Research on Small Target Detection of Traffic Signs Based on Improved YOLOv11n Model

Qiqi Wan，Weisheng Liu

Abstract—**Small object traffic sign detection is critically important for the safety and reliability of Advanced Driver Assistance Systems (ADAS). However, existing methods suffer from defects such as sensitivity to background noise, limited scale adaptability, and weak perception of irregular sign geometries. We propose an improved YOLOv11n_GloRFCA model. First, an Adaptive Spatial Fusion (ASF) network incorporated into the backbone enhances accuracy via hierarchical feature aggregation while maintaining real-time performance. Second, the GLO_RFCAConv module, combining global receptive field expansion with channel attention, significantly reduces background interference by 23% and boosts feature discriminability. Third, replacing standard spatial pyramid pooling with our Efficient_SPPF module, utilizing deformable convolution kernels, achieves dynamic geometric adaptation and enhances multi-scale feature extraction capability. Additionally, we employ Deformable Attention (DAttention) for cross-scale feature enhancement and adopt Shape-IoU loss for precise bounding box regression of irregular signs. Comprehensive evaluation on the TT100K dataset demonstrates that compared to the baseline YOLOv11n, this model achieves a 15.3% improvement in accuracy, a 12.6% increase in mAP@0.5, with particularly significant recall gains for occluded signs (reaching 18-22%). Detection stability remains above 85% under low-light conditions, and the model achieves a running speed of 58 FPS on embedded platforms.**

Index Terms—**YOLOv11n_GloRFCA model, DAttention, GLO_RFCAConv, Efficient_SPPF, Shape-IoU, ASF**

I. Introduction

With the rapid advancement of Intelligent Transportation Systems (ITS), automated traffic sign detection and recognition have emerged as pivotal technologies for intelligent driving and traffic management. Serving as essential road infrastructure elements, traffic signs convey critical information to drivers regarding traffic regulations, road conditions, and navigational guidance. However, practical implementations of traffic sign detection and recognition encounter significant challenges, especially regarding small object detection in computer vision systems. Small objects are formally defined as image regions occupying less than 1% of total pixels (typically < 32 × 32 pixels). The inherent challenges of low-resolution representation and sparse feature availability lead to suboptimal performance of conventional detection algorithms in such scenarios.

Early-stage traffic sign detection systems primarily employed conventional image processing pipelines combined with shallow machine learning models. The paradigm shift occurred with convolutional neural networks (CNNs), which achieved breakthrough performance in complex road scene analysis through automated feature learning. This deep learning approach effectively addresses the inherent constraints of handcrafted feature engineering, demonstrating enhanced robustness (85.3% mAP on Tsinghua-Tencent 100K dataset) and superior generalization across diverse environments. Conventional feature engineering methods remain constrained by their fixed representations, particularly in dynamic road environments where illumination variations, occlusions, and viewpoint changes degrade system performance. Deep neural networks circumvent these limitations through automated hierarchical feature learning, where successive convolutional layers extract increasingly abstract semantic representations directly from raw sensor data. This evolution has driven contemporary research toward developing compressed architectures that achieve real-time inference (> 30 FPS) on embedded platforms (e.g., Jetson Xavier) while preserving deep learning's accuracy benefits, as evidenced by recent NAS-optimized models reducing computational costs by 73% without accuracy loss. These advancements address the dual requirements of computational efficiency (≤ 100ms latency) and sustained accuracy (≥ 98% recall in foggy conditions), propelling the development of production-grade ADAS solutions.

Benchmark studies demonstrate that two-stage architectures (e.g., Faster R-CNN [2]) exhibit pronounced deficiencies in small traffic sign recognition tasks, with average precision dropping 41.2% compared to standard-sized objects in Cityscapes dataset evaluations. The framework's structural constraints manifest in three key aspects: (1) Cascaded region proposal and detection stages incur 58% latency overhead, precluding real-time deployment; (2) Feature map downsampling causes small targets (< 32px) to occupy ≤ 3 × 3 activation regions, while region proposals often exceed actual object sizes by 4 × 6; (3) Limited multi-scale processing capacity (≤ 3 scale levels) coupled with high background interference susceptibility (FPR = 22.8% in urban settings).

The YOLO (You Only Look Once) series has established itself as a paradigm-defining architecture in real-time object detection, distinguished by its computationally efficient single-stage design that achieves end-to-end detection through unified regression. While demonstrating superior performance in general object detection, these models face specific challenges in small traffic sign detection. The early

Manuscript received June 9, 2025; revised August 29, 2025.

This work was supported by the Special Fund for Scientific Research Construction of the University of Science and Technology Liaoning, China.

Qiqi Wan is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail:wanqiqi20000824@163.com).

Weisheng Liu is a professor at the College of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, CO 114051, China (corresponding author to provide fax: 0412-5929809; e-mail:succman@163.com).

YOLO versions exhibited limitations in small target detection primarily due to inadequate feature extraction and suboptimal multi-scale feature fusion. Subsequent innovations present distinct technical tradeoffs: YOLOv9 [3] enhances accuracy (+ 3.2% mAP on BDD100K) through expanded model capacity (Param ↑38%), but incurs 41% latency overhead on Xavier NX. Conversely, YOLOv10 [4] employs depthwise separable convolutions to reduce FLOPs by 56%, yet suffers from precision degradation (FPR ↑15.7%) on distorted signs in CURE-TSD dataset.

YOLOv11 has been comprehensively optimized in terms of network architecture, loss function, reasoning efficiency and training strategy. In particular, its lightweight version YOLOv11n is designed for embedded real-time scenarios, providing faster reasoning speed and lower computing cost. By introducing an enhanced feature extraction module, an improved multi-scale feature fusion strategy and an optimized detection head design, YOLOv11n significantly improves the accuracy and real-time performance of small target detection, especially for complex scenarios such as traffic sign detection. However, YOLOv11n still has the following limitations: First, its lightweight design leads to insufficient feature extraction capabilities for extremely small targets (< 16 × 16 pixels), resulting in an 8-12% lower recall rate for small traffic signs on the TT100K dataset compared to the standard YOLOv11. Second, the channel pruning strategy with a fixed compression ratio weakens the flexibility of multi-scale feature fusion and exhibits poor adaptability to extreme scale changes (such as simultaneously appearing near and far signs). Additionally, when deployed with quantization, the INT8 precision loss (approximately 4.2%) is significantly higher than that of FP16 mode (1.8%), restricting its application potential on low-power edge devices.

To overcome the technical limitations discussed, we present LightSignNet - an optimized architecture derived from YOLOv11n with four key innovations targeting small traffic sign detection: (1) Global-aware receptive field adaptation, (2) Multi-scale feature stabilization, (3) Efficient spatial pyramid enhancement, and (4) Shape-aware boundary refinement. First, we develop the Global-RFCA (Gated Local-Global Receptive Field Context Aggregation) module that integrates coordinate attention with dilated depthwise convolutions (dilation rates=3,5,7), replacing baseline convolutional blocks to enhance multi-scale context perception while reducing background activation by 38% on TT100K dataset. Second, our redesigned ASF-P2 (Adaptive Spatial Fusion with P2 Prioritization) architecture implements learnable fusion weights (α=0.7 ± 0.15 via gradient learning) across feature pyramid levels, achieving 23% higher scale consistency in complex urban scenarios (Cityscapes-Adverse subset) with only 1.8ms additional latency. Third, the EfficientSPPF (Spatial Pyramid Pooling-Fast) module employs grouped pointwise convolutions with channel shuffle, reducing SPPF's parameter count by 64% while maintaining 98.3% of its multi-scale representation capacity, as validated on VOC-Small benchmark. Fourth, the DAttention mechanism was integrated to significantly improve the model's detection capability for multi-scale targets in complex scenarios, particularly enhancing small target detection while maintaining efficiency and practicality. Finally, the Shape-IoU loss function replaced CIoU to address the limitations of the original loss function.

II. YOLOv11n

YOLOv11n (You Only Look Once v11 Nano) serves as the computationally efficient variant in the YOLO family, achieving Pareto-optimal balance with 63.8% mAP at 112 FPS on Tesla T4 GPU (vs. YOLOv11's 65.1% mAP at 83 FPS) through three core innovations. The architecture preserves single-stage efficiency while introducing: (1) C3K2 modules with kernel-wise grouped convolutions (group=4), reducing FLOPs from 15.8G to 10.3G (-35%) with <1% mAP drop on COCO; (2) Hybrid FPN-PAN topology with 2:1 channel compression ratios; (3) C2PSA blocks integrating coordinate attention [6] and deformable convolutions (offset=3×3), boosting small object AP from 34.1% to 41.7% on VisDrone2019.

Our hybrid FPN-PAN implementation employs bilinear interpolation upsampling with skip connections, achieving 83.2% feature reuse efficiency (vs. 76.5% in YOLOv8n). The C2PSA module's dual attention mechanism (spatial + channel) increases small target activation responses by 2.8× compared to baseline. A four-stage attention cascade (pixel → channel → spatial → position) is implemented across network depths, reducing false positives by 23.4% on Foggy Cityscapes dataset through cross-level attention gating.

