Proceedings of Machine Learning Research vol 278:1–9, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

An Improved YOLOv11 Algorithm For Traffic Sign Detection

Qianlong Chen 202430310263@STU.SHMTU.EDU.CN
School of information engineering Shanghai Maritime University Shanghai, China
Changming Zhu* CMZHU@SHMTU.EDU.CN
School of information engineering Shanghai Maritime University Shanghai, China
Hengbin Li 202430310137@STU.SHMTU.EDU.CN
School of information engineering Shanghai Maritime University Shanghai, China
**Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

As an important part of the Intelligent Transportation System (ITS), traffic sign recognition is of great significance for ensuring driving safety and improving traffic efficiency. This paper proposes a method that incorporates three modules, namely Adaptive Spatial Feature Fusion (ASFF), Spatial and Channel Synergistic Attention (SCSA), and Omni-Dimensional Dynamic Convolution (ODConv), into YOLOv11, aiming to enhance the performance of traffic sign detection. By enhancing the adaptability of feature scales through ASFF, optimizing feature extraction and fusion with SCSA, and strengthening the convolution operation with ODConv, the research results effectively improve the recognition accuracy and speed of various road signs on complex roads. Experimental results show that this integrated model outperforms the original YOLOv11 model and other comparative models in the traffic sign detection task, providing a more effective detection solution for the intelligent transportation field.

Keywords: Traffic Sign Detection; YOLOv11; SCSA; ODConv; ASFF

1. Introduction

Computer vision and deep learning technologies have developed so rapidly in the past few years that they have revolutionized intelligent transportation systems, especially in the field of traffic sign recognition (Yan et al., 2022). Traffic sign detection is an important component of modern intelligent transportation systems, aiming to enhance road safety and efficiency and improve overall traffic management. As road networks become increasingly complex and the number of vehicles steadily grows, traffic sign detection systems must be robust and flawless. Traffic signs are crucial for guiding and managing the flow of vehicles and for conveying key information to road users to ensure safety (Kaur et al., 2024).

Currently, convolutional neural networks used for object detection of traffic signs can be divided into single-stage and two-stage processes. The single-stage process is represented by the YOLO series of algorithms (Redmon et al., 2016; Redmon and Farhadi, 2017), while the two-stage process is represented by the R-CNN algorithm (Zhu et al., 2018). In traffic sign recognition, the YOLO series of algorithms are mainly employed, because the inference speed of the YOLO algorithm is faster than that of the two-stage algorithms (Luo et al., 2023).

On this basis, this paper integrates the SCSA (Si et al., 2025), ODConv (Li et al., 2022), and ASFF (Liu et al., 2019) modules into the YOLOv11 network. The aim is to enable YOLOv11 to make better use of the advantages of each module and improve its application ability in traffic sign recognition. Through experiments conducted on the traffic sign dataset, the effectiveness and superiority of the improved model have been verified, providing more powerful technical support for the development of intelligent transportation systems.

2. Related Method

2.1. YOLOv11

YOLOv11 is a new generation of real-time object detection algorithm launched by Ultralytics in 2024. Its core innovation lies in the adoption of the optimized C3k2 module and the novel C2PSA attention mechanism. As shown in Figure 1, C3k2 reconstructs the feature fusion structure through cross-layer connections and a lightweight attention mechanism, significantly enhancing the multi-scale feature representation ability. It improves the detection performance in scenarios involving small objects and occlusions. Meanwhile, it achieves higher accuracy and convergence efficiency at a relatively low computational cost. The C2PSA, on the other hand, conducts collaborative modeling of attention in both the channel and spatial dimensions. It strengthens the global dependency relationships and suppresses background interference. Its lightweight design takes into account the requirements of detection robustness and real-time performance in complex scenarios. This algorithm continues the three-stage architecture of Backbone-Neck-Head. Through modular innovation, it has made breakthrough progress in aspects such as feature extraction and information interaction. It has achieved good results in terms of detection accuracy, environmental adaptability, and deployment efficiency, and has become a technical benchmark with important academic significance and application prospects.

