BGC-YOLO: A Feature Fusion-Based Algorithm for Traffic Sign Detection

Shuo Cui[1,2], YingZhao Xue[1,2], and ZeKai Liu[1,2]
[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong, China
[2] Shanxi Provincial Key Laboratory of Intelligent Optimization Computing and Blockchain Technology, Jinzhong, China

Abstract—With the development of intelligent transportation systems, the automatic detection of traffic signs has become a key task in assisted driving and unmanned driving perception systems. In view of the problem that traffic signs are small in scale in images and their accuracy is affected by complex environments, this paper constructs a BGC-YOLO target detection algorithm based on YOLOv11. First, by introducing the bidirectional feature fusion structure BiFPN, the interactive expression of multi-scale features is enhanced. Secondly, the global-local spatial attention mechanism GLSA is combined to improve the model's perception of detail information and contextual semantics. Finally, the content-aware upsampling module CARAFE is used to optimize the feature reconstruction process and effectively retain the key information of small targets. The experimental results on the CCTSDB2021 traffic sign dataset show that the improved model achieves a good balance between accuracy and efficiency, with an increase of 1.4% in mAP@0.5 compared to the original model, and maintains a low computational overhead, which is practical.

Index Terms—Traffic sign detection, YOLOv11, feature fusion, object detection

I. INTRODUCTION

Traffic signs, as the core carrier of road information, play a vital role in ensuring the safe driving of vehicles. They are also an indispensable key link in realizing autonomous driving technology. Due to the wide variety of traffic signs, they often appear in complex and changing background environments. In addition, the system requires real-time detection results, making automatic detection and recognition of traffic signs a very challenging task.

Early traffic sign detection methods mainly rely on artificially designed features such as color and shape to achieve classification and recognition. For example, Bahlmann[1] proposed a method that uses color, shape and motion information for traffic sign detection; Li H[2] combined color segmentation and robust shape matching with a new method and used support vector machines for classification. Although these traditional methods have achieved results to a certain extent, they generally rely on specific manual feature design for different traffic signs and are easily affected by environmental noise, resulting in poor robustness. To overcome these limitations, researchers began to introduce deep learning models into traffic sign detection tasks[3]. Compared with traditional methods, deep learning-based models have become the mainstream technical path in the field of traffic sign detection in recent years due to their higher recognition accuracy and stronger anti-interference ability.

At present, the object detection methods based on deep learning are mainly divided into two categories: two-stage detection algorithms represented by the R-CNN[4] series and single-stage detection algorithms represented by YOLO[5]. The two-stage method usually generates candidate regions first and then performs classification and regression. Although it performs well in detection accuracy, it is relatively slow due to the complex process. In contrast, the single-stage algorithm omits the step of generating candidate regions, which can achieve faster detection speed and is suitable for real-time applications. However, it still has certain shortcomings in detection accuracy, especially in processing small objects.

In order to further break the limitations of traditional convolutional architecture in modeling long-distance dependencies and object relationships, the DETR model was proposed[6], which introduced the Transformer architecture to build a new end-to-end object detection framework. DETR transforms the object detection task into a set prediction problem, no longer relying on candidate region generation or non-maximum suppression, and realizes the modeling of global image information through the self-attention mechanism. This method shows significant advantages in modeling object relationships and complex semantic contexts, and is particularly suitable for optimizing object position and category prediction in dense scenes. However, DETR still has shortcomings in convergence speed and small target detection, which has prompted the proposal of a series of improved variants to balance detection accuracy and training efficiency.

Aiming at the problem of missed detection of small targets when the span of traffic signs is large, as well as the problem of

This research was funded by the Shanxi Provincial Science and Technology Strategic Research Special Key Project (Project No.: 202304031401011) and the Shanxi Provincial Basic Research Plan (Free Exploration) Project (Project No.: 202403021222276). The corresponding author is Shuo Cui (email: cs811767@163.com). The author YingZhao Xue (email: 18366902381@163.com) and ZeKai Liu(email: 15536368230@163.com) are from the School of Computer Science and Technology, Taiyuan Normal University.

false detection in complex environments, this paper takes the YOLOv11n model as the basic architecture from the perspective of improving detection accuracy and robustness, comprehensively considers the detection speed and deployment efficiency of the model, and proposes a BGC-YOLO traffic sign detection model. The work of this paper is as follows:

1) In order to enhance the multi-scale feature fusion capability, the BiPFN feature fusion network is introduced. Through richer bidirectional paths and cross-layer connections, the feature expression capability of targets of different scales, especially small target traffic signs, is improved.
2) The GLSA attention mechanism is introduced in the Neck part to enhance the information selectivity of the model in the feature fusion process. GLSA pays attention to local details and global context at the same time. By weighted selection of semantic features at different levels, it effectively improves the model's perception of the edge and shape details of traffic signs and improves the accuracy of target recognition under complex background interference.
3) The lightweight and efficient CARAFE module is used to replace the original nearest neighbor interpolation method. CARAFE achieves more accurate high-resolution feature reconstruction through content-aware reconstruction mechanism, effectively preserving the detailed information of small target traffic signs.

II. Related Works

A. R-CNN Series Object Detectors

The R-CNN family has had a significant impact on the evolution of deep learning-based object detection frameworks. The original R-CNN framework was proposed by Girshick et al. in 2014. It proposed a two-stage detection process: using selective search to generate region proposals, each region is independently passed through a CNN to extract features, and then classified and bounding box regression is performed. Although R-CNN shows high detection accuracy, its computational efficiency is low due to the redundant forward propagation of thousands of regions, which poses a challenge for real-time applications.

