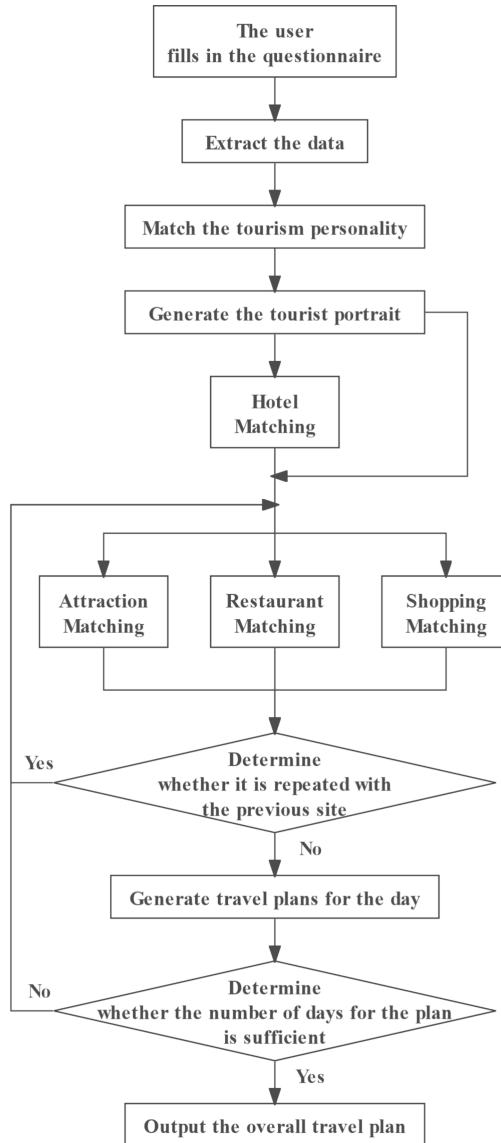


Fig. 3. Program process

the user's tourism personality type and calculating a suitable travel plan, *ResultActivity* is the result display. By obtaining the final result, the recommended travel plan of the user is displayed for the user to choose. Attraction, Hotel and Restaurant are data models, storing the information of hotels, attractions and restaurants required by the software and providing access and modification methods.

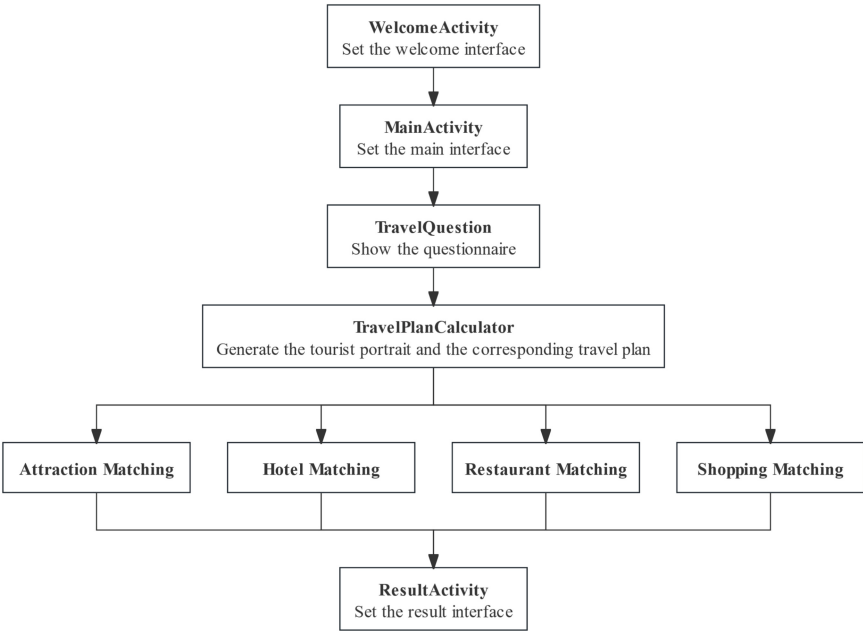


Fig. 4. Software framework

5 Conclusion

In view of the lack of individuation of existing travel recommendation services and generalization of travel route planning, this paper creatively proposed the concept and classification of “tourism personality” based on tourist portraits and tourism resource portraits, and proposed a travel plan recommendation algorithm based on bipartite graph and realized by software, so as to bring tourists high-quality tourism services and personalized experience to meet diversified needs. Which also assist tourists to make plans and improve decision-making efficiency. This paper finally helps to bring positive impact on tourism and economic progress and promote the construction of smart cities.

References

1. The State Council General Office issued a notice on the “Measures for Releasing Tourism Consumption Potential and Promoting High-Quality Development of the Tourism Industry”. Chinese State Council Gazette (29), 14–17 2023. (in Chinese)
2. Ai, J.: Research on personalized tourist routine recommendation based on improved collaborative filtering technology. Mod. Electron. Technol. **42**(23), 182–186 (2019). (in Chinese)
3. Hwang, S.Y., Yang, W.S.: On-tour attraction recommendation in a mobile environment. In: 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 661–666 (2012)

4. Yuan, Z., Wang, G.: Research on visitor profiling analysis method based on multi-source data in scenic and historic areas-taking Wudang mountain scenic and historic area as an example. *Architect. Cult.* (11), 72–75 (2022). (in Chinese)
5. Zhang, Z.: A study on user profiling of virtual tourism community based on ontology. Jilin University (2020). (in Chinese)
6. Zeng, J., Li, F., He, X., et al.: Fused collaborative filtering with user preference, geographical and social influence for point of interest recommendation. *Int. J. Web Serv. Res. (IJWSR)* **16**(4), 40–52 (2019)
7. Wang, S.: Building a smart reading recommendation system based on user profiles. *Libr. Sci. Res.* (4), 92–96 (2018). (in Chinese)
8. He, Y., Feng, Y., Zhao, S., et al.: Research on content recommendation based on Zhihu Q&A community – taking Logistics topic as an example. *Data Anal. Knowl. Discov.* **2**(09), 42–49 (2018). (in Chinese)
9. Li, T., Yan, D., Yang, X.: Research on microblog recommendation based on emotion weighted association rules. *Data Anal. Knowl. Discov.* **4**(04), 27–33 (2019). (in Chinese)
10. Li, J.: Collaborative Filtering Recommendation of Online Learning Resources for E-commerce Logistics Talent Training. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 351–366 (2024)
11. Abbasi-Moud, Z., Vahdat-Nejad, H., Sadri, J.: Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Syst. Appl.* **167**, 114324 (2021)



Trajectory Optimization for Mining Robots Based on an Improved Chicken Swarm Optimization Algorithm

Jianjian Yang^{1,2(✉)}, Teng Zhang^{1,2}, Guanghui Zhao³, Qi Cheng^{1,2},
Kangshuai Chen^{1,2}, Zhixiang Jiao^{1,2}, Shaowen Zheng¹, and Zifan Zhao¹

¹ School of Mechanical and Electrical Engineering, China University of Mining and Technology, Beijing 100083, China
yangjiannedved@163.com

² Inner Mongolia Research Institute, China University of Mining and Technology, Ordos 017000, Inner Mongolia Autonomous Region, China

³ National Energy Quasi Energy Group Limited Liability Company, Jungar Banner 010300, Inner Mongolia Autonomous Region, China

Abstract. In response to the problem of prolonged operation time caused by inefficient and suboptimal trajectories of electric shovels during excavation, this paper proposes a robotic arm trajectory planning method based on an Improved Chicken Swarm Optimization (ICSO) algorithm. A trajectory planning experiment is conducted, where sequences of time and position in joint space are used to construct interpolation curves via quintic polynomial functions, with time as the optimization objective and joint angular velocity and acceleration of the robotic arm as constraints. The position update methods for the rooster, hen, and chick within the traditional Chicken Swarm Optimization (CSO) algorithm are refined, incorporating a series of new position update operations to form the ICSO algorithm. Comparative simulations between the ICSO and traditional CSO algorithms show that the improved algorithm significantly reduces the motion time of the robotic arm, effectively enhancing operational efficiency.

Keywords: Trajectory optimization · Chicken Swarm Optimization · Improved Chicken Swarm Optimization · Motion time

1 Introduction

Large electric shovels are essential equipment in open-pit mining loading operations and play a critical role in enhancing mining efficiency. Technological advancements and innovations in electric shovels have thus become important drivers of progress in the mining industry [1–3]. The performance and productivity of these shovels are directly linked to ore extraction rates and cost-effectiveness. Currently, shovel operations rely mainly on highly trained drivers. However, increasing evidence indicates that human factors significantly impact shovel maintenance costs and reliability [4, 5]. Operational errors, inaccurate judgments, and fatigue can reduce excavation efficiency and may even

lead to safety incidents. To address these issues, the concept of robotic excavation has gained attention, where advanced control algorithms and sensor technologies enable the automation and intelligent operation of shovels. Compared to traditional human-operated methods, robotic excavation offers higher precision and stability, potentially improving excavation efficiency, reducing maintenance costs, and mitigating risks associated with human factors.

Effective trajectory planning is fundamental for robotic excavation. Dunbabin M et al. [6] designed an auxiliary system for cable shovels, helping drivers plan excavation trajectories and predict bucket fill. Jaemyung et al. [7] introduced a bucket trajectory planning algorithm based on deep learning with collision-avoidance algorithm, generating retracable trajectories for the excavator's controller that achieve the desired excavation volume and time without collisions. Yoo et al. [8] applied B-spline functions to parameterize excavator displacement and established a relationship between maximum pump flow and actuator motion, yielding an energy-optimized excavation trajectory. Chen Guangling et al. [9] proposed a multi-objective trajectory planning method based on the NSGA algorithm to achieve energy-efficient real-time operation of intelligent shovels. Bi Qiushi et al. [10] developed a two-stage multi-objective genetic algorithm to optimize cable shovel trajectory planning. Fu Tao et al. [11], addressing control lag and low precision in shovel trajectory tracking, proposed a tracking control strategy enhanced by deep learning. Using an LSTM neural network, they modeled the inherent inverse dynamics of the shovel system, transforming the optimal planned trajectory into a reference trajectory that minimizes tracking errors as a control system input. Compared with other earthmoving machinery, mining shovels have larger bucket capacities and operate under more complex excavation conditions, making their excavation performance more challenging to predict. Therefore, trajectory planning and optimization are critically important.

This paper primarily focuses on establishing mining robot kinematics, generating precise excavation trajectories using polynomial interpolation, and incorporating the CSO algorithm. Through experimental validation, we introduce Cauchy mutation, Cuckoo Search algorithm, and directed mutation strategies into CSO, achieving a 17.09% reduction in total time compared to the traditional CSO planning approach.

2 Related Work

Robot kinematics describes the positional, velocity, and acceleration changes over time from a geometric point of view, including forward kinematics and inverse kinematics [12]. Forward kinematics calculates the position and orientation of the end-effector based on the joint angles, whereas inverse kinematics determines the precise joint angles required to achieve specific end-effector positions and orientations [13].

2.1 D-H Parameter Modeling

Kinematic modeling methods include geometric modeling and D-H parameter modeling, with D-H modeling being widely used for its general applicability to both serial and parallel robots. In this study, the D-H modeling method was adopted. As shown in

Fig. 1, the relationship between the end-effector and the base of the robotic arm was established by setting the shovel's base center as the origin, and constructing the initial coordinate system. The D-H linkage coordinate systems for each joint of the shovel were constructed, where represents the base rotation angle, and represents the slewing angle of the pushing mechanism.

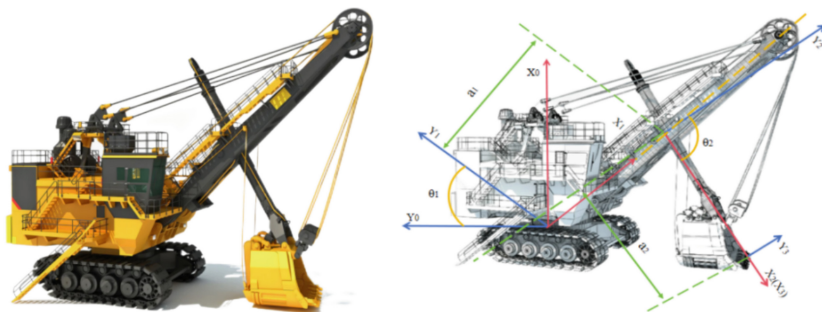


Fig. 1. D-H Coordinate System of Electric Shovel

The D-H parameters corresponding to this model are shown in Table 1.

Table 1. D-H Model Parameters of the Electric Shovel

Joint	Joint Offset d_n/m	Link Length a_n/m	Twist Angle $\alpha_n/(^{\circ})$	Joint Angle $\theta_n/(^{\circ})$
1	0	a_1	0	θ_1
2	0	a_2	α_2	θ_2
3	0	0	0	0

With the joint angles θ_i ($i = 1 \sim 2$) known, forward kinematics is used to calculate the position and orientation of the end-effector based on the given joint angles. This process involves transforming coordinates from frame i to $i-1$ using a transformation matrix ${}^{i-1}T_i$. Any given frame i can be converted precisely and effectively to frame $i-1$ through a series of four relative transformations. The D-H transformation matrix ${}^{i-1}T_i$ is defined as follows:

$${}^{i-1}T_i = T_{RZ-1}(\theta_i)T_Z(d_i)T_X(a_i)T_{RX}(\theta_i) \tag{1}$$

Based on the established D-H coordinate system, the following relationship holds:

$${}^0T = \begin{bmatrix} c\theta_1 & -s\theta_1 & 0 & 0 \\ s\theta_1 & c\theta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} {}^2T = \begin{bmatrix} c\theta_2 & -s\theta_2 & 0 & a_1 \\ s\theta_2 & c\theta_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} {}^1T = \begin{bmatrix} 1 & 0 & 0 & a_2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Through sequential multiplication, the transformation matrix of the bucket teeth endpoint with respect to the base coordinate system can be derived:

$$\begin{aligned}
 {}^0_3T &= {}^0_1T {}^1_2T {}^2_3T \\
 &= \begin{bmatrix} c\theta_{12} - s\theta_{12} & -c\theta_1 s\theta_2 - s\theta_1 c\theta_2 & 0 & a_2(c\theta_{12} - s\theta_{12}) + a_1 c\theta_1 \\ s\theta_1 c\theta_2 + c\theta_1 s\theta_2 & -s\theta_{12} + c\theta_{12} & 0 & a_2(s\theta_1 c\theta_2 + c\theta_1 s\theta_2) + a_1 c\theta_1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3)
 \end{aligned}$$

2.2 Inverse Kinematics Solution

Inverse kinematics involves calculating the precise set of joint angles in joint space necessary to achieve a given position and orientation of the end-effector in Cartesian space [14]. This reverse calculation is fundamental for robotic control. In the trajectory planning and control of mining robots, the process is carried out in joint space; therefore, it is essential to apply inverse kinematics modeling based on the bucket's position and orientation to obtain the exact inverse kinematics solution [15].

To solve for θ_1 , Eq. (3) is left-multiplied by ${}^0_1T^{-1}(\theta_1)$, yielding:

$${}^0_1T^{-1}(\theta_1) {}^0_3T = {}^1_3T \quad (4)$$

Thus, we have:

$$\begin{bmatrix} c\theta_1 & s\theta_1 & 0 & 0 \\ -s\theta_1 & c\theta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = {}^1_3T \quad (5)$$

Additionally,

$${}^1_3T = {}^1_2T {}^2_3T = \begin{bmatrix} c\theta_2 - s\theta_2 & 0 & a_2 c\theta_2 + a_1 \\ s\theta_2 & c\theta_2 & 0 & a_2 s\theta_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

By setting corresponding values (2,3) on both sides of Eqs. (5) and (6), we obtain:

$$-s\theta_1 a_x + c\theta_1 a_y = 0 \quad (7)$$

which simplifies to:

$$\theta_1 = A \tan 2(a_y, a_x) \quad (8)$$

Next, equating the (1,1) terms on both sides of Eqs. (5) and (6) allows us to find θ_2 :

$$\theta_2 = \arccos(c\theta_1 n_x + s\theta_1 n_y) \quad (9)$$

This solution provides the inverse kinematic solution for the robotic shovel arm.

2.3 Excavation Trajectory Generation

During excavation, the bucket must move smoothly from the starting point to the end-point, a process that involves critical factors such as velocity and acceleration. To achieve precise trajectory planning and improve operational efficiency, this paper applies quintic polynomial interpolation [16]. This method uses polynomial functions to fit curves between points, achieving smooth transitions between points, thus optimizing trajectory planning and execution.

The mathematical expression for quintic polynomial interpolation is:

$$Q(t) = a_5t^5 + a_4t^4 + a_3t^3 + a_2t^2 + a_1t + a_0 \quad (10)$$

The corresponding joint velocity and acceleration are represented as:

$$V = \dot{Q}(t) = 5a_5t^4 + 4a_4t^3 + 3a_3t^2 + 2a_2t + a_1 \quad (11)$$

$$a = \ddot{Q}(t) = 20a_5t^3 + 12a_4t^2 + 6a_3t + 2a_2 \quad (12)$$

To analyze the rotation joint of the boom arm around the pushing center, we consider a rotating joint as an example. The start time of the motion trajectory within a specified time interval is denoted as t_0 , and the endpoint as t_1 . At t_0 and t_1 , the rotation angles of the boom arm are θ_0 and θ_1 , respectively, with corresponding angular velocities $\dot{\theta}_0$ and $\dot{\theta}_1$. Based on these parameters, the following system of equations is constructed:

$$\begin{cases} Q(t_0) = \theta_0 \\ Q(t_1) = \theta_1 \\ \dot{Q}(t_0) = \dot{\theta}_0 \\ \dot{Q}(t_1) = \dot{\theta}_1 \\ \ddot{Q}(t_0) = \ddot{\theta}_0 \\ \ddot{Q}(t_1) = \ddot{\theta}_1 \end{cases} \quad (13)$$

By solving (13) along with (10), (11), and (12), we derive:

$$\begin{cases} a_0 = \theta_0 \\ a_1 = \dot{\theta}_0 \\ a_2 = \frac{\ddot{\theta}_0}{2} \\ a_3 = \frac{20(\theta_1 - \theta_0) + (8\dot{\theta}_1 + 12\dot{\theta}_0)(t_1 - t_0) - (3\ddot{\theta}_0 - \ddot{\theta}_1)(t_1 - t_0)^2}{2(t_1 - t_0)^3} \\ a_4 = \frac{-30(\theta_1 - \theta_0) + (14\dot{\theta}_1 + 16\dot{\theta}_0)(t_1 - t_0) + (3\ddot{\theta}_0 - \ddot{\theta}_1)(t_1 - t_0)^2}{2(t_1 - t_0)^3} \\ a_5 = \frac{12(\theta_1 - \theta_0) - 6(\dot{\theta}_1 + \dot{\theta}_0)(t_1 - t_0) - (\ddot{\theta}_0 - \ddot{\theta}_1)(t_1 - t_0)^2}{2(t_1 - t_0)^3} \end{cases} \quad (14)$$

Substituting the values obtained in (14) back into Eqs. (10), (11), and (12) yields the final expression for the quintic polynomial.

3 Chicken Swarm Optimization

3.1 Traditional Chicken Swarm Optimization (CSO)

The Chicken Swarm Optimization (CSO) algorithm, a global optimization algorithm, simulates the behavioral traits and intra-group hierarchy of a chicken swarm. The core concept is to divide chickens into roosters, hens, and chicks based on their fitness values, with each type of chicken following specific movement rules to compete for food [17]. Roosters play the role of explorers, moving further from their current positions to search for potentially better solutions. They generally have higher mobility and larger search radii. Hens focus on refining their search within the current solution space to improve precision and seek local optima, while chicks follow the roosters or hens, learning their search strategies and gradually enhancing their own capabilities. As a global optimization algorithm, CSO has shown unique advantages and potential in solving complex optimization problems. By simulating the natural behavior traits observed in chicken swarms, CSO effectively balances exploration and exploitation, achieving efficient global search and local optimization.

The rooster, exhibiting the highest fitness value, acts as the group leader and enjoys a broader foraging area. The position update formula for a rooster is as follows:

$$x_{i,j}(t+1) = x_{i,j}(t) \times [1 + \text{randn}(0, \sigma^2)] \quad (15)$$

$$\sigma^2 = \begin{cases} 1, & f_i < f_k \\ \exp(\frac{f_k}{|f_i| + \varepsilon}), & f_i \geq f_k, k \in [1, N_R], k \neq i \end{cases} \quad (16)$$

where k represents another rooster k distinct from i , $k \neq i$; $\text{randn}(0, \sigma^2)$ is a Gaussian distribution with mean 0 and variance σ^2 ; f is each rooster's optimal fitness value; and ε is a small positive constant to prevent division by zero.

Among the group, hens represent the majority, and their position updates are as follows:

$$x_{i,j}(t+1) = x_{i,j}(t) + S_1 \text{rand}[x_{r1,j}(t) - x_{i,j}(t)] + S_2 \text{rand}[x_{r2,j}(t) - x_{i,j}(t)] \quad (17)$$

$$S_1 = \exp(\frac{f_i - f_{r1}}{[|f_i| + \varepsilon]}) \quad (18)$$

$$S_2 = \exp(f_{r2} - f_i) \quad (19)$$

where rand denotes a random number within the interval $[0,1]$; r_1 is the rooster in the same group as hen i ; r_2 is another chicken (either a rooster or hen) from the optimization group but different from i and r_1 ; $r_1 \neq r_2$.

For chicks, their position updates are defined by:

$$x_{i,j}(t+1) = x_{i,j}(t) + FL \times [x_{m,j}(t) - x_{i,j}(t)] \quad (20)$$

where $x_{m,j}(t)$ denotes the mother hen's position that chick i follows during the t^{th} iteration, and FL is the chick's adjustment parameter, a random value between 0 and 2.

3.2 Improved Chicken Swarm Optimization (ICSO)

The update equations in the CSO algorithm are more conducive to maintaining population diversity but lead to issues such as low solution precision, weak local exploitation, and slower convergence in the later stages [18]. An effective position update mechanism improves solution precision; however, if misused, it can cause individuals in the population to over-rely on optimal solutions, reducing population diversity and increasing the risk of getting trapped in local optima. To address this, some researchers introduced the global optimal position into rooster updates [19].

The rooster update process essentially involves individual mutation. To counter the decreasing diversity of the population during algorithm iterations, we introduced Cauchy mutation as a mutation operator and modified the rooster position update as $\max t/2$. This approach enhances individual mutation capability in later stages, ensuring a progressive increase in mutation variance to counteract declining diversity. Furthermore, we applied a linear decrement strategy to accelerate optimization convergence rate in the later stages, as defined below:

$$x_{i,j}^{G+1} = \begin{cases} \frac{\max t - t}{\max t} \times x_{i,j}^G \times (1 + \text{randn}(0, \sigma^2)) \\ \frac{\max t - t}{\max t} \times x_{i,j}^G \times (1 + \text{cauchy}(0, \sigma^2)) \end{cases} \quad (21)$$

The preferential random walk process appears in many algorithms, with the Cuckoo Search algorithm being a typical example. Preferential random wandering is inherently a weighted factor-driven process. To improve the accuracy of hen and chick foraging behavior simulation, Eq. (23) was specifically introduced to enhance the emphasis on individual learning processes. By thoroughly drawing on and leveraging the position information of each individual, each member of the group demonstrates a stronger memory characteristic in foraging, allowing a more accurate simulation of chicken foraging behavior.

The hen position update equation in ICSO is:

$$x_{i,j}^{G+1} = w(t) \times x_{i,j}^G + S_1 R(x_{r1}^G - x_{i,j}^G) + S_2 R(x_{r2}^G - x_{i,j}^G) \quad (22)$$

$$w(t) = w_{\min} + (w_{\max} - w_{\min}) \times \exp(-k \times (\frac{t}{\max t})^2) \quad (23)$$

Through analysis of Eq. (23), it is confirmed that maintaining a relatively higher weight value in the early stages of the optimization process effectively broadens the search scope, thereby ensuring superior global optimization performance. In the later stages of optimization, appropriately reducing the inertia weight helps to improve the algorithm's local exploitation capability, facilitating faster convergence. This weighting adjustment effectively balances the algorithm's global and local search capabilities, significantly boosting algorithm stability.

For chicks following hens, to enhance global search ability, a directed mutation strategy is adopted. The chick position update in ICSO is as follows:

$$x_{i,j}^{G+1} = \text{best} + F \times (x_{m,j}^G - x_{i,j}^G) \quad (24)$$

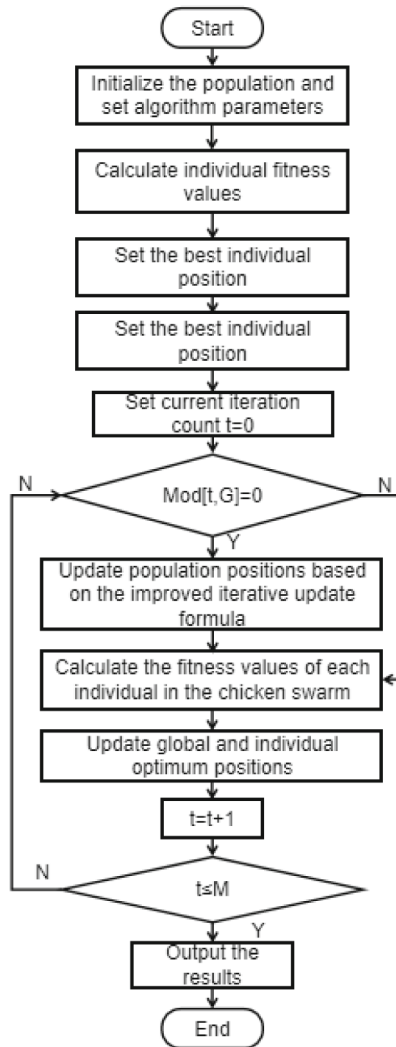


Fig. 2. Flowchart of the Chicken Swarm Optimization Algorithm

3.3 Algorithm Steps

Figure 2 presents the CSO algorithm flowchart. The CSO algorithm follows these steps:

Step 1: Set the initial population size N , specify spatial dimensions D , determine the maximum number of iterations T , and establish key parameters such as NR , NH , NC , and NM ;

Step 2: Evaluate each individual's fitness value, identify the best individual position, and initialize the iteration counter to 0, i.e. $t = 0$;

Step 3: Determine if the remainder of the current iteration count t divided by G is zero. If it is zero, proceed; otherwise, reorder the individuals within the group based on fitness and adjust the hierarchy accordingly;

Step 4: Update the position of each individual according to the improved iteration formula;

Step 5: Re-evaluate the fitness values and update the individual and global optima;

Step 6: Increment t by 1, i.e. $t = t + 1$. If the termination condition is met, proceed to Step 7; otherwise, return to Step 3 and continue iterating;

Step 7: Upon completing all iterations, output the final global optimum.

3.4 Performance Testing

To evaluate the performance of the ICSO algorithm, we conducted comparative simulations between the CSO and ICSO algorithms on the MATLAB platform. In this test, the population size was set to 50, and the maximum number of iterations was limited to 100 to ensure accuracy and reliability of the results. The Rosenbrock function was chosen as the test function, a well-known non-convex function used to evaluate optimization algorithms' performance. It is defined as follows:

$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2 \quad (25)$$

The Rosenbrock function, known for its complexity and prominence in optimization research across mathematics, computer science, and engineering, has contours forming a distinctive parabolic shape, which introduces additional complexity to the optimization process. These parabolic contours are not simple parabolas; rather, they exhibit curvature and distortions, making it easy for algorithms to become trapped in local optima rather than reaching the global optimum.

The global minimum of the Rosenbrock function lies in a “banana-shaped” valley. This shape often requires the algorithm to traverse multiple local optima before arriving at the true global minimum, making the Rosenbrock function a challenging optimization problem for testing algorithm performance.

From the comparison between subfigures (a) and (b) in Fig. 3, it is evident that, given the same number of iterations, the convergence performance of the traditional CSO algorithm is markedly inferior to that of the improved ICSO. This difference is apparent not only in the numerical convergence results but also in the convergence speed. Data from the figures clearly indicate that, as the number of iterations increases, CSO gradually approaches the optimal solution, but its convergence process is relatively slow and prone to becoming trapped in local optima. In contrast, ICSO demonstrates faster convergence and superior global search capabilities, achieving a near-optimal or global solution within the same number of iterations, whereas CSO does not reach this level.

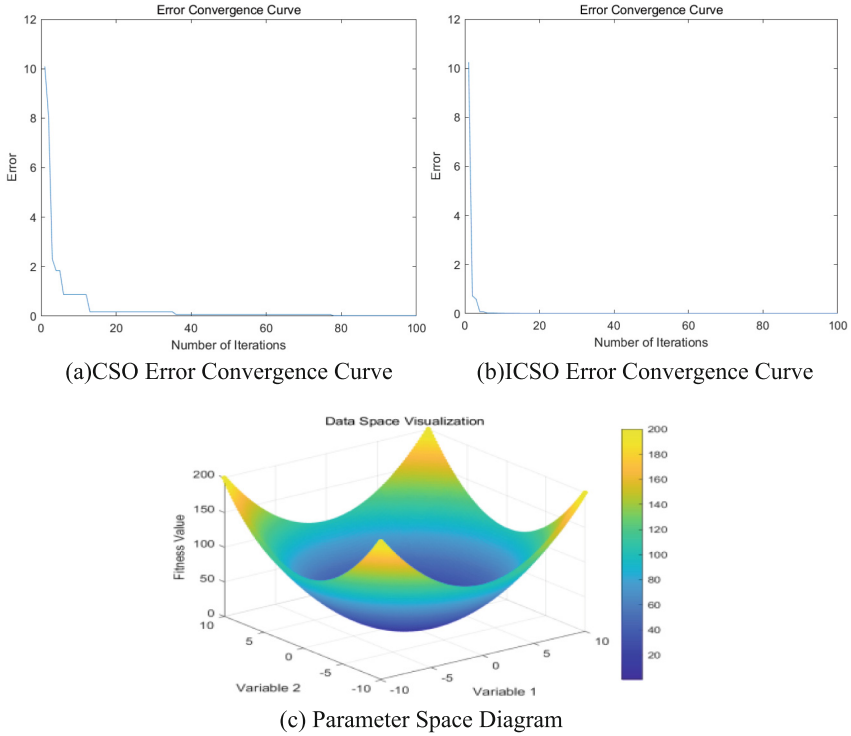


Fig. 3. Performance Testing of CSO and ICSO

4 Experimental Simulation

With the ICSO algorithm's performance validated in Sect. 3.4, we applied it to the shovel arm of the electric shovel by constructing a kinematic model using MATLAB's Robotics Toolbox. Given the six endpoint positions in Cartesian space, the robotic arm was designed to follow a trajectory through these points. The coordinates of the six endpoints are presented in Table 2.

In MATLAB, we employed quintic polynomial interpolation to optimize the movement trajectory across these six trajectory points, comparing the performance of CSO and ICSO optimizations. The experiment was set with 100 iterations, a population size of 50, and a rooster-to-hen-to-chick ratio of 1:2:2. Figures 4 and 5 present the results, showing the changes in joint angles, angular velocity, and angular acceleration over time for different schemes.

After ICSO optimization, the changes in the robotic arm's joint angles exhibit faster response times compared to the traditional CSO method. Under identical movement tasks, the ICSO-optimized robotic arm adjusts its joint angles more rapidly, significantly reducing the time required to complete the entire movement. The specific operation times are shown in Table 3.

Table 2. Endpoint Position Coordinates

Endpoint	Endpoint Position Coordinates		
	x	y	z
1	1	0.5	1
2	0.5	−0.5	1.2
3	−0.2	0.8	0.5
4	−0.8	−0.3	1
5	0.6	0.2	0.8
6	−0.4	0.6	1

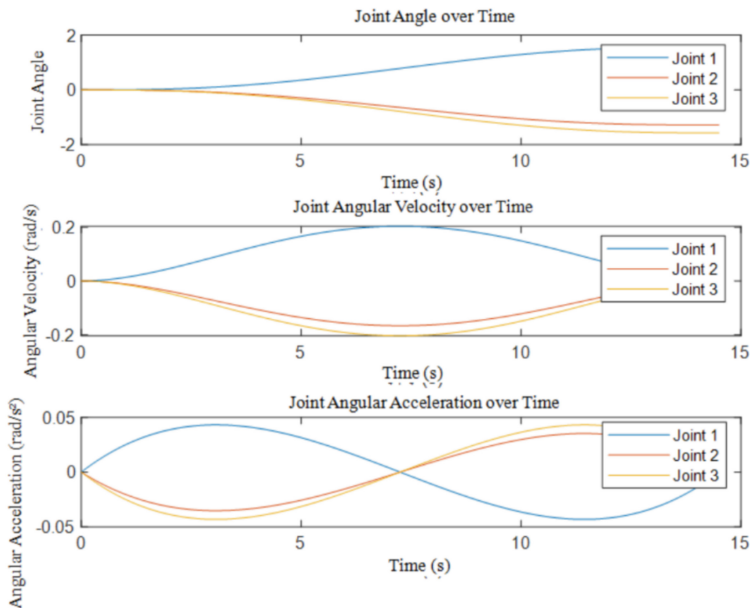


Fig. 4. CSO Optimization Curves

From the detailed data comparison in Table 3, it is clear that the ICSO algorithm significantly reduces the overall motion time of the robotic arm trajectory compared to the traditional CSO planning method, achieving a reduction of 17.09%. Consequently, the ICSO algorithm proposed in this paper finds more optimal solutions through a more intelligent search strategy and more efficient computation, achieving an optimization of the robotic arm’s motion trajectory curve and effectively lowering its operational time.