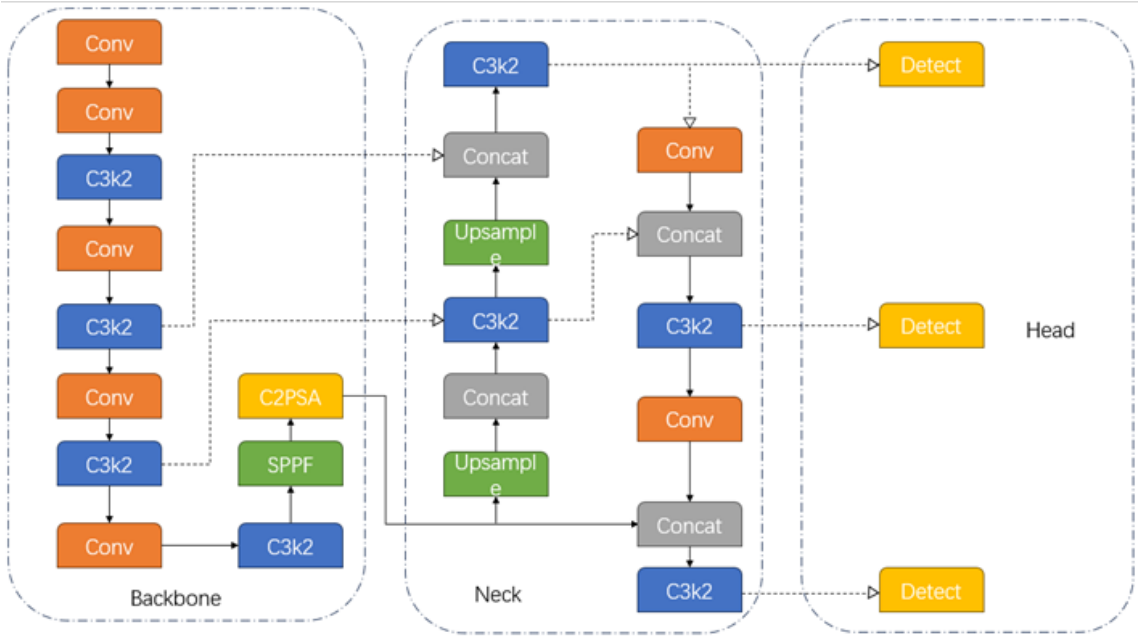

Figure 1: The YOLOv11 network structure.

2.2. SCSA

To improve the traffic sign detection performance of YOLOv11, this paper introduces the Spatial and Channel Cooperative Attention (SCSA) module, which makes use of the collaborative effect of spatial and channel attention across multiple semantic levels. The SCSA is composed of the Shared

Multi-semantic Spatial Attention (SMSA) and the Progressive Channel Self-Attention (PCSA). The SCSA establishes an attention mechanism simultaneously in both the spatial and channel dimensions, constructing a "dual-channel" collaborative attention pattern. In YOLOv11, this mechanism can assist the model in focusing more precisely on the traffic sign areas, suppressing background interference. Especially in complex traffic scenarios (such as when the signs are occluded or the background elements are similar), it significantly enhances the pertinence and effectiveness of feature extraction.

Shared Multi-semantic Spatial Attention (SMSA): As shown in the upper part of Figure 2, SMSA extracts spatial features through multi-scale depthwise separable 1D convolutions (convolution kernels with different heights and widths). In this way, it can extract local and overall information separately. To reduce the number of parameters and improve computational efficiency, the convolutional layers share weights, and Group Normalization (GN) is used to integrate the information from different branches to generate a spatial attention map. GN helps to maintain the independence of sub-features and avoids the batch dependence of Batch Normalization (BN) during small-batch training.

Progressive Channel Self-Attention (PCSA): As shown in the lower part of Figure 2, based on the spatially enhanced features provided by SMSA, PCSA reduces the computational complexity through progressive compression while retaining the key information. On this basis, it uses the channel self-attention mechanism to explicitly model the channel relationships, realizing the adaptive reweighting of channel features. In this way, it highlights the channel features most relevant to traffic sign detection, reduces the differences among various semantic information, and improves the accuracy of traffic sign recognition..

