Received 22 November 2023, accepted 22 December 2023, date of publication 25 December 2023,
date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347352

RESEARCH ARTICLE

CR-YOLOv8: Multiscale Object Detection in Traffic Sign Images

LU JIA ZHANG[1], **JIAN JUN FANG JR.**[2], **YAN XIA LIU JR.**[2], **HAI FENG LE**[1], **ZHI QIANG RAO JR.**[2], **AND JIA XIANG ZHAO**[2]

[1]Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China
[2]College of Urban Rail Transit and Logistics, Beijing Union University, Beijing 100101, China

Corresponding author: Jian Jun Fang Jr. (jianjun@buu.edu.cn)

This work was supported by the Beijing Natural Foundation under Grant L221015.

ABSTRACT Due to the large-scale changes of different forms of traffic signs and the rapid speed of vehicles, the detection accuracy and real-time performance of general object detectors are greatly challenged, especially the detection accuracy of small objects. In order to solve this problem, a multi-scale traffic sign detection model CR-YOLOv8 is proposed based on the latest YOLOv8. In the feature extraction stage, the attention module is introduced to enhance the channel and spatial features, so that the network can learn the key information of the small objects more easily. The RFB module is introduced in the feature fusion stage, which improves the feature diversity with less computational overhead and improves the network's ability to detect multi-scale objects. By improving the loss function to enable the model to effectively balance multi-scale objectives during training, the model generalization ability is improved.The experimental results on TT100k dataset show that compared with the baseline network, the average detection accuracy of the improved method is increased by 2.3 %, and the detection accuracy of small objects is increased by 1.6 %, which effectively reduces the detection accuracy gap among different scales.

INDEX TERMS Traffic sign recognition, traffic sign recognition, YOLOv8.

I. INTRODUCTION

Road traffic sign detection is the most basic and critical component of Intelligent Transportation Systems (ITS) and driverless systems [1]. Demand for intelligence has accelerated the technological development of computer vision, intelligent traffic safety systems, and other related fields. To improve the safety of vehicles and pedestrians on the road, the accuracy and detection efficiency of traffic sign detection must be continuously improved. However, in real-world scenarios, owing to the different uses and types of traffic signs (as shown in Figure 1), there are large differences in the scales of traffic signs captured by imaging devices. At the same time, to ensure that high-speed vehicles have sufficient braking distance, traffic signs at a greater distance need to be identified as early as possible, which further increases the difficulty of detecting traffic signs and is receiving increasing attention.

In recent years, object detection techniques driven by deep learning architectures have developed rapidly and achieved fruitful results. Most state-of-the-art object detectors use convolutional neural networks (CNN) for image feature extraction, which can be categorized into two-stage object detectors and one-stage object detectors. Among them, R-CNN [2] is the most representative two-stage object detection method and has laid a solid foundation for the development of subsequent two-stage algorithms. The remaining two-stage object detection algorithms are Cascade R-CNN [3], Pv-RCNN [4], sparse R-CNN [5] and others [6], [7], [8], [9]. The above algorithms first generate candidate regions, and then perform classification and regression. This method is designed flexibly and covers a wide range of work, but due to the need to generate a large number of candidate regions, it increases computational complexity and reduces detection speed. In order to reduce the high computational cost and improve the slow detection speed caused by obtaining higher detection accuracy, researchers discarded the object generation phase and investigated one-stage object-detection algorithms. Typical one-stage object detection methods are

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Tucci.

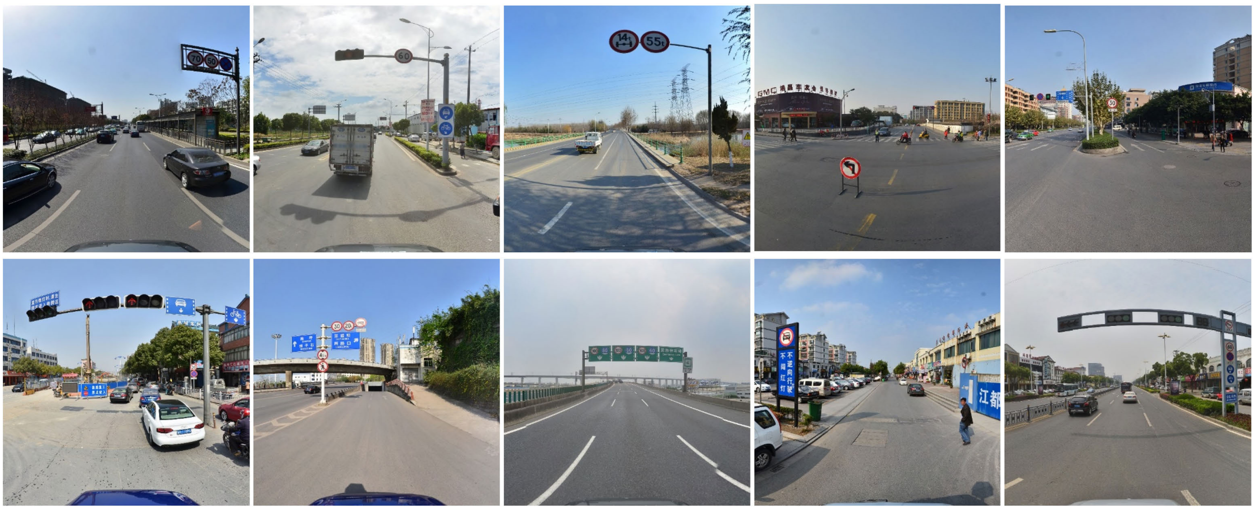


FIGURE 1. Examples of the size, color, and shape of traffic signs.

SSD [10], M2det [11], You Only Look Once (YOLO) [12], [13], [14], [15], [16] series and their variants [17], [18], [19], [20], [21].

