



ARTICLE

YOLO-PBE: An Improved YOLOv11 Vehicle Detection Algorithm for Complex Traffic Scenes

Yixiang Wan and Wenqiu Zhu*

School of Computer Science and Artificial Intelligence, Hunan University of Technology, Zhuzhou, China

*Corresponding Author: Wenqiu Zhu. Email: zwq@hut.edu.cn

Received: 23 April 2026; Accepted: 27 May 2026; Published: 23 July 2026

ABSTRACT: Addressing the two critical challenges of missed detection of distant small targets and difficulty in identifying occluded targets under complex road conditions, this paper proposes YOLO-PBE, an improved high-precision vehicle detection model based on YOLOv11n. First, to tackle the fine-grained feature loss caused by conventional strided convolutions during downsampling, we add a high-resolution P2 detection layer and introduce SPD-Conv, a lossless spatial-to-depth feature transformation technique, for feature extraction. By preserving complete pixel-level information, the model's perception accuracy for distant small vehicles is enhanced. For feature fusion, we design an improved BiFPN incorporating a Ghost module. Specifically, by pruning redundant fusion nodes, replacing computationally expensive operations with inexpensive linear transformations, and employing an adaptive weighted fusion mechanism, the model optimizes feature fusion efficiency while ensuring precise semantic alignment across multi-scale features. Finally, we construct a deeply embedded C2f-EMA module that seamlessly incorporates an efficient multi-scale attention mechanism into the C2f unit. Through cross-dimensional feature interaction and pixel-wise reweighting, the module effectively suppresses background noise interference, strengthens the reconstruction ability of damaged features, and improves robustness in detecting occluded and overlapping vehicles. Experimental results demonstrate that compared with YOLOv11n, the mean Average Precision (mAP@0.5) improves by 5.8 and 4.2 percentage points on the UA-DETRAC and KITTI datasets, respectively. This study provides a new technical approach for vehicle detection in complex traffic scenarios.

KEYWORDS: Traffic scenario; vehicle detection; YOLOv11; attention mechanism; multi-scale feature fusion

1 Introduction

With the rapid development of autonomous driving and intelligent transportation systems, high-precision and high-real-time vehicle detection technology has become a key foundation for ensuring traffic safety and improving road traffic efficiency. However, in real-world road environments, due to significant variations in vehicle scale and frequent occlusions, along with background interference, current detection algorithms face numerous challenges in complex traffic scenarios [1].

To address these challenges, scholars both domestically and internationally have conducted extensive explorations in directions such as multi-scale modeling, feature fusion, and attention enhancement. For example, in multi-scale object detection, Liu et al. [2] proposed a multi-scale feature fusion network that improved the model's sensitivity to small objects by adding detection branches and expanding the receptive field. However, its additional computational overhead limits the deployment of this method on embedded devices. In feature fusion mechanisms, Lin et al. [3] proposed FPN, which enhanced the

fusion of features at different scales through a top-down pyramid structure. Liu et al. [4] further proposed PANet, which optimized cross-scale information transfer through bidirectional paths. However, in traffic scenarios, distant small targets and nearby large targets exhibit significant differences in semantic strength and spatial resolution. The element-wise addition fusion method in PANet tends to dominate the fusion process with strong semantic features, diluting the fine-grained features of small-scale vehicles. To improve the representation of key regions, Woo et al. [5] proposed the CBAM attention mechanism, and Hou et al. [6] proposed Coordinate Attention, both of which enhanced the model's ability to focus on target regions to some extent. However, in high-dynamic and highly occluded complex traffic scenarios, CBAM's channel-spatial serial modeling approach struggles to achieve synergistic optimization, while Coordinate Attention's coordinate axis encoding method easily introduces background noise, leading to insufficient feature calibration for incomplete vehicle targets and limited robustness.

As the mainstream detection algorithm in the field of object detection, the YOLO series continues to evolve with high efficiency. Although the latest version, YOLOv11, performs excellently on general benchmarks, direct migration to complex traffic scenarios still presents the following limitations: First, its backbone network relies on traditional stride convolutions for downsampling, causing irreversible loss of shallow-layer high-resolution spatial details when processing distant small vehicles, which restricts the extraction of geometric and texture features of small targets [7]. Second, the standard PANet feature pyramid it adopts implements bidirectional information flow, but the fusion process depends on fixed topology element-wise addition, failing to adaptively adjust the contributions of features at different scales, easily causing small target features to be overwhelmed. Finally, when facing severe occlusions or background interference, its core C2f module lacks an explicit cross-dimensional feature calibration mechanism, making it difficult to effectively suppress noise and reconstruct discriminative representations of incomplete vehicle targets [8].

To overcome the shortcomings of existing methods, this paper proposes an improved vehicle detection model for complex traffic scenarios, YOLO-PBE, based on the YOLOv11 framework. The main innovations are as follows:

1. A small object perception branch based on SPD-Conv and a P2 high-resolution detection head were constructed. To address the issue of distant small vehicle features being easily lost, traditional stride convolution downsampling was abandoned, utilizing SPD lossless transformation technology to preserve original pixel information, and introducing a high-resolution P2 layer, which enhances the model's ability to capture the geometric contours and texture features of small-scale objects.
2. Designed a lightweight BiFPN feature pyramid integrated with the Ghost module. To address the contradiction between real-time detection and multi-scale adaptability, redundant nodes were pruned and the Ghost module was introduced to reduce computational burden, while an adaptive weighting mechanism was utilized to resolve feature semantic misalignment, achieving efficient multi-scale vehicle feature interaction and flow.
3. Proposed the C2f cross-dimensional feature interaction module with embedded EMA mechanism (C2f-EMA). To address occlusion and background interference in complex road conditions, the EMA attention mechanism is deeply integrated into the C2f unit, effectively suppressing background noise and significantly enhancing the robustness of identifying incomplete vehicle targets under occlusion through cross-spatial dimension feature calibration and denoising reconstruction.

2 Related Work

Deep learning-based object detection technology is a key driving force behind the development of computer vision. Among them, the single-stage YOLO series, with its synergistic advantages of speed and

accuracy, has received widespread attention and become the preferred solution for real-time visual tasks, subsequently being deployed in a variety of real-time applications. Based on its significant advantages, this study selects the latest YOLOv11-n as the baseline model. However, when directly applied to complex traffic environments, it was found that the model still has limitations in handling small vehicle targets, severe occlusion situations, and complex background interference. This section will take this baseline model and the problems it encounters in real-world scenarios as a starting point to conduct a comprehensive and systematic review of related research.

2.1 Research Status of Object Detection Algorithms

The YOLO series, as a typical example of single-stage object detection algorithms, demonstrates continuous progress and breakthroughs in detection accuracy and inference speed. Since YOLOv3 established the foundational framework for multi-scale detection, to YOLOv5 achieving significant engineering application breakthroughs, and then YOLOv8 prompting a shift in design philosophy, until YOLOv11 conducting deeper network optimization, each generation's development iteration represents a key advancement in the technical path of this series [9].