SMSA employs multi-scale, depth-shared 1D convolutions and Group Normalization (GN) to extract multi-semantic spatial information, so as to avoid semantic interference.

$$X = DWConvld\,(X) \tag{1}$$

This can optimize channel features and reduce semantic differences. PCSA further optimizes features using progressive compression and channel-wise single-head self-attention.

The complete SCSA module is represented as:

$$SCSA\,(X) = PCSA\,(SMSA\,(X)) \tag{2}$$

These formulas and detailed explanations are helpful for comprehensively describing the integration of the SCSA module in the YOLOv11 architecture and its working principle, thus improving the performance of traffic sign detection.

2.3. ODConv

Under the framework of YOLOv11, traditional convolution has limited adaptability to traffic signs with different shapes and postures. However, ODConv can dynamically adjust the parameters of the convolution kernel according to the local structure and semantic information of the input features. In this way, it enhances the model's ability to extract features of traffic signs in complex situations such as deformation and rotation. Therefore, we propose to replace some convolutional layers with Omni - Dimensional Dynamic Convolution (ODConv). Based on the multi - dimensional attention mechanism, ODConv studies the attention weights of convolutional kernels from four

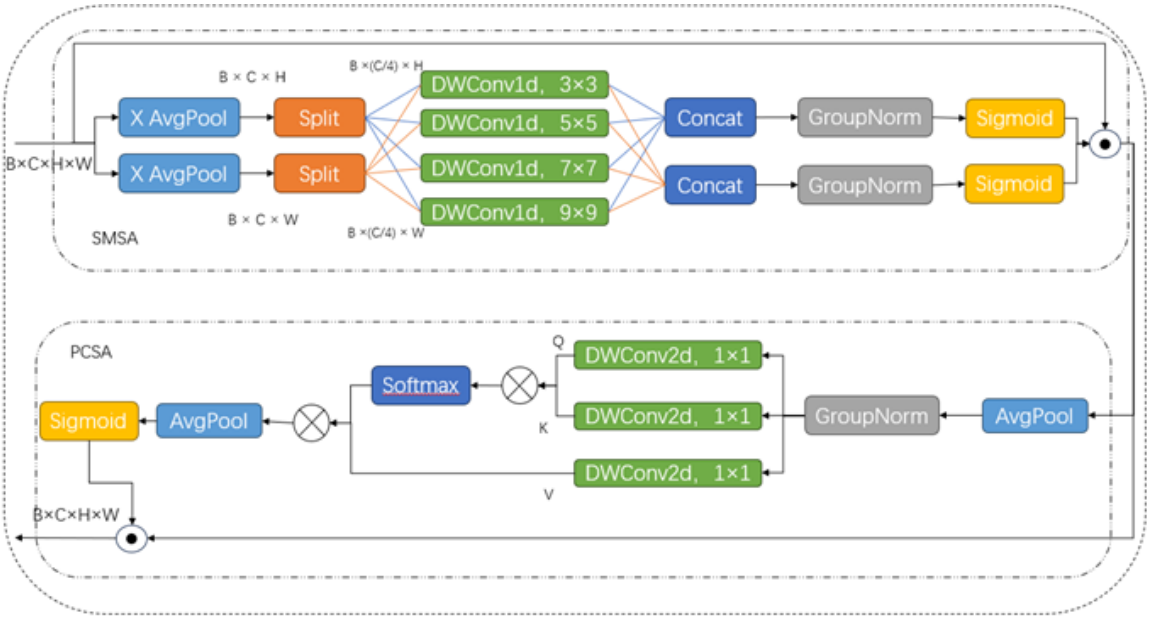

Figure 2: The Framework Diagram of SCSA.

aspects: space, input, output, and the number of kernels. The formula is as follows:

$$y = (\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + ... + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n) * x \tag{3}$$

Here, x and y are the input and output feature maps. W_i is the i-th conv kernel, with α_{wi}, α_{si}, α_{ci}, α_{fi} being attention params in different dims. This mechanism adjusts conv kernels based on input features, boosting distinguishability and capturing context. $\odot$ denotes multiplications in different dims. The algorithm tunes kernel dims dynamically, enhancing feature discriminability. ODConv captures rich context via diff conv operations on spatial pos, input channels, filtering, and kernel level.