To overcome these limitations, Fast R-CNN[7] was born, which processes the entire image only once through the convolutional backbone network. Then, region of interest (RoI) pooling is used to map region proposals to feature maps, which significantly improves speed and reduces memory usage. Faster R-CNN[8] further improves on this by introducing a region proposal network to generate region proposals directly from shared convolutional features, thereby building an end-to-end trainable detection system with state-of-the-art accuracy and higher efficiency.

Later advances, such as Mask R-CNN[9], extended Faster R-CNN by adding parallel branches, demonstrating the adaptability of the R-CNN family to more complex visual tasks. Other variants, such as Cascade R-CNN[10], Libra R-CNN[11], and R-FCN[12], further optimized multi-stage training, balanced feature representation, and fully convolutional reasoning to improve detection performance (both precision and recall).

Overall, the R-CNN family represents the foundational paradigm for two-stage object detection, known for its strong accuracy and scalability. However, computational complexity and inference speed remain limiting factors for real-time and resource-constrained applications, such as autonomous driving or embedded traffic sign detection systems.

B. YOLO Series Object Detectors

The YOLO family represents one of the most influential research directions in the field of real-time object detection. Unlike two-stage detectors such as R-CNN, the YOLO family adopts a single-stage end-to-end framework that can directly predict the object category and bounding box in the entire image in a single network transmission. This unified architecture significantly improves the inference speed, making YOLO particularly suitable for real-time applications such as autonomous driving and video surveillance.

The first version of YOLO, YOLOv1[13], was proposed by Redmon et al. It defines object detection as a regression problem, dividing the input image into a fixed grid and predicting the bounding box and category probability based on each grid cell. Although YOLOv1 exhibits impressive speed, it has poor localization accuracy and has difficulty detecting small or clustered objects.

To overcome these limitations, YOLOv2[14] introduced anchor boxes, batch normalization, and a new backbone network, which significantly improved accuracy without sacrificing speed. YOLOv3[15] further improved this performance by using multi-scale prediction and a deeper backbone network (Darknet-53), achieving a good balance between detection performance and inference speed for objects of different sizes.

YOLOv4[16], developed by Bochkovskiy et al., strikes a balance between accuracy and deployability, incorporating a variety of modern training strategies such as cross-stage partial connections (CSP), Mish activation function, and CIoU loss function. It achieves state-of-the-art performance on the COCO benchmark and has good generalization ability and efficiency.

The advent of YOLOv5[17] marked the transition of the model to a PyTorch-based implementation, which makes the model more widely adopted and easier to customize. YOLOv5 introduced a series of lightweight models and emphasized the feasibility of practical deployment by focusing on speed, scalability, and compatibility with various platforms.

YOLOv6[18], YOLOv7[19], and YOLOv8[20] further pushed the boundaries. YOLOv6 improves the neck and head design for industrial applications. YOLOv7 proposes an extended E-ELAN module and auxiliary head to improve detection accuracy and convergence speed. YOLOv8, developed by Ultralytics, focuses on unified tasks (detection, segmentation, pose estimation), adopts anchor-free detection head and decoupled head design to achieve better performance on different data sets.

The evolution from YOLOv9 to YOLOv13 continues to push lightweight object detection technology to break through the

limits. YOLOv9[21] optimizes feature extraction and gradient flow through programmable gradient information (PGI) and general efficient layer aggregation network (GELAN), achieving millisecond-level response on edge devices; YOLOv10[22] eliminates NMS dependence with an end-to-end architecture and combines spatial channel decoupling and downsampling technology to achieve a new benchmark for real-time detection on edge devices; YOLOv11[23] introduces a dynamic detection head to significantly improve the ability to parse complex scenes while maintaining its lightweight; YOLOv12[24] achieves global semantic modeling with extremely low computational cost by relying on regional attention (A2) and Flash Attention mechanisms; the latest YOLOv13[25] breaks the constraints of traditional architecture with HyperACE and FullPAD, and demonstrates excellent energy efficiency in scenarios such as drone inspection and smart wearables, promoting the development of target detection technology towards a more edge and real-time direction. In summary, the YOLO family has evolved from a basic real-time detector to a highly optimized family of robust models that achieve an excellent balance between speed and accuracy. Due to their simplicity, scalability, and computational efficiency, these detectors have become the cornerstone of many object detection systems.

C. DETR Series Object Detectors

The introduction of the Transformer architecture has brought a new research paradigm to the field of object detection. The most representative work is the DETR model proposed by Carion. DETR first applied the Transformer encoder-decoder structure to the object detection task, innovatively transformed the detection problem into a set prediction task, and omitted the candidate region generation and non-maximum suppression (NMS) modules commonly used in traditional methods. It uses the self-attention mechanism to model the global features of the image, and has good end-to-end trainability and structural simplicity. However, DETR has problems such as slow convergence speed and insufficient detection ability for small targets, which limits its wide deployment in practical applications.

To address these shortcomings, researchers have made many improvements to the DETR model and formed multiple variants, forming the DETR series of detectors. Among them, Deformable DETR[26] introduces a sparse multi-scale deformable attention mechanism, which enables the model to focus on local key positions, effectively improving the convergence speed and small target detection performance. Conditional DETR[27] uses a content-based dynamic query vector to enhance the model's adaptability to target semantics. DAB-DETR[28] introduces the idea of dynamic anchor boxes in the target query mechanism and improves positioning accuracy by iteratively optimizing the reference box position. Furthermore, DN-DETR[29] adopts a noise learning strategy to improve training stability and matching efficiency by introducing positive and negative sample noise. DINO[30], which combines the above optimization strategies, achieves a dual improvement in detection accuracy and training efficiency, and achieves excellent performance on multiple benchmark datasets. In addition, H-DETR[31] further enhances the local perception ability of the model by introducing the fusion of convolutional features and Transformer features, which is particularly suitable for dense small target scenes.