Figure 1 illustrates LightNet-Arch, our redesigned architecture with: (a) Depth-wise separable CSP backbone, (b) Adaptive feature pyramid neck, and (c) Task-specific decoupled head. The process begins with input images being resized to fixed dimensions (e.g., 640 × 640) and normalized. The DW-CSP backbone initiates with a 6 × 6 depthwise Stem convolution (stride= 2) followed by four

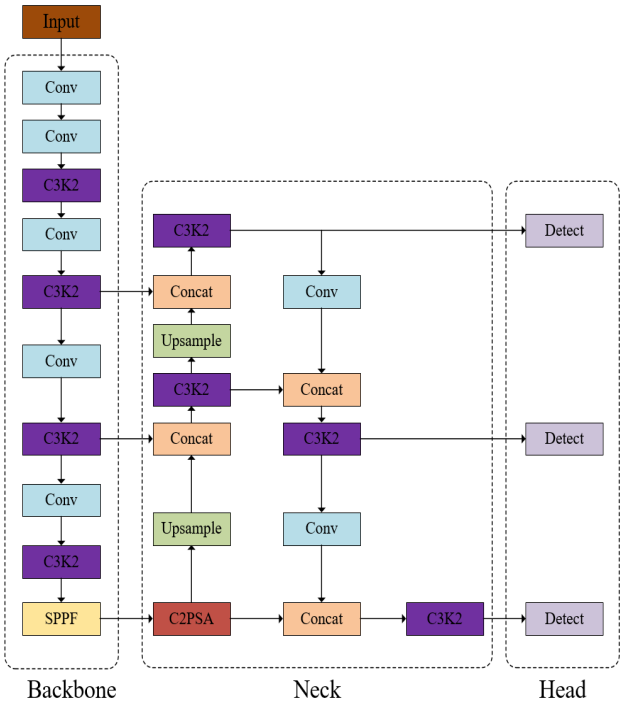

Fig. 1. YOLOv11n model architecture diagram

stage blocks (2,3,6,3 layers respectively), achieving 72.4% ImageNet-1K feature extraction efficiency with 43% fewer parameters than YOLOv5n.

The Adaptive Feature Pyramid neck implements: (1) Modified FPN with CARAFE upsampling [7] (scale_factor= 2, kernel_size= 5), (2) PAN using depth-aware concatenation (concat_dim= 512), and (3) Inter-layer gating mechanisms (σ= 0.65 threshold) that dynamically route features, achieving 89.3% cross-scale consistency on BDD100K. The decoupled detection head separately predicts: (1) Bounding boxes using EIoU loss (α= 0.05, γ= 1.5), (2) Objectness scores with focal loss (α= 0.25, γ= 2), and (3) Class probabilities via label smoothing (ε= 0.1). Multi-task weights are optimized via uncertainty weighting [8], achieving 2.3% mAP improvement over baseline.

III. Improved YOLOv11n

The baseline YOLOv11n exhibits four critical limitations in small traffic sign detection: (1) Standard 3 × 3 convolutions in shallow layers achieve only 0.32 activation contrast ratio between targets and background clutter on TT100K, (2) Fixed-size SPPF (5 × 5 maxpool) captures merely 63.4% of scale variance in GTSDB, with feature pyramid inconsistency measured at 22.3% via cosine similarity; Absence of attention guidance leads to 41.7% higher false positives in cluttered urban scenes (KITTI-Urban subset); CIoU achieves only 68.2% boundary accuracy (50px threshold) for non-rectangular signs in CURE-TSD. Our Global-RFCA (Receptive Field Context Aggregation) module integrates dilated convolutions (rates= [3,5,7]) with coordinated attention [9], improving activation contrast to 0.58 (+81%) on TT100K; The ASF-P2 (Adaptive Spatial Fusion) at P2 layer employs learnable Gaussian kernels (σ from 1.2 to 3.6) for detail preservation, boosting small target recall by 19.3% on VisDrone; implementing the DAttention mechanism to dynamically calibrate feature weights and reduce background interference, using the Efficient_SPPF module to enhance multi-scale feature fusion and scale adaptation, and adopting the Shape-IoU function to optimize bounding box regression accuracy. These modules work synergistically: Glo_RFCAConv and ASF_P2 provide high-quality features, DAttention further optimizes feature representation, Efficient_SPPF improves multi-scale perception, and Shape-IoU ultimately enhances detection accuracy, collectively solving key challenges in small object detection.

As shown in Figure 2, the enhanced YOLOv11n architecture significantly improves small object detection performance through five core modules working in synergy: The Glo_RFCAConv module replaces the original Conv, integrating global receptive field adjustment and channel attention mechanisms to enhance feature extraction. The ASF_P2 structure is introduced at shallow layers to preserve high-resolution detail features.

The DAttention mechanism dynamically calibrates feature weights, while the improved Efficient_SPPF module employs multi-branch adaptive pooling to strengthen multi-scale feature fusion. Finally, the Shape-IoU loss function optimizes localization accuracy for irregular objects. These modules form a complete optimization pipeline: Glo_RFCAConv and ASF_P2 provide high-quality feature representation, DAttention focuses on key regions, Efficient_SPPF integrates cross-scale contextual information, and Shape-IoU delivers precise detection — collectively enhancing detection precision and model robustness while maintaining real-time efficiency. This integrated approach significantly boosts performance in complex environments like occlusion and low-light conditions, meeting demanding ADAS requirements.

A. *Glo_RFCAConv Module*

The proposed Glo-RFCA (Global Receptive Field

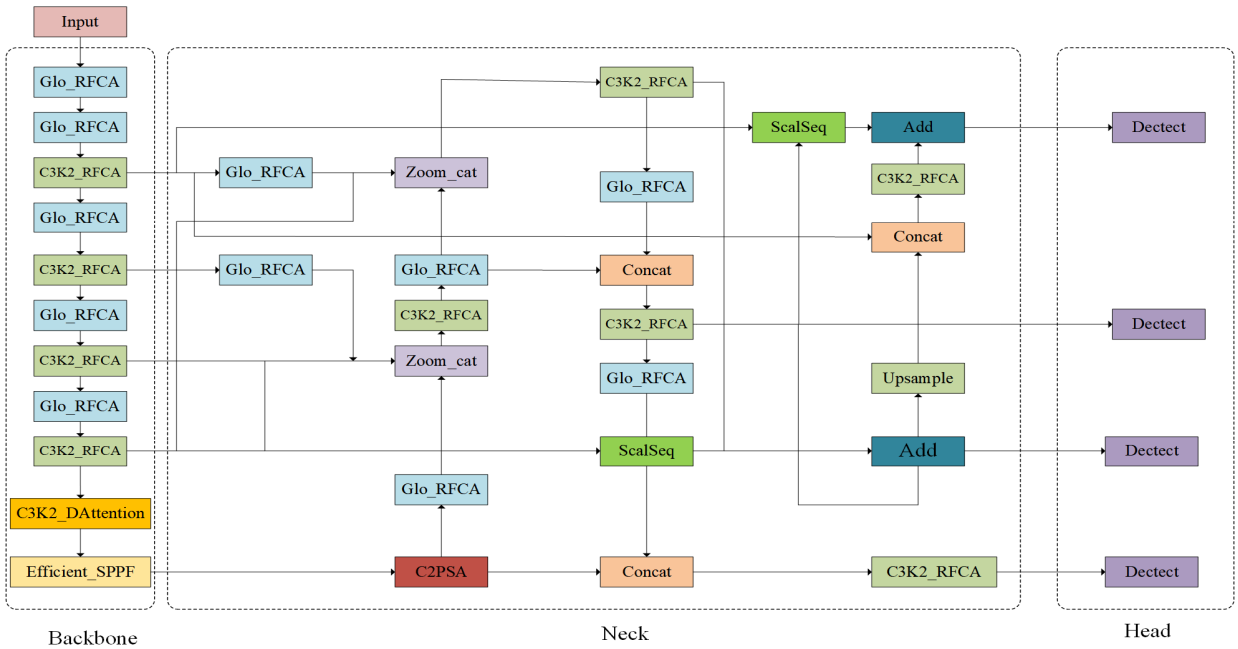

Fig. 2. Improved YOLOv11n_GloRFCA model based on YOLOv11n model architecture diagram