As in Figure 3, GAP on input x in the channel direction yields a c_{in} length feature vector. It enters an FC layer and splits into four branches. Their outputs are $k \times k$, $c_{in} \times 1$, $c_{out} \times 1$ and $n \times 1$. Softmax or Sigmoid generates normalized attention params α_{s_i}, α_{c_i}, α_{f_i} and α_{w_i}.

2.4. ASFF

This experiment embeds an adaptive spatial feature fusion model into YOLOv11 to boost detection. Traffic sign scales vary widely, with small ones hard to detect. YOLOv11 balances speed and accuracy well, yet FPN may introduce errors in feature fusion, undermining detection accuracy.

This experiment adds the ASFF component (Figure 4). In traffic sign detection, for small-scale signs, the ASFF can enhance the weight of high-resolution features in the lower layers. For large-scale signs, it pays more attention to the fusion of high-level semantic features, thus improving the model's ability to detect objects of different scales. The ASFF algorithm conducts adaptive fusion at multiple feature levels, eliminating the problem of inconsistency. Firstly, the multi-layer feature maps output by the FPN are adjusted to the same spatial dimension through upsampling

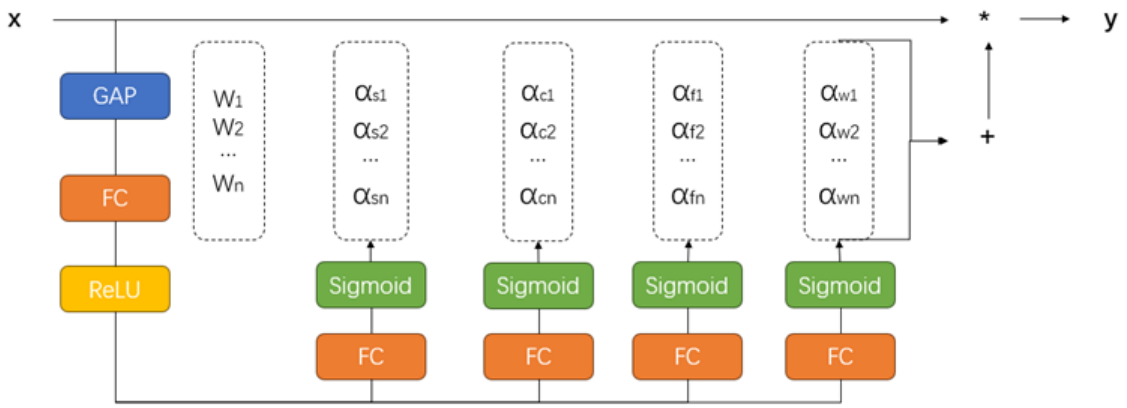

Figure 3: The Structure Diagram of ODConv.

(1x1 convolution + bilinear interpolation) and downsampling (stride=2, 3x3 convolution). Then, the weights learned by using the 1x1 convolution and the Softmax function are used to adaptively fuse the feature maps. The process is:

$$y_{ij}^{l} = \alpha_{ij}^{l} \cdot x_{ij}^{1\to l} + \beta_{ij}^{l} \cdot x_{ij}^{2\to l} + \gamma_{ij}^{l} \cdot x_{ij}^{3\to l} \tag{4}$$

Here, y_{ij}^{l} represents the fused feature vector at position (i, j) in level l; $x_{ij}^{n\to l}$ is the feature vector adjusted from level n to level l; α_{ij}^{l}, β_{ij}^{l}, and γ_{ij}^{l} are the fusion weights corresponding to position (i, j), calculated via the softmax function.

$$\alpha_{ij}^{l} = \frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}} + e^{\lambda_{\beta_{ij}}^{l}} + e^{\lambda_{\gamma_{ij}}^{l}}} \tag{5}$$

Here, $\lambda_{\alpha_{ij}}^{l}$, $\lambda_{\beta_{ij}}^{l}$ and $\lambda_{\gamma_{ij}}^{l}$ are 1x1 - conv - predicted weights. They control feature layer contributions. These spatially - varying weights let the network adaptively pick features, reducing inconsistency, thus boosting traffic sign detection accuracy, robustness and precision.

Finally, we obtained an improved detection algorithm based on YOLOv11. The structure diagram is shown in Figure 5.

