Traffic Sign Recognition Using Multi-Scale Enhanced Residual Network and Deformable YOLO in Intelligent Transportation

Zhuhui Ye
School of Transportation and Information, Shaanxi College of Communications Technology, Xian, 710018, China
E-mail: yzh20241026@163.com

Keywords: intelligent transportation, image enhancement network, logo recognition, multi-scale convolution, YOLO

Received: January 14, 2025

With the continuous development of intelligent transportation systems, traffic sign recognition (TSR) under complex scenarios such as low resolution and poor lighting has become a critical research focus. This study proposes a two-stage TSR framework that combines a Multi-scale Enhanced Channel Residual Network (MECRN) for image enhancement and a Deformable Convolutional YOLO-based detection network (PP-YOLO-DCN). In the enhancement stage, MECRN integrates dense connections, multi-scale convolution, and channel attention mechanisms to improve image clarity and detail preservation. In the recognition stage, deformable convolutions and depthwise separable convolutions are introduced into the PP-YOLO framework to enhance the detection of small, irregular traffic signs while reducing computational complexity. Experimental results show that the MECRN achieves a peak signal-to-noise ratio (PSNR) of 31.7 dB and a structural similarity index (SSIM) of 0.897. In low-light scenarios, the learned perceptual image patch similarity (LPIPS) reaches 0.185, indicating superior visual restoration. The PP-YOLO-DCN model attains a mean average precision (mAP@0.5) of 0.91 and 0.86 under dense multi-target and adverse weather conditions, respectively, with real-time performance of over 40 FPS. Compared with baseline methods, the proposed framework significantly improves recognition accuracy and robustness in challenging traffic environments, providing effective technical support for intelligent transportation applications.

Povzetek: Prispevek uvaja dvostopenjski sistem za prepoznavanje prometnih znakov z izboljšano natančnostjo in robustnostjo, ki združuje MECRN za izboljšavo slik in deformabilni PP-YOLO-DCN za zaznavo.

1 Introduction

The rapid growth of the global intelligent transportation industry, coupled with the accelerated pace of urbanization and a substantial rise in transportation demand, has spurred ongoing exploration into efficient traffic management and safety enhancement technologies [1]. Traffic sign recognition (TSR) serves as a vital element within smart transportation networks, holding a crucial position in improving the accuracy of autonomous driving and traffic monitoring [2-3]. However, the precision of traffic sign identification faces significant challenges. Especially in low light environments such as nighttime and severe weather, low-quality images often make it difficult for traditional algorithms to accurately capture landmark information, affecting the detection efficiency and safety of the system [4-5]. In this context, combining image enhancement with TSR has become an effective way to enhance the robustness of intelligent transportation systems [6]. Fu et al. proposed a weak light image enhancement method based on brightness attention mechanism and generative adversarial network for image analysis of smart cities. The brightness attention mechanism was used to forecast the light distribution of low light images, guiding the enhancement network to adaptively improve image quality in various brightness regions [7]. Cheng X et al. raised an improved generative adversarial network algorithm that combined attention mechanism and multi-scale feature fusion to handle the problem of low resolution and blurry details in highway images resulting from complex weather conditions. The algorithm introduced strategies to increase attention to high-frequency region information and local discrimination. The outcomes indicated that the nighttime enhancement effect of the algorithm was raised by 12.89% [8]. Hu et al. proposed a joint image to image conversion enhancement method to address the issue of decreased accuracy in facial recognition of drivers in intelligent transportation systems due to multiple degradations of facial images. This method designed a fast diagonal symmetry pattern to generate a large number of degraded/clear image pairs as training data. In addition, the proposed dual residual block enhanced the network's ability to learn facial detail features [9]. Chenmin et al. raised a dehazing algorithm that combined division of sky zones and enhancement of transparency. First, the image was divided into regions through segmentation, and the transparency of the sky

region was further optimized. Subsequently, the simulated annealing algorithm was used to enhance the transmittance parameters and adaptively optimize the brightness and contrast of the image. The outcomes indicated that the algorithm could validly improve the subjective and objective effects of TSR [10].

Image enhancement technology has made significant progress in improving the clarity and details of low-quality traffic images, providing a good data foundation for TSR. However, achieving efficient and accurate TSR requires further reliance on advanced detection and classification methods. Ferencz et al. raised a TSR system grounded on convolutional neural networks (CNNs) to meet the widespread application needs of TSR in the area of computer vision. The system used deep CNNs to classify 43 categories of road signs in the TensorFlow framework. The results indicated that the model exhibited high accuracy on the hold out dataset [11]. Zhu Y et al. studied how to overcome environmental factors to achieve accurate and efficient TSR based on the trend of autonomous vehicles gradually maturing. The study evaluated the performance of TSR using You Only Look Once v5 (YOLOv5) and single shot multi-box detector (SSD) model. Experiments showed that YOLOv5 outperformed SSDs in both recognition accuracy and speed [12]. Min W et al. raised a TSR approach grounded on interpretation of scene semantics and architectural limitations. By constructing a spatial relationship model between traffic signs and surrounding objects, and combining it with an improved model for semantic analysis, the raised multi-scale dense-connected object detector was tested on two benchmark datasets for TSR, with accuracies of 92.8% and 99.90% [13]. Abdel Salam et al. raised an instantaneous image improvement CNN for the various TSR datasets, addressing the issues of significant environmental impact and poor real-time performance. The experiment was tested on traffic sign benchmarks in Germany, Belgium, and Croatia, with recognition rates of 99.75%, 99.25%, and 99.55% [14]. A summary of the existing studies is shown in Table 1.

Table 1: Summary of existing studies

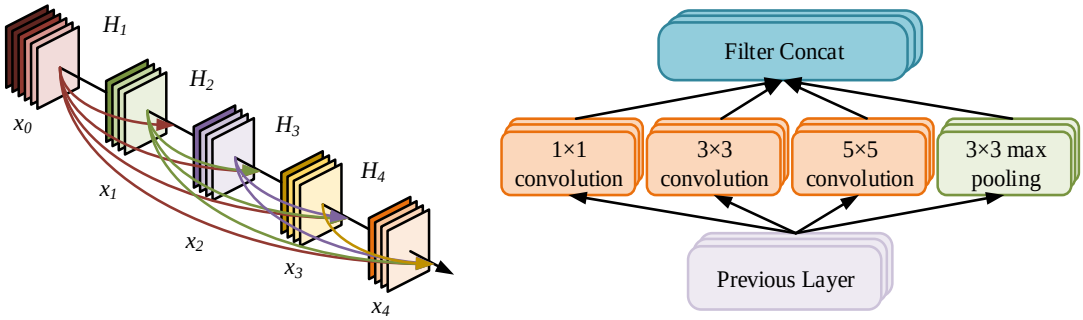
Author	Method description	Results	Limitations
Fu J et al. [7]	Brightness attention mechanism + GAN-based enhancement	Significant improvement in brightness and quality in low light	Not integrated with detection; no sign recognition
Cheng X et al. [8]	GAN with attention and multi-scale fusion	Nighttime enhancement improved by 12.89%	Weak for small-object detection
Hu C et al. [9]	Image-to-image translation + dual residual blocks	Improved facial image quality under degradation	Focused on face images, not traffic signs
Chenmin N et al. [10]	Dehazing via sky segmentation and transmittance optimization	Enhanced contrast and subjective clarity in traffic images	High dependency on sky region segmentation
Ferencz C et al. [11]	CNN-based multi-class sign classification (43 classes)	High classification accuracy on hold-out dataset	Poor robustness to blur and distortion
Zhu Y et al. [12]	Comparison of YOLOv5 and SSD for detection	YOLOv5 outperformed SSD in accuracy and speed	Limited in low-light and small-object scenarios
Min W et al. [13]	Scene modeling + multi-scale dense detection	Accuracy of 92.8% and 99.90% on two datasets	Strong reliance on image clarity and resolution
Abdel-Salam R et al. [14]	Real-time image-enhanced CNN (RIECNN)	Recognition rate of 99.75%, 99.25%, and 99.55% on three datasets	Lacks structural flexibility in complex backgrounds