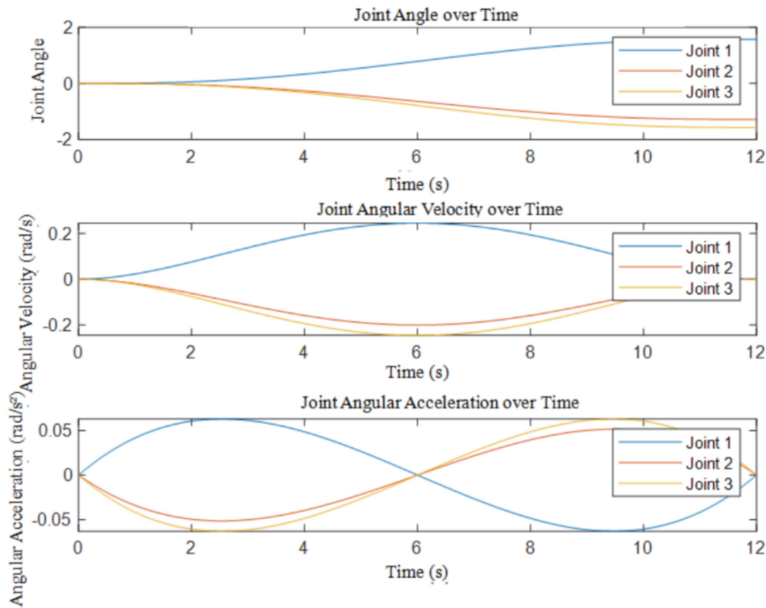


Fig.5. ICSO Optimization Curves

Table 3. Robotic Arm Motion Times for Different Schemes

Trajectory	1	2	3	4	5	6	Total Time (s)
CSO	2.06	1.23	3.12	1.56	3.45	3.09	14.51
ICSO	1.94	0.88	2.86	1.21	2.75	2.39	12.03

5 Conclusion

This paper presents an improved Chicken Swarm Optimization (ICSO) algorithm, which enhances global search capabilities by incorporating Cauchy mutation in rooster position updates to enable faster convergence to high-quality solutions. Additionally, by adapting the Cuckoo Search algorithm for hen position updates, search efficiency is increased, accelerating the convergence toward the global optimum. Furthermore, chick position updates were improved using a directed mutation strategy, maintaining randomness and exploration throughout the learning process and avoiding premature convergence to local optima. These comprehensive position update optimizations significantly improve convergence speed and global search capabilities. In the application of robotic arm trajectory optimization, the ICSO method was tested and compared with CSO trajectory optimization results. Experimental results demonstrate that the ICSO-based trajectory optimization method significantly reduces robotic arm motion time, achieving a 17.09% reduction in total movement time, thereby effectively reducing the robotic arm's operational time.

References

1. Sun, W., Li, E., Wang, X., Guo, Z., Li, X.: Optimal excavation trajectory planning for intelligent excavators. *J. Dalian Univ. Tech.* **58**(03), 246–253 (2018). (in Chinese)
2. Wei, B., Gao, F., Chen, J., He, J., Wu, S., Song, Q.: Mechanics performance of three-degree-of-freedom excavating mechanism of an electric shovel. *Proc. Inst. Mech. Eng. C J. Mech. Eng. Sci.* **225**(6), 1443–1457 (2011)
3. Khorzoughi, M.B., Hall, R.: A study of digging productivity of an electric rope shovel for different operators. *Minerals* **6**(2), 48 (2016)
4. Topno, S.A., Lalit, K., Umre, B.: Energy efficiency assessment of electric shovel operating in opencast mine. *Energy* **230**, 120703 (2021)
5. Fang, Y., et al.: Research on path planning and trajectory tracking of an unmanned electric shovel based on improved APF and preview deviation fuzzy control. *Machines* **10**(8), 707 (2022)
6. Dunbabin, M., Corke, P.: Autonomous excavation using a rope shovel. *Field Serv. Robots* **25**, 555 (2006)
7. Huh, J., et al.: Deep learning-based autonomous excavation: a bucket-trajectory planning algorithm. *IEEE Access* **11**, 38047–38060 (2023)
8. Yoo, S., Park, C.G., You, S.H., Lim, B.: A dynamic-based optimal trajectory generation for controlling an automated excavator. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **224**(10), 2109–2119 (2010)
9. Chen, G., Zhang, T., Fu, T., Wang, L., Song, X.: Multi-objective optimal excavation trajectory planning for intelligent electric shovels based on NSGA-II. *Mod. Manufact. Eng.* (02), 142–149 (2024). (in Chinese)
10. Bi, Q., Wang, G., Wang, Y., Yao, Z., Hall, R.: Digging trajectory optimization for cable shovel robotic excavation based on a multi-objective genetic algorithm. *Energies* **13**(12), 3118 (2020)
11. Fu, T., Zhang, T., Cui, Y., Song, X.: Deep learning-enhanced trajectory tracking control of intelligent mining electric shovels. *J. Mech. Eng.* 1–10 (2024). (in Chinese)
12. Wang, M., Luo, J., Yuan, J., Walter, U.: Coordinated trajectory planning of dual-arm space robot using constrained particle swarm optimization. *Acta Astronaut.* **146**, 259–272 (2018)
13. Ye, H., Wang, D., Wu, J., Yue, Y., Zhou, Y.: Forward and inverse kinematics of a 5-DOF hybrid robot for composite material machining. *Robot. Comput.-Integr. Manufact.* **65**, 101961 (2020)
14. Rubio, J.D., Aquino, V., Figueroa, M.: Inverse kinematics of a mobile robot. *Neural Comput. Appl.* **23**(1), 187–194 (2013)
15. Sun, Z., Liu, S., Wang, Y., Li, H.: Time-optimal and pulse-optimal trajectory planning of excavators. *Mach. Des. Manufact.* **01**(5), 304–307+313 (2024). (in Chinese)
16. Gai, R., Du, X.: Research on high-quality quintic polynomial speed planning algorithm. *Mach. Des. Manuf.* 1–7 (2024). (in Chinese)
17. Ma, L., Li, C., Si, Q.: Trajectory planning for mechanical arm based on improved chicken swarm optimization algorithm. *Comb. Mach. Tools Autom. Process. Technol.* **12**, 61–66 (2023). (in Chinese)
18. Wang, K., Li, Z., Zhang, K.: Mutation chicken swarm optimization based on nonlinear inertia weight. In: 2017 3rd IEEE International Conference on Computer and Communications, Chengdu, pp. 2206–2211. IEEE (2017)
19. Liang, X., Kou, D., Wen, L.: An improved chicken swarm optimization algorithm and its application in robot path planning. *IEEE Access* **8**, 49543–49550 (2020)



Evaluation of the Energy-Saving Ability of Pulse-and-Glide Mode in Vehicles: A Coupled Simulation Approach

Honggang Li¹(✉), Changhui Yan², and Jinjun Sun²

¹ College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin 150000, China

lhg@nefu.edu.cn

² School of Civil Engineering and Transportation, Northeast Forestry University, Harbin 150000, China

{ych, jinjunsun2001}@nefu.edu.cn

Abstract. The growing concerns over the energy crisis and climate change have led to increasing demands for fuel economy in automobiles, which has accelerated the adoption of the (Pulse-and-Glide) PnG mode as a energy-saving technology based on speed control. While numerous studies have validated the energy-saving potential of the PnG mode, there is still a lack of quantitative analysis and identification of the key factors influencing its energy-saving performance, as well as optimization strategies for the PnG mode. This study uses a specific vehicle model as the subject, establishing a coupled simulation platform with CarSim and MATLAB/Simulink. The energy-saving effects of the PnG mode are analyzed through various control variables. The results show that, compared to the constant speed mode, the PnG mode achieves significant energy-saving in the low-to-medium speed range, with the energy-saving effect diminishing as the average speed increases. Additionally, idle fuel consumption is identified as one of the key factors affecting the energy-saving performance of the PnG mode. Therefore, during the G-phase, it is recommended to disengage the transmission system from the wheels and ensure that the engine stops fuel consumption, thereby maximizing the energy-saving effect of the PnG mode.

Keywords: Speed control · Energy-saving · PnG mode

1 Introduction

Due to the continuous depletion of fossil fuels and the intensification of the greenhouse effect, which has led to growing expectations and requirements regarding the energy consumption of transportation vehicles [1]. However, fossil fuels still account for 98% of the energy used in transportation [2]. Against this backdrop, researchers have proposed various methods to enhance fuel efficiency [3, 4]. One of the primary approaches to improving fuel economy involves reducing aerodynamic drag, decreasing vehicle

weight, and enhancing powertrain efficiency, which are achieved by modifying the vehicle's body structure or powertrain system [5]. Compared to modifying the vehicle itself, empowering the vehicle through CAV (Connected and Automated Vehicle) is another important approach [6, 7]. However, due to the limitations in the widespread adoption and cost of autonomous driving technologies, as well as the significant influence of driver control on the vehicle's motion state, some studies have focused on optimizing driver control strategies as a means to improve fuel economy [8, 9]. In this field, the Pulse-and-Glide (PnG) mode is considered an effective method for improving fuel economy through speed control [10, 11].

PnG is an energy-efficient driving technique based on periodic control, which optimizes fuel consumption during different driving phases by periodically adjusting the vehicle's power input, thereby enhancing energy utilization efficiency. Additionally, numerous studies have validated the energy-saving effects of the PnG mode during vehicle operation. For example, Cao et al. [12] formulated an optimization problem aimed at minimizing fuel consumption by considering variables such as initial speed, duration of the PnG mode, and output power under the PnG mode. They developed a corresponding vehicle dynamics model and solved the optimization problem using numerical analysis methods. The results of their study indicated that, compared to the constant speed (CS) mode, the PnG mode offers better energy-saving performance. Similarly, Xu et al. [13] formulated an optimal control problem aimed at minimizing fuel consumption per 100 km and developed a longitudinal vehicle control model specific to the PnG mode. They then applied the pseudospectral method to convert the optimization problem into a nonlinear programming problem for solution. Their results also demonstrated the energy-saving advantages of the PnG mode over the CS mode. Building upon theoretical validation, some researchers have further investigated the energy-saving effects of the PnG mode through simulation methods. Sohn et al. [14] first derived the energy-saving mechanism of the PnG mode, then constructed a vehicle simulation model with acceleration and its duration as control parameters, and conducted tests. Based on the simulation results, Sohn found that, compared to the CS mode, the PnG mode could achieve a fuel saving rate of over 30% when the average vehicle speed was below 40 km/h. Additionally, Kim et al. [15] conducted simulations on the energy-saving effects of the PnG mode by constructing a vehicle powertrain model. The simulation results showed that, at an average speed of 100 km/h, the PnG mode was more energy-saving than the CS mode. Furthermore, when idle fuel consumption was excluded, the PnG mode reduced fuel consumption by an additional 3.36% compared to when idle fuel consumption was included.

Although numerous studies have verified the feasibility of the PnG mode in achieving energy-saving effects through speed control, most of these are limited to theoretical analysis and lack sufficient data to support a quantitative examination of the influencing factors. While simulations can also demonstrate the energy-saving potential of the PnG mode, the results are constrained by the simulation conditions, such as average speed and acceleration, which may be set too high or differ excessively. As a result, the energy-saving effects of the PnG mode obtained in these simulations exhibit certain limitations. For example, when considering speed as a factor, a more comprehensive and practical range of speeds should be considered, such as high-speed segments represented by

highways [16], low-speed segments represented by urban roads [17], and intermediate-speed segments. To this end, this study is based on a coupled traffic simulation approach established with CarSim and MATLAB/Simulink, utilizing a specific vehicle model as the subject. The control strategy of the PnG mode is integrated into the simulation, and the energy-saving potential of the PnG mode, along with its key influencing factors, is investigated through various variables such as speed fluctuation. Based on this analysis, an optimized control strategy for the PnG mode is proposed to advance its practical application in road traffic.

2 Method

2.1 Design of Simulation Scheme

To fully leverage the advantages of simulation analysis in assessing and optimizing the fuel consumption characteristics of a vehicle under the PnG mode, and to provide more comprehensive and accurate simulation results, this study employs a coupled simulation approach using CarSim and MATLAB/Simulink.

The Setup of Carsim

The vehicle model selected is an SUV, and its main relevant parameters are shown in Table 1.

Table 1. Main parameters of vehicle

Project	Parameters
Vehicle Mass (kg)	1661
Engine Rated Power (kW)	150
Final Drive Ratio	4.1
Number of Transmission Gears	6
Transmission Type	Manual
Frontal Area (m ²)	2.8
Air Density (kg/m ³)	1.206
Drag Coefficient	0.3
Tire Specification	225/60 R18

The selected engine has a rated power of 150 kW, and the relationship between engine speed, output torque, and fuel consumption performance is shown in Fig. 1.

The selected transmission is a 6-speed manual gearbox, with a shift duration of 0.15 s. The gear ratios and related parameters of the transmission are shown in Table 2.

The Setup of MATLAB/Simulink

The MATLAB/Simulink simulation platform mainly consists of two parts: the interface

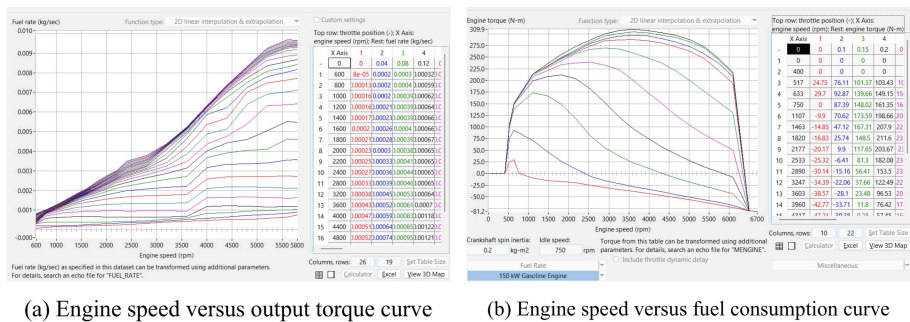


Fig. 1. Engine speed vs output torque and fuel consumption diagram

Table 2. Transmission gear ratios and main performance parameters

Transmission Gear	Final Drive Ratio	Moment of inertia	Transmission efficiency
1	3.82	0.037	0.92
2	2.26	0.034	0.92
3	1.64	0.042	0.95
4	1.25	0.04	0.95
5	1.06	0.04	0.98
6	0.92	0.04	0.99

module and the control system. The communication between Simulink and CarSim is primarily achieved through the CarSim S-Function module. The configuration of the CarSim interface module in Simulink for this simulation platform is shown in Fig. 2(a).

The control system in Simulink is mainly divided into upper-level control and lower-level control. The upper-level controller takes the input parameters of average speed and speed fluctuation amplitude (PnG mode speed fluctuation) as inputs, and then the PnG mode control module outputs the desired control acceleration. The lower-level controller, on the other hand, outputs control parameters to the vehicle in CarSim through the inverse drive control module, thereby achieving the desired acceleration specified by the upper-level control.

The design of the upper-level controller mainly includes determining the optimal acceleration during the P-phase of PnG mode, as well as the design of the PnG mode control module.

The maximum number of iterations is set to 10, and the PSO iterative calculation is performed to obtain the optimal acceleration values for different speeds, as shown in Fig. 2(d). The PnG mode control module, as shown in Fig. 2(b), is mainly used to control the simulated vehicle in CarSim to operate in the specified PnG mode.

Since the PnG mode discussed in this chapter does not involve the braking situation of the brake, the design focus of the lower-level controller is concentrated on the reverse drive control module. Based on the MAP diagram of the engine output torque, the

corresponding throttle opening MAP diagram can be obtained, as shown in Fig. 2(e). The reverse drive control model of the lower-level controller in this simulation platform is shown in Fig. 2(c).

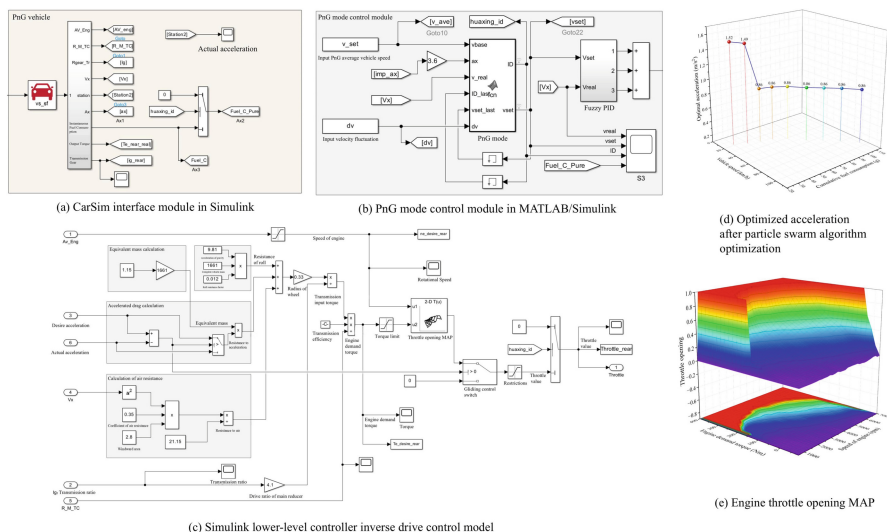


Fig. 2. Coupled traffic simulation construction

2.2 The Design of PnG Mode Control Strategy

In order to deeply investigate the energy-saving potential of the PnG mode under different driving speeds and speed fluctuation amplitudes, a series of comprehensive coupled traffic simulations were designed. Within this simulation framework, the driving speed range of the PnG mode is set from 30 to 120 km/h, with increments of 5 km/h. Meanwhile, the speed fluctuation amplitude (defined as the difference between the maximum and minimum speeds during the operation of the PnG mode) is set between 2 and 20 km/h, with increments of 2 km/h. This speed range fully covers the complete operating range of the vehicle, from low-speed urban driving to the upper speed limit on highways. A total of 190 sets of simulations were implemented, and each set of simulations ensured that the vehicle completed more than seven stable PnG running segments. Data acquisition starts after the first stable PnG segment in the simulation flow to exclude possible data fluctuations in the initial segment and ensure the stability of the recorded data. In addition, the running data of at least six stable PnG segments were recorded continuously for subsequent analysis. In the G-phase of PnG mode, a strategy of complete power interruption is adopted. The fuel consumption data is collected in two subsets: one subset includes the idling fuel consumption during gliding, while the other subset excludes this portion of fuel consumption.

3 Simulation results and discussion

3.1 Fuel Consumption of 100km

Through the organization and analysis of the couple traffic simulation result data, the fuel consumption and fuel saving rate under different average PnG speeds and varying speed fluctuation amplitudes are presented in Fig. 3. For convenience of description, three speed ranges are defined: low-speed range (30–60 km/h), mid-speed range (60–90 km/h), and high-speed range (90–120 km/h). The detailed analysis is as follows:

As shown in Fig. 3(a), for vehicles in which the engine remains at idle during the G-phase, under the PnG mode, it is observed that when the vehicle speed varies between 30 km/h and 120 km/h, the fuel consumption curve per 100 km consistently exhibits a concave shape, regardless of the speed fluctuation amplitude. Specifically, the fuel consumption per 100 km is relatively high in the low-speed and high-speed ranges, while it is comparatively lower in the mid-speed range. The fuel consumption trends under different speed fluctuation amplitudes show consistency. In addition, the trend of the curves indicates that as the amplitude of speed fluctuations increases under the PnG mode, the fuel consumption per 100 km of the vehicle decreases at the same average PnG speed. However, as the amplitude of speed fluctuations continues to increase, the rate of decrease in fuel consumption gradually slows down. It can be observed that when the speed fluctuation amplitude reaches approximately 10 km/h, the fuel consumption curves per 100 km begin to converge.

If the fuel consumption during the G-phase with the engine at idle under the PnG mode is excluded, the fuel consumption per 100 km of the vehicle at different average PnG speeds and speed fluctuation amplitudes can be obtained, as shown in Fig. 3(b). From the trend of the curves in the figure, it is clearly observed that the vehicle's fuel consumption per 100 km increases steadily with the rise in average PnG speed across low, medium, and high-speed ranges. Unlike the case that includes idle fuel consumption, no pronounced dip in the fuel consumption curve is observed, indicating that the fuel consumption during idle significantly influences the overall fuel consumption at low and moderate speeds. Moreover, the impact of speed fluctuation amplitude on the overall fuel consumption per 100 km is not significant, with noticeable differences only occurring in the high-speed range, particularly when the speed fluctuation amplitude is small (e.g., 2 km/h and 4 km/h). This finding further highlights the varying effects of idle fuel consumption on the vehicle's fuel efficiency across different speed ranges.

Figure 3(c) and 3(d) respectively present the vehicle fuel consumption per 100 km under different PnG speed fluctuation amplitudes and average speed conditions, both with and without idle fuel consumption. By analyzing the data in Fig. 8, the following trends can be observed:

With the increase in PnG speed fluctuation amplitude, the vehicle fuel consumption per 100 km generally exhibits a pattern of initial reduction followed by stabilization at all average speeds. Specifically, when idle fuel consumption is included, the impact of speed fluctuation amplitude on fuel consumption becomes negligible for low and medium-speed conditions once the fluctuation amplitude exceeds 12 km/h. In contrast, when idle fuel consumption is excluded, the influence of speed fluctuation amplitude on fuel consumption is further diminished, with no significant impact on fuel consumption

across low, medium, and high-speed ranges when the fluctuation amplitude reaches or exceeds 10 km/h.

Further comparative analysis of the curves in the two figures reveals that the curve including idle fuel consumption is denser than the one excluding it. This indicates that idle fuel consumption contributes significantly to fuel consumption per 100 km in the low and medium-speed ranges, while its impact is relatively smaller at high speeds. The primary reason for this phenomenon is that idle fuel consumption is relatively stable and less affected by fluctuations in vehicle speed. As vehicle speed increases, overall fuel consumption rises, thereby reducing the proportion of idle fuel consumption in total fuel consumption. Therefore, it can be concluded that the impact of idle fuel consumption on vehicle fuel economy exhibits significant variation across different speed ranges.

3.2 Fuel Saving Rate

The fuel saving rate curve is derived by comparing the fuel consumption per 100 km under the PnG mode with that under the corresponding constant-speed driving conditions. Figure 3(e)–3(h) illustrate the fuel saving rate under different average speeds in the PnG mode and various speed fluctuation amplitudes, respectively. The following trends can be observed from the figures:

- 1) As the average speed in the PnG mode increases, the fuel saving rate exhibits a significant decreasing trend, regardless of whether idle fuel consumption is included in the gliding phase (see Fig. 3(e) and 3(f)). Under both conditions, with and without idle fuel consumption, the fuel saving rate decreases to a minimum of -47.7% and -30.0%, respectively. This indicates that at higher speeds, the energy-saving benefits of the PnG mode are less pronounced compared to lower speeds, and beyond a certain threshold, fuel consumption may even increase.
- 2) In speed ranges below an average of 60 km/h, the fluctuation in fuel saving rate is primarily influenced by the gear selection. Some average speeds in the PnG mode fall within the shift point range, indicating that, especially in the low-to-mid speed range, fuel saving rate is highly sensitive to gear selection. Above 60 km/h, the two curves (with and without idle fuel consumption) exhibit similar trends (see Fig. 3(e) and 3(f)), suggesting that in this speed range, the impact of gear configuration on fuel saving rate diminishes.
- 3) When the average speed reaches 100 km/h, a peak appears in the fuel saving rate curve (see Fig. 3(e) and 3(f)). In the high-speed range, the gear has already shifted to the highest gear and remains unchanged. This phenomenon is likely related to the engine operating within its optimal efficiency range at this speed.
- 4) As the amplitude of speed fluctuations increases, the energy-saving efficiency shows an upward trend (see Fig. 3(g) and 3(h)). However, in the high-speed range, smaller speed fluctuations lead to additional fuel consumption, which is higher than that of steady-speed driving. This is likely due to the greater driving resistance at high speeds, where small fluctuations in speed are often insufficient to bring the engine into its optimal efficiency operating range. Additionally, this may also be related to the fact that vehicles with discrete gear ratios are not designed with appropriately matched gears for the high-speed range.

5) The fuel efficiency begins to show negative values when the average speed and the range of speed fluctuations exceed certain thresholds (as shown in Fig. 3(e)-3(h)). This indicates that the fuel consumption of vehicles operating in the PnG mode exceeds that of the corresponding steady-speed condition, thereby no longer resulting in energy-saving.

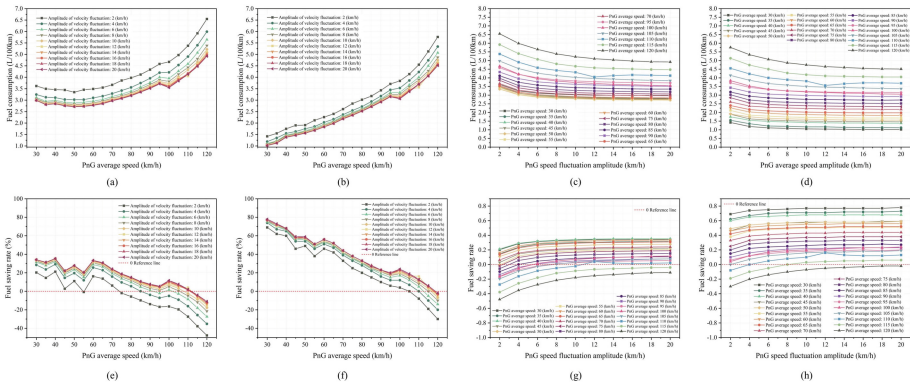


Fig. 3. Fuel saving rate at different PnG speed fluctuation magnitudes

In summary, if the vehicle can achieve zero fuel consumption during the G-phase, the overall energy-saving performance of the PnG mode in the low-to-medium speed range will be significantly improved. However, as the vehicle speed increases, the energy-saving effect gradually diminishes compared to steady-speed operation. Therefore, to fully exploit the energy-saving potential of the PnG mode, it is essential to implement a strategy that disconnects the powertrain from the wheels during the G-phase and ensures that the engine stops consuming fuel during this period. Furthermore, with regard to speed fluctuation amplitude, it is not always beneficial for it to be as large as possible. From the perspectives of actual energy-saving, ride comfort, and safety, the optimal balance should be sought within the constraints of these factors.

4 Conclusion

This study established a coupled traffic simulation integrating CarSim and MATLAB/Simulink. Using this approach, simulations and analyses were conducted to evaluate vehicle fuel consumption under various average speeds and speed fluctuation ranges in the PnG mode. The simulation results indicate that, compared to CS mode, the PnG mode achieves a significant energy-saving effect at medium to low speeds. However, as the average PnG speed increases, the energy-saving benefit gradually diminishes. To maximize the energy-saving effect of the PnG mode, the key lies in minimizing the idle fuel consumption during the G-phase. If the vehicle can achieve zero fuel consumption during the G-phase, such as by disengaging the powertrain from the wheels, the overall energy-saving efficiency of the PnG mode can be significantly improved.

References

1. Shieh, S.Y., Ersal, T., Peng, H.: Pulse-and-glide operations for hybrid electric vehicles in the car-following scenario. *IEEE Trans. Veh. Technol.* **72**(8), 9922–9937 (2023)
2. Cao, W., Yuno, T., Kawabe, T.: Pulse and glide strategy analysis based on engine operating point during pulse mode. *Eur. J. Control.* **65**, 100629 (2022)
3. Li, H., Wang, L., Yang, M., et al.: Collaborative effects of vehicle speed and illumination gradient at highway intersection exits on drivers' stress response capacity. *Accid. Anal. Prev.* **209**, 107829 (2025)
4. Li, H., Wang, L., Bie, Y.: Dynamic illumination method for rural highway intersections with traffic flow changes. *Transp. Res. Rec. J. Transp. Res. Board* **2678**(7), 977–991 (2023)
5. Kim, J., Ahn, C.: Real-time speed trajectory planning for minimum fuel consumption of a ground vehicle. *IEEE Trans. Intell. Transp. Syst.* **21**(6), 2324–2338 (2020)
6. Fu, C., Lu, Z., Ding, N., et al.: Distance headway-based safety evaluation of emerging mixed traffic flow under snowy weather. *Physica A* **642**, 129792 (2024)
7. Lu, Z., Ding, N., Gao, J., et al.: Safety benefits evaluation of mixed traffic flow with connected and automated vehicles under snowy conditions. In: 2023 7th International Conference on Transportation Information and Safety (ICTIS), pp. 380–385 (2023)
8. Evans, L.: Driver behavior effects on fuel consumption in urban driving. *Hum. Factors J. Factors Ergon. Soc.* **21**(4), 389–398 (1979)
9. Mierlo, J.V., Maggetto, G., Van de Burg, E., et al.: Driving style and traffic measures-influence on vehicle emissions and fuel consumption. *Proc. Inst. Mech. Eng. Part D J. Autom. Eng.* **218**(1), 43–50 (2004)
10. Li, S.E., Hu, X., Li, K., et al.: Mechanism of vehicular periodic operation for optimal fuel economy in free-driving scenarios. *IET Intel. Transport Syst.* **9**(3), 306–313 (2015)
11. Li, S.E., Peng, H.: Strategies to minimize fuel consumption of passenger cars during car-following scenarios. *Proc. Inst. Mech. Eng. Part D J. Automobile Eng.* **226**(3), 419–429 (2012)
12. Cao, W., Kawabe, T., Yuno, T.: Analysis of PnG strategy focusing on the effect of the duration of pulse mode. *IFAC-PapersOnLine* **54**(10), 424–429 (2021)
13. Xu, S., Li, S.E., Zhang, X., et al.: Fuel-optimal cruising strategy for road vehicles with step-gear mechanical transmission. *IEEE Trans. Intell. Transp. Syst.* **16**(6), 3496–3507 (2015)
14. Sohn, C., Andert, J., Jolovic, D.: An analysis of the tradeoff between fuel consumption and ride comfort for the pulse and glide driving strategy. *IEEE Trans. Veh. Technol.* **69**(7), 7223–7233 (2020)
15. Kim, T., Lee, W., Park, D., et al.: Energy-saving strategy for speed cruise control using pulse and glide driving. *Int. J. Precis. Eng. Manuf.-Green Technol.* **10**, 1553–1564 (2023)
16. Ding, N., Lu, Z., Jiao, N., et al.: Quantifying effects of reverse linear perspective as a visual cue on vehicle and platoon crash risk variations in car-following using path analysis. *Accid. Anal. Prev.* **159**, 106215 (2021)
17. Lu, Z., Ding, N., Lu, L., et al.: Optimizing signal timing of the arterial-branch intersection: a fuzzy control and nonlinear programming approach. *Asian J. Control* **24**(6), 2952–2968 (2022)



Research on High-Confidence Simulation Testing Methods for AI-Based End-to-End Autonomous Driving Vehicles

ZhangJie Meng^{1,2} , Deng Pan^{1,2}, Qiang Zhang^{1,2}, Bin Lu¹, Jinying Zhou³, Pan Xiao¹, Bing He¹, and YiChao Tang¹

¹ China Automotive Engineering Research Institute Co., Intelligent Connected Technology of CAERI Co., Ltd., Chongqing, China
Jaden@caeri.com.cn

² State Key Laboratory of Intelligent Vehicle Safety Technology, Chongqing, China

³ China Automotive Engineering Research Institute Co. Ltd., Chongqing, China

Abstract. In the current development of intelligent driving technology, high-level autonomous driving systems are undergoing significant transformations, transitioning from traditional rule-based approaches to AI-driven end-to-end methods. This paper presents a comprehensive study on high-confidence simulation testing methods for AI-based end-to-end autonomous vehicles, focusing on innovative methodologies that enhance the accuracy and robustness of testing. The study emphasizes three core aspects: the use of 3D Gaussian-based synthetic data to generate high-fidelity test scenarios that improve model performance in complex and rare long-tail driving situations; advanced physical-level sensor modeling that simulates sensor behavior under various environmental and weather conditions, including rain, fog, and high-glare sunlight; and the development of an axle-coupled Vehicle-in-the-Loop (ViL) system by China Automotive Engineering Research Institute (CAERI) to precisely simulate longitudinal and lateral road loads without the need for vehicle steering modifications. The proposed methods address key challenges in replicating real-world driving conditions in a controlled environment, ensuring reliable performance validation for autonomous systems. The results demonstrate that these comprehensive simulation approaches significantly enhance the reliability, consistency, and validity of test outcomes, supporting more robust development and deployment of intelligent driving technology. This research introduces innovative practices that lay a strong foundation for the future development of high-confidence simulation systems.

Keywords: Autonomous Driving · High-Confidence Simulation · End-to-End · 3D Gaussian Model · Sensor Modeling · Vehicle-in-the-Loop

Supported by Automated Test and Evaluation Technology and Software Development for Intelligent Vehicles CSTB2023TIAD-STX0029.

© Beijing Paiké Culture Commu. Co., Ltd. 2025

J. Liu et al. (Eds.): AIAT 2024, LNEE 1393, pp. 444–453, 2025.

https://doi.org/10.1007/978-981-96-3973-1_47

1 Introduction

Amid the rapid development of intelligent driving technology, high-level autonomous driving systems are undergoing significant transformations. These systems have evolved from the traditional rule-based approach to an AI-driven, end-to-end (E2E) model. This shift is fueled by the rapid advancements in AI technology, particularly deep learning and large-scale models that have enhanced computational power and data processing capabilities. The rule-based approach, which relies heavily on engineers manually coding extensive sets of rules to ensure the vehicle's safety and stability across various scenarios, proves insufficient when faced with complex and dynamic real-world traffic conditions [1]. It falls short of meeting the increasingly complex and diverse demands of driving scenarios.