3. Results and Analysis

3.1. Experimental Configuration

The operating system used in this experiment is Windows 11. The GPU used is an NVIDIA GeForce RTX 4060, and the CPU is an AMD R7 7735H. The system is equipped with 16GB of memory. In the software environment, PyTorch 2.0.0, torchvision 0.15.1, Python 3.11.4, and CUDA 11.8 (along with cudnn 8.9) are employed..

3.2. Dataset

The traffic sign dataset used in this experiment is selected from the universe dataset. This dataset encompasses various factors such as different lighting conditions (obvious road surface reflection

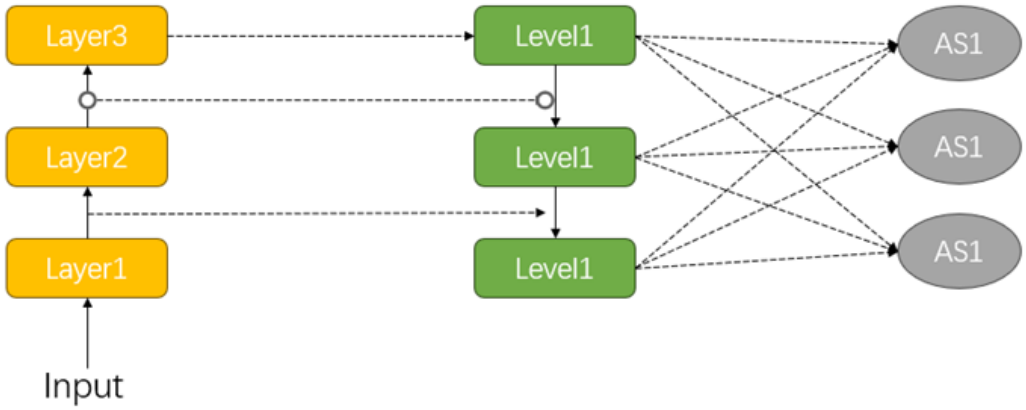

Figure 4: The Structure Diagram of ASFF.

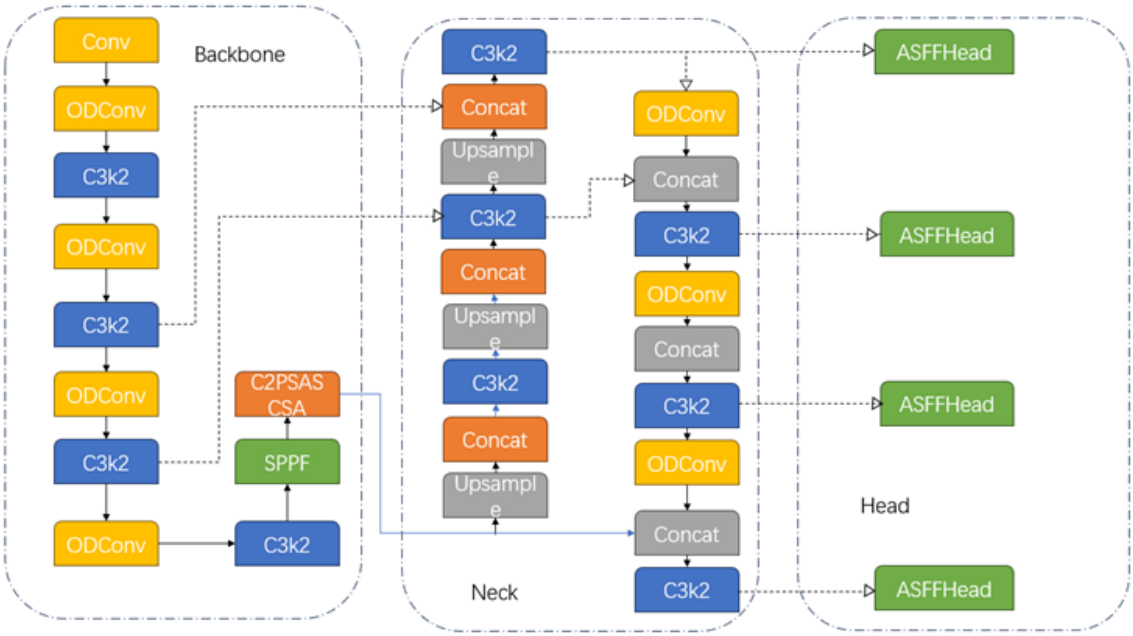

Figure 5: Network Framework Diagram of the Improved YOLOv11.