Overall, the DETR series of methods gradually make up for the limitations of the original model by introducing multi-scale features, dynamic query and auxiliary training mechanisms while maintaining the simplicity of structure and global modeling capabilities. It has become one of the key directions of Transformer architecture research in the field of target detection, and has shown broad application prospects in tasks such as autonomous driving, remote sensing image analysis, and traffic sign detection.

III. METHODS

A. BGC-YOLO Overview

In order to effectively improve the model's ability to detect multi-scale small targets in complex traffic environments, this paper proposes an improved detection model based on YOLOv11, BGC-YOLO. Its overall method introduces structural optimization and module integration to enhance performance from three dimensions: feature fusion, attention mechanism, and upsampling strategy. BGC-YOLO aims to solve problems such as complex background interference, insufficient expression of multi-scale features, and low accuracy in small target recognition.

First, in the feature fusion network, BGC-YOLO uses BiPFN (Bidirectional Pyramid Feature Network) to replace the PANet structure in the original YOLOv11. BiPFN introduces a bidirectional feature transfer mechanism to establish a more sufficient information flow between high-level semantic features and low-level detail features, thereby enhancing the model's ability to detect targets of different scales, especially small-sized traffic signs.

Secondly, the GLSA (Global-Local Selective Attention) attention mechanism is introduced in the Neck part to improve the model's ability to select information during feature fusion. Compared with the traditional attention module, GLSA integrates global context and local detail features, which can effectively highlight the key areas related to the target under complex backgrounds, and improve the discriminability and robustness of feature expression.

In addition, the model uses the CARAFE module to replace the original nearest neighbor interpolation or deconvolution method in the upsampling stage. CARAFE adaptively reconstructs spatial features through the content-aware dynamic convolution kernel generation mechanism, effectively retains the image structure details, and improves the restoration quality of small targets in high-resolution feature maps. This not only improves the model's perception of fine-grained targets, but also alleviates the common information loss problem in the upsampling process to a certain extent, which helps to improve the overall detection performance.

In summary, BGC-YOLO integrates three modules, BiPFN, GLSA and CARAFE, on the basis of the YOLOv11 framework,

and significantly enhances the model's robustness in complex backgrounds and the detection accuracy of small targets while maintaining the original detection speed advantage. The organic integration of the three in structure enables the model to complement each other in multi-scale feature modeling, contextual information attention, and detail feature reconstruction, building a more efficient, lightweight and expressive detection framework.

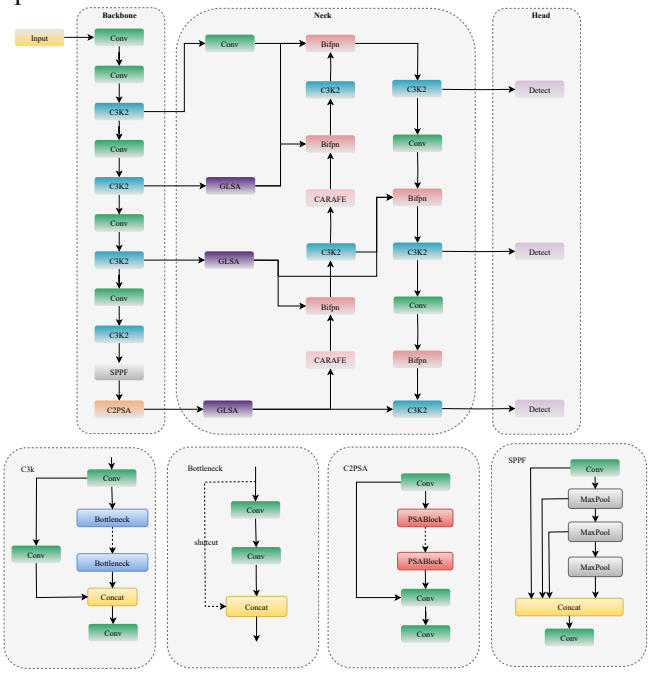


Fig. 1. BGC-YOLO structure

B. BiPFN

Although PANet can enhance the semantic transmission capability of features at different levels, it still suffers from problems such as insufficient feature flow and incomplete semantic fusion in small target detection scenarios. To address the problems of PANet in traffic sign detection tasks, such as low efficiency in small target feature transmission, insufficient fusion of upper and lower layer information, and lack of dynamic feature control capabilities, BiFPN[32] was used to replace the original PANet structure.

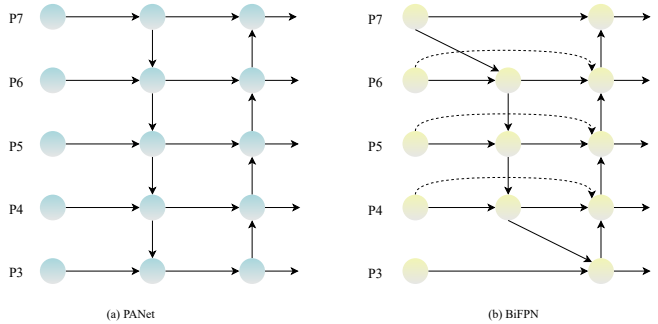


Fig. 2. BiFPN structure

As shown in Figure 2, BiFPN achieves bidirectional information flow between features at different levels by constructing bidirectional paths from top to bottom and from bottom to top, effectively enhancing the ability to integrate high-level semantics with low-level details. At the same time, the introduced learnable weighting mechanism allows the model to dynamically assign the importance of different feature layers according to task requirements, improving the flexibility and accuracy of feature fusion. Compared with the static fusion method of PANet, BiFPN has a streamlined structure and further improves the perception of multi-scale traffic signs, especially small-sized targets, significantly improving the detection performance in complex traffic scenarios.