In contrast, the E2E approach, powered by large-scale AI models, learns driving logic and decision-making directly from vast amounts of data, demonstrating superior generalization and adaptability. By training on massive driving datasets, these models can autonomously extract critical driving features and respond in real time under complex conditions [9]. This approach not only enhances the system's robustness but also significantly simplifies the engineering design process, allowing high-level autonomous driving systems to perform more effectively in diverse and challenging scenarios.

One of the key challenges in transitioning to an E2E model lies in addressing long-tail scenarios. Long-tail scenarios refer to rare yet safety-critical situations encountered during driving, such as complex intersections or driving in extreme weather conditions [12]. The system's ability to handle these scenarios directly impacts overall driving safety and reliability. To ensure that E2E autonomous driving systems can operate safely and efficiently under all circumstances, high-confidence simulation testing systems are indispensable, and Vehicle-in-the-Loop (ViL) simulation technology plays a crucial role in this context.

ViL simulation is a method that places a real vehicle within a virtual testing environment for comprehensive evaluation. Unlike purely virtual simulations, ViL testing integrates simulation scenarios directly onto the physical vehicle, enabling the system to process real-time data from a simulated environment under near-realistic driving conditions. This method effectively bridges the gap between virtual and physical testing, allowing the system to be tested in complex and dynamic scenarios with greater precision.

The primary advantages of ViL simulation lie in its high fidelity and reliability, making it particularly suited for validating system performance in extreme and high-risk scenarios. Through ViL simulation, the entire perception, decision-making, and execution chain of an autonomous system can be tested, providing precise data support for various long-tail and complex scenarios that might be encountered in real-world driving. This simulation approach allows repeated testing of complex situations without risking human or vehicle safety, offering high-confidence test results that facilitate comprehensive system optimization. This, in turn, accelerates the development of E2E autonomous driving technology and enhances its reliability and safety in real-world driving conditions.

In terms of high-confidence analysis, we draw the following conclusions:

1. For AI-based E2E autonomous driving systems, we innovatively propose the use of 3D Gaussian model-based synthetic data to generate high-fidelity test scenarios, supporting precise validation in complex driving environments.
2. For autonomous driving systems facing complex environmental conditions, we employ physical-level sensor modeling to simulate sensor behavior in various environments and weather conditions, improving the realism and confidence of simulation tests.
3. For testing complex vehicle dynamics, we propose a high-confidence dynamics modeling approach (e.g., axle-coupled ViL testing) to accurately simulate vehicle dynamics and road load interactions, ensuring that test results are more aligned with real-world driving conditions, thus enhancing overall system test confidence.

These methods provide a comprehensive analysis and theoretical foundation for high-confidence ViL testing, ensuring the stability and safety of E2E autonomous driving systems across a wide range of complex scenarios.

2 Related Work

In the current field of simulation testing for intelligent driving systems, researchers are continually exploring new methods to enhance the realism and confidence of these tests. With the ongoing evolution of intelligent driving technology, traditional commercial rendering engines no longer meet the demand for generating high-fidelity scenes. Recent research has increasingly shifted toward the use of synthetic data to create high-quality simulation scenarios for end-to-end intelligent driving testing. Among these, the use of 3D Gaussian models for data generation has emerged as a cutting-edge technique. By employing 3D Gaussian distributions or voxel-based models, these methods are capable of rendering realistic driving scenarios that include both static and dynamic object details, especially in rare and complex long-tail scenarios. For instance, the Decoupled Hybrid Gaussian Projection (DHGS) method models road and non-road elements separately and utilizes signed distance fields (SDF) to achieve geometrically precise road surfaces. This approach ensures high fidelity while maintaining multi-view consistency, enhancing the representation of scenes from complex viewpoints. Another method, the Visual-Only Dynamic Gaussian (VDG), independently optimizes dynamic and static elements in scene simulations, allowing for the modeling of time-dependent behaviors such as vehicle and pedestrian trajectories. The VDG model leverages self-supervised learning techniques and monocular depth maps to improve the realism and detail of dynamic scenes. These synthetic data generation techniques significantly increase the diversity and realism of training and testing data, greatly enhancing the robustness and adaptability of intelligent driving systems in complex scenarios.

To further improve the confidence level of perception system testing, researchers have gradually turned to physical-level sensor modeling. This approach focuses on modeling the original physical data of sensors, such as optical characteristics, radar reflections, and LiDAR point cloud responses [7]. By thoroughly modeling sensor behavior under different environmental conditions [18], it is possible to simulate the performance of real sensors in complex environments. In recent years, studies have combined physical modeling with deep learning and multi-sensor fusion to increase the system's stability

when facing uncertainties in single-sensor data. For example, some research has applied deep learning algorithms to optimize the fusion of radar and LiDAR data, thus improving system positioning and perception capabilities in dynamic and complex traffic environments. Additionally, sensor modeling methods capture external noise characteristics [4], transmission delays, and other factors [10], further enhancing the reliability and realism of simulation testing.

Vehicle-in-the-Loop (ViL) testing is an essential method for verifying the performance and reliability of intelligent driving systems [2]. ViL integrates dynamometers and high-fidelity simulation environments to simulate road loads accurately, providing high-confidence testing for vehicles under various driving conditions. Research indicates that ViL testing can capture the dynamic characteristics of vehicles and their behavior in complex traffic scenarios. For instance, Wu et al. proposed a ViL architecture integrated with dynamometers to test the energy efficiency of hybrid and electric vehicles in simulated traffic congestion scenarios, demonstrating significant potential for energy optimization [5]. Furthermore, the ViL method can dynamically adjust the dynamometer's load to simulate various driving conditions, such as highways, urban roads, and emergency braking situations, thus providing solid data support for system validation under extreme scenarios [6]. The advantage of ViL testing lies in its high fidelity and the accurate validation of system responses in real-world environments [7]. This simulation method greatly enhances the stability and reliability of testing across diverse environments [23, 24], providing robust assurance for high-confidence verification of intelligent driving systems.

3 Method

China Automotive Engineering Research Institute (CAERI) has proposed a high-confidence simulation testing solution to evaluate AI-based end-to-end autonomous vehicles. This paper provides a detailed description of how to implement high-confidence simulation testing methods for end-to-end autonomous driving. The proposed approach primarily includes the generation of synthetic data using 3D Gaussian models, physical-level sensor modeling, and the integration of Vehicle-in-the-Loop (ViL) simulation testing to validate the overall performance and stability of the system (Fig. 1).

3.1 Synthetic Data Generation Based on 3D Gaussian Models

To enhance the realism and diversity of scenarios in end-to-end autonomous driving simulation testing, we adopt a 3D Gaussian model-based synthetic data generation method. This approach simulates various objects and dynamic scenes in the environment, producing high-fidelity simulation scene data. The core of the 3D Gaussian model lies in representing scene elements as Gaussian distribution ellipsoids, enabling quick computation and efficient rasterized rendering during the rendering and training process. Initialization: Sparse point clouds and camera poses are generated using an incremental Structure-from-Motion (Sfm) algorithm, such as COLMAP, to provide initial conditions for the 3D Gaussian distributions. Gaussian Modeling: Each object in the scene is modeled using 3D Gaussian distributions, creating ellipsoidal Gaussian kernels that

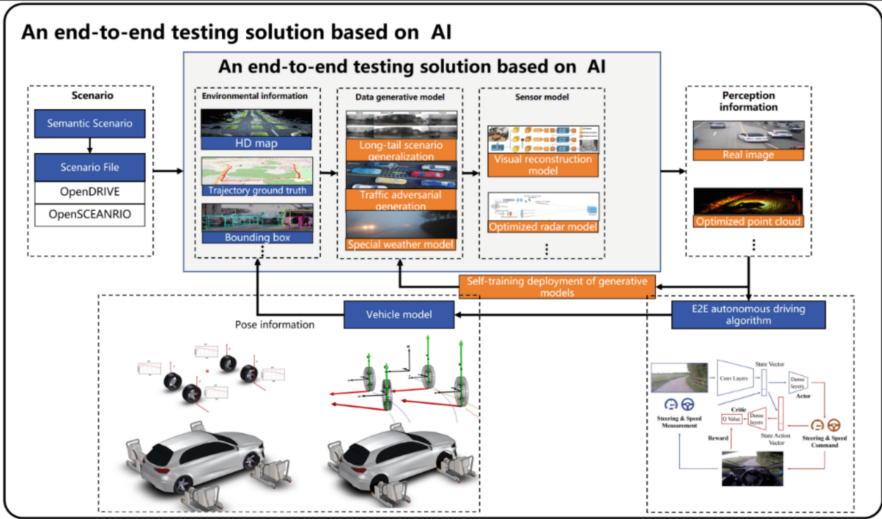


Fig. 1. End-to-End testing solution based on AI

represent the geometric structure and surface properties of objects. Rendering and Synthesis: Multi-view scene generation is efficiently achieved using a differentiable renderer, ensuring consistency and high fidelity of the scene across different perspectives (Fig. 2).

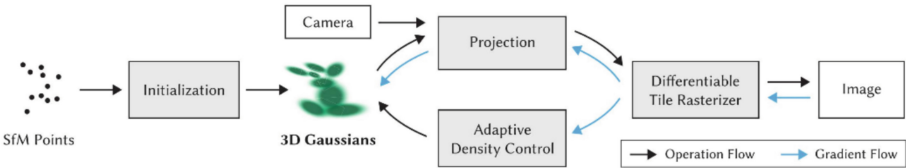


Fig. 2. Process of 3D gaussians

3.2 Physical Sensor Modeling

In the development of high-confidence simulation testing for end-to-end autonomous driving systems, physical-level sensor modeling plays a crucial role. This modeling aims to accurately replicate the processes of signal emission, reflection, and reception to simulate the real-world performance of sensors such as cameras, LiDAR, and radar systems. This section details the methods used to construct high-fidelity sensor models and the factors involved in simulating real-world conditions. We illustrate this with examples of three key types of sensors: cameras, LiDAR, and millimeter-wave radar. **Camera Simulation:** Camera modeling involves replicating the imaging characteristics of real cameras, including lens properties, image distortion, and various optical effects. Creating a realistic camera model requires. **Lens Imaging Principles**:** Simulating the

path of light through convex lenses to ensure that parallel rays focused through the lens converge at the focal point and rays passing through the lens center remain unbent. Image Distortion and Aberration: Real cameras often exhibit image distortions and chromatic aberrations, which must be accurately simulated in virtual environments. This involves adding post-processing steps to the perfectly rendered images in simulations to replicate real camera effects such as glare, exposure adjustments, and white balance modifications. This level of detail ensures that the simulated camera closely mimics the behavior and characteristics of its real-world counterpart, enhancing the reliability of the simulation tests.

Principle of Light Propagation and Imaging

$$\frac{1}{f} = \frac{1}{u} + \frac{1}{v}$$

where f is the focal length of the lens, u is the object distance, and v is the image distance.

The mathematical expression for image distortion:

$$\gamma_{\text{distorted}} = \gamma(1 + k_1\gamma^2 + k_1\gamma^4 + k_1\gamma^6)$$

where: $\gamma_{\text{distorted}}$ is the radial distance of a distorted image point from the center of the image. γ is the original radial distance of the point, k are distortion coefficients that represent the degree of distortion.

LiDAR simulation involves the emission, propagation, and reception of laser pulses to capture distance and reflectance information. The process of physical-level LiDAR modeling includes the following key steps:

Laser Emission and Reflection: This step simulates the behavior of laser pulses as they interact with various surfaces. It involves calculating the angles of reflection and refraction and modeling the attenuation of signal strength over different materials and distances. This allows for accurate representation of how laser pulses react when encountering different types of surfaces, whether smooth, rough, or absorbing. **Ray Tracing Techniques:** Ray tracing is used to simulate the path of laser pulses as they travel through the environment and interact with objects. This method ensures that both single and multiple reflections (multi-bounce effects) are captured, which enhances the fidelity of point cloud data. By tracking how laser pulses reflect off surfaces and change direction, ray tracing contributes to a more realistic depiction of the environment. **Echo Data and Noise Simulation:** The LiDAR model can simulate single and multiple echo returns, replicating how real-world LiDAR sensors detect objects at various distances and with different levels of reflectivity. To represent the imperfections of real sensors, random noise and data loss are added to the simulation. This can include noise from atmospheric conditions, signal interference, or hardware limitations, which introduces variability and challenges similar to those encountered in real-world scenarios. These detailed simulations are crucial for creating realistic test environments that mirror the performance of actual LiDAR sensors, enabling comprehensive testing and validation of autonomous vehicle perception systems.

3.3 Vehicle-in-the-Loop (ViL) Simulation Testing

This paper adopts Vehicle-in-the-Loop (ViL) simulation technology to validate the performance of intelligent driving systems under real driving conditions. ViL testing integrates a real vehicle into a virtual environment for comprehensive assessment, combined with a dynamometer to precisely simulate road loads, thereby verifying the system's dynamic characteristics and environmental responses (Fig. 3).



Fig. 3. Vehicle-in-the-Loop

Longitudinal Load Simulation: The axle-coupled ViL system achieves precise longitudinal load simulation, primarily through accurate modeling of the interaction between the tires and the road surface. By utilizing the Magic Formula, the system can accurately simulate the longitudinal force of the tires under different driving conditions. The Magic Formula is defined as:

$$F_x = D \sin(\text{Carctan}(B\kappa - E(B\kappa - \arctan(B\kappa))))$$

where F_x represents the longitudinal force, κ the slip ratio, and B, C, D, E are parameters related to tire characteristics. This formula is used to model the mechanical behavior of the tires at various slip ratios, allowing the system to replicate the load response of a vehicle under different acceleration and braking conditions accurately. The use of this model enables the axle-coupled system to apply and measure high-precision longitudinal loads on the vehicle, thereby achieving testing that closely mirrors real-world road conditions in a laboratory setting.

Lateral Load Simulation: For lateral load simulation, the system employs the Ackermann steering model to ensure precise reproduction of load distribution during vehicle turning. The Ackermann steering model describes the relationship between the steering angles of the inner and outer wheels during a turn, ensuring smooth cornering without causing skidding. The basic equation for this model is:

$$\tan(\delta_{\text{inner}}) = R - 2W/L, \tan(\delta_{\text{outer}}) = R + 2W/L$$

where δ_{inner} , δ_{outer} are the steering angles of the inner and outer wheels, L is the wheel-base, R is the turning radius, and W is the track width of the vehicle. This formula ensures that the inner and outer wheels are adjusted to the correct steering angles according to the turning radius in the simulation, achieving realistic lateral load reproduction.

In the axle-coupled ViL system, integrating the Ackermann steering model enables precise lateral load application without affecting the normal steering operation of the vehicle. The system uses high-precision control algorithms to adjust the load distribution in real time for different steering angles, matching the lateral force characteristics observed in actual driving scenarios.

4 Conclusion

In this study, we presented a comprehensive approach to high-confidence simulation testing for AI-based end-to-end autonomous driving systems. The methods outlined include the use of 3D Gaussian model-based synthetic data generation, physical-level sensor modeling, and the integration of Vehicle-in-the-Loop (ViL) simulation technology. These techniques collectively address the key challenges of replicating realistic driving scenarios, ensuring robust sensor behavior under varied environmental conditions, and accurately simulating vehicle dynamics. The implementation of these methods has shown to improve the reliability, robustness, and accuracy of testing, providing essential support for the development of intelligent driving systems capable of handling complex and long-tail scenarios. By combining advanced data generation, detailed sensor modeling, and ViL testing, this research lays the foundation for future advancements in simulation methods that enhance the overall safety and performance of autonomous vehicles. Future work will focus on further refining these approaches and integrating additional real-world variables to enhance simulation fidelity and testing effectiveness.

References

1. Beglerovic, H., Metzner, S., Horn, M.: Challenges for the validation and testing of automated driving functions. In: Proceedings of the AMAA 2017 Conference (2017)
2. Hwang, T., et al.: Development of HILS systems for active brake control systems. In: 2006 SICE-ICASE International Joint Conference (2006)
3. Fodor, D., Enisz, K.: Vehicle dynamics based ABS ECU verification on real-time hardware-in-the-loop simulator. In: 2014 16th International Power Electronics and Motion Control Conference and Exposition (2014)
4. Moore, J.D., Molen, G.M.: SIL, HIL, and vehicle fuel economy analysis of a pre-transmission parallel PHEV. In: 2013 World Electric Vehicle Symposium and Exhibition (EVS27) (2013)
5. Hendriks, F., Tideman, M., Pelders, R., Bours, R., Liu, X.: Development tools for active safety systems: Prescan and VeHIL. In: Proceedings of 2010 IEEE International Conference on Vehicular Electronics and Safety (2010)
6. Schyr, C., Brissardt, A.: Drivingcube – a novel concept for validation of powertrain and steering systems with automated driving. In: Proceedings of the 13th International Symposium on Advanced Vehicle Control (AVEC2016) (2016)
7. Fayazi, S.A., Vahidi, A.: Vehicle-in-the-loop (ViL) verification of a smart city intersection control scheme for autonomous vehicles. In: 2017 IEEE Conference on Control Technology and Applications (CCTA) (2017)

8. Wang, F.-Y.: Parallel control: a method for data-driven and computational control. *J. Internet Technol.* **39**(04) (2017)
9. Wang, F.-Y., Zheng, N., Cao, D., Martinez, C., Li, L., Liu, T.: Parallel driving in CPSS: a unified approach for transport automation and vehicle intelligence. *IEEE/CAA J. Automatica Sinica* **4**(04) (2017)
10. Su, Z., Liu, F., Yang, R., Wang, F.-Y.: Architecture of marine environmental parallel monitoring system based on the Lanhai information network. *Acta Automatica Sinica* **46**(08) (2020)
11. Liu, T., Tian, B., Ai, Y., Li, L., Wang, F.-Y.: Parallel reinforcement learning: a framework and case study. *IEEE/CAA J. Automatica Sinica* **5**(04) (2018)
12. Duan, Y., Lv, Y., Zhang, J., Zhao, X., Wang, F.-Y.: Deep learning for control: the state of the art and prospects. *IEEE/CAA J. Automatica Sinica* **42**(05) (2016)
13. Wang, F.-Y.: Parallel system methods for management and control of complex systems. *Control Decis.* **19**(05) (2004)
14. Xiong, G., et al.: Overview of research and application progress of parallel control and management. In: *Proceedings of the 2nd China Aerospace Security Conference* (2017)
15. Wang, F.-Y., Wei, Q.: Intelligent control: from learning control to parallel control. *Control Theory Appl.* **35**(07) (2018)
16. Duan, Y., Lv, Y., Zhang, J., Wang, F.-Y.: Research status and prospects of deep learning in the control field. *Acta Automatica Sinica* **42**(05) (2016)
17. Wang, F.-Y., Zhang, J., Wei, Q., Zheng, X., Li, L.: PDP: parallel dynamic programming. *IEEE/CAA J. Automatica Sinica* **4**(01) (2017)
18. Wei, Q., Li, H., Wang, F.-Y.: Parallel control for continuous-time linear systems: a case study. *IEEE/CAA J. Automatica Sinica* **7**(4) (2020)
19. Lu, J., Wei, Q., Liu, Y., Zhou, T., Wang, F.-Y.: Event-triggered optimal parallel tracking control for discrete-time nonlinear systems. *IEEE Trans. Syst. Man Cybern. Syst.* (2021)
20. Wei, Q., Wang, L., Lu, J., Wang, F.-Y.: Discrete-time self-learning parallel control. *IEEE Trans. Syst. Man Cybern. Syst.* (2020)
21. Wang, F.-Y.: Parallel control and digital twins: control theory revisited and reshaped. *Chin. J. Intell. Sci. Technol.* **2**(3) (2020)
22. Lu, J., Wei, Q., Wang, F.-Y.: Parallel control for optimal tracking via adaptive dynamic programming. *IEEE/CAA J. Automatica Sinica* **7**(6) (2020)
23. Siegl, S., et al.: Vehicle-in-the-loop at testbeds for ADAS/AD validation. *ATZelectronics Worldwide* **16**(7) (2021)
24. Pacejka, H.B., Bakker, E.: The magic formula tyre model. *Veh. Syst. Dyn.* **21** (1992)
25. Meng, Z., Zhao, S., Chen, H., Hu, M., Tang, Y., Song, Y.: The vehicle testing based on digital twins theory for autonomous vehicles. *IEEE J. Radio Freq. Identif.* **6**, 710–714 (2022). <https://doi.org/10.1109/JRFID.2022.3211565>
26. Zhou, H., Wang, R., Weng, D., Wang, R., Qiao, Y.: Adjust train schedules to minimize passenger waiting times during disruptions. *Smart Resilient Transp.* **6**(1), 2–20 (2024). <https://doi.org/10.1108/SRT-10-2023-0009>
27. Zengqing, W., Xie, Z.Y., Yiling, J.: Location and segmentation of important railway signs based on improved segmentation. *Smart Resilient Transp.* **6**(1), 21–31 (2024). <https://doi.org/10.1108/SRT-10-2023-0010>
28. Oliveira, R.C.N., Zhang, Z.: A conceptual non-stop operation strategy for urban rail transit: a case study of São Paulo's monorail line. *Smart Resilient Transp.* **6**(1), 32–49 (2024). <https://doi.org/10.1108/SRT-09-2023-0006>

29. Zhang, K., Huang, A.: Portraying passenger travel patterns for Beijing public transit system with user profiling method. *Smart Resilient Transp.* **6**(1), 50–71 (2024). <https://doi.org/10.1108/SRT-11-2023-0014>
30. Kang, D., Kirkpatrick, S.W., Zhang, Z., Liu, X., Bian, Z.: Freight train derailment severity prediction: a physics-informed one-dimensional model. *Smart Resilient Transp.* **6**(1), 72–92 (2024). <https://doi.org/10.1108/SRT-10-2023-0008>



A Star Structure Algorithm Based on GNSS/CORS and Experimental Analysis

Zhu Chao^(✉)

Eastern Navigation Service Center of Maritime Safety Administration of China,
Shanghai 200090, China
541762219@qq.com

Abstract. The GPS network differential technology, which based on CORS (Continuous Operation Reference Stations), has a wide coverage, high reliability. And now, it has become the representative of the most high-precision level in real-time kinematic positioning. It decreases correlative errors, such as atmosphere delay, satellite orbit errors, through integrating data from continuous operation stations. Then generates the RTCM correction, and broadcasts the information by wireless communication technology. As to VRS (Virtual Reference Station)/RTK technology, the structure of CORS fiducial station networks was analyzed. Traditionally, based on the principles of Delaunay triangulation, it had realized automatic networking between stations, and dynamic updating such as point inserting and deleting, and also the problem of selecting the best reference station in mobile positioning was solved. But, this sort of triangular network also has some shortcomings. Such as, it only adopts one baseline estimation, and redundant observation information has not yet been fully tapped, this would go against large-scale reference station network initialization. Also, it has less independent baseline information, this would affect the interpolation precision of the mathematic model, especially in some areas where have large elevation difference. The third is that, in some areas where outside the basic solution unit, the correction accuracy would have fallen dramatically with the increase of the distance. Aiming at the shortcomings of VRS positioning algorithm based on traditional triangulation network, a new VRS localization algorithm based on star network is proposed. This paper establishes a method of GNSS/CORS baseline solution based on star structure, and presents a method of ambiguity solution, and further research about the ionospheric and tropospheric correction algorithms are analyzed respectively, and two experiments are carried out to verify and analyze the calculation and correction of network ambiguity in the new network and traditional triangulation network. Based on the self-development software, experiments were implemented through actual GNSS data. These experiments show that the proposed algorithm model based on star structure can significantly accelerate the time of ambiguity fixed, up to 45%, and has higher accuracy and reliability in generating corrections. These research results have great theoretical significance for the wide application of GNSS/CORS.

Keywords: CORS · Network differential · Algorithm · Ambiguity resolution · Atmospheric correction

1 Introduction

The network real-time differential positioning technology represented by GNSS/CORS is a comprehensive positioning system integrating Internet, wireless communication, computer network management and GNSS positioning technology [1, 2]. It makes the establishment of reference station network service system become the latest trend of the application and development of GNSS technology. The biggest feature of the virtual reference station technology is to achieve high-precision RTK positioning by virtualizing the original phase observations of the virtual reference station [3, 4]. Due to the limited accuracy of the pseudo-range code itself, relying solely on the virtual pseudo-range to achieve differential positioning is obviously unable to meet the accuracy requirements. Therefore, the generation of high-precision virtual carrier phase observations has become the key to the realization of the entire system algorithm [5–7]. In this paper, the algorithm model of VRS virtual phase observation value generation is deeply studied for medium-distance centimeter-level network RTK positioning. It is applied to the self-developed GNSS virtual reference station system software, and based on a continuous operation reference station test network. The experimental analysis was carried out to realize the centimeter-level RTK positioning service in the region.

2 Construction of Star Structure Algorithm Model

When positioning a single user, it is not necessary to combine the data of all base stations to process. Professor Lambert also believes that at least three base stations are needed to achieve network differential positioning [8, 9], and in most cases this is also an ideal number. Among the many two-dimensional planar network structures, the triangular network is the representative and widely used network structure [10, 11].

The so-called star network is centered on a base station in the network, and the remaining base stations are connected to form a baseline edge to form a radial star network structure, as shown in Fig. 1. In the star network structure, according to the spatial position relationship between the user mobile station and the central reference station, the corresponding error correction model is established. One of the reference stations is selected as the main reference station, and the remaining reference stations are used as auxiliary reference stations. This structure includes all the base stations and baselines in the network, the flow chart of Star structure network CORS/VRS positioning realization is shown in Fig. 2. With the increase of the number of base stations and the increase of distance, the correlation of system errors between baselines is greatly weakened, and the interpolation accuracy and reliability will be reduced by using the uncorrelated baseline data for model interpolation, which will affect the quality of corrections. At the same time, the processing of a large amount of data will undoubtedly increase the solution burden of the system control center. To solve this problem, we can use the method of dividing the base station network into several subnets, and select several base stations as the center of the star network, so that the mobile station users can select the nearest center and its nearby base stations to form a solution unit, which ensures the accuracy of error interpolation.

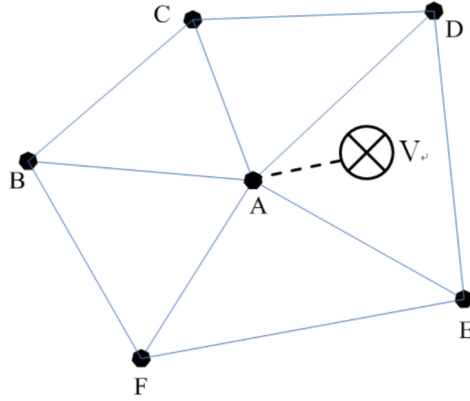


Fig. 1. A CORS/VRS star structure diagram.

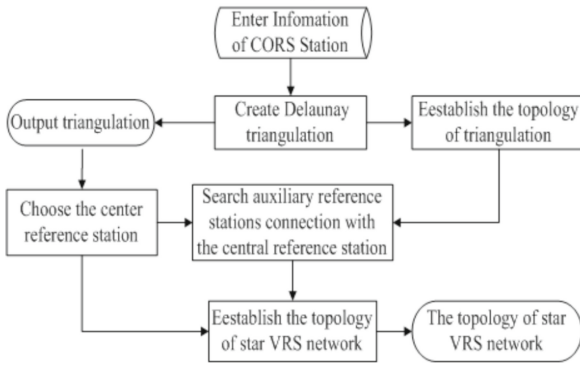


Fig. 2. Flow chart of Star structure network CORS/VRS positioning realization.

3 Ambiguity Resolution and Differential Correction Calculation

In the VRS system, with the increase of the baseline length, the spatial correlation of errors such as ionosphere and troposphere is weakened [12, 13]. It is difficult to fix the ambiguity only by using the LAMBDA algorithm due to the influence of a large number of residuals, and the speed is also difficult to guarantee [14–16]. In this paper, the wide-lane ambiguity is calculated by using the pseudo-range wide-lane, and then the L1 ambiguity floating-point solution is solved by using the ionosphere-free constraint combination. Finally, the LAMBDA algorithm is used to search and fix the L1 and L2 ambiguity.

Firstly, the L1 and L2 pseudo-range and carrier phase observations are combined to form a double-difference pseudo-range wide-lane combined observation:

$$\Delta \nabla NP = \Delta \nabla \varphi_1 - \Delta \nabla \varphi_2 - \frac{1}{\lambda_P} \frac{f_1 \Delta \nabla Q_1 + f_2 \Delta \nabla Q_2}{f_1 + f_2} + \varepsilon_P \quad (1)$$

where, $\Delta\nabla N_P$ is double difference wide lane ambiguity, $\Delta\nabla\varphi_1$, $\Delta\nabla\varphi_2$ are L1 and L2 double difference carrier phase observations, $\lambda_P = c/(f_1 - f_2)$ is the combined wavelength of wide lane, $\Delta\nabla Q_1$, $\Delta\nabla Q_2$ are L1, L2 band double difference pseudorange, f_1, f_2 are the frequencies of L1 and L2 bands, ε_P is combined noise,

using the characteristic that λ_P is a long wavelength, the double-difference wide-lane ambiguity can be quickly fixed.

At the same time, the free combination of L1 and L2 without ionosphere can effectively eliminate 2/3 of the ionospheric delay, which can greatly improve the speed and accuracy of ambiguity fixing.

$$\Delta\nabla N1 = \frac{1}{\lambda_n} \left\{ \lambda_P \left(\frac{f_1}{f_1 + f_2} \Delta\nabla\varphi_{L1} - \frac{f_2}{f_1 + f_2} \Delta\nabla\varphi_{L2} \right) - \Delta\nabla\rho - \Delta\nabla O - \Delta\nabla T \right\} - \frac{f_2}{f_1 - f_2} \Delta\nabla N_P \quad (2)$$

where, $\Delta\nabla N1$ is L1 band double difference ambiguity, $\lambda_n = c/(f_1 + f_2)$ is the wavelength of narrow lane, $\Delta\nabla\rho$ is the star distance of double difference station, which can be obtained from the known coordinates, $\Delta\nabla O$ is double difference orbit error, $\Delta\nabla T$ is double-difference tropospheric delay, in this way, the floating point solution can be obtained.

Finally, the LAMBDA algorithm is used to search and fix the L1 and L2 ambiguity. Due to the influence of ionosphere, tropospheric residuals and other errors, the fluctuation of floating-point solution is still relatively large. Therefore, in order to ensure the speed and reliability of the search, special attention should be paid to the construction of the weight matrix in the LAMBDA algorithm.

4 Generation of Virtual Phase Observations

The multipath effect and the influence of measurement noise are neglected according to the above mentioned, so that the double difference phase observation equation of the main reference station A and the virtual reference station R between the satellite pairs (m, n) can be written as:

$$\lambda \Delta\nabla\varphi_{AR}^{mn} = \Delta\nabla\rho_{AR}^{mn} + \lambda \Delta\nabla N_{AR}^{mn} + \Delta\nabla O_{AR}^{mn} + \Delta\nabla T_{AR}^{mn} - \Delta\nabla I_{AR}^{mn} = \lambda(\Delta\varphi_A^{mn} - \Delta\varphi_R^{mn}) \quad (3)$$

where, $\Delta\varphi_A^{mn}$ is the single difference of the phase observation value of the main reference station A between the satellite pairs (m, n) , that can be calculated directly from the carrier phase observation value. Since the coordinates of the reference station are accurately known, the coordinates of the virtual reference station can be obtained according to the single point positioning, so the double difference of the station-star distance between the main reference station A and the virtual reference station R can also be obtained. Double differential integer ambiguity $\Delta\nabla N_{AR}^{mn}$ has been solved in the above, $\Delta\nabla O_{AR}^{mn} + \Delta\nabla T_{AR}^{mn} - \Delta\nabla I_{AR}^{mn}$ has also been calculated.

Thus, the only unknown quantity in the above equation can be solved:

$$\Delta\varphi_R^{mn} = \Delta\varphi_A^{mn} - (\Delta\nabla\rho_{AR}^{mn} + \Delta\nabla T_{AR}^{mn} - \Delta\nabla I_{AR}^{mn})/\lambda - \Delta\nabla N_{AR}^{mn} \quad (4)$$

and $\Delta\varphi_R^{mn}$ is the virtual reference station phase single difference observation value.

Following Eq. (3), the double difference phase observation equation between the main reference station A and the user mobile station F in the satellite pair can be established:

$$\lambda \Delta \nabla \varphi_{AF}^{mn} = \Delta \nabla \rho_{AF}^{mn} + \lambda \Delta \nabla N_{AF}^{mn} + \Delta \nabla O_{AF}^{mn} + \Delta \nabla T_{AF}^{mn} - \Delta \nabla I_{AF}^{mn} = \lambda (\Delta \varphi_A^{mn} - \Delta \varphi_F^{mn}) \quad (5)$$

By subtracting (3) and (5), we can get:

$$\lambda (\Delta \varphi_R^{mn} - \Delta \varphi_F^{mn}) = (\Delta \rho_R^{mn} - \Delta \rho_F^{mn}) + \lambda (\Delta N_R^{mn} - \Delta N_F^{mn}) + (\Delta T_R^{mn} - \Delta T_F^{mn}) - (\Delta I_R^{mn} - \Delta I_F^{mn}) \quad (6)$$

Usually, the distance between the virtual reference station R and the user mobile station F is close, so it can be approximately considered that:

$$\Delta T_R^{mn} = \Delta T_F^{mn}, \Delta I_R^{mn} = \Delta I_F^{mn} \quad (7)$$

Then Eq. (6) can be written as:

$$\Delta \nabla \rho_{FR}^{mn} = \lambda (\Delta \nabla \varphi_{FR}^{mn} - \Delta \nabla N_{FR}^{mn}) \quad (8)$$

The above equation is the double difference observation equation between the virtual reference station R and the rover station F, so that the position of the user rover station can be calculated according to the conventional GNSS/CORS positioning method, and then the virtual reference station network differential positioning can be realized.