under strong light, low brightness of signs in low light, and severe shadows in some areas of the signs due to backlighting), weather conditions (clear vision on sunny days, signs possibly being obscured by raindrops or having reflection on rainy days, and blurred signs in foggy weather), and traffic flow (dense vehicles around the signs during high traffic flow, relatively moderate traffic during medium flow, and relatively empty scenes during low traffic flow). There are a total of 5,152 photos in this dataset, including 4,121 images in the training set, 515 photos in the test set, and 516

photos in the validation set. It contains 14 categories such as no parking, speed limit of 30, red light, etc.

3.3. Experimental and Results and Analysis

In the research of traffic sign detection algorithms, we introduced key metrics to comprehensively evaluate model performance, including mean average precision (mAP) and floating - point operations per second (FLOPs). Average precision (AP) measures the recognition ability of a category by quantifying the area under the precision - recall (P - R) curve. True positive (TP) represents correctly - identified traffic sign bounding boxes, false positive (FP, not TF as in your text) indicates cases wrongly classified as positive samples, and false negative (FN) refers to undetected real signs. AP is a metric for the recognition accuracy of a single class, while mAP is a comprehensive evaluation of the average accuracy across all classes, where n is the total number of traffic sign classes.

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \int_0^1 P\left(R\right) dR \tag{8}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{9}$$

The model parameters (Params) refer to the total number of parameters in the model during the training process, including weights, bias values, and so on. This indicator measures the spatial complexity and size of the model. The model's computational power (GFLOPs) focuses on the number of floating-point operations performed by the model during a single forward pass. Generally, it is quantified in terms of billions of floating-point operations per second. This indicator directly reflects the complexity and processing efficiency of the model at the computational level.

In the traffic sign detection model, the Intersection over Union (IoU) measures the overlap between the predicted bounding box and the actual bounding box, and the IoU threshold is used to determine whether the prediction result is correct. When the IoU reaches the given threshold, our prediction is considered correct. The model is evaluated using mAP50 (IoU = 0.5) and mAP50-95 (IoU ranging from 0.5 to 0.95). During the model training process, we set the image size to 640×640 and apply the gradient descent optimization algorithm. To alleviate the problem of overfitting, we dynamically adjust the learning rate and the weight decay coefficient. Through continuous verification, we finally determine that the initial learning rate is set to 0.0001, the weight decay coefficient is 0.0005, the batch size is 8, and we complete 200 training epochs.

We also compared different object detection algorithms. As shown in Table 1, although in terms of precision and the number of parameters, YOLOv11 is inferior to YOLOv5, in terms of recall rate, mAP50, mAP50-95, and the number of floating-point operations per second (FLOPs), YOLOv11 outperforms YOLOv5. When compared with YOLOv8, YOLOv11 surpasses YOLOv8 in every aspect. Therefore, we finally selected YOLOv11 as the base model.

We conducted a series of ablation experiments to evaluate the contributions of each module. As shown in Table 2, The SCSA has increased the precision by 0.026, the mean average precision

Table 1: Comparative experiments of the models.

	Precision	Recall	mAP50	mAP50-95	Parameters	GFLOPs
YOLOv5	0.819	0.719	0.802	0.473	2505674	7.1
YOLOv8	0.791	0.729	0.795	0.469	3008378	8.1
YOLOv11	0.781	0.741	0.805	0.477	2584882	6.3