Although current TSR methods have achieved promising results under standard conditions, significant challenges remain under various forms of image degradation. Most existing studies fail to systematically identify the specific limitations of different degradation types and their direct impacts on recognition performance. In low-light environments, reduced brightness and signal-to-noise ratio hinder traditional CNNs from extracting critical textures of small targets, leading to increased false detections and omissions. In blurry images, the loss of edge sharpness disrupts spatial structure modeling, especially affecting the localization of irregular-shaped signs. In adverse weather conditions such as fog or rain, weakened contrast and increased background noise further degrade performance, particularly for models lacking robust multi-scale feature fusion capabilities. Moreover, many methods focus solely on either image enhancement or recognition, lacking an end-to-end co-optimization mechanism that ensures both visual quality and detection accuracy. To address these issues, this study proposes a dual-stage framework combining enhancement and detection. The first stage introduces a Multi-scale Enhanced Channel Residual Network (MECRN), which integrates multi-scale convolution, channel attention (CA), and dense residual connections to restore structural clarity and contrast. The second stage constructs an improved PaddlePaddle-You Only Look Once model based on Deformable Convolutional Networks and Depthwise Separable Convolutions (PP-YOLO-DCN), enhancing robustness in detecting small, irregular, and multi-scale signs. This framework achieves stable and accurate recognition across diverse degraded traffic scenarios.

2 Methods and materials

2.1 Design of image enhancement model based on MECRN

In intelligent transportation systems, the clarity of images will directly affect the accuracy of TSR, especially in conditions such as long distance and severe weather, where images often appear blurry or low resolution, affecting the recognition effect [15]. The widely-used Super-Resolution CNN (SRCNN) achieves direct mapping from low resolution images to high-resolution images through an end-to-end CNN structure, greatly improving the reconstruction effect [16-17]. However, in complex traffic scenarios, SRCNN shows limitations in handling fine textures and multi-scale features. As the network deepens, it tends to suffer from gradient vanishing and feature redundancy, which adversely affect training stability and reconstruction quality. To address these issues, this study aims to design an image enhancement model capable of adapting to various types of degradation, while improving the accuracy and robustness of subsequent TSR. The specific research questions include: (1) How can structurally clear and detail-rich images be effectively restored under conditions such as low illumination, blurring, and compression? (2) Can multi-scale feature fusion and CA mechanisms enhance the model's ability to represent hierarchical texture information? (3) When used as a preprocessing module for detection, can the enhancement model significantly improve overall recognition performance? To answer these questions, this study introduces a strategy that combines dense connections and multi-scale convolution into the SRCNN framework, aiming to better meet the image enhancement needs of intelligent transportation systems and provide high-quality inputs for TSR. The structure of dense connections and multi-scale convolution is illustrated in Figure 1.



(a) Dense connections (b) Inception module

Figure 1: Densely connected and multi-scale convolutional structures

Figure 1 (a) and Figure 1 (b) respectively show dense connections and multi-scale convolution structures. In Figure 1 (a), the output of each layer is not only directly passed to subsequent layers, but also connected to deeper feature maps. By integrating shallow and deep features through cross layer connections, the propagation of information flow is enhanced, which can successfully mitigate the issue of gradient dissipation. In Figure 1 (b), the Inception module is a classic example of multi-scale convolution, which achieves multi-scale feature extraction through convolution kernels (CKs) and max pooling layers at different scales of 1×1, 3×3, and 5×5. Finally, by combining this multi-scale information through Concat, the network can maintain sparsity while maintaining high computational efficiency and adaptability. In dense connections, H represents a set of nonlinear operations, including convolutions, activation functions, etc. The input of each layer is the cumulative output of the current layer and all previous layers, where the input expression is shown in equation (1).

$$x_N = H_N([x_0, x_1, ..., x_{N-1}]) \quad (1)$$

In equation (1), N means the total number of layers. x_N is the input, and $[x_0, x_1, ..., x_{N-1}]$ represents the concatenation of all features from layer 0 to layer $N-1$. H_N performs non-linear transformations such as convolution and activation on these features. In the multi-scale convolution structure, the outputs of each branch are doing Concat in the channel dimension to obtain the output of the Inception module, as shown in equation (2).

$$x_{inception} = Concat(x_{1\times1}, x_{3\times3}, x_{5\times5}, x_{pool}) \quad (2)$$

In equation (2), $x_{1\times1}$, $x_{3\times3}$, and $x_{5\times5}$ represent the features extracted from the input feature map x using 1×1, 3×3, and 5×5 CKs. $x_{5\times5}$ means the final output feature map. x_{pool} represents the maximum pooling operation and subsequent convolution to obtain characteristics from the given feature map x. To further enhance multi-scale feature representation and information modeling, a Multi-Scale Residual Feature Attention (MSRFA) module is designed based on the structure in Figure 1. Compared with conventional multi-scale fusion methods such as Feature Pyramid Network (FPN) and Atrous Spatial Pyramid Pooling (ASPP), MSRFA is more lightweight and capable of preserving multi-scale details while enhancing the response to key regions through residual connections and

CA. This design improves image enhancement performance in challenging scenarios such as low light and blur, making it more suitable for intelligent transportation applications. The structure of MSRFA is shown in Figure 2.

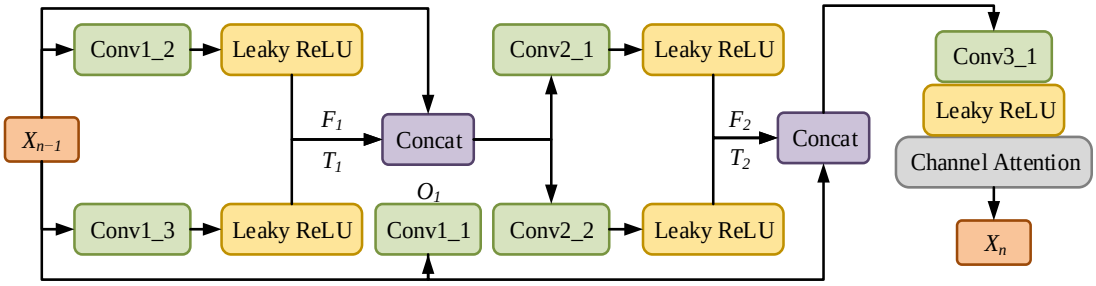


Figure 2: MSRFA structure

As shown in Figure 2, the MSRFA module is an optimized implementation based on the concepts of multi-scale convolution and dense connections presented in Figure 1. The module first extracts initial features from the input feature map through two parallel branches with different kernel sizes (Conv1_2 and Conv1_3), capturing multi-scale features F_1 and T_1. These two branches represent a simplified version of the Inception structure in Figure 1(b), retaining 3×3 and 5×5 convolutions to enhance medium and large receptive field feature extraction, while discarding the 1×1 convolution and max-pooling branches to reduce computational complexity and improve structural detail modeling. In addition, a 1×1 convolution (Conv1_1) is introduced to complement local feature extraction, forming the shallow feature extraction stage together with Conv1_2 and Conv1_3. All three branches produce 64 output channels. The concatenated features are then passed through Conv2_1 and Conv2_2 (3×3 and 5×5 convolutions) to extract deeper multi-scale features, with the output dimension remaining at 64. Finally, Conv3_1 applies a 1×1 convolution to compress the fused intermediate features into a 256-dimensional representation, which is further refined by a CA mechanism to enhance responses to key regions. This enables joint modeling of global context and local details, producing the final enhanced feature map.