Traffic sign detection [22] is a subfield of object detection techniques for detecting and localizing traffic-sign instances in digital images or video frames. However, when the above detection methods are applied directly to traffic sign detection in real-world scenarios, the results are usually unsatisfactory. Owing to the differences in the uses and types of traffic signs, as well as the differences in the scales of traffic signs captured by imaging devices owing to the characteristics of vehicle travel, a universal detector cannot fully extract all scales of traffic sign features, and thus cannot satisfy the needs of vehicles for highly accurate detection of objects at multiple scales. At the same time, the object detection of the vehicle mobile terminal not only requires high detection accuracy of objects of different scales but also has a high demand for the speed of identification. A higher processing speed in the traffic-sign detection stage can provide more time for subsequent decision-making and subsequent operation of the vehicle. Therefore, a detector with both detection accuracy and speed is essential to ensure the safety and efficiency of road traffic. Compared to two-phase object detectors, one-phase object detectors are more in line with the need for detection speed in traffic-sign detection. To ensure that the detector can effectively detect traffic signs on mobile devices with limited computational resources, the state-of-the-art one-stage object detection network YOLOv8s is chosen as the benchmark in this paper. YOLOv8s combines the advantages of fast detection speed and a small number of network parameters. However, the detection effect of YOLOv8s on traffic signs at different scales needs to be improved, especially when the detection accuracy of small-scale objects is quite different from that of large-scale objects, which reduces the detection accuracy of large-scale objects, and the detection accuracy of large-scale objects is quite different from that of large-scale objects. In particular, the detection accuracy of small-scale objects differs greatly from that of large-scale objects, and narrowing the gap between the detection accuracies of large, medium, and small scales is the key issue that this study aims to address.

To enhance the detection effect of YOLOv8s on multiscale traffic signs, this study proposes a CR-Yolov8 (Convolutional Block Attention Module and Receptive Field Block-You Only Look Once Version 8) network structure based on YOLOv8s. CR-Yolov8 detection network, that is, meets the requirements of mobile devices on the size of the deployment model, and simultaneously improves the detection accuracy of the detector for multi-scale objects to meet real-time requirements. The specific contributions of this study are as follows.

1) The Convolutional Block Attention Module (CBAM) lightweight attention module is introduced in to the backbone network to reduce the effect of information loss due to down-sampling by focusing on spatial and positional information and to improve the sensitivity of the network to traffic sign information.

2) Expanding the sensory field to obtain higher-resolution features by fusing the Receptive Field Block (RFB) module improves the feature diversity of the lightweight model and enhances the network's ability to learn multi-scale features.

3) Optimize the model bounding box regression loss function and gradient gain allocation strategy to enhance the robustness to object-scale changes. Compared with advanced traffic sign detection, the method proposed in this study exhibits good performance.

The remainder of this paper is organized as follows. In Section II, related research work is presented, including the development and research results in the field of traffic sign detection and the YOLOv8 network structure. In Section III, the details of the proposed method, which achieves accurate recognition and localization of traffic signs at different scales, are presented. In Section IV, the results of the experiments are presented and compared with those of previous methods.

Finally, in the Section V, the paper will summarize the main conclusions of this study and propose future research directions.

II. RELATED WORK

In this section, we will review the early detection methods in the field of traffic sign detection and the advanced research results in recent years, and introduce the YOLOv8 network structure to lay the foundation for the subsequent research.

A. TRAFFIC SIGN DETECTION

Automatic identification and detection of traffic signs on the road through image processing algorithms is an indispensable key link in driverless and ITS systems, providing an important basis for subsequent actions.In the realm of traffic sign detection, there is a considerable variation in the size, color, and shape of signs, significantly augmenting the complexity of the detection task. Consequently, the exploration of algorithms resilient to the diverse nature of traffic signs stands as a pivotal and enduring challenge in this domain [23]. In order to achieve the functions of indication, warning, restriction, and guidance, traffic signage is usually distinguished from its surroundings by eye-catching colors to enhance the identifiability of the signs. Traffic signs can be roughly distinguished from other objects by color (red, yellow, or blue) and shape (triangle, circle, rectangle, or polygon). Thus, early traffic sign detection can be classified into color and shape based processing methods. Color-based detection methods [24], [25], [26], [27] perform traffic sign extraction through RGB (red, green, and blue) color space or HSI (hue, saturation, and intensity) color space. The detection performance of the above methods decreases drastically when faced with lighting changes and signage fading. The shape-based detection method [28], [29], [30] employs manually crafted features and classifiers for object detection and recognition. However, these detection algorithms exhibit sensitivity to external environmental conditions, including occlusion, deformation, and scale differences, which can substantially impact the detection performance. These scenarios are prevalent in practical application scenarios [31].

Accurate and efficient detection can effectively extract the traffic signs in the image, provide a reliable detection result for subsequent processing and analysis, and ensure the safe driving of vehicles and well-organized traffic. In recent years, traffic sign detection techniques based on deep learning architectures have been widely studied and rapidly developed, with outstanding performance on publicly available traffic sign datasets. These methods driven by data exhibit robust adaptability, demonstrating effectiveness in effectively managing background noise and fluctuations in external environmental conditions [31]. Wang et al. [22] enhanced the internal correlation between location information and channel information by adding coordinate attention to the neck, effectively fused shallow feature representation and deep semantic information, and improved the network's ability in complex environments for occluded traffic signs. Wang et al. [32] mitigated the loss of contextual information due to feature channel reduction by introducing an attention module and a feature enhancement module to improve the network's sensitivity to traffic signs. Yuan et al. [33] incorporated a path aggregation module into the feature pyramid (FPN) structure to further enhance the encoding and decoding parts by adding horizontal connections to the spatial information, effectively enhancing the network's feature representation of traffic signs under normal weather. Liang et al. [34] combined the coordinate attention module with the backbone network ResNeSt and constructed the feature pyramid for multi-scale detection [35], which enhanced the network's ability to extract shallow texture and contour information and enabled the extracted features to focus on traffic sign information, thereby improving the detection accuracy. Wang et al. [36] proposed a BANet network using bi-directional attention, which reduces the impact of shallow information loss caused by expanding the receptive field and improves the performance of small object detection and localization in traffic scenarios. Cao et al. [37] applied Swin Transformer to the neck structure of YOLOv5s, which makes the network more concerned with contextual spatial information and, combined with coordinated attention, improves the detection of traffic objects in real traffic scenarios. Overall, deep learning-based detection methods show high application potential and research value in traffic scenarios. All of the above methods have shown a good improvement in detection accuracy. However, practical application scenarios require the model to be deployed on mobile devices for fast and accurate detection of traffic signs of different sizes and shapes in real environments. Therefore, the model is required to improve the detection performance of multi-scale objects with a small number of parameters and a fast detection speed. That is to say, while meeting the real-time requirements of the vehicle-mobile terminal, the detection network should focus on narrowing the gap between the detection accuracy of different scales so as to avoid the occurrence of good detection performance only for a specific scale. In order to achieve more accurate and fast traffic sign detection, it is still necessary to further study and optimize the deep learning algorithm and strive to achieve, under the premise of meeting the detection speed, the ability to accurately detect objects of multiple scales so that the detection network has a better generalization of the frequent changes in scale. Based on the above needs, the latest single-stage detector YOLOv8s network is used as the basis in this paper, and the detection speed and the number of network parameters can well meet the needs of practical applications, on the basis of which the difference between the detection performance of different scales objects is narrowed.