Addressing the limitations of early detectors when processing multi-scale targets, YOLOv3 employs the Darknet-53 architecture to build foundational capabilities, while YOLOv5 achieves a leap from experimental to engineering practicality by integrating a modular CSPNet backbone network with adaptive anchors, significantly enhancing the flexibility of model deployment. As task complexity increases, YOLOv8 discards the traditional anchor box mechanism in favor of a more streamlined anchor-free design and a task-cooperative architecture, markedly improving the model's generalization ability. As the latest achievement, YOLOv11 focuses on refined structural improvements, optimizing gradient paths and feature fusion mechanisms to enhance feature expression while controlling computational complexity, establishing a performance benchmark for subsequent real-time detection research.

Building on this foundation, research improvements targeting specific complex scenarios further expand the applicability of YOLO. For example, MLE-YOLO [10] introduces multi-stage Transformer branches and hybrid aggregation networks in YOLOv11 to effectively suppress precipitation interference and improve small object perception. YOLO-ACE [11], based on YOLOv10, introduces additive convolutional gate linear units combined with a dual distillation strategy, achieving a high-efficient balance between accuracy and speed while significantly reducing parameters. ZZ-YOLO [12] proposes the global edge information transfer (GEIT) module and integrates LAMP pruning technology, significantly enhancing the ability of lightweight models to capture vehicle edge features. YOLO-SCDI [13] optimizes YOLOv8 using the C2f-SCSA module and dynamic detection heads, effectively addressing the detection challenges of multi-scale targets and dense occlusions in complex urban environments, providing an improved solution for multi-scale vehicle perception. These detailed technical upgrades enable the improved algorithms to achieve a new level of balance between accuracy and efficiency when processing specific tasks.

2.2 Challenges in Object Detection in Complex Traffic Scenarios

Vehicle detection in complex traffic scenarios is a core challenge faced by autonomous driving and intelligent transportation systems, with its performance largely limited by the robustness of object detection models. In recent years, single-stage detectors represented by the YOLO series have been widely applied in practical systems due to their excellent balance between speed and accuracy. As the latest version of this series, YOLOv11 demonstrates outstanding performance in general object detection tasks through refined structural optimizations. However, when dealing with highly challenging long-distance small-scale vehicles in complex traffic environments and severely occluded targets, its detection accuracy still needs

further improvement [14]. This limitation has prompted researchers to strengthen the baseline model at multiple levels.

At the feature pyramid architecture level, traditional FPN and its enhanced version PANet have established the basic framework for multi-scale feature fusion. However, they typically treat all input features equally, without fully considering the varying contributions of features at different resolutions during fusion. To address this, the bidirectional feature pyramid network (BiFPN) introduces learnable feature weights and bidirectional cross-scale connections, achieving more efficient multi-scale fusion and significantly enhancing the model's ability to recognize vehicles with drastic scale variations. Nevertheless, when handling semantic alignment in deep networks, this structure still faces issues of feature information loss or noise interference. In terms of feature enhancement, attention mechanisms have been proven to effectively improve the model's focus on key regions. The efficient multi-scale attention mechanism (EMA) achieves richer multi-scale feature representation by incorporating parallel subnetwork structures and cross-dimensional interaction strategies, without compromising channel dimension. Although the efficient multi-scale attention mechanism offers higher accuracy than traditional serial attention mechanisms (such as CBAM), its stability in capturing discriminative features remains room for further optimization in high-density occlusion scenarios. Additionally, for the issue of missed detections of small targets, introducing a high-resolution shallow P2 detection layer is considered a direct and effective strategy [15]. This layer preserves richer spatial details and positional information, compensating for the shortcomings of high-level features. However, the additional computational overhead generated, as well as the semantic misalignment between deep and shallow features, often limits further improvements in both operational efficiency and detection accuracy.

In summary, although BiFPN, EMA, and P2 detection layers each excel in feature fusion efficiency, discriminative feature refinement, and preservation of low-level details, respectively, overcoming the structural redundancy and computational burden that may arise from simply stacking these modules, and addressing the adaptability challenges of multi-scale features during fusion, remain practical challenges for current model improvements. Therefore, how to organically and synergistically integrate these technologies into the latest YOLOv11 framework, and to design a vehicle detection algorithm that comprehensively optimizes feature granularity, fusion efficiency, and feature selection mechanisms, remains an issue awaiting in-depth research [16]. The research presented in this paper is based on the aforementioned technical bottlenecks and optimization opportunities.

3 Improve the YOLO-PBE Framework

This study uses YOLOv11 as the baseline model and collaboratively improves the feature extraction structure, multi-scale feature fusion paths, and detection head configuration of the model to address issues such as large vehicle scale variations and the easy missed detection of small targets at long distances in complex traffic scenes. The improvement work mainly includes: reconstructing the C2f module in the backbone network with an embedded EMA mechanism to enhance the focus on key information during the feature expression stage. In the neck network, a lightweight bidirectional feature pyramid structure integrating the Ghost module is constructed to achieve efficient and adaptive fusion of cross-scale features. In the multi-scale detection layer, a high-resolution P2 shallow detection head is introduced to improve the model's perception of small-scale vehicle targets.

Fig. 1 shows the overall architecture of the improved YOLOv11-PBE network. This overall design balances model real-time performance while enhancing the comprehensive detection performance of the model in actual complex traffic environments, particularly improving the detection accuracy of small-scale vehicle targets and occluded vehicles.

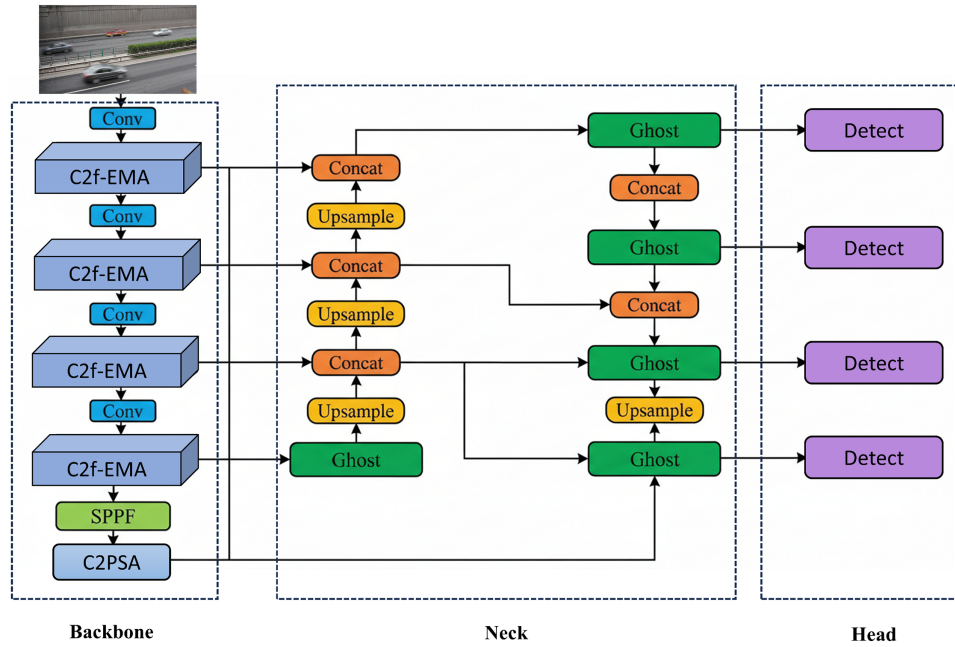


Figure 1: YOLO-PBE structure frame diagram.

3.1 P2 Micro-Target Detection Head Design

In complex traffic monitoring and on-board vision scenarios, vehicle targets exhibit extremely large scale variations. Specifically, for distant vehicles with small pixel dimensions, the original P3 detection head in YOLOv11 architecture (based on 8x downsampling) suffers from significant representational bottlenecks. Traditional strided convolution often discards a large amount of critical pixel information during downsampling, which leads to the physical-level loss of geometric contours of small targets, as well as fine-grained features like taillight textures in deep networks, resulting in irreversible feature blurring phenomena.

To address this challenge, this study compares multiple technical approaches to enhance small object perception capabilities. Although deformable convolution (DCN) can adjust receptive field shapes through adaptive offsets, and dynamic convolution can enhance expression through content-aware weights, both are based on traditional strided downsampling and cannot fundamentally solve the signal loss problem under the Nyquist-Shannon sampling theorem constraints. In contrast, SPD-Conv (Space-to-Depth Convolution) transforms downsampling into a deterministic pixel rearrangement through spatial-to-depth conversion, completely avoiding the filtering of small signals by convolution stride. For distant vehicles with extremely low pixel ratios in traffic scenarios, this lossless transformation can preserve key feature representations for subsequent detection heads [17].

Specifically, for the downsampling process of the P2 high-resolution branch, we introduce the SPD-Conv lossless transformation technology. The SPD module does not directly filter information through convolution kernels, but instead, for the weak signal characteristics of small targets in traffic monitoring, slices the spatial dimensions of feature maps and rearranges them to map to channel dimensions. Assuming the input feature map is $\mathbf{X} \in \mathbb{R}^{(S \times S \times C_1)}$, SPD transformation reorganizes it into $\mathbf{X}' \in \mathbb{R}^{\frac{S}{2} \times \frac{S}{2} \times 4C_1}$. Its mathematical process can be described as:

$$\mathbf{X}'_{x,y,z} = \mathbf{X}_{2x+(z \bmod 2), 2y+[(z \bmod 4)/2], [z/4]} \quad (1)$$

On this basis, we constructed a high-resolution P2 detection head (corresponding to $4\times$ downsampling), and cascaded the SPD-processed shallow P2 feature details with the strong semantic features of deep P3 at the channel level to form a complementary mechanism: P2 layer provides precise localization and edge information, P3 layer provides high-level semantic context required for category discrimination, thus significantly enhancing the model's sensitivity to extremely small-scale vehicles.

To balance accuracy and real-time performance, this model adopts a decoupled detection head design and introduces depth-wise separable convolutions in the classification branch to reduce computational burden. Although the introduction of the P2 layer theoretically doubles the feature map area, thanks to the parameter-free nature of SPD-Conv and the lightweight detection head structure, the overall computational overhead only slightly increases, and inference speed remains basically at a practical level. More importantly, this design significantly improves small object detection performance, while maintaining high frame rates, achieving substantial improvements in both the detection rate and localization accuracy of distant small vehicles [18].

This model adopts a decoupled detection head design, separating the classification and regression tasks. The classification branch uses depth-wise separable convolutions (DWConv) to balance computational efficiency and feature extraction capability. The improved four detection head (P2, P3, P4, P5) structure is shown in Fig. 2.

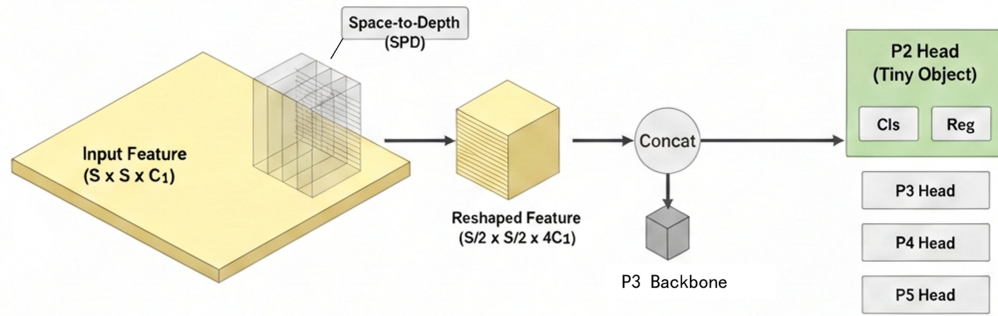


Figure 2: SPD-based P2 small-target layer.

The integration of the P2 detection head not only heightens the model's sensitivity to minute vehicle targets but also strengthens its multi-scale detection capabilities through synergy with other detection heads. Consequently, this yields a comprehensive improvement in detection accuracy and stability within complex traffic scenarios while preserving real-time performance.

3.2 Lightweight BiFPN Feature Fusion Design

Real-time detection in traffic scenarios heavily relies on the computational resources of edge devices, while the model needs to possess the ability to process multi-scale targets. Although the PANet structure adopted by the original YOLOv11 has basic multi-scale fusion capabilities, its simple feature addition strategy ignores the varying contributions of different input features, and the numerous standard convolutions within it introduce a considerable amount of parameter redundancy. For this reason, this paper proposes a lightweight bidirectional feature pyramid for traffic scenarios, the structure of which is shown in Fig. 3.

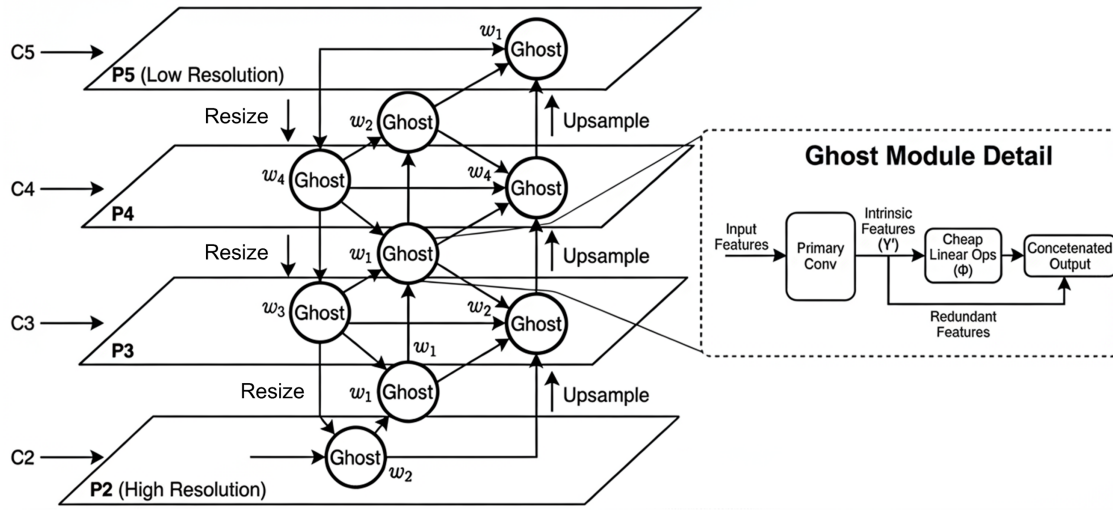


Figure 3: Ghost-enhanced lightweight BiFPN.

Based on an analysis of feature flow efficiency, this study optimizes and refines the standard BiFPN topology by eliminating redundant intermediate connection nodes, constructing a more compact bi-directional cross-scale connection path. This design aims to reduce computational redundancy in feature transmission, improve inference efficiency, and make the model more suitable for the real-time requirements of traffic scenarios while maintaining its multi-scale feature fusion capabilities. Furthermore, to further reduce computational burden, we introduce the lightweight recombination strategy of the Ghost Module at the fusion nodes [19]. The Ghost module generates core features through a small number of convolutions and applies inexpensive linear transformations to generate redundant features, assuming the input features are $\mathbf{X} \in \mathbb{R}^{h \times w \times c}$. The Ghost module first generates m intrinsic feature maps $\mathbf{Y}'$ through standard convolution, then applies s cheap linear transformations $\mathbf{Y}'$ to each Φ to generate the final output features:

$$\mathbf{Y}_{i,j} = \Phi_{i,j}(\mathbf{Y}'_i), \forall i = 1, \dots, m; \quad j = 1, \dots, s. \quad (2)$$

By applying this module, the floating-point operation count (FLOPs) in the feature extraction process can be effectively reduced. To address the semantic misalignment issue caused by significant target scale differences in traffic scenarios, TL-BiFPN employs an adaptive weighted fusion mechanism, assigning learnable weights w_i to the features of each layer from P2 to P5. Unlike the conventional BiFPN's generic setup, this paper applies this mechanism to the pruned topology structure, aiming to enable the network to automatically learn and enhance feature layers containing key target information while suppressing background noise. The weighted fusion formula is as follows:

$$\mathbf{O} = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot \mathbf{I}_i. \quad (3)$$

Among them, $\mathbf{O}$ is the fusion of output features, $\mathbf{I}_i$ is the i layer input feature, $\epsilon = 0.0001$ is the numerical stability constant. This mechanism gives the network feature selection attention, effectively suppressing the interference of background noise features.

3.3 EMA Feature Reconstruction Module Design

In urban road environments, vehicles are often occluded by trees, streetlights, or adjacent vehicles, leading to incomplete key features. Traditional attention mechanisms are typically stacked as independent modules at the end of the network, making it difficult to promptly repair damaged features at the early stage of feature extraction. To address this, this paper proposes the Deeply Embedded C2f-EMA module, which deeply integrates the EMA mechanism into the core C2f unit of YOLO, replacing the original BottleNeck structure to enhance the discriminative power and expressive ability of features. Its structure is shown in Fig. 4.

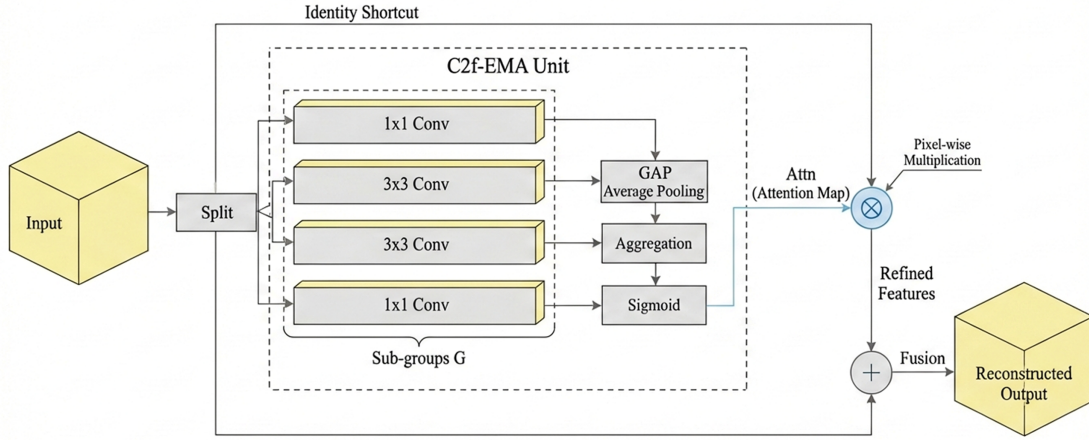


Figure 4: EMA-embedded C2f module.

To obtain rich contextual information at low computational cost, this module first employs a multi-scale parallel grouping strategy. Assuming the input features are $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, it divides them into multiple subgroups along the channel dimension. Each subgroup is processed by parallel convolutional branches of 1×1 and 3×3 to simultaneously extract local spatial details and global channel contextual information. On this basis, the module introduces a cross-spatial dimension feature calibration mechanism: after the interaction and fusion of the aforementioned dual-branch features, global average pooling encodes spatial information [20]. This design captures long-range dependencies across dimensions, establishing an effective mapping between spatial positions and channel features. The core attention weight generation process can be expressed as:

$$\mathbf{M} = \sigma(\mathcal{F}_{agg}(\mathbf{X}_{1 \times 1} \odot \mathbf{X}_{3 \times 3})) \quad (4)$$

Among them, $\mathbf{X}_{1 \times 1}$ and $\mathbf{X}_{3 \times 3}$ respectively represent dual-branch features, $\odot$ represents cross-dimensional interaction aggregation, $\mathcal{F}_{agg}$ is the spatial information encoding function, and σ is the Sigmoid activation function.

Finally, the generated attention weights $\mathbf{M}$ perform pixel-level re-weighting operations on the original features ($\mathbf{Y} = \mathbf{X} \odot \mathbf{M}$). This process enhances the concentration of feature extraction, enabling the network to focus more on the target object and reduce interference from irrelevant background features. Meanwhile, cross-dimensional feature verification optimizes the robustness of vehicle detection in complex occlusion scenarios, helping to improve the detection probability of incomplete vehicle targets.

4 Experimental Results and Analysis

The experiment was conducted on the Windows 11 operating system, with the programming language being Python 3.9.23. The deep learning framework used was PyTorch 2.5.1, and the CUDA version was 12.1. The hardware environment configuration included the 12th generation Intel Core i5-14600K processor and the NVIDIA RTX 3060 graphics card. Both the baseline YOLOv11-n and the improved model were loaded with the same version of the official COCO pre-trained weights, with the new modules randomly initialized while maintaining consistent initialization strategies. All experiments were completed under the same hyperparameter settings to ensure the comparability and reproducibility of the results. Table 1 summarizes the experimental environment configuration, and Table 2 provides the specific training parameters.

Table 1: Experimental environment configuration table.

Item	Specification
Operating System	Windows 11
CPU	Intel Core i5-14600K
GPU	NVIDIA GeForce RTX 3060
Programming Language	Python 3.9.23
Deep Learning Framework	PyTorch 2.5.1
CUDA Version	12.1

Table 2: Training parameter table.

Parameter	Value
Input image size	640 × 640
Batch size	16
Epochs	100
Initial learning rate	0.0005
Weight decay	0.001
Pretrained weights	yolov11n.pt

4.1 Experimental Datasets

The experiment uses the UA-DETRAC vehicle detection dataset, typical scene samples of which are shown in Fig. 5. This dataset is a large-scale dataset widely used in vehicle recognition and tracking tasks, with its material collected from 24 different locations of road overpasses in Beijing and Tianjin, China. It provides fine annotations for over 14,000 vehicle instances, with a total of approximately 1.21 million object bounding boxes. Vehicles in the dataset are labeled into four categories: cars, buses, trucks, and others, and it includes various typical weather and lighting conditions such as sunny, cloudy, rainy, and nighttime. Following conventional practices, the dataset is divided into training, validation, and test sets at a 7:2:1 ratio. The main technical challenge of this dataset lies in the presence of many small-sized vehicles and severe occlusions between vehicles [21].

The KITTI autonomous driving dataset is selected for model generalization performance validation. This dataset was jointly created by the Karlsruhe Institute of Technology and the Toyota Technological Institute at Chicago, covering diverse road conditions such as urban, suburban, and highway environments. It also includes complex factors like lighting changes, shadow interference, partial vehicle occlusion, and

different traffic densities, making it one of the most influential open benchmark datasets in the field of autonomous driving [22]. Through cross-evaluation on both the UA-DETRAC and KITTI datasets, the robustness of the improved model in different shooting perspectives and actual traffic scenarios can be more objectively verified.

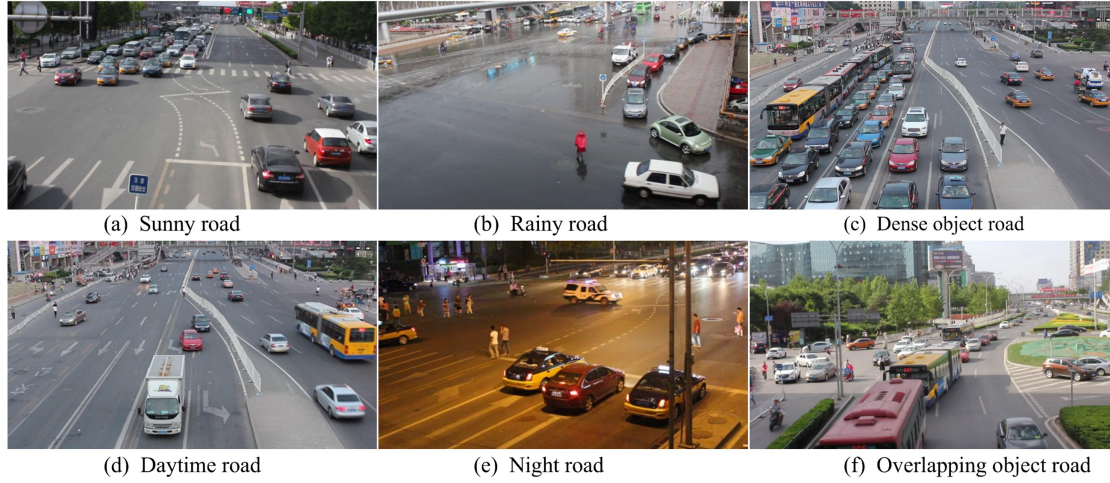


Figure 5: UA-DETRAC dataset samples.

4.2 Evaluation Metrics

Multiple data augmentation techniques were used during training to improve the model's generalization ability, while the input image size was uniformly adjusted to 640×640 pixels. Model performance evaluation adopted mean average precision (mAP@0.5), precision, and recall as the main evaluation indicators.

Precision is used to measure the accuracy of the model's detection results. Its definition is the proportion of actual vehicle targets among all targets judged by the model as vehicles. This metric focuses on the reliability of the model's prediction results.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

Among them, TP refers to the number of targets correctly detected, and FP indicates incorrect detections that do not match any real labeled targets (false positives) or repeated detections of the same target.

Recall rate is used to measure the model's ability to cover real vehicles, defined as the proportion of actual vehicles successfully identified by the model among all real existing vehicles. This metric reflects the model's ability to find all real vehicles in the dataset.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

Among them, FN represents the number of real labeled targets that the model fails to detect.

Mean Average Precision (mAP) is an indicator for comprehensively evaluating the overall performance of the model, which comprehensively considers the overall performance of precision and recall at different decision thresholds. This indicator first calculates the average precision for each detected category, i.e., the area covered by the precision-recall curve, and then takes the arithmetic mean of the average precision for all categories [23]. Among them, mAP@0.5 refers to the mean average precision calculated when the

Intersection over Union (IoU) threshold is set to 0.5, which is a widely adopted core evaluation standard in the field of object detection.

$$\text{mAP} = \frac{1}{N} \sum AP_i \quad (7)$$

Among them, N is the total number of target categories. This paper primarily adopts $\text{mAP}@0.5$ (mean mAP when the IoU threshold of intersection union ratio is 0.5) as the key metric for evaluating the overall detection performance of the model.

4.3 Experimental Results

To verify the effectiveness of the proposed improved algorithm, we conducted training and comparison between the YOLO-PBE model and the original YOLOv11-n model on the UA-DETRAC dataset, and the performance comparison curve is shown in Fig. 6.

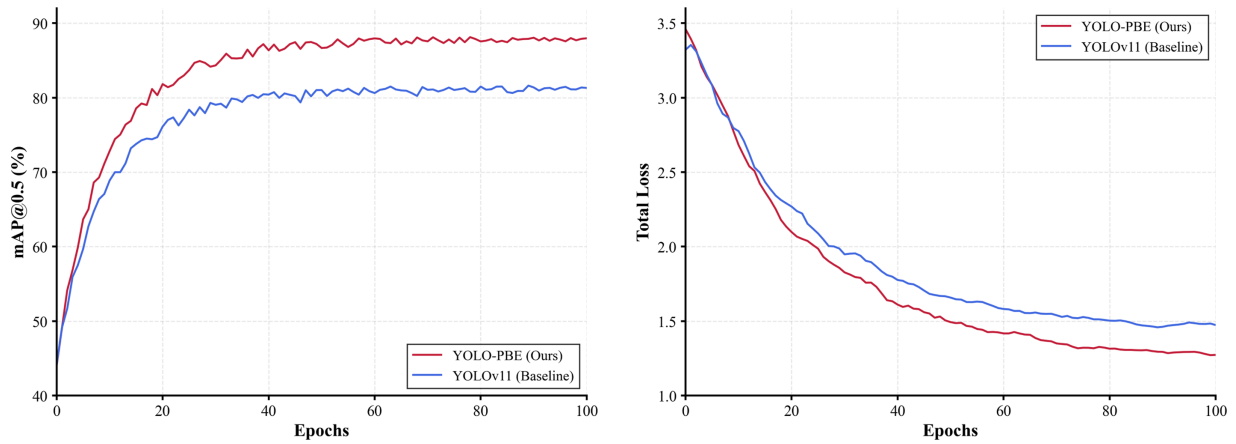


Figure 6: Performance comparison between YOLO-PBE and original YOLOv11-n.

From the perspective of the $\text{mAP}@0.5$ accuracy curve, the proposed YOLO-PBE model consistently outperformed the original YOLOv11-n model throughout the training period. In the early stages of training, the slope of the red line was significantly higher than that of the blue line, indicating that the improved model has faster feature learning and convergence speed. As training approached stability, the $\text{mAP}@0.5$ of YOLO-PBE finally stabilized at around 87%, showing a clear improvement compared to the original model's approximately 81%. This demonstrates that the improved module introduced in this paper has more efficient capabilities in extracting and fusing features, thereby significantly enhancing the model's target detection accuracy.

From the Loss curve perspective, both sets of models exhibited good convergence trends. During most of the training phase, the loss values of the improved YOLO-PBE model were lower than those of the original model, and its descent was smoother. Finally, the convergence loss of YOLO-PBE stabilized at around 1.3, lower than the original model's 1.5. The lower loss values and smoother curve further confirm that the improved model has higher efficiency in fitting target features.

4.4 Ablation Experiments

To investigate the contribution of each module design to the overall performance, we conducted ablation studies on the UA-DETRAC dataset, using YOLOv11n as the baseline model. The ablation results are shown in Table 3.

Table 3: Ablation results of different improved module combinations.

Experiment	Ghost-BiFPN	C2f-EMA	SPD-P2	Params (M)	FLOPs (G)	Precision (%)	Recall (%)	mAP@0.5 (%)	FPS
1				3.01	8.2	78.5	76.2	81.8	145.3
2	✓			3.05	7.5	79.8	77.5	83.1	131.5
3		✓		3.15	8.6	80.5	76.8	83.5	131.8
4			✓	3.52	10.3	78.8	79.0	84.3	132.7
5	✓	✓		3.20	7.9	81.2	79.3	85.2	127.5
6	✓		✓	3.58	9.8	80.5	81.2	86.5	128.4
7		✓	✓	3.66	10.7	81.0	80.5	86.0	127.7
8	✓	✓	✓	3.75	11.2	83.6	83.2	87.6	124.6

By introducing the Ghost-BiFPN module alone, mAP@0.5 improved from the baseline's 81.8% to 83.1%, while the computational cost was optimized to 7.5G, proving the advantage of weighted bidirectional feature pyramids in maintaining lightweight characteristics while enhancing feature fusion. When the C2f-EMA attention mechanism was added alone, the parameter count slightly increased to 3.15 M, with accuracy surpassing 80%, and mAP@0.5 rising to 83.5%. Notably, after introducing the SPD-P2 tiny object detection layer alone, recall improved by 2.8 percentage points, and mAP@0.5 significantly increased to 84.3%, effectively enhancing the model's ability to recognize small objects.

Based on the improvement strategies proposed in this paper, the model that ultimately integrated the Ghost-BiFPN, C2f-EMA, and SPD-P2 modules (Experiment 8) demonstrated the best performance. The parameter count and computational cost of this model were 3.75M and 11.2G, respectively, with detection accuracy reaching 83.6%, recall improving to 83.2%, and mAP@0.5 achieving the highest value of 87.6%, representing increases of 5.1%, 7.0%, and 5.8% over the baseline model, respectively. Although FPS was slightly affected, dropping from 145.3 to 124.6, this speed fully meets real-time detection requirements. These results fully demonstrate the advantages of collaborative work among the modules, achieving a significant improvement in detection accuracy.

4.5 Comparison Experiment

To verify the effectiveness and superiority of the YOLO-PBE improved model proposed in this paper, comparative experiments were conducted on the UA-DETRAC dataset under the same experimental environment and training strategies, comparing it with various classical and mainstream object detection algorithms, including the two-stage detection algorithm Faster R-CNN [24], the single-stage detection algorithm SSD, as well as YOLOv5-s, YOLOv8-n, YOLOv11-n, ZZ-YOLOv11, ACE-YOLO, and RT-DETR-R18. The comparative results of the models are shown in Table 4.

Table 4: Performance comparison results of different models.

Method	Precision (%)	Recall (%)	mAP@0.5 (%)	Params (M)	FLOPs (G)	FPS
Faster R-CNN	79.5	78.2	79.4	41.3	145.2	8.5
SSD	74.5	72.1	75.6	24.5	31.4	75.2
YOLOv5-s	76.5	74.2	78.5	7.2	16.4	92.0
YOLOv7-t	75.0	73.5	77.2	6.1	13.1	112.5
YOLOv8-n	78.6	75.8	80.6	3.1	8.7	105.2
YOLOv9-t	78.5	76.1	81.3	2.2	7.1	128.3
YOLOv10-n	78.3	76.5	81.4	2.4	6.6	132.3
YOLOv11-n	78.5	76.2	81.8	3.0	8.2	145.3
RT-DETR-R18	80.8	81.5	84.3	20.5	46.3	92.8
ZZ-YOLOv11	80.6	79.5	84.1	2.1	4.4	162.5
ACE-YOLO	81.8	80.9	85.5	2.5	5.3	132.8
YOLO-PBE	83.6	83.2	87.6	3.7	11.2	124.6

From Table 4, it can be observed that YOLO-PBE performs the best in terms of detection accuracy. On the mAP@0.5 metric, YOLO-PBE achieved 87.6%, which is an improvement of 8.2%, 12.0%, 9.1%, 10.4%, 7.0%, 5.8%, and 3.3%, respectively compared to Faster R-CNN, SSD, YOLOv5-s, YOLOv7-t, YOLOv8-n, YOLOv11-n, and RT-DETR-R18. Meanwhile, on the Precision and Recall metrics, YOLO-PBE reached 83.6% and 83.2%, respectively, both surpassing the other models involved in the comparison, indicating that the model has significant advantages in detection accuracy and stability.

In terms of model complexity, YOLO-PBE has 3.5M parameters and 8.4G FLOPs. It effectively controls the model size and computational load while maintaining high detection accuracy. Compared to Faster R-CNN, YOLO-PBE significantly reduces the number of parameters and computational complexity while improving detection accuracy. Compared to lightweight models such as ZZ-YOLOv11 and ACE-YOLO, YOLO-PBE demonstrates a clear advantage in accuracy. Compared to the latest YOLOv11-n model, YOLO-PBE still improved by 5.8% on the mAP@0.5 metric, reflecting the effectiveness of the proposed improvement strategy.

Overall, YOLO-PBE achieves a good balance between detection accuracy, model size, and inference speed, demonstrating strong accuracy and robustness. It is particularly suitable for real-time and high-precision object detection tasks in traffic scenarios requiring high accuracy and real-time performance.

4.6 Generalization Ability Evaluation

To verify the generalization capability of YOLO-PBE in complex traffic environments, this paper conducted comparative experiments with mainstream object detection algorithms on the KITTI autonomous driving dataset. The results, as shown in the Table 5, demonstrate that YOLO-PBE exhibits the best adaptability in handling outdoor complex scenes, achieving an mAP@0.5 of 82.5%, significantly surpassing traditional models Faster R-CNN (71.2%) and mainstream lightweight models YOLOv11-n (75.8%), with respective accuracy improvements of 11.3% and 6.7%. Additionally, compared to RT-DETR-R18 based on the Transformer architecture, YOLO-PBE achieves a 3.3% higher accuracy with only 17% of its parameter count and an 81.8% reduction in computational load, fully validating the efficiency of the improved structure in feature capture and semantic fusion [25].

Table 5: KITTI dataset generalization experiment results.

Method	Precision (%)	Recall (%)	mAP@0.5 (%)	Params (M)	FLOPs (G)	FPS
Faster R-CNN	72.1	70.4	71.2	41.3	145.2	8.5
SSD	66.5	63.2	65.8	24.5	31.4	75.2
YOLOv5-s	70.2	68.1	71.4	7.2	16.4	92.0
YOLOv7-t	69.4	67.5	70.1	6.1	13.1	112.5
YOLOv8-n	73.5	71.2	74.2	3.1	8.7	105.2
YOLOv9-t	74.1	72.0	75.2	2.2	7.1	128.3
YOLOv10-n	74.3	72.5	75.3	2.4	6.6	132.3
YOLOv11-n	74.8	73.4	75.8	3.0	8.2	145.3
RT-DETR-R18	76.2	77.1	79.2	20.5	46.3	92.8
ZZ-YOLOv11	78.4	76.9	79.8	2.1	4.4	162.5
ACE-YOLO	79.1	78.2	80.9	2.5	5.3	132.8
YOLO-PBE	79.5	78.8	82.5	3.7	11.2	124.6

To more intuitively demonstrate the model's detection performance, Fig. 7 provides some visualization results of YOLO-PBE and YOLOv11-n on the KITTI test set. From the figure, it can be seen that whether in dense urban scenes with severe occlusions (as shown in Fig. 7a) or open road scenes containing objects of different scales (as shown in Fig. 7b), YOLO-PBE can accurately detect small target vehicles at a distance. The bounding boxes generated by the model closely align with the target contours, demonstrating excellent localization capabilities. These visualization results are consistent with the quantitative data in Table 5, proving the powerful generalization ability of the YOLO-PBE model and its robustness in complex scenarios.

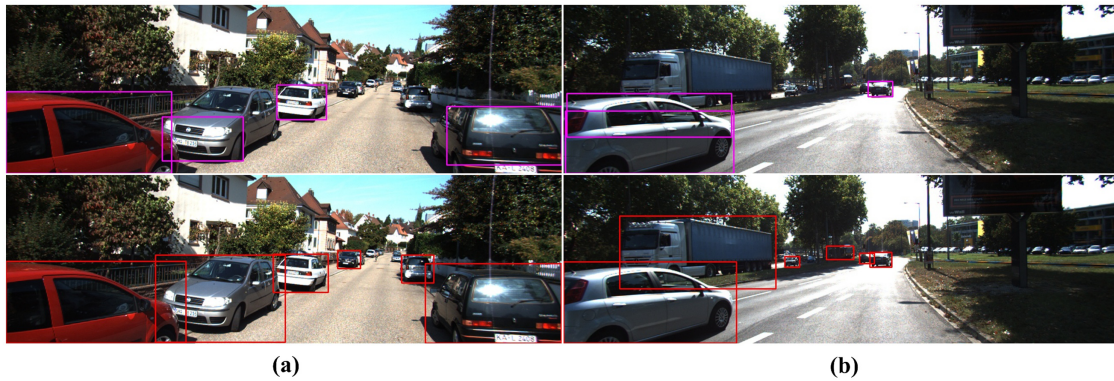


Figure 7: Visualization results of vehicle detection on the KITTI dataset: (a) Urban scenario with dense vehicles. (b) Open road scenario containing targets at various scales. Purple boxes (YOLOv11-n) and red boxes (YOLO-PBE in this paper).

In terms of inference speed and model size, YOLO-PBE also demonstrates excellent balance for practical application scenarios. Although its parameter count (3.5M) and computational load (8.4G) are slightly higher than the lightweight YOLOv9-t to achieve higher accuracy in dynamic environments, its inference speed of 124.6 FPS far exceeds real-time requirements, and it maintains leading performance in both Precision and Recall metrics. Table 5 presents the generalization performance comparison of YOLO-PBE with current mainstream algorithms under equally complex scene conditions. Overall, YOLO-PBE achieves significant accuracy breakthroughs with minimal computational cost, enhancing detection accuracy for unknown test

data while maintaining high real-time performance, making it more suitable for practical deployment in complex dynamic scenarios.

4.7 Visual Analysis

To comprehensively evaluate the effectiveness and robustness of the improved vehicle detection model YOLO-PBE proposed in this paper under complex traffic scenarios, this paper conducts a visual analysis of the detection results based on the UA-DETRAC dataset. The UA-DETRAC dataset covers representative complex traffic environments such as daytime, nighttime, rainy days, and scenarios with dense vehicles and occlusions, providing a reliable experimental foundation for verifying the model's comprehensive performance under different lighting conditions, weather changes, and target distributions [26].

Fig. 8 shows a comparison of the detection results between the original YOLOv11-n baseline model and the improved model proposed in this paper under typical complex traffic scenarios. From the visual results, it can be observed that under normal daylight conditions, as shown in Fig. 8a, the improved model can more accurately detect distant small-scale vehicles, significantly reducing the missed detection of small targets compared to the baseline model. This indicates that the improved network structure has stronger capabilities in fine-grained feature extraction. In nighttime or low-light scenarios, as shown in Fig. 8b, the improved model still maintains stable detection performance under complex background interference, with more complete localization of vehicle targets, fully demonstrating better adaptability to low-light environments.

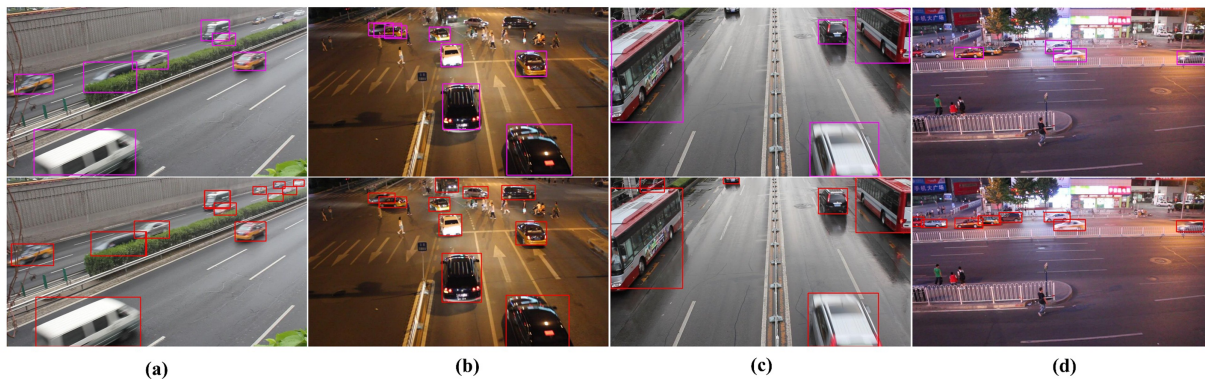


Figure 8: Comparison of detection results between YOLOv11-n and YOLO-PBE: (a) Daytime scene with distant small-scale vehicles. (b) Nighttime scene with low-light conditions. (c,d) Traffic scenarios with partial occlusions. Purple boxes (YOLOv11-n) and red boxes (YOLO-PBE in this paper).

Additionally, in scenarios with dense vehicle distribution and local occlusions, as shown in Fig. 8c, the improved model exhibits stronger differentiation between adjacent vehicles and can still provide reasonable prediction bounding boxes for partially occluded targets, while the baseline model is more prone to missed detections or inaccurate localization in the same scenario. This reflects the improved model's higher robustness in multi-scale feature fusion and contextual information construction, effectively mitigating detection challenges caused by occlusions and scale variations in complex traffic environments.

Comprehensive analysis of the results from the visual analysis above reveals that the improved model YOLO-PBE demonstrates superior detection performance in typical complex scenarios such as small target detection, low-light conditions, complex background interference, and vehicle occlusions. This qualitative analysis result aligns with the quantitative evaluation metrics from the ablation experiments and comparative

experiments described earlier, further intuitively validating the effectiveness and practical value of the proposed method in vehicle detection tasks under complex traffic scenarios.

5 Conclusion

This paper focuses on the core challenges of easily missing small targets in distant scenes and the difficulty of identifying occluded targets in complex traffic scenarios, proposing an improved detection model YOLO-PBE that balances high accuracy and lightweight design. By reconstructing the feature extraction and fusion paths, the model effectively establishes a new balance between computational efficiency and feature expression capability.

Specifically, to address the bottleneck of feature loss for small vehicles in deep networks, this study innovatively introduces the SPD-Conv lossless transformation technology and constructs a P2 high-resolution perception layer, preserving the fine-grained pixel information of targets from a physical level and fundamentally solving the feature blurring issue caused by traditional stride convolutions. On this basis, through the sparse connection and linear transformation strategies of Ghost-BiFPN, the model significantly reduces computational redundancy in multi-scale fusion, combined with the pixel-level reweighting mechanism of the deep-embedded C2f-EMA module, effectively suppressing occlusion and background noise, and significantly enhancing the robustness of the model under complex road conditions.

Experimental results on the UA-DETRAC dataset strongly validate the effectiveness of the above improvement strategies: YOLO-PBE achieves an mAP@0.5 of 87.6% while maintaining real-time inference performance, an improvement of 5.8 percentage points over the original YOLOv11. This significant accuracy improvement indicates that the proposed improved architecture successfully breaks through the performance barriers of traditional detectors in both small target vehicle detection and occluded target identification. Additionally, comparative experimental results on the generalization dataset KITTI further demonstrate that YOLO-PBE maintains strong detection capabilities under cross-scenario conditions, achieving an optimal balance between accuracy and speed, and validating the method's good generalization ability and practical application potential.

Although this method performs excellently in traffic scenarios, there is still room for improvement in deployment efficiency on resource-constrained embedded devices. Future research will focus on hardware-aware model compression and explore multi-modal fusion techniques to address complex all-weather environments, further promoting the large-scale application and implementation of high-precision vehicle detection algorithms on intelligent traffic edge devices.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and methodology: Yixiang Wan and Wenqiu Zhu; Validation: Yixiang Wan; Formal analysis: Yixiang Wan; Writing—original draft preparation: Yixiang Wan; Writing—review and editing: Wenqiu Zhu; Supervision: Wenqiu Zhu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The UA-DETRAC and KITTI datasets supporting the findings of this study are openly available in public repositories at <https://detrac-db.rit.albany.edu/> and <http://www.cvlibs.net/datasets/kitti/>. Other data and materials are available from the first author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao R, Tang S, Bin Supeni EE, Bin Abdul Rahim S, Fan L. A review of object detection in traffic scenes based on deep learning. *Appl Math Nonlinear Sci.* 2024;9(1):20240322. doi:10.2478/amns-2024-0322.
2. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. arXiv:1512.02325. 2015.
3. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. arXiv:1612.03144. 2016.
4. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. arXiv:1803.01534. 2018.
5. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. arXiv:1807.06521. 2018.
6. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. arXiv:2103.02907. 2021.
7. Wei Z, Zou Y, Xu H, Wang S, Wei Z, Zou Y, et al. Small object detection in traffic scenes for mobile robots: challenges, strategies, and future directions. *Electronics.* 2025;14(13):2614. doi:10.1109/icspcc62635.2024.10770469.
8. Yang Z, Pun-Cheng LSC. Vehicle detection in intelligent transportation systems and its applications under varying environments: a review. *Image Vis Comput.* 2018;69(11):143–54. doi:10.1016/j.imavis.2017.09.008.
9. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. arXiv:1506.02640. 2015.
10. Du D, Bi M, Xie Y, Liu Y, Qi G, Guo Y. MLE-YOLO: a lightweight and robust vehicle and pedestrian detector for adverse weather in autonomous driving. *Digit Signal Process.* 2026;168:105628. doi:10.1016/j.dsp.2025.105628.
11. Xie Y, Du D, Bi M. YOLO-ACE: a vehicle and pedestrian detection algorithm for autonomous driving scenarios based on knowledge distillation of YOLOv10. *IEEE Internet Things J.* 2025;12(15):30086–97. doi:10.1109/JIOT.2025.3569735.
12. Zhang Z, Zhang Z, Li G, Xia C. ZZ-YOLOv11: a lightweight vehicle detection model based on improved YOLOv11. *Sensors.* 2025;25(11):3399. doi:10.3390/s25113399.
13. Wu L, Cao W. YOLO-SCDI: a vehicle detection algorithm based on an improved YOLOv8. *J Electron Inf Technol.* 2025;47(10):3967–78. (In Chinese). doi:10.11999/JEIT250238.
14. Zheng Z, Zhao J, Fan J, Bai R, Zhao J, Liu J. A complex roadside object detection model based on multi-scale feature pyramid network. *Sci Rep.* 2025;15(1):15992. doi:10.1038/s41598-025-99544-1.
15. Li X, Chen J, Sun Y, Lin N, Hawbani A, Zhao L. YOLO-vehicle-pro: a cloud-edge collaborative framework for object detection in autonomous driving under adverse weather conditions. arXiv:2410.17734. 2024.
16. Xu H, Lai S, Li X, Yang Y. Cross-domain car detection model with integrated convolutional block attention mechanism. arXiv:2305.20055. 2023.
17. Qiu J, Zhang W, Xu S, Zhou H. DP-YOLO: a lightweight traffic sign detection model for small object detection. *Digit Signal Process.* 2025;165(1):105311. doi:10.1016/j.dsp.2025.105311.
18. Khalili B, Smyth AW. SOD-YOLOv8—enhancing YOLOv8 for small object detection in traffic scenes. arXiv:2408.04786. 2024.
19. Doherty J, Gardiner B, Kerr E, Siddique N. BiFPN-YOLO: one-stage object detection integrating Bi-directional feature pyramid networks. *Pattern Recognit.* 2025;160(7):111209. doi:10.1016/j.patcog.2024.111209.
20. Gao Y, Li Z, Wang Y, Zhu S. A Novel YOLOv5_ES based on lightweight small object detection head for PCB surface defect detection. *Sci Rep.* 2024;14(1):23650. doi:10.1038/s41598-024-74368-7.
21. Wen L, Du D, Cai Z, Lei Z, Chang MC, Qi H, et al. UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking. *Comput Vis Image Underst.* 2020;193(9):102907. doi:10.1016/j.cviu.2020.102907.
22. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*; 2012 Jun 16–21; Providence, RI, USA. New York, NY, USA: IEEE; 2012. p. 3354–61.
23. Oksuz K, Cam BC, Akbas E, Kalkan S. Localization recall precision (LRP): a new performance metric for object detection. arXiv:1807.01696. 2018.
24. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. arXiv:1506.01497. 2015.

25. Zhang M, Zhang Z. Research on vehicle target detection method based on improved YOLOv8. *Appl Sci*. 2025;15(10):5546. doi:10.3390/app15105546.
26. Xu L, Li P, Peng Q, Zhao Y, Zhu L, Yuan S. Vehicle detection on roads based on Yolov5 with multi-scale feature fusion. *Array*. 2025;28(19):100522. doi:10.1016/j.array.2025.100522.