5 Experimental Analysis

The ultimate goal of establishing a VRS network differential system is to achieve high-precision positioning. Therefore, the important indicator of the evaluation of the carrier phase virtual observation algorithm model should be the point error after the ambiguity is fixed, that is, positioning accuracy. In this paper, the network differential positioning experiment is carried out based on the network platform of a CORS system as shown in Fig. 3, and the self-developed virtual reference station software. The reference station network includes six reference stations, which constitute a VRS network covering this area. The longest distance between the reference stations is 107 km, and the shortest distance is 56 km. The reference station and the data processing center are connected through the private line, forming a network as a whole. The data acquisition time is September 10, 2023, with a 5 s sampling interval and continuous observation for 4 h.

5.1 The Reliability and Stability of the Star Structure Algorithm

In order to verify the reliability and stability of the algorithm model, a single epoch solution is used as a benchmark to study the relative dispersion of other epoch solutions, that is, the internal coincidence accuracy of real-time positioning is obtained.

From Table 1 and Fig. 4, it can be seen that the difference between the results of repeated positioning at the same point is small, indicating that the system positioning is stable. At the same time, using the algorithm model proposed in this paper, the ambiguity fixing time is greatly improved in the multi-baseline mode, which is more than 45% higher than that in the single-baseline mode.

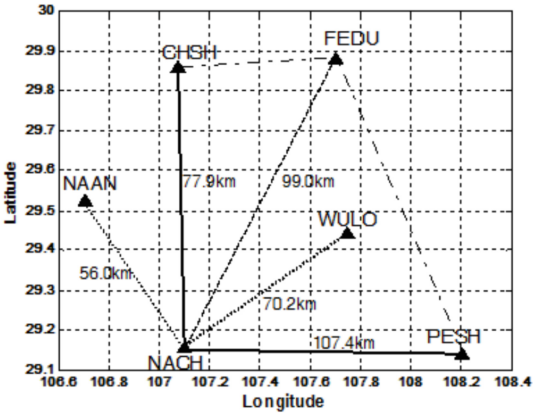


Fig. 3. A CORS/VRS network distribution map.

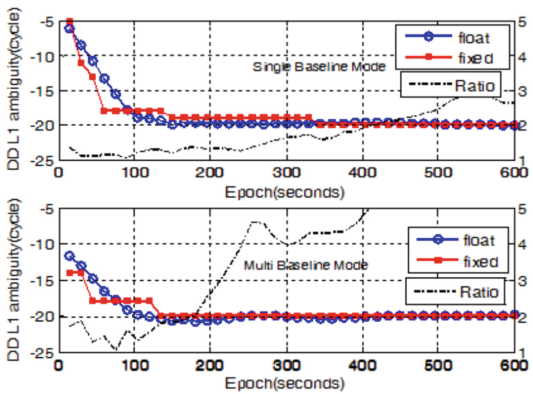


Fig. 4. Comparison of fixed time of ambiguity resolution.

Table 1. Ambiguity fixed time under different baseline modes/s.

Baseline	N-N	N-C	N-F	N-W	N-P
Single-line Mode	135	267	301	388	454
Multi-line Mode	123	132	147	169	187

5.2 Accuracy and Precision of the Star Structure Algorithm

In order to verify the correctness of the algorithm model and make an objective evaluation of the positioning accuracy, a high-level GNSS known control point in the network is selected as the verification point. Therefore, by comparing the differential positioning

results of the virtual reference station with the known coordinates, the external coincidence accuracy of the real-time positioning of the virtual reference station system can be reflected.

Table 2. Accuracy and precision of Atmospheric correction/m.

Statistic	PRN1		PRN2	
	LIM	SIM	LIM	SIM
Max	0.048	0.022	0.088	0.035
Mean	0.016	0.010	0.027	0.009
Std	0.018	0.011	0.034	0.012

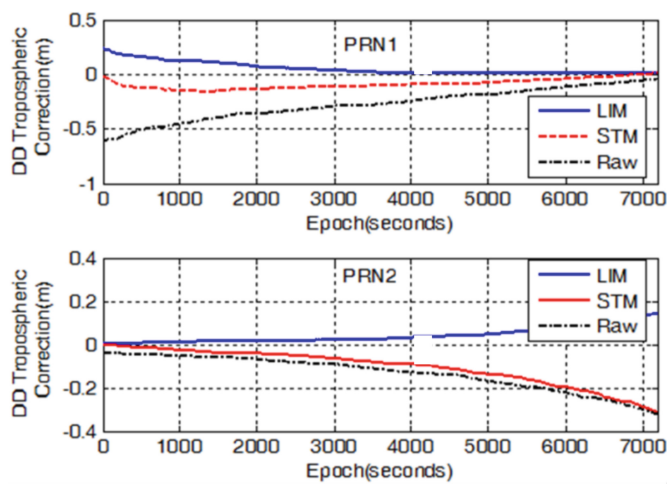


Fig. 5. Comparison of atmospheric delay bias between two satellites.

From Table 2 and Fig. 5, the virtual phase observations generated by the star structure algorithm model are used for positioning, it can be seen that, the SIM accuracy of the star structure algorithm is higher than 1cm, and the STM accuracy is higher than LIM. Thus, centimeter-level real-time positioning is realized.

6 Conclusion

In this paper, the algorithm model of virtual carrier phase observation and correction generation in differential positioning of virtual reference station network is proposed, and the detailed derivation is given. At the same time, in order to verify the correctness of the algorithm and the positioning accuracy of network difference, the experimental analysis is carried out based on a continuous operation reference station network combined with the self-developed GNSS/CORS system software. The results show that the

comprehensive error interpolation effect between reference stations is better, and the positioning accuracy in the horizontal direction and vertical direction are better than traditional triangulated network model.

References

1. Landau, L., Vollath, U., Chen, X.M.: Virtual reference station systems. *J. Glob. Position. Syst.* **1**(2), 137–143 (2002)
2. Hu, G.R., Khoo, H.S.: Development and assessment of GPS VRS for RTK positioning. *J. Geodesy* **77**, 292–302 (2003)
3. Lambrou, E., Kanellopoulos, N.: Check and calibration of a single GNSS receiver by using the VRS RTN positioning method. *Measurement* 221–225 (2018)
4. Gumus, K., Selbesoglu, M.O., Celik, C.T.: Accuracy investigation of height obtained from Classical and Network RTK with ANOVA test. *Measurement* 135–143 (2016)
5. Zhang, R.C., Gao, C.F., Wang, Z.B., et al.: Ambiguity resolution for long baseline in a network with BDS-3 quad-frequency ionosphere-weighted model. *Remote Sensing* **14**(7), 1654 (2022)
6. Wang, P.X., Liu, H., Yang, Z.X., et al.: Evaluation of network RTK positioning performance based on BDS-3 new signals system. *Remote Sensing* **14**(1), 2 (2021)
7. Zhang, Y.Z., Kubo, N., Chen, J.P., et al.: Contribution of QZSS with four satellites to multi-GNSS long baseline RTK. *J. Spat. Sci.* **65**(1), 41–60 (2020)
8. Cui Lili, X., Wenchao, S.H., et al.: The research on triplefrequency geometry-free and ionosphere-free cycle slip detection and repair method. *Sci. Technol. Eng.* **20**(3), 904–912 (2020)
9. Cui Lili, X., Wenchao, Z.Z., et al.: The detection and repair algorithm of triple-frequency cycle slip under strong ionospheric activity. *Sci. Technol. Eng.* **20**(36), 30–35 (2020)
10. Li, Z.Y.J.M.: Time-series InSAR ground deformation monitoring: atmospheric delay modeling and estimating. *Earth-Sci. Rev.* **192**, 258–284 (2019)
11. Shen, L., Hooper, A., Elliott, J.: A spatially varying scaling method for InSAR tropospheric corrections using a high-resolution weather model. *J. Geophys. Res. Solid Earth* (2019). <https://doi.org/10.1029/2018JB016189>
12. Jiang, W.: Challenges and opportunities of GNSS reference station network. *Acta Geodaetica et Cartographica Sinica* **46**(10), 1379–1388 (2017)
13. Yang, Y., Lu, M., Han, C.: Some Notes on Interoperability of GNSS. *Acta Geodaetica et Cartographica Sinica* 253–259 (2016)
14. Pan, S.G., Wang, Q., Ke, F.Y., et al.: Method for integer ambiguity resolution in GPS network RTK. *J. Southeast Univ. (English Edition)* **25**(4), 491–495 (2009)
15. Hu, G., Abbey, D.A., Castleden, N., Featherstone, W.E., et al.: An approach for instantaneous ambiguity resolution for medium to long-range multiple reference station networks. *GPS Solut.* **9**(1), 1–11 (2005)
16. Li, C., Lijie, Li, Y., et al.: Ionospheric disturbance delay correction method for reference station network over mid & low latitude region. *Sci. Surv. Mapp.* **34**(4), 27–29 (2009)



Research on Test Scenarios Coverage for the Behaviour Safety of Autonomous Driving Vehicles

Yu Tang^{1,2,3}, Yingzhi Xiong³(✉), Junfu Huang^{1,2}, Qiang Zhang^{1,2}, Jinying Zhou^{1,2},
Xiaona He², and Kan Yi³

¹ State Key Laboratory of Intelligent Vehicle Safety Technology, Chongqing 401122, China

² China Automotive Engineering Research Institute Co., Ltd., Chongqing 400044, China

³ Intelligent Connected Technology of CAERI Co., Ltd., Chongqing 401122, China
xiongyingzhi@caeri.com.cn

Abstract. Simulation test can reduce time costs by accelerating testing and quickly improve test scenario coverage, especially edge scenarios with low probability of occurrence. How to reasonably and accurately define and calculate the coverage of autonomous driving (AD) simulation test scenarios is a very important issue. This article proposes a scenario enumeration method for AD behavior safety, and defines test scenario coverage from four levels, namely functional scenario theoretical coverage, functional scenario construction coverage, logical scenario construction coverage, and concrete scenario parameter coverage. The first two layers represent the entire traffic environment with typical limited functional scenarios. The last two layers represent behavioral patterns with reasonable parameter ranges. Based on the data-driven method, the completeness of coverage is proved and the adequacy of simulation verification are ensured. This research can be used to evaluate the coverage of simulation scenarios, prove the completeness of the simulation scenario database, and guide the development and testing of autonomous driving systems (ADS).

Keywords: Autonomous driving · Scenario coverage · Simulation test · Behavior safety

1 Introduction

Globally, traffic accidents cause more than 1 million casualties every year [1]. Studies show human error accounts for 75% of all traffic accidents, and that number exceeds 90% in the United States [2]. AD technology aims to improve the safety and convenience of vehicles. As AD technology, regulations, and standards continue to improve, levels 3 and higher [3] autonomous vehicles (AVs) will gradually appear on the road. But before that, how to test the safety of AVs is a problem that must be solved.

The safety of AVs mainly includes functional safety, safety of the intended functionality, behavior safety, human-computer interaction safety, cybersecurity, etc. According to ISO 26262 [4], functional safety focuses on unreasonable risks caused by failures

of electronic and electrical systems, and the goal is to reduce the probability of danger caused by these risks to a sufficiently low level. Safety of the intended functionality is first defined in ISO 21448 [5], focusing on hazards and risks caused by performance insufficiencies or reasonably foreseeable misuse. For example, in extreme weather conditions such as heavy rain and fog, the perception sensors do not malfunction, but the AVs can not drive as expected. Functional safety and safety of the intended functionality both focus on low-probability events. During the entire life cycle of AVs, most of the traffic scenarios AVs deal with are normal, such as leading vehicles braking and neighboring vehicles cutting in. Scenarios such as serial collisions and dropped goods are only met with a very small probability. Safety in scenarios with low occurrence probability is important, but more important is how to define and evaluate the safety in normal traffic scenarios. Only AVs that meet behavioral safety requirements can be recognized by the government, users and the general public.

The world forum for harmonization of vehicle regulations (WP.29) released the “Framework Document on Automated/autonomous Vehicles” [6], aiming to identify key principles for the safety and security of AVs of levels 3 and higher. The safety principle it defines is that autonomous vehicle systems, under their operational design domain (ODD), shall not cause any traffic accidents resulting in injury or death that are reasonably foreseeable and preventable. The “reasonably foreseeable” mentioned above refers to scenarios that are reasonable and not extremely likely to occur under the ODD of AVs. The “preventable” means a collision that can be avoided by a competent and attentive human driver. If AVs can avoid collisions to the level of a competent and attentive human driver in reasonably foreseeable scenarios under ODD, AVs can be judged to meet the safety principles defined in WP.29. The AVs need to avoid reasonably foreseeable and preventable accidents under ODD, which has become an industry consensus on the capability requirements of the AVs under normal conditions. This is also the core of the behavior safety.

The method for determining parameter range of the reasonably foreseeable scenarios can refer to standard ISO 34502 [7, 8]. Based on the driving behavior data of human drivers in real-world traffic, the distribution of scenario parameters is statistically obtained, thereby determining the parameter range of reasonably foreseeable scenarios. The method for determining the level of the competent and attentive driver’s ability to avoid a collision may refer to the regulation UN R-157 [9]. The regulation defines an emergency braking model for the competent and attentive human driver, and requires AVs to achieve the above-mentioned driver’s capability to minimize the risk of collision. By applying the collision avoidance model of human driver in the reasonably foreseeable simulation scenarios, the parameter set where no collision occurs in the simulation is the reasonably foreseeable and preventable, as shown in Fig. 1.

Safety tests of AVs include road test, track test and simulation test. Road test faces problems such as long cycle, high cost, and unsafety. A study by the Rand [10] shows that to prove that AVs are safer than human drivers, 100 vehicles need to be tested continuously for 400 years. Track test also faces the high cost and unsafe problems. At the same time, track testing is difficult to cover a large number of test cases. Simulation test can avoid the safety risks of real vehicle test, especially testing in dangerous scenarios.

Besides, accelerated test can reduce test time costs and quickly increase scenario coverage, especially edge scenarios with low probability of occurrence. Simulation test has become an indispensable method of AVs' safety test.

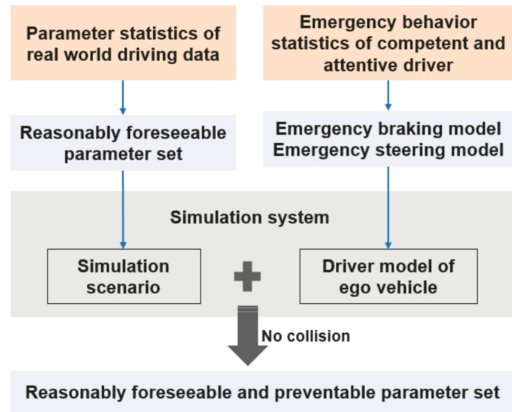


Fig. 1. Flow chart of determining reasonably foreseeable and preventable scenario parameter set.

Simulation test is mainly divided into two types: scenario-based test and continuous test. Scenario-based test [11] uses short-duration, fragmented, and designed scenarios to test the capacity of AVs to deal with specific weather, roads, and traffic behaviors. Continuous test [12, 13] uses scenarios of large map and complex traffic flow to test the capacity of AVs in real-world traffic. Since the scenario-based test is more targeted at testing and verifying AD functions, this method has become one of the most widely used methods in access and evaluation test of AVs. It is important to evaluate whether the coverage of simulation test scenarios meets the safety requirements of AVs.

The PEGASUS project [14] proposes a standardized process for scenario-based simulation test, which divides simulation scenarios into functional scenario, logical scenario, and concrete scenario. Many simulation scenario designers use the number of scenarios as an indicator to measure the coverage of scenario database, which mainly includes the number of functional scenarios, logical scenarios and concrete scenarios. Because the dimension of scenario generalization is too broad, it is difficult to judge that the greater the number of scenarios, the higher the coverage of scenarios. The total test duration or total test mileage are also common indicators to measure scenario coverage. However, the scenarios meet by AVs in the above tests are highly similar, and the total test time or mileage cannot truly reflect the real scenario coverage. In addition to the above two basic coverage indicators, [15] uses the critical indicator function to identify key scenarios in the scenario parameter space and estimate the coverage. Critical indicators such as time to collision (TTC) lack a strict correspondence with the value ranges of real-world scenario parameters, and some scenarios corresponding to specific TTC values will not appear in the real world. In [16], the integral value of the probability distribution function of the scenario parameter is used as the coverage indicator of the scenario parameter. The probability distribution function of scenario parameters is difficult to obtain accurately, which affects the calculation accuracy of the

final coverage. The study [17] proposes the use of statistical methods to evaluate the action differences and coverage of ADS in multiple simulation tests. Horel J B, et al. [18] provide a automatic method to generate a massive relevant critical scenarios based on a formal model and test purpose. The paper [19] proposed a approach to simplify unnecessary ADS testing scenarios.

Section 2 introduces the key problems in detail that this article focuses on solving, and provide an overview of the scenario coverage methodology proposed in this article. Section 3 introduces the scenario coverage calculation method based on behavior safety test, including the behavior set and determination method of behavior safety scenario, as well as the multi-level scenario coverage calculation method. Section 4 gives an example of scenario coverage calculation based on the ADS ODD defined in UN R-157. Section 5 summarizes the above content, gives conclusions, and looks forward to future research work.

2 Problem Statement

The scenarios for autonomous driving are endless and can never be completely covered. Scenario coverage is a relative value, and its calculation must have a reasonable universal set of scenarios as the denominator. However, current scenario coverage researches mainly focus on absolute values. Some evaluate the number of scenarios, and some serve as cost indicators for finding key or dangerous scenarios. It is difficult to truly evaluate whether the completed test scenarios for AVs truly meet the safety requirements. If not, how far away are the scenarios from the safety requirements? In addition, current scenario coverage researches focus on the parameter values of concrete scenarios, while it is insufficient for higher-level coverage of functional and logical scenarios. Finally, current scenario coverage is mostly used to solve the long-tail problem of AD, especially focusing on scenario types with a low probability of occurring. As for the types of scenarios with high probability, there is no reasonable standard to define what capability level AVs must reach to be considered safe.

This article focuses on the scenario coverage of behavior safety which has little attention but high importance. By defining the scenario coverage of behavior safety, a reference standard is provided for the capability baseline of AVs in normal traffic scenarios. First, this article proposes a method of extracting corresponding behavior sets and behavior safety scenarios based on the ODD of the ADS. The method can comprehensively identify the universal set of behavior safety scenarios within ODD, which can be used as the denominator for coverage calculation. Secondly, this article proposes a multi-level scenario coverage calculation method to comprehensively evaluate the coverage of functional scenarios, logical scenarios, and concrete scenarios. Finally, for parameter coverage of concrete scenario, this paper proposes different coverage calculation methods based on the reasonably foreseeable and preventable parameter set. The method can adaptively adjust the neighborhood size of parameter calculation according to the criticality of the scenario, ensuring the feasibility of coverage calculation of discrete parameter values (see Fig. 2).

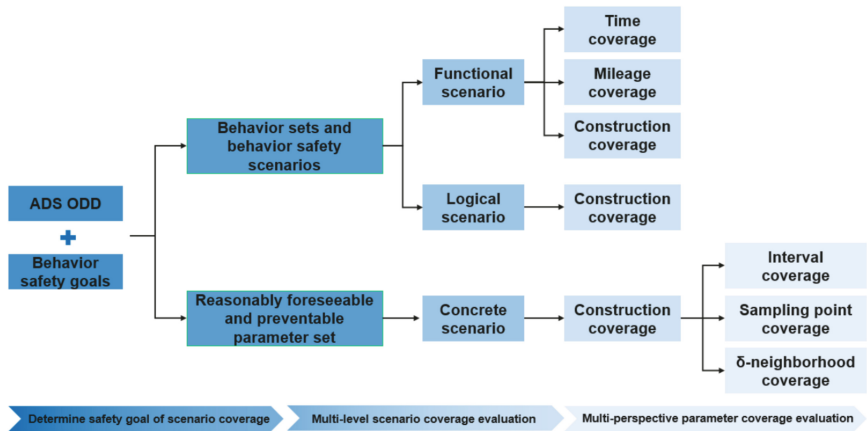


Fig. 2. Flow chart of assessing behavior safety scenario coverage.

3 Methodology of Scenario Coverage

3.1 Behavior Set and Behavior Safety Scenario

Currently, when designing simulation scenarios, considered factors include road, weather, object type, object behavior, etc. Since the generalization range of each factor is too large, the number of scenarios can be easily led to an explosion. Not only does this make a complete test take too much time, but the test fails to focus on key safety issues of the ADS. Since behavior safety focuses on reasonably foreseeable normal traffic scenarios, the risks in such scenarios mainly come from the behavior interaction between AVs and objects. Therefore, based on typical road, weather, and object types, the safety of behavior interactions between AVs and objects can be focused on. In this way, the complex dimensionality of scenarios is reduced to the level of behavior interaction. Only by determining the behavior set of AV and object, the behavior safety can be completely verified. Furthermore, reasonable and effective scenario coverage indicators are used to prove that limited and reasonable behavior safety scenarios can fully represent the real traffic environment, thereby proving the behavior safety of AVs.

When determining the behavior set of AV and object, the ODD must first be determined. ODD defined in the UN R-157 represents the current capability level of AD. The following will take the ODD mentioned above as an example to introduce the behavior set. The ODD is designed to operate at a speed of 0-60km/h of AV. The road environment is a closed road, such as main road of highway or urban expressway. The function is automated lane keeping (ALK) and does not include active lane changing.

In this ODD, objects include vehicles such as small vehicles, large vehicles, etc., and pedestrians. If the object is a vehicle, the behaviors include steering behavior in the lateral direction and speed behavior in the longitudinal direction. Lateral steering behaviors include lane keeping, cut-in, and cut-out. Longitudinal speed behaviors include stationary, constant speed, acceleration, and deceleration. Since the ODD does not include intersections, there is no crossing, left turn, right turn and U-turn in the steering behavior.

Under the premise of same object vehicle's behavior, if object vehicle is in different positions of ego vehicle, whether the risks of behavior safety will occur and the types of risks will be different. Therefore, in addition to the behaviors of the object vehicle, the position relationship between object vehicle and ego vehicle also needs to be considered. The position of the object vehicle is shown in Fig. 3. The positions numbered by "1"–"8" are all possible position relationships between the object vehicle and ego vehicle. The positions numbered by "+1", "+2", and "+3" are the other object vehicles that locate in front of the object vehicles numbered by "1", "2", and "3" respectively. As shown in Fig. 3, behavioral safety risks will arise when the leading vehicle cuts out to meet another vehicle, the leading vehicle brakes, and the neighboring vehicle cuts in.

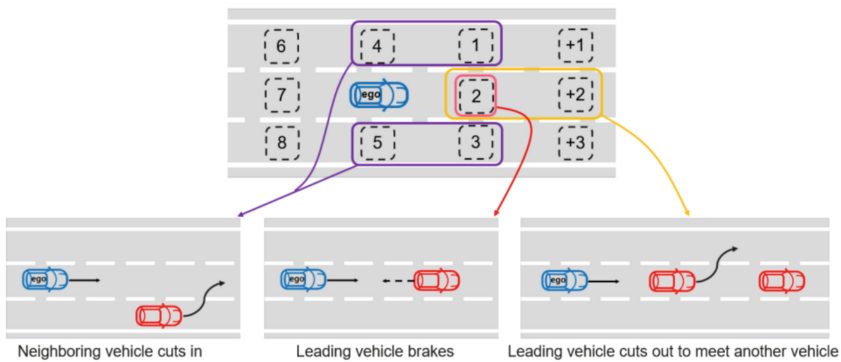


Fig. 3. Identify risks of behavior safety through locations and behaviors of object vehicle.

By traversing behaviors and positions of the object vehicle, the behavior safety scenarios when the object is a vehicle can be determined, as shown in Table 1 below. The scenarios marked by ● are all likely to have behavior safety risks, including the vehicle decelerating at position "2", the vehicle cutting in at position "1" at a constant speed, the vehicle cutting out at position "2" at a constant speed and then encountering a stationary vehicle, etc.

For example, it is obvious that the vehicle at position "2" is stationary or slowing down, which may easily cause risks. In addition, when the vehicle at position "2" is traveling at low constant speed, or when accelerating from a standstill, the higher speed of the ego vehicle may also cause risks. It is worth mentioning that only the most typical steering behaviors and speed behaviors are listed in the table. For example, leading vehicle constant speed cutting out is the most typical cut-out behavior. Acceleration cutting out or deceleration cutting out are both low-probability events in actual traffic. In addition, braking after cutting in and cancellation during the cutting in are all small-probability events and do not fall within the scope of reasonably foreseeable normal traffic scenarios that concerned by behavior safety. If the object is a pedestrian, in an ODD of a closed road, pedestrians walking along, cutting in, and crossing the road are common scenarios that are easily leading to behavior safety risks. The method for determining behavior safety scenarios is similar to that where the object is a vehicle. No need to go into details here.

3.2 Multi-level Scenario Coverage Calculation Method

For simulation test of AVs, the current mainstream approach is to determine functional scenarios based on AV product functions and test requirements firstly. Then design and generalize the corresponding logical scenarios and concrete scenarios. Finally, the AV is simulated and tested through concrete scenarios. When calculating the coverage of simulation test scenarios, it is also necessary to calculate separately for the three levels of scenarios. Multi-level scenario coverage calculation can more comprehensively reflect whether the construction of the simulation scenario library is complete.

After the above behavior safety scenarios are determined, corresponding functional scenarios, logical scenarios and concrete scenarios are designed respectively to form a universal set of respective scenarios. On the one hand, it is necessary to use the coverage indicators of functional scenario to prove whether the designed functional scenarios can fully represent the real traffic environment. On the other hand, it is necessary to use the coverage indicators of logical scenarios and concrete scenarios to prove whether the designed logical scenarios and concrete scenarios can fully cover the behavior set and scenario parameters. Simulation scenario builders can use multi-level coverage indicators to evaluate the coverage of existing scenario categories and quantities to prove the completeness of the simulation scenario library and guide the direction of ADS development and testing.

The scenario coverage indicators designed in this article are divided into four levels, including functional scenario theoretical coverage, functional scenario construction coverage, logical scenario construction coverage, and concrete scenario parameter coverage. The first two layers represent the entire traffic environment with typical limited functional scenarios. The last two layers represent behavioral patterns with reasonable parameter ranges. The specific definitions of scenario coverage at each level are as follows.

3.2.1 Functional Scenario Theoretical Coverage

Since the functional scenario only focuses on the behavior of the ego vehicle and the object vehicle, for the functional scenario of behavior safety in ODD of closed roads, the vehicle behaviors only in Table 1 need to be focused on. It is easy to see that the universal set of functional scenarios about behavior safety includes 11 functional scenarios such as the object vehicle decelerating at position, the object vehicle cutting in at a constant speed, the object vehicle cutting out at a constant speed and then encountering a stationary vehicle.

Functional scenario theoretical coverage includes time coverage and mileage coverage. The two aim to prove that the limited number of behavior safety function scenarios extracted above can cover most real road traffic scenarios in the time and mileage dimensions. For the universal set of real road traffic scenarios, it is necessary to collect vehicle behavior data in real road traffic based on the ODD of closed road. Sensors such as millimeter-wave radar, lidar, and camera can be installed on the data collection vehicle to obtain objects and behaviors through perception fusion algorithms. Data collection must be time-sufficient and continuous. Only in this way can the validity of the data be ensured and the coverage of functional scenarios be calculated truly and effectively.

According to the identified functional scenarios of behavior safety, functional scenarios matching and slicing are performed on the real traffic data with a total duration of T and a total mileage of D . For example, [slice 1 of vehicle decelerating, slice 2 of vehicle decelerating, slice 1 of vehicle cutting in at a constant speed, slice 1 of vehicle accelerating and cutting in, slice 2 of vehicle accelerating and cutting in, slice 1 of vehicle cutting out at a constant speed and then encountering a stationary vehicle...]. The time duration of each scenario slice is $[t_1, t_2, t_3, t_4, t_5, t_6, \dots]$. The mileage duration of each scenario slice is $[d_1, d_2, d_3, d_4, d_5, d_6, \dots]$. The time coverage C_{f_t} of the identified functional scenarios in the universal set of real traffic scenarios is:

$$C_{f_t} = \sum_i \frac{t_i}{T} \quad (1)$$

The mileage coverage C_{f_d} of the identified functional scenarios in the universal set of real traffic scenarios is:

$$C_{f_d} = \sum_i \frac{d_i}{T} \quad (2)$$

3.2.2 Functional Scenario Construction Coverage

Functional scenario construction coverage is designed to evaluate whether the constructed functional scenarios of behavior safety are complete. The key to the coverage calculation is to reasonably identify the universal set of functional scenarios and determine the number of scenarios in the universal set. If the number of scenario categories in the universal set of functional scenarios of behavior safety is N_f , and the number of functional scenario categories identified and implemented by the scenario builder is n_f , then the functional scenario construction coverage C_{f_c} is:

$$C_{f_c} = \frac{n_f}{N_f} \quad (3)$$

$$C_{p_c} = \frac{\prod_i r_i}{\prod_i R_i} \quad (4)$$

The above calculation method ignores the correlation between multiple parameter values and is a simplified version of the calculation method. This calculation method is simple and easy to accomplish, and can reflect the coverage of concrete scenario parameters within an appropriate error range. Since the real parameter value ranges are irregular, the value ranges can be sampled according to a certain value step. In this way, the reasonably foreseeable and preventable value range becomes the set of value sampling points. If the number of sampling points in the set of reasonably foreseeable and preventable parameter sampling points is P_c , and the number of identified and realized parameter sampling points of concrete scenario is p_c , then the concrete scenario parameter coverage C_{p_s} is:

$$C_{p_s} = \frac{p_c}{P_c} \quad (5)$$

The above method is to take the discrete values in a certain step within a continuous parameter value range, and then to calculate scenario coverage based on whether discrete values of parameters are tested. In the actual test process, it is easy to occur that the discrete value points in the standard cannot be aligned with that in the actual test, due to the different sampling step size. To solve this problem, the δ -neighborhood of the sampling point can be determined with the value in the standard as the center and a certain length as the radius of the neighborhood. If the discrete parameter value of actual test falls within the δ -neighborhood of a certain standard parameter value, the standard parameter value can be considered to have been tested. If the number of sampling points in the set of reasonably foreseeable and preventable parameter sampling points is $P_{c,\delta}$, and among them, there are $p_{c,\delta}$ neighborhoods with concrete scenarios that have been identified and implemented in them, then the concrete scenario parameter coverage $C_{p,\delta}$ is:

$$C_{p,\delta} = \frac{p_{c,\delta}}{P_{c,\delta}} \quad (6)$$

It is worth mentioning that the radius of the δ -neighborhood of a sampling point is related to the danger level of the scenario parameter value. The higher the danger level of the scenario, the smaller the radius of the δ -neighborhood. The purpose of this is that for parameter values in dangerous scenarios, a small difference in values will lead to completely different responses and test results of the AV, so the test should be as detailed as possible. For scenarios with a low level of danger, the radius of the δ -neighborhood can be slightly larger to improve test efficiency while ensuring test accuracy.

4 Verification of Scenario Coverage

The following will take the behavior set and behavior safety scenarios mentioned in Sect. 3.1 as an example to give the verification of the multi-level scenario coverage calculation method.

4.1 Functional Scenario Theoretical Coverage

Since behavior safety focuses on reasonably foreseeable traffic scenarios, the time coverage and mileage coverage of the identified function scenarios just need to reach above 90% or 99%. If the AVs have higher requirements, the target values of time coverage and mileage coverage can be appropriately increased. When selecting functional scenarios, scenario builder should first sort scenarios according to proportion of occurrence in real traffic, giving priority to scenarios with a high proportion of occurrence until the cumulative time coverage or mileage coverage reaches the target, as shown in Fig. 4.

4.2 Functional Scenario Construction Coverage

Through the recognition method of behavior safety function scenarios, the object behaviors are traversed and unreasonable scenarios are eliminated. For example, the vehicle cannot cut in while the vehicle is stationary. Then a universal set of behavior safety

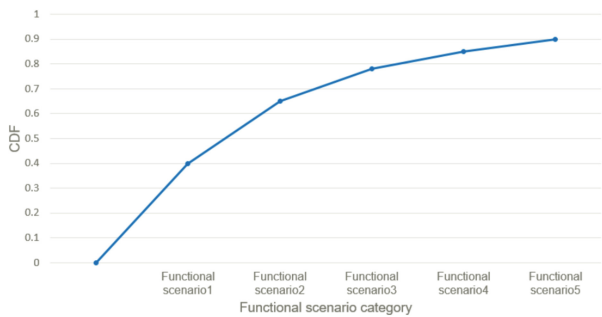


Fig. 4. CDF curve of functional scenario time/mileage coverage.

function scenarios can be constructed. The universal set of functional scenarios about behavior safety obtained in Table 1 includes 11 functional scenarios such as the object vehicle decelerating at position, the object vehicle cutting in at a constant speed, the object vehicle cutting out at a constant speed and then encountering a stationary vehicle. If 8 functional scenarios have been implemented, then the functional scenario construction coverage is $8/11 = 72.7\%$.

4.3 Logic Scenario Construction Coverage

Taking the constant speed cut-in functional scenario mentioned above as an example, the universal set of logical scenarios corresponding to this functional scenario need to be identified. As shown in Fig. 5, under the typical road topology condition in ODD of closed road, there are three different positions for the ego vehicle to locate in the lanes. Then traverse all possible position relationships between the object vehicle and the ego vehicle. The positions of the object vehicle are the positions marked by numbers “1”–“8” in Fig. 5. Then, there are four cut-in situations marked by ①–④ for the object vehicle constant speed cutting in. In addition, whether there is a leading vehicle at position “2” in front of ego vehicle also has a non-negligible impact on the scenario.