B. INTRODUCTION TO THE YOLOV8 NETWORK ARCHITECTURE

The YOLO family of algorithms in computer vision detection stands out among the many detectors owing

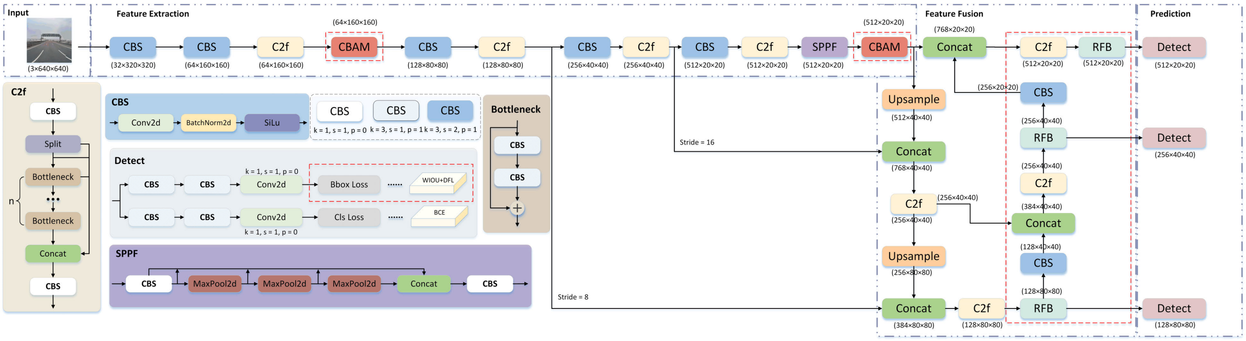


FIGURE 2. The architecture of improved YOLOv8 network structure. Among them, k is the size of the convolution kernel, s is the stripe, and p is padding; CSPBottleneck with 2 conversions (C2f) is used for concatenating different feature maps and other operations; CBS consists of Conv, BatchNorm, and SiLu activation functions for feature extraction; Spatial Pyramid Pooling Fast (SPPF) is used to increase feature diversity.

to its excellent balance of detection accuracy and speed. As a typical algorithm in one-stage detection networks, the YOLO series of detectors can quickly and reliably recognize objects in images. In the field of traffic sign detection, the real-time detection performance of YOLO is highly valuable. YOLOv8, as the latest SOTA model of the YOLO series, has better detection accuracy and speed than other versions. Based on the different requirements of the detection scenarios, there are five versions based on the scaling factor: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Starting from the fact that the network deployed on vehicle-mounted mobile platforms needs to be lightweight enough and the traffic sign detection needs to be highly accurate and responsive, this was selects YOLOv8s as YOLOv8s with a memory size of only 11.2M, which meets the deployment requirements of the in-vehicle mobile platform and is also the best choice for detection accuracy and response speed.

The YOLOv8s network consists of four parts: image input, feature extraction network, feature fusion module, and detection head. YOLOv8 preprocesses the data in the input stage, and the processing method continues the way that YOLOv5 enhances the data using Mosaic, Mixup, random perspective, and HSV augmentation. Inspired by the extended efficient layer aggregation networks(E-ELAN) module in the YOLOv7 network, the C2f module is proposed, and the combination of the three modules, CBS, C2f, and SPPF. In the feature fusion stage, a feature pyramid is constructed using the PAFPN structure to fully fuse shallow and deep feature information. Inspired by the YOLOX network detection head, the YOLOv8 detection head adopts a decoupling head structure to separate classification and positioning tasks.

YOLOv8 is the most advanced one-stage object detector, which integrates many current advanced detection methods from the practical needs of traffic sign detection, and will be based on YOLOv8s subsequent research.

III. THE PROPOSED METHOD

Traffic signs photographed in road environments vary greatly in scale, resulting in a limited amount of information about smaller objects contained in the extracted image features, which further exacerbates the difference in the accuracy of multi-scale object detection. In order to reduce the difference in the network's accuracy for multi-scale traffic signs, this paper proposes CR-Yolov8 for multi-scale traffic sign detection based on the state-of-the-art one-stage detector YOLOv8s for improvement.The main content of this section is the improvement method of YOLOv8, including (a) adding a lightweight CBAM attention mechanism in the feature extraction network, which trades a negligible computational overhead for accuracy improvement. (b) Incorporating the RFB module in the feature fusion module to enhance the feature diversity of the lightweight network model (c) Optimize the gradient gain allocation strategy through the Wise Intersection over Union (WIOU) loss function to enhance the adaptability of the detector to multiple-scale object changes. The improved network structure is shown in Figure 2.

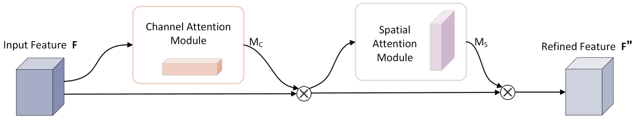


FIGURE 3. CBAM attention module.