It is worth noting that Conv3_1 not only integrates the preceding multi-scale features but also employs 1×1 convolution to compress channels and fuse spatial information, enabling the model to retain fine-grained local textures while introducing global contextual awareness. Combined with the CA mechanism, this design allows the model to simultaneously focus on key local regions and overall structural patterns, enhancing both perceptual consistency and structural fidelity in the enhanced image.

Firstly, the initial feature extraction expression is shown in equation (3).

$$O_1, F_1, T_1 = \Gamma(\omega_{1\times1} * X_{n-1}), \Gamma(\omega_{3\times3} * X_{n-1}), \Gamma(\omega_{5\times5} * X_{n-1}) \quad (3)$$

In equation (3), X_{n-1} represents the input feature map of the previous layer. $\omega_{1\times1}$, $\omega_{3\times3}$ and $\omega_{5\times5}$ respectively represent the weights of different convolutions. Γ represents the combination of convolution and activation function Leaky ReLU. O_1, F_1 and T_1 respectively represent feature information extracted through 1×1, 3×3, and 5×5 convolutions. Similarly, after deep feature fusion, the final output feature fusion expression is shown in equation (4).

$$X_n = C(\omega_{1\times1} * [T_2, F_2, O_1, X_{n-1}]) \quad (4)$$

In equation (4), $[T_2, F_2, O_1, X_{n-1}]$ represents concatenating the multi-scale features F_2 and T_2 obtained from two convolutions with the initial feature O_1 and input feature X_{n-1}, integrating the multi-scale information of the intermediate layer. C represents the CA mechanism. X_n is the final output feature map. Finally, grounded on the improved SRCNN network by integrating MSRFA, the MECRN image enhancement model structure is shown in Figure 3.

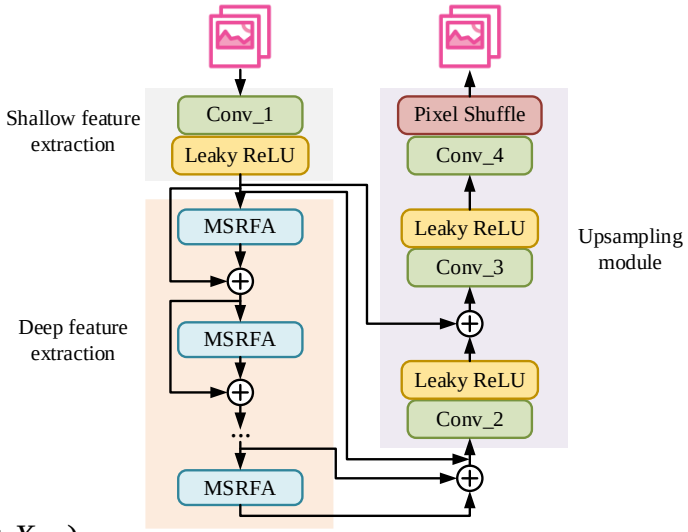


Fig. 3. MECRN structure diagram

In Figure 3, the network structure of the MECRN

image enhancement model is divided into shallow and deep feature extraction modules, and upsampling modules. Firstly, the shallow feature extraction module extracts initial features from the input low resolution image through convolutional layers Conv1 and Leaky ReLU activation functions. Secondly, the deep feature extraction module includes multiple MSRFAs, which capture different levels of detail information in the image through multi-scale feature fusion and CA mechanism. Meanwhile, skip connections have been added between each MSRFA module to preserve shallow information and reduce gradient dissipation. Finally, the upsampling module enhances the spatial resolution of features through the Pixel Shuffle layer, generating the final super-resolution output image.

2.2 Design of TSR model based on PP-YOLO

After constructing the MECRN-based image enhancement model, a TSR model is further developed based on the enhanced high-quality images. The study mainly focuses on the following questions: (1) Can deformable convolution (DCN) enhance the model's feature adaptability to irregular traffic signs? (2) Can depthwise separable convolution (DSC) reduce computational cost while maintaining detection accuracy? (3) Can the integration of image enhancement and detection improve recognition robustness in complex scenarios? The TSR model is based on PP-YOLO, and is a lightweight target detection algorithm based on YOLO architecture [18-19]. PP-YOLO inherits the efficient real-time detection capability of the YOLO series, and achieves a good balance between computational efficiency and accuracy by introducing more efficient model structures and feature optimization strategies. Its network structure is shown in Figure 4 [20].

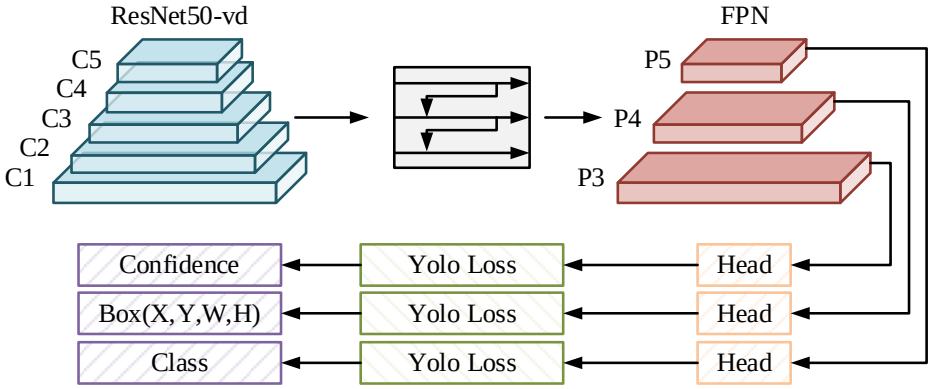


Figure 4: PP-YOLO network structure

In Figure 4, PP-YOLO includes three parts: feature extraction network, FPN, and detection head. ResNet50-vd, as its feature extraction network, convolves the input image layer by layer to extract feature maps of different levels C1 to C5, where C5 represents the depth features of the highest level. Secondly, the FPN module is used to fuse the feature maps of P3, P4, and P5 at different scales. Subsequently, the feature maps output by FPN are passed to the detection head, and each layer of feature map corresponds to a detection head. The YOLO Loss serves the purpose of computing the classification loss, bounding box loss, and confidence loss. Finally, the network outputs the Class, Box, and Confidence of each target.

Due to the fact that traffic signs often exhibit different shapes and scales due to factors such as perspective, distance, and lighting, to raise the feature extraction capability of the model for irregular and small-scale traffic signs in complicated traffic scenes, a DCN is introduced into PP-YOLO. DCN can dynamically adjust the sampling position of the CK, so that the network can focus on the key areas of the target, thereby capturing the detailed features of traffic signs more flexibly and efficiently. DCN is a key component of DCN, as shown in Figure 5.