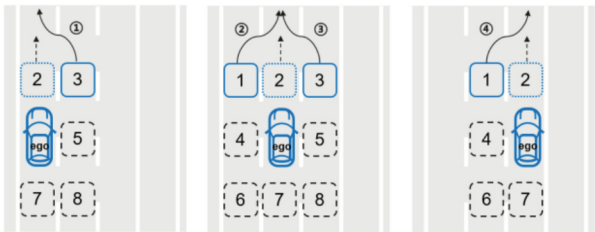


Fig. 5. Different logical scenarios corresponding to constant speed cutting in.

By traversing 4 different cut-in scenarios and with or without a leading vehicle, 8 logical scenarios can be obtained. Remove highly similar scenarios, where scenario ① is similar to scenarios ③, and scenarios ② is similar to scenarios ④. Therefore, the

cut-in scenarios ① and ④, or ② and ③ need to be reserved. Finally, a universal set of logical scenarios containing 4 scenarios is obtained, as shown in Fig. 6. If 3 of the logical scenarios have been implemented, the logical scenario construction coverage of the functional scenario is $3/4 = 75\%$.

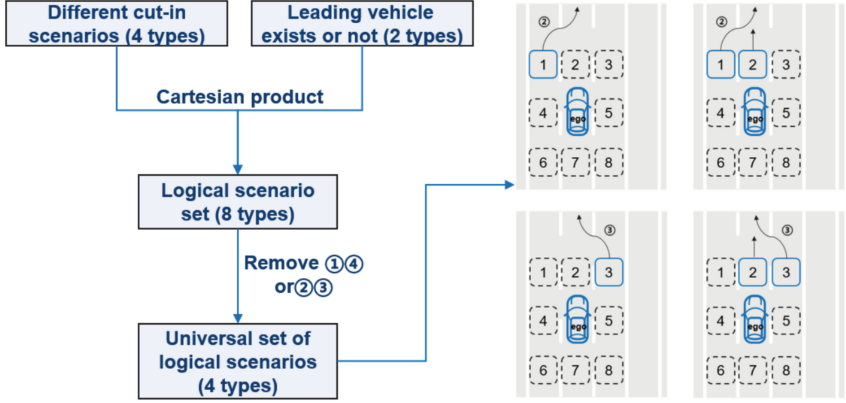


Fig. 6. Universal set of logical scenarios corresponding to constant speed cutting in.

4.4 Concrete Scenario Parameter Coverage

Taking the constant speed cut-in scenario as an example, the scenario has two key parameters, namely the cut-in distance between the object vehicle and the ego vehicle and the cut-in duration. As shown in Fig. 7, the reasonably foreseeable and preventable parameter value range of the scenario is $[l_{d3} - l_{d1}, l_{t3} - l_{t1}]$, which are the value ranges of cut-in distance and cut-in time respectively. Within the above value range, the concrete scenario parameter value range that the scenario builder has implemented is $[l_{d3} - l_{d2}, l_{t3} - l_{t2}]$. Assuming that the cut-in distances l_{d1} , l_{d2} , and l_{d3} are 10 m, 20 m, and 50 m respectively, and the cut-in times l_{t1} , l_{t2} , and l_{t3} are respectively 3 s, 5 s, and 15 s, then the concrete scenario parameter coverage C_{p-c} is:

$$\begin{aligned}
 C_{p-c} &= \frac{(l_{d3} - l_{d2}) \times (l_{t3} - l_{t2})}{(l_{d3} - l_{d1}) \times (l_{t3} - l_{t1})} \\
 &= \frac{(50 - 20) \times (15 - 5)}{(50 - 10) \times (15 - 3)} \\
 &= 62.5\%
 \end{aligned} \tag{7}$$

If the coverage is calculated using sampling points method, as shown in Fig. 7, the number of sampling points in the set of reasonably foreseeable and preventable parameter sampling points is $P_c = 150$, and the number of identified and realized concrete scenario parameter sampling points is $p_c = 108$, then the concrete scenario parameter coverage C_{p-s} is:

$$C_{p-s} = \frac{p_c}{P_c} = \frac{108}{150} = 72\% \tag{8}$$

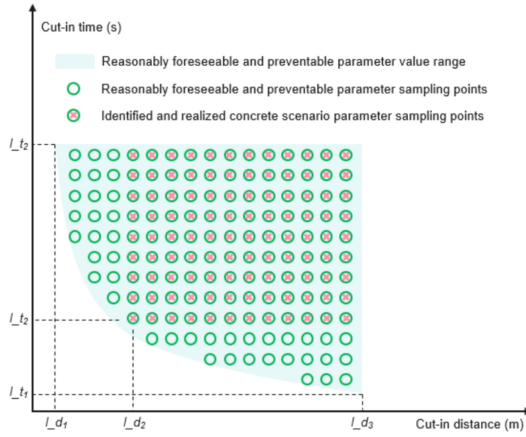


Fig. 7. Constant speed cut-in scenario parameter coverage based on intervals and sampling points.

If the coverage is calculated based on the δ -neighborhood method, as shown in Fig. 8, the number of sampling points in the set of reasonably foreseeable and preventable parameter sampling points $P_{c_\delta} = 11$. The scenarios near the parameter boundary is more dangerous, so the δ -neighborhood of the sampling point is smaller. Among them, there are $p_{c_\delta} = 8$ neighborhoods with concrete scenarios that have been identified and implemented in them, then the concrete scenario parameter coverage C_{p_δ} is:

$$C_{p_\delta} = \frac{p_{c_\delta}}{P_{c_\delta}} = \frac{8}{11} = 72.7\% \quad (9)$$

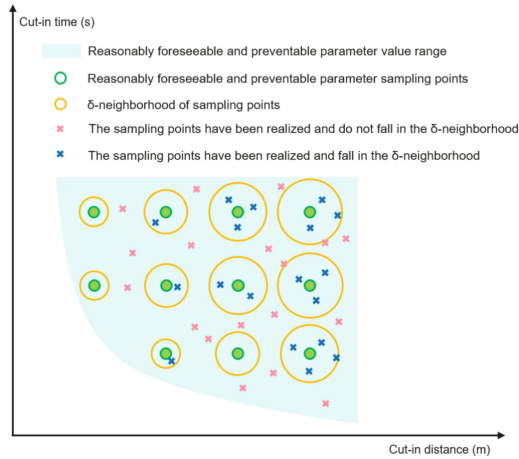


Fig. 8. Constant speed cut-in scenario parameter coverage based on δ -neighborhood.

5 Conclusions

This article proposes a scenario enumeration method for AD behavior safety, and defines test scenario coverage from four levels, namely functional scenario theoretical coverage, functional scenario construction coverage, logical scenario construction coverage, and concrete scenario parameter coverage. For parameter coverage of concrete scenario, this paper proposes different coverage calculation methods based on the reasonably foreseeable and preventable parameter set. The method can adaptively adjust the neighborhood size of parameter calculation according to the criticality of the scenario, ensuring the feasibility of coverage calculation of discrete parameter values. This research can be used to evaluate the coverage of simulation scenarios, prove the completeness of the simulation scenario database, and guide the development and testing of ADS.

Acknowledgments. This paper was funded by Research and Application of Key Technologies for Evaluation of City Navigation Pilot System (CNP, No. 0001KTCP20230440) and China-SAE Young Elite Scientists Sponsorship Program Project (2023–2025, No. QK-20240316-008).

References

1. Navet, N., Simonot-Lion, F.: *Automotive Embedded Systems Handbook*. CRC Press (2017)
2. Chen, S., Leng, Y., Labi, S.: A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information. *Comput.-Aided Civil Infrastruct. Eng.* **35**(4), 305–321 (2020)
3. SAE J3016_201806, Taxonomy and Definitions for Terms Related Driving Automation System for On-road Motor Vehicles (2018). https://www.sae.org/standards/content/j3016_201806/
4. ISO 26262, Road Vehicles - Functional safety. Technical Committee ISO/TC 22/SC 32 (2018). <https://www.iso.org/standard/68383.html>
5. ISO 21448, Road Vehicles - Safety of the Intended Functionality. Technical Committee ISO/TC 22/SC 32 (2022). <https://www.iso.org/standard/77490.html>
6. ECE/TRANS/WP.29, Framework Document on Automated/Autonomous Vehicles (2019). <https://unece.org/DAM/trans/doc/2019/wp29/WP29-177-19e.pdf>
7. ISO 34502, Road vehicles-Scenario-based Safety Evaluation Framework for Automated Driving Systems (2022). <https://www.iso.org/standard/78951.html>
8. Nakamura, H., Muslim, H., Kato, R., et al.: Defining reasonably foreseeable parameter ranges using real-world traffic data for scenario-based safety assessment of automated vehicles. *IEEE Access* **10**, 37743–37760 (2022)
9. ECE/TRANS, UN Regulation No. 157 - Automated Lane Keeping Systems (ALKS) (2021). <https://unece.org/transport/documents/2021/03/standards/un-regulation-no-157-automated-lane-keeping-systems-alks>
10. Kalra, N.: *Challenges and Approaches to Realizing Autonomous Vehicle Safety*. Santa Monica, RAND (2017)
11. Riedmaier, S., Ponn, T., Ludwig, D., et al.: Survey on scenario-based safety assessment of automated vehicles. *IEEE Access* **8**, 87456–87477 (2020)
12. Li, N., Oyler, D.W., Zhang, M., et al.: Game theoretic modeling of driver and vehicle interactions for verification and validation of autonomous vehicle control systems. *IEEE Trans. Control Syst. Technol.* **26**(5), 1782–1797 (2017)

13. Feng, S., Yan, X., Sun, H., et al.: Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nat. Commun.* **12**(1), 748 (2021)
14. PEGASUS. PEGASUS Research Project (2019). <http://www.pegasusprojekt.de/en/home>
15. Hungar, H.: A concept of scenario space exploration with criticality coverage guarantees, pp. 293–306. Springer (2020)
16. Weissensteiner, P., Stettinger, G., Khastgir, S., et al.: Operational design domain-driven coverage for the safety argumentation of automated vehicles. *IEEE Access* **11**, 12263–12284 (2023)
17. Laurent, T., Klikovits, S., Arcaini, P., et al.: Parameter coverage for testing of autonomous driving systems under uncertainty. *ACM Trans. Softw. Eng. Methodol.* **32**(3), 1–31 (2023)
18. Horel, J.B., Laugier, C., Marsso, L., et al.: Using formal conformance testing to generate scenarios for autonomous vehicles, pp. 532–537. IEEE (2022)
19. Arcaini, P., Zhang, X.Y., Ishikawa, F.: Less is more: simplification of test scenarios for autonomous driving system testing, pp. 279–290. IEEE (2022)



Research on Charging Location of New Energy Vehicles and Mobile Charging Vehicles Under Multi-objective Coordination

Yongqi Diao¹ and Ying Lv²(✉)

¹ School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

² School of Systems Science, Beijing Jiaotong University, Beijing 100044, China
ylv@bjtu.edu.cn

Abstract. Mobile charging vehicles can further alleviate the mileage anxiety of new energy vehicle users. However, due to the high cost of mobile charging vehicles dispatching, mobile charging vehicles face certain challenges in promotion and popularization. In order to solve the above problems, this study proposes a service strategy for mobile charging vehicles to serve multiple users at the same time. Considering the user request of new energy vehicles and the operating cost of mobile charging vehicles, a multi-objective coordinated mobile charging vehicle scheduling optimization strategy is proposed. By constructing a multi-objective optimization model including the service cost of the mobile charging vehicle, the user's circumference cost and the waiting cost, the charging demand of the new energy vehicle user, the operation restriction of the mobile charging vehicle and the user's time window constraint are integrated into the path optimization model of the mobile charging vehicle and the new energy vehicle. A method for finding non-dominant extreme points in the target space is designed to generate the frontier surface of the law of the vital few. The experimental analysis of the coupling and mutual exclusion of each target shows that the service strategy can reduce the charging cost while meeting the user request of the new energy vehicle.

Keywords: Urban traffic · path optimization of mobile charging vehicle · pareto front · new energy vehicles · multi-objective optimization

1 Introduction

The mobile charging vehicle's flexibility and mobility enable it to be precisely dispatched according to the actual needs of the expressway service area. When the charging pile in the service area fails or the utilization rate is too high, the mobile charging vehicle can quickly replenish the energy supply capacity. At the same time, for some remote expressway service areas, the construction of fixed charging piles may be relatively lagging behind, and the mobile charging vehicle can play its unique advantages to provide reliable energy security for new energy vehicles in these areas. In short, the mobile charging vehicle has great application potential in the expressway service area, which will provide strong support for the wide promotion and sustainable development of new energy vehicles.

The current research on mobile charging vehicles mainly focuses on mobile charging vehicle scheduling and mobile charging vehicle travel routes. First, for the study of mobile charging vehicle scheduling, many scholars have proposed a variety of scheduling algorithms and strategies, aiming to provide users with more efficient energy replenishment services. Atmaja et al. [1] designed a scheduling algorithm to dispatch mobile charging vehicles to fixed charging stations or directly to the user's location to provide energy replenishment services; Tang et al. [2] based on a first-come, first-served service strategy, scheduling mobile charging vehicles, and first-come, first-served. The difference is that the electric vehicles in the first-come, first-served service are charged according to their order on the virtual queue, and the earliest available charging time window for the user is selected; Zhang et al. [3] based on the historical order data of the fixed charging station, introduce a reservation model for mobile charging services, and dispatch mobile charging vehicles to provide charging services for users who need electric vehicles; Liu et al. [4] consider the charging time of all users and the travel distance of the mobile charging vehicle as the goal, and propose a greedy algorithm and a global optimization algorithm based on the earliest-end first-served scheduling. The mobile charging vehicle is dispatched to the location where the charging demand is greatest; in order to dispatch the mobile charging vehicle to the appropriate location of the road network, Moghaddam et al. [5] proposed a meta-heuristic solution based on the ant colony optimization algorithm; Riboac et al. [6] proposed a mobile charging vehicle operation mode, which temporarily arranges the mobile charging vehicle in different locations according to the charging demand, and abstracts the mobile charging problem into nonlinear optimization and equivalent linear optimization problems.

The second is related research on the driving route of mobile charging vehicles. Qureshi et al. [7] introduced the optimization problem of minimizing the charging cost of mobile charging vehicles under the premise of satisfying the driving constraints of mobile charging stations and the cut-off time of charging requests, and proposed a heuristic algorithm to solve this problem; Cui et al. [8] Considering the operating cost of charging stations, a mobile charging vehicle service problem with time windows and multiple modes of service was introduced, and it was expressed as a 0–1 mixed integer linear model for solving; Kabir et al. [9] expressed the mobile charging vehicle travel route problem as an integer linear programming problem, and determined the best route for each mobile charging station with the goal of maximizing the number of electric vehicles served; Li et al. [10] designed a mobile charging vehicle charging based on charging factors such as charging time estimation, travel time prediction and charging convenience. Navigation scheme to determine the best driving route and parking location; Jeon et al. [11] proposed a different type of routing problem for mobile charging vehicles, which are dispatched to overloaded fixed charging stations through the resulting optimal path to reduce the waiting time of electric vehicles rather than traveling to the location of the user's electric vehicle.

At present, in the field of energy replenishment of new energy vehicles, although mobile charging vehicles are regarded as innovative solutions, most of the existing studies are limited to mobile charging vehicle scheduling under the scenario where the user's charging location is determined, and less attention is paid to the situation where the user negotiates the charging location with the mobile charging vehicle. In fact, the

negotiated charging location can be regarded as a state where the charging location is uncertain. In this case, the mobile charging vehicle can comprehensively consider the needs of multiple users. When the charging location is uncertain, the mobile charging vehicle is no longer only based on the fixed user charging location for scheduling, but can be dynamically planned according to the real-time location, power demand, and surrounding environment of multiple users. In view of this, this paper proposes a charging location selection model that considers the negotiation of charging locations between users and mobile charging vehicles, and verifies the rationality of the model through an example, opening up a new way to solve the energy replenishment problem of new energy vehicles.

2 Problem Description and Model Construction

2.1 Problem Description

The mobile charging operator has a fixed number of mobile charging vehicles to meet all the charging needs of P -bit users. Each user sends the current power, departure time, departure place, destination and arrival time to the mobile charging operator. The mobile charging operator designates the same charging location for multiple users according to the user request information, and sends the mobile charging vehicle to the designated charging location. The user completes the power supply at the designated charging location. The goal is how to specify a suitable charging location for the user and plan the service path of the mobile charging vehicle, so that the mobile charging operator has the lowest total cost of the mobile charging vehicle and the charging cost for the user under the premise of meeting the charging needs of all users.

To facilitate the establishment of the model, the following assumptions are made.

- (1) The user's charging demand information is known, and the user will go to the charging location designated by the mobile charging operator for energy replenishment.
- (2) Before the mobile charging car departs, it receives charging requests from P users, and no new charging requests will be generated on the way. There is only one mobile charging car site, and the mobile charging car departs from the site and finally returns.
- (3) Do not consider the acquisition cost of the mobile charging vehicle, but only consider the driving cost and waiting cost during a single charging service.
- (4) Mobile charging vehicles are of the same model, with the same driving speed, service capability, and charging speed.
- (5) Mobile charging vehicles and user vehicles do not consider uncertainties such as traffic jams and temporary road restrictions during driving, that is, the vehicles travel at a uniform speed during driving. The driving costs between nodes are known, and road congestion is not taken into account.

2.2 Model Construction

Table 1. Definition of variables and parameters.

symbol	definition
P	User set, user serial number $p \in P$
M	Mobile charging car collection, mobile charging car serial number $m \in M$
K	Charging point collection, collection point serial number $k \in K$
S	Time period collection, time period serial number $s \in S$
N_p	Number of users requesting mobile charging services
(i_O^p, t_O^p)	The departure time node of user p
(i_D^p, t_D^p)	The arrival time node of user p
E_0^p	The initial charge of user p
L_p	The lowest cost of travel for users to reach their destination
E^+	The charging rate of electric vehicles
E^-	The discharge rate of electric vehicles
E_0	The battery capacity of an electric vehicle
MCN	The maximum number of users allowed to travel to the same charging assembly point during the same time period
MCM	The maximum number of time periods in which a mobile charging vehicle can provide energy replenishment services
α	User request satisfaction rate
c_{pk}^1	The driving cost of user p from the starting point to the charging point k
c_{pk}^2	The driving cost for user p to reach the destination from the charging location k
t_k^s	The mobile charging cart arrives at the charging location k during period s
c_m	The driving cost of mobile charging car m
pc_m	The dispatch cost of mobile charging vehicle m
ucm	The penalty cost of mobile charging vehicles failing to reach the maximum number of simultaneous services
wc_p	The waiting cost incurred by user p due to the need to charge
wc_m	The waiting cost of the mobile charging cart m waiting for the user to arrive
X_{pk}^s	If the user's electric car goes to point k and charges during period s
Y_{mk}^s	If the mobile charging cart m arrives at the charging assembly point k and provides charging service during the s period

According to the above parameters and variables, the optimization goal of this paper is to minimize the charging cost of the user, the driving cost and the waiting cost of the

mobile charging vehicle. The user charging cost mainly includes the user bypass cost and the user waiting cost. When the user goes to the charging location designated by the mobile charging operator, the actual driving distance may be greater than the shortest path of OD , and the distance difference between the two is the user bypass cost. If the user allows the mobile charging vehicle to arrive at the charging location in advance before the mobile charging vehicle arrives at the designated charging location, the user will wait for the mobile charging vehicle to arrive, and the difference between the time the user arrives at the charging location and the arrival time of the mobile charging vehicle is the user waiting cost (Table 1).

$$\min f_1 = \sum_{s \in S} \sum_{p \in P} \sum_{k \in K} (c_{pk}^1 + c_{pk}^2 - L_p) \cdot X_{pk}^s + \sum_{p \in P} wc_p \quad (1)$$

The mobile charging car will incur a certain driving cost when going to the charging location, and its size will change with the change of the charging location. At the same time, the mobile charging car will also arrive at the charging location in advance, and the mobile charging car will also have waiting costs. Due to the service strategy of specifying charging locations with multiple users, it is impossible to provide charging services according to the pre-set maximum number of users serving users at the same time, which will increase the average cost of the mobile charging car serving a single user, which will generate part of the penalty cost.

$$\min f_2 = \sum_{m \in M} (wc_m + c_m + pc_m) + \left(MCN \cdot \sum_{s \in S} \sum_{m \in M} \sum_{k \in K} Y_{mk}^s - \sum_{s \in S} \sum_{p \in P} \sum_{k \in K} X_{pk}^s \right) \cdot ucm \quad (2)$$

$$\sum_{s \in S} \sum_{k \in K} X_{pk}^s \leq 1, \forall p \in P \quad (3)$$

$$\sum_{p \in P} X_{pk}^s \leq \sum_{m \in M} Y_{mk}^s \cdot MCN, \forall s \in S, \forall k \in K \quad (4)$$

$$\sum_{k \in K} Y_{mk}^s \leq 1, \forall m \in M, \forall s \in S \quad (5)$$

$$\sum_{s \in S} \sum_{k \in K} Y_{mk}^s \leq MCM, \forall m \in M \quad (6)$$

$$\sum_{s \in S} \sum_{p \in P} \sum_{k \in K} X_{pk}^s = P \cdot \alpha \quad (7)$$

$$\sum_{s \in S} \sum_{k \in K} c_{pk}^1 \cdot X_{pk}^s \leq \frac{E_0^p}{E^-}, \forall p \in P \quad (8)$$

$$\sum_{s \in S} \sum_{k \in K} c_{pk}^2 \cdot X_{pk}^s \leq \frac{E_0}{E^-}, \forall p \in P \quad (9)$$

$$t_O^p + \sum_{s \in S} \sum_{k \in K} c_{pk}^1 \cdot X_{pk}^s + \frac{\left(E_0 - E_0^p + \sum_{s \in S} \sum_{k \in K} c_{pk}^1 \cdot X_{pk}^s \cdot E^- \right)}{E^+} + \sum_{s \in S} \sum_{k \in K} c_{pk}^2 \cdot X_{pk}^s \leq t_D^p, \forall p \in P \quad (10)$$

$$(s - t) \cdot X_{pk}^s \leq (t_O^p + c_{pk}^1) \cdot X_{pk}^s \leq (s + t) \cdot X_{pk}^s, \forall p \in P, \forall s \in S, \forall k \in K \quad (11)$$

Among them, the constraint (3) indicates that each user can go to a maximum of one charging assembly point for power replenishment; the constraint (4) indicates that each charging assembly point can serve a maximum of MCN users; the constraint (5) indicates that each mobile charging vehicle can go to a maximum of one charging assembly point to provide charging services in the same time period; the constraint (6) indicates that each mobile charging vehicle can provide charging services in a maximum of MCM time periods; the constraint (7) indicates that the user's demand satisfaction rate must reach at least α ; the constraint (8) indicates that the user's initial power can reach the charging assembly point; the constraint (9) indicates that the user can travel to the destination after power replenishment; the constraint (10) indicates that the user can reach the destination within the specified time; the constraint (11) indicates that the user and the mobile charging vehicle can reach the destination. The maximum waiting time for each other does not exceed t .

3 Example Analysis

3.1 Example Settings

As mentioned above, the time for the excitation wave to travel back and forth is conIn order to verify the effectiveness of the model and algorithm, this paper uses *Sioux Falls network* to test, and uses Gurobi 9.1.1 to solve the model. The mobile charging car site and all user departure and destination are distributed at the node location. All nodes can become charging collection points. At present, there are 100 users who need to be served, and the number of mobile charging cars is 20. The parallel service capability and maximum number of services of a single mobile charging car are 3. The user's departure, destination, departure time, arrival time, initial power and maximum battery capacity are assumed to be known. Mobile charging operators have a fixed number of mobile charging cars to meet all the charging needs of P users.

3.2 Example Results and Analysis

In order to use the solution method of the law of the vital few frontier generation method based on the target space introduced in this paper to output the relevant service information of the mobile charging car, the obtained law of the vital few frontier is shown in Fig. 1. It can be seen from the results that the driving cost of the mobile charging car, the cost of waiting for the user to arrive and the penalty cost of not providing the charging service are between 902 yuan and 1105 yuan; the cost of detouring and waiting for the user is between 258 yuan and 381 yuan. The cost of the mobile charging car and the user

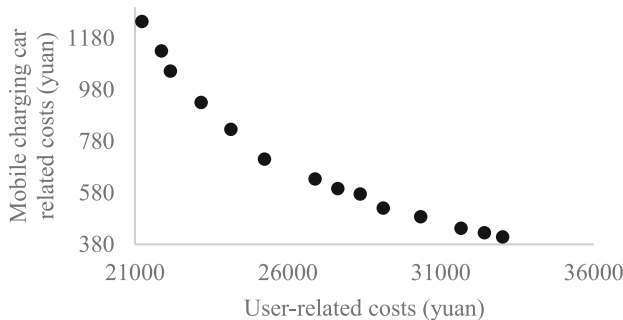


Fig. 1. The Pareto front

cost show a more obvious negative correlation trend. And when the user cost is around 280 yuan, the cost of the mobile charging car changes significantly.

When the sum of the driving cost of the mobile charging car, the cost of waiting for the user to arrive, and the penalty cost of not providing charging service is around 900 yuan, the user’s detour cost and waiting cost remain at a high level; when the cost of the mobile charging car is around 930 yuan, the user’s cost decreases significantly with the increase in the cost of the mobile charging car; when the cost of the mobile charging car is around 1100 yuan, the user’s detour cost and waiting cost remain at a low level. It can be seen that by adjusting the cost of the mobile charging car slightly near 930 yuan, the user’s cost can be significantly reduced. In other words, dynamically adjusting the user cost around 270 yuan can significantly reduce the cost of the mobile charging car (Fig. 2).

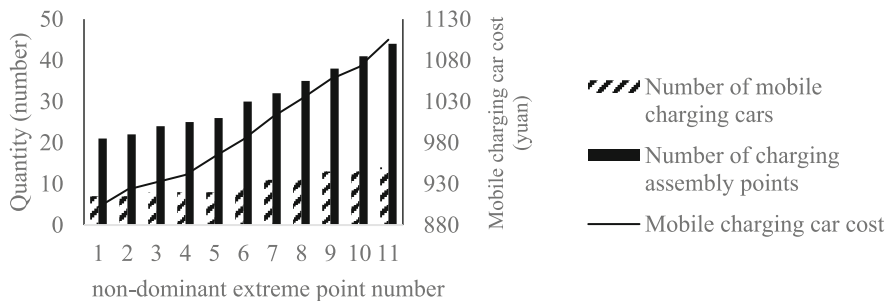


Fig. 2. Mobile charging vehicle information

By analyzing the information of the mobile charging vehicles corresponding to each non-dominant extreme point, it can be seen that when the number of mobile charging vehicles is used reaches a minimum, although it can effectively reduce the service cost of the mobile charging vehicle, this practice is often accompanied by a significant increase in the user’s detour cost and waiting cost. Users may need to travel a longer distance to the charging assembly point, or face a longer waiting time to obtain charging services.

By moderately increasing the number of mobile charging vehicles dispatched, the user's detour cost and waiting cost can be significantly reduced.

Through the above analysis, it can be found that the other objective function can maintain basic stability while the cost of a single objective function continues to decline. However, when the cost of a single objective is reduced to a certain threshold, the revenue of another objective will drop sharply. In the selection process of charging assembly points, factors such as geographical location, parking resources, and user request distribution of each assembly point should be fully considered to ensure that the value of a single target cost is within a reasonable range.

4 Conclusions

This paper preliminarily discusses the problem of determining the charging location of mobile charging vehicles and new energy vehicles, and designs corresponding examples for theoretical analysis. However, there are certain limitations in the construction of the current model, mainly reflected in focusing only on the mobile charging vehicle as a power supply method, and assuming that the user request is a fixed state. Although this simplification is convenient for model construction and analysis, in practical application scenarios, the user's charging demand is often affected by various factors, showing more complex dynamic characteristics.

Acknowledgments. This work was funded by National Natural Science Foundation of China (No. 72471026).

References

1. Atmaja, T.D., Mirdanies, M.: Electric vehicle mobile charging station dispatch algorithm. *Energy Procedia* **68**, 326–333 (2015)
2. Tang, P., He, F., Lin, X., et al.: Online-to-offline mobile charging system for electric vehicles: strategic planning and online operation. *Transp. Res. Part D: Transp. Environ.* **87**, 102522 (2020)
3. Zhang, X., Cao, Y., Peng, L., et al.: Mobile charging as a service: a reservation-based approach. *IEEE Trans. Autom. Sci. Eng.* **17**(4), 1976–1988 (2020)
4. Liu, Q., Li, J., Sun, X., et al.: Towards an efficient and real-time scheduling platform for mobile charging vehicles. In: *Algorithms and Architectures for Parallel Processing: 18th International Conference, ICA3PP 2018, Guangzhou, China, 15–17 November 2018, Proceedings, Part III* 18, pp. 402–416. Springer (2018)
5. Moghaddam, V., Ahmad, I., Habibi, D., et al.: Dispatch management of portable charging stations in electric vehicle networks. *ETransportation* **8**, 100112 (2021)
6. Răboacă, M.S., Băncescu, I., Preda, V., et al.: An optimization model for the temporary locations of mobile charging stations. *Mathematics* **8**(3), 453 (2020)
7. Qureshi, U., Ghosh, A., Panigrahi, B.K.: Scheduling and routing of mobile charging stations to charge electric vehicles in a smart-city. In: *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 1–7. IEEE (2020)
8. Cui, S., Yao, B., Chen, G., et al.: The multi-mode mobile charging service based on electric vehicle spatiotemporal distribution. *Energy* **198**, 117302 (2020)

9. Kabir, M.E., Sorkhoh, I., Moussa, B., et al.: Joint routing and scheduling of mobile charging infrastructure for v2v energy transfer. *IEEE Trans. Intell. Veh.* **6**(4), 736–746 (2021)
10. Li, G., Sun, Q., Boukhatem, L., et al.: Intelligent vehicle-to-vehicle charging navigation for mobile electric vehicles via VANET-based communication. *IEEE Access* **7**, 170888–170906 (2019)
11. Jeon, S., Choi, D.H.: Optimal energy management framework for truck-mounted mobile charging stations considering power distribution system operating conditions. *Sensors* **21**(8), 2798 (2021)



Analyzing User Needs of Autonomous Transportation Systems with Activity Theory and Knowledge Graph

Lai Zheng^(✉), Chunrong Bao, and Wei Wei

School of Transportation Science and Engineering, Harbin Institute of Technology,
Harbin 150000, China
zhenglai@hit.edu.cn

Abstract. Thanks to the progress of emerging technologies such as autonomous driving and vehicle networking, China's transportation system has been continuously and rapidly developing towards autonomy. Autonomous Transportation System (ATS) will integrate the new concepts of modern transportation on a basis of traditional Intelligent Transportation System (ITS), and is committed to achieving the great transportation concepts of collaborative services, intelligent travel and comprehensive transportation. This paper uses the decomposition model of activity theory to elicit the user needs of ATS from multiple dimensions and constructs a knowledge graph network of ATS user needs to understand and analyze. Combining the content of activity theory and the structure results of the knowledge graph for user needs, the core components of the transportation system and complex interaction among them are studied, which helps to expand the user needs of ATS from two dimensions of technology promotion and policy guidance, thereby promoting the development of ATS.

Keywords: ATS · User Needs · Activity Theory · Knowledge Graph

1 Introduction

At present, the field of autonomous transportation system (ATS) architecture is in the construction [1], and some scholars have conducted research on the user need framework of autonomous transportation systems and achieved phased results. Xu et al. modeled the ATS logic framework for road autonomous driving scenarios and conducted reliability analysis and evaluation [2]; Zhang et al. used complex network theory to divide autonomous transportation systems between generations [3].

User needs are the foundation and service objectives for the construction of an ATS architecture. Researches can effectively guide the development and evolution of transportation systems. Based on this background, Zheng et al. utilized need reuse technology to establish a relatively complete list of user needs for ATS applicable to China's current stage of development, which provides a data base for researches [4, 5].

Dimitrakopoulos et al. mentioned in their book that the model of activity theory may be applied to the researches of complex traffic user needs [6]; Through the framework of activity theory, complex traffic scenarios can be approached from the perspective of various activity elements [7, 8]. In addition, the knowledge graph has been widely used in various fields in recent years. Nandi and Kou combined with the personalized needs system functions recommended the appropriate knowledge graph [9, 10]. Sakurai based on users' subscription on the music platform realized the recommendation to meet various user preferences through the knowledge graph constructed [11].

This paper is based on the idea of introducing activity theory into traffic system by Dimitrakopoulos [6], combined with the ATS user need list by Zheng [4], to construct a knowledge graph network of ATS user needs. By visualizing user needs with the help of a knowledge graph, it can help to intuitively understand the impact mechanisms of various components in complex transportation systems.

2 Activity Theory in ATS User Needs

2.1 Activity Theory

Activity theory is a philosophical framework that aims to explain the process of integration and interaction between humans and the external environment. The purpose is to guide activities into actions, which in turn constitute complex operations. The structure diagram of human activities is shown in Fig. 1.

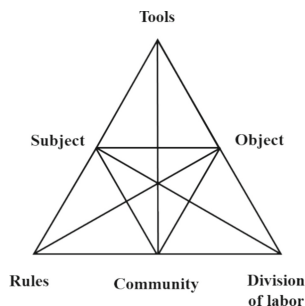


Fig. 1. Diagram of the correlation matrix

This activity structure includes six elements: subject, object, community, tools, rule and division of labor: Subject is the individual or organization initiating the activity; Object is the object of behavior; Community is a participant who shares common goals; Tools are the means to achieve goals; Rules are conditional restrictions; The division of labor represents the proportion of work. By utilizing the decomposition concept of activity content, complex activity systems can be divided and analyzed.

