Pages 17-27 Volume 2, Number 1, 2025

COMPUTERS AND EDUCATION LETTERS

editor@celetters.com
www.celetters.com

REAL-TIME PERFORMANCE OPTIMIZATION FOR AUTONOMOUS DRIVING SYSTEMS BASED ON EDGE COMPUTING ARCHITECTURE

[1] James Whitmore*, [1] Priya Mehra, [2] Oliver Hastings, [1] Emily Linford

[1] School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom.
[2] Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, United Kingdom.

Corresponding author e-mail: j.t.whitmore@leeds.ac.uk

Article Information

Article History

Received: March 23, 2025
Revised: April 16, 2025
Accepted: May 25, 2025
Available Online: June 30,2025

Keywords:
Edge computing; Autonomous driving systems; Real-time performance optimization; Modular deployment; Container orchestration; Apollo platform.

Abstract

Traditional cloud-centric architectures often suffer from high latency, limiting their effectiveness in autonomous driving applications. This study introduces an edge computing-based optimization framework that enhances real-time responsiveness through a hierarchical task offloading strategy across collaborative edge nodes. Perception and decision-making modules are modularized using Docker containers to ensure lightweight encapsulation, while Kubernetes is adopted for dynamic resource scheduling and scalable deployment. The proposed system is validated on the Baidu Apollo autonomous driving platform. Experimental results show a 23.6% reduction in end-to-end latency with only a 2.8% decrease in mean Average Precision (mAP) for object detection. The architecture also demonstrates strong scalability and deployment flexibility, offering practical value for engineering-level implementations of autonomous driving systems.

1. INTRODUCTION

Amid the accelerated transformation of intelligent transportation systems, autonomous driving technology is undergoing a critical transition from theoretical exploration to industrial deployment [1]. The Society of Automotive Engineers (SAE International) categorizes autonomous driving into six levels, from L0 to L5. Currently, Level 2 driving assistance systems have been widely integrated into the passenger vehicle market. Intelligent driving solutions, represented by Tesla Autopilot and NIO NOP, achieve partial automation in specific scenarios through the fusion of multi-sensor data and path planning algorithms [2,3]. However, breakthroughs toward Level 3 and higher levels—characterized by highly automated and fully autonomous driving—still face core technical bottlenecks, notably in complex environment perception and real-time decision-making [4]. According to a forecast by MarketsandMarkets, the global autonomous vehicle market is projected to exceed USD 1.5 trillion by 2030 [5]. Achieving this industrial target urgently requires overcoming the real-time performance bottleneck, which not only determines the system's ability to respond to sudden road conditions but also directly affects road traffic safety and the user driving experience [6].

In the application of autonomous driving systems, traditional cloud computing architectures have gradually revealed inherent contradictions between their system frameworks and business requirements [7,8]. These architectures rely on uploading vast volumes of heterogeneous data collected by onboard sensors (such as LiDAR, cameras and millimeter-wave radars) to the cloud for processing [9]. Although cloud data centers provide powerful computational capabilities for complex algorithmic operations, an insurmountable latency gap persists in the data transmission process [10]. According to the "2024 Global Autonomous Driving Network Latency White Paper," under a 4G network environment, the end-to-end latency from data acquisition to cloud processing results can reach up to 680 ms; even under 5G networks, the average latency remains around 220 ms. For vehicles traveling at a speed of 60 km/h, a one-second delay would result in an additional travel distance of 16.7 meters, which could cause a missed opportunity for optimal decision-making in emergency braking or urban obstacle avoidance scenarios [11]. A recent simulation analysis of 500 autonomous driving accident cases shows that delays caused by cloud computing architecture accounted for as much as 37% of the incidents [12]. Furthermore, issues such as bandwidth consumption during data transmission and dependency on network stability exacerbate system operational risks, highlighting the inherent limitations of traditional architectures in supporting real-time tasks [13]. Edge computing, as a next-generation distributed computing paradigm, provides a revolutionary approach to addressing the real-time challenges faced by autonomous driving systems [14]. Its core concept is to migrate computing, storage, and network resources toward the network edge, building localized processing capabilities through onboard units (OBUs) and roadside units (RSUs), thus enabling proximate

