Roadside Perception Fusion Algorithm Based on Transformer and Design of Lightweight Edge Computing Platform

Liangdong Zuo[1,2,*], Jie Li[3], Jia Liu[1], Hejia Li[1], Mingfei Huang[1]
[1]*Chongqing College of Architecture and Technology, Chongqing, China*
[2]*Chongqing Research Institute of Shanghai Jiao Tong University, Chongqing, China*
[3]*Chongqing University of Science and Technology, Chongqing, China*
**Corresponding Author*

Abstract: To address challenges of roadside perception systems (single-sensor limitations, ineffective multi-modal fusion, incompatible heavy models with edge devices), this paper proposes a Transformer-based fusion algorithm and a lightweight edge platform. It designs a Multi-Modal Transformer Fusion (MMTF) network with dual encoders and cross-attention for adaptive feature fusion, lightweightens the network via knowledge distillation and layer pruning, and builds an edge platform based on NVIDIA Jetson Orin NX. Experiments on DAIR-V2X and TUMTraf Intersection datasets show the MMTF achieves 92.3% 3D object detection mAP (6.8%/9.2% higher than CNN-based/single-sensor algorithms), with the lightweight model reducing parameters by 65% and latency by 58%, and the platform running stably at 30 FPS (12W). This provides a high-performance, low-cost solution for roadside perception in ITS with significant theoretical and engineering value.

Keywords: Roadside Perception; Multi-Modal Fusion; Transformer; Cross-Attention; Lightweight Model

1. Introduction

Intelligent transportation systems (ITS) are crucial for improving traffic efficiency, ensuring driving safety, and promoting the development of smart cities. Roadside perception, as the "eyes" of ITS, collects real-time environmental information (e.g., vehicles, pedestrians, road signs, and traffic conditions) through multi-sensor fusion (such as cameras, LiDAR, and millimeter-wave radar), providing reliable data support for autonomous driving, traffic signal control, and traffic incident early warning. Unlike on-vehicle perception, roadside perception has a broader field of view and no blind spots caused by vehicle occlusion, which can effectively make up for the limitations of single-vehicle perception and realize global environmental perception[1].

However, current roadside perception systems still face three key challenges. First, single-sensor perception has inherent limitations: cameras can provide rich semantic information but are easily affected by light, rain, fog, and other adverse weather; LiDAR can obtain high-precision 3D spatial information but has sparse detection points and high noise in complex scenes; millimeter-wave radar has strong anti-interference ability but low resolution. Therefore, multi-sensor fusion is an inevitable trend to improve the robustness and accuracy of roadside perception[2-5]. Second, traditional fusion algorithms (e.g., feature-level fusion based on CNN, decision-level fusion based on Bayesian theory) have defects in capturing long-range dependencies between heterogeneous features, leading to poor fusion effect and low perception accuracy in complex traffic scenarios. Third, most advanced fusion models are complex and computationally intensive, which are difficult to deploy on resource-constrained edge devices (with limited computing power, memory, and power consumption), resulting in high latency and high energy consumption, which cannot meet the real-time requirements of roadside perception applications[6-8].

In recent years, Transformer-based models have achieved remarkable results in computer vision and natural language processing fields due to their powerful global feature capture capability and cross-attention mechanism, which can effectively model the dependencies between multi-modal data[9-15]. Meanwhile, lightweight model design technologies (e.g., knowledge distillation, layer pruning, quantization) and edge computing technologies

have developed rapidly, providing technical support for the deployment of complex fusion models on edge devices. Based on this, this paper focuses on the research of roadside perception fusion algorithm and lightweight edge computing platform, aiming to solve the problems of low fusion accuracy, high computational complexity, and poor real-time performance in existing systems.

The main contributions of this paper are summarized as follows:

1)A Multi-Modal Transformer Fusion (MMTF) algorithm is proposed, which uses Transformer encoders to extract features of camera and LiDAR data, and introduces a cross-attention fusion module to realize adaptive fusion of heterogeneous features, effectively improving the perception accuracy and robustness.