A. ATTENTION MECHANISM

The attention mechanism is a method used to simulate the characteristics and behavior of the human perception system, which enables the model to selectively focus on important information by assigning different weights to different features to achieve the purpose of focusing on the information of the object. In the fields of machine learning and deep learning, attention mechanisms are widely used. It flexibly expresses features according to the importance of each input feature, effectively helping information flow through the network and enabling the model to adapt better to complex tasks and changing input data.

In the feature extraction phase, spatial and channel features are equally important for generating feature maps. The CBAM attention module [38] improves the expressive ability of the model to learn important features by generating

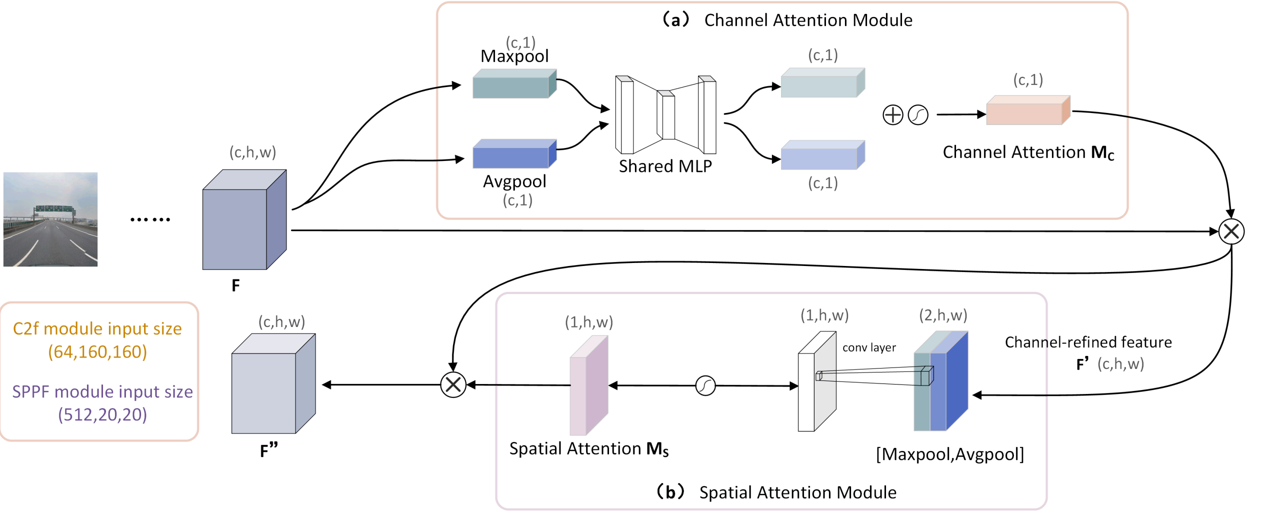


FIGURE 4. Structure of channel attention module and spatial attention module.CBAM is added at two locations in the feature fusion section. The input to the CBAM module after the C2f module is a feature matrix with a channel count of 64,160×160 and the input to the CBAM module after the SPPF module is a feature matrix with a channel count of 512,20×20.

attention feature maps in both the channel and spatial dimensions and multiplying them with the original input feature maps with adaptive adjustment. CBAM assigns varying weights to different features based on their importance, flexibly expressing features. It selectively focuses on crucial details, aiming to concentrate on target object information. The specific process is shown in Figure 3, where the intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$ is the input and then outputs the one-dimensional channel attention map $M_C \in \mathbb{R}^{C\times 1\times 1}$ and the two-dimensional spatial attention map $M_S \in \mathbb{R}^{1\times H\times W}$ in sequence, and the attention is calculated as shown in Equation (1).

$$\begin{aligned} F' &= M_C(F) \otimes F \\ F'' &= M_S(F') \otimes F' \end{aligned} \tag{1}$$

The CBAM Attention Module consists of two submodules: the Channel Attention Module and Spatial Attention Module. The Channel Attention Module utilizes the internal relationship between feature channels to produce a channel attention graph. As shown in Figure 4(a), the spatial information is compressed by average pooling and maximum pooling operations to improve the computational performance of the channel attention features, followed by the generation of two contextual feature maps, which finally generate a one-dimensional channel attention map $M_C \in \mathbb{R}^{C\times 1\times 1}$ through the multi-layer perceptron (MLP) and hidden layer. The spatial attention module uses the internal relations of the feature space to produce a spatial attention graph. As shown in Figure 4(b), average pooling and maximum pooling are performed in the channel dimension, and then they are concatenated to obtain the feature map. Finally, the 2D spatial attention map $M_S \in \mathbb{R}^{1\times H\times W}$ is obtained after the convolution operation. The use of signage and vehicle driving characteristics leads to a large difference in the pixel percentage of captured traffic signs in the image, resulting in effective feature information extracted by the backbone network that cannot take into account the three scales of large, medium, and small traffic signs. Introducing the CBAM module enables the network to focus on both channel and spatial features, comprehensively capturing traffic sign-related characteristics. This provides a more valuable feature representation, enhancing the recognizability of these signs.

CR-Yolov8 incorporates the CBAM attention mechanism into the feature extraction part after C2f and SPPF modules. After the C2f module, the CBAM produces a 160×160 feature matrix with 64 output channels. With the CBAM module, the network focuses more on capturing features related to traffic signs, providing a more valuable representation of the features, and enhancing the recognizability of these signs. The CBAM after the SPPF adaptive output module generates a feature matrix with a channel count of 512, 20×20, enhances the channel features and spatial features, and outputs the results to the concat and up-sampling modules for the subsequent feature fusion stage. The model, by introducing CBAM, in which spatial attention and channel attention are used for joint processing, reduces the impact caused by the loss of information in the sampling process based on the original model and improves the network's detection performance for multi-scale objects.