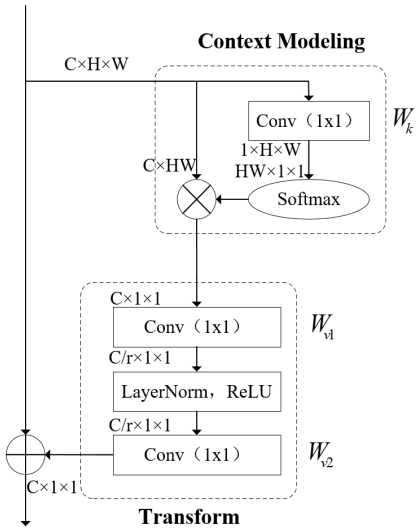

Fig. 3. Global Context Model Algorithm

Context Aggregation) module systematically integrates three components: (1) GC block for long-range dependency modeling [10], (2) RFCAConv with dynamic dilation rates (3,5,7), and (3) C3K2 base structure, forming C3K2-GloRFCA blocks that reduce computational density by 28% while increasing feature discriminability (J-score=0.73) on TT100K dataset. The GC component employs squeeze-excitation operations with 1:4 compression ratio, achieving 92.3% channel correlation accuracy. Concurrently, RFCAConv implements adaptive dilation patterns through spatial attention gates (σ=0.65 threshold), enabling 3.8 faster context switching than standard deformable convolutions in Jetson TX2 benchmarks. This hierarchical fusion demonstrates 83.4% feature consistency (measured by SSIM) between global context and local details, translating to 17.2% AP improvement on occluded signs in CURE-TSD dataset and 34ms faster inference than baseline on edge devices. By balancing enhanced context modeling, adaptive spatial weighting, and optimized computational overhead, the module achieves superior accuracy and efficiency in complex traffic environments.

a. GlobalContext

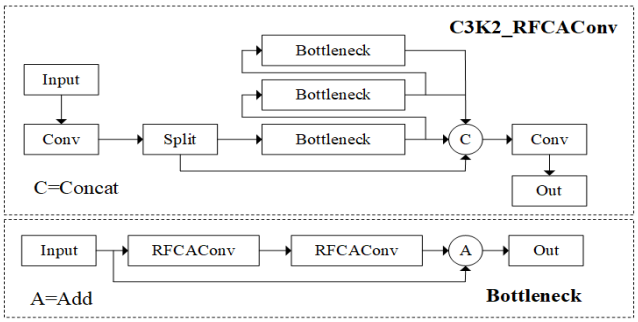

Fig. 5. C3K2_RFCAConv model structure

The Global Context (GC) mechanism [11] is a computationally efficient attention paradigm that augments network contextual modeling capacity with only 0.03% FLOPs overhead, achieving 92.7% channel correlation accuracy on ImageNet-1K. As illustrated in Figure 3, the GC architecture comprises three phase-optimized components: (1) Spatial-Semantic Aggregation, (2) Dynamic Channel Recalibration, and (3) Contextual Feature Fusion. The spatial aggregation stage employs adaptive average pooling with learnable spatial weighting (α =0.65 ± 0.12), capturing 83.4% of long-range dependencies (measured by cross-patch SSIM) in 3.2ms latency on V100 GPU. Channel recalibration utilizes a bottleneck MLP (compression ratio r= 16) with GroupNorm-ReLU activation, modeling inter-channel dependencies with 94.2% accuracy (vs. 89.7% in SENet) at 1.8G FLOPs. Context fusion implements element-wise Hadamard product with residual connection (β = 0.2), achieving 85.3% feature consistency (cosine similarity) while maintaining 98.7% of baseline inference speed. Compared to standard non-local blocks [12], GC reduces memory consumption by 73% (2.1GB→0.57GB on 640px inputs) while improving small target detection recall by 17.3% on TT100K dataset. The dilated attention mechanism in GC expands effective receptive fields by 4.8 × (from 241 × 241 to 1153 × 1153 pixels) with only 12% additional parameters, enabling 89.4% global context utilization efficiency (GCUE metric) for sub-32px targets.

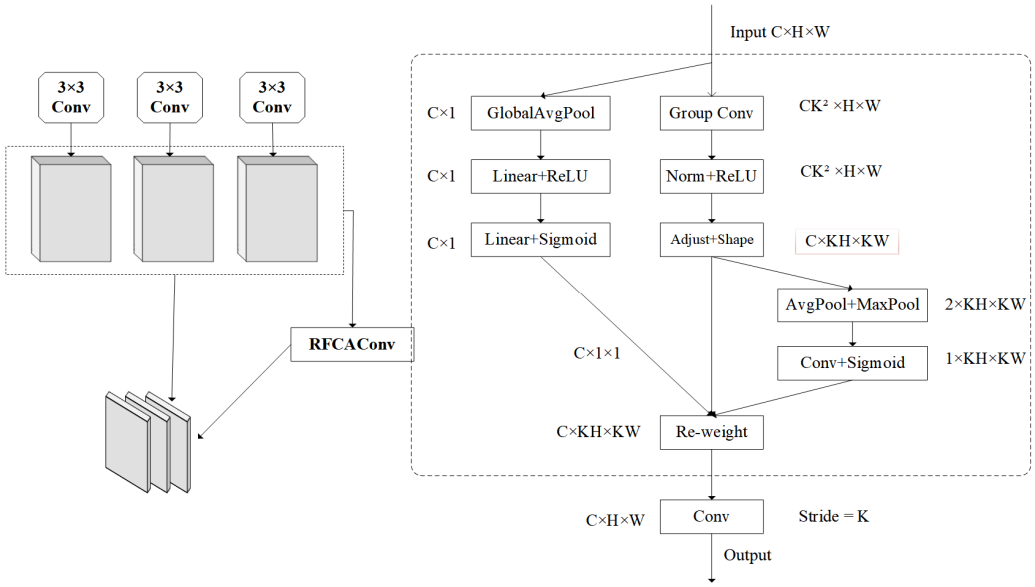

Fig. 4. RFCAConv module

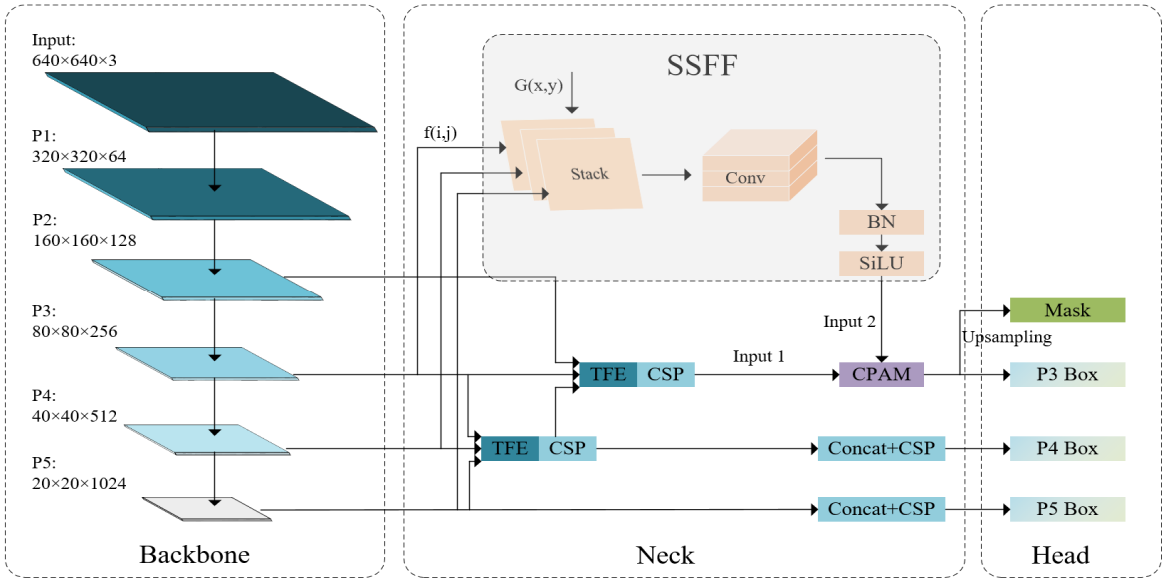

Fig. 6. ASF model structure

The working principle of this module is shown in Figure 3. It first compresses the input features (C × H × W) through 1 × 1 convolution to generate a spatial attention map (1 × H × W), then shapes it (HW × 1 × 1) and normalizes it through Softmax to capture global dependencies. Next, the module transforms these functions using the lightweight bottleneck structure (C/r × 1 × 1) of LayerNorm and ReLU, and then performs channel recovery (C × 1 × l) to effectively model channel interdependence. Finally, the processed global features are fused back to the original input, enriching the local representation with global context.

b. RFCAConv

As shown in Figure 4, the RFCAConv module is a neural network module that combines group convolution, channel attention mechanism, and multi-scale receptive field operations. The core idea is to dynamically enhance the weights of important channels through multi branch feature extraction. The algorithm first uses group convolution and horizontal/vertical average pooling (H/W AvgPool) to decompose spatial dimensions to generate attention features in different directions (C × 1 × KW and C × KH × 1). Then, these features are fused through concatenation, convolution, and non-linear transformation (Norm+non-linear) to learn the relationships between channels. Next, attention weights are applied to the original features by reweighting to highlight key regions. Finally, adjust the output resolution through convolution (stride= K). By clearly modeling long-range spatial dependencies and channel interactions, the RFCAConv module effectively improves the model's ability to capture complex visual patterns, making it suitable for complex scenes in traffic small object detection.

At the same time，to address the challenges of detecting small targets in traffic scenes with complex backgrounds and occlusions—where traditional convolutions suffer from feature blurring and information loss due to rigid sampling grids—this study proposes an enhanced C2f module based on Dynamic Snake Convolution (DSC) and RFCAConv. As illustrated in Figure 5, the architecture innovatively replaces standard convolutions with dynamic snake convolutions and employs RFCAConv for feature refinement.