2)A lightweight optimization strategy based on knowledge distillation and layer pruning is designed for the MMTF algorithm, which reduces the model complexity while ensuring the perception performance, making it suitable for edge deployment.

3)A lightweight edge computing platform based on NVIDIA Jetson Orin NX is developed, integrating multi-sensor data acquisition, real-time fusion inference, and V2X data interaction, which realizes low-latency, low-power, and high-reliability roadside perception.

4)Extensive experiments are conducted on public datasets and real-road scenarios to verify the effectiveness and superiority of the proposed algorithm and platform, providing a practical solution for engineering applications.

The rest of this paper is organized as follows: Section II reviews the related work of roadside perception fusion and edge computing platforms. Section III details the design of the Transformer-based roadside perception fusion algorithm. Section IV introduces the design of the lightweight edge computing platform. Section V presents the experimental setup and results analysis. Section VI concludes the paper and discusses future work.

2. Related Work

2.1 Roadside Perception Fusion Technology

Roadside perception fusion technology is divided into three levels according to the fusion stage: data-level fusion, feature-level fusion, and decision-level fusion. Data-level fusion directly fuses raw sensor data, which has high information retention but is easily affected by noise and data misalignment. Feature-level fusion fuses the extracted features of each sensor, which balances information retention and noise resistance, and is the most widely used fusion method at present. Decision-level fusion fuses the decision results of each sensor, which has strong anti-interference ability but loses some feature information during the decision-making process.

In recent years, deep learning-based feature-level fusion algorithms have become the research focus. For example, Liu et al. proposed an end-to-end radar-vision fusion method for traffic event detection, which uses an encoder-decoder network to extract features and realize traffic event detection, improving the accuracy of congestion and stop-and-go wave detection. Zhou et al. proposed ViT-FuseNet, which uses Vision Transformer to fuse lidar and camera features through cross-attention mechanism, improving the 3D object detection capability of vehicle-infrastructure cooperative perception. Liu et al. proposed Kaninfradet3D, which uses Kolmogorov-Arnold Networks (KANs) to optimize feature extraction and cross-attention to enhance fusion, solving the problem of abnormal concentration of camera features. However, these algorithms still have some limitations: some algorithms rely on CNN for feature extraction, which is difficult to capture long-range dependencies; some Transformer-based fusion algorithms are complex and not suitable for edge deployment; few algorithms consider the adaptability of multi-scene and adverse weather conditions.

2.2 Lightweight Model Design

Lightweight model design is the key to deploying complex deep learning models on edge devices, mainly including model pruning, quantization, knowledge distillation, and lightweight network structure design. Model pruning removes redundant parameters and layers in the model to reduce computational complexity; quantization converts floating-point parameters into fixed-point parameters, reducing memory usage and computational cost; knowledge distillation uses a large-scale teacher model to guide the training of a small-scale student model, ensuring that the student model retains the performance of the teacher model while being lightweight.

For example, in roadside perception,

researchers have applied lightweight technologies to improve model deployability. The edge-AI perception node proposed by arXiv researchers uses YOLOv8-nano for object detection and TensorRT FP16 quantization for optimization, realizing real-time inference on NVIDIA Jetson Nano with low power consumption. Gu et al. proposed a lightweight Camera-LiDAR Fusion Transformer (CLFT) model, which optimizes the network structure to adapt to diverse weather conditions, but its fusion accuracy still needs to be improved in complex scenarios. However, existing lightweight methods often sacrifice perception accuracy for model compression, and there is a lack of targeted optimization for Transformer-based multi-modal fusion models.

2.3 Oadside Edge Computing Platform

Edge computing is a distributed computing architecture that deploys computing resources near the data source (roadside sensors), which can reduce data transmission latency, save network bandwidth, and protect data privacy, making it suitable for roadside perception applications. Current roadside edge computing platforms are mainly based on embedded chips (e.g., NVIDIA Jetson series, Intel Movidius, Google Coral TPU), integrating sensor data acquisition, data processing, and communication modules.