(mAP50) when the Intersection over Union (IoU) is 0.5 by 0.010, and the mean average precision (mAP50-95) within the IoU range from 0.5 to 0.95 by 0.001. The ODConv has enhanced the precision by 0.028, the mAP50 by 0.002, and the mAP50-95 by 0.001. The ASSF has boosted the precision by 0.003, the mAP50 by 0.016, and the mAP50-95 by 0.013. Compared with the original model, the performance of our improved YOLOv11 model has been significantly enhanced. Specifically, the precision of this model reaches 0.825, which is an increase of 0.044 compared with the original version; the recall rate is 0.752, representing an improvement of 0.011. In addition, the mAP50 of this model when the IoU is 0.5 and the mAP50-95 within the IoU range from 0.5 to 0.95 are 0.834 and 0.497 respectively. Although the number of parameters of YOLOv11 has increased and the floating-point calculations have gone up, the improved YOLOv11 performs better in terms of precision, recall rate, mAP50, and mAP50-95, and its detection ability is more accurate. Generally speaking, the improved YOLOv11 has witnessed various improvements compared with the original model.

Table 2: Ablation Experiment

	Precision	Recall	mAP50	mAP50-95	Parameters	GFLOPs
YOLOv11	0.781	0.741	0.805	0.477	2584882	6.3
SCSA	0.807	0.725	0.815	0.478	2537010	6.3
ODConv	0.809	0.718	0.807	0.478	3574017	7.2
ASFF	0.784	0.78	0.821	0.490	4054408	13.3
SCSA+ODConv+ASFF	0.825	0.752	0.834	0.497	6070179	11.5

4. Conclusion

In this study, we propose an enhanced YOLOv11 model specifically designed for traffic sign detection, which integrates three advanced modules: SCSA (Synergistic Channel and Spatial Attention), ODConv (Omni-Dimensional Dynamic Convolution), and ASFF (Adaptively Spatial Feature Fusion). By embedding these modules into the backbone and neck of the YOLOv11 architecture, the model's capacity for feature representation has been significantly improved. The SCSA module strengthens attention across both spatial and channel dimensions, enhancing the model's focus on relevant features. The ODConv module introduces dynamic and flexible convolution operations across multiple dimensions, boosting the network's adaptability to complex patterns. The ASFF module facilitates more effective multi-scale feature fusion, leading to better localization and recognition accuracy. Comprehensive experiments conducted on a benchmark traffic sign dataset demonstrate the superior performance of the proposed model. Compared with the original YOLOv11, the improved version achieves higher detection accuracy while maintaining relatively fast inference speed, striking a good balance between precision and efficiency. Furthermore, a series of abla-

tion studies confirm the individual and combined contributions of the SCSA, ODConv, and ASFF modules, highlighting their effectiveness in enhancing the model's robustness and generalization capabilities. These findings provide valuable insights for the further optimization and design of object detection frameworks in real-world applications.

Acknowledgments

This work is supported by National Natural Science Foundation of China (CN) [62276164, 61602296], 'Science and technology innovation action plan' Natural Science Foundation of Shanghai [22ZR1427000], and Shanghai Oriental Talent Program-Youth Program. The authors would like to thank their supports.

References

A. Kaur, V. Kukreja, N. Thapliyal, M. Aeri, R. Sharma, and S. Hariharan. An improved yolov8 model for traffic sign detection and classification. In *2024 3rd International Conference for Innovation in Technology (INOCON)*, pages 1–5, 2024. doi: 10.1109/INOCON60754.2024.10511576.

C. Li, A. Zhou, and A. Yao. Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947*, 2022. doi: 10.48550/arXiv.2209.07947.

S. Liu, D. Huang, and Y. Wang. Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*, 2019. doi: 10.48550/arXiv.1911.09516.

S. Luo, C. Wu, and L. Li. Detection and recognition of obscured traffic signs during vehicle movement. *IEEE Access*, 11:122516–122525, 2023. doi: 10.1109/ACCESS.2023.3329068.

J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. doi: 10.1109/CVPR.2017.690.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.

Y. Si, H. Xu, X. Zhu, et al. Scsa: Exploring the synergistic effects between spatial and channel attention. *Neurocomputing*, 634, 2025. doi: 10.1016/j.neucom.2025.129866.

W. Yan, G. Yang, W. Zhang, et al. Traffic sign recognition using yolov4. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, page 0, 2022. doi: 10.1109/ICSP54964.2022.9778657.

Y. Zhu, M. Liao, M. Yang, and W. Liu. Cascaded segmentation-detection networks for text-based traffic sign detection. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):209–219, 2018. doi: 10.1109/TITS.2017.2768827.