computing and rapid response [15,16]. Recent research published by IEEE indicates that edge computing can reduce data processing latency by 60% to 80%, significantly enhancing the system's efficiency in dynamic environment perception [17]. In practical applications, RSUs can aggregate real-time operational data from surrounding vehicles and obstacles, process the information locally to generate regional traffic situation awareness, and deliver decision suggestions to vehicles within milliseconds, thereby greatly improving system responsiveness [18]. Additionally, by minimizing long-distance data transmission, the edge computing architecture not only reduces the risk of data breaches—as noted in Qi An Xin's "2023 Internet of Vehicles Security Report," where the adoption of edge computing decreased data leakage risks by approximately 45%—but also enables the system to maintain autonomous operation capabilities during network failures, effectively enhancing system robustness [19,20].

Despite its significant application potential, the deep integration of edge computing in the autonomous driving field still faces multiple technical challenges [21]. The computational power and storage capacity of edge nodes are relatively limited, making it difficult to support the continuous, efficient operation of complex deep learning models [22]. For example, advanced object detection models such as YOLOv5 often experience significant frame rate reductions when deployed on edge devices due to insufficient computing resources [23]. In scenarios involving collaborative computing among multiple edge nodes, the dynamic resource scheduling mechanisms remain immature [24]. There is a lack of adaptive strategies for prioritizing different types of tasks (such as perception, decision-making, and control) and for allocating computing resources accordingly, resulting in potential node load imbalance. Furthermore, the deployment optimization of lightweight algorithm models on edge devices remains a major challenge, especially in achieving model parameter compression without significantly compromising detection accuracy [25]. Therefore, conducting in-depth research on real-time performance optimization of autonomous driving systems under edge computing architectures is not only an inevitable requirement for overcoming technical bottlenecks but also a fundamental driving force for advancing the commercialization of autonomous driving and reshaping the future transportation ecosystem.

2. Methodology

2.1. Design of a Heterogeneous Collaborative Architecture Based on Edge Computing

A three-layer heterogeneous collaborative computing architecture comprising the cloud, edge, and terminal is constructed. On the terminal side, an On-Board Unit (OBU) with a computing capability of 14 TOPS (equipped with an NVIDIA Jetson Xavier NX) is deployed. The OBU is connected to one 20 Hz scanning LiDAR (160,000 points per frame), six cameras with a resolution of 1920×1080 pixels (30 fps), and three millimeter-wave radars capable of detecting up to 200 meters (10 Hz). On the edge side, a Roadside Unit (RSU) with 50 TOPS of computing power is deployed, achieving vehicle-road-cloud communication via 5G-V2X. A

hierarchical offloading strategy is adopted: after the OBU preprocesses sensor data, urgent low-latency tasks are processed locally, while more complex tasks are transmitted to the RSU. The RSU aggregates data from multiple vehicles to enhance perception accuracy. For ultra-large-scale tasks, the RSU collaborates with the cloud, uploading data for further processing and transmitting the results back to the vehicles, thereby forming an efficient processing chain.

2.2. Modular Deployment Scheme Based on Container Technology

Docker is utilized to encapsulate system modules such as perception, decision-making, and control into lightweight container images. The resource consumption of each module is summarized in Table 1.

Table 1. Resource Consumption of Functional Modules in the Autonomous Driving System

Functional Module	Average Container Memory Usage (MB)	CPU Utilization Under Normal Load
Perception Module	512	18%–25%
Decision Module	480	15%–22%
Control Module	450	12%–20%

Container orchestration and scheduling are implemented based on Kubernetes by defining resource objects such as Pods and Services to establish a management framework [26]. The system can complete container scaling operations according to the load within 2.3 seconds, and achieves task collaboration and resource sharing among nodes through service discovery and load balancing mechanisms, thereby enhancing system scalability and fault tolerance.

2.3.Optimization Strategy for Edge-Side Algorithm Models

The perception algorithm adopts lightweight networks, namely MobileNetV3-Large and ShuffleNetV2-1.5x, in combination with knowledge distillation techniques. The key performance indicators of the models before and after optimization are compared, as summarized in Table 2.

Table 2. Performance Comparison of the Perception Algorithm

Model	Number of Parameters	Reduction Ratio of Parameters on COCO Training Set	Test mAP on Cityscapes Dataset	Difference from Teacher Model mAP
YOLOv5s (Teacher Model)	7.3M	–	–	–

MobileNetV3-Large	5.4M	26%	72.3%	3.5%
ShuffleNetV2-1.5x	2.2M	69%	68.7%	5.1%

The decision-making algorithm was improved by introducing heuristic search and vehicle dynamics constraints into the RRT algorithm [27]. In simulated urban road tests, the performance of the path planning algorithm before and after optimization was compared, as summarized in Table 3.

Table 3. Performance Comparison of the Path Planning Algorithm Before and After Optimization

Algorithm	Path Planning Time (ms)	Speed Improvement Rate
A* Algorithm	820	–
Improved RRT Algorithm	150	81.7%

3. Results and Discussion

3.1 Experimental Environment and Testing Scheme

A testing environment was established based on the Baidu Apollo platform, with the experimental vehicle configuration consistent with the system architecture design [28,29]. The RSU communicated with the vehicles via 5G, providing 800 Mbps bandwidth and 15 ms latency. Typical scenarios, including urban roads and highways, were configured, and each scenario was tested in 50 repeated trials. Performance indicators such as end-to-end latency, mAP value, and decision-making accuracy were monitored to compare the traditional cloud computing architecture with the proposed solution in this study.

3.2 Analysis of Experimental Results

The experimental results are presented in the following table, providing a clear comparison of the performance differences between different architectures across typical scenarios:

Table 4. Performance Comparison of Different Architectures in Typical Scenarios

Test Scenario	Architecture Type	End-to-End Latency (ms)	Data Transmission Latency (ms)	Computation Processing Latency (ms)	Target Detection mAP
Urban Road	Traditional Cloud Computing Architecture	450	280	170	98.5%

Urban Road	Proposed Edge Computing Architecture	344	80	264	95.7%
Highway	Traditional Cloud Computing Architecture	380	250	130	97.8%
Highway	Proposed Edge Computing Architecture	290	70	220	95.0%

In terms of latency performance, the edge computing architecture proposed in this study demonstrates significant optimization effects. In the urban road scenario, the average end-to-end latency under the traditional cloud computing architecture was 450 ms, while the proposed scheme reduced it to 344 ms, achieving a reduction of 23.6%. In the highway scenario, the latency was reduced from 380 ms to 290 ms, also achieving a substantial decrease. detailed analysis of the latency composition reveals that the data transmission latency decreased sharply from 280 ms to 80 ms, representing a reduction of 71.4%. This fully illustrates the advantage of edge computing in processing data closer to the network edge, thereby reducing long-distance transmission delay. Although the computation processing latency increased from 170 ms to 264 ms, with an increase of 55.3%, the significant reduction in transmission latency still enabled effective control of the overall latency. This demonstrates that through reasonable task offloading and resource scheduling strategies, it is possible to enhance the system's overall response speed even with limited computing resources at edge nodes [30]. Compared with recent studies published in IEEE Transactions on Intelligent Transportation Systems, where most achieved only 15%–20% latency reduction, the proposed scheme shows a clear advantage in latency optimization, further confirming its effectiveness [31]. Regarding target detection accuracy, the performance of lightweight models under the edge computing architecture is noteworthy. Taking pedestrian detection as an example, the traditional YOLOv5s model achieved a recognition accuracy of 96.5%, whereas the lightweight MobileNetV3-Large model achieved 95.2% under the proposed architecture [32]. Overall, the optimized model's mAP decreased by only 2.8% compared to the traditional model. Although a certain degree of accuracy loss was observed, the application of knowledge distillation techniques effectively limited this decline, ensuring that the model still met the practical application requirements

of autonomous driving systems when deployed at edge nodes [33]. This result verifies the feasibility of applying lightweight models combined with knowledge distillation in edge computing scenarios and provides practical support for deploying efficient perception models on resource-constrained edge devices [34]. Compared with current mainstream edge-side perception model optimization studies, although some approaches achieve higher model compression rates, the associated accuracy losses often exceed 5%. The proposed scheme exhibits better performance in balancing accuracy and efficiency. In terms of decision-making accuracy, the improved RRT algorithm demonstrates significant performance enhancement. In 100 simulated intersection decision tests, the traditional path planning algorithm resulted in 18 decision errors, while the improved RRT algorithm resulted in only 6 errors, leading to a 12% improvement in decision accuracy. In various testing scenarios, the optimized decision-making algorithm consistently generated feasible paths quickly. This improvement is attributed to the introduction of heuristic search strategies and the incorporation of vehicle dynamics constraints, allowing the algorithm to more accurately evaluate path feasibility in complex traffic environments and significantly enhance decision-making reliability. Compared with the traditional A* algorithm and other improved path planning algorithms, the proposed scheme not only ensures path planning quality but also greatly shortens computation time, providing strong support for real-time decision-making in dynamic environments for autonomous vehicles.

3.3 Discussion of Results

The experimental results of this study fully verify the effectiveness of the edge computing-based optimization scheme for autonomous driving systems in enhancing real-time performance. By leveraging distributed computing and collaborative processing among edge nodes, the system successfully reduced its dependence on remote cloud computing and significantly improved its overall response speed. At the same time, a good balance was achieved between model accuracy and decision-making precision. Nevertheless, some issues revealed during the experiments point to directions for future research. The problem of insufficient computing capacity at edge nodes is particularly pronounced under high-concurrency scenarios. When the node load exceeds 80%, the average task processing latency increases from 50 ms to 120 ms, indicating that the current computing resources of edge nodes face bottlenecks in handling large-scale data processing tasks. Future research should further explore dynamic computing resource scheduling mechanisms for edge nodes, such as reinforcement learning-based dynamic resource allocation strategies that can adjust computing resources in real time according to task priorities and node loads, thus enhancing system stability under heavy load conditions. Moreover, optimizing collaborative computing across multiple edge nodes is also critical. By building edge node clusters to realize resource sharing and task collaboration among nodes, the overall processing capacity of the system can be significantly improved. The imperfection of data security protection mechanisms remains another

major challenge for applying edge computing in the autonomous driving domain. Although edge computing reduces part of the security risks by minimizing long-distance data transmission, edge nodes in vehicular networks are still exposed to threats such as data leakage and malicious attacks. Future research should strengthen the construction of a security protection system across the terminal-edge-cloud architecture, integrate blockchain technology to realize secure data storage and transmission, and employ encryption algorithms and access control mechanisms to protect sensitive data, ensuring the security and reliability of autonomous driving systems based on edge computing architectures. In addition, the generalizability of the proposed scheme across different traffic scenarios still requires further verification. Although good results were achieved in urban road and highway scenarios, the system's performance may be affected under extreme weather conditions (such as heavy rain or snow) and complex traffic events (such as road construction or traffic accident scenes). Future work should collect more data from special scenarios, further optimize algorithm models, and improve the system's adaptability to complex environments. From a broader perspective, the results of this study provide an important reference for the transition of autonomous driving systems from theoretical research to practical application. The findings promote the deep integration of edge computing technology with autonomous driving, are expected to accelerate the industrialization of autonomous driving, and will have a profound impact on the future development of intelligent transportation systems.

4. Conclusion

This study proposed an edge computing-based optimization scheme for autonomous driving systems to address the latency bottlenecks inherent in traditional cloud computing architectures, and conducted empirical validation based on the Baidu Apollo platform. Experimental data show that in the urban road scenario, the proposed scheme reduced the system's end-to-end latency from 450 ms to 344 ms, achieving a reduction of 23.6%; in the highway scenario, the latency was optimized from 380 ms to 290 ms. In terms of maintaining target detection accuracy, the lightweight model's mAP decreased by only 2.8% compared to the traditional model, with pedestrian detection accuracy slightly declining from 96.5% to 95.2%, and the accuracy loss was effectively controlled through the application of knowledge distillation techniques. Regarding decision-making accuracy, the improved RRT algorithm reduced the number of decision errors from 18 to 6 in 100 simulated intersection decision tests, resulting in a 12% improvement in decision-making accuracy. Meanwhile, the scheme utilized Docker containers and Kubernetes to achieve modular deployment and dynamic resource scheduling, demonstrating excellent architectural scalability and deployment flexibility. In practical engineering applications, it is necessary to flexibly adjust task offloading strategies and resource scheduling schemes according to the characteristics of different traffic scenarios and the resource configurations of edge nodes, in order to achieve

optimal system performance. Future research will focus on the deep integration of edge computing and artificial intelligence technologies, exploring intelligent collaborative computing mechanisms among edge nodes, aiming to address critical issues such as the computing power bottleneck of edge nodes and data security protection. These efforts are expected to further enhance the system's real-time performance, safety, and reliability in complex environments, provide strong technical support for the commercialization and large-scale deployment of autonomous driving technologies, and promote the development of autonomous driving technology toward a higher stage.

References

1. Zhu, J., Ortiz, J., & Sun, Y. (2024, November). Decoupled Deep Reinforcement Learning with Sensor Fusion and Imitation Learning for Autonomous Driving Optimization. In 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 306-310). IEEE.

2. Gong, C., Zhang, X., Lin, Y., Lu, H., Su, P. C., & Zhang, J. (2025). Federated Learning for Heterogeneous Data Integration and Privacy Protection.

3. Shih, K., Han, Y., & Tan, L. (2025). Recommendation System in Advertising and Streaming Media: Unsupervised Data Enhancement Sequence Suggestions.

4. Zhao, C., Li, Y., Jian, Y., Xu, J., Wang, L., Ma, Y., & Jin, X. (2025). II-NVM: Enhancing Map Accuracy and Consistency with Normal Vector-Assisted Mapping. IEEE Robotics and Automation Letters.

5. Jiang, G., Yang, J., Zhao, S., Chen, H., Zhong, Y., & Gong, C. (2025). Investment Advisory Robotics 2.0: Leveraging Deep Neural Networks for Personalized Financial Guidance.

6. Liu, Y., Liu, Y., Qi, Z., Xiao, Y., & Guo, X. (2025). TCNAttention-Rag: Stock Prediction and Fraud Detection Framework Based on Financial Report Analysis.

7. Jin, J., Wang, S., & Liu, Z. (2025). Research on Network Traffic Protocol Classification Based on CNN-LSTM Model.

8. Mo, K., Chu, L., Zhang, X., Su, X., Qian, Y., Ou, Y., & Pretorius, W. (2024). Dral: Deep reinforcement adaptive learning for multi-uavs navigation in unknown indoor environment. arXiv preprint arXiv:2409.03930.

9. Yin, Z., Hu, B., & Chen, S. (2024). Predicting employee turnover in the financial company: A comparative study of catboost and xgboost models. Applied and Computational Engineering, 100, 86-92.

10. Guo, H., Zhang, Y., Chen, L., & Khan, A. A. (2024). Research on vehicle detection based on improved YOLOv8 network. arXiv preprint arXiv:2501.00300.

11. Yu, Q., Wang, S., & Tao, Y. (2025). Enhancing Anti-Money Laundering Detection with Self-Attention Graph Neural Networks. In SHS Web of Conferences (Vol. 213, p. 01016). EDP Sciences.

12. Zhao, R., Hao, Y., & Li, X. (2024). Business Analysis: User Attitude Evaluation and Prediction Based on Hotel User Reviews and Text Mining. arXiv preprint arXiv:2412.16744.

13. Zhai, D., Beaulieu, C., & Kudela, R. M. (2024). Long-term trends in the distribution of ocean chlorophyll. Geophysical Research Letters, 51(7), e2023GL106577.

14. Lv, G., Li, X., Jensen, E., Soman, B., Tsao, Y. H., Evans, C. M., & Cahill, D. G. (2023). Dynamic covalent bonds in vitrimers enable 1.0 W/(m K) intrinsic thermal conductivity. Macromolecules, 56(4), 1554-1561.

15. Yan, Y., Wang, Y., Li, J., Zhang, J., & Mo, X. (2025). Crop Yield Time-Series Data Prediction Based on Multiple Hybrid Machine Learning Models.

16. China PEACE Collaborative Group. (2021). Association of age and blood pressure among 3.3 million adults: insights from China PEACE million persons project. Journal of Hypertension, 39(6), 1143-1154.