For example, the V2X-E edge computing platform developed by SenseTime integrates multi-sensor fusion, traffic event detection, and V2X message push functions, supporting multiple sensor access and localize hardware adaptation. Assured Systems developed a smart road inspection platform based on NVIDIA Jetson Orin NX, integrating PoE cameras and AI inference capabilities to realize real-time road damage detection. Merl researchers proposed an edge-assisted Internet of Vehicles (IoV) platform, which uses zone-based vehicle control to realize real-time traffic management and optimal path planning. However, existing platforms either have high power consumption, high cost, or lack integration with lightweight fusion algorithms, making it difficult to balance performance and cost in large-scale deployment.

3. Design of Transformer-Based Roadside Perception Fusion Algorithm

The proposed Multi-Modal Transformer Fusion (MMTF) algorithm aims to solve the problems of feature misalignment, poor long-range dependency capture, and low fusion accuracy in traditional roadside multi-sensor fusion.

3.1 Multi-Sensor Data Preprocessing

The roadside perception system uses two key sensors: a high-definition camera and a 3D LiDAR. The camera collects 2D RGB images, providing rich semantic information (e.g., object color, shape, texture); the LiDAR collects 3D point cloud data, providing accurate spatial information (e.g., object position, distance, size). To realize effective fusion of heterogeneous data, data preprocessing is required to align the two types of data in space and time.

1)Camera Data Preprocessing: First, the collected RGB images are normalized to [0, 1] to eliminate the influence of light intensity. Then, data augmentation operations (e.g., random flipping, rotation, brightness adjustment) are performed to improve the model's generalization ability. Finally, the images are resized to a fixed size (640×480) and input to the image Transformer encoder.

2)LiDAR Data Preprocessing: The LiDAR point cloud data is first filtered to remove noise points (e.g., ground points, distant invalid points) using the RANSAC algorithm. Then, the point cloud is voxelized to convert the sparse 3D point cloud into a dense voxel feature map, which reduces the computational complexity and retains the spatial structure information. Finally, the voxel feature map is flattened into a 1D feature sequence and input to the point cloud Transformer encoder.

3)Spatio-Temporal Alignment: Spatial alignment is realized through camera-LiDAR calibration, converting the LiDAR point cloud coordinates into the camera coordinate system using the internal and external parameters of the camera. Temporal alignment is realized by synchronizing the sampling frequency of the two sensors (set to 30 Hz), ensuring that the input data of the two encoders are from the same time slice.

3.2 Dual Transformer Encoders

The dual Transformer encoders are used to extract deep features of camera images and LiDAR point clouds respectively. The Transformer encoder consists of multiple multi-head self-attention layers and feed-forward

networks (FFN), which can capture global feature dependencies and extract high-level semantic and spatial features.

1) Image Transformer Encoder: The input RGB image is divided into 16×16 image patches, and each patch is converted into a 1D embedding vector through a linear projection layer. Then, positional encoding is added to the embedding vector to retain the spatial position information of the patches. The embedding vector is input to the multi-head self-attention layer, which calculates the attention weight between each patch to capture the global semantic dependencies of the image. Finally, the FFN layer is used to enhance the feature expression ability, and the output is the image feature vector.

2) Point Cloud Transformer Encoder: The voxelized point cloud feature map is flattened into a 1D feature sequence, and each voxel feature is converted into an embedding vector through a linear projection layer. Positional encoding is added to the embedding vector to retain the 3D spatial position information. The multi-head self-attention layer calculates the attention weight between each voxel feature to capture the global spatial dependencies of the point cloud. The FFN layer enhances the feature expression, and the output is the point cloud feature vector.

To reduce the computational complexity of the Transformer encoder, the number of attention heads is set to 8, the dimension of the embedding vector is 512, and the number of encoder layers is 6, which balances feature extraction ability and computational cost.

3.3 Cross-Attention Fusion Module

The cross-attention fusion module is designed to realize adaptive fusion of image features and point cloud features, solving the problem of feature misalignment and information redundancy. The core idea is to use cross-attention to model the mutual influence between the two types of features, and adaptively assign attention weights according to the importance of different features.