2.2 Autonomous Transportation Systems

ATS, based on the combination logic of autonomous perception, learning, decision and response, aim to be safe, efficient, convenient, green, and economical.

Perception is the process in which a transportation system perceives the world through sensors or communication technologies; Learning is the process in which the transportation system processes and analyzes the received data; Decision making is the process in which ATS forms an optimized execution plan; Response is the ability of the transportation system to respond to received instructions to complete transportation services, representing the results of the operation of ATS. The four components of perception, learning, decision-making and response complement each other to form a complex functional network of the ATS.

In order to better guide the development of ATS, it is necessary to design a system framework, to support the technological research. User needs are the important basis for the construction of the framework of ATS. The results of researches about user needs can provide a basis for the architecture design of it.

2.3 User Needs Elicitation Framework

Using activity theory, the complex traffic scene under ATS can be dealt with from the perspective of various activity elements in the activity system, the components of the complex traffic system can be decomposed.

Considering ATS as the activity system for researching, the subject is the user of systems, namely the driver; The main objectives of ATS are safety, efficiency, convenience, green, and economy; The community refers to other traffic participants in the autonomous transportation system, except for drivers. The subject and community are the users of the researches; Tools are intelligent devices used by user entities, such as communication devices, roadside learning devices; The rules include traffic regulations, potential social norm, and relevant policies; The division of labor refers to the proportion of division of labor undertaken by different users.

For the activity system under the ATS framework, the drivers and other traffic participants are the users of ATS. So we integrate and simplify the decomposition mode of activity system into three action units: using action unit, constraining action unit and division of labor action unit, as shown in the Fig. 2.

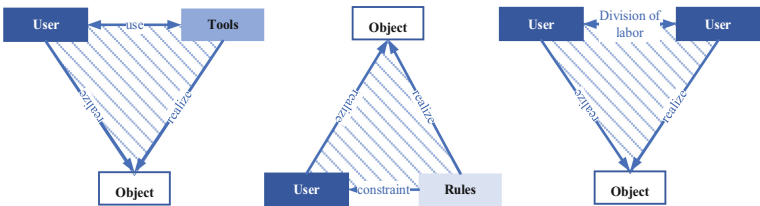


Fig. 2. Action decomposition of ATS

3 Statistical Analysis of ATS User Needs

3.1 ATS User Needs List

The user need list of ATS provides a detailed description of need content, including 294 comprehensive user needs with the tool using that may arise during the development of China's transportation system, which is suitable for the decomposition mode of three action units, as shown in the Table 1.

Table 1. Specific example of user needs list

Number	User needs description
1	Drivers should be able to access traffic regulation information
2	Drivers should be able to access meteorological information
3	Drivers should be able to determine whether the condition of the motor vehicle is normal
4	Drivers should be able to move quickly while ensuring safety
5	Drivers should be able to receive instructions from the vehicle or passengers
6	Drivers should be able to detect other vehicles in the vicinity of the vehicle and determine their driving status

According to the division mode of three units above, focusing on the use of tools, the division of labor among users and the constraints of rules in the needs, dividing the user needs in the list into activity elements.

For example, for the need that “drivers should be able to obtain meteorological information”. It can be extracted the drivers as users, the communication equipment as the tool. The division of labor between drivers and traffic information suppliers may be involved, and the implementation of requirements will be constrained by laws, policies, and potential social norms.

3.2 Judgment of Important Nodes in User Needs

The interaction relationship among user needs is essentially a weighted directed complex network. The various activity elements in transportation activity system are the nodes in the complex network of user needs.

However, the interaction relationships of the transportation system are complex, and many activity elements with low connectivity and low weight have little impact on the formation and development of the ATS user needs network, greatly disrupting the researches vision, especially for users and tools. Therefore, it is particularly important to identify these activity elements and their high-frequency interaction relationships.

Through extensive researches, the main indicators for identifying important nodes in the weighted directed complex networks include node strength, weighted coreness, weighted closeness centrality [12, 13]. The node strength represents the frequency of

contact between this node and other nodes. The more important a node is, the more frequent it is with other nodes, the higher the node strength. The weighted coreness represents the distance between a node and the core element in the network, and the closer it is to the core element, the higher the weighted coreness. For nodes, there are formulas for their node strength and weighted coreness.

$$s_i = \sum_{j=1}^n w_{ij} \tag{1}$$

where w_{ij} is the weight of the directed path from node i to node j , s_i is the point strength of the node i .

$$k_i^W = (k_i^\alpha s_i^\beta)^{\frac{1}{\alpha+\beta}} \tag{2}$$

where k_i is the degree of the node i , α, β are the adjustable parameters, $\alpha = 0.4, \beta = 0.6$, k_i^W is the weighted coreness of the node i .

The weighted coreness of nodes for users and tools are shown in the Table 2.

Table 2. Calculation results of weighted coreness of activity elements

Name	Type	Weighted coreness
Transportation center operators	User	58.785
Driver	User	55.853
Traffic manager	User	50.650
Transportation information supplier	User	45.295
Transport infrastructure manufacturer and operator	User	33.767
Emergency manager	User	23.143
Transport infrastructure maintainer	User	18.379
Traffic information collector	User	12.723
Transport infrastructure status monitor	User	10.506
Traffic product service provider manager	User	7.579
Environmental and climate information center operator	User	7.579
Transportation product service provider	User	7.339
Environmental climate information collector	User	6.507
Traffic facility manager	User	5.800
Vehicle repair and maintenance provider	User	3.565
Environmental climate information provider	User	3.000
Communication equipment	Tool	50.063
Management department response equipment	Tool	20.366

(continued)

Table 2. (continued)

Name	Type	Weighted coreness
Management department learning equipment	Tool	17.741
Roadside perception equipment	Tool	10.048
Management department decision-making equipment	Tool	8.871
Information storage management equipment	Tool	7.114
Information publishing equipment	Tool	6.507
Vehicle response equipment	Tool	5.474
Vehicle perception equipment	Tool	4.660
Maintenance equipment	Tool	4.076
Monitoring equipment	Tool	3.214
Vehicle learning equipment	Tool	2.930
Vehicle decision-making equipment	Tool	1.516

4 Knowledge Graph Building of ATS User Needs

4.1 Description of Knowledge Entity Relationships

The selected activity elements are shown in the Table 3, which will be used to construct the knowledge graph.

4.2 Analysis of ATS User Needs Based on Knowledge Graph

The ATS user need knowledge graph is shown in the Fig. 3. According to the construction result, tools are the core means for users to achieve goal orientation. Users have the high use of various tools, with the most frequent use of communication equipment and management department response equipment, far exceeding the use of other devices. It shows that information exchange and effective response to decision-making are the most important parts of the ATS system function. Communication equipment is mainly used by drivers, traffic center operators and traffic information suppliers, while management department response equipment is mainly used by traffic center operators and traffic managers. The use of vehicle response devices belongs to drivers, the use of management department learning equipment belongs to traffic center operators. The usage proportion of other equipment is relatively low.

Table 3. Selected knowledge entities for ATS

Type	Name of knowledge entities
User	Transportation center operators
	Driver
	Traffic manager
	Transportation information supplier
	Transport infrastructure manufacturer and operator
	Emergency manager
	Transport infrastructure maintainer
	Traffic information collector
	Transport infrastructure status monitor
Tool	Communication equipment
	Management department response equipment
	Management department learning equipment
	Roadside perception equipment
	Management department decision-making equipment
	Information storage management equipment
	Information publishing equipment
	Vehicle response equipment
Rule	Traffic regulation
	Policy
	Potential social norm

Rules are important constraints that restrict user behavior in transportation systems, setting boundaries for user activities and behavioral norms, promoting the orderly operation of autonomous transportation systems under social and cultural constraints. Almost all needs need to be constrained by laws, policies, and potential social norms.

The division of labor among users is the core operating mechanism for the development of transportation systems towards collaboration and autonomy, representing the orderly allocation of transportation resources and the organic combination of regulatory measures, which plays an important role in promoting the achievement of goals in autonomous transportation systems.

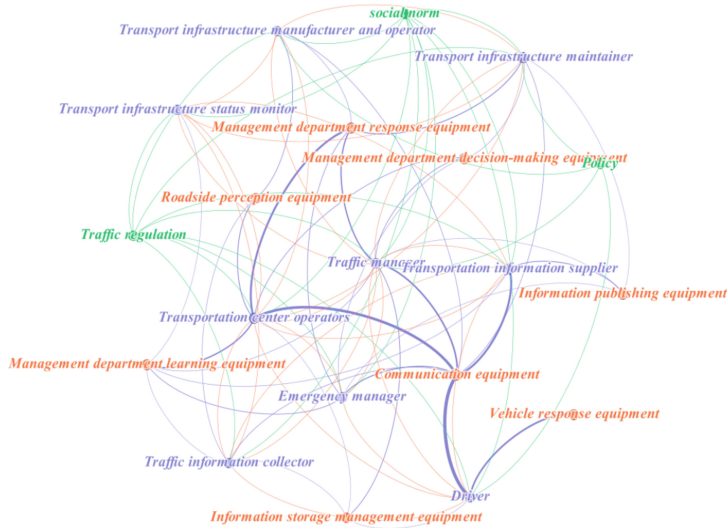


Fig. 3. Knowledge graph of user needs of ATS

5 Conclusion

The paper applies activity theory to ATS, analyzes the user needs of ATS from different dimensions and visualizes the results with knowledge graph, to show the interaction relationship between the user needs activity elements. The main research objective is to expand the application field of activity theory and knowledge graph in transport systems and further promote the construction of self-driving transport systems.

The paper improves the original decomposition model of activity theory, dividing the content of user needs into three action units: using action unit, constraining action unit and division of labor action unit, and discovers two core tools of ATS under the framework of activity theory through the combination of activity theory and complex network theory: communication equipment and management department response equipment. And the knowledge graph visualizes the interaction between users and helps to show the process of information collection and transfer, data processing and analysis, decision making and implementation.

However, there are still some limitations that need to be overcome in future work. Firstly, there is a certain degree of subjectivity in identifying elements according to the user need list. With the development of technology, the identification of elements should become more accurate; Secondly, with the changes of ATS service levels, some user needs will also evolve, and further researches is needed on the evolution of user needs.

Acknowledgements. The support from National Key R&D Program of China (2020YFB1600400) is acknowledged.

References

1. Guo, J.J., Bilal, M., Qiu, Y.Y.: Survey on digital twins for internet of vehicles: fundamentals, challenges, and opportunities. *Digital Commun. Netw.* **10**(2), 237–247 (2024)
2. Xu, G.M., Liu, X.Y., Zhong, L.H.: Reliability analysis and evaluation of autonomous transportation system logic architecture. *J. Railway Sci. Eng.* **10**(14), 1–9 (2022). (in Chinese)
3. Zhang, L.M., Xiao, Y., Jiang, S., Xiong, C.: Researches on the evolution model of autonomous transportation system based on complex networks. In: *Proceedings of the 16th China Intelligent Transportation Annual Meeting*, pp. 141–148. China Intelligent Transportation Systems Association, Changsha (2021). (in Chinese)
4. Zheng, L., Wei, W., Cai, M.: Eliciting user needs of intelligent transportation systems in the context of autonomous driving: an application of activity theory. In: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 755–760. IEEE, Macau (2022)
5. Zheng, L., Wei, W., Cai, M.: Researches on user need of intelligent transportation system for autonomous transportation. *Technol. Econ. Areas Commun.* **24**(02), 1–7 (2022). (in Chinese)
6. Dimitrakopoulos, G., Uden, L., Varlamis, L.: *The Future of Intelligent Transport Systems*, 1st edn. Elsevier, Netherlands (2020)
7. Li, Y.: Leontievs activity theory. *J. Psychol.* **02**, 233–241 (1979). (in Chinese)
8. Lv, J.J., Liu, F., Shi, X.L.: Analysis of the development and application of activity theory. *Modern Educ. Technol.* **01**, 8–14 (2007). (in Chinese)
9. Nandi, S., Dutta, A.: Local closeness gravity model to identify the vital nodes in complex networks. In: *2024 16th International Conference on Communication Systems & Networks*, pp. 864–872. IEEE, Bengaluru (2024)
10. Kou, Y.Q., Guo, Y.X.: Knowledge map construction and demand mining of automobile media user evaluation. *Chin. Automobile* **10**, 8–16 (2024). (in Chinese)
11. Sakurai, K., Togo, R.: Explainable artist recommendation based on reinforcement knowledge graph exploration. In: *International Workshop on Advanced Imaging Technology*, vol. 12177, pp. 91–96 (2022)
12. Zhang, Y., Xiao, G.: How to implement a knowledge graph completeness assessment with the guidance of user requirements. *J. Syst. Eng. Electron.* **35**(3), 679–688 (2024)
13. Liu, Q., Li, Y., Duan, H.: Overview of knowledge graph construction technology. *Comput. Res. Develop.* **53**(03), 582–600 (2016)



Dynamic Performance Prediction and Fluctuation Analysis Using Momentum Modeling in Competitive Environments

Enxin Gu^(✉)

Beijing Jiaotong University, Beijing, China
22251100@bjtu.edu.cn

Abstract. Accurate prediction of performance fluctuations in competitive environments is essential for optimizing strategies and decision-making processes. This study introduces a computational framework employing Principal Component Analysis (PCA) and regression techniques to evaluate dynamic performance and momentum impact. The proposed model quantifies the influence of critical factors, integrating data-driven insights to predict outcome variations. Experimental results indicate a strong correlation between identified momentum shifts and performance success, with the fluctuation prediction model achieving over 95% accuracy. The framework's applicability is further extended to diverse competitive scenarios, demonstrating its robustness and generalizability in predictive modeling of performance dynamics. The analysis results indicate that the above models have good universality.

Keywords: Decision Tree · AHP-CRITIC · Factor Analysis · Pearson Correlation

1 Introduction

The accurate prediction of performance fluctuations in competitive environments has gained increasing attention in academic and practical fields. Performance variations during a match provide valuable data for analyzing influencing factors, which can guide the optimization of strategies. With advancements in computational modeling, researchers are developing methods to simulate match dynamics, evaluate player performance, and predict outcomes [1–3]. These models offer significant potential to provide insights for optimizing training regimens and enhancing overall performance.

This study addresses the following problems:

Develop a model to determine which player is performing better at a given moment and provide a quantitative and visual representation of their performance. Analyze the role of “momentum” in tennis matches using the model and explain the results in a way that can convince tennis coaches. Combine match data to establish a fluctuation prediction model, identify the key factors affecting performance deviations. Extend the model to a general domain, evaluate its predictive performance and applicability.

2 Data

Dataset Wimbledon_*featured_matches* contains data for every point after the first two rounds of the 2023 Wimbledon men’s singles. Each row consists of the competition identification code, player name, elapsed time from the start of the competition to the current point, the number of sets, innings, and scores in the competition, as well as the competition activity data of the two players. And the dataset *data_dictionary* contains an explanation of the competition data of the two contestants.

speed_mph i.e. speed of serve is one of the important indicators for measuring athlete fatigue. We found some missing data, so interpolation processing is needed first. We initially planned to utilize the correlation between these four variables. We initially planned to use the correlation between variables *speed_mph*, *serve_width*, *serve_depth*, *return_depth* to regress *speed_mph*, but found that the missing data for *return_depth* was also very severe. Therefore, we ultimately decided to first estimate the missing value of *speed_mph* based on decision trees using data *serve_width* and *serve_depth*, and estimate the remaining missing values using cubic spline interpolation.

In order to comprehensively understand tennis culture and conduct subsequent model universality research, we collected additional data in addition to the original data to comprehensively understand tennis. We also listed the data sources and purposes in Table 1 as required.

Table 1. Data source collation.

Data source website	Data Usage
http://itf.uberflip.com/	The distribution of tennis players
https://www.wtatennis.com/rankings/singles/	Current women’s tennis rankings

3 Method

3.1 Performance Evaluation Model for Men’s Singles Tennis

To evaluate which player performs better at a given moment during a tennis match, we constructed a model based on Principal Component Analysis (PCA). Early in the match, there might be insufficient data to accurately assess player performance, so we selected “game” as the minimum unit for analysis. This approach accounts for the tennis scoring system, where serving often gives a player a significant advantage. Therefore, two separate models—one for the server and one for the receiver—were constructed using PCA to evaluate player performance more effectively [4].

The performance of an athlete on the tennis court is a comprehensive reflection of their various technical applications and on-site state. In order to evaluate the on-site performance of men’s tennis matches, we need to determine indicators that can represent men’s tennis skills. Based on dataset Wimbledon_*featured_matches*, we obtained a

total of 12 original indicators through literature search, data analysis, and other means, as follows [5]: ace(X_1), double_fault(X_2), serve_1st_ok(X_3), serve_1st_won(X_4), serve_2nd_won(X_5), serve_speed(X_6), break_pt_won(X_7), break_pt_missed(X_8), net_pt_won(X_9), winner(X_10), unf_err(X_11), distance_run(X_12).

For the server scoring model, we extracted the following 10 relevant indicators from the original indicators for analysis: X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_10. For the receiver scoring model, we extracted the following 6 relevant indicators from the original indicators for analysis: X_7, X_8, X_9, X_10, X_11, X_12.

Using factor analysis, we assessed whether the selected indicators were suitable for PCA. The Kaiser-Meyer-Olkin (KMO) and Bartlett's sphericity tests indicated that the dataset was appropriate for factor analysis. We then extracted three common factors, which explained 80.38% of the variance in player performance, ensuring minimal loss of information from the original data. Based on the above analysis, we have constructed formula Eq. (1) as the scoring model for the server/receiver on the tennis court at a specific moment:

$$\begin{cases} F_i = \sum_i \alpha_i x_i \\ F_t = \sum_i w_i F_i \\ \forall i = 1, 2, \dots, 12 \\ \alpha, w_i \in R \end{cases} \quad (1)$$

3.2 Momentum Indicator Selection and Analysis in Tennis Matches

In the previous section, we selected 12 technical indicators to evaluate the overall performance of players in each game. However, when analyzing the influence of “momentum” on players, the focus shifts from the game level to the point-by-point basis within each match. We introduced psychological factors such as whether the player hit a winning shot, whether the shot was made during the opponent's service game, consecutive wins, service errors, and receiving errors. These psychological factors were represented by binary variables ($x_{1t}^i, x_{2t}^i, x_{3t}^i, x_{4t}^i, x_{5t}^i$ respectively). Using these variables, the performance of a player at any given moment is calculated as follows:

$$P_t^i = \omega_1 x_{1t}^i + \omega_2 x_{2t}^i + \omega_3 x_{3t}^i h(c_3^i) - \omega_4 x_{4t}^i h(c_4^i) - \omega_5 x_{5t}^i h(c_5^i) \quad (2)$$

When determining the specific weight index values, we use a combination of AHP and CRITIC methods, and the weight formula is shown in formula:

$$\omega_{\text{total}} = \alpha * \omega_{\text{AHP}} + (1 - \alpha) \omega_{\text{CRITIC}} (0 \leq \alpha \leq 1) \quad (3)$$

Based on the experience on formula Eq. (3), we set $\alpha = 0.6$, the weights obtained based on AHP and CRITIC methods are shown in Table 2.

The weight result obtained by substituting formula Eq. (3) is shown below. To consider the impact of continuous actions on the player's momentum, we define $h(c_j^i)$:

$$h(c_j^i) = (1 + m_j^i)^{c_j^i - 1}, \quad j = 1, 2, 3 \quad (4)$$

Table 2. Momentum factor meight

Notation	AHP(60%)	CRITIC(40%)
ω_1	0.074	0.0526
ω_2	0.195	0.224
ω_3	0.462	0.171
ω_4	0.195	0.355
ω_5	0.074	0.197

According to formula Eq. (8), when c_j^i takes values of 1 and 2 respectively, it represents that player i has won one small game and won two consecutive games. This definition makes the consecutive wins have a continuous impact on the player’s state, which is more in line with reality. We take the initial values of m_1^i, m_2^i, m_3^i as 0.5 (we will explore their effects in subsequent sensitivity analysis). In addition to exploring the impact of psychological factors on players’ momentum, we also focus on the important role that physical condition plays in the game’s momentum, and introduce the concept of fatigue factor based on running distance.

$$S_t^i = \frac{\sum_{t_w \in t} S_{t_w}^i}{\sum_{t_w \in T} S_{t_w}^i} \tag{5}$$

Among them, T represents the duration of the entire game, and t represents the specific duration of the start of the game at a certain moment. As stated in hypothesis 1, when t is set to 0, it represents the beginning of the game and the player’s fatigue factor is 0. When t = T, the game is over and the player’s fatigue factor is set to 1, which is consistent with the actual situation. And the performance of players undergoes unstable changes as the game progresses, but the momentum of players is the result of the superposition of their performance from the beginning of the game to a certain moment and the interaction between the current level of fatigue. Based on this, we have defined the final “momentum function” as shown in formula Eq. (6):

$$M_t^i = \left(1 - \omega_s S_t^i\right) * \sum_{t_w \in t} P_{t_w}^i \tag{6}$$

Among them, M_t^i represents the momentum of the i-th player at time t, and ω_s represents the fatigue coefficient (here we take $\omega_s = 0.5$).

3.3 Game Fluctuation Prediction Model and Key Indicators

In competitive tennis matches, the flow of the game often shifts between players, with momentum fluctuating throughout the match. Problem 3 requires us to establish a fluctuation prediction model to determine which factors most significantly influence these momentum shifts and advantage deviations. Traditional metrics such as serve scoring rate or serve speed may not fully capture the overall dynamics of the game. To address

this, we used single-factor analysis and stepwise discriminant analysis to build a model that predicts match outcomes. Given the inherent advantage held by the serving player, we developed separate models for the server and receiver to improve predictive accuracy.

To identify the key factors that influence match outcomes, we first tested the correlation between 12 technical indicators, ensuring that the variables were independent enough to conduct stepwise discriminant analysis. This approach allowed us to isolate the most relevant factors affecting match fluctuations, ensuring the stability of the prediction model.

When establishing a prediction model for the winner of a server, we conducted stepwise discriminant analysis on the following 10 relevant indicators from the original indicators: ace (X1), double_fault (X2), serve_1st_ok (X3), serve_1st_won(X4), serve_2nd_won(X5), serve_speed (X6), net_pt_won (X9), winner (X10), unf_err(X11), distance_run (X12). Four factors were ultimately selected. According to the basic principle of stepwise discriminant analysis and combined with practical situations, these four technical statistical items with standardized coefficients are the key factors for winning men's tennis singles matches (Table 3).

Table 3. Server-classification function coefficient.

factor	loss	win
serve_1st_ok	0.508	− 0.096
serve_1st_won	− 0.3093	0.585
serve_2nd_won	− 1.738	0.329
unf_err	0.663	− 0.125
constant	− 4.576	− 0.271

It can be seen that when the player is in the position of the server, the factor with the strongest impact on game fluctuations is serve_2nd_won, which is the “second serve scoring rate” Based on the discriminant results of the non-standardized discriminant function, establish the non-standardized discriminant equation for the server winning the game as follows:

$$Y = -0.096X_3 + 0.585X_4 + 0.329X_5 - 0.125X_{11} - 0.271 \quad (7)$$

Among them, M_t^i represents the momentum of the i -th player at time t , and ω_s represents the fatigue coefficient (here we take $\omega_s = 0.5$).

Similarly, when establishing a prediction model for the winner of a catcher, we conducted a stepwise discriminant analysis on the following six relevant indicators from the original indicators: break_pt_won (X7), break_pt_missed (X8), net_pt_won (X9), winner (X10), unf_err (X11), distance_run (X12). Two factors were ultimately selected (see Table 4).

From the table, it can be seen that when the player is in the position of the receiver, the factor with the strongest impact on game fluctuations is break_pt_won, which is

Table 4. Server-classification function coefficient.

factor	loss	win
break_pt_won	− 2.713	14.342
break_pt_missed	− 0.186	0.983
constant	− 0.727	− 17.313

the success rate of breaking. Based on the results of the non-standardized discriminant function, establish the non-standardized discriminant equation for the winner of the catcher’s game as follows:

$$Y = 14.342X_7 + 0.983X_8 - 17.313 \tag{8}$$

3.4 Model Evaluation and Generalization

Accuracy test of fluctuation prediction model: The volatility prediction model has high accuracy. To verify the predictive performance of the volatility prediction model, we normalized the specific data in the Wimbledon_featured_matches on a per round basis, resulting in a total of 1188 pieces of data. Then import it into the volatility prediction model for verification.

Model generalization testing: Considering that the dataset Wimbledon_featured_matches only provides data on every point after the first two rounds of the 2023 Wimbledon men’s singles, due to differences in factors such as male and female sexual physiology, there may be different performances in competition performance evaluation, momentum, and situation fluctuations.

In order to extend the model from special fields to general fields, evaluate the predictive performance and applicability of the model, we collected data from the top 64 women’s singles tournaments of the US Open from 2018 to 2021 and beyond for statistical analysis, and imported them into the tennis singles technical score model and tennis field fluctuation prediction model established in 2.1 and 2.3 for calculation.

4 Results

4.1 Performance Evaluation Model for Men’s Singles Tennis

According to the score function coefficients of male tennis technical indicators (see Table 5), we can obtain the scoring model formula for the three factors of the server as follows:

$$\begin{aligned} F_{s1} &= 0.181X_1 - 0.092X_2 + 0.253X_3 + 0.268X_4 - 0.110X_5 + 0.189X_6 + 0.091X_9 + 0.195X_{10} - 0.362X_{11} - 0.361X_{12} \\ F_{s2} &= -0.063X_1 + 0.294X_2 + 0.181X_3 - 0.003X_4 - 0.448X_5 + 0.257X_6 - 0.095X_9 - 0.191X_{10} + 0.357X_{11} + 0.026X_{12} \\ F_{s3} &= 0.429X_1 + 0.357X_2 - 0.229X_3 - 0.106X_4 + 0.163X_5 + 0.061X_6 - 0.304X_9 + 0.221X_{10} + 0.126X_{11} - 0.421X_{12} \end{aligned} \tag{9}$$

From the above analysis, it can be seen that the variance contribution rates of the three common factors are 50.8%, 27.9%, and 21.2%, respectively. Therefore, the tennis singles technical score model for the server is:

$$F_{st} = \frac{(0.323F_{s1} + 0.178F_{s2} + 0.135F_{s3})}{0.636} \quad (10)$$

The tennis singles score model for the receiver can be obtained as follows:

$$F_{rt} = \frac{(0.279F_{r1} + 0.193F_{r2} + 0.171F_{r3})}{0.643} \quad (11)$$

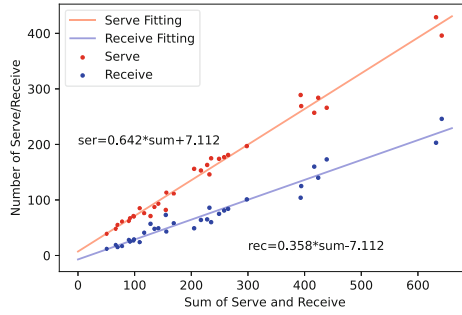


Fig. 1. Score Regression Equation

If the difficulty of serving and receiving is the same, then the slopes of both curves should be around 0.5. Now we use regression curve fitting to obtain slopes of 0.642 and 0.358 corresponding to serving and receiving scores, respectively. Therefore, when balancing the contribution of the two scoring methods to the player's scoring performance, we take the weights of serving and receiving scores as w_{st} and w_{rt} , respectively, as 0.358 and 0.642 (Fig. 1). From this, we can construct a scoring function to evaluate the performance of players as follows:

$$F = w_{st}x_{st}F_{st} + w_{rt}x_{rt}F_{rt} \quad (12)$$

4.2 Momentum Indicator Selection and Analysis in Tennis Matches

We chose a match to visualize whether “momentum” affects player scores, using each scoring store as the independent variable, the player's “momentum” at each moment as the independent variable, and the player's score at that moment as the dependent variable to draw a discrete graph as shown below [7].

It can be seen that there is a clear relationship between a player's “momentum” and their score. Player X has been falling behind and his momentum is negative. The turning point occurred in the x -th inning, and the turning point of the ball began to rise. To prove that the outcome of this game was not accidental, but an objective law, we calculated the

Pearson correlation coefficient between the “momentum” of each game and the score, as shown in formula Eq. (13):

$$r = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{13}$$

Among them, X and Y represent the player’s momentum and score.

Respectively, μ_X and μ_Y represent the mean, σ_X and σ_Y represent variance. The relationship between the value of r and the correlation size is shown in Table 5. By calculation, it can be concluded that $r = 0.85$. According to Table 5, there is a significant correlation between the two variables. Therefore, it can be concluded that a player’s “momentum” affects their score and plays an important role in the game.

Table 5. Pearson correlation coefficient evaluation

The value range of r	Correlation evaluation
0.0–0.2	No correlation
0.2–0.4	Weak correlation
0.4–0.8	Moderate correlation
0.8–1.0	Strong correlation

4.3 Game Fluctuation Prediction Model and Key Indicators

Based on the dataset data_dictionary dataset, we extracted relevant data from two matches, Alejandro Davidovich and Fokina Holger Rune, Daniil Medvedev and Marton Fucsovics, and processed them to obtain the relevant indicators required for the volatility prediction model. These indicators were imported into the model for calculation, and the visualization results of the volatility are shown below (Fig. 2).

The horizontal axis represents the number of innings in the game; The vertical axis represents the result obtained from non-standardized discriminant function discrimination of the serving party in the game, denoted as Y; The alternating appearance of two-colored scattered dots indicates that in the same game, two players alternate as serving players.

The closer the point is to the high indicator in the victory center of gravity level chart, the more likely the player is to win the competition, that is, the player gains an advantage at this time; On the contrary, the opponent gains an advantage. When the predicted winning player changes, it can be seen as a fluctuation in the competition and a deviation in the advantage.

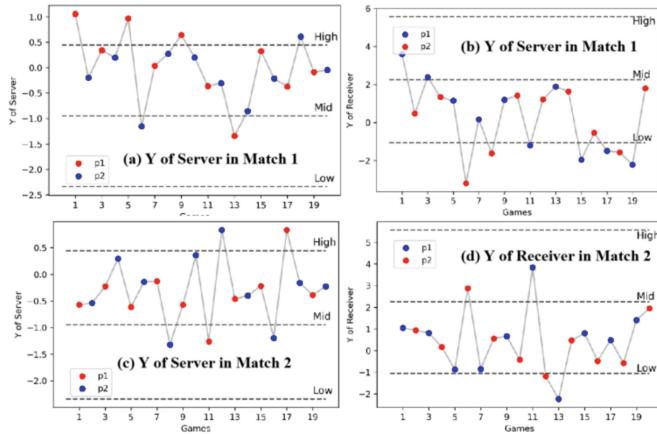


Fig. 2. Visualization of the Volatility

5 Conclusion

In this study, we developed a comprehensive model to evaluate tennis player performance, analyze momentum shifts, and predict match fluctuations based on data from the 2023 Wimbledon Men's Singles Tournament. Our approach effectively addresses the inherent advantages of serving, provides robust momentum analysis through a weighted indicator system, and offers high predictive accuracy for match outcomes.

By differentiating between serving and receiving roles, the model accounts for a key aspect of tennis that significantly impacts player performance. The inclusion of both subjective and objective weighting methods for evaluating momentum further strengthens the model's credibility. The model validation and its 95% accuracy rate underscore its potential for practical applications in coaching and performance analysis.

Looking ahead, the incorporation of additional factors, such as tournament significance and psychological data, will further enhance the model's predictive power. This model lays the foundation for a more nuanced understanding of player dynamics on the court, and with further refinement, it holds the promise of becoming an essential tool for optimizing performance in competitive tennis [8–10].

References

1. Constantinou, A., Fenton, N.: Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowl.-Based Syst.* **124**, 93–104 (2017)
2. Li, W., Lin, K., Fan, Z.: Analysis of player performance and game fluctuation based on LM-BP neural network and SVM. In: 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT). IEEE, pp. 330–338 (2024)
3. Guhan, T., Revathy, N.: EMLARDE tree: ensemble machine learning based random decorrelated extra decision tree for the forest cover type prediction. *SIViP* **18**, 8525–8536 (2024)
4. Gauriot, R., Page, L.: Does success breed success? A quasi-experiment on strategic momentum in dynamic contests. *Econ. J.* **129**(624), 3107–3136 (2019)

5. Zhang, Y.: The winning factors of the hard court competition for world-class men's singles tennis players. *J. Beijing Sport Univ.* (2009).(in Chinese)
6. Martello, S., Toth, P.: Linear assignment problems. In: *North-Holland Mathematics Studies*. North-Holland, vol. 132, pp. 259–282 (1987)
7. Meier, P., Flepp, R., Ruedisser, M., et al.: Separating psychological momentum from strategic momentum: evidence from men's professional tennis. *J. Econ. Psychol.* **78**, 102269 (2020)
8. Lin, B., Liu, C., Miao, D.: An improved decision tree algorithm based on boundary mixed attribute dependency. *Appl. Intell.* **54**, 2136–2153 (2024)
9. Alfonso, M., Menayo, R.: Induced variability during the tennis service practice affect the performance of every tennis player individually and specifically. *Euro. J. Hum. Mov.* **43**, 86–101 (2019)
10. Almarashi, A.M., Daniyal, M., Jamal, F.: A novel comparative study of NNAR approach with linear stochastic time series models in predicting tennis player's performance. *BMC Sports Sci. Med. Rehabil.* **16**(1), 28 (2024)



Behavior Prediction of Inland River Ferries Based on Encounter Scenarios

Jian Li^{1,4}, Yuanqiao Wen^{1,2,3}(✉), Liang Huang^{1,2,3}, Ahui Xu^{1,4}, and Yamin Huang^{1,2,3}

¹ State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology (WUT), Wuhan, China

wenyqwhut@foxmail.com

² Sanya Science and Education Innovation Park, Wuhan University of Technology, Wuhan, China

³ Intelligent Transportation Systems Research Centre, WUT, Wuhan, China

⁴ School of Navigation, WUT, Wuhan, China

Abstract. In inland river waters characterized by heavy traffic and a complex navigation environment, ferries frequently need to cross navigational channels due to the nature of ship activity. These situations are particularly challenging in the face of complex and dynamically changing encounter scenarios, which involve significant risks and uncertainties, thereby posing substantial challenges to the safety of navigation in these waters. To address these challenges, this paper proposes a scenario-constrained graph convolutional network (SC-GCN) for ferry behavior prediction. The model comprises an encoder module, a scene-labeling attention module, and a decoder module. The scene-labeling module employs expert knowledge, such as proficient boating techniques, to recognize and classify ferry encounter scenarios, thereby constructing scene labels. The encoder module subsequently extracts spatiotemporal interaction features of the ferry using a combination of GCN layers, LSTM layers, and scene label information. The decoder module then retains contextual information and adjusts the output as predictions are generated sequentially. To validate the SC-GCN model, Automatic Identification System (AIS) data from the Wuhan section of the Yangtze River was used to classify ferry encounter scenarios. The results demonstrate that the proposed scene-constrained graph convolutional network effectively predicts ferry behavior in encounter scenarios.