C. GLSA

In order to further improve the model's perception of small-target traffic signs in complex traffic scenes, this paper introduces the GLSA[33] module for feature preprocessing before the feature fusion network BiFPN. Traditional attention mechanisms such as SE and CBAM have performed well in improving model feature selectivity, but most of them focus on a single scale or spatial channel and lack unified modeling of global context and local details. GLSA can enhance the expressiveness of the input features before fusion, so that the BiFPN operation can integrate multi-scale information based on more discriminative features, thereby achieving an overall grasp of the large-scale semantic structure and full retention of small-scale sign details, and enhancing the model's detection performance in multi-scale environments.

The GLSA module combines the advantages of local spatial attention (LSA) and global spatial attention (GSA). As shown in Figure 3, LSA focuses on the spatial detail information of the traffic sign area, especially has a stronger response to small differences such as pixel-level edges and shapes, and effectively improves the model's feature sensitivity to distant, blurred or partially occluded targets; while GSA strengthens the structural semantic understanding of the entire image by modeling the long-distance dependency between pixels in the image, and suppresses redundant background textures and external noise interference. The two work together through a cross-scale fusion mechanism, enhancing the global context modeling capability while retaining local fine information, significantly improving the model's ability to discriminate traffic signs of different sizes and semantic levels.

Specifically, GLSA first splits the input feature map along the channel dimension to obtain two sub-features, which are input into the GSA branch and the LSA branch respectively. The GSA branch models long-distance pixel relationships and supplements the missing semantic context information in local features; the LSA branch focuses on local key areas, enhances the detail expression ability of features, and alleviates the problem of small target information being diluted in deep features. Subsequently, GLSA concatenates the outputs of the two branches into fused features in the channel dimension, and then compresses the channel through 1×1 convolution to generate the final output features. This processing method not only improves the diversity and accuracy of feature expression, but also avoids a significant increase in the amount of calculation, ensuring the adaptability of the module to real-time detection tasks. The calculation formula of GLSA is as follows:

$$X_0, X_1 = \text{Split}\ ^\circ X \tag{1}$$

$$Att_G(X_0) = Soft\max(Transpose(Conv_{1\times1}(X_0))) \tag{2}$$

$$GSA(X_0) = MLP(Att_G(X_0) \otimes X_0) + X_0 \quad (3)$$

$$Att_L(X_1) = Sigmoid(Conv_{1\times1}(DWconv_{3\times3}(Conv_{1\times1}(\cdot))_{\times3}) + X_1) \quad (4)$$

$$LSA(X_1) = Att_L(X_1) \otimes X_1 + X_1 \quad (5)$$

$$Y = Conv_{1\times1}(Contact(GSA(X_0), LSA(X_1))) \quad (6)$$

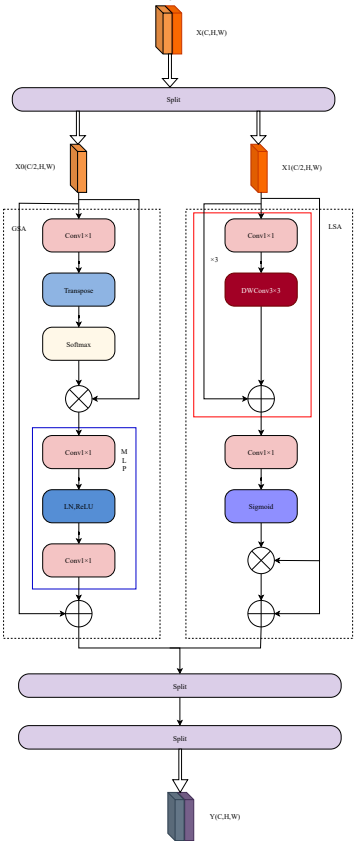


Fig. 3. GLSA structure

D. CARAFE

In the task of traffic sign detection, images often contain complex backgrounds, dynamic interference (such as lighting changes, rainy and foggy weather, occlusions, etc.) and small-scale targets. Conventional upsampling methods easily lead to blurred and discontinuous feature edges, which in turn causes the feature information of small targets to be lost during the upsampling process, affecting the detection accuracy. In order to solve the above problems, a lightweight upsampling module CARAFE[34] is introduced to replace the traditional upsampling operator to enhance the feature reconstruction capability and the ability to retain contextual information.

As shown in Figure 4, the CARAFE module mainly consists of two parts: an upsampling kernel prediction module and a feature reorganization module. In the upsampling kernel prediction module, a small convolution kernel is first used to compress the input feature map to reduce the computational complexity; then the compressed features are processed by the content encoder to generate a reorganization weight matrix at the corresponding position, which is normalized and used as an adaptive upsampling convolution kernel. Then, the feature reorganization module uses the convolution kernel to reorganize the original low-resolution feature map to complete the generation of a high-resolution feature map. Different from traditional interpolation methods that upsample based on fixed geometric rules, CARAFE adopts a content-based dynamic convolution method to capture contextual information within a larger receptive field. It can flexibly adjust the upsampling strategy according to the semantic and texture characteristics of specific image areas, thereby effectively retaining detail information.