B. RECEPTIVE FIELD BLOCK

The Receptive Field Block (RFB) module [39] aims to enhance the feature diversity of lightweight networks by emulating the human visual system. This is achieved through the extraction of features from input feature maps via multiple receptive field sizes. The architecture of the RFB consists of two primary components: the Multi-Branch Convolutional Layer and the Dilated Convolutional Layer,as the structure is shown in Figure 5. A multi-branch convolutional layer is used to simulate different sizes of population receptive fields (pRF). Using the Inception [40] structure, the bottleneck structure of each branch undergoes 1×1 convolution for dimensionality reduction and is combined with an n×n convolutional layer for reducing the number of channels for feature mapping to emphasize the importance of the focus region and enhance the network's sensitivity to spatial changes. The dilated convolutional layer simulates the correlation between the size of the pRF and the eccentricity of the human visual system. RFB achieves a larger receptive field by employing convolution kernels of different sizes and dilated convolutions with various rates. Leveraging the diversity and complementarity of different features, it enhances the feature diversity of the model. CR-Yolov8 adds the RFB module to the feature fusion stage to enhance the feature diversity of the network before detecting the header. The issue of information loss and reduced resolution, typically caused by down-sampling, is mitigated while maintaining nearly identical computational overhead.

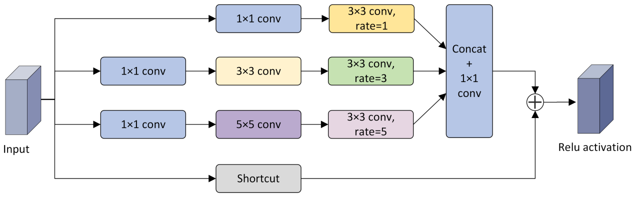


FIGURE 5. RFB structure diagram.

C. LOSS FUNCTION

The loss function of bounding box regression (BBR) helps the model optimize its ability to lock the object location information by calculating the difference between the predicted bounding box and the real bounding box at different scales. Traffic sign detection requires the detector to be robust to scale changes in an object. The bounding box loss function in the optimization process, by gradually adjusting the predicted position of the bounding box, makes the model better adapted to the scale change of the object to improve the model's ability to perceive multi-scale objects.The Complete Intersection over Union (CIOU) [41] in the original YOLOv8 network cannot efficiently measure the difference between the object frame and the Anchor, which leads to slow convergence and inaccurate localization in model optimization. Compared with CIOU, Wise intersection over union(WIOU) [42] optimizes the gradient gain allocation strategy so that the model can well balance the learning of large, medium, and small objects during the training process, which improves the overall performance of the detector. Hence, WIOU is employed to substitute CIOU in CR-Yolov8. The WIOU loss function can be defined as shown in equation(2):

$$\mathcal{L}_{WIOU} = rR_{WIOU}\mathcal{L}_{IOU}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}}$$

$$R_{WIOU} = exp\left(\frac{\left((x-x_{gt})^2+(y-y_{gt})^2\right)}{\left(W_g^2+H_g^2\right)^*}\right)$$

$$\mathcal{L}_{IOU} = 1 - IOU \tag{2}$$

where x and y are the center coordinates of the anchor frame; x_{gt} and y_{gt} are the coordinates of the center point of the object frame;W_g and H_g are the sizes of the minimum enclosing frames; and the gradient gain r is dynamically adjusted by means of α, the δ hyperparameter, and nonmonotonic focusing factor β. $\mathcal{L}_{IOU}$ is the IOU loss function.

WIOU enables YOLOv8 to equalize the learning of multi-scale objects during the training process and improve the ability to localize the targets through a dynamic allocation strategy. By introducing non-monotonicity, the model treats high-quality and low-quality examples equally during training. This means that as the loss increases, the gradient gain does not follow a monotonic pattern. While reducing the gradient gain for low-quality anchor boxes, it also decreases the gradient gain for high-quality ones. This ensures the model treats high and low-quality examples equally, aiming to stably learn effective features throughout the training process.

IV. EXPERIMENTS AND RESULTS

In this section, the experimental dataset and evaluation metrics are presented along with a more detailed description of the experimental setting, analysis of results, and ablation experiments.

A. DATASETS AND EVALUATION INDICATORS

1) DATASET

Traffic sign detection and localization are indispensable components of ITS and autonomous driving. In generalized object detection datasets such as MS COCO [43] (e.g., COCO2014), Pascal VOC [44], [45] (e.g., VOC2007, VOC2012), and ImageNet [46] (e.g., ILSVRC2014), even though they contain traffic sign images, detectors trained on generalized benchmarks cannot learn well the feature information of traffic signs. To better meet the detection needs of real-world environments, CR-Yolov8 used TT100k [47] as a benchmark to train and validate the improved network.

The TT100K dataset [47] is a large-scale traffic data benchmark jointly compiled by Tsinghua and Tencent Labs, covering street scenes of several cities in China and multiple lighting and weather conditions. The dataset contained 221 traffic sign types, such as speed limit, warning, and no passing, totaling 100,000 images and 30,000 instances. In addition, the TT100k dataset provides a large number of scene images, including city roads, highways, and rural roads, which can satisfy different application requirements. Nearly half of the instances in the TT100K dataset were severely underrepresented, resulting in an extremely unbalanced data distribution. Therefore, in this study, the dataset is processed according to the article [47], and only 45 categories with instance counts of 50 or more. The resolution of the processed data is 640 × 640 RGB image,, with 5962 images in the training set and 2979 images in the test set.

TABLE 1. Experimental results of different algorithms on TT100k.