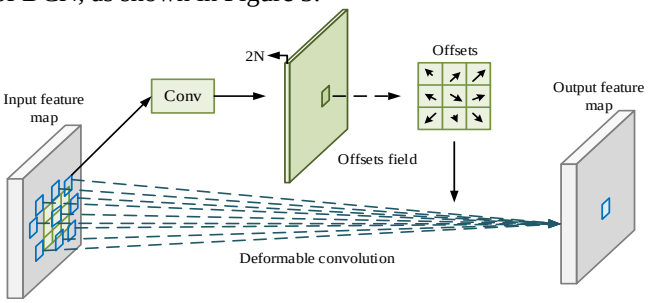


Figure 5: Schematic diagram of DCN

As illustrated in Figure 5, traditional convolution operations utilize fixed sampling positions arranged in a regular grid pattern (e.g., 3×3 centered on a pixel). However, in real-world traffic scenes, objects often exhibit complex variations such as deformation, occlusion, and tilt, making fixed sampling insufficient for capturing critical structural information. To address this limitation, DCN introduces learnable offsets that allow each sampling point in the convolutional kernel to dynamically adjust its position based on the input feature map. These offsets are generated by an independent

convolutional branch and added to the original regular grid positions to produce new, content-adaptive sampling locations. This mechanism enables the kernel to focus on structurally variant regions, such as object boundaries, curved contours, or partially missing areas, thereby enhancing the model's capability to represent fine-grained details. Bilinear interpolation is used to extract features from non-integer sampling coordinates, ensuring spatial continuity in the output.

Equations (5) to (7) describe the three core processes of DCN: offset calculation, non-uniform sampling, and weighted feature aggregation. Specifically, Equation (5) defines the adaptive adjustment of sampling positions. Equation (6) reconstructs features at fractional locations through bilinear interpolation. Equation (7) aggregates all sampled features to produce the final output map. Compared to standard convolution, DCN increases the flexibility of spatial modeling and improves adaptability to irregularly shaped objects. This structure is selected due to its strong compatibility with YOLO-based architectures and its effectiveness in handling small targets, partial occlusions, and geometric deformations, which are common challenges in TSR scenarios. Firstly, the offset calculation is shown in equation (5).

$$\Delta p_n = f_{offset}(x) \tag{5}$$

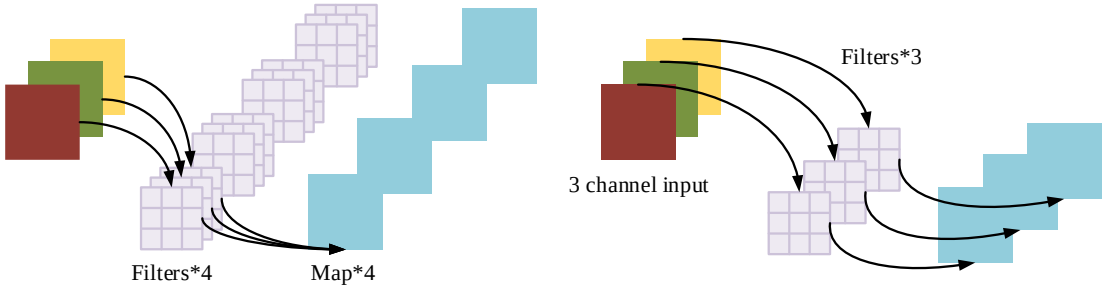
In equation (5), Δp_n represents the dynamic offset of each sampling point, and f_{offset} represents the convolution function used to generate the offset. Subsequently, the variability convolution operation is shown in equation (6).

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{6}$$

In equation (6), $y(p_0)$ represents the value of the output feature map at position p_0, and $x(p_0 + p_n + \Delta p_n)$ means the value of the input feature map at the position of the dynamic sampling point. $\omega(p_n)$ is the weight of the CK, consistent with traditional convolution operations. Finally, the overall output feature map of the variable convolution is represented by equation (7).

$$Y = f_{deform}(X, \Delta P) = \sum_{p_0 \in X} \sum_{p_n \in R} \omega(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{7}$$

In equation (7), Y is the output feature map after variable convolution. f_{deform} is the DCN operation, and X means the input feature map. ΔP is the offset field. $p_0 \in X$ represents traversing all positions of the input feature map. Considering that increasing the depth of convolution will lead to an increase in parameter count and computational overhead, this study attempts to introduce DSC. By decomposing the standard convolution into deep convolution and 1×1 convolution, it can effectively reduce the computational and parameter count, and improve the computing speed. The structure is shown in Figure 6.



(a) Regular convolution (b) DSC

Figure 6: Schematic diagram of depth-wise separable convolution

Figures 6(a) and 6(b) illustrate the operations of standard convolution and DSC, respectively. In Figure 6(a), for a 5×5 input with three channels, four 3×3 convolutional kernels are applied. Each kernel performs a 3×3 convolution across all three input channels, and the results are summed to produce one output feature map, generating a total of four output channels. This operation involves simultaneous computation across spatial and channel dimensions, with computational cost increasing rapidly with the number of channels and kernels. In contrast, Figure 6(b) shows that DSC decomposes the convolution into two steps: First, depthwise convolution is applied separately to each input channel using a single corresponding kernel, maintaining the same number of output channels. Second, a 1×1 pointwise convolution is used to fuse information across channels. This design significantly reduces the number of multiplications and parameters while retaining spatial feature extraction and inter-channel interaction capabilities.

Compared to standard convolution, DSC offers higher efficiency and practicality in TSR tasks. On one hand, these tasks demand real-time inference, and the high computational cost of standard convolution in the channel dimension often becomes a bottleneck during early-stage feature extraction. DSC alleviates this issue by reducing the parameter count and computational load, thereby accelerating inference. On the other hand, traffic signs are typically small objects with clear structural

boundaries. DSC preserves spatial detail sensitivity through depthwise convolution and enhances semantic representation via pointwise convolution, improving the model's ability to extract edge and texture features. Therefore, the structure depicted in Figure 6(b) is highly suitable for integration into lightweight object detection models. The calculation of deep convolution is shown in equation (8).

$$y_{c,i,j} = \sum_{m=0}^{K-1}\sum_{n=0}^{K-1} x_{c,i+m,j+n} \cdot \omega_{c,m,n} \tag{8}$$

In equation (8), $y_{c,i,j}$ means the value of the output feature of channel c at (i, j). $x_{c,i+m,j+n}$ is the value of channel c at $(i+m, j+n)$, and $\omega_{c,m,n}$ is the weight of the $K \times K$ CK on channel c. Subsequently, the point by point convolution calculation is shown in equation (9).

$$z_{k,i,j} = \sum_{c=0}^{C-1} y_{c,i,j} \cdot v_{k,c} \tag{9}$$

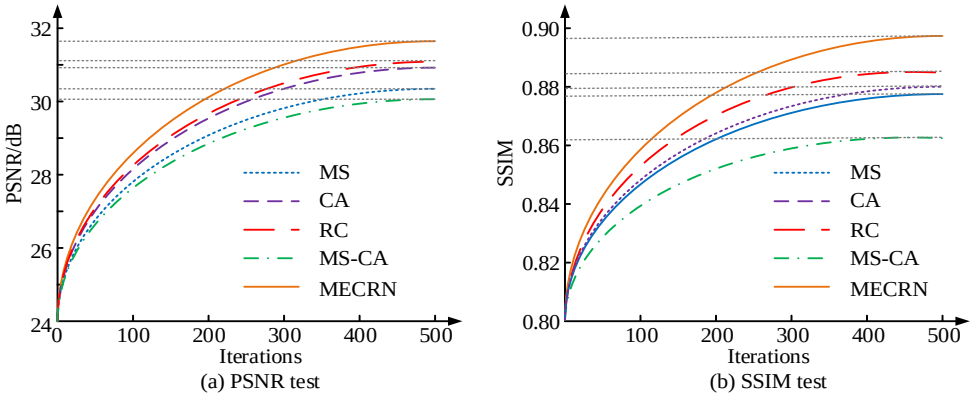
In equation (9), $z_{k,i,j}$ means the value of channel k at (i, j). $v_{k,c}$ is the weight of the stationary convolution, applied to $y_{c,i,j}$ and output channel k. Therefore, based on the above improvements, in the PP-YOLO-DCN model, the input image is first subjected to enhancement preprocessing. Secondly, the ResNet50-vd network is applied to obtain characteristics layer by layer, and the recognition adaptability is improved through FPN multi-scale fusion. DCN dynamic sampling is applied to capture details at key locations, computational efficiency is optimized through DSC, and ultimately the category, position, and confidence of traffic signs are output through detection heads to achieve efficient recognition in complex scenes.

3 Results

3.1 Performance testing of MECRN image enhancement model

The experiment utilized high-capacity calculation devices to guarantee effective training and evaluation of image enhancement and object recognition models. The experimental environment used Ubuntu operating system, Python as the programming language, combined with PyTorch deep learning framework and OpenCV image processing library to implement the algorithm. In terms of hardware configuration, the device was equipped with Intel Core i9 processor, NVIDIA RTX 3090 graphics card, and 128GB of memory. The Set14 dataset was used for training and testing in the image enhancement stage, as it contains structurally complex and texture-rich images suitable for evaluating detail restoration. As a general-purpose enhancement module, MECRN does not depend on the semantic content of the image. Its ability to improve visual clarity directly supports downstream TSR tasks, demonstrating strong generalizability. For TSR, the TT100K dataset was adopted, featuring real-world road scenes with small, blurred, and occluded signs, making it ideal for testing model robustness. All images were normalized and resized to 224×224 for enhancement and 640 × 640 for detection. Data augmentation such as brightness jittering, rotation, and random cropping was applied to improve generalization. For hyper-parameters, the enhancement model used the Adam optimizer with a learning rate of 1e-4, batch size of 16, and 500 epochs. The detection model used SGD with a learning rate of 0.001, batch size of 32, and 300 epochs. The loss functions included a combination of L1 and SSIM for enhancement, and multi-task YOLO loss for detection.

Firstly, as MECRN was composed of multiple modules, the ablation test outcomes are in Figure 7.



(a) PSNR test (b) SSIM test

Figure 7: Ablation test results

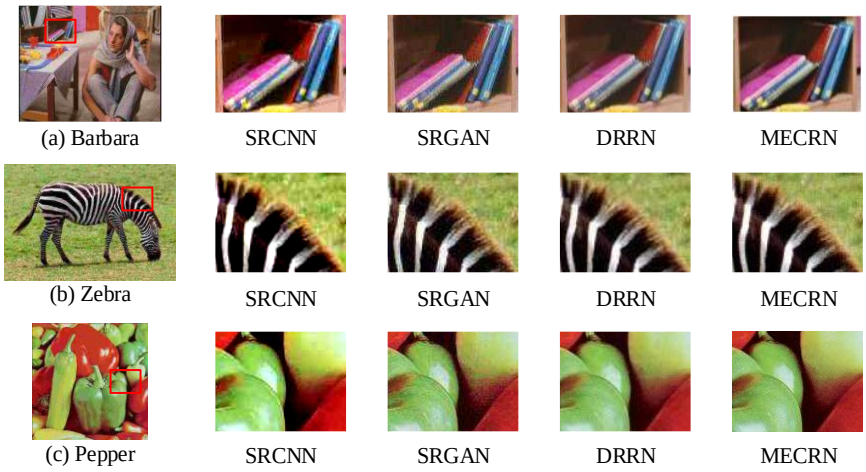


Figure 8: Qualitative analysis results

Figures 7 (a) and 7 (b) show the test results of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) as they vary with the number of iterations, respectively. PSNR reflects reconstruction accuracy, while SSIM measures structural fidelity—both widely used in image enhancement tasks. MS, CA, RC, and MS-CA respectively represent the removal of multi-scale convolution, CA mechanism, residual connection, multi-scale convolution, and attention mechanism. In Figure 7 (a), when the number of iterations was 500, the PSNRs of removing MS, CA, RC, MS-CA, and the complete MECRN model were 30.4 dB, 30.9 dB, 31.1 dB, 30.1 dB, and 31.7 dB, respectively. In Figure 7 (b), the SSIMs without MS, CA, RC, MS-CA, and the complete MECRN model were 0.876, 0.879, 0.883, 0.862, and 0.897, respectively. The results demonstrated that each component of MECRN played a vital role in performance enhancement. RC helped maintain structural consistency and facilitates shallow-to-deep feature propagation, improving model stability. MS expanded the receptive field and strengthened the model's ability to capture fine textures and edges, which was particularly beneficial in upscaling and deblurring scenarios. CA enhanced the response to dark and critical semantic regions while suppressing redundant features. The largest performance drop occurred when both MS and CA were removed, highlighting their complementary effect in structural reconstruction and perceptual quality. These components together formed the core of the MSRFA module, and their integration enabled MECRN to achieve superior enhancement results through both independent effectiveness and mutual reinforcement.

Subsequently, Super Resolution CNN (SRCNN), Super Resolution Generative Adversarial Network (SRGAN), and Deep Recursive Residual Network (DRRN) were selected as comparison models. Three images were extracted from Set14, and then processed using models. The restoration effects of each model on image details and textures were examined. The qualitative assessment outcomes are in Figure 8.

Figure 8 shows the qualitative analysis results of super-resolution processing using different models. The enhancement effect of SRCNN was relatively poor, with blurry image details, especially in images with rich details such as Zebra and Pepper. The main reason was that SRCNN had a simple structure and lacked sufficient ability to capture details. SRGAN introduced generative adversarial networks, which could improve details to a certain extent, but there was still room for improvement in its performance on complex textures. DRRN effectively improved image clarity and edge details through recursive residual connections, but still had limitations in some multi-scale feature processing. In contrast, the MECRN model combined multi-scale convolution and CA mechanisms, which could more comprehensively capture multi-scale details and local features, resulting in the best texture and structure restoration effect and the highest clarity of the image. MECRN exhibited better enhancement effects in areas with rich details, such as Zebra stripes and Pepper surface textures. Finally, to simulate scenes in the real world where image quality is often affected by various factors, low lighting, low resolution, noise interference, and image blurring conditions were set. The test outcomes are in Table 2.