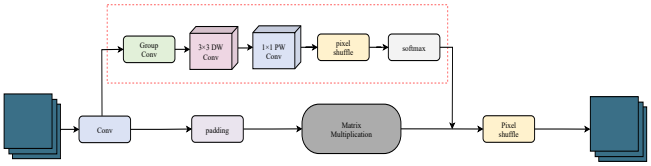


Fig. 4. CARAFE structure

CARAFE has obvious advantages in improving small target detection capabilities for targets such as traffic signs, which have small scales, fine edges, and complex shapes. Its large receptive field and dynamic content perception mechanism help to enhance the discriminability of high-resolution features and reduce semantic ambiguity caused by upsampling. In addition, CARAFE has a lightweight structural design and low computational overhead. It can improve overall detection performance without significantly increasing model complexity, and is suitable for traffic scenarios that require both real-time performance and accuracy.

IV. Experiments

A. Implementation Details

We conducted extensive experiments on the CCTSDB2021[35] dataset. All experiments were performed on an NVIDIA GeForce RTX 4090 GPU. Our network was trained for 200 epochs using the stochastic gradient descent (SGD) optimizer with a momentum of 0.937, a weight decay of 0.0005, a batch size of 32, and an initial learning rate of 0.01.

In order to comprehensively and objectively evaluate the detection performance of various algorithms in traffic scenarios, this experiment uses multiple indicators as performance evaluation criteria, including precision, recall, mean average precision (mAP), and floating point operations (FLOPs). In addition, FLOPs is used to measure the total computational effort of the model during forward reasoning, which is an important indicator for judging the computational overhead and complexity of the model.

B. Ablation experiment

In order to verify the effectiveness of the module designed in this paper, an ablation experiment was carried out with the original YOLOv11n network as the baseline. The experimental results are shown in Table 1.

TABLE I
ABLATION EXPERIMENT

BiFPN	GLSA	CARAFE	mAP@0.5	GFLOPs	Params (M)
			0.779	6.3G	2.59
√			0.781	6.3G	1.92
	√		0.782	8.6G	3.73
		√	0.782	6.6G	2.72
√	√		0.789	6.7G	2.07
√	√	√	0.792	7.0G	2.19

From the experimental results in Table 1, it can be seen that after adding the BiFPN module, by constructing a bidirectional feature fusion path, the interaction between multi-scale semantic information and detail features is strengthened, and the perception of small target traffic signs is significantly improved. Without increasing the amount of calculation, the mAP@0.5 is increased to 0.781, and the model parameters are reduced from 2.59M to 1.92M, showing stronger feature utilization efficiency and lightweight structure. Secondly, after introducing the GLSA module, local attention is used to enhance the perception of key area details, and global attention is used to model contextual relationships, which significantly enhances the model's ability to distinguish targets under complex backgrounds. This module improves the detection accuracy to 0.782. After further integrating the CARAFE upsampling module into the network, the feature map is reconstructed through content-aware dynamic convolution kernels, which improves the semantic information loss problem caused by traditional interpolation. The mAP@0.5 also reaches 0.782, the computational cost is 6.6 GFLOPs, and the parameter volume is also controlled at 2.72M, showing a good balance between efficiency and performance.

When BiFPN and GLSA modules are combined at the same time, the detection accuracy of the model is improved to 0.789, which further verifies the complementary role of bidirectional feature fusion and attention mechanism in multi-scale object recognition. When BiFPN, GLSA and CARAFE are used together, the model mAP@0.5 is improved to 0.792, the calculation amount is controlled at 7.0 GFLOPs, and the parameter amount is 2.19M, which shows that the combination achieves better detection performance while maintaining low resource consumption, reflecting the effectiveness and practicality of the overall structural design.

In order to evaluate the model more comprehensively, this paper introduces the precision-recall curve to make a detailed comparison of the models. Compared with a single indicator, the PR curve can fully reflect the detection trade-off relationship of the model under different confidence thresholds, and is especially suitable for analyzing the dynamic changes between false alarm rate and missed detection rate in small targets and complex background scenes.

As shown in Figures 4 and 5, the overall distribution of the PR curve of BGC-YOLO is significantly better than the baseline YOLOv11 model. The curve is smoother and overall close to the upper right corner, indicating that it maintains a high precision and recall rate. This performance fully verifies the role of the module integration (BiPFN, GLSA and CARAFE) designed in this paper in improving the target feature expression and selection capabilities in complex scenes.