Method	Input size	Params(M)	FPS	mAP	mAP@0.5	AP_S	AP_M	AP_L
SSD [10]	300×300	26.07	24	0.318	0.346	0.028	0.237	0.561
M2det [11]	512×512	147	12	-	0.466	0.030	0.320	0.657
YOLOv3 [12]	640×640	59.48	27	0.389	0.579	0.389	0.445	0.429
Wang et al. [32]	608×608	8.04	95	-	0.651	0.415	0.578	0.582
YOLOv5-HC [48]	640×640	89.5	-	-	0.790	0.671	0.921	0.951
YOLOv5s	640×640	7.2	156	0.610	0.813	0.721	0.796	0.824
YOLO-SG [49]	640×640	4.0	131	-	0.758	0.456	0.618	0.637
YOLOv8s	640×640	11.2	119	0.638	0.846	0.743	0.819	0.841
CR-YOLOv8(Ours)	640×640	14.6	103	**0.651**	**0.869**	**0.759**	**0.821**	**0.844**

2) EVALUATION METRICS

For object detection, the evaluation criteria generally include Precision, Recall, and Mean Average Precision (*mAP*), as shown in equations(3) to(5). where TP represents the number of correct detections by the detector, FP represents the number of detector localization errors, FN represents the number of false and missed detections by the detector, Precision represents the accuracy of the algorithm with respect to the results of the detection, Recall expresses the algorithm's ability to check the entirety of the algorithm, and *mAP* measures the comprehensive performance of the object detection algorithm in multiple categories.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{5}$$

CR-Yolov8 aims to enhance the performance of traffic signs at multiple scales. For traffic sign detection, the detection accuracy, detection speed, and model parameter size have a significant impact on practical application scenarios. Therefore, in addition to the commonly used object detection evaluation metrics, FPS, AP_S, AP_M, and AP_L were also used to comprehensively evaluate the processing speed of the model and the detection accuracy of objects at different scales. In addition, parametric params of the network model are used to determine whether they meet the deployment requirements.

B. RESULT ANALYSIS AND ABLATION EXPERIMENT

The experimental environment for this experiment was a LINUX x86_64 operating system, Xeon(R) Gold 6230R, graphics card NVIDIA Tesla A100 40G×2, 128G GB RAM, and an experimental platform with CUDA 11.3, PyTorch 1.11, and Python 3.8. The network uses pre-trained weights on ImageNet for transfer learning during training.The initial learning rate of the model is 0.01, the bactch size is 32, and the epoch is 200.

CR-YOLOv8 was compared with other state-of-the-art traffic sign detection algorithms to validate the effectiveness in the field of multi-scale traffic sign detection. The evaluation was carried out on TT100k. The detection performance of the network was evaluated by comparing the parameters of the model, frames per second (FPS),mean average precision (mAP), and detection accuracy for small, medium, and large sized (AP_S, AP_M, and AP_L) objects. The experimental results are shown in Table 1.

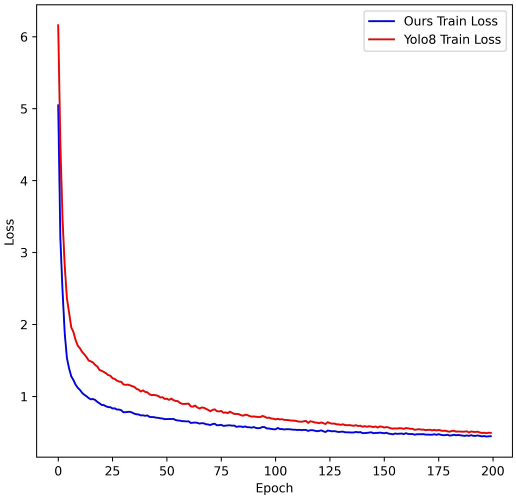


FIGURE 6. Comparison of training losse between CR-YOLOv8 and YOLOv8s models.

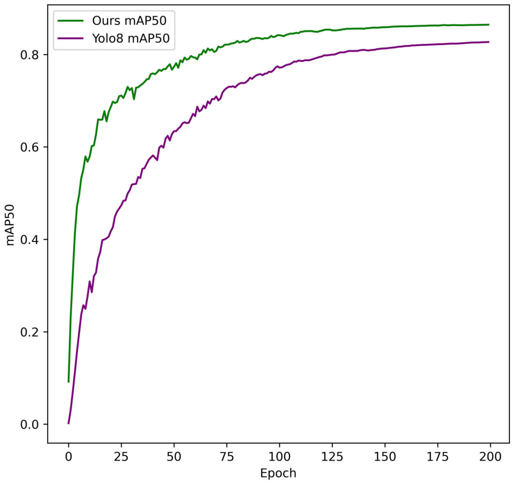
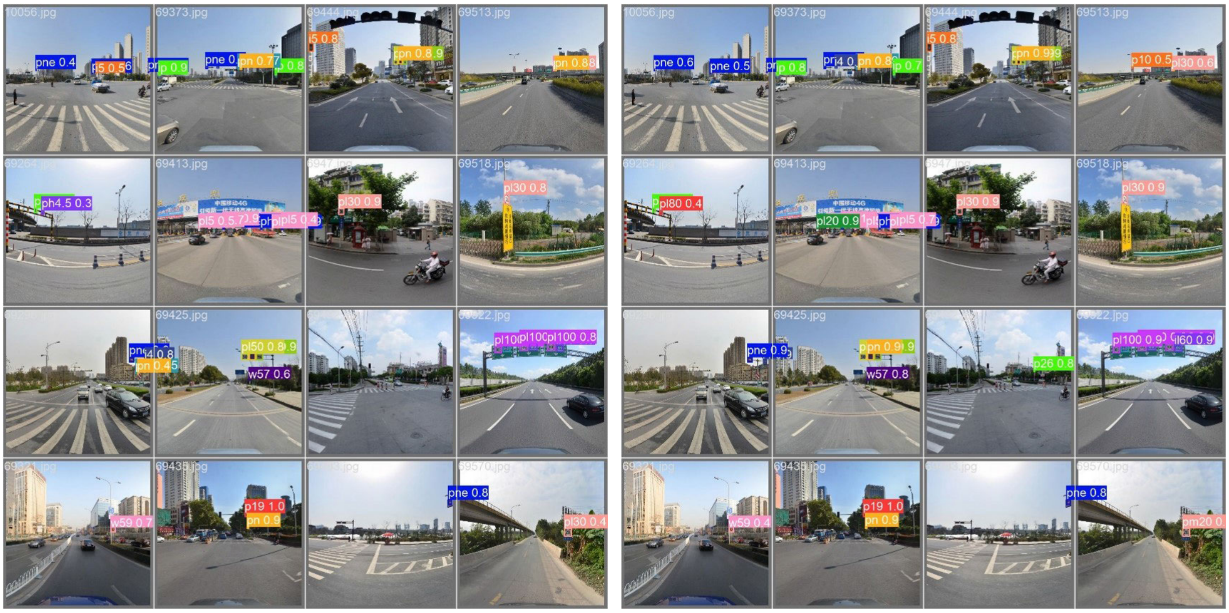


FIGURE 7. Comparison of mAP between CR-YOLOv8 and YOLOv8s models.