Keywords: Encounter Scenarios · Ferry Behavior · Ais Data

1 Introduction

Inland navigation, epitomized by the Yangtze River, is characterized by high traffic density, narrow channels, and a complex navigational environment. Within this challenging context, ferries frequently cross heavily trafficked waterways, creating intricate encounter scenarios with other vessels. This makes the safety and predictability of ferry navigation a priority for waterway management. Extracting encounter scenarios and

forecasting ferry behaviors in these situations is crucial for enhancing situational awareness in complex waterways and reducing collision risk. In behavior prediction research, trajectory forecasting models primarily employ deep learning techniques and physics-based trajectory projection. These models aim to predict vessels' future trajectories and behaviors based on historical data and current conditions. While physics-based models can leverage vessel handling and environmental parameters to forecast ferry behavior, they often suffer from low accuracy due to the dynamic and uncertain nature of environmental factors such as hydrology, complicating real-time prediction. Conversely, deep learning-based approaches effectively utilize AIS data to predict ferry behavior but often overlook the surrounding scene information, interaction dynamics with neighboring ships, and lack interpretability. Addressing these limitations can improve the precision and applicability of behavior prediction models in real-world scenarios.

To address these challenges, this paper presents a scene-constrained graph convolutional neural network (SC-GCN) model for predicting ferry behavior, as illustrated in Fig. 1. The model first establishes scene labels by classifying ferry encounter scenarios based on expert knowledge, including advanced navigation skills. These scene labels, along with AIS data from the ferry's encounter scenarios, are then fed into the encoder module to extract spatiotemporal interaction features. The decoder module subsequently preserves contextual information throughout the prediction process, adjusting outputs step by step to generate accurate behavior predictions. This approach enhances the model's predictive performance and provides a more nuanced understanding of ferry behavior in complex navigational environments.

The rest of this paper is organized as follows: Sect. 2 provides an overview of the current research status on encounter scene recognition and ferry behavior prediction. Sections 3 and 4 detail the proposed ferry encounter scene recognition method and the graph convolutional neural network (GCN) model for ferry behavior prediction based on scene constraints. Section 5 presents the validation of the proposed algorithms and models using AIS data from ferry encounter scenarios in the Wuhan section of the Yangtze River, followed by an analysis of the experimental results. Finally, Sect. 6 summarizes the conclusions of the study.

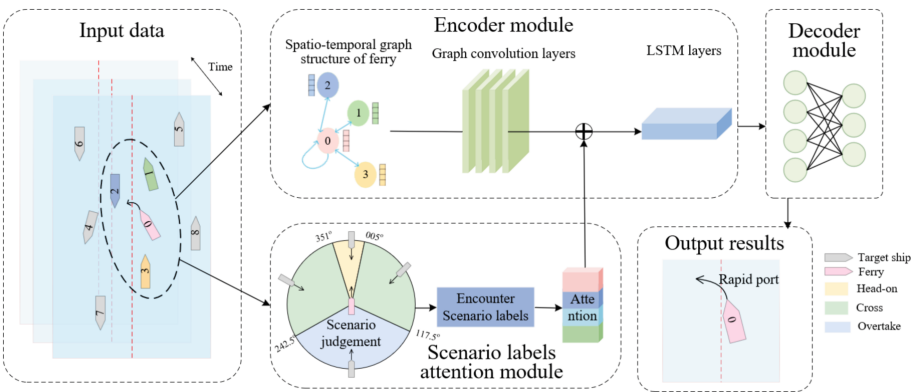


Fig. 1. SC-GCN model architecture

2 Related Work

2.1 Encounter Identification

A ship encounter refers to a navigational situation in which two vessels approach each other from different directions, posing a potential risk of collision. According to the International Regulations for Preventing Collisions at Sea (COLREGs) and the Regulations of the People's Republic of China for Collision Avoidance on Inland Waterways (RPCIW), ship encounters can be classified into three primary types: overtaking, crossing, and meeting. Previous studies have utilized various methodologies, including adherence to regulatory rules and the spatial-temporal characteristics of ship encounters, to identify such scenarios.

In the existing literature, several approaches have been proposed to detect and classify ship encounters. For instance, in [1], Automatic Identification System (AIS) data were utilized to calculate the distance of closest point of approach (DCPA) and time to closest point of approach (TCPA) between ferries and other vessels to identify potential encounters. Similarly, in [2], the authors screened AIS trajectory data using temporal and spatial overlap, calculated the relative values of state characteristics between two vessels, and overlapped their domains to identify encounter scenarios. In [3], the velocity obstacle method was employed to identify encounter scenarios.

However, it is important to note that vessels do not always exhibit behavior that complies with collision avoidance regulations in practice. To address this challenge, several researchers have employed traditional machine learning algorithms to identify and classify ship encounters while considering compliance with navigational rules. For example, in [4], the authors utilized the relative distance between two vessels, heading differences, and vessel state characteristics as inputs to a support vector machine (SVM), followed by Bayesian filtering to classify encounter scenarios. Additionally, [5] proposed an unsupervised hierarchical density-based spatial clustering algorithm to recognize and classify ship encounters based on distance similarity. However, this approach faced limitations in effectively handling scenarios without significant distance similarity.

To overcome these challenges, recent studies have increasingly leveraged deep learning methods to learn and extract patterns of ship encounters directly from data. In [6], a Convolutional Neural Network (CNN) was employed to recognize encounter scenarios, although no classification of the identified scenes was performed. In [7], the authors proposed a semi-supervised convolutional encoding-decoding network to both identify and classify encounter scenarios, providing a more comprehensive solution.

2.2 Behavior Prediction

Accurate and effective prediction of ship behavior is essential to ensure the safety of maritime navigation. Currently, ship behavior prediction methods can be broadly categorized into three approaches: those based on dynamic equations, traditional intelligent algorithms, and deep learning techniques.

Regarding behavior prediction based on kinetic equations, literature [8] proposes a Hidden Markov Model (HMM)-based method for predicting ferry behavior by analyzing the dynamic crossing process of ferries. In [9], a mathematical model of a ship's sailing trajectory is constructed based on the principles of inertial positioning, utilizing environmental and positional data during the voyage to predict ship behavior. Literature [10] employs a semantic network to represent ferry berthing and de-berthing behaviors, which is subsequently used to construct a dynamic Bayesian network for behavior prediction. However, methods based on dynamic equations have limited capacity for predicting complex behaviors and providing real-time predictions.

To address these limitations, some researchers have turned to traditional intelligent algorithms for ship behavior prediction. Literature [11] extracts navigational features such as position (longitude, latitude), speed over ground (SOG), and course over ground (COG) from AIS data, which are then input into a Back Propagation (BP) neural network to predict ship behaviors. Literature [12] further improves on the BP neural network by introducing a double delta function to mitigate prediction errors caused by abrupt changes in ship heading, particularly near zero degrees.

In terms of behavior prediction using deep learning techniques, literature [13] presents a hybrid approach that integrates a ferry encounter scene recognition algorithm with a kinematic neural network model to predict ferry trajectories. Additionally, literature [14] uses an improved Seq2Seq model to predict the short-term behavior sequences of ships. In this paper, a ferry behavior prediction method based on encounter scenarios, incorporating expert knowledge, is proposed and applied to inland waterways.

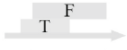
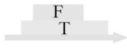
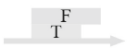
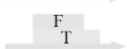
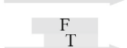
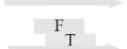
3 Encounter Scenario Identification

3.1 AIS Data Processing and Spatial and Temporal Analyses

The AIS data utilized in this study comprises the information transmitted and received by the ship Automatic Identification System (AIS). Due to factors such as signal fluctuations of shipboard positioning equipment, the AIS data contains numerous anomalies, including noise and redundant entries. To enhance the accuracy of subsequent ferry behavior prediction, it is essential to pre-process the AIS trajectory data. Firstly, anomalous data such as duplicate entries, abnormal speeds, steering irregularities, and drift points are removed from the AIS trajectory data. Secondly, since this study focuses on ship behavior within ferry encounter scenarios, trajectory data with speeds below 2 knots, such as during berthing and departure, are excluded. Subsequently, a maximum distance threshold of 1 km and a maximum time interval of 30 min are set to segment the ship trajectories. Finally, to address the issue of sparse and uneven trajectory distribution, a linear interpolation method is applied based on the results of trajectory segmentation to improve the completeness of the trajectory data.

Encounter scenarios represent the spatiotemporal evolution of ships, and therefore, prior to identifying encounter situations, it is necessary to filter out vessel data that overlaps spatiotemporally with the ferry in the study area. For spatial overlap, based on the characteristics of inland waterways, ferry navigation, and the ship domain model, a distance threshold of $D = 500\text{ m}$ [2] is selected to identify vessels spatially overlapping with the ferry. For temporal overlap, the ALLEN temporal model identifies nine possible temporal overlap relationships between the ferry and target vessels, with an overlap duration $\Delta T > 1\text{ s}$, as shown in Table 1. Here, t_{F1} and t_{Fn} represent the start and end times of the ferry, while t_{T1} and t_{Tn} denote the start and end times of the target vessel. This spatiotemporal filtering process effectively eliminates a significant amount of trajectory data that does not involve potential encounter scenarios.

Table 1. Time overlap relationship

Time overlap diagram	Mathematical constraint relations
	$t_{T1} < t_{F1} < t_{Tn} < t_{Fn}$
	$t_{T1} < t_{F1} < t_{Fn} = t_{Tn}$
	$t_{T1} < t_{F1} < t_{Fn} = t_{Tn}$
	$t_{T1} = t_{F1} < t_{Tn} < t_{Fn}$
	$t_{T1} = t_{F1} < t_{Tn} = t_{Fn}$
	$t_{T1} = t_{F1} < t_{Fn} < t_{Tn}$
	$t_{F1} < t_{T1} < t_{Tn} < t_{Fn}$
	$t_{F1} < t_{T1} < t_{Tn} = t_{Fn}$
	$t_{F1} < t_{T1} < t_{Fn} < t_{Tn}$

3.2 Ferry Encounter Scenario Identification

In accordance with the International Regulations for Preventing Collisions at Sea, 1972 (COLREGs) and the Inland Waterway Collision Avoidance Rules of the People’s Republic of China, this study categorizes ship encounter scenarios into three primary types: overtaking, crossing, and head-on encounters. Figure 2 illustrates these encounter scenarios, with a defined encounter radius of $R = 500\text{ m}$. Table 2 summarizes the distinguishing characteristics of each encounter type for ferries. The ferry’s heading is denoted by C_F , speed by V_F , Distance of Closest Point of Approach (DCPA), and Time to Closest Point of Approach (TCPA), with reference to the target vessel’s position at the moment of closest approach. Additionally, B_T represents the relative bearing of the target vessel in relation to the ferry, R signifies the distance between the two vessels, ΔC indicates the heading difference, while C_T and V_T refer to the target vessel’s heading and speed, respectively.

Table 2. Constraints on each encounter scenario

Classification of encounter scenarios	Constraint relations
Ferry overtaking target vessel	$\begin{cases} 000^\circ \leq B_T \leq 027.5^\circ \\ 297.5^\circ \leq B_T \leq 360^\circ \\ V_F > V_T * \cos \Delta C \\ R \leq 400m \end{cases}$
Target vessel overtaking ferry	$\begin{cases} 117.5^\circ \leq B_T \leq 242.5^\circ \\ V_F < V_T * \cos \Delta C \\ R \leq 400m \end{cases}$
Head-on	$\begin{cases} 351^\circ \leq B_T \leq 360^\circ \\ 000^\circ \leq B_T \leq 005^\circ \\ 174^\circ \leq C_T - C_F \leq 186^\circ \\ TCPA > 0 \\ DCPA \leq 400m \end{cases}$
	$\begin{cases} 171^\circ \leq B_T \leq 185^\circ \\ 174^\circ \leq C_T - C_F \leq 186^\circ \\ TCPA > 0 \\ DCPA \leq 400m \end{cases}$
Cross to starboard	$\begin{cases} 000^\circ \leq B_T \leq 117.5^\circ \\ 000^\circ \leq C_T - C_F \leq 180^\circ \\ TCPA > 0 \\ DCPA \leq 400m \end{cases}$
Cross to port	$\begin{cases} 242.5^\circ \leq B_T \leq 360^\circ \\ 000^\circ \leq C_T - C_F \leq 180^\circ \\ TCPA > 0 \\ DCPA \leq 400m \end{cases}$

4 Scenario-Constrained Graph Convolutional Network Model

4.1 Definition of the Problem

Ship behavior refers to the activities of a vessel that occur as a result of both internal and external influences. In this study, ferry behavior is defined by the spatial and temporal characteristics of the continuous trajectory points of a ferry that maintains a consistent motion state over a sustained period. Based on the stop/move conceptual model, the

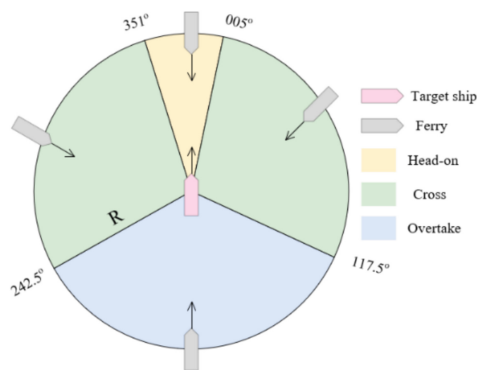


Fig. 2. Schematic diagram of ferry encounter scenarios

motion state of the ferry is categorized into two primary conditions: moving and stationary. Given that this paper focuses on the ferry behavior within a rendezvous scenario, only the moving state is considered. The ferry behavior is further classified into various categories, including uniform ahead, acceleration, deceleration, straight ahead, port, and starboard, by modeling motion parameters such as speed, heading, and acceleration, as illustrated in Fig. 3. For analysis, ferry behavior is broken down into two scalar components: heading and speed.

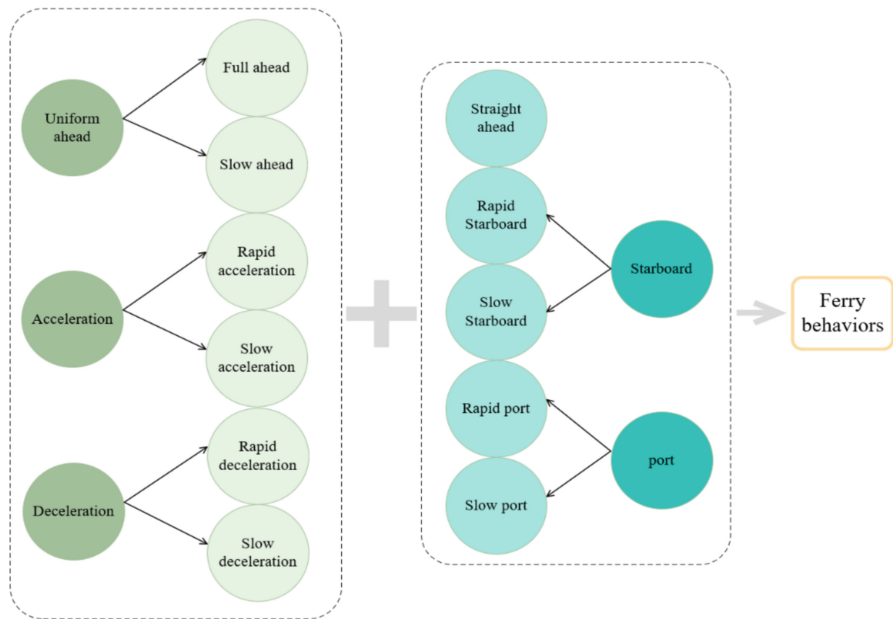


Fig. 3. Behavioral composition of the ferry

The sequence of behavioral features of a ferry can be represented as a series of trajectory points, each containing key parameters such as the ship's latitude, longitude, speed, heading, and other relevant data. This can be mathematically expressed as Eq. (1), where T_F denotes the complete sequence of behavioral features of the ferry, comprising n trajectory points. The i point in the trajectory is represented by Eq. (2), where t denotes time, and x , y , v , and c represent the ship's latitude, longitude, speed, and heading, respectively.

$$T_F = [P_{F1}, \dots, P_{Fn}] \quad (1)$$

$$P_{Fi} = [t_{Fi}, x_{Fi}, y_{Fi}, v_{Fi}, c_{Fi}] \quad (2)$$

For the sequence of ferry behavioral features, the i trajectory point is characterized by the following parameters: the corner change value $\Delta\alpha_{Fi} = c_{Fi} - c_{Fi-1}$, where $\Delta\alpha_{F\varepsilon} (> 0)$ represents the corner threshold; the steering acceleration $\alpha_{Fa} = \frac{\Delta\alpha_{Fi}}{t_{Fi} - t_{Fi-1}}$, with a steering acceleration threshold $\alpha_{Fa\varepsilon} (> 0)$; the velocity change value $\Delta v_{Fi} = v_{Fi} - v_{Fi-1}$, with a velocity threshold $v_{F\varepsilon} > 0$; the velocity difference threshold $\Delta v_{F\varepsilon} (> 0)$; and the acceleration $v_{Fa} = \frac{\Delta v_{Fi}}{t_{Fi} - t_{Fi-1}}$, with an acceleration threshold $v_{Fa\varepsilon} (> 0)$. These parameters are used to define and identify different behavioral patterns of the ferry, as summarized in Table 3.

In the ferry encounter scenario, the observed set of behavioral feature sequences $L_{obs} \in \mathcal{R}^{N \times P_{0N} \times T_{obs}}$ represents the behavioral feature sequences of N vessels within the observation period T_{obs} . The goal is to predict the ferry's behavioral feature sequences $L_{pred} \in \mathcal{R}^{1 \times P_0 \times T_{pred}}$ over the prediction horizon T_{pred} , where P_0 denotes the ferry's behavioral feature sequence, given as $\bar{P}_0^t = (\bar{x}_0^t, \bar{y}_0^t, \bar{v}_0^t, \bar{c}_0^t)$, for each time step $t \in \{1, \dots, T_{pred}\}$. It is assumed that the behavioral features follow a two-dimensional Gaussian distribution $\mathcal{N}(\mu_t, \sigma_t, \rho_t)$. Therefore, the predicted behavioral feature sequence for the ferry within the prediction period T_{pred} is modeled as a behavioral feature vector $\bar{P}_0^t = (\bar{x}_0^t, \bar{y}_0^t, \bar{v}_0^t, \bar{c}_0^t)$, which also follows a two-dimensional Gaussian distribution $\mathcal{N}(\mu_t, \sigma_t, \rho_t)$.

$$Loss(\mathcal{W}) = - \sum_{t=1}^{T_{pred}} \log \left(P \left(\bar{P}_0^t | \bar{\mu}_t, \bar{\sigma}_t, \bar{\rho}_t \right) \right) \quad (3)$$

The sequence of behavioral features is then predicted by minimizing the negative log-likelihood of the predicted sequence, with the objective function expressed in Eq. (3). The ferry's behavioral feature sequences are trained using this minimization-based approach to optimize the prediction accuracy.

4.2 Scenario-Constrained Graph Convolutional Network (SC-GCN) Model

The SC-GCN model proposed in this study consists of five main components: 1) Input Data: Graph data, such as the neighbor matrix, is constructed by extracting the features of ships in the encounter scenario from processed AIS data; 2) Encoder Module: The input feature matrix and neighbor matrix are encoded using graph convolutional layers

Table 3. Constraints on different behavioral characteristics of ferry

Ferry behavior	Behavioral characteristics constraints
Straight ahead	$ \Delta\alpha_{Fi} \leq \Delta\alpha_{F\varepsilon}$
Slow port	$ \Delta\alpha_{Fi} \geq \Delta\alpha_{F\varepsilon} \cap$ $\Delta\alpha_{Fi} < 0 \cap \alpha_{Fa} \leq \alpha_{Fa\varepsilon}$
Rapid port	$ \Delta\alpha_{Fi} \geq \Delta\alpha_{F\varepsilon} \cap$ $\Delta\alpha_{Fi} < 0 \cap \alpha_{Fa} > \alpha_{Fa\varepsilon}$
Slow Starboard	$ \Delta\alpha_{Fi} \geq \Delta\alpha_{F\varepsilon} \cap$ $\Delta\alpha_{Fi} > 0 \cap \alpha_{Fa} \leq \alpha_{Fa\varepsilon}$
Rapid Starboard	$ \Delta\alpha_{Fi} \geq \Delta\alpha_{F\varepsilon} \cap$ $\Delta\alpha_{Fi} > 0 \cap \alpha_{Fa} > \alpha_{Fa\varepsilon}$
Full ahead	$ \Delta v_{Fi} \leq \Delta v_{F\varepsilon} \cap v_{Fi} > v_{F\varepsilon}$
Slow ahead	$ \Delta v_{Fi} \leq \Delta v_{F\varepsilon} \cap v_{Fi} \leq v_{F\varepsilon}$
Rapid deceleration	$ \Delta v_{Fi} > \Delta v_{F\varepsilon} \cap$ $\Delta v_{Fi} < 0 \cap v_{Fa} > v_{Fa\varepsilon}$
Slow deceleration	$ \Delta v_{Fi} > \Delta v_{F\varepsilon} \cap$ $\Delta v_{Fi} < 0 \cap v_{Fa} \leq v_{Fa\varepsilon}$
Rapid acceleration	$ \Delta v_{Fi} > \Delta v_{F\varepsilon} \cap$ $\Delta v_{Fi} > 0 \cap v_{Fa} > v_{Fa\varepsilon}$
Slow acceleration	$ \Delta v_{Fi} > \Delta v_{F\varepsilon} \cap$ $\Delta v_{Fi} > 0 \cap v_{Fa} \leq v_{Fa\varepsilon}$

and LSTM layers; 3) Scenario Labelling Attention Module: Scene labeling information in the ferry behavior feature sequence is extracted using an attention mechanism, which adjusts the results obtained from the graph convolution in the encoder module by weighting the scene label information; 4) Decoder Module: The decoder module is responsible for mapping the context information, extracted by the attention mechanism, to the final predicted output; 5) Output Results: This component processes the final prediction results, evaluates the model's performance, and ensures interpretability. The model architecture is depicted in Fig. 1.

4.3 Encoder Module

- (1) Graph Convolution Layer: To capture the interactions between ships in the ferry encounter scenario, a spatio-temporal correlation graph is constructed, as illustrated in Fig. 4, followed by a graph convolution operation. This allows the dynamic interaction between the ferry and its neighboring vessels to be modeled as a directed graph structure $G = (V, E)$, where $V^t = \{V_0^t, \dots, V_m^t\}$ represents the set of nodes at time t , and $E^t = \{e_{o,i}^t\}_{(i=0,\dots,m)} \cup \{e_{i,0}^t\}_{(i=1,\dots,m)}$ represents the set of edges. Here, v_0 denotes the ferry node, and $v_i (i \neq 0)$ denotes the ferry's neighboring ships. The inputs to the graph convolution layer are the feature matrices H^t and adjacency

matrices A^t of the nodes, while the output represents the interaction feature information h_c in the ferry encounter scenario. The feature matrix at time t is given by Eq. (4), and the adjacency matrices are represented in Eqs. (5) and (6). In this study, x_0^t represents the longitude feature of the node, y_0^t denotes the latitude feature, v_0^t is the speed feature, and c_0^t corresponds to the heading feature of the node. Additionally, $d(V_0^t, V_j^t)$ represents the Euclidean distance from the ferry to a neighboring ship, while $d(V_i^t, V_0^t)$ represents the Euclidean distance from the neighboring ship to the ferry. In this paper, a star-directed graph with self-connections is considered. Spatially, the influence of neighboring ships on the ferry decreases as the distance increases. Therefore, the weights of the edges in the adjacency matrix are taken as the inverse of the Euclidean distances between the nodes, and the weight of the self-connected edges is set to 1.

$$H^t = \begin{bmatrix} V_0^t \\ V_1^t \\ \dots \\ V_m^t \end{bmatrix} = \begin{bmatrix} x_0^t & y_0^t & v_0^t & c_0^t \\ x_1^t & y_1^t & v_1^t & c_1^t \\ \dots & \dots & \dots & \dots \\ x_m^t & y_m^t & v_m^t & c_m^t \end{bmatrix} \quad (4)$$

$$A^t = \begin{bmatrix} e_{0,0}^t & e_{0,1}^t & \dots & e_{0,m-1}^t & e_{0,m}^t \\ e_{1,0}^t & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ e_{m-1,0}^t & 0 & \dots & 0 & 0 \\ e_{m,0}^t & 0 & \dots & 0 & 0 \end{bmatrix} \quad (5)$$

$$e_{i,j}^t = \begin{cases} 1, i = j = 0 \\ \frac{1}{d(V_0^t, V_j^t)}, j = 1, \dots, m \\ \frac{1}{d(V_i^t, V_0^t)}, i = 1, \dots, m \end{cases} \quad (6)$$

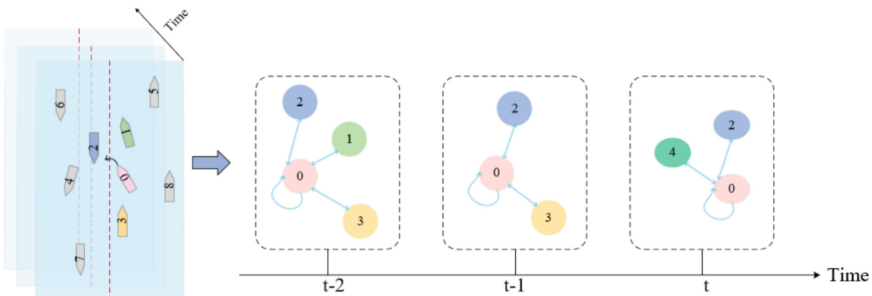


Fig. 4. Spatio-temporal correlation diagram of the ferry in the encounter scenario

- (2) LSTM Layer: This layer primarily utilizes the interaction feature information output from the graph convolution layer to capture the ferry's behavioral information in the encounter scenario over an extended time period. The structure of the LSTM unit consists of the forget gate f_t , input gate i_t , and output gate o_t , as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (7)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

where x_t represents the input at time t , h_{t-1} is the hidden state at the previous time step, and σ denotes the sigmoid activation function. W is the coefficient matrix, and b is the bias term. The updated memory cells and the final output of the LSTM layer are then expressed by Eqs. (10), (11), and (12). Here, c_{t-1} is the memory cell at the previous time step, c_t is the memory cell at the current time step, and h_t is the updated hidden state, which also serves as the final output of the LSTM layer.

$$q_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (10)$$

$$c_t = f_t \times c_{t-1} + i_t \times q_t \quad (11)$$

$$h_t = o_t * \tanh(c_t) \quad (12)$$

4.4 Scenario Labelling Attention Module

The scene label attention module provides additional contextual information for behavioral prediction, allowing the model to better understand the current sailing state of the ferry. This enables more accurate predictions of the ferry's future behavior, thereby enhancing both the model's prediction accuracy and interpretability. The calculation formulas for the attention weights are presented in Eqs. (13), (14), and (15). Where h_t is the input to the scene label attention module, containing the features of each node in 7 dimensions at time t , and e_t represents the attention weight corresponding to h_t . W is a learnable weight matrix used for the linear transformation of the input features, while b denotes the bias coefficients that are trained. α_t is the attention weight corresponding to h_t after applying the softmax function, and c represents the final output, which combines the context vector of all time step information.

$$e_t = v^T \cdot \tanh(W \cdot h_t + b) \quad (13)$$

$$\alpha_t = \text{softmax}(e_t) \quad (14)$$

$$c = \sum_t \alpha_t h_t \quad (15)$$

4.5 Decoder Module

The decoder module is primarily implemented using an LSTM decoder, which is capable of effectively retaining contextual information and adjusting the output during the step-by-step prediction generation process. This makes it particularly well-suited for the task of generating complex behavioral feature sequences, as described in this paper. The specific process is detailed in Eq. (16), Eq. (17), and Eq. (18). Here, h_T^{enc} and c_T^{enc} represent the hidden state and cellular states of the encoder at the final time step, which are passed to the decoder LSTM as the initial state of the decoder. At each time step t , the input to the decoder is the output from the previous step, while x_t is the input at the current time step. y_t is the final output of the decoder LSTM after passing through the fully connected layer, with W and b denoting the weights and biases of the fully connected layer, respectively.

$$h_0^{dec} = h_T^{enc}, c_0^{dec} = c_T^{enc} \quad (16)$$

$$(h_t^{dec}, c_t^{dec}) = LSTM(x_t, (h_{t-1}^{dec}, c_{t-1}^{dec})) \quad (17)$$

$$y_t = w \cdot h_t^{dec} + b \quad (18)$$

5 Experiments and Results

This study focuses on the confluence of the Wuhan section of the Yangtze River and the Han Jiang River, which features a complex ship traffic environment and numerous encounter scenarios (see Fig. 5, left). The ferry with MMSI number ‘413932548’ is chosen as the research object, with its activity trajectory data shown on the right side of Fig. 5. The ferry frequently traverses the confluence waters, experiencing complex encounter scenarios and rapid changes, making behavior prediction in such scenarios highly significant. Model training and algorithm validation were conducted using three days of ferry AIS data (from 1 to 3 January 2021) along with target vessel AIS data from encounter scenarios. After preprocessing, trajectory segmentation, interpolation, and spatio-temporal overlap screening, 108,723 ferry trajectory data points and 153,529 target vessel trajectory data points were obtained (see Fig. 5). The model was trained with a batch size of 32, 1000 training epochs, Adam optimizer, and a learning rate of 0.001, using the Pytorch framework and an NVIDIA GeForce RTX 3050 graphics card.

5.1 Statistical Analysis of Encounter Scenarios

Based on the extracted ship trajectory data, a total of 41,489 ferry encounter scenarios were identified, accounting for 38.2% of all ferry data. Of these, 16,025 scenarios (38.6%) were classified as one type of encounter, 23,744 scenarios (57.2%) as another, and 1,720 scenarios (4.1%) as a third type, as shown in Fig. 6. Ships operating in inland rivers primarily follow the upper and lower waterways, with ferries mainly crossing the channel for round-trip activities between fixed piers. As a result, encounter scenarios are relatively rare. The statistics indicate that the proportion of encounter scenarios is lowest, while crossing encounter scenarios are the most common, aligning with the typical behavior of ferries.

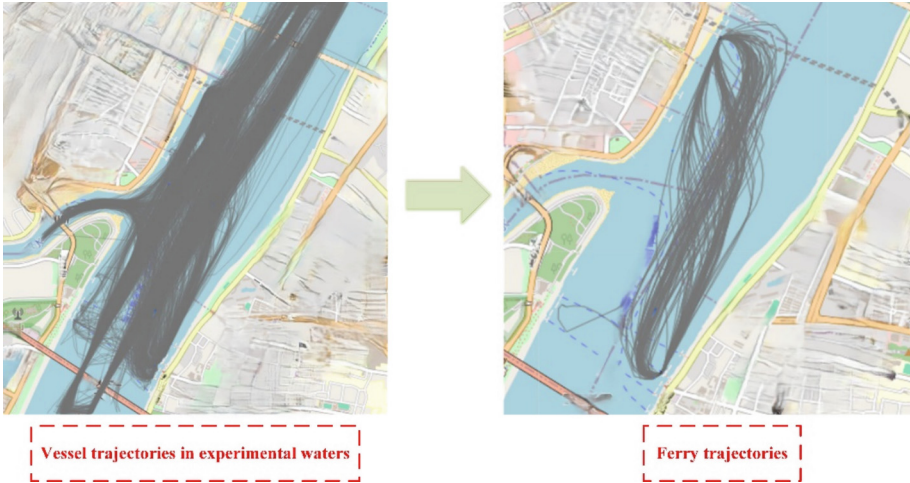


Fig. 5. Vessel trajectories in experimental waters

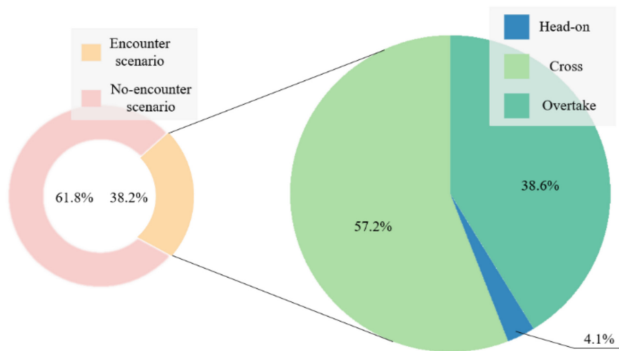


Fig. 6. Ferry Encounter Scenario Statistics

5.2 Analysis of Predicted Results

In this study, MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) are used as evaluation metrics for the prediction model, where smaller values of MAE and RMSE indicate better model performance. The dataset consists of the ferry encounter scene data for MMSI number ‘413932548’, with the SC-GCN model trained using an 80% training set and a 20% test set. The model’s loss function, shown in Fig. 7, decreases as the number of iterations increases, with lower loss values corresponding to higher accuracy. As shown in Fig. 8, during the initial phase (first 20 epochs), both the training and test losses decrease rapidly, indicating that the model quickly learns the data’s basic patterns. In the later stages (after about 100 epochs), both losses stabilize, and the gap between them is small, suggesting good model fitting without significant overfitting.