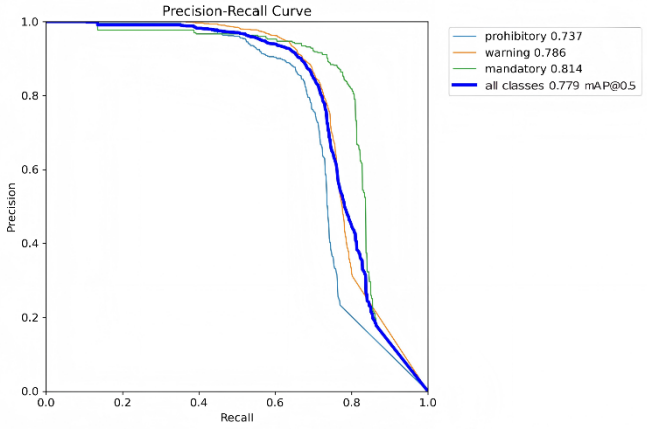


Fig. 4. YOLOV11 PR curve

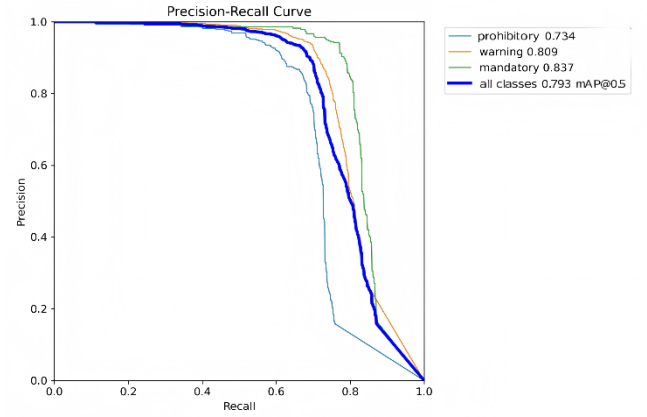


Fig. 5. BGC-YOLO PR curve

C. Comparative experiment

In order to comprehensively verify the effectiveness of the algorithm in this paper, this section conducts a comparative experiment with the current mainstream traffic sign detection algorithm. The selected comparison algorithms include SSD, Faster R-CNN, YOLOv3, YOLOv5n, YOLOv7, YOLOv8 and YOLOv10. The comparison results are shown in Table 2.

TABLE II
COMPARATIVE EXPERIMENT

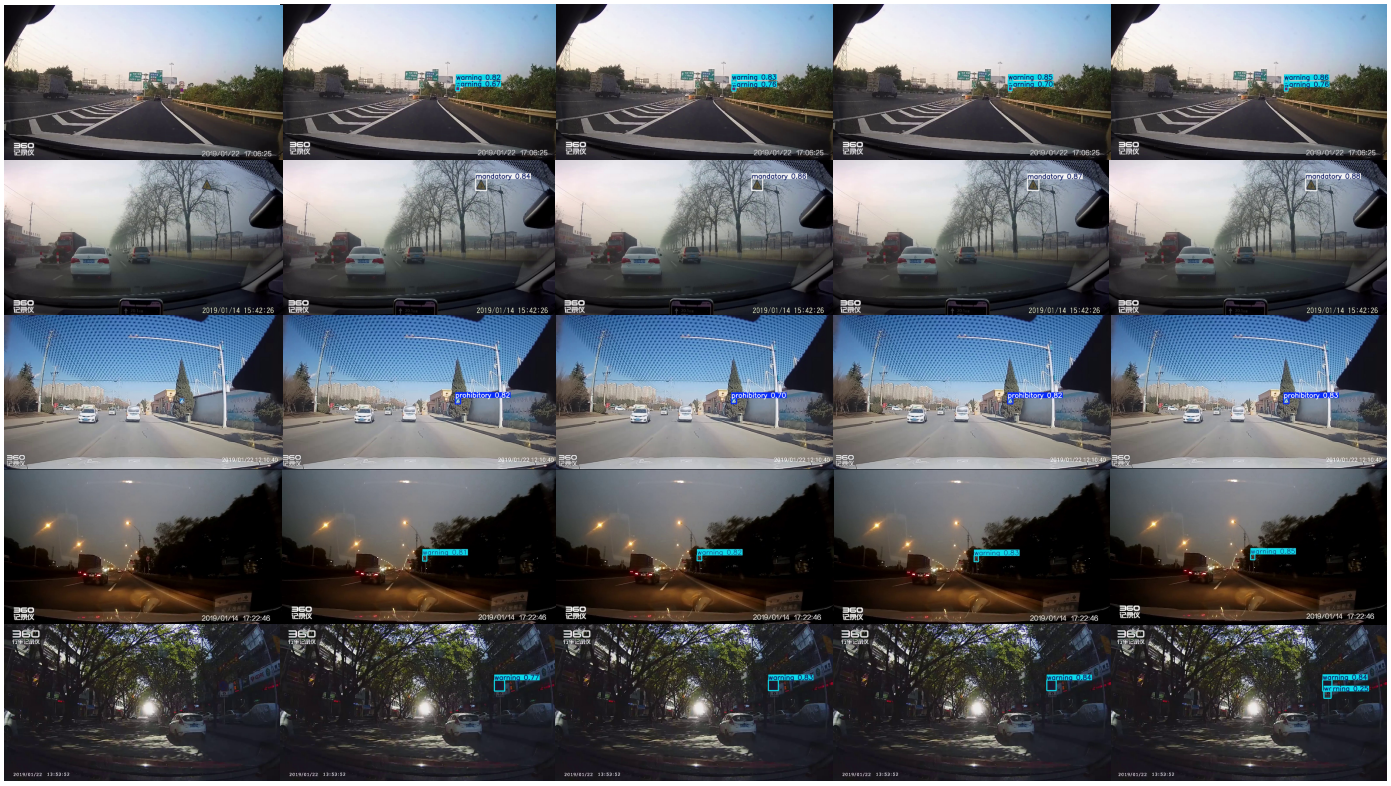
Model	P	R	mAP@0.5	GFLOPs	Params(M)
SSD	0.865	0.277	0.492	15.4G	25.0
Faster RCNN	0.848	0.550	0.566	92.2G	41.6
YOLOv3	0.846	0.427	0.505	5.1G	61.7
YOLOv5n	0.864	0.694	0.775	7.1G	2.5
YOLOv7-Tiny	0.865	0.684	0.764	13.2G	6.0
YOLOv8n	0.879	0.706	0.782	8.1G	3.0
YOLOv10n	0.871	0.713	0.791	6.5G	2.27
YOLOv11n	0.866	0.708	0.779	6.3G	2.59
YOLOv12	0.883	0.692	0.779	6.3G	2.56
CGS-Ghost YOLO[36]	0.824	0.614	0.68	16.6G	-
Hyper-YOLO[37]	0.875	0.702	0.78	9.5G	3.62
Ours	0.896	0.68	0.793	7.0G	2.19