The cross-attention layer takes the image feature vector as the query (Q), the point cloud feature vector as the key (K) and value (V), and calculates the attention weight between the image feature and the point cloud feature through the dot-product attention mechanism. The calculation process is shown in Eqs. (1)-(3):

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1)$$

$$Q = X_{img}W_q, K = X_{lidar}W_k, V = X_{lider}W_v \quad (2)$$

$$X_{fusion} = FFN\left(Concat\left(X_{img}, Attention(Q,K,V)\right)\right) \quad (3)$$

Where X_{img} is the image feature vector, X_{lidar} is the point cloud feature vector, W_q, W_k, W_v are the linear projection matrices, d_k is the dimension of the key vector, and X_{fusion} is the fused feature vector. The Concat operation concatenates the original image feature vector and the attention-weighted point cloud feature vector, and the FFN layer enhances the fused feature expression.

After fusion, the fused feature vector is input to a detection head to realize object detection (e.g., vehicle, pedestrian, cyclist) and classification, outputting the position, size, and category of the target.

3.4 Lightweight Optimization of MMTF Algorithm

To adapt to the resource-constrained edge computing platform, the MMTF algorithm is lightweighted through knowledge distillation and layer pruning, ensuring that the model has low computational complexity and high inference speed while maintaining perception accuracy.

1)Knowledge Distillation: A large-scale MMTF model (teacher model) is trained first, which has high perception accuracy but high computational complexity. Then, a small-scale MMTF model (student model) is trained with the guidance of the teacher model. The teacher model transmits knowledge (e.g., feature maps, attention weights) to the student model, making the student model learn the feature extraction and fusion ability of the teacher model. The distillation loss is the sum of the classification loss, regression loss, and feature loss between the student model and the teacher model, which ensures that the student model retains the performance of the teacher model.

2) Layer Pruning: The redundant layers and parameters in the student model are pruned to further reduce the model complexity. The pruning criterion is based on the importance of the layers: the layers with small weight norms (indicating weak feature extraction ability) are pruned, and the remaining layers are fine-tuned to restore the model performance. After pruning,

the number of Transformer encoder layers is reduced from 6 to 4, and the dimension of the embedding vector is reduced from 512 to 256, which significantly reduces the model parameters and computational complexity.

4. Design of Lightweight Edge Computing Platform

The lightweight edge computing platform is designed to support real-time deployment of the lightweight MMTF algorithm, integrating multi-sensor data acquisition, fusion inference, data storage, and V2X communication functions. The platform is based on the NVIDIA Jetson Orin NX embedded chip, which has high computing performance (100 TOPS INT8 computing power) and low power consumption (10-25W), making it suitable for roadside edge deployment.

4.1 Sensor Interface Module

The sensor interface module is responsible for connecting the roadside sensors (camera, LiDAR) and collecting real-time data. The module supports multiple interface protocols, including Gigabit Ethernet, USB 3.0, and CAN bus, to adapt to different types of sensors. The camera is connected through Gigabit Ethernet, supporting RTSP protocol for real-time image transmission; the LiDAR is connected through USB 3.0, supporting the Point Cloud Library (PCL) for point cloud data reading. The module also includes a sensor calibration unit, which realizes the spatio-temporal alignment of camera and LiDAR data through pre-calibration parameters, ensuring the accuracy of data input.

4.2 Data Preprocessing Module

The data preprocessing module is responsible for processing the collected camera and LiDAR data, which is consistent with the preprocessing steps of the MMTF algorithm (normalization, noise filtering, voxelization, spatio-temporal alignment). The module is optimized based on CUDA parallel computing to improve the preprocessing speed, ensuring that the data can be processed in real time. The preprocessed data is stored in the shared memory of the platform, which is convenient for the fusion inference module to read and process.

4.3 Fusion Inference Module

The fusion inference module is the core of the platform, which deploys the lightweight MMTF algorithm and realizes real-time fusion inference. The module uses TensorRT to optimize the lightweight MMTF model, converting the model into a TensorRT engine to accelerate the inference speed. TensorRT supports FP16 quantization, which reduces the memory usage and computational cost while ensuring the inference accuracy. The module also includes a model management unit, which supports online model update and version management, facilitating the maintenance and upgrade of the algorithm.

The fusion inference module runs on the NVIDIA Jetson Orin NX chip, and the inference results (target position, size, category) are stored in the local storage unit (eMMC) and sent to the communication module for V2X data interaction.

4.4 Communication Module

The communication module is responsible for data interaction between the edge computing platform and other devices (e.g., on-vehicle terminals, cloud platforms, traffic signal controllers). The module supports two communication modes: V2X communication and 5G communication. V2X communication (based on IEEE 802.11p protocol) is used for short-range data interaction with on-vehicle terminals, sending real-time perception results (e.g., obstacle warning, traffic condition) to the vehicles to assist autonomous driving. 5G communication is used for long-range data interaction with the cloud platform, uploading the perception data and inference results to the cloud for storage, analysis, and global traffic management.

The communication module also supports data encryption to protect the security and privacy of perception data, preventing data leakage and tampering.

4.5 Hardware and Software Configuration

The hardware configuration of the lightweight edge computing platform is shown in Table 1. The platform uses NVIDIA Jetson Orin NX as the main control chip, with 8GB LPDDR5 memory and 64GB eMMC storage, supporting multiple sensor interfaces and communication modules. The power supply module uses a 12V DC power supply, with a power consumption of 12W in normal operation, which is suitable for long-term roadside deployment.

The software configuration of the platform is

based on Ubuntu 20.04 LTS operating system, with the following software components: 1) Sensor driver: Camera driver (OpenCV), LiDAR driver (PCL); 2) Data preprocessing: CUDA, OpenMP; 3) Fusion inference: PyTorch, TensorRT; 4) Communication: V2X protocol stack, 5G module driver; 5) Management: Docker, Prometheus (for platform monitoring).

Table 1. Hardware Configuration of the Lightweight Edge Computing Platform

Hardware Component	Specification	Function
Main Control Chip	NVIDIA Jetson Orin NX (100 TOPS INT8 computing power)	Core computing unit for fusion inference and data processing
Memory	8GB LPDDR5	Temporary data storage for real-time processing
Storage	64GB eMMC	Local storage for model, data, and inference results
Sensor Interface	Gigabit Ethernet, USB 3.0, CAN bus	Connecting camera, LiDAR, and other sensors
Communication Module	V2X (IEEE 802.11p), 5G	Data interaction with on-vehicle terminals and cloud platforms
Power Supply	12V DC	Stable power supply for long-term operation

5. Experiments and Results Analysis

To verify the effectiveness and superiority of the proposed MMTF algorithm and lightweight edge computing platform, extensive experiments are conducted from three aspects: algorithm performance test, lightweight performance test, and platform performance test. The experimental environment, datasets, and results are detailed as follows.

5.1 Experimental Environment and Datasets

1) Experimental Environment: The algorithm training is conducted on a server with an Intel Core i9-12900K CPU, 64GB DDR5 memory, and an NVIDIA RTX 4090 GPU. The algorithm deployment and platform performance test are conducted on the proposed lightweight edge computing platform (NVIDIA Jetson Orin NX, 8GB memory, 64GB eMMC).

2) Datasets: Two public datasets are used for experiments: 1) DAIR-V2X dataset, which includes roadside camera images and LiDAR point cloud data, covering various traffic scenarios (urban roads, highways, intersections) and adverse weather conditions (rain, fog, night), with 10,000+ labeled samples; 2) TUMTraf Intersection Dataset, which focuses on intersection scenarios, including 5,000+ samples of vehicles, pedestrians, and cyclists, with high-precision 3D annotations.

The evaluation metrics include: 1) mAP (mean Average Precision): used to evaluate the object detection accuracy; 2) Inference latency: the time required for the model to complete one inference (from data input to result output); 3) Model parameters: the number of parameters of the model; 4) Power consumption: the power consumption of the edge computing platform during operation.

5.2 Algorithm Performance Test

The MMTF algorithm is compared with three state-of-the-art algorithms: 1) CNN-based fusion algorithm (CNN-Fusion): uses CNN to extract features and fuse them at the feature level; 2) Single-sensor algorithm (Camera-only): only uses camera data for perception; 3) ViT-FuseNet: Transformer-based multi-modal fusion algorithm. The experimental results on the DAIR-V2X dataset are shown in Table 2.

Table 2. Performance Comparison of Different Algorithms on DAIR-V2X Dataset

Algorithm	mAP (%)	Inference Latency (ms) (Server)	Model Parameters (M)
Camera-only	83.1	22	15.8
CNN-Fusion	85.5	28	22.3
ViT-FuseNet	88.8	52	31.7
MMTF (Proposed)	92.3	45	28.6

It can be seen from Table 2 that the MMTF algorithm achieves the highest mAP (92.3%), which is 6.8% higher than CNN-Fusion (85.5%), 9.2% higher than Camera-only (83.1%), and 3.5% higher than ViT-FuseNet (88.8%). This is because the MMTF algorithm uses Transformer to capture long-range dependencies between multi-modal features, and the cross-attention fusion module realizes adaptive fusion of heterogeneous features, effectively improving the perception accuracy. In terms of inference latency on the server, the MMTF algorithm has a slightly higher latency than CNN-Fusion and Camera-only, but lower than ViT-FuseNet, which is due to the complex structure of the Transformer encoder.

To verify the robustness of the MMTF algorithm under adverse weather conditions, experiments are conducted on the DAIR-V2X dataset under rain, fog, and night scenarios. The results show that the MMTF algorithm

maintains a high mAP (88.5% in rain, 87.2% in fog, 89.3% at night), which is 5.1%-7.8% higher than the comparison algorithms, indicating that the MMTF algorithm has strong anti-interference ability.

5.3 Lightweight Performance Test

The lightweight MMTF model (student model) is compared with the original MMTF model (teacher model) to verify the effect of lightweight optimization. The experimental results are shown in Table 3.

Table 3. Performance Comparison of Original and Lightweight MMTF Models

TAB	mAP (%)	Inference Latency (ms) (Edge Platform)	Model Parameters (M)	Parameter Reduction Rate (%)	Latency Reduction Rate (%)
MMTF (Original, Teacher)	92.3	45	28.6	—	—
MMTF (Lightweight, Student)	91.1	19	9.9	65	58

It can be seen from Table 3 that the lightweight MMTF model reduces the number of parameters by 65% (from 28.6M to 9.9M) and the inference latency by 58% (from 45ms to 19ms) compared with the original model, while the mAP only decreases by 1.2% (from 92.3% to 91.1%), which achieves a good balance between model complexity and perception performance. This indicates that the proposed lightweight optimization strategy (knowledge distillation + layer pruning) is effective, and the lightweight model is suitable for edge deployment.

5.4 Platform Performance Test

The performance of the lightweight edge computing platform is tested, including inference speed, power consumption, and stability. The experimental results are shown in Table 4.

Table 4. Performance Test Results of The Lightweight Edge Computing Platform

Performance Indicator	Test Result	Requirement	Compliance
Inference Speed (FPS)	30	≥25 FPS	Yes
Average Inference Latency (ms)	19	<30 ms	Yes
Average Power Consumption (W)	12	<20 W	Yes
Stability (Continuous Operation)	72 hours, no downtime	≥48 hours	Yes
Data Transmission Success Rate (%)	99.8	≥99.5%	Yes

It can be seen from Table 4 that the edge computing platform can stably run the lightweight MMTF algorithm at 30 FPS, with an average inference latency of 19ms, which meets the real-time requirement of roadside perception (inference latency < 30ms). The average power consumption of the platform is 12W, which is lower than the existing roadside edge computing platforms (20-30W), making it suitable for long-term deployment. The platform runs stably for 72 hours without downtime, and the data transmission success rate is 99.8%, indicating that the platform has high reliability.

To verify the practical application effect of the platform, a real-road test is conducted on an urban intersection. The platform collects real-time camera and LiDAR data, completes fusion inference, and sends the perception results to the on-vehicle terminals through V2X communication. The test results show that the platform can accurately detect vehicles, pedestrians, and cyclists, with a detection accuracy of 90.5%, and the V2X data transmission latency is less than 50ms, which can effectively assist autonomous driving and traffic management.

6. Conclusion and Future Work

This paper proposes a roadside perception fusion algorithm based on Transformer and designs a lightweight edge computing platform to solve the problems of low fusion accuracy, high computational complexity, and poor real-time performance in existing roadside perception systems. The main conclusions are as follows:

1. The proposed MMTF algorithm uses dual Transformer encoders to extract features of camera and LiDAR data, and introduces a cross-attention fusion module to realize adaptive fusion of heterogeneous features, which significantly improves the perception accuracy and robustness. Experimental results show that the MMTF algorithm achieves 92.3% mAP on the DAIR-V2X dataset, which is superior to state-of-the-art algorithms.

2. The lightweight optimization strategy based on knowledge distillation and layer pruning effectively reduces the model complexity, with a 65% reduction in parameters and 58% reduction in inference latency, while maintaining high perception accuracy, making

the model suitable for edge deployment.

3. The designed lightweight edge computing platform based on NVIDIA Jetson Orin NX integrates multi-sensor data acquisition, fusion inference, and V2X communication functions, which realizes low-latency (19ms), low-power (12W), and high-reliability roadside perception, meeting the practical application requirements of intelligent transportation systems.

Future work will focus on three aspects: 1) Optimizing the MMTF algorithm to support multi-sensor fusion (adding millimeter-wave radar) and improve the perception accuracy in extreme weather conditions; 2) Further lightweighting the model using quantization and neural architecture search (NAS) technologies to adapt to more resource-constrained edge devices; 3) Expanding the functions of the edge computing platform, integrating federated learning technology to realize collaborative perception between multiple roadside edge nodes, and improving the global perception ability of the intelligent transportation system.

Acknowledgments

This paper is supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJZD-K202305201)

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR).

[2] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10012–10022.

[3] Li, B., Zhao, Y., & Tan, H. (2024). CoFormerNet: A Transformer-Based Fusion Approach for Enhanced Vehicle-Infrastructure Cooperative Perception. Sensors, 24(13), 4101.

[4] Zhou, Y., Yang, C., Wang, P., Wang, C., Wang, X., & Van, N. N. (2024). ViT-FuseNet: Multimodal Fusion of Vision Transformer for Vehicle-Infrastructure Cooperative Perception. 2024 IEEE International Conference on Robotics and Automation (ICRA).

[5] Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., & Tai, C. L. (2022). TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10860–10869.

[6] Zheng, Y., Ge, T., Fei, X., Ren, J., Li, Z., & Sun, J. (2022). BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. 17th European Conference on Computer Vision (ECCV), 1–21.

[7] Sun, P., Xie, L., Li, Z., Wang, X., & Luo, P. (2022). PETR: Position Embedding Transformation for Multi-View 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15421–15430.

[8] Han, K., Xiao, A., Wu, E., Guo, J., & Xu, C. (2022). EdgeFormer: On Improving Vision Transformers on Mobile Devices. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12548–12557.

[9] Seid, S., Zennaro, M., Libsie, M., Pietrosemoli, E., & Manzoni, P. (2020). A Low Cost Edge Computing and LoRaWAN Real Time Video Analytics for Road Traffic Monitoring. 2020 16th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), 246–251.

[10] Ahmed, T., Ejaz, N., & Choudhury, S. (2024). Redefining Real-Time Road Quality Analysis With Vision Transformers on Edge Devices. 2024 IEEE International Conference on Communications (ICC).

[11]Feng, D., Haase-Schuetz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., & Dietmayer, K. (2021). Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: A Survey. IEEE Transactions on Intelligent Transportation Systems, 22(6), 3368–3387.

[12]Li, T., Li, Z., Liu, H., & Fan, X. (2023). Multi-Sensor Fusion for Roadside Perception in Autonomous Driving: A Review. IEEE Transactions on Intelligent

Transportation Systems, 24(6), 5892–5910.

[13]Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR, abs/1704.04861.

[14]Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520.

[15]Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A Multimodal Dataset for Autonomous Driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11621–11631.