Table 2: Multi-index test outcomes under different environmental conditions

Environment condition	Metric	SRCNN	SRGAN	DRRN	MECRN
Low light	PSNR	27.45	28.32	29.54	30.75
	SSIM	0.812	0.826	0.846	0.871
	LPIPS	0.347	0.282	0.232	0.185
Low resolution	PSNR	26.31	27.12	28.66	29.91
	SSIM	0.804	0.819	0.836	0.861
	LPIPS	0.362	0.297	0.241	0.201
Noise interference	PSNR	25.81	26.53	27.92	29.21
	SSIM	0.791	0.806	0.826	0.851
	LPIPS	0.371	0.321	0.271	0.223
Blurred image	PSNR	24.91	25.71	27.13	28.51
	SSIM	0.776	0.791	0.811	0.842
	LPIPS	0.386	0.343	0.293	0.253
Compression artifacts	PSNR	25.35	26.18	27.45	28.68
	SSIM	0.788	0.812	0.821	0.842
	LPIPS	0.365	0.328	0.283	0.236

Table 2 presents the quantitative evaluation results of different models. The Learned Perceptual Image Patch Similarity (LPIPS) is a metric for assessing perceptual image quality, where lower values indicate that the enhanced image is more perceptually similar to the ground truth. Unlike PSNR and SSIM, which mainly measure pixel-level differences and structural similarity, LPIPS focuses on perceptual consistency in deep feature space. This makes it more effective in capturing the restoration of edges, textures, and fine details, which is particularly valuable for improving feature extraction and discrimination in TSR. Under different environmental conditions, the MECRN surpassed other models in terms of PSNR, SSIM, and LPIPS metrics. Under low light conditions, MECRN's PSNR, SSIM, and LPIPS were 30.75, 0.871, and 0.185, respectively, demonstrating its excellent low light enhancement ability. This was because MECRN combined multi-scale convolution and CA mechanisms, which could better capture dark details. Under low resolution conditions, its PSNR and SSIM were 29.91 and 0.861, respectively, indicating its excellent amplification capability. For noise interference and blurry images, MECRN performed well in denoising and deblurring, with LPIPS of 0.223 and 0.253 respectively, indicating that it could effectively restore visual perception quality. In addition, under compression distortion conditions, MECRN effectively suppressed compression artifacts, and SSIM reached 0.840. MECRN exhibited stronger robustness and adaptability in complex environments.

The advantage of MECRN in LPIPS performance mainly stems from its effective modeling of image details and perceptual consistency. The multi-scale convolution captures texture and edge information at different scales, helping to restore key details in upscaling and blurring scenarios. The CA mechanism enhances responses to dark regions and important areas, reducing interference from irrelevant features. Dense connections facilitate the transfer of shallow features to deeper layers, improving structural and detail reconstruction, thereby reducing LPIPS and enhancing overall visual quality.

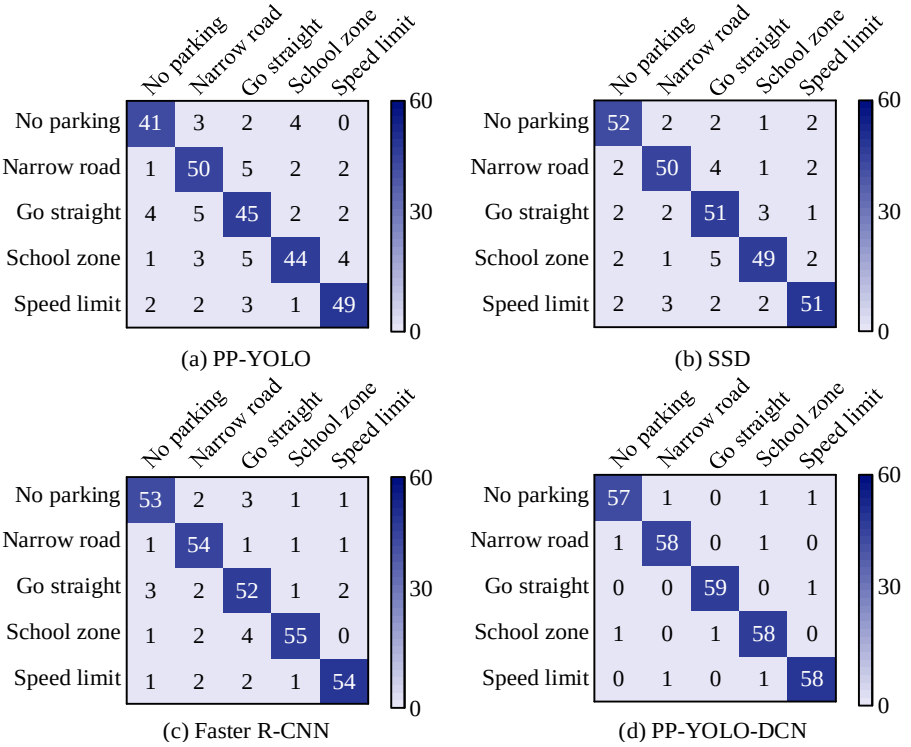
3.2 Experimental analysis of PP-YOLO-DCN object recognition model

The experimental environment was the same as the previous section, and the comparison models were PP-YOLO, the classic SSD, and the Faster Regional CNN (Faster R-CNN). The dataset was the TT100K traffic dataset, which contains 100000 real road scene images in China, covering more than 100 different traffic signs, suitable for small object detection and TSR in complex scenes. Firstly, to thoroughly assess the sorting performance of the recognition model in TSR, especially in similar category and multi-category detection, the results of the confusion matrix for each model are shown in Figure 9.

As shown in Figure 9, PP-YOLO-DCN outperformed other models in classification accuracy across various traffic sign categories. In particular, categories such as "go straight" and "school zone," which share similar shapes and subtle edge differences, exhibited 4 and 3 false detections respectively in PP-YOLO and SSD. Faster R-CNN also showed confusion in the "speed limit" category. In contrast, PP-YOLO-DCN demonstrated minimal misclassification across all categories. This performance improvement could be attributed to the architectural differences, especially the incorporation of DCN. Unlike standard convolution, DCN enabled dynamic adjustment of sampling positions based on the input features, allowing the model to better align with object boundaries and internal structures. This structural adaptability enhanced the model’s ability to distinguish fine-grained differences in small or partially occluded signs. The improvements

observed in the confusion matrix directly reflected the effectiveness of DCN in handling complex and visually similar traffic sign categories. Subsequently, to evaluate the feature focusing capability of the model in TSR, gradient-weighted class activation mapping (Grad CAM) visualization technology was applied to display the attention regions of the model when detecting traffic signs, in order to intuitively analyze the capture effect and detail attention of each model on key features. The results are shown in Figure 10.

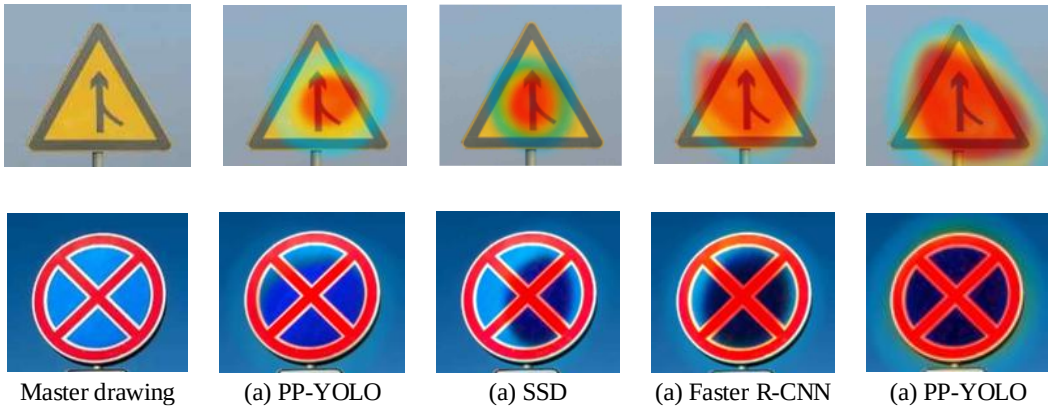
As shown in Figure 10, the Grad-CAM visualizations revealed clear differences in attention distribution among the models. PP-YOLO mainly focused on the central region of the sign, with weak responses along the edges, which may lead to localization errors. SSD showed a more uniform attention map but lacked distinct activation for critical visual cues such as arrows and cross lines. Faster R-CNN demonstrated strong attention in both central and boundary areas, indicating more balanced feature extraction. PP-YOLO-DCN exhibited the most precise and concentrated attention, effectively covering both the center and the boundaries of the signs, particularly highlighting detailed patterns and contours. This was directly related to the DCN mechanism. By learning spatial offsets, the convolutional kernels could shift their sampling positions to better align with object boundaries and local structures. This enhanced the model's ability to extract detailed features from curved edges, fine textures, and variable shapes. Under conditions such as low resolution or partial occlusion, this precise sampling enabled more complete structural representation, reducing false detections and omissions in TSR. Finally, considering the presence of TSR in complex scenarios in practical applications, the test results of each model in different environments are in Table 3.



(a) PP-YOLO (b) SSD

(c) Faster R-CNN (d) PP-YOLO-DCN

Figure 9: Confusion matrix test



Master drawing (a) PP-YOLO (a) SSD (a) Faster R-CNN (a) PP-YOLO

Figure 10: Heatmap visualization results

Table 3: Model performance testing in complex scenarios

Model	Scenario	mAP@0.5	FPS	Processing latency/ms	Robustness score
PP-YOLO	Dense multi-target	0.82	45	22	0.74
	Dynamic background	0.78	47	21	0.70
	Adverse weather	0.75	43	24	0.66
	Low light	0.72	40	26	0.63
	Motion blur	0.71	42	25	0.61
SSD	Dense multi-target	0.83	30	34	0.70
	Dynamic background	0.76	32	32	0.65
	Adverse weather	0.72	29	35	0.60
	Low light	0.69	28	36	0.57
	Motion blur	0.68	31	33	0.55
Faster R-CNN	Dense multi-target	0.87	15	66	0.81
	Dynamic background	0.84	16	64	0.75
	Adverse weather	0.82	14	69	0.70
	Low light	0.78	13	71	0.68
	Motion blur	0.76	14	68	0.64
PP-YOLO-DCN	Dense multi-target	0.91	42	25	0.85
	Dynamic background	0.88	44	23	0.82
	Adverse weather	0.86	40	27	0.78
	Low light	0.84	39	28	0.75
	Motion blur	0.82	41	26	0.73

As shown in Table 3, PP-YOLO-DCN demonstrated significant advantages in complex traffic scenarios such as dense multi-target and adverse weather conditions, achieving mAP@0.5 values of 0.91 and 0.86 respectively, outperforming other models in detection accuracy. Under low-light and motion-blurred conditions, its robustness scores reached 0.75 and 0.73, indicating strong resistance to environmental interference. Additionally, the model maintained a stable frame rate across all tested scenarios, ranging from 39 to 44 FPS, ensuring reliable real-time detection performance. In Table 3, the robustness score was introduced to comprehensively evaluate a model's ability to maintain performance under challenging conditions. It considered multiple factors, including the drop in mAP, changes in false detection rates, and fluctuations in processing latency. This metric reflected the overall stability and interference resistance of the model. In contrast, although Faster R-CNN achieved relatively high accuracy in some scenarios, its frame rate remained below 20 FPS, resulting in high latency and limited real-time applicability. Meanwhile, PP-YOLO and SSD exhibited weaker robustness, with scores of 0.63/0.61 and 0.57/0.55 in low-light and motion blur conditions, respectively, showing greater sensitivity to visual degradation. Overall, PP-YOLO-DCN achieved a strong balance among accuracy, efficiency, and robustness, making it well-suited for deployment in complex real-world traffic environments.

PP-YOLO-DCN maintained high FPS while improving accuracy due to its optimized architecture. DSCs reduced parameters and computation, while DCNs enhanced adaptability to complex shapes with minimal impact on speed. Additionally, the lightweight one-stage design of PP-YOLO ensured strong real-time performance, achieving a good balance between precision and efficiency.

To further verify that the proposed PP-YOLO-DCN model maintains high detection accuracy while ensuring computational efficiency, this study compared it with mainstream YOLO variants (YOLOv3, YOLOv5-M, and YOLOv5-L) in terms of model parameters, computational cost (GFLOPs), and inference speed (FPS). The results are shown in Table 4.

Table 4: Comparison of algorithmic complexity and performance

MODEL	PARAMETERS (M)	COMPUTATION (GFLOPS)	FPS (640 × 640)	MEAN ACCURACY (MAP@0.5)
YOLOV3	61	118	30	0.87
YOLOV5-M	44	98	45	0.88
YOLOV5-L	77	109	35	0.89
PP-YOLO-DCN	47	86	42	0.91

As shown in the table, PP-YOLO-DCN achieved a favorable trade-off between detection accuracy and computational efficiency. With 47 million parameters and 86 GFLOPs, it was lighter than YOLOv5-L and YOLOv3, and significantly reduced computation compared to YOLOv5-L (by approximately 21%). In terms of inference speed, PP-YOLO-DCN reached 42 FPS, slightly below YOLOv5-M but notably faster than YOLOv3 and YOLOv5-L. Overall, the model demonstrated strong performance in terms of accuracy, size, and real-time capability, making it well-suited for practical deployment in TSR scenarios.

4 Discussion

In TSR tasks, image quality is closely tied to detection accuracy. Although prior studies have explored both image enhancement and object detection independently, many existing methods still face limitations when applied to complex real-world scenarios. Compared with these approaches, the proposed MECRN enhancement module and PP-YOLO-DCN detection model demonstrated superior overall performance and adaptability across diverse conditions. In terms of image enhancement, Fu J et al. [7] proposed a method that integrated a brightness attention mechanism with a GAN to enhance image brightness in low-light environments. However, the method was not integrated with downstream recognition tasks, leading to a disconnect between enhancement quality and detection performance. Cheng X et al. [8] employed multi-scale fusion and attention mechanisms to improve nighttime image quality but showed limited capability in restoring small-object details. In contrast, the proposed MECRN model utilized multi-scale convolution to extract texture information at various scales, incorporated CA to enhance the response to dark regions and target areas, and introduced dense connections to improve feature reuse and training stability. As a result, MECRN improved traditional metrics such as PSNR and SSIM while significantly reducing perceptual distortion measured by LPIPS, ultimately providing higher-quality inputs for detection models.