From the comparison results in Table 2, it can be seen that the BGC-YOLO model proposed in this paper outperforms the existing mainstream target detection algorithms in multiple performance indicators. In terms of detection accuracy, BGC-YOLO reaches 0.793mAP@0.5, which is better than YOLOv5n, YOLOv8n, YOLOv10n, CGS-Ghost YOLO and Hyper-YOLO. At the same time, the accuracy (Precision) and recall (Recall) of the model are 0.896 and 0.680 respectively, and the overall detection performance shows stronger stability and reliability. In terms of model efficiency, the computational complexity of BGC-YOLO is controlled at 7.0 GFLOPs, and the number of parameters is 2.19M, showing good lightweight characteristics, which is suitable for traffic sign detection tasks with high requirements for real-time performance. Compared with the classic SSD and Faster R-CNN, BGC-YOLO has improved its accuracy by 30.1% and 22.7% respectively while significantly reducing the number of parameters and computation, demonstrating its obvious advantages in small target detection and adaptability to complex scenes. Compared with YOLOv5n, YOLOv8n and YOLOv7-Tiny, BGC-YOLO has improved its accuracy by 1.8%, 1.1% and 2.9% respectively while keeping the model size controllable. In particular, compared with the basic model YOLOv11n, although the computational complexity has only increased by 0.7G, the mAP has increased by 1.4%, showing the effective improvement brought by the improvement of the module structure. In summary, BGC-YOLO not only performs well in the mAP@0.5 indicator, but also maintains reasonable control in terms of model complexity, fully verifying its feasibility in actual traffic sign detection scenarios.

D. Visualization

In order to more intuitively demonstrate the difference in detection effect between the BGC-YOLO model and the YOLOv11 and YOLO12 models, Figure 6 gives the visualization results of different methods.

As can be seen from Figure 6, compared with the lower performance scores of YOLOv11, YOLOv12 and Hyper-YOLO, the BGC-YOLO model not only achieves accurate positioning and recognition of traffic signs, but also maintains a high detection accuracy. This comparison result confirms that BGC-YOLO has a detection advantage, and its positioning and recognition performance have been effectively improved.



(a) Original image (b) YOLOv11 (c) YOLOv12 (d) Hyper-YOLO (e) BGC-YOLO

Fig. 6. Visual detection effect comparison

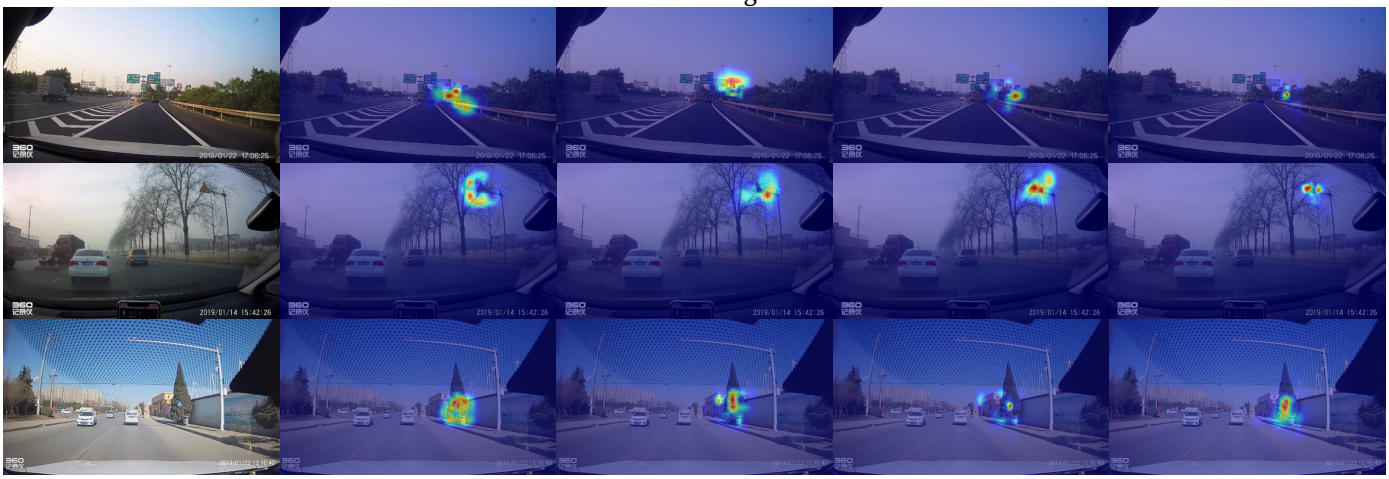
Figure 7 shows the heat map comparison results. It can be observed from the figure that in the high-resolution class activation area, BGC-YOLO responds more strongly to the target area and the brightness distribution is more concentrated, indicating that it is more sensitive in extracting target features. Especially in complex environments with more background interference, BGC-YOLO can more effectively suppress irrelevant information and improve the accuracy of target detection. This advantage makes BGC-YOLO more practical in

intelligent transportation systems, especially in tasks dealing with complex road scenes.

E. summary

The main innovation of this paper lies in its structural integration and collaborative design in the YOLOv11 architecture. Different from the previous research that introduced attention mechanism or feature enhancement module separately, BGC-YOLO emphasizes the systematic optimization of functional complementarity and coupling strategy between modules: BiFPN is used to enhance cross-scale semantic flow, GLSA strengthens the feature selection mechanism, and CARAFE improves the ability to restore details during upsampling. This three-in-one collaborative structural design helps to form a more stable detection performance in complex background and small target detection scenarios.

In order to verify the effectiveness of this integration strategy, this paper designs a series of ablation experiments and comparative experiments to evaluate the specific impact of each module on the model performance when introduced separately and in combination. The experimental results show that the complete BGC-YOLO model is superior to the model configuration that only introduces any one of the modules in multiple evaluation indicators, and shows a better balance between precision and recall in comparison with other YOLO improvement methods (such as CGS-Ghost YOLO) that adopt similar strategies, especially in the small target detection task.



(a) Original image (b) YOLOv11 (c) YOLOv12 (d) Hyper-YOLO (e) BGC-YOLO

Fig. 7. Heatmap comparison

V. Conclusion

In order to solve the problems of small target recognition difficulty, easy loss of feature information and complex background interference in traffic sign detection, a BGC-YOLO detection framework is proposed based on the YOLOv11 model. By introducing BiFPN to achieve efficient multi-scale feature fusion, the GLSA module is used to improve the model's perception of local details and global semantics, and the CARAFE upsampling operator is combined to enhance the quality of feature reconstruction, thereby significantly improving the detection accuracy while ensuring the computational efficiency of the model. Experiments have shown that BGC-YOLO performs better than multiple mainstream models, with mAP@0.5 reaching 0.793, while maintaining low computational complexity and parameter quantity, and has good real-time performance and deployment feasibility. In future work, we will focus on further optimizing the attention mechanism and feature fusion structure to enhance the robustness of the model in complex scenarios, including extreme weather conditions and occlusions. We are also committed to exploring lightweight model compression and acceleration strategies to enable real-time deployment on edge devices with limited computing resources.

References

[1] Bahlmann, Claus, et al. "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information." *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*. IEEE, 2005.

[2] Li, Haojie, et al. "A novel traffic sign detection method via color segmentation and robust shape matching." *Neurocomputing* 169 (2015): 77-88.

[3] Dolatyabi, Parya, Jacob Regan, and Mahdi Khodayar. "Deep Learning for Traffic Scene Understanding: A Review." *IEEE Access* (2025).

[4] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[5] Jiang, Peiyuan, et al. "A Review of Yolo algorithm developments." *Procedia computer science* 199 (2022): 1066-1073.

[6] Carion, Nicolas, et al. "End-to-end object detection with transformers." *European conference on computer vision*. Cham: Springer International Publishing, 2020.

[7] Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

[8] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks."

Advances in neural information processing systems 28 (2015).
[9] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.
[10] Cai, Zhaowei, and Nuno Vasconcelos. "Cascade r-cnn: Delving into high quality object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[11] Pang, Jiangmiao, et al. "Libra r-cnn: Towards balanced learning for object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
[12] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." *Advances in neural information processing systems* 29 (2016).
[13] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[14] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[15] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).
[16] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).
[17] Jocher, Glenn, and Ayush Chaurasia. "yolov5. github repository." *2020-06-09)[2021-07-09].[Article]* (2020).
[18] Li, Chuyi, et al. "YOLOv6: A single-stage object detection framework for industrial applications." *arXiv preprint arXiv:2209.02976* (2022).
[19] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.
[20] Sohan, Mupparaju, Thotakura Sai Ram, and Ch Venkata Rami Reddy. "A review on yolov8 and its advancements." *International Conference on Data Intelligence and Cognitive Informatics*. Springer, Singapore, 2024.
[21] Wang, Chien-Yao, I-Hau Yeh, and Hong-Yuan Mark Liao. "Yolov9: Learning what you want to learn using programmable gradient information." *European conference on computer vision*. Cham: Springer Nature Switzerland, 2024.
[22] Wang, Ao, et al. "Yolov10: Real-time end-to-end object detection." *Advances in Neural Information Processing Systems* 37 (2024): 107984-108011.
[23] Khanam, Rahima, and Muhammad Hussain. "Yolov11: An overview of the key architectural enhancements." *arXiv preprint arXiv:2410.17725* (2024).
[24] Tian, Yunjie, Qixiang Ye, and David Doermann. "Yolov12: Attention-centric real-time object detectors." *arXiv preprint arXiv:2502.12524* (2025).
[25] Lei, Mengqi, et al. "YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception." *arXiv preprint arXiv:2506.17733* (2025).
[26] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." *arXiv preprint arXiv:2010.04159* (2020).
[27] Meng, Depu, et al. "Conditional detr for fast training convergence." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
[28] Liu, Shilong, et al. "Dab-detr: Dynamic anchor boxes are better queries for detr." *arXiv preprint arXiv:2201.12329* (2022).
[29] Li, Feng, et al. "Dn-detr: Accelerate detr training by introducing query denoising." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
[30] Zhang, Hao, et al. "Dino: Detr with improved denoising anchor boxes for end-to-end object detection." *arXiv preprint arXiv:2203.03605* (2022).
[31] Jia, Ding, et al. "Detrs with hybrid matching." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.
[32] Chen, Jun, et al. "Effective feature fusion network in BIFPN for small object detection." *2021 IEEE international conference on image processing (ICIP)*. IEEE, 2021.
[33] Hu, Xudong, et al. "GLSANet: Global-local self-attention network for remote sensing image semantic segmentation." *IEEE Geoscience and Remote Sensing Letters* 20 (2023): 1-5.
[34] Wang, Jiaqi, et al. "Carafe: Content-aware reassembly of features." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
[35] Zhang, Jianming, et al. "CCTSDB 2021: a more comprehensive traffic sign detection benchmark." *Human-centric Computing and Information Sciences* 12 (2022).
[36] Zhao, H., and Y. B. Feng. "Research on traffic sign detection based on CGS-Ghost YOLO." *computer engineering* 49.12 (2023): 194-204.
[37] Feng, Yifan, et al. "Hyper-yolo: When visual object detection meets hypergraph computation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).