Although the test loss is slightly higher than the training loss, it remains within an acceptable range. Overall, the model learns efficiently, with both training and test losses reaching low values, indicating good model performance and generalization to the test data.

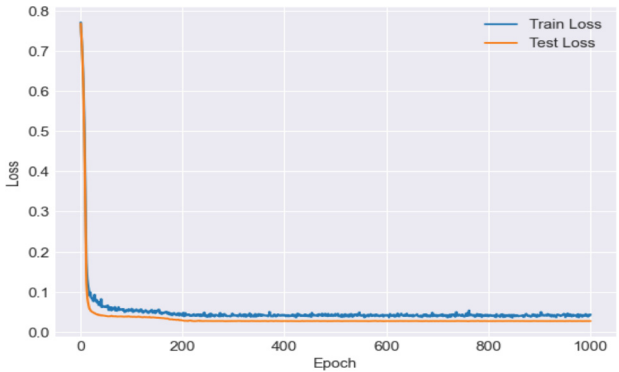


Fig. 7. Ferry Encounter Scenario Statistic

Due to the short duration and frequent encounters in the ferry encounter scenario, this paper randomly selects ferry encounter data from the dataset, specifically 180s of ferry data along with neighboring ship data, to assess the accuracy and generality of the model. The results are shown in Fig. 8-a. As observed in Fig. 8-a, the green trajectory line represents a high match between the predicted and actual trajectories, demonstrating that the SC-GCN model can effectively learn the complex dynamic interaction patterns in ferry encounter scenarios and accurately predict the ferry’s behavior. Figure 8-b presents the predicted and actual heading and speed values for 80 randomly selected time steps during the model testing phase. For heading prediction, the model aligns well with the actual values, particularly in periods of stable change. However, in the time intervals of 0–10, 30–40, and 50–60, the heading changes in the actual data are irregular and large, leading to poorer prediction accuracy in these sharply changing periods. Despite this, the general trend remains consistent with the actual behavior. Regarding speed prediction, the model maintains a small error range throughout most time periods, performing particularly well in areas with smoother speed changes, where it accurately captures the speed characteristics of the ferry.

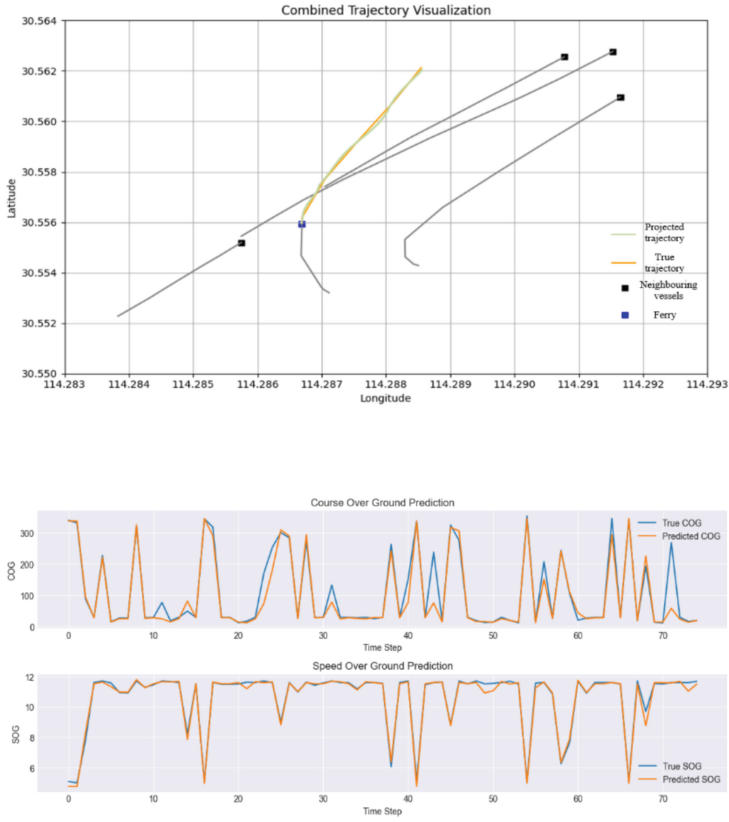


Fig. 8. (a) Visualization of trajectory prediction results (b) Visualization of heading and speed prediction results

6 Conclusion

This paper proposes a scene-constrained graph convolutional network (SC-GCN) model to predict ferry behavior in complex inland waterway encounter scenarios. The SC-GCN model achieves high-precision predictions by analyzing the spatial and temporal features of the encounter scene, constructing semantic labels, and extracting vessel behavior features from AIS data. It efficiently integrates scene information and ferry voyage data, modeling the dynamic interactions between the ferry and neighboring vessels through the graph convolutional layer. The Scene Label Attention Module further enhances the model by weighting different scene features, improving its ability to predict ferry behavior in various scenarios. Experimental results show that the SC-GCN model excels at capturing the interactions between ferries and other vessels, as well as dynamically changing behaviors in complex navigation areas, contributing to enhanced ferry navigation safety. Future research could focus on modeling interactions between different vessel types, especially in multi-vessel encounters and emergency avoidance scenarios, to improve the model's handling of unexpected events. Additionally, incorporating more environmental

factors, such as hydrological and meteorological data, could further enhance prediction accuracy in complex, dynamic environments.

Acknowledgement. This work is supported by the Research Program of Sanya City through Grant No. 2022KJCX36, the National Science Foundation of China (NSFC) through Grant No. 52072287.

References

1. Cai, M.Y., Zhang, J.F., Zhang, D., Yuan, X.L., Soares, C.G.: Collision risk analysis on ferry ships in Jiangsu Section of the Yangtze River based on AIS data. *Reliab. Eng. Syst. Safety* **215**, 107901 (2021)
2. Zhang, J.F., You, B., Cai, M.Y., Huang, L., Zhang, D.: Identification of behavior of ferry boat crossing channel and modeling of collision risk. *Navig. Chin.* **46**(2), 4653–4659 (2023). (in Chinese)
3. Yuan, X.L., Zhang, D., Zhang, J.F., Zhang, M.Y., Soares, C.G.: A novel real-time collision risk awareness method based on velocity obstacle considering uncertainties in ship dynamics. *Ocean Eng.* **220**, 108436 (2020)
4. Ma, J., Li, W.K., Zhang, C.W., Zhang, Y.: Ship encounter situation recognition by processing AIS data from traffic intersection waters. *Navig. Chin.* **44**(1), 68–74 (2021). (in Chinese)
5. Brian, M., Perera, L.P.: An AIS-based deep learning framework for regional ship behavior prediction. *Reliab. Eng. Syst. Safety* **215**, 107819 (2021)
6. Chen, X., Liu, Y.C., Kamalasudhan, A., Zhang, X.Y.: A ship movement classification based on Automatic Identification System (AIS) data using convolutional neural network. *Ocean Eng.* **218**, 108182 (2020)
7. Chen, X., Liu, Y.C., Kamalasudhan, A., Zhang, X.Y., Chen, J.H.: A semi-supervised deep learning model for ship encounter situation classification. *Ocean Eng.* **239**, 109824 (2021)
8. Cheng, T.T., Wu, Q., Wu, B., Zhang, M.Y.: Prediction of rationality for ferry crossing behavior based on multi hidden Markov model. In: 2019 5th International Conference on Transportation Information and Safety (ICTIS), Liverpool, UK, pp. 1382–1388 (2019)
9. Li, R.: Design of mathematical model for predicting ship's navigation trajectory. *Ship Sci. Technol.* **2020**(8), 13–15 (2020). (in Chinese)
10. Wen, Y.Q., Zhang, Y.M., Huang, L., Zhou, C.H., Xiao, C.S., Zhang, F.: Mechanism of ship behavior dynamic reasoning based on semantics. *Navig. Chin.* **42**(3), 34–39. (in Chinese)
11. Zhen, R., Jin, Y.X., Hu, Q.Y., Shi, C.J., Wang, S.Z.: Vessel behavior prediction based on AIS data and BP neural network. *Chin. Navig.* **40**(2), 6–10 (2017). (in Chinese)
12. Gao, T.H., Xu, L., Jin, L.J., Ge, B.: Vessel trajectory prediction considering difference between heading and data changes. *J. Transp. Syst. Eng. Inf. Technol.* **21**(1), 90–94 (2021). (in Chinese)
13. Zhu, M.D., Han, P.H., Wang, C.L., Robert, S., Zhang, H.X., Li, G.Y.: An AIS data-driven hybrid approach to ship trajectory prediction. *IEEE Trans. Syst. Man Cybern. Syst.* **55**(1), 96–109 (2024)
14. You, L., Han, X.W., He, Z.W., Xiao, S.Y., He, D., Pan, X.M.: Short-term AIS trajectory sequence prediction model based on improved Seq2Seq. *Comput. Sci.* **47**(9), 169–174 (2020). (in Chinese)



A Real-Time Railway Intrusion Detection and Multi-factor Collaborative Risk Assessment Method

Genwang Peng^{1,2} , Yang Gao^{1,2} , Zhiwei Cao¹ , Yong Qin^{1,3} ,
and Jianyuan Guo²

¹ State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University,
Beijing 100044, China

{penggenwang, gao_yang, zhiwei, yqin}@bjtu.edu.cn

² School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
jyguo@bjtu.edu.cn

³ Key Laboratory of Railway Industry of Proactive Safety and Risk Control,
Beijing Jiaotong University, Beijing 100044, China

Abstract. The safety risk due to intrusion in railways is a huge hazard that threatens the safety of train operations. Existing real-time detection algorithms can quickly and accurately detect intrusion. Since different intrusion scenarios have different levels of threat to train operation safety, drivers need to adopt different risk response strategies to avoid risks. There is a lack of a timely and reliable risk assessment method. Quantitative calculation of intrusion risk for different intrusion situations to help drivers make optimal decisions. This paper presents a real-time railway intrusion detection and multi-factor collaborative risk assessment method. This method consists of two parts: intrusion detection and risk assessment. Firstly, intrusion detection in the whole scene is realized by object detection algorithm and semantic segmentation algorithm. Secondly, this method establishes a risk assessment system that evaluates the risk of intrusion based on two major risk factors: the spatial location of the foreign object in relation to the track area, and the category of the foreign object. Experiments have proved that the intrusion detection algorithm proposed in this method can quickly and accurately realize the railway intrusion detection. The risk assessment system proposed in this method can scientifically, efficiently and real-time realize timely risk assessment of train operation.

Keywords: Intrusion detection · Semantic segmentation · On-board video · Risk assessment · Railway safety

1 Introduction

The safety risks caused by intrusion in railway seriously affect the efficiency and safety of railway systems [1, 2]. It can cause trains to be delayed and even lead to train scrapping or casualties. According to statistics, by the end of 2023, the operating mileage of China's

railway has reached 159,000 km. By combining fast and accurate real-time intrusion detection algorithms with a scientific and efficient risk assessment system, trains can be guaranteed to travel safely on such a large-scale railway network.

Currently, the intrusion detection is mostly based on real-time intelligent algorithms. The object detection algorithms are developed to be more mature and can quickly recognize foreign objects in the environment in front of the train. But for the complex and changing railway scene, there is still room for improvement in long-distance detection accuracy. In addition, most of the existing research focuses on intrusion detection, with less research on subsequent risk assessment, but the risk assessment link is essential. Timely and accurate risk assessment results can assist the driver to deal with the risk at the first time, which can effectively guarantee the operation safety of the train. Finally, current research lacks a rational risk assessment system. To ensure the accuracy of risk assessment, there is a need to develop an assessment system. The system can scientifically and comprehensively consider various risk factors and feedback accurate risk assessment results.

2 Related Work

2.1 Railway Intrusion Detection

There have more cases of intrusion detection through deep learning algorithms at domestic and international level, and the existing object detection algorithms have been developed more maturely. Ma Y et al. [3] proposed a high-speed railway intrusion detection network based on enhanced YOLOv3. This network improves the detection of pictures but reduces the detection speed. Cao Z et al. [4] proposed an effective railway intrusion detection method using dynamic intrusion region and lightweight neural network. Overall, current intrusion detection algorithms still have room to grow in terms of balancing algorithmic speed and accuracy.

2.2 Railway Scene Segmentation

Earlier track region segmentation mainly relied on traditional image processing and Hough transform methods. The image is first preprocessed, then extracted using employing Canny edge detection algorithm and finally the track is identified by Hough transform [5]. In recent years, most of the semantic segmentation methods based on deep learning are used to segment the track regions. Cao Z et al. [6] proposed efficient dual-stream fusion network (EDFNet), a lightweight semantic segmentation algorithm, for understanding railway scenes. Lian L et al. [7] proposed a dense multi-scale fusion network (DMFNet), a lightweight and precise semantic segmentation framework tailored for railway scenes to solve the issue of semantic sample imbalance. Due to the complex and changing environment of train operation, the track direction is not regular. There is still room for improving the accuracy of semantic segmentation algorithms in complex scenes.

2.3 Railway Risk Assessment

There are fewer research cases for railway foreign object intrusion risk level assessment at home and abroad. Zhang Z et al. [8] proposed a YOLO-based method for railway obstacle intrusion detection and risk assessment. Dagvasumberel A et al. [9] built a railway risk detection system using deep learning to analyses on-board surveillance images to detect and locate foreign objects in a railway scene. In general, there is a lack of an assessment system that scientifically and comprehensively quantifies the level of risk.

3 Method

3.1 Object Detection

The foreign object detection model in this paper uses the YOLOv5s model. Among the several versions of the YOLOv5 model, the YOLOv5s model has the smallest number of parameters and the fastest detection speed. As shown in Fig. 1, after the image is input to the neural network of YOLOv5s, the backbone network is first responsible for feature extraction. After that, the neck network is responsible for feature fusion, and finally the head network outputs the detection results. The neck network of YOLOv5s consists of FPN and PAN [10] structures. FPN enhances se-mantic information of shallow feature maps by up-sample. PAN enhances the spatial detail information of deep feature maps by down-sample. The head network of YOLOv5s consists of three different sized detection heads for detecting targets of different sizes.

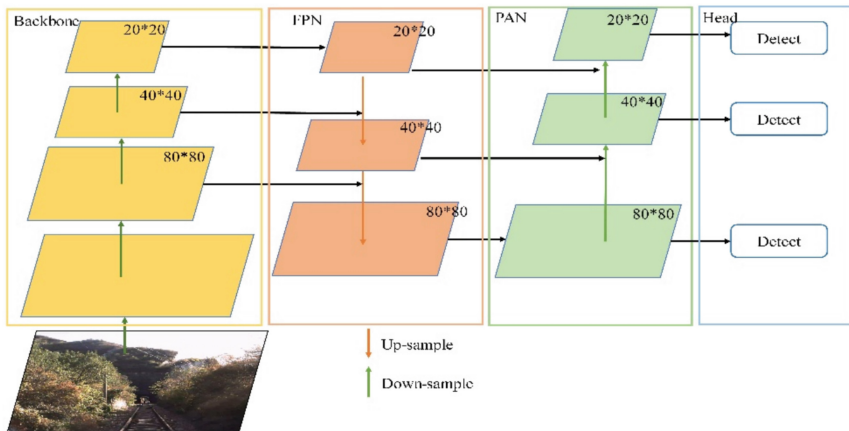


Fig. 1. YOLOv5s network structure diagram.

3.2 Semantic Segmentation

In this paper, the real-time semantic segmentation model STDC [11] model is used to segment the track region. The network structure of STDC model is shown in Fig. 2. In contrast to the two-branch structure of the BiSeNet model, the STDC model replaces the spatial path of BiSeNet by introducing the detail guidance module. This module allows shallow networks to encode spatial information in the low level features without an extra time-consuming path. The STDC model retains BiSeNet's context path to learn deep semantic information. Finally, the image segmentation results are obtained by fusing the deep semantic information and shallow spatial detail information through ARM and FFM modules. The loss function of the model consists of binary cross-entropy loss function and Dice loss function.

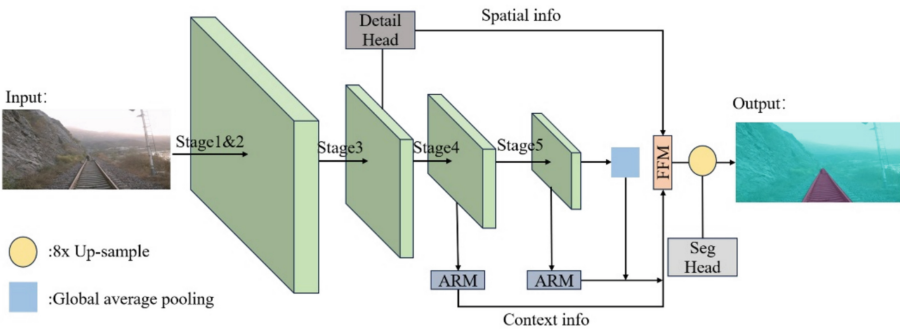


Fig. 2. STDC network structure.

3.3 A Real-Time Railway Intrusion Detection and Multi-factor Collaborative Risk Assessment Method

The methodology proposed in this paper for risk assessment is shown in Fig. 3. The level of risk arising from railway intrusion is influenced by two main factors. The two factors are the spatial location of the foreign object in relation to the track area and the type of foreign object. The former mainly affects the probability of a risk turning into an accident, while the latter affects the severity of the accident. This paper constructs the intrusion assessment system based on these two major risk factors.

The concrete process is to use the object detection algorithm to detect whether there is a foreign object in the whole image. If there is a foreign object, the image is introduced into the semantic segmentation algorithm. Segmentation of track area using the segmentation algorithm and at the same time widen the track area to form a warning area. When a foreign object is located within the warning area, the foreign object is considered to be a threat to the safety of train operations. Calculate the pixel distance between the object and the centerline of the track. Based on this distance and the category of the object, calculate the risk value of different objects in the image. Divide the image into different levels of intrusion risk based on the highest risk value. The formula for calculating the risk value of foreign objects is shown in Eq. (1) and Eq. (2).

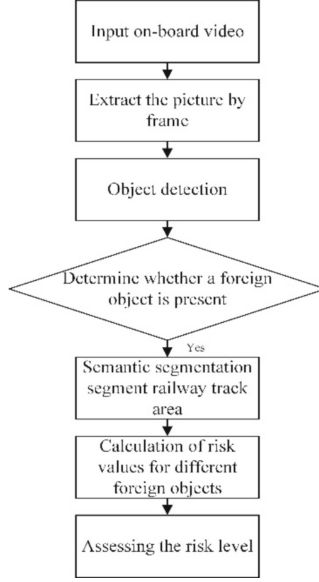


Fig. 3. Risk Assessment Flowchart.

$$R_i = \frac{P_i}{L_i} \quad (1)$$

$$R_{\max} = \max\{R_i\} \quad (2)$$

where R_i is the risk value of the i th foreign object in the image. P_i is the risk factor of the category corresponding to the i th foreign object. $R_{\max}$ is the highest risk value. L_i is the result after classifying the pixel distance between the center point of the i th foreign object and the center line of the track to $[0,4]$. The formula for L_i is shown in Eq. (3).

$$L_i = \frac{4L'_i}{L_{\max}} \quad (3)$$

where L'_i is the pixel distance between the center point of the i th foreign object and the center line of the track. $L_{\max}$ is the pixel distance from an image boundary point on the same level as the center point of the foreign object to the center line of the track.

The magnitude of the risk factor for foreign objects is based on the type of foreign object. In this paper, the categories of foreign objects are categorized as person, large foreign objects, floating objects and small foreign objects. Combining the degree of threat to the train when intrusion, and the degree of negative impact after the accident. In this paper sets the risk factors of different foreign objects as shown in Table 1.

The size of the security risk level is based on the level of the risk value. In this paper, the risk level is divided into three categories: High, Medium and Low. The specific division basis is shown in Table 2.

Table 1. Risk factors for different types of foreign objects.

Type	Risk factors
Person	4
Large foreign objects	3
Floating objects	2
Small foreign objects	1

Table 2. Basis for classifying different risk levels.

Risk Value	Risk Level
$\text{Risk Value} \geq 100$	High
$10 \leq \text{Risk Value} < 100$	Medium
$0 \leq \text{Risk Value} < 10$	Low

4 Experiments

4.1 Datasets

In this paper captures video of the foreign object intruding into the railway from the viewpoint of an on-board camera. 1788 images were extracted frame by frame from the video. These images cover the track in different scenarios and four types of foreign object: person, stone, box, and float. Where stone corresponds to large foreign objects, box to small foreign objects, and float to floating objects. In this paper, the intrusion detection dataset and the track segmentation dataset are produced by utilizing these images. The ratio of training, validation and test sets in both datasets is 8:1:1.

4.2 Object Detection

The dataset in this section uses the intrusion detection dataset in Sect. 4.1. The experimental environment configuration is shown in Table 3. The model training effect is tested after the model training using the test set. The visualization results of the test are shown in Fig. 4. From the visualization results, it can be seen that YOLOv5s can accurately detect all kinds of foreign objects in the image in different scenes.

The final test results of the model are shown in Table 4. According to the test results, the Precision of YOLOv5s reaches 87.5% and Recall reaches 80.6%, the model has high detection accuracy. In addition, YOLOv5s detects an image at FPS is 107 frames/second on the GPU of GeForce RTX3090, which enables real-time detection. The experiments show that the performance of the YOLOv5s model is capable of the task of railway intrusion detection.

Table 3. Configuration of the experimental environment.

Name	Parameter
System	Ubuntu 20.04
Graphics card	NVIDIA GeForce RTX 3090
Environment configuration	Python 3.8.17 + Pytorch 2.0.0 + CUDA 11.7



Fig. 4. YOLOv5s detection results.

Table 4. Model accuracy test results.

Class	Images	Instances	Precision	Recall	F1
All	179	526	87.5%	80.6%	83.9%
Person	179	262	84.0%	82.1%	83.0%
Stone	179	154	89.6%	89.7%	89.6%
Float	179	24	90.0%	74.9%	81.8%
Box	179	86	86.5%	75.6%	80.7%

4.3 Sematic Segmentation

The dataset in this section uses the track segmentation dataset in Sect. 4.1. The experimental environment is configured as shown in Table 5. After the STDC model training, the results of the test are shown in Fig. 5. From the figure, it can be seen that the model can accurately segment different track regions from the image in different scenes.

Table 5. Configuration of the experimental environment.

Name	Parameter
System	Ubuntu 20.04
Graphics card	NVIDIA GeForce RTX 3090
Environment configuration	Python 3.9.18 + Pytorch 1.8.0 + CUDA 11.1

As shown in Table 6, the final MIoU of the STDC model is as high as 98.5%, of which the IoU of the track area is 97.1%. It indicates that the STDC model can complete the division of the track area with high accuracy. The speed of the model was tested with an FPS of 16.5 frames/second, consuming approximately 0.06 s per image detected. The experiment proves that the STDC model also does a better job of splitting the track area in terms of speed.

Table 6. Model Test Results.

IoU _{track}	MIoU	PA _{track}	MPA	FPS
97.1%	98.5%	98.8%	99.3%	16.5



Fig. 5. Segmentation Effect Comparison Chart.

4.4 Intrusion Detection and Risk Assessment

In this paper, 489 images are selected to test the proposed intrusion detection and risk assessment method. These images include four types of foreign object, namely, person, stone, box, and float, in different scenarios with different alignments. The visualization results of the risk assessment are shown in Fig. 6.

As can be seen from the figure, the present method can accurately detect the intrusion phenomenon of different scenarios. It is also possible to calculate the risk value of different foreign objects and output the risk level based on the largest risk value. There are two persons in each of the last two images, both located near the track and further away from the track. Since the latter is outside the warning area delineated herein, it does not threaten the safety of train operations. So, the detection box and the calculated risk value for the person were not output.



Fig. 6. Visualization of the results of the risk assessment.

5 Conclusion

This paper presents a real-time railway intrusion detection and multi-factor collaborative risk assessment method. The method was tested to be able to be applied in real railway scenarios. This method can proactively detect intrusion in the area in front of the train in real time and output the risk level in time. It is conducive to assisting the driver in driving and ensuring the safe operation of the train.

Acknowledgments. This work is supported by the National Key R&D Program of China. (Contract No. 2023YFB3211103).

References

1. Qin, Y., Cao, Z., Sun, Y.: Research on active safety methodologies for intelligent railway systems. *Engineering* **27**(08), 266–279 (2023)
2. Cao, Z., Qin, Y., Jia, L.: Railway intrusion detection based on machine vision: a survey, challenges, and perspectives. *IEEE Trans. Intell. Transp. Syst.* **25**(07), 6427–6448 (2024)
3. Ma, Y., Fang, J., Zhao, J.: Research on the application of YOLO v3 in railway intruding objects recognition. In: 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 583–586. IEEE, Dalian (2022)
4. Cao, Z., Qin, Y., Xie, Z.: An effective railway intrusion detection method using dynamic intrusion region and lightweight neural network. *Measurement* **191**, 110564 (2022)
5. Peng, F., Chen, W., Mao, B.: Rail-framework abstraction based on canny edge detection and poly-connection. *J. Chin. Railway Soc.* **34**(02), 52–57 (2012). (in Chinese)
6. Cao, Z., Gao, Y., Bai, J.: Efficient dual-stream fusion network for real-time railway scene understanding. *IEEE Trans. Intell. Transp. Syst.* **25**(08), 9442–9452 (2024)
7. Lian, L., Cao, Z., Qin, Y.: Densely multi-scale fusion network for lightweight and accurate semantic segmentation of railway scenes. *IEEE Trans. Instrum. Meas.* **73**, 1–11 (2024)
8. Zhang, Z., Chen, P., Huang, Y.: Railway obstacle intrusion warning mechanism integrating YOLO-based detection and risk assessment. *J. Ind. Inf. Integr.* **38**, 100571 (2024)
9. Dagvasumberel, A., Myagmardulam, B., Myagmar, B.: Railroad near-miss occurrence detection and risk estimation system with data from camera using deep learning. In: 2021 5th International Conference on Imaging, Signal Processing and Communications (ICISPC), pp. 83–87. IEEE, Kumamoto (2021)

10. Liu, S., Qi, L., Qin, H.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768. IEEE, Salt Lake City (2018)
11. Fan, M., Lai, S., Huang, J.: Rethinking BiSeNet for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9716–9725. IEEE/CVF, Nashville (2021)

Author Index

B

Bai, Wei 179
Bai, Wu 268
Bai, Yingjie 277
Bao, Chaoming 188
Bao, Chunrong 486

C

Cao, Zhiwei 521
Chao, Zhu 454
Chen, Changyu 117
Chen, Huihui 148
Chen, Jinduo 360
Chen, Jingyuan 69
Chen, Kangshuai 422
Chen, Linying 127
Chen, Min 303
Chen, Sidao 387
Chen, Xinqiang 46
Chen, Yang 198
Cheng, Qi 422
Cui, Chenghao 171
Cui, Mengjie 221

D

Dao, Gong 229
Deng, Hua 337
Deng, Ming 9
Deng, Shuaiqing 94
Deng, Xingdong 289
Diao, Yongqi 477
Dong, Yixuan 329

F

Fan, Lingling 312
Fang, Haoda 209
Fu, Chuanyun 179
Fu, Minxue 171
Fu, Yuling 349

G

Gao, Yang 521
Gong, Dao 1
Gu, Enxin 495
Gu, Nan 139
Guo, Fei 188
Guo, Feng 319
Guo, Jianyuan 521
Guo, Tao 78
Guo, Zexian 414

H

Han, Bing 46
Han, Dezhi 46
He, Bing 444
He, Xiaona 462
Hu, Xinkai 27
Huang, Junfu 462
Huang, Lei 369
Huang, Liang 505
Huang, Wenqian 117
Huang, Yamin 127, 155, 505
Huang, Yan 396
Huang, Zhiqiu 329

J

Ji, Honghai 312
Jia, Wanting 360
Jia, Xinzhaoh 209
Jiang, Gege 69
Jiang, Shuyan 277
Jiang, Yanxuan 27
Jiao, Zhixiang 422
Jin, Haiwei 117

K

Kong, Xianghao 360
Kou, Lei 319

L

Lai, Qingying 268
 Lan, Junyi 1
 Lei, Zihao 94
 Li, Changan 239
 Li, Haijun 396
 Li, Honggang 435
 Li, Huazhi 198
 Li, Hui 360
 Li, Huiqiang 239
 Li, Jian 505
 Li, Jiyang 188
 Li, Kaopeng 53
 Li, Sha 9
 Li, Ya 78
 Li, Yingqi 18
 Li, Yongzhong 378
 Li, Yuheng 378
 Li, Zihua 69
 Li, Zinuo 268
 Lin, Ke 277
 Lin, Yuemin 414
 Liu, Chengli 148
 Liu, Haodong 139
 Liu, Hu 9
 Liu, Huanfeng 35
 Liu, Jie 127, 155
 Liu, Jinzhao 179
 Liu, Jun 268
 Liu, Lu 139
 Liu, Shida 312
 Liu, Xiao 337
 Liu, Xiaoming 61
 Liu, Yongqiang 78
 Liu, Yu 268
 Liu, Zimin 94
 Lu, Bin 444
 Lu, Kai 277, 289
 Lu, Qiong 209
 Lu, Zhaoyou 179
 Luo, Zhouyang 414
 Lv, Ying 477

M

Ma, Xiaoning 53, 107
 Ma, Yong 148
 Meng, ZhangJie 444
 Meng, Zhangxuan 94
 Muriatoy, Lootha Loserian 329

N

Nie, Lei 221

P

Pan, Deng 444
 Peng, Daiwen 9
 Peng, Genwang 521
 Peng, Zhouhua 139

Q

Qiao, Xiaoshu 239
 Qin, Baoxin 239
 Qin, Yong 521

R

Ren, Kang 188

S

Shi, Qirui 312
 Song, Anning 35
 Song, Lihang 155
 Su, Junsheng 319
 Su, Yu 94
 Sun, Baodong 239
 Sun, Dingxiang 329
 Sun, Jinjun 435
 Sun, Siqi 53, 107
 Sun, Xiao 229
 Sun, Xiaohui 303, 349

T

Tan, Yunzhou 369
 Tang, Ruiyu 319
 Tang, YiChao 444
 Tang, Yu 462
 Tang, Zongdi 127
 Tian, Zhengda 369

W

Wang, Bing 18, 209, 337
 Wang, Endian 69
 Wang, Gaihui 329
 Wang, Hanfei 35
 Wang, Haoliang 329
 Wang, Haoyue 259
 Wang, Jianqiang 396
 wang, Li 312
 Wang, Peiran 53, 107
 Wang, Qi 378

Wang, Zhangyu 198

Wang, Ziqi 69

Wei, Dan 360, 369

Wei, Jingfeng 289

Wei, Wei 486

Wei, Yuhua 404

Wen, Guangrui 94

Wen, Yuanqiao 127, 155, 505

Wu, Jiang 53, 107

Wu, Mengxi 18

Wu, Xinkai 251

Wu, Yaoting 289

Wu, Yong 337

Wu, Yuanlong 86

Wu, Zhongdai 148

Wu, Ziyang 35

X

Xiao, Pan 444

Xiao, Yao 69

Xiong, Chanzhen 86

Xiong, Xinyan 27

Xiong, Yingzhi 462

Xiong, Zhongxia 251

Xu, Ahui 155, 505

Xu, Jiaxue 139

Xu, Ming 117

Xue, Qianshu 171

Xun, Jing 171

Y

Yan, Changhui 435

Yan, Wei 319

Yang, Jianjian 422

Yang, Jie 27

Yang, Rudong 117

Yang, Yongsheng 46

Yao, Lvqi 369

Yao, Ziyang 251

Yi, Kan 462

Yu, Guizhen 198

Z

Zeng, Kun 289

Zhang, Daiyu 188

Zhang, Han 46

Zhang, Jiaxin 303

Zhang, Ke 337

Zhang, Liang 61

Zhang, Peng 148

Zhang, Qiang 444, 462

Zhang, Shilin 259

Zhang, Teng 422

Zhang, Weidong 387

Zhang, Xuejun 387

Zhang, Zheng 69

Zhang, Zhifen 94

Zhang, Zishuo 35

Zhang, Zundong 259

Zhao, Fei 198

Zhao, Guanghui 422

Zhao, Hongzheng 319

Zhao, Xingya 337

Zhao, Xuqiang 319

Zhao, Ying 396

Zhao, Zifan 422

Zheng, Jingqi 349

Zheng, Lai 486

Zheng, Shaowen 422

Zheng, Yiwen 46

Zhou, Huijuan 61

Zhou, Jinying 444, 462

Zhou, Zhijie 277

Zou, Dan 53, 107