In terms of detection, Zhu Y et al. [12] compared YOLOv5 and SSD and found that YOLOv5 achieved a better balance between speed and accuracy. However, both models exhibited performance drops in low-light and small-object scenarios. While Faster R-CNN delivered high accuracy, it suffered from heavy computational cost and high latency, making it unsuitable for real-time deployment. The proposed PP-YOLO-DCN introduced DCN to enable dynamic sampling, enhancing the model’s ability to capture object boundaries and complex structures. This proved especially effective in degraded conditions such as blur and occlusion, where accurate localization of small targets was challenging. Additionally, the integration of DSC reduced redundant computation, achieving a good trade-off between accuracy and inference speed. The proposed model outperformed others in terms of robustness and real-time performance under complex conditions. In summary, MECRN and PP-YOLO-DCN demonstrated strong adaptability to low-quality image conditions and effectively addressed the trade-off between accuracy and efficiency in TSR, offering promising potential for real-world deployment.

5 Conclusion

In response to the difficulty of recognizing traffic signs in complex traffic scenarios such as low light and long distance, the MECRN image enhancement model and PP-YOLO-DCN object recognition model were proposed to improve the image quality and detection accuracy of intelligent transportation systems. In the ablation test, the PSNR of the complete MECRN model was 31.7 dB, and the SSIM was 0.897, confirming the effectiveness of the study in improving MECRN. In qualitative analysis, MECRN had the best image enhancement effect and the highest clarity. Under low light conditions, the PSNR, SSIM, and LPIPS of MECRN were 30.75, 0.871, and 0.185, respectively, indicating good low light enhancement effect. The PP-YOLO-DCN model had the lowest number of false positives and false negatives in confusion matrix testing. In Grad CAM testing, the heatmap of PP-YOLO-DCN covered the edges and internal details of the logo, especially with high attention to key features. Under conditions of dense multi-target and adverse weather conditions mAP@0.5 reached 0.90 and 0.86 respectively, the FPS remained above 40 frames per second. The outcomes indicated that the raised method attained a high degree of recognition accuracy in various complex scenarios.

Although the proposed model demonstrated strong performance in TSR, achieving high accuracy, real-time speed, and robustness across challenging conditions such as low light and motion blur, there remains room for further improvement. First, the current evaluation was conducted on high-performance GPUs, and the model's inference efficiency on edge devices has yet to be

validated. Second, the experiments were primarily based on the TT100K dataset, lacking cross-domain generalization testing on international benchmarks such as GTSRB. Finally, while the model exhibited relatively high robustness scores, it has not been systematically evaluated under extreme conditions such as severe occlusion, heavy noise, or adversarial perturbations. Future work will focus on improving model adaptability through edge deployment, cross-dataset validation, and robustness analysis under more challenging real-world scenarios.

References

[1] Hasanvand M, Nooshyar M, Moharamkhani E, Selyari A. Machine learning methodology for identifying vehicles using image processing. Artificial Intelligence and Applications, 2023, 1(3):170-178. https://doi.org/10.47852/bonviewAIA3202833

[2] Muhammad K, Ullah A, Lloret J, Del Ser J, de Albuquerque V. H. C. Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(7): 4316-4336. https://doi.org/10.1109/TITS.2020.3032227

[3] Liu X, Yan W Q. Traffic-light sign recognition using Capsule network. Multimedia Tools and Applications, 2021, 80(10):15161-15171. https://doi.org/10.1007/s11042-020-10455-x

[4] Shivayogi A B, Dharmendra N C M, Ramakrishna A M, et al. Real-time traffic sign recognition using deep learning. Pertanika J Sci Technol, 2022, 31(1):137-148. https://doi.org/10.1007/s11042-022-12163-0

[5] Yucong S, Shuqing G. Traffic sign recognition based on HOG feature extraction. Journal of Measurements in Engineering, 2021, 9(3):142-155. https://doi.org/10.21595/jme.2021.22022

[6] Luo Q, Zheng W. Pre-Locator Incorporating Swin-Transformer Refined Classifier for Traffic Sign Recognition. Intelligent Automation & Soft Computing, 2023, 37(2):2227-2246. https://doi.org/10.32604/iasc.2023.040195

[7] Fu J, Yan L, Peng Y, Zheng K, Gao R, Ling H. Low-light image enhancement base on brightness attention mechanism generative adversarial networks. Multimedia Tools and Applications, 2024, 83(4):10341-10365. https://doi.org/10.1007/s11042-023-15815-x

[8] Cheng X, Zhou J, Song J, Zhao X. A highway traffic image enhancement algorithm based on improved GAN in complex weather conditions. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(8):8716-8726. https://doi.org/10.1109/TITS.2023.3258063

[9] Hu C, Liu Y, Xu L, Xiao X, Lu X, Yang W, Liu P. Joint image-to-image translation for traffic monitoring driver face image enhancement. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(8):7961-7973. https://doi.org/10.1109/TITS.2023.3258634

[10] Chenmin N, Marsani M F, Shan F P. Traffic image dehazing based on sky region segmentation and transmittance optimization. Journal of Intelligent & Fuzzy Systems, 2024, 46(1):1005-1017. https://doi.org/10.3233/JIFS-233433

[11] Ferencz C, Zöldy M. Neural Network-based Multi-Class Traffic-Sign Classification with the German Traffic Sign Recognition Benchmark. Acta Polytechnica Hungarica, 2024, 21(7):203-220.

[12] Zhu Y, Yan W Q. Traffic sign recognition based on deep learning. Multimedia Tools and Applications, 2022, 81(13):17779-17791. https://doi.org/10.1007/s11042-022-12163-0

[13] Min W, Liu R, He D, Han Q, Wei Q, Wang Q. Traffic sign recognition based on semantic scene understanding and structural traffic sign location. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(9):15794-15807. https://doi.org/10.1109/TITS.2022.3145467

[14] Abdel-Salam R, Mostafa R, Abdel-Gawad A H. RIECNN: real-time image enhanced CNN for traffic sign recognition. Neural Computing and Applications, 2022, 34(8):6085-6096. https://doi.org/10.1007/s00521-021-06762-5

[15] Jeya R, Krishnan G R, Babu C R. Traffic Sign Classification and Recognition using LE-NET Technique. International Journal of Early Childhood Special Education, 2022, 14(3):544-551. http://doi.org/10.9756/INT-JECSE/V14I1.221001

[16] Chen Q, Zhang J, Li B. Research on 3D MFL testing of wire rope based on empirical wavelet transform and SRCNN. Journal of Vibroengineering, 2022, 24(4):779-792. http://doi.org/10.9756/INT-JECSE/V14I1.221001

[17] Nguyen N B, Doan V S, Pham M N, et al. SRCNN: Stacked-Residual Convolutional Neural Network for Improving Human Activity Classification Based on Micro-Doppler Signatures of FMCW Radar. Journal of Electromagnetic Engineering and Science, 2024, 24(4):358-369. https://doi.org/10.1371/journal.pone.0308045

[18] Chen C, Yu J, Lin Y, Lai, F, Zheng G, Lin Y. Fire detection based on improved PP-YOLO. Signal, Image and Video Processing, 2023, 17(4):1061-1067. https://doi.org/10.1007/s11760-022-02312-1

[19] Zhang G, Zhang J. High-Precision Photogrammetric 3D Modeling Technology Based on Multi-Source Data Fusion and Deep Learning-Enhanced Feature Learning Using Internet of Things Big Data. Informatica, 2025, 49(11).

[20] Gu X, Dai S. IRF-HTID-BO-LSTM: Classification Model of Curve Shape Index for Mountainous Highways and Intelligent Traffic Incident Detection Method. Informatica, 2025, 49(11). https://doi.org/10.31449/inf.v49i11.7449