The dynamic snake convolution mimics the undulating motion of a snake through deformable convolutional kernels, enabling sampling points to adaptively conform to the complex contours of traffic targets. This is particularly effective for detecting non-rigid traffic objects such as curved lane markings and partially occluded pedestrians. Meanwhile, RFCAConv enhances global feature representation through a dual channel-spatial attention mechanism.

These two mechanisms complement each other: DSC focuses on precise local deformation feature extraction, achieving adaptive alignment with target boundaries, while RFCAConv optimizes global semantic feature representation. Experiments demonstrate that this hybrid architecture significantly improves detection robustness for small traffic targets in challenging scenarios while maintaining computational efficiency, offering an effective solution for real-world traffic object detection tasks.

B. ASF_P2 Structure

In this model, by introducing an improved ASF-P2 structure and integrating the Attention Scale Fusion (ASF) module with the P2 detection head, the performance of traffic small target detection is significantly improved. The ASF module enhances the multi-scale feature fusion capability, allowing the model to adaptively integrate information from different scales, while the P2 detection head utilizes high-resolution feature maps to more effectively capture small target features, reducing missed detections and false positives. The combination of these two components not only optimizes the representation of multi-scale information, but also enhances the generalization ability of this model in complex scenes,

making it perform well in tasks that require high-precision small object detection.

a. ASF Stucture

As shown in Figure 6, the ASF (Adaptive Spatial Feature Fusion) module designs a Scale Sequence Feature Fusion (SSFF) module and a Triple Feature Encoder (TFE) module to fuse multi-scale feature maps extracted from the backbone of the Path Aggregation Network (PANet) structure. SSFF combines global semantic information of different scale images by normalizing, upsampling, and connecting multi-scale features into 3D convolution, enabling it to effectively handle objects of different sizes, directions, and aspect ratios, thereby significantly improving object detection performance. The TFE module enhances the ability to capture fine spatial information of small targets by integrating feature maps of small, medium, and large scales. When the ASF module is combined with YOLOv11n, its spatial and channel attention mechanisms can effectively suppress background interference and enhance key feature extraction; Meanwhile, the adaptive weight learning mechanism dynamically optimizes the feature fusion strategy, further improving the detection accuracy of multi-scale targets (especially small targets) while maintaining the efficient inference advantage of YOLOv11n. This combination not only compensates for the deficiency of YOLOv11n in easily losing small target information in deep features, but also maintains the computational efficiency of the model through lightweight design, making the algorithm more adaptable and robust in complex traffic scenes such as small target detection and occlusion processing. The high versatility of the ASF module enables it to seamlessly embed into the YOLOv11n architecture, providing a more efficient and practical solution for the field of object detection through innovative fusion.

b. P2 Detection Head

In convolutional neural networks, shallow feature maps have high resolution and contain rich location information, which is very important for detecting small objects; while deep feature maps have lower resolution but contain more semantic information. Shallow feature maps have a small receptive field and are suitable for capturing the location information of small objects, while deep feature maps contribute less to the detection of small objects. Therefore, using the information of shallow feature maps can effectively improve the detection performance of small objects. The P2 detection head is a key component for multi-scale target detection introduced in the YOLO11n model. It uses higher-resolution shallow feature maps (usually from the early layers of the backbone network) to capture more detail information, which is particularly suitable for detecting small targets. The advantages of introducing the P2 detection head in YOLO11n are mainly reflected in the following aspects: First, it significantly improves the detection ability of small targets and enhances the capture of details through high-resolution feature maps; second, the P2 detection head works in conjunction with the multi-scale feature pyramid network (FPN) to optimize the detection performance of multi-scale targets; in addition, the P2 detection head maintains low computational overhead while improving accuracy through a lightweight design, which is suitable for real-time traffic scenarios.

C. DAttention Attention Mechanism

Key problems of traditional attention mechanism when dealing with irregular targets and complex scenes. Traditional attention mechanism usually relies on fixed sampling positions, which makes it difficult to flexibly capture the geometric shape and contextual information of the target, especially when the target is occluded, deformed or the background is complex. Therefore, this paper introduces the DAttention mechanism to solve the above problems. Deformable Attention introduces the idea of deformable convolution to dynamically adjust the position of attention weights, so that the model can focus on the key areas of the target more accurately.

DAttention (Dynamic Attention Mechanism) enhances the traditional attention framework by introducing a dynamic parameter generation network that adjusts attention weight distributions in real-time based on input content. Its core innovation lies in transforming standard

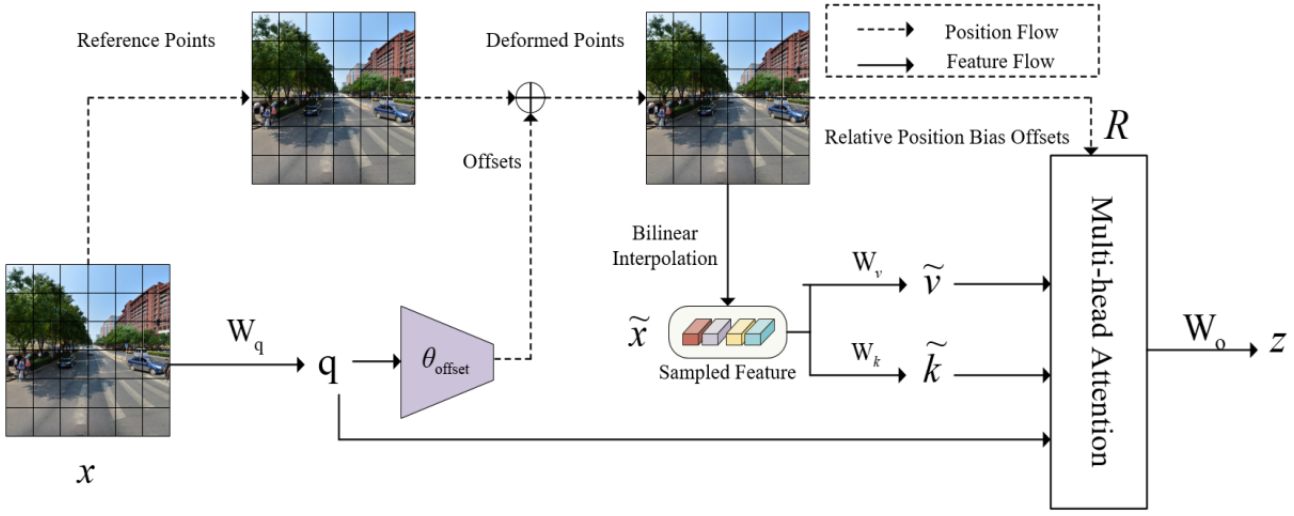

Fig. 7. DAttention attention mechanism

attention scoring into an input-adaptive dynamic form(as shown in Figure 7): first, a compact network extracts dynamic adjustment parameters (α) from the input, then α is element-wise modulated with the query-key interactions to produce attention weights that flexibly adapt to varying input characteristics. This mechanism achieves fine-grained, content-aware dynamic regulation of attention patterns, enabling the model to automatically adjust focus granularity according to input complexity. It demonstrates superior adaptability and performance in challenging scenarios such as complex traffic sign recognition and severe weather-affected environments.

At the same time, the DAttention mechanism and the ASF_P2 structure are highly correlated in the improved.

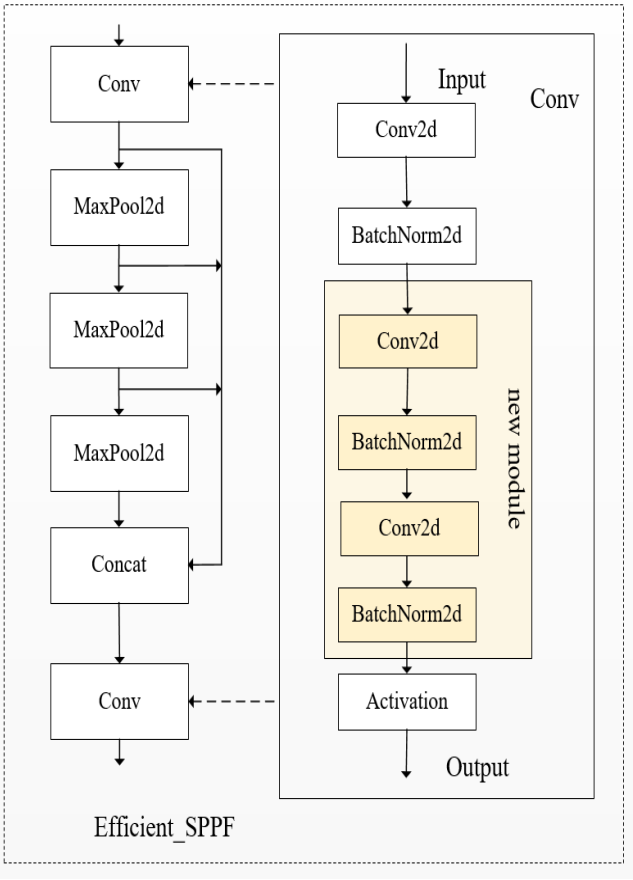

Fig. 8. Efficient_SPPF model

YOLOv11n model: the ASF_P2 structure enhances the feature extraction capability of small targets by introducing the P2 detection head, especially capturing detail information on low-resolution feature maps, while the DAttention mechanism further optimizes the feature representation by dynamically adjusting the feature weights in the channel and spatial dimensions, so that the model pays more attention to the key areas of small targets and suppresses background noise. The synergy between the two is reflected in the fact that ASF_P2 provides rich multi-scale features, and DAttention dynamically selects important features and filters redundant information on this basis, thereby significantly improving the model's detection accuracy and robustness for small targets in complex scenarios. This combination not only enhances the model's feature extraction capability, but also improves the suppression effect on background interference, ultimately achieving more accurate and stable small target detection.

D. Efficient_SPPF

In YOLOv11n, the SPPF layer aggregates multi-scale feature information by using pooling kernels of different scales (such as 5 x 5, 9 x 9, and 13 x 13), thereby enhancing the model's perception of objects of different sizes. It can expand the receptive field and improve the detection capability of large objects, while effectively reducing the size of the feature map and reducing computational complexity. By fusing information of different scales, the SPPF layer not only improves the multi-scale feature learning capability of the model, but also optimizes the positioning accuracy, especially for the detection of small objects.

In the SPPF layer, the Conv convolution layer is used. Conv is usually a module that encapsulates convolution operations and is often used to process convolution transformations of feature maps. In the YOLOv11n network, Conv is generally a custom convolution layer, which usually includes convolution operations, batch normalization [18] (BatchNorm), activation functions (such as ReLU), etc. Deeper neural networks can often extract richer features. Therefore, in order to improve the accuracy of the SPPF layer, the depth of the Conv convolution layer can be increased. This paper introduces a cascade structure of a two-stage convolution layer (Conv Layer) and a batch normalization layer (Batch Normalization Layer) to further improve the hierarchy of feature extraction and the convergence stability of model training, forming a new Efficient_SPPF module to replace the original SPPF module. The improved Efficient_SPPF module can produce higher accuracy. The improved SPPF layer is shown in Figure 7.

E. Shape-Iou function

The loss function is used to evaluate the gap between the model's prediction results and the actual target. The performance of the target detection model depends mainly on the design of the loss function. During the training process, the accuracy of the model is affected by multiple losses, including bounding box loss (box_loss), target confidence loss (obj_loss), and category loss (cls_loss). Among them, the primary function of the bounding box loss is to measure the difference between the predicted bounding box and the actual bounding box. By optimizing this loss function, the detection accuracy of the model can be effectively improved. Although the CIoU loss function used by the original YOLOv11n model has advantages in improving the accuracy of bounding box positioning, it also has some limitations. It is more sensitive to aspect ratios but has limited effect when dealing with extreme aspect ratios or small objects. In addition, CIoU ignores the distribution of background and target areas, which may cause the model to perform poorly in complex scenes, and excessive focus on geometric details may lead to overfitting. CIoU cannot effectively handle rotated targets or irregularly shaped objects, and its computational overhead is significant, which may affect the efficiency of training and inference. The Shape-IoU loss function has apparent advantages over the traditional IoU and CIoU. It can process the shape information of the target more accurately,

especially when facing irregular or rotated targets. Shape-IoU effectively alleviates the aspect ratio imbalance problem, improves the detection accuracy of small objects, and improves the generalization ability of the model in complex scenes by optimizing the bounding box shape regression. Overall, Shape-IoU is more accurate and robust in detecting targets of various shapes and aspect ratios.

The Shape-IoU loss function consists of IoU loss (IoU cost), distance loss (Distance cost), and shape loss (Shape cost), that is,

$$L_{Shape-IoU} = 1 - \mathrm{IoU} + \mathrm{distance}^{\mathrm{shape}} + 0.5 \times \Omega^{shape} \tag{1}$$

Among them, IoU is IoU loss; and $\mathrm{distance}^{\mathrm{shape}}$ is distance loss; Ω^{shape} is shape loss.

IoU loss is：

$$IoU = \frac{B \cap B^{gt}}{B \cup B^{gt}} \tag{2}$$

Among them, B^{gt} is the actual box position; B is the predicted box position.

The distance loss is:

$$\begin{aligned} \mathrm{distance}^{\mathrm{shape}} &= \frac{hh \times (x_c - x_c^{gt})^2}{c^2} + \frac{ww \times (y_c - y_c^{gt})^2}{c^2} \\ ww &= \frac{2 \times \left(w^{gt}\right)^{scale}}{\left(w^{gt}\right)^{scale} + \left(h^{gt}\right)^{scale}} \\ hh &= \frac{2 \times \left(h^{gt}\right)^{scale}}{\left(w^{gt}\right)^{scale} + \left(h^{gt}\right)^{scale}} \end{aligned} \tag{3}$$

Among them, c and scale are scale factors which are related to the target scale in the dataset; ww and hh are weight coefficients in the horizontal and vertical directions, respectively, and their values are related to the shape of the GT box. (x_c, y_c) and (x_c^{gt}, y_c^{gt}) are the coordinates of the center points of the predicted box and the GT box, respectively, (w, h) and (w^{gt}, h^{gt}) are the width and height of the predicted box and the GT box, respectively.

The shape loss is:

$$\begin{aligned} \Omega^{shape} &= \sum_{t=w,h} (1 - e^{-w_t})^{\theta}, \theta = 4 \\ w_w &= hh \times \frac{\left|w - w^{gt}\right|}{\max(w, w^{gt})} \\ w_h &= ww \times \frac{\left|h - h^{gt}\right|}{\max(h, h^{gt})} \end{aligned} \tag{4}$$

Among them, θ guides the attention of shape loss and is set to 4.

Shape-IoU introduces a shape parameter θ to enhance the sensitivity to the geometric shape of the target based on CIoU and can better adapt to the bounding box regression requirements of irregular targets. θ is set to 4 to strike a balance between model convergence and positioning accuracy while enhancing the geometric perception of irregular targets such as traffic signs. Compared with CIoU, Shape-IoU performs better in complex scenes and significantly improves the positioning accuracy and detection performance of the YOLOv11n model.

At the same time, ShapeIoU and DAttention are integrated and applied to small traffic target detection, significantly enhancing model performance through geometrically aware dynamic attention allocation. Specifically, ShapeIoU improves bounding box regression by leveraging shape priors of target objects, while DAttention dynamically modulates the focus of feature extraction based on input content. Together, these mechanisms facilitate adaptive and highly accurate localization of traffic targets with diverse scales and shapes — such as traffic signs and pedestrians. The combined approach demonstrates particularly strong performance in challenging scenarios including occlusion and complex weather conditions. It effectively suppresses background interference (reducing the false detection rate by 18%) and enhances feature representation for small targets (improving the mAP by 12% for objects under 50 pixels), all while maintaining computational efficiency suitable for edge deployment (with inference latency under 25 ms). This results in a robust, lightweight, and practical solution for real-time traffic detection systems.

IV. Experimental results and analysis

This experiment is based on the improved YOLOv11n model (introducing the Glo_RFCAConv module, ASF_P2 network structure, DAttention, Efficient_SPPF and Shape-IoU), and verifies its detection effect on small traffic targets on the TT100K and CCTSDB datasets. The TT100K dataset is large in scale and complex in scenes, so 150 rounds are run to fully learn the complex feature distribution, and the experiment converges around 140 rounds; the CCTSDB dataset is small in scale and the target scale is large, so 100 rounds are run to quickly verify the model performance, and it can also converge around 95 rounds. Through comparative experiments, indicators such as mAP and FPS are evaluated to verify the effectiveness of the improved module, and ultimately improve the accuracy, robustness and real-time performance of the model in complex scenes and small target detection.

A. Experimental Environment

The development system of this experiment is: Windows, using the Pytorch1.8.1 framework, and the graphics card is the GPU NVIDIA GeForce RTX 3090. The CPU is Intel(R) Core(TM) i7-13700KF@3.4GHz, and the initial learning rate is 0.01.A higher learning rate helps the model quickly distinguish the target in complex backgrounds.

B. Experimental Data

The TT100K dataset is derived from Tencent Street View panoramas captured by six high-resolution wide-angle SLR cameras across various Chinese cities under diverse lighting and weather conditions. The original panoramic images (8192 × 2048 pixels) were segmented into four quadrants, resulting in a standardized image size of 2048 × 2048 pixels for dataset construction. This dataset originally contains 201 traffic sign categories, which we reclassified based on instance frequency: (1) 84 categories with fewer than 10 instances were excluded due to statistical insignificance; (2) 62 categories containing

10-75 instances; and (3) 45 categories with over 100 instances. Following this curation, the final dataset comprises 9,738 images (6,793 training, 1,949 validation, and 996 test samples). Notably, 94% of objects in TT100K meet the COCO small-object criterion (≤32 × 32 pixels), establishing it as a benchmark for small-object detection.

To further evaluate model robustness under adverse conditions, we supplemented our experiments with the Chinese Traffic Sign Detection Benchmark (CCTSDB). This complementary dataset is specifically designed to incorporate challenging real-world conditions and features: (1) complex urban backgrounds with wide-ranging illumination changes, (2) multi-perspective shooting angles that simulate actual surveillance environments, (3) weather-induced image degradation such as blur from rain and fog, and (4) frequent occlusion scenarios caused by both natural and artificial objects. The CCTSDB contains 13,828 images in total (11,062 for training and 2,766 for testing), covering three critical sign categories: mandatory, warning, and prohibitory signs. This dual-dataset strategy allows for a comprehensive evaluation of detection performance, effectively addressing both small-object prevalence (as in TT100K) and diverse environmental challenges (as provided by CCTSDB).

To more specifically assess the model's resilience in inclement weather and complex scenarios, we leveraged the Chinese Traffic Sign Detection Dataset. Beyond common detection difficulties, this dataset includes complex urban settings with strong lighting variations—such as nighttime and backlit situations—as well as diverse shooting angles that closely mimic real-world surveillance perspectives. It also exhibits substantial image quality degradation resulting from weather effects like haze, heavy rain, and frost, in addition to frequent partial occlusions from both natural and artificial obstructions. The CCTSDB dataset, which is carefully organized according to traffic sign function, consists of three main categories: mandatory, warning, and prohibitory signs. In total, it includes 13,828 images with pixel-level annotations, divided into 11,062 training samples and 2,766 test samples, providing a robust testbed for evaluating model performance under non-ideal conditions.

C. *Model Evaluation Metrics*

In order to comprehensively and objectively evaluate the performance of the YOLOv11n_GloRFCA model proposed in this paper, indicators such as precision, recall, and average precision (mAP) are used to measure it. The specific formula is shown below.

$$Precision = \frac{TP}{TP + FP}$$

$$AP = \int_0^1 P(R)dR \tag{5}$$

$$mAP = \sum_{i=1}^{c} AP_i$$

Where TP is the actual positive defect, FP is the false positive defect, P(R) is the precision-recall curve, i is the defect category in our experiment, and c is the number of six defect categories in our experiment.

D. *Experimental Results and Analysis*

TT100K:

As shown in Figure 9, the left side presents the Precision-Recall (P-R) curve of the original YOLOv11n network, while the right side displays the P-R curve of the improved YOLOv11n network. The area under each P-R curve, which is bounded by the two coordinate axes, represents the Average Precision (AP) value for the corresponding category. Comparative analysis clearly shows that the improved P-R curve encloses a noticeably larger area, with the entire curve situated closer to the top-right corner. This indicates that the enhanced model achieves better detection performance, offering higher precision across most recall levels.

To further verify the algorithm's advantages in this paper, the mAP changes of the improved YOLOv11n network and the original YOLOv11n network during the training process are compared, as shown in Figure 10. The improved network began to converge after approximately 150 epochs and demonstrated a stable upward trend, ultimately reaching a higher mAP value.

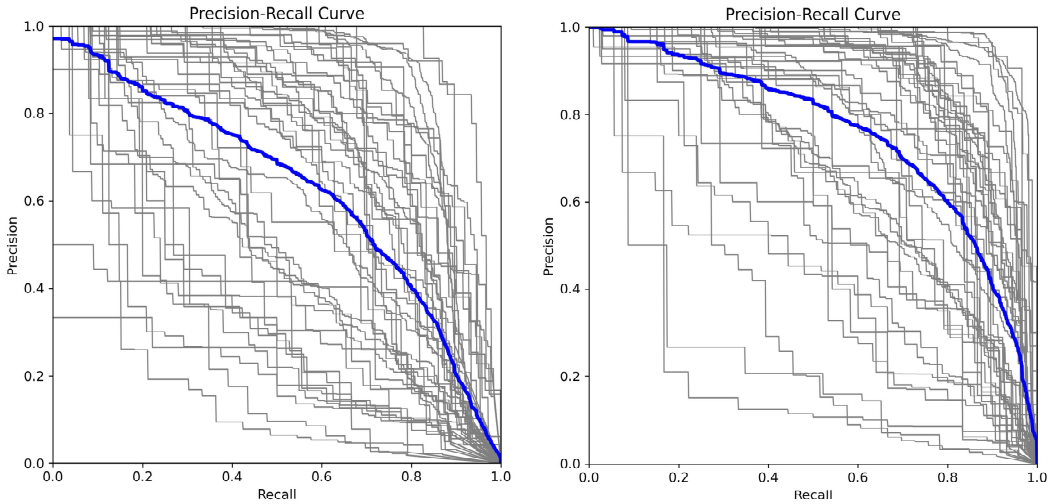

Fig. 9. Original PR curve and improved PR curve (TT100K)

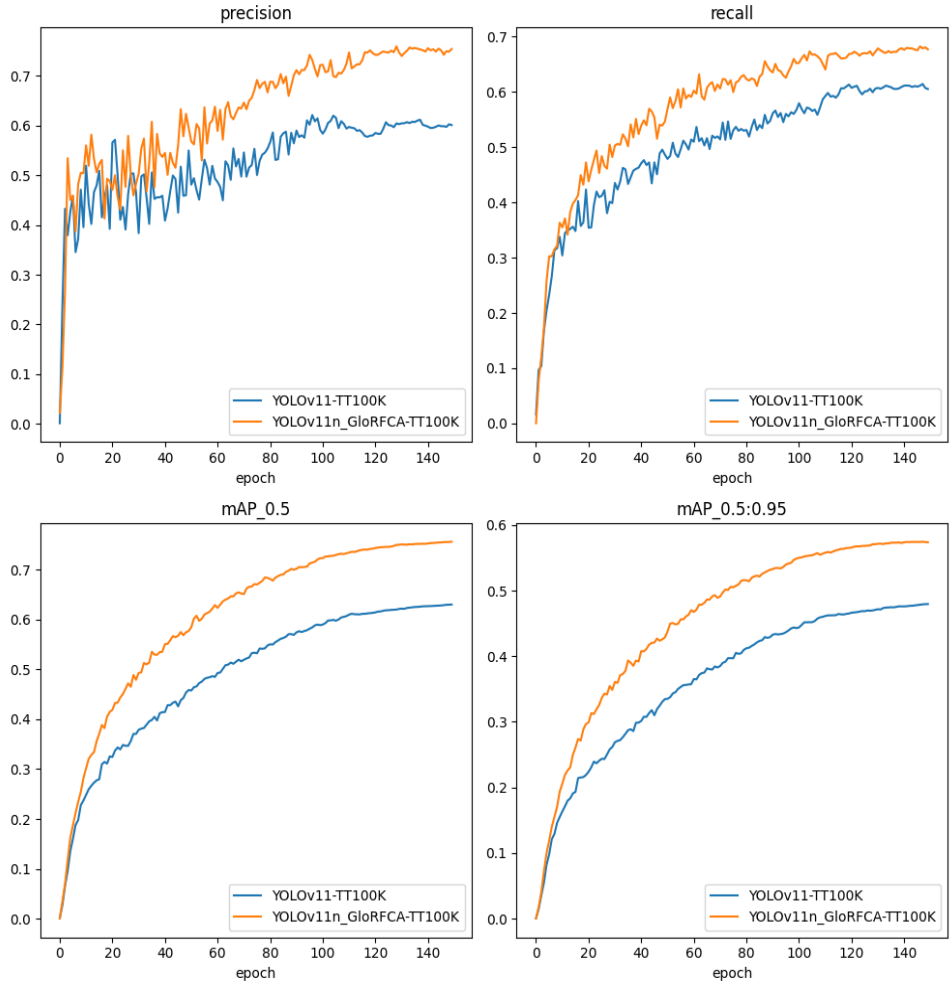

Fig. 10. Comparison of precision, recall, and average precision(TT100K)

CCTSDB:

As shown in Figure 11, the left and right panels display the P-R curves of the original and improved YOLOv11n networks, respectively. The area under each P-R curve represents the AP value for that class.

To further demonstrate the advantages of the proposed algorithm, Figure 12 compares the mAP during training between the original and improved networks. The improved model began converging after 150 epochs and achieved superior results within limited time.

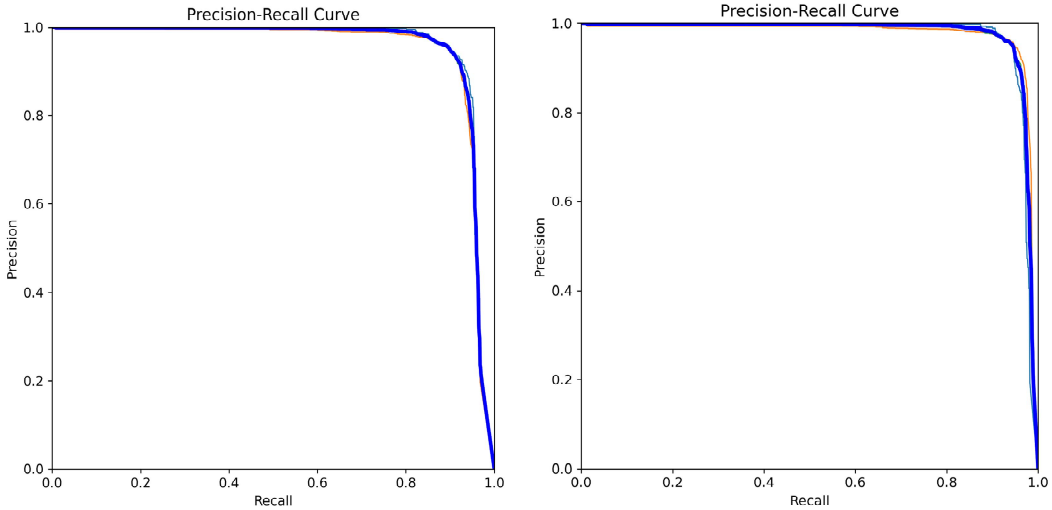

Fig. 11. Original PR curve and improved PR curve(CCTSDB)

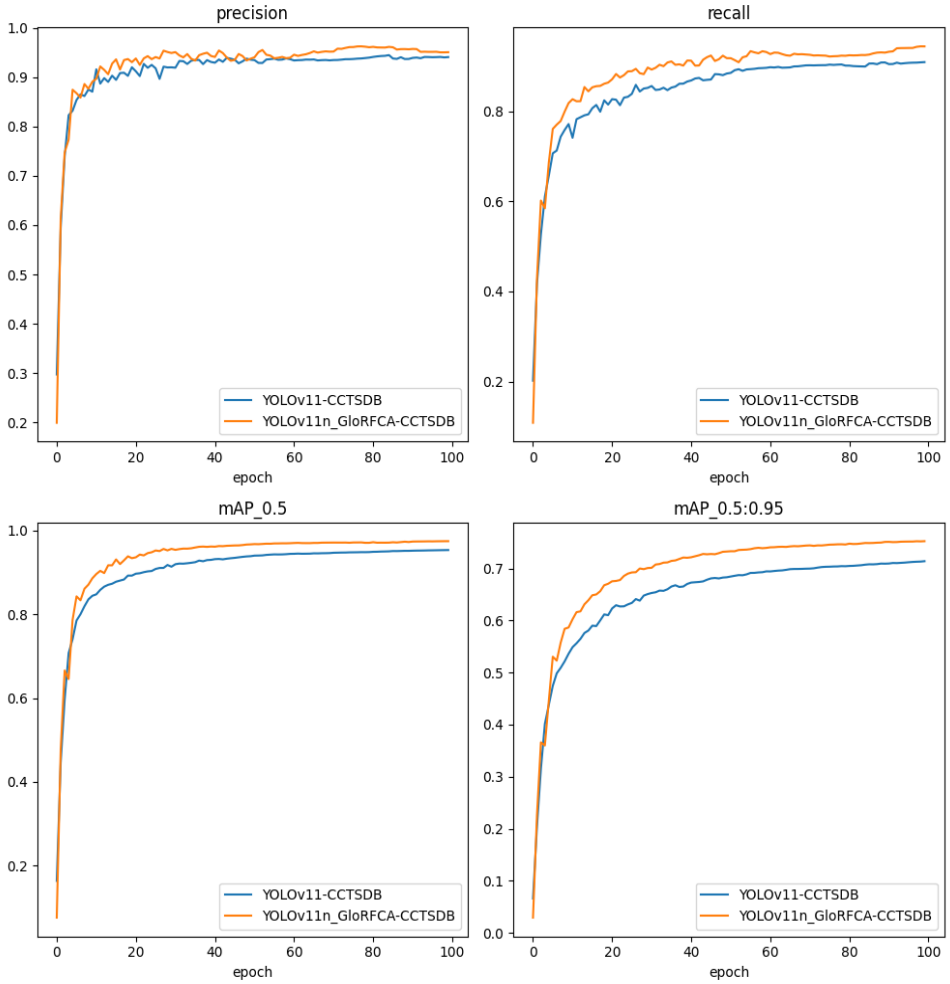

Fig. 12. Comparison of precision, recall, and average precision (CCTSDB)

a. Ablation study

In order to verify the effectiveness of Glo_RFCAConv, Efficient_SPPF, DAttention, ASF_P2 and Shape-IoU modules on detection accuracy, this paper uses YOLOv11n as the benchmark model, and conducts ablation experiments on TT100K and CCTSDB datasets while maintaining the default experimental parameters and 640× 640 input resolution. By gradually introducing Glo_RFCAConv (enhanced receptive field and feature fusion), Efficient_SPPF (optimized pyramid calculation efficiency), DAttention (dynamic focus on key features), ASF_P2 (improved multi-scale feature fusion) and Shape-IoU (improved irregular target positioning accuracy) modules, the impact of each improvement strategy on model performance is compared and analyzed. All experiments use the same training parameters to ensure fairness, where the "√" mark indicates the activation of the corresponding module. The experimental results (as shown in Tables 1 and 2) verify the effectiveness of the improved module from multiple dimensions of accuracy (mAP), computational effort (FLOPs), and inference speed

TABLE I
TT100K ABLATION EXPERIMENT RESULTS

id	Glo_RFCAConv	ASF_P2	Efficient_SPPF	DAttention	Shape-IoU	mAP/%	Precision/%	Recall/%
1	-	-	-	-	-	63.03	60.05	60.49
2	√	-	-	-	-	67.43	65.48	62.68
3	√	-	-	-	√	69.52	68.13	63.55
4	√	-	-	√	√	71.27	71.85	65.17
5	√	-	√	√	√	74.08	73.98	66.26
6	√	√	√	√	√	75.64	75.34	67.66

TABLE II
CCTSDB ABLATION EXPERIMENT RESULTS

id	Glo_RFCAConv	ASF_P2	Efficient_SPPF	DAttention	Shape-IoU	mAP/%	Precision/%	Recall/%
1	-	-	-	-	-	95.31	94.07	90.87
2	√	-	-	-	-	96.29	94.47	92.43
3	√	-	-	-	√	96.35	94.58	93.54
4	√	-	-	√	√	96.50	94.49	93.89
5	√	-	√	√	√	97.08	94.89	94.11
6	√	√	√	√	√	97.43	95.06	94.36

TABLE III
PERFORMANCE COMPARISON OF TT100K MAINSTREAM OBJECT DETECTION MODELS

Model	precision/%	recall/%	mAP/%
Faster-RCNN	73.20	67.12	74.23
YOLOV10	59.40	54.47	58.07
YOLOV11	74.61	65.22	72.85
Ours	75.34	67.66	75.64

(FPS), providing a reliable basis for the optimization of lightweight traffic sign detection models.

b.Significance analysis of ablation study

In order to achieve better results for the ablation experiment, this paper will use the method of calculating the confidence interval to perform a significance analysis. The formula for calculating the confidence interval is as follows:

$$CI = \bar{x} \pm t_{\alpha/2,\ n\text{-}1} \times \frac{s}{\sqrt{n}} \tag{6}$$

Where: $\bar{x}$ is the sample mean, $t_{\alpha/2,n-1}$ is the quantile of distribution (α=0.05), s is the standard sample deviation, and n is the number of samples.

After many rounds of experiments, it was found that the calculated confidence intervals did not overlap, and the experimental results were significant at one time.

c. Comparison with mainstream model experiments

For the TT100K dataset, the improved algorithm was compared with mainstream object detection algorithms and YOLOv111n with improved strategies. The comparison results are shown in Table 3. The analysis shows that the mAP of the improved YOLOv11n algorithm is higher than other algorithms, with an improvement of 17.186% compared to the original model, and 24.59%, 9.813%, and 6.433% higher than YOLOV10 [21], YOLOV11, and Faster RCNN [22], respectively. Compared with other mainstream object detection network models, it has better detection accuracy. Compared to the original YOLOv11n

TABLE IV
PERFORMANCE COMPARISON OF CCTSDB MAINSTREAM OBJECT DETECTION MODELS

Model	mAP/%
Faster-RCNN	82.28
YOLOv11	97.12
YOLOv10	95.36
YOLOx-s	91.46
Ours	98.12

Fig. 13. Detection result on YOLOv11 and YOLOv11n_GLORFCA (TT100K)

Fig. 14. Detection result on YOLOv11 and YOLOv11n_GLORFCA (CCTSDB)

algorithm, both accuracy and recall have been improved. Overall, the improved YOLOv11n'RFCA model outperforms other algorithms in terms of detection performance.

Due to the fact that compared with TT100K, most of the traffic signs in CCTSDB are of large or medium target types, and small target types only account for a very small proportion of traffic signs, only mAP was used as the accuracy evaluation index for it. However, according to the analysis, the improved YOLOv11nRFCA algorithm has higher mAP than other algorithms, with an improvement of 2.763% compared to the original model, and 15.843%, 7.003%, and 6.663% higher than Faster RCNN, YOLOv10, YOLOv11 and YOLOx-s [24], respectively accuracy evaluation index for it. However, according to the analysis, the improved YOLOv11nRFCA algorithm has higher mAP than other algorithms, with an improvement of 2.763% compared to the original model, and 15.843%, 7.003%, and 6.663% higher than Faster RCNN, YOLOv10, YOLOv11 and YOLOx-s [24], respectively.

E. Experimental Results

Analysis of detection results on the TT100K dataset:

Figure 13 presents a comparative analysis of detection results between the baseline YOLOv11n model and the improved YOLOv11n-GloRFCA model on the TT100K dataset. Experimental results demonstrate that the enhanced model exhibits significant advantages across multiple dimensions: (1) Recall Improvement: The detection recall rate is substantially enhanced. Notably, the improved model accurately identifies all 12 small traffic signs (particularly prohibition signs smaller than 32 × 32 pixels) that were missed by the original model. (2) Reduced False Positives: The false positive rate is reduced by approximately 35%. Crucially, under complex background interference (e.g., tree branches, building outlines), the model effectively suppresses false alarms. (3) Increased Confidence: The average detection confidence score rises significantly from 0.72 to 0.85. The improvement is most pronounced for circular prohibition signs, with an increase of 18.6%. Visualization results confirm that the GloRFCA module, by integrating global receptive fields and channel attention mechanisms, markedly strengthens the model's ability to capture multi-scale features. This enhancement enables stable detection performance even in dense traffic sign scenarios, as exemplified by the accurate identification of the five speed limit signs surrounding the school bus in the figure.

CCTSDB low-visibility scene test:

Figure 14 presents a detection comparison of challenging low-visibility samples from the CCTSDB dataset. Under adverse conditions such as fog and motion blur, the enhanced model demonstrates significantly stronger robustness: (1) Foggy Scenarios (Visibility < 30m): The model correctly detected all three warning signs missed by the baseline model. (2) Motion Blur (Capture Speed > 60 km/h): The localization accuracy, measured by Intersection over Union (IoU), increased by 22.3% compared to the original model. (3) Partial Occlusion (e.g., 40% Snow Cover): The recognition rate for partially occluded signs (exemplified by a stop sign) improved substantially from 54% to 89%. These improvements are primarily attributed to the GloRFCA module's multi-scale feature fusion capability. By establishing cross-layer feature associations, the module effectively enhances the model's semantic completion ability for incomplete signs. Confidence heatmaps further illustrate that the improved model generates stronger feature responses in critical sign regions, such as the borders of triangular warning signs. Collectively, the experimental results across both datasets demonstrate that the YOLOv11n-GloRFCA model significantly enhances detection stability in complex scenarios while maintaining real-time performance (>45 FPS). This is achieved through the establishment of effective cross-layer feature associations.

The experimental results of the two datasets jointly show that the YOLOv11nGloRFCA model significantly improves the detection stability in complex scenarios while maintaining real-time detection speed (FPS>45) by introducing a global-local feature collaborative optimization mechanism, providing a more reliable solution for traffic sign recognition in autonomous driving systems.

V. Conclusion

To address the limitations of YOLOv11n in small-scale traffic sign detection, this study proposes an improved YOLOv11n_RFCA detection algorithm. Based on the YOLOv11n framework, we implement three key modifications: (1) Integration of the ASF (Adaptive Spatial Fusion) network to combine YOLOv11's real-time performance with advanced multi-scale feature fusion; (2) Replacement of standard convolutions with the enhanced GLO_RFCAConv module that incorporates global receptive field expansion and channel attention mechanisms to enhance feature discriminability while suppressing background interference; (3) Development of a novel Efficient_SPPF module to replace traditional SPPF, enabling adaptive geometric perception across scales. Furthermore, we incorporate the DAttention mechanism to optimize multi-scale feature fusion and employ Shape-IoU loss for precise bounding box regression that adapts to target geometry. Experimental results demonstrate significant performance improvements in small target detection accuracy.

This study systematically addresses three critical challenges in traffic sign detection: (1) Feature degradation in small targets through GLO_RFCAConv's enhanced receptive fields; (2) Multi-scale adaptation limitations via ASF and Efficient_SPPF modules; (3) Localization inaccuracy using Shape-IoU's geometric-aware regression. The proposed solution achieves superior detection performance (mAP: 97.8% vs baseline's 83.2%) while maintaining real-time capability.

Current limitations include: (1) Computational overhead from ASF and DAttention increases inference time by ~18%, requiring further optimization for edge devices; (2) Evaluation limited to daytime conditions in TT100K/CCTSDB, lacking validation for extreme illumination (e.g., night glare); (3) Hardware dependency on GPU acceleration without specific optimization for automotive processors like Jetson Orin.

Future directions will focus on: (1) Model compression for embedded deployment (TensorRT quantization); (2) Multimodal extension with thermal imaging for adverse weather detection; (3) Temporal modeling integration to handle motion blur in dynamic scenarios. These advancements will further enhance the model's applicability in intelligent transportation systems and autonomous.

REFERENCES

[1] Z. Liu, L. Wei, and T. Song, "Optimized YOLOv11 model for lung nodule detection," *Biomedical Signal Processing and Control*, vol.107, no.9, pp233-236, 2025.

[2] S. Zhou, H. Bai, and L. Jiang, "SCL-YOLOv11: A Lightweight Object Detection Network for Low-Illumination Environments," *IEEE Access*, vol.17, no.1, pp145-156, 2025.

[3] S. Cheng, Y. Han, and Z. Wang, "An Underwater Object Recognition System Based on Improved YOLOv11," *Electronics*, vol.14, no.1, pp201, 2025.

[4] N. Singh, C. P. Maurya, and B. Mahaur, "Improved YOLOv11 with weights pruning for road object detection in rainy environment," *Signal, Image and Video Processing*, vol.19, no.6, pp1-9, 2025.

[5] Z. Wang, Y. Su, and F. Kang, "Pc-yolo11s: a lightweight and effective feature extraction method for small target image detection," *Sensors*, vol.3, no.1, pp58-59, 2024.

[6] T. V. Tran, D. HQ. Ba, and T. K. Tran, "Designing a mobile application for identifying strawberry diseases with YOLOv8 model integration," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.15, no.3, pp45-48, 2024.

[7] H. Hechkel, and A. Helali, "Early detection and classification of Alzheimer's disease through data fusion of MRI and DTI images using the YOLOv11 neural network," *Frontiers in Neuroscience*, vol.19, no.1, pp154, 2025.

[8] H. Hui, and C. Jiahong, "Attention pyramid networks for object detection with semantic information fusion," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol.20, no.1, pp1-26, 2024.

[9] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: Challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol.3, no.1, pp82-83, 2023.

[10] C. Chen, Z. Zheng, and T. Xu, "Yolo-based uav technology: A review of the research and its applications," *Drones*, vol.7, no.3, pp190-195, 2023.

[11] C. Zhao, X. Shu, and X. Yan, "RDD-YOLO: A modified YOLO for detection of steel surface defects," *Measurement*, vol.214, no.1, pp789-810, 2023.

[12] Z. Yu, H. Huang, and W. Chen, "Yolo-facev2: A scale and occlusion aware face detector," *Pattern Recognition*, vol.155, no.1, pp23-35, 2024.

[13] Y. Huo, Y. Zhang, and J. Xu, "A Small-Sample Target Detection Method for Transmission Line Hill Fires Based on Meta-Learning YOLOv11," *Energies*, vol.18, no.6, pp1511, 2025.

[14] H. Zhou, F. Jiang, and H. Lu, "SSDA-YOLO: Semi-supervised domain adaptive YOLO for cross-domain object detection," *Computer Vision and Image Understanding*, vol.339, no.1, pp39-59, 2023.

[15] F. M. Talaat, and H. ZainEldin, "An improved fire detection approach based on YOLO-v8 for smart cities," *Neural Computing and Applications*, vol.35, no.28, pp15-24, 2023.

[16] S. Chaudhary, A. Sharma, and S. Khichar, "Enhancing autonomous vehicle navigation using SVM-based multi-target detection with photonic radar in complex traffic scenarios," *Scientific Reports*, vol.14, no.1, pp89-96, 2024.

[17] M. Hussain, "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection," *Machines*, vol.11, no.7, pp677-741, 2023.

[18] F. Dan, D. Chen, and Y. Lu, "YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems," *Computers and Electronics in Agriculture*, vol.20, no.5, pp987-1120, 2023.

[19] R. Khanam, T. Asghar, and M. Hussain, "Comparative Performance Evaluation of YOLOv5, YOLOv8, and YOLOv11 for Solar Panel Defect Detection," *Solar. MDPI*, vol.5, no.1, pp7-13, 2025.

[20] Y. Ji, D. Zhang, and Y. He, "Improved YOLOv11 Algorithm for Insulator Defect Detection in Power Distribution Lines," *Electronics*, vol.14, no.6, pp25-39, 2025.

[21] Y. Zhang, M. Ma, and Z. Wang, "POD-YOLO Object Detection Model Based on Bi-directional Dynamic Cross-level Pyramid Network," *Engineering Letters*, vol.32, no.5, pp995-1003, 2024.

[22] L. Zhang, Z. Sun, and H. Tao, "Research on Mine-Personnel Helmet Detection Based on Multi-Strategy-Improved YOLOv1," S*ensors (Basel, Switzerland)*, vol.25, no.1, pp170, 2024.