From the Table 1, it can be seen that the performance of the model proposed CR-YOLOv8 model has been improved compared to state-of-the-art traffic sign detection networks.



(a) Original

(b) CR-YOLOv8(Ours)

FIGURE 8. Visualization results of traffic sign detection on the TT100k dataset. (a) shows an example of detection on the benchmark network YOLOv8s. (b) is an example of detection on the CR-YOLOv8.

The accuracy was improved by 2.3 percentage points compared to that of YOLOv8s. In terms of AP_S, AP_M, and AP_L, the proposed model has different degrees of improvement compared to the YOLO series of algorithms. The highest improvement is in AP_S, which is 1.6 percentage points higher, and the improvement effect is obvious in the performance of small-scale traffic-sign detection. Despite the slight increase in the number of parameters it does not have a significant impact on the performance, and this task focuses more on improving the detection accuracy of multi-scale object, where the effect of the improvement is more pronounced at this order of magnitude. This proves that CR-YOLOv8 effectively shortens the difference in the detection accuracy of objects at different scales and successfully optimizes the multi-scale detection effect. Figure 6 and Figure 7 show a comparison of the training process between the CR-YOLOv8 model and the YOLOv8s model. It can be seen that the CR-YOLOv8 model converges faster and the training process is smoother than YOLOv8s, indicating that CR-YOLOv8 has better stability and reliability in traffic sign detection tasks.

Ablation experiments were performed on TT100k to verify the effects of the CBAM, RFB, and loss functions on network performance. As shown in Table 2, network improvement increased the accuracy of multiscale traffic sign detection. To address the problem of information loss in the feature extraction stage of the network, the CBAM module is added at the bottom of the backbone network, which makes the network more focused on effective feature information, and the detection accuracy is improved by 0.6 percentage points. It can be seen from Table 3 that the CBAM module is more effective than other attention mechanisms. To enhance the feature diversity of the lightweight network, the RBF module was added to the feature fusion part to improve the network accuracy by 1.2 percentage points.By improving the loss function, the model is well balanced to learn multi-scale objects during the training process and has better adaptability to model scale changes. It can be seen from Table 4 that compared with CIOU, intersection over Union (IOU) and generalized intersection over Union (GIOU), WIOU has the best effect on improving the network performance, which improves the detection accuracy by 1.4 percentage points.The detection network incorporating the three modules improves by 2.3 percentage points compared to the baseline network, which effectively improves the performance of multi-scale traffic signs.

TABLE 2. Ablation experiments on TT100k.

Experimental programmes	CBAM	RFB	WIOU	mAP@0.5
0				0.846
1	✓			0.852(+0.006)
2		✓		0.858(+0.012)
3			✓	0.860(+0.014)
4	✓	✓		0.863(+0.017)
5	✓		✓	0.861(+0.015)
6		✓	✓	0.866(+0.020)
7	✓	✓	✓	0.869(+0.023)

Figure 8 shows the visualization results of the benchmark network and the improved network on TT100k, which shows that CR-YOLOv8 network has a higher detection accuracy for traffic signs of different sizes and shapes, and the objects not detected by the benchmark network on the left side are also improved in the method on the right side. CR-YOLOv8 network can effectively improve the different-scale sign

TABLE 3. Comparison of results of different attention mechanisms.

Experimental programmes	Recall	Precision	mAP@0.5
Baseline	0.751	0.856	0.846
+SE	0.749	0.863	0.849(+0.003)
+CA	0.747	0.871	0.849(+0.003)
+CBAM	0.760	0.878	**0.852(+0.006)**

TABLE 4. Performance of each bounding box loss.

Experimental programmes	Recall	Precision	mAP@0.5
CIOU	0.751	0.856	0.846
IOU	0.747	0.825	0.831
GIOU	0.736	0.861	0.837
WIOU	**0.759**	**0.867**	**0.860**

detection accuracy, narrow the multi-scale detection accuracy imbalance problem, and is more suitable for multi-scale traffic sign detection

V. CONCLUSION

In summary, this paper presents a multi-scale traffic sign detection model CR-Yolov8 based on a single-stage detection network. The proposed approach effectively mitigates the information loss problem caused by down-sampling, thereby preserving richer object information in the high-level feature map. Therefore, the network is better equipped to focus on and learn important features within the feature map. The experimental results on TT100k dataset demonstrate that the CR-Yolov8 significantly improves the performance of traffic sign detection across large, medium, and small scales, and alleviates the imbalance of precision measurement among multi-scale traffic signs. Considering the inherently complex and dynamic nature of the traffic sign detection environment, future research should explore strategies to maintain or even enhance detection performance under adverse weather conditions such as fog, snow, and low-light scenarios.

REFERENCES

[1] D. J. Edwards, J. Akhtar, I. Rillie, N. Chileshe, J. H. K. Lai, C. J. Roberts, and O. Ejohwomu, "Systematic analysis of driverless technologies," *J. Eng., Des. Technol.*, vol. 20, no. 6, pp. 1388–1411, Dec. 2022.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mali, Jun. 2014, pp. 580–587.

[3] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[4] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10526–10535.

[5] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14449–14458.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[8] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[11] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2Det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Art. Intel. (AAAI)*, vol. 33, Jan. 2019, pp. 9259–9266.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[13] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, and W. Nie, "YOLOV6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[15] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOx: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[16] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," 2021, *arXiv:2105.04206*.

[17] J. Yu and W. Zhang, "Face mask wearing detection algorithm based on improved YOLO-v4," *Sensors*, vol. 21, no. 9, p. 3263, May 2021.

[18] D. Wu, S. Lv, M. Jiang, and H. Song, "Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments," *Comput. Electron. Agricult.*, vol. 178, 2020, Art. no. 105742.

[19] C.-Y. Wang, A. Bochkovskiy, and H. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13024–13033.

[20] H. Gong, T. Mu, Q. Li, H. Dai, C. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, H. Li, X. Lang, Z. Li, and B. Wang, "Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images," *Remote Sens.*, vol. 14, no. 12, p. 2861, Jun. 2022.

[21] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[22] G. Wang, K. Zhou, L. Wang, and L. Wang, "Context-aware and attention-driven weighted fusion traffic sign detection network," *IEEE Access*, vol. 11, pp. 42104–42112, 2023.

[23] M. S. Mohammed, A. Al-Dhamari, W. Saeed, F. N. Al-Aswadi, S. A. M. Saleh, and M. N. Marsono, "Motion pattern-based scene classification using adaptive synthetic oversampling and fully connected deep neural network," *IEEE Access*, vol. 11, pp. 119659–119675, 2023.

[24] Y. Ishizuka and Y. Hirai, "Segmentation of road sign symbols using opponent-color filters," in *Proc. ITSWC*, vol. 18, 2004, p. 22.

[25] K. Baba and Y. Hirai, "Real-time recognition of traffic signs using opponent color filters," in *Proc. 14th World Congr. Intell. Transp. Syst. (ITS)*, Oct. 2007, p. 12.

[26] A. de la Escalera, L. E. Moreno, M. A. Salichs, and J. M. Armingol, "Road traffic sign detection and classification," *IEEE Trans. Ind. Electron.*, vol. 44, no. 6, pp. 848–859, 1997.

[27] J. Miura, T. Kanda, and Y. Shirai, "An active vision system for real-time traffic sign recognition," in *Proc. IEEE Intell. Transp. Syst.*, Oct. 2000, pp. 52–57.

[28] A. Ellahyani, M. El Ansari, R. Lahmyed, and A. Trémeau, "Traffic sign recognition method for intelligent vehicles," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 35, no. 11, pp. 1907–1914, 2018.

[29] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma, "A hybrid vehicle detection method based on Viola-Jones and HOG + SVM from UAV images," *Sensors*, vol. 16, no. 8, p. 1325, Aug. 2016.

[30] A. Møgelmose, D. Liu, and M. M. Trivedi, "Detection of US traffic signs," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3116–3125, Jun. 2015.

[31] X. Pan, T. Yang, Y. Xiao, H. Yao, and H. Adeli, "Vision-based real-time structural vibration measurement through deep-learning-based detection and tracking methods," *Eng. Struct.*, vol. 281, Apr. 2023, Art. no. 115676.

[32] J. Wang, Y. Chen, Z. Dong, and M. Gao, "Improved YOLOv5 network for real-time multi-scale traffic sign detection," *Neural Comput. Appl.*, vol. 35, no. 10, pp. 7853–7865, Apr. 2023.

[33] X. Yuan, A. Kuerban, Y. Chen, and W. Lin, "Faster light detection algorithm of traffic signs based on YOLOv5s-A2," *IEEE Access*, vol. 11, pp. 19395–19404, 2023.

[34] T. Liang, H. Bao, W. Pan, and F. Pan, "Traffic sign detection via improved sparse R-CNN for autonomous vehicles," *J. Adv. Transp.*, vol. 2022, pp. 1–16, Mar. 2022.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[36] S.-Y. Wang, Z. Qu, C.-J. Li, and L.-Y. Gao, "BANet: Small and multi-object detection with a bidirectional attention network for traffic scenes," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105504.

[37] Y. Cao, C. Li, Y. Peng, and H. Ru, "MCS-YOLO: A multiscale object detection method for autonomous driving road environment recognition," *IEEE Access*, vol. 11, pp. 22342–22354, 2023.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[39] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 385–400.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–7.

[41] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Feb. 2020, pp. 12993–13000.

[42] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

[44] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[45] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[46] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[47] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2110–2118.

[48] J. Xu, Y. Huang, and D. Ying, "Traffic sign detection and recognition using multi-frame embedding of video-log images," *Remote Sens.*, vol. 15, no. 12, p. 2959, Jun. 2023.

[49] Y. Han, F. Wang, W. Wang, X. Li, and J. Zhang, "YOLO-SG: Small traffic signs detection method in complex scene," *J. Supercomput.*, pp. 1–22, Jul. 2023. [Online]. Available: https://doi.org/10.1007/s11227-023-05547-y

LU JIA ZHANG was born in Beijing, China, in 1996. She received the bachelor's degree in computer science and technology from the Smart City College, Beijing Union University, in 2019. She is currently pursuing the master's degree with the School of Robotics, Beijing Union University. Her research interests include image recognition and deep learning and applications.

JIAN JUN FANG JR. received the B.Eng. degree in mechanical engineering from Huazhong Agricultural University, in 1993, and the Ph.D. degree in mechanical engineering from China Agricultural University, in 1998. He is currently a Full Professor with the Department of Transportation Engineering, Beijing Union University, Beijing, China. His research interests include intelligent transportation, special-purpose robots, machine vision, and deep learning.

YAN XIA LIU JR. received the Ph.D. degree from the School of Automation and Electrical Engineering, University of Science and Technology Beijing, in 2013. She is currently a Professor with the College of Urban Rail Transit and Logistics, Beijing Union University. Her current research interests include pattern recognition, computer vision, deep learning, and intelligent instruments.

HAI FENG LE received the bachelor's degree from the College of Urban Rail Transit and Logistics, Beijing Union University, in 2019. He is currently pursuing the master's degree with the School of Robotics, Beijing Union University. His research interests include deep learning and applications and computer graphics.

ZHI QIANG RAO JR. received the Ph.D. degree in engineering from the Wuhan University of Technology, in 2011. He is currently an Associate Professor with the College of Urban Rail Transit and Logistics, Beijing Union University. His current research interests include rail traffic safety, deep learning, and digital image processing.

JIA XIANG ZHAO received the B.S. degree in transportation from the North China University of Science and Technology. He is currently pursuing the master's degree with the College of Urban Rail Transit and Logistics, Beijing Union University. His research interests include intelligent transportation systems and deep learning.