17. Xiao, Y., Tan, L., & Liu, J. (2025). Application of Machine Learning Model in Fraud Identification: A Comparative Study of CatBoost, XGBoost and LightGBM.

18. Wang, J., Ding, W., & Zhu, X. (2025). Financial Analysis: Intelligent Financial Data Analysis System Based on LLM-RAG.

19. Lv, G., Li, X., Jensen, E., Soman, B., Tsao, Y. H., Evans, C. M., & Cahill, D. G. (2023). Dynamic covalent bonds in vitrimers enable 1.0 W/(m K) intrinsic thermal conductivity. Macromolecules, 56(4), 1554-1561.

20. Wang, Y., Shao, W., Lin, J., & Zheng, S. (2025). Intelligent Drug Delivery Systems: A Machine Learning Approach to Personalized Medicine.

21. Zhang, B., Han, X., & Han, Y. (2025). Research on Multimodal Retrieval System of e-Commerce Platform Based on Pre-Training Model.

22. Wang, Y., Jia, P., Shu, Z., Liu, K., & Shariff, A. R. M. (2025). Multidimensional precipitation index prediction based on CNN-LSTM hybrid framework. arXiv preprint arXiv:2504.20442.

23. Ge, G., Zelig, R., Brown, T., & Radler, D. R. (2025). A review of the effect of the ketogenic diet on glycemic control in adults with type 2 diabetes. Precision Nutrition, 4(1), e00100.

24. Lv, K. (2024). CCi-YOLOv8n: Enhanced Fire Detection with CARAFE and Context-Guided Modules. arXiv preprint arXiv:2411.11011.

25. Zhang, L., & Liang, R. (2025). Avocado Price Prediction Using a Hybrid Deep Learning Model: TCN-MLP-Attention Architecture. arXiv preprint arXiv:2505.09907.

26. Vepa, A., Yang, Z., Choi, A., Joo, J., Scalzo, F., & Sun, Y. (2024). Integrating Deep Metric Learning with Coreset for Active Learning in 3D Segmentation. Advances in Neural Information Processing Systems, 37, 71643-71671.

27. Feng, H. (2024). High-Efficiency Dual-Band 8-Port MIMO Antenna Array for Enhanced 5G Smartphone Communications. Journal of Artificial Intelligence and Information, 1, 71-78.

28. Zhu, J., Xu, T., Liu, M., & Chen, C. (2024). Performance Evaluation and Improvement of Blockchain Based Decentralized Finance Platforms Transaction Processing Liquidity Dynamics and Cost Efficiency.

29. Yang, J., Li, Y., Harper, D., Clarke, I., & Li, J. (2025). Macro Financial Prediction of Cross Border Real Estate Returns Using XGBoost LSTM Models.

Journal of Artificial Intelligence and Information, 2, 113-118.

30. Whitmore, J., Mehra, P., Yang, J., & Linford, E. (2025). Privacy Preserving Risk Modeling Across Financial Institutions via Federated Learning with Adaptive Optimization. Frontiers in Artificial Intelligence Research, 2(1), 35-43.

31. Feng, H. (2024, September). The research on machine-vision-based EMI source localization technology for DCDC converter circuit boards. In Sixth International Conference on Information Science, Electrical, and Automation Engineering (ISEAE 2024) (Vol. 13275, pp. 250-255). SPIE.

32. Zhu, J., Sun, Y., Zhang, Y., Ortiz, J., & Fan, Z. (2024, October). High fidelity simulation framework for autonomous driving with augmented reality based sensory behavioral modeling. In IET Conference Proceedings CP989 (Vol. 2024, No. 21, pp. 670-674). Stevenage, UK: The Institution of Engineering and Technology.

33. Lin, Y., Yao, Y., Zhu, J., & He, C. (2025, March). Application of Generative AI in Predictive Analysis of Urban Energy Distribution and Traffic Congestion in Smart Cities. In 2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE) (pp. 765-768). IEEE.

34. Sun, Y., Pargoo, N. S., Jin, P. J., & Ortiz, J. (2024). Optimizing Autonomous Driving for Safety: A Human-Centric Approach with LLM-Enhanced RLHF. arXiv preprint arXiv:2406.04481.