scientific reports

OPEN

# A lightweight network for traffic sign detection via multiple scale context awareness and semantic information guidance

Chenjie Du[1,2], Siyu Su[2], Chenwei Lin[2], Yingbiao Yao[2], Ran Jin[1]✉ & Xinhua Hong[1]

**Traffic sign detection, as a critical branch of object detection, plays an essential role in both assisted driving and autonomous driving technologies. In this paper, we propose MASG-Net, a lightweight detection network designed to improve the accuracy and efficiency of traffic sign detection. First, we introduce a channel attention mechanism into MobileNetV3 to create a novel E-block structure and design E-mobilenet, a lightweight backbone network, to replace the backbone in YOLOv4-tiny, significantly enhancing feature extraction while reducing parameters. Second, we propose a multi-scale dilated convolution spatial pyramid pooling (MDSPP) module to expand the receptive field of feature maps, enabling the network to capture multi-scale contextual information effectively. Finally, a semantic information guidance (SIG) module is introduced to leverage deep semantic information to guide shallow feature layers, improving the detection of small traffic signs and enhancing robustness against cluttered backgrounds. Experimental results on the CCTSDB, GTSDB and TT100K datasets demonstrate that MASG-Net achieves superior detection performance, particularly for small and challenging traffic signs, while maintaining high efficiency with an inference speed of 203.6 FPS. These results highlight MASG-Net's potential for real-time traffic sign detection in practical applications.**

**Keywords** Traffic sign detection, Lightweight, E-block structure, Multi-scale dilated convolution, Semantic information guidance

With the rapid development of smart cities, safe and reliable intelligent transportation systems have become an urgent demand. Intelligent transportation systems can provide real-time traffic information and signage to help drivers avoid accidents and dangerous situations. In addition, the development of autonomous driving technology will also reduce accidents caused by human driving in the future. As a branch of object detection, traffic sign detection is an indispensable part of automatic driving technology in intelligent transportation systems[1–4]. It has great practical value for ensuring safe vehicle driving, alleviating traffic congestion, and building smart cities[5–7]. Despite significant progress in object detection techniques, traffic sign detection remains a challenging task due to several factors. First, traffic signs are often small in size and may appear blurry or dim, particularly in low-light conditions or adverse weather. This makes it difficult for detection models to extract sufficient features for accurate recognition. Second, traffic signs are frequently surrounded by cluttered backgrounds, such as trees, buildings, or other road elements, which can confuse detection models and lead to false positives. Third, achieving a balance between high detection accuracy and computational efficiency is a persistent challenge, especially for real-time applications in resource-constrained environments, such as embedded systems in vehicles.

Nowadays, traffic sign detection[8] are mainly divided into one-stage and two-stage algorithms. The principles of the two algorithms are different. The two-stage detector generally classifies the candidate regions, whereas the one-stage detector uses a regression method, which can directly give the detection results for the input image. R-CNN[9–11] series are the classic representative two-stage detectors and have achieved very good detection results. However, traffic sign recognition requires the network to have high detection accuracy and fast detection speed[12–15]. The detection speed of this series of algorithms is slow and is not suitable for detecting traffic signs. The one-stage algorithms represented by the SSD[16] and YOLO series[17–23] have a faster detection speed and lower complexity of the network model, allowing real-time target detection with a higher detection speed and better accuracy[24,25]. Compared with the R-CNN series algorithm, the YOLO series algorithm can detect objects more

[1]College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, China. [2]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China. ✉email: ran.jin@163.com

efficiently in the way of single-stage detection. However, in some complex scenes or small target detection, the YOLO series algorithm may have a certain performance loss[26–28].

Compared to high-precision detectors, YOLOv4-tiny may have slightly lower detection accuracy, but it performs well in terms of training efficiency and detection speed and is suitable for applications requiring real-time target detection, such as real-time video analysis, traffic monitoring, face recognition, etc[29–32]. It can perform object detection in images or videos in a relatively short period of time, providing immediate feedback. Thus, this paper conducts in-depth research on traffic sign detection using YOLOv4-tiny and finds three problems as follows: (1) In YOLOv4-tiny, the backbone network struggles to automatically prioritize important features and suppress irrelevant ones, leading to a gradual decline in the model's discrimination ability when interference persists. (2) YOLOv4-tiny mainly uses single-branch ordinary convolution with a fixed kernel size, leading to a uniform receptive field. This restricts the extracted features, making complex detection tasks difficult due to the absence of multi-scale capabilities. (3) When using YOLOv4-tiny to identify traffic signs, the accuracy is hindered by the relatively small size of the signs, low resolution, unclear features, and other objective factors. This often results in missed detections and false positives, reducing the effectiveness of small target recognition.

Addressing the aforementioned issues, we introduce MASG-Net, an end-to-end lightweight detection approach, grounded in multi-scale awareness and semantic guidance. First, we introduce an ultra-lightweight channel attention mechanism into MobileNetV3 to create a novel E-block structure. Based on this structure, we design E-mobilenet, a lightweight backbone network that significantly improves feature extraction while reducing the number of parameters, making it suitable for real-time applications. To address the limitations of small feature maps in capturing sufficient information for small targets, we propose the multi-scale dilated convolution spatial pyramid pooling (MDSPP) module. This module expands the receptive field of the feature map, enabling the network to capture global and local context information more effectively. Further, we introduce the semantic information guidance (SIG) module, which leverages deep semantic information to guide the shallow feature layer. This design enhances the distinction between traffic signs and their backgrounds, reducing the negative impact of cluttered environments and improving detection performance for small and blurry signs. The ablation study indicates that the integrated application of E-mobilenet, MDSPP and SIG tends to outperform their independent usage. In contrast to many mainstream traffic sign detection algorithms, the main innovations of this paper are detailed as follows:

(1) A new backbone feature extraction network, E-mobilenet, is designed by enhancing MobileNetV3's lightweight cell structure with a channel attention mechanism. This backbone replaces YOLOv4-tiny's backbone, improving feature extraction efficiency while maintaining a lightweight design.
(2) The proposed MDSPP module incorporates multi-scale dilated convolutions to provide rich multi-scale receptive field information. This design addresses the problem of information loss caused by large-scale pooling operations, enhancing the network's ability to capture global context.
(3) The introduction of the SIG module enhances the detection of small traffic signs by leveraging deep semantic information to guide the shallow feature layer. This module improves the model's resistance to cluttered backgrounds and preserves critical semantic information for small target detection.

The overall structure of this paper is as follows. We first introduce the research work related to this experiment, and then detail the innovations in this paper in the MASG-Net section. The "Experiments" section provides comparative experiments, as well as qualitative and quantitative analysis of test results. Finally, the work of this paper is summarized.

## Related work

### Two-stage detectors

The two-stage detectors are to generate target candidate boxes by a regional proposal network (RPN), and then classify and regression these candidate boxes to get the final detection results.

In 2014, Girschick proposed the R-CNN, which surpassed YannLecun's contemporaneous end-to -end OverFeat[33] in terms of performance. In 2015, SPP-Net[34] added a spatial pyramid pool structure[35] between the convolutional layer and the fully connected layer, which not only ensured performance, but also greatly improved detection speed. In 2016, Fast R-CNN algorithm is proposed. The algorithm scales each feature matrix by ROI-Pooling[36] layer to a 7×7 feature map, and then flattens the feature map through a series of fully connected layers to get the prediction result. In addition, Kaiming He and Girshick of Microsoft Research proposed Faster R-CNN algorithm and proposed RPN, which can share the feature information extracted by a convolutional neural network throughout the network process, saving computing costs and solving the problem of slow generation of positive and negative sample candidate frames by Fast R-CNN algorithm[37] The Mask R-CNN[38] algorithm added a branch fully convolutional network (FCN)[39] layer on the basis of border recognition for semantic mask recognition.

The main difference between the two-stage detectors lies in the specific structure and optimization mode of RPN and the target classification regression network[40,41]. The two-stage detector usually has high detection accuracy, but the detection speed is relatively slow[42]. Thus, it is suitable for scenarios that require high detection accuracy, such as medical image analysis and security checks.

### One-stage detectors

The one-stage detector extracts the advanced features of the image through the convolutional network, and then fuses the feature map to complete the object detection and classification[43]. Currently, one-stage detectors mainly include the YOLO series, SSD, RefineDet[44], etc.

The SSD algorithm used the weighted sum of data enhancement, positioning, and confidence losses to train the model, which was faster, but the training was difficult, resulting in low algorithm accuracy. Shifeng Zhang et al. proposed the RefineDet detection method, which uses two-stage regression to improve detection accuracy and realized end-to-end multi-task training. The YOLO series includes multiple one-stage detection algorithms. YOLOv1[17] is the first algorithm in the YOLO series, and YOLOv4[20] combined various performance enhancing modules, making it one of the models with better detection performance in the YOLO series. YOLOX[21] conducted classification and regression separately, which increased the complexity of the model. YOLOv6[22] introduced the new frame regression loss function of SIoU[45] to improve the training speed and regression accuracy. The YOLOv7[23] network adopted a feature pyramid network and an improved backbone network to achieve more accurate and faster target detection, but with higher requirements for device performance.

YOLO-NAS[46] is the latest algorithm in the YOLO series, which employs the neural architecture search to achieve a balance between accuracy and computational complexity. However, YOLO-NAS is still in the research stage and has not been widely applied and verified. The YOLO tiny series are the lightweight versions of YOLO, featuring a smaller model size and faster detection speed, suitable for real-time detection on mobile terminals and other scenarios with limited computing resources[47–50]. YOLO tiny series includes YOLOv3-tiny[51], YOLOv4-tiny[52], and YOLOv7-tiny[23] three versions. Although YOLOv3-tiny has a very fast inference speed, the detection accuracy is relatively low. YOLOv7-tiny is the latest YOLO-tiny series and introduces several improvements over YOLOv7. These changes optimize its detection speed and model size, but may also result in a slight loss in detection accuracy. For example, when the intersection over union is larger, the detection accuracy of the YOLOv7-tiny network is lower. Therefore, among these three models, YOLOv4-tiny is the most mature lightweight model, which has the advantages of small model size and high detection accuracy.

## Attention mechanism

The attention mechanism is used to simulate human visual attention. In a deep learning model, it can automatically learn to assign different attention weights to different parts of the input, thereby improving the model's ability to understand and express the input[53–55]. SENet[56] pays attention to the information on the channel using adaptive weights, in which only a relatively small full-connection layer is introduced, so the number of parameters is relatively small. CBAM[57] is improved and proposed to obtain useful information from both space and channel. However, the introduction of CBAM module will increase the complexity of the network, resulting in increased computing and memory requirements in the training process, and thus increasing the time cost of training. The coordinate attention (CA) mechanism[58] is applicable to scenes with spatial dimensions. ECANet[59] proposed by the author of this paper is a relatively efficient channel attention mechanism, which is suitable for models with high detection efficiency requirements. For scenes with larger feature map sizes, ECANet can consider both the channel dimension and the spatial dimension of attention, which is more efficient.

Shuffle attention[60] that combines channel attention and spatial attention. It improves the feature representation capability and network performance by grouping, calculating, and applying attention to the input feature map. Efficient local attention (ELA)[61] is a lightweight attention mechanism using 1D convolution and group-normalized feature enhancement. The essence of scaled dot-product attention[62] is to quantify the similarity between the query and the key through the dot product, then assign attention weights through softmax, and weighted sum the value vectors according to these weights to form a context-sensitive representation of each position in the input sequence.

In this article, we introduce the ECANet structure in the MobilenetV3 cell structure to form a new backbone network of the E-mobilenet. Unlike the SE block, our module uses a lightweight design that minimizes computational overhead, making it more suitable for real-time applications. Compared to the ECA module, which focuses on local channel interactions, our module incorporates a broader context to enhance feature extraction for small and blurry traffic signs.

# Methodology

## Overall structure

The network structure of the proposed MASG-Net is shown in Fig. 1, and its improvements mainly include three points. Firstly, we propose a new backbone feature extraction network E-mobilenet, which is based on the MobileNetV3[63] and ECANet. Secondly, we proposed a new multi-scale dilated convolution spatial pyramid pooling structure. Finally, we introduce a semantic information guidance (SIG) module to enhance tiny sign detection by leveraging deep semantic information to guide the shallow feature layer.

We find that the parameters of the backbone network CSPdarknet53_tiny account for the majority of the parameters of YOLOv4-tiny. Therefore, in order to reduce the model size of YOLOv4-tiny, it is necessary to reduce the number of parameters of its backbone network CSPdarknet53_tiny. As we know, MobileNetV3 is an ultra-lightweight cnn model for mobile devices and has a small model size. Thus, we first replace the CSPdarknet53_tiny of YOLOv4-tiny with MobileNetV3.

Then, we integrate the ECANet attention mechanism into the MobileNetV3 model. ECANet overcomes the contradiction between performance and complexity to learn effective channel attention in a more efficient way by employing local cross-channel interactions that significantly reduce the complexity of the network model while maintaining performance. To improve the detection accuracy of the model for small targets, we propose to add the MDSPP module after the E-mobilenet. Because the deep feature output by the backbone network contains limited information for small targets due to the small size of the feature map, the receptive field size can be effectively enhanced to obtain more global information after the MDSPP, enabling the network to extract more abundant features. Taking a step further, we propose the SIG module, which enhances the semantic information of the shallow feature layer, improving the distinction between the target and the background and reducing the negative impact of complex backgrounds on detection performance. This design also significantly

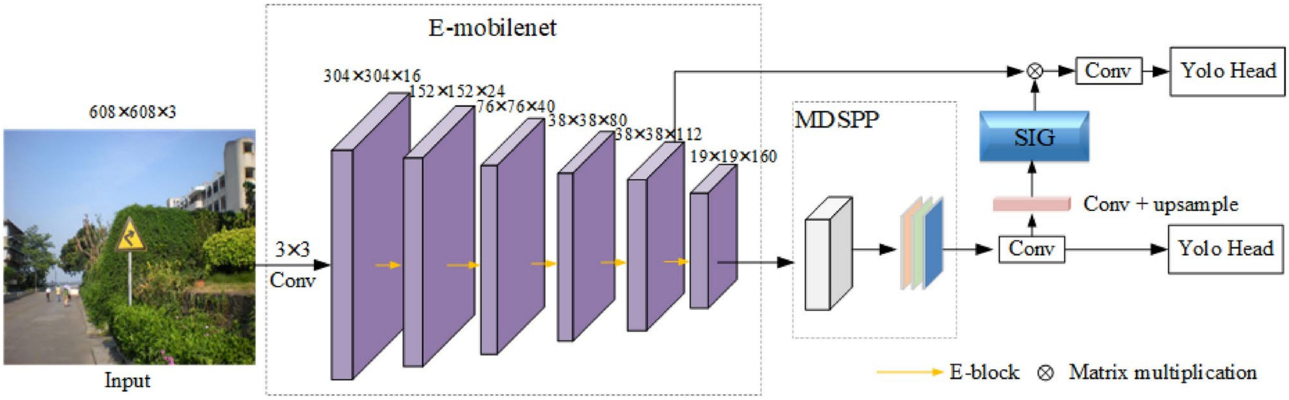


**Fig. 1**. Network architecture of the MASG-Net.

| Input | Operator(size) | *t* | out | ECANet | NL | *s* |
|---|---|---|---|---|---|---|
| 608×608×3 | Conv2d, 3 × 3 | – | 16 | – | HS | 2 |
| 304×304×16 | E- block, 3 × 3 | 1 | 16 | – | RE | 1 |
| 304×304×16 | E- block, 3 × 3 | 4 | 24 | – | RE | 2 |
| 152×152×24 | E- block, 3 × 3 | 3 | 24 | – | RE | 1 |
| 152×152×24 | E- block, 5 × 5 | 3 | 40 | ✓ | RE | 2 |
| 76×76×40 | E- block, 5 × 5 | 3 | 40 | ✓ | RE | 1 |
| 76×76×40 | E- block, 5 × 5 | 3 | 40 | ✓ | RE | 1 |
| 76×76×40 | E- block, 3 × 3 | 6 | 80 | – | HS | 2 |
| 38×38×80 | E- block, 3 × 3 | 2.5 | 80 | – | HS | 1 |
| 38×38×80 | E- block, 3 × 3 | 2.3 | 80 | – | HS | 1 |
| 38×38×80 | E- block, 3 × 3 | 2.3 | 80 | – | HS | 1 |
| 38×38×80 | E- block, 3 × 3 | 6 | 112 | ✓ | HS | 1 |
| 38×38×112 | E- block, 3 × 3 | 6 | 112 | ✓ | HS | 1 |
| 38×38×112 | E- block, 5 × 5 | 6 | 160 | ✓ | HS | 2 |
| 19×19×160 | E- block, 5 × 5 | 6 | 160 | ✓ | HS | 1 |
| 19×19×160 | E- block, 5 × 5 | 6 | 160 | ✓ | HS | 1 |

**Table 1**. E-mobilenet network structure.

retains important semantic information in small traffic sign targets. MASG-Net is very suitable for deployment on resource limited vehicle terminal devices for traffic sign recognition due to its high detection accuracy and real-time performance.

## E-mobilenet structure

In order to carry out feature extraction of input images more efficiently on the premise of ensuring low complexity of the model, we design an ultra-lightweight backbone network structure E-mobilenet as shown in Table 1.

The third column represents the proportion of the number of channels in the E-block that are up-dimensioned and then down-dimensioned in the inverse residual structure. The fourth column represents the number of channels in the feature layer output after the second column of operations. The sixth column NL represents the type of non-linear activation function, HS and RE are the h-swish and RELU6 activation function, respectively. H-swish function has the characteristics of no upper bound, lower bound, smooth and non-monotonic. The seventh column, parameter *s*, represents the step size used for each convolution or E-block structure. Moreover, the definitions of ReLU6 and h-swish activation function are as follows:

$$ReLU6\left(x\right) = min\left(max\left(x, 0\right), 6\right) \tag{1}$$

$$h - swish\left(x\right) = x \cdot \frac{ReLU6\left(x+3\right)}{6} \tag{2}$$

The E-block, which is shown in Fig. 2, adopts a backward residual structure with a linear bottleneck and includes three convolution layers: 1 × 1 convolution to reduce the dimension, 3 × 3 convolution to extract features, and 1 × 1 convolution to restore the dimension. Moreover, ECANet is integrated into the E-block to improve its performance.

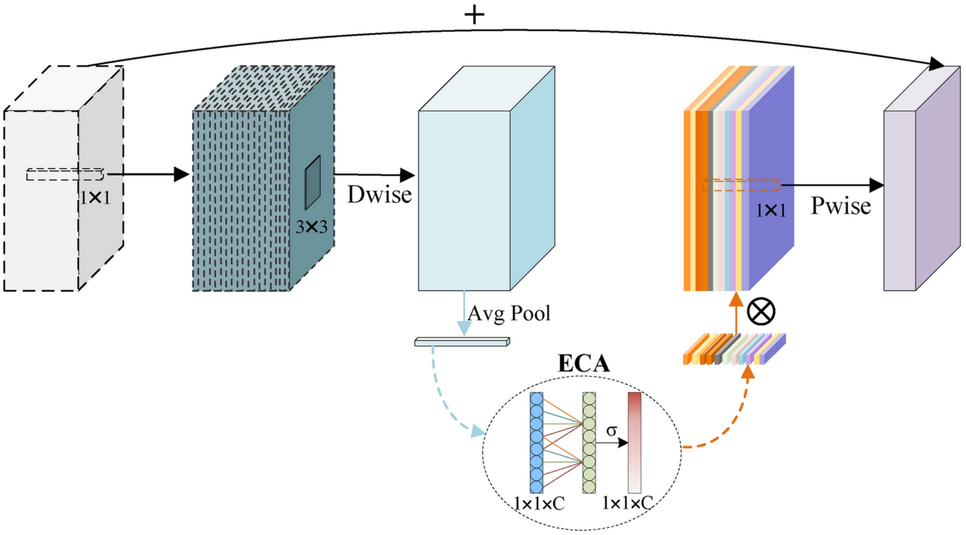


**Fig. 2**. Improved cell structure E-block.

The attention mechanism in the original MobilenetV3 is implemented in the same way as SENet, which employs two fully connected layers to capture nonlinear cross-channel interactions. However, this mechanism has two defects: it cannot capture the attention in the spatial dimension and two fully connected layers will increase the number of network parameters. Therefore, we introduced ECANet into the MobilenetV3 cell structure to form a new ultra-lightweight cell structure E-block. ECANet cancels the two fully connected layers and the feature extraction is carried out directly through a one-dimensional convolution to obtain the weight of each dimension. This way can make the weight learning process more simple and direct. The convolution kernel size of this one-dimensional convolution is obtained by adaptive calculation and represents the coverage of local cross-channel interactions. The weight sharing means that each set of convolution uses exactly the same weight, which greatly reduces the number of parameters. Specifically, the number of parameters is reduced from the original SENet's $2C^2/r$ to *k*, where *C* is the number of channels, *r* is the dimensionality reduction hyperparameter and *k* is the convolution kernel size. In addition, given the channel dimension *C*, *k* can be adaptively determined as:

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{3}$$

where *odd* indicates that the value is odd, $\gamma$ is set to 2, and *b* is set to 1.

## Multi-scale dilated convolution spatial pyramid pooling

The backbone network of YOLOv4-tiny primarily relies on single-branch ordinary convolution with a fixed kernel size, leading to a deterministic and uniform receptive field. This limitation results in extracted features with limited information, making it challenging to handle complex detection tasks due to the lack of multi-scale capabilities.

Based on the SPP structure, we proposed a multi-scale dilated convolution spatial pyramid pooling (MDSPP) structure. The SPP structure is essentially a multi-scale pooling, which extracts multi-scale pooling information for the same feature layer. Since the input feature layer is relatively fixed with respect to the original image receptive field, the enhancement of the receptive field by the structure is not obvious after the fusion of multi-scale pooling information. In order to better enrich the receptive field scale and improve the feature extraction capability of the network, dilated convolutions with different hole rates are introduced in each branch of the SPP structure. This structure was designed to increase the receptive field of the feature map, thus helping the network capture context information at more scales. The specific structure is shown in Fig. 3.

## Semantic information guidance

MDSPP first divides the input feature layer into three main branches for three different scales of the dilated convolution, each with a convolution kernel of size 3 × 3, but with dilation rates of 1, 3 and 5, respectively. By using dilated convolution, the MDSPP module avoids the need for multiple large kernels, which would increase the number of trainable parameters. This design aligns with the ultra-lightweight nature of MASG-Net, ensuring that the model remains compact and suitable for real-time applications in resource-constrained environments. Then the extracted features with different receptive fields are passed through the feature pyramid pooling layer, where the pooling pyramid is also divided into three branches and is articulated after the dilated convolution, with the maximum pooling window size of 5, 9 and 13, respectively. The output of these three branches is then connected to obtain the final output of the structure. Based on the pyramid pool structure, MDSPP forms a

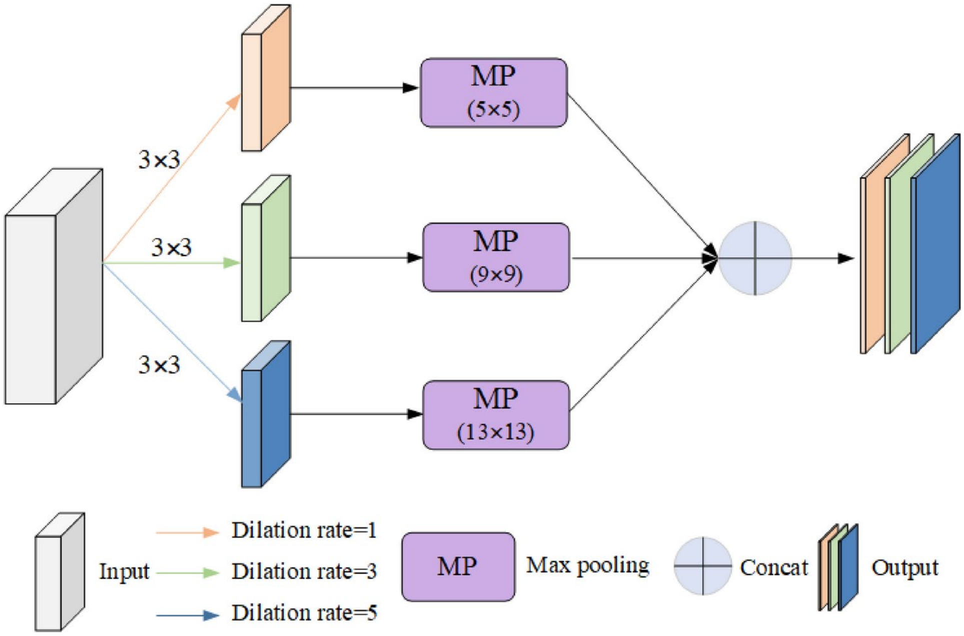


**Fig. 3**. The structure diagram of the proposed MDSPP module.

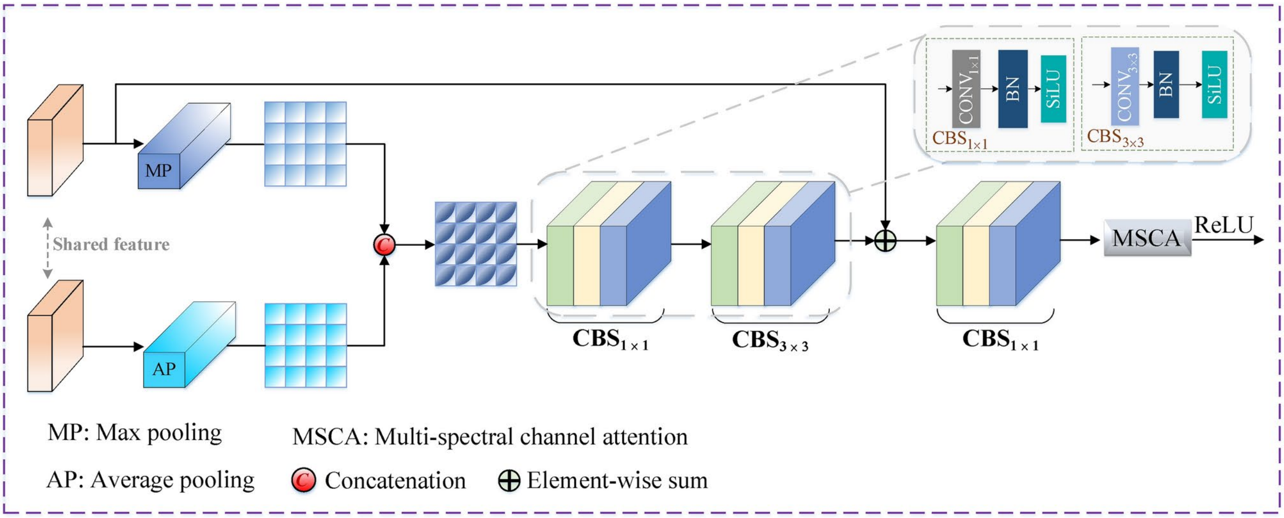


**Fig. 4**. The structure diagram of the proposed SIG module.

feature extraction structure that can greatly enhance the receptive field. It can effectively enhance the feature extraction capability of the network without greatly increasing the complexity of the network model.

When using YOLOv4-tiny to identify traffic signs, the accuracy is hindered by the relatively small size of the signs, low resolution, unclear features, and other objective factors. This often results in missed detections and false positives, reducing the effectiveness of small target recognition. Drawing from recent research on defect detection[64], we propose a SIG module that utilizes deep feature layers to guide shallow feature layers. By refining the semantics of the shallow feature layer, the influence of complex backgrounds on detection performance is reduced, and the semantic details of small traffic sign targets are effectively preserved. Additionally, traffic signs are typically small targets, and details about small targets are richer in shallow features due to higher spatial resolution in the shallow layer. Infusing semantic information into these shallow features can enhance and highlight the information representation of these small targets. For instance, the distinct shape and color of a traffic sign can be accentuated by the crucial semantic details from high-level features, aiding the network in accurately identifying small targets during detection.

The detailed structure of the SIG is depicted in Fig. 4. The workflow of SIG proceeds as follows. Initially, the deep output features undergo max pooling and average pooling. Furthermore, the output feature is represented as

$$f^{'} = cat(\varphi_{MP}(f_n), \varphi_{AP}(f_n)) \tag{4}$$

where *cat* denotes cascading operation, $\varphi_{MP}$ and $\varphi_{AP}$ refer to max pooling and average pooling operations, respectively. Taking a step further, we combine the features of the two branches to encompass more detailed global information. Following, the CBS module with a 1 × 1 convolution adjusts the channel count, while the CBS module with a 3 × 3 convolution enhances local context. The residual edge is then added element-wise to the deep feature map:

$$f^{''} = \Phi(\xi\left\{\beta\left\{Conv_{3\times3}\left\{\xi\left\{\beta[Conv_{1\times1}(f^{'})]\right\}\right\}\right\}\right\} + f_n) \tag{5}$$

where $\beta$ is the batch normalization (BN), $\xi$ represents the SiLU activation function, and $\Phi$ is the element-wise sum operation. The features then pass through a CBS block with a 1 × 1 convolution and a multi-spectral channel attention (MSCA)[65] to obtain the deep feature map's weight, which is activated by a modified ReLU function:

$$Y_n = \tau\left\{GAP\left\{\sigma_1\left\{\beta[Conv_{1\times1}(f^{''})]\right\}\right\}\right\} \tag{6}$$

where $\tau$ is the ReLU activation function, and MSCA denotes the multi-spectral channel attention operation. The MSCA mechanism dynamically adjusts the weights of different feature channels, enabling the network to focus on the most informative channels while suppressing irrelevant or redundant ones. This dynamic weighting process enhances the network's ability to learn target-specific features, which is particularly important for small and complex objects like traffic signs.

Through the above procedures, the SIG module leverages deep semantic information from the backbone network to guide the shallow feature layer. This design enhances the distinction between traffic signs and complex backgrounds, improving the detection of small traffic signs and reducing false positives caused by cluttered environments. Unlike traditional feature fusion methods, the SIG module explicitly strengthens the semantic information in shallow layers, which is critical for detecting small and dim traffic signs.

## Experiments

### Settings

In order to verify the effectiveness of the proposed E-mobilenet, MDSPP, and SIG modules, several comparative tests are conducted in this section. The experimental environment and the parameter settings are shown in Table 2.

Moreover, during the training process, the current training weight file is saved in time after the end of each epoch. At the same time, the change of the loss function is observed during the network training process. The model is tested when the loss function tends to be stable, indicating that the model has converged. At last, in order to eliminate the randomness of experimental results, the average of the model weights of 20 epochs after stabilization is taken for validation.

### Dataset and evaluation metrics

*Dataset*

(1) *CCTSDB dataset*: The Chinese traffic sign database (CCTSDB)[66] is produced by Zhang Jianming's team of Hunan Key Laboratory of Integrated Transportation Big Data Intelligent Processing of Changsha University of Science and Technology. Up to now, 15,734 images have been uploaded, including nearly 40,000 traffic sign targets. The current labeling data is divided into three categories: Indication sign, prohibition sign, warning sign. In this paper, the CCTSDB data set is divided into CCTSDB_l and CCTSDB_s according to the size of traffic signs in the image. Among them, 11,4735 images with large traffic signs were divided into CCTSDB_l dataset, and the remaining 4000 images with small traffic signs constituted CCTSDB_s dataset.

(2) *GTSDB dataset*: The German traffic sign detection benchmark (GTSDB)[67] is a standard dataset for traffic sign detection, featuring 900 high-resolution images of 43 common German traffic sign types. It includes

| | | |
|---|---|---|
| Experimental environment | Platform | Cuda11.7 |
| | Framework | Pytorch1.13.1 |
| | GPU | NVIDIA RTX 3070Ti |
| | Memory size | 8G |
| Parameter settings | Input_shape | 608*608 |
| | lr | 1e−5 |
| | IOU | 0.3 |
| | Batch_size | 64 |
| | Freeze_epoch | 50 |
| | Unfreeze_epoch | 150 |

**Table 2.** Experimental environment and parameter settings.

diverse scenes with varying weather, lighting, and challenges like partial occlusion, making it ideal for testing detection algorithm robustness in real-world applications. Widely used in autonomous driving and intelligent transportation research, GTSDB is a key benchmark in traffic sign detection.

(3) *TT100K dataset*: The Tsinghua-Tencent 100K (TT100K)[68] is a large-scale traffic sign detection and recognition benchmark with over 100,000 high-resolution images and 221 types of traffic signs commonly found on Chinese roads. Featuring significant class imbalances and challenging scenarios like occlusion, blur, and lighting variations, it is widely used to assess target detection and classification algorithms, making it a key resource in autonomous driving and intelligent transportation research.

We divided the data set into a 7:3 ratio of training sets and validation sets, and these images contained vehicle information in each scene and traffic signs at each location. The authenticity and universality of the data set are guaranteed. In addition, in order to verify the effectiveness of the improved ultra-lightweight and high-precision network structure in practical applications, we use mobile phones to shoot images of real scenes inside and around the campus. These scenes include traffic sign images under different circumstances, covering different angles, different lighting conditions, and different distances. The actual application scenario is simulated more realistically, which is helpful in evaluating the performance of the improved network structure in a complex environment.

### *Evaluation metrics*

When evaluating the target detection model, the accuracy and speed are generally measured. The accuracy evaluation index mainly includes four kinds:

- *Precision (Pr)*: represents the proportion of samples classified as positive that are truly positive.

$$Pr = \frac{TP}{TP + FP}. \tag{7}$$

- *Recall (Re)*: represents the proportion of samples that are correctly classified as positive in the true positive category.

$$Re = \frac{TP}{TP + FN}. \tag{8}$$

- *F1 score (F)*: the accuracy rate and recall rate are considered comprehensively, and it is the harmonic average of the two.

$$F = \frac{2 \times Pr \times Re}{Pr + Re}, \tag{9}$$

where TP, FP and FN are true positive examples, false positive examples and false negative examples, respectively.

- *mAP*: represents the average of the accuracy rates for all classes. In addition, AP represents the average accuracy of a single class, corresponding to the area under the precision recall curve, and mAP represents the average accuracy across all categories. The size of the mAP must be in the range [0,1], and the larger the better.

$$AP = \int_0^1 P\left(R\right) dR, \tag{10}$$

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i, \tag{11}$$

where *P*, *R*, *P*(*R*), *C* and $AP_i$ represent the accuracy rate, the recall rate, the precision recall curve, the total number of classes and the *AP* value of class *i*, respectively.

- *Frame per second (FPS)*: which represents how many images are recognized per second.

## Results and analysis

### *Quantitative comparison with state-of-the-arts*

In order to comprehensively compare the performance of MASG-Net and the other current mainstream networks, Table 3 shows their results of mAP and FPS on CCTSDB_s and their model size. Compared with the large complex network SSD_512 and YOLOv4, MASG-Net still has a certain gap in detection accuracy but is significantly ahead in terms of detection speed and model size. The YOLOv4-tiny+AFPN+RFB network[72] is built by adding adaptive feature pyramid networks (AFPN) and receptive field block (RFB) modules to YOLOv4-tiny. Compared to YOLOv4-tiny+AFPN+RFB, MASG-Net reduces the number of model parameters and improves the detection accuracy and speed. Compared to the latest YOLOv7-tiny, the detection accuracy of the network

| Network | mAP (%) | Total params (M) | FPS |
|---|---|---|---|
| SSD_512 | 96.5 | 24.7 | 38.4 |
| YOLOv4 | 95.9 | 64.0 | 39.1 |
| YOLOv4-tiny | 91.4 | 6.1 | 197.3 |
| YOLOv4-tiny+AFPN+RFB | 93.2 | 9.1 | 145.7 |
| YOLOv7-tiny | 94.1 | 6.2 | 256.0 |
| MASG-Net | **94.2** | **5.6** | **203.6** |

**Table 3**. The performance comparisons of various models on the CCTSDB_s dataset. Significant values are in bold.

| Network | Precision (%) | Recall (%) | F1-score (%) | mAP (%) | FPS |
|---|---|---|---|---|---|
| YOLOv4-tiny | 91.7 | 74.2 | 82.0 | 80.2 | 197.3 |
| MobileNetV3* | 89.6 | 72.4 | 80.1 | 78.9 | 192.1 |
| E-mobilenet* | 93.6 | 83.3 | 88.2 | 85.4 | 223.9 |
| E-mobilenet*+SPP | 94.1 | 83.8 | 88.7 | 86.5 | 221.4 |
| E-mobilenet*+MDSPP | 94.5 | 87.7 | 91.0 | 89.8 | 215.7 |
| Ren et al.[69] | 83.6 | 77.3 | 80.3 | 81.5 | 61 |
| Tang et al.[70] | 87.3 | 84.1 | 85.6 | 84.4 | 22.3 |
| Zhang et al.[71] | 98.7 | 90.5 | 94.4 | 92.7 | 29.6 |
| Yao et al.[72] | 93.5 | 82.4 | 87.6 | 86.8 | 145.7 |
| MASG-Net | 95.1 | 90.3 | 92.6 | 90.8 | 203.4 |

**Table 4**. The performance comparisons of various models on the GTSDB dataset. *indicates to highlight architectural modifications and their impact on performance metrics.

|  |  |  | AP(%) |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Networks | mAP(%) | Params(M) | i5 | io | p11 | p140 | p150 | pn | pne | po |
| SSD_512 | 68.3 | 24.7 | 81.5 | 74.4 | 65.8 | 72.1 | 70.1 | 70.7 | 84.3 | 44.3 |
| YOLOv4-tiny | 58.8 | 6.1 | 88.7 | 74.6 | 53.1 | 61.1 | 63.1 | 69.4 | 86.7 | 39.1 |
| YOLOv7-tiny | 54.9 | 6.2 | 90.0 | 70.2 | 64.7 | 38.4 | 49.1 | 90.0 | 89.8 | 38.6 |
| YOLOv8n | 70.4 | 3.1 | 89.4 | 81.3 | 78.5 | 67.2 | 65.5 | 86.6 | 89.5 | 60.0 |
| MASG-Net | 68.6 | 5.6 | 88.0 | 83.8 | 71.3 | 63.4 | 68.4 | 84.3 | 89.4 | 53.0 |

**Table 5**. The performance comparisons of various models on the TT100K dataset.

is similar, but MASG-Net uses a lightweight backbone network and still leads in model size. In addition, we can also see that lightweight models, such as MASG-Net and YOLOv4-tiny, have significantly faster detection speed than complex models, such as SSD and YOLO. Therefore, the proposed MASG-Net has superior comprehensive performance for traffic sign detection applications.

In GTSDB, we have compared MASG-Net with several state-of-the-art methods published in recent years, including both lightweight and high-accuracy detection networks commonly used for traffic sign detection tasks. The comparison models includes YOLOv4-tiny, Ren et al.[69], Tang et al.[70], Zhang et al.[71] and Yao et al.[72]. From the detection results in Table 4, it can be seen that MASG-Net has achieved suboptimal performance in multiple indicators, and the mAP has reached 90.8%. In GTSDB dataset, the scale and angle of traffic signs in the image may change due to the shooting distance, camera angle of view or the installation position of the sign, especially the detection of small targets at long distances or oblique angles. To address this challenge, MASG-Net introduces a global-local perception module that can simultaneously capture long-distance dependent global information and fine-grained local features. This module enhances the model's understanding of the macroscopic structure and microscopic features of traffic signs, thereby improving detection accuracy. MASG-Net achieves 203.4 FPS, combining high detection accuracy with impressive inference speed.

To further assess the effectiveness of the proposed algorithm, it is compared with five target detection models SSD, YOLOv4-tiny, YOLOv7-tiny, and YOLOv8n[73] on the TT100K dataset. Table 5 presents the results, where Params indicates the total parameters needed for model training, and mAP evaluates overall detection accuracy across all categories, reflecting the model's performance. Experimental results show that our algorithm achieves the second highest mAP among all compared models, reaching 68.6%, which is superior to most other models in detection accuracy. From the specific data in Table 5, it can be seen that the proposed algorithm performs best in the io and pl50 detection tasks, with detection accuracies reaching 83.8% and 68.4% respectively, achieving the best results. At the same time, MASG-Net achieved the second best performance on the p11, pl40 and po detection tasks. In addition, for the detection of other traffic sign categories, the proposed algorithm also maintained a high accuracy and achieved the second overall ranking of the mAP indicator.

To better show the lightweight and effectiveness of the proposed method, we draw a figure in which an x-axis indicates the number of parameters and the y-axis is the performance of different methods. In Fig. 5, we compare the performance of MASG-Net with preceding algorithms, including SSD, YOLOv4-tiny, YOLOv4, YOLOv7-tiny, and YOLOv4-tiny+AFPN+RFB. As shown in Fig. 5, the MASG-Net achieves a superior balance between lightweight design and high accuracy, outperforming other models with fewer parameters. This result shows that the proposed algorithm can achieve superior detection performance while reducing the number of parameters, fully reflecting the balance between computational efficiency and accuracy, and showing high practical application value.

*Qualitative results*

We visualize the detection effects in specific scenarios of different methods on the CCTSDB dataset, as shown in Fig. 6. YOLOv4, YOLOv4-tiny, YOLOv7-tiny and MASG-Net are respectively used to detect the public data set. It can be seen that the proposed MASG-Net achieves high detection accuracy. This is because MASG-Net accurately identifies traffic signs by capturing fine-grained features and memorizing long-term contexts. Additionally, it includes MDSSP and SIG modules to minimize external interference and ambiguous detections. The visual comparisons clearly illustrate MASG-Net's superior ability to detect small and difficult-to-recognize traffic signs while maintaining fewer false positives in cluttered backgrounds.

To better verify the generalizability of MASG-Net, we use mobile devices to photograph real scenes such as traffic signs and surrounding roads on campus, as shown in Fig. 7. YOLOv4, YOLOv4-tiny, YOLOv7-tiny and MASG-Net are, respectively, used to detect the random scene. By observing the detection results of the network model, the improved model not only effectively improves the probability of judging the prediction box as a certain type, but also the position of the prediction box is more accurate than that given by YOLOv4, YOLOv4-tiny and YOLOv7-tiny, and the center point of the prediction box basically coincides with the center point of the traffic sign. On the whole, MASG-Net has achieved significant improvement in the detection effect of small targets and dim and fuzzy traffic sign pictures with insufficient light. The practicability and robustness of MASG-Net have been verified by testing the environmental pictures of traffic signs around the campus taken by mobile phones. It is proved that MASG-Net has strong generalization ability in real road environment scene.

On the TT100K dataset, we have performed a visual comparative analysis of the baseline models YOLOv4-tiny, YOLOv7-tiny, and the proposed algorithm MASG-Net, as shown in Fig. 8. These results showcase detection outputs for various challenging scenarios, such as small, blurry, and occluded traffic signs, as well as signs in low-light environments. The comparison results show that the proposed algorithm outperforms the YOLOv4-tiny and YOLOv7-tiny models in detecting various types of traffic signs, including i5, io, p11, pl40, pl50 and pn. The prediction box generated by the proposed MASG-Net has a higher degree of match with the actual sign area, especially in the case of complex background, partial occlusion, aging, defacement and other reasons that lead to missing information, showing stronger robustness.

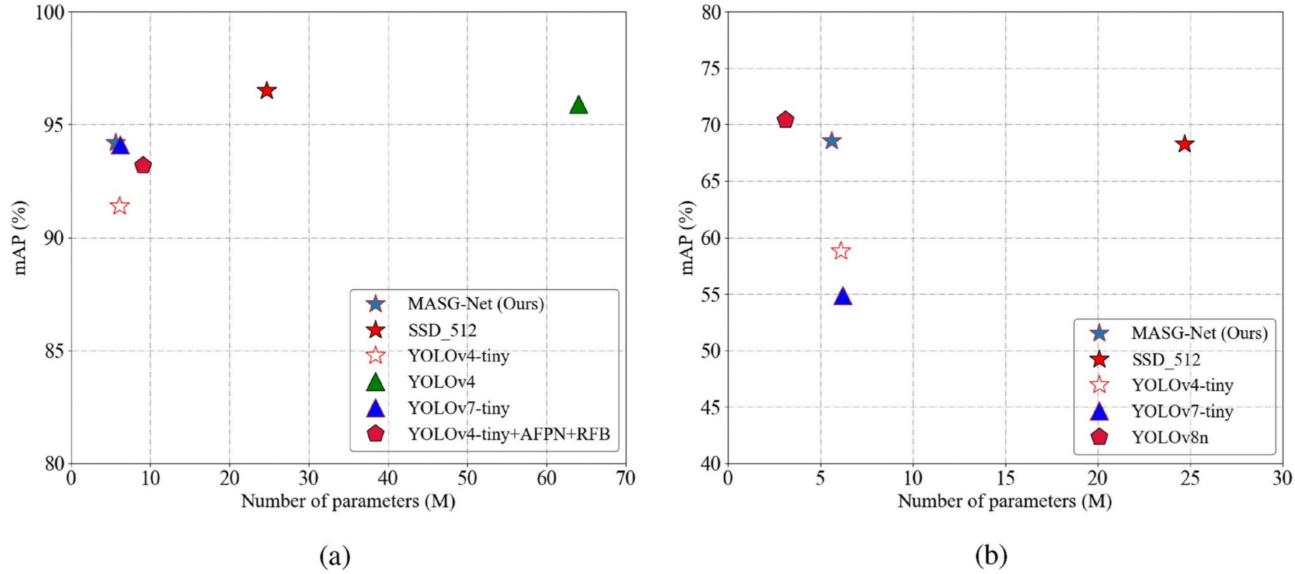
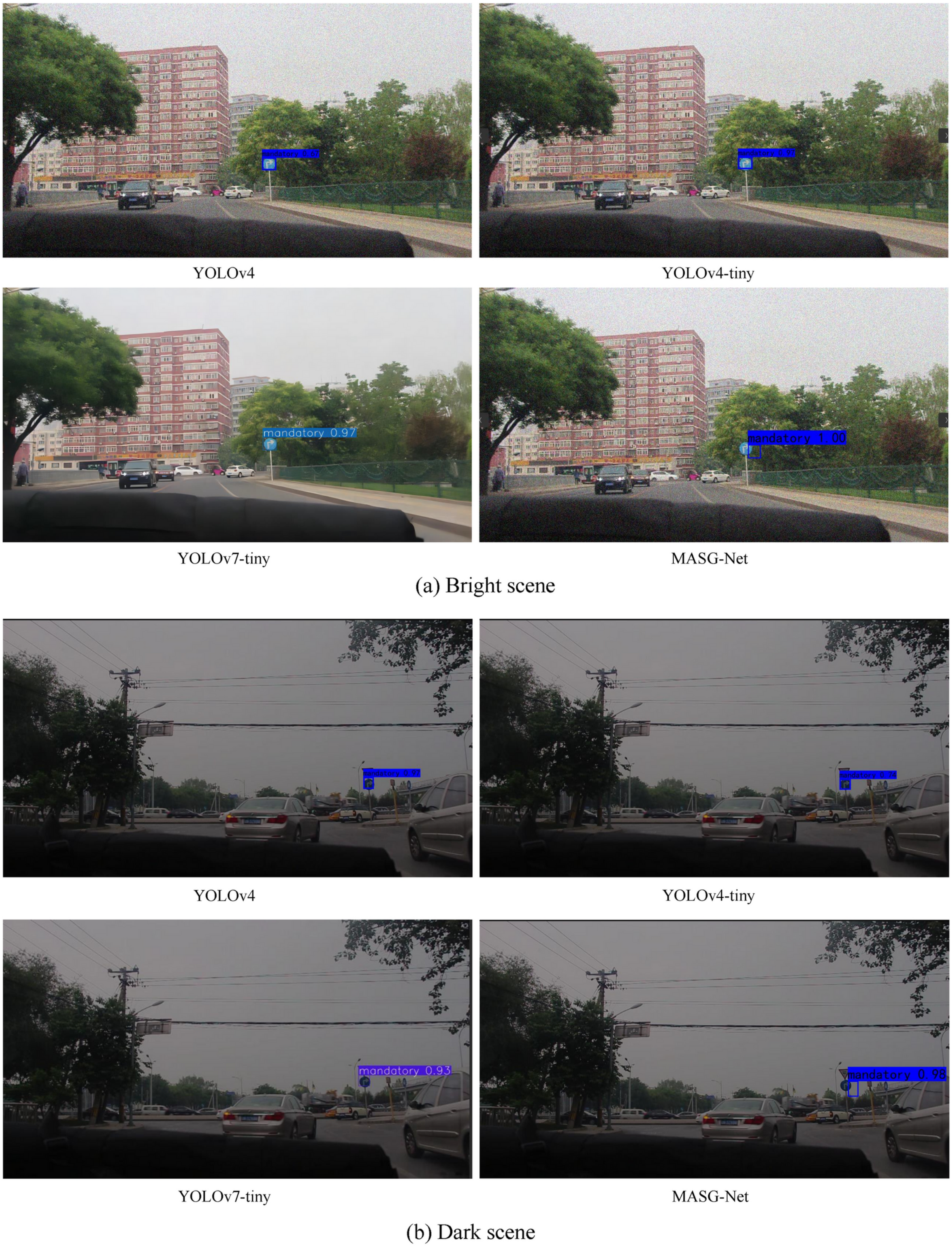


**Fig. 5**. MASG-Net and existing methods computational complexity analysis in terms of number of parameters and mAP. (**a**) and (**b**) show the test results using the CCTSDB_s and TT100K, respectively.



YOLOv4 YOLOv4-tiny

YOLOv7-tiny MASG-Net

(a) Bright scene

YOLOv4 YOLOv4-tiny

YOLOv7-tiny MASG-Net

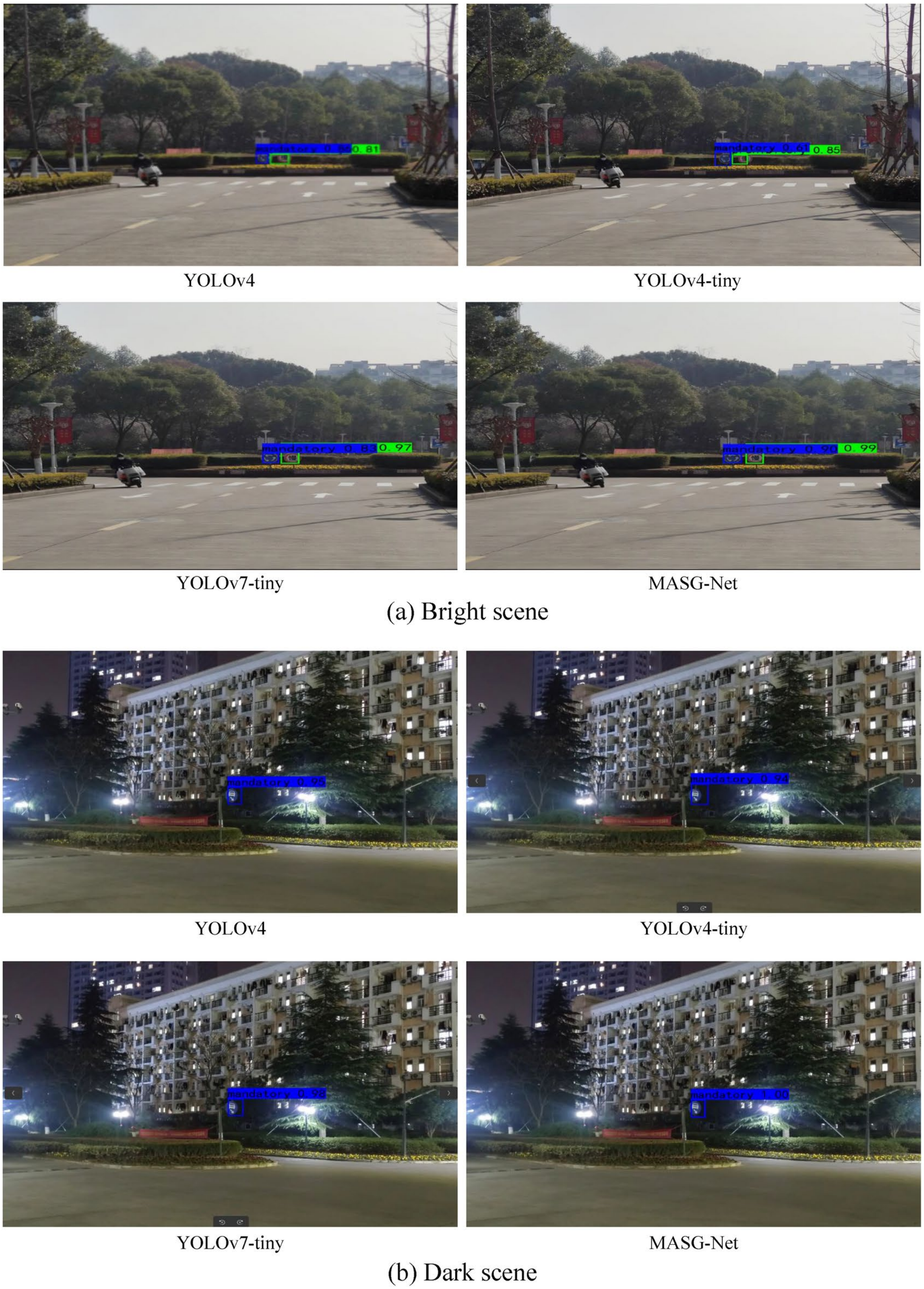
(b) Dark scene

**Fig. 6**. Visualization results of different algorithms on the CCTSDB dataset.

*Ablation study*

In order to evaluate MASG-Net more systematically and comprehensively, its trained models on the CCTSDB dataset and its subset CCTSDB_s and CCTSDB_l were used for ablation experimental tests, and the final results are shown in Table 6.

From the performance in Table 6 on CCTSDB_s, it can be seen that the performance of MASG-Net have been significantly improved compared with the original network. The new backbone network E-mobilenet brings obvious accuracy improvement, and the average accuracy index mAP is increased from 91.4 to 92.4%. This



YOLOv4 YOLOv4-tiny

YOLOv7-tiny MASG-Net

(a) Bright scene

YOLOv4 YOLOv4-tiny

YOLOv7-tiny MASG-Net

(b) Dark scene

**Fig. 7**. Visualization results of different algorithms on self-shot real scenes.

verifies the feature extraction ability of the backbone network E-mobilenet. This design resolves the performance-complexity trade-off by using local cross-channel interactions, significantly reducing network complexity while maintaining performance. By introducing multi-scale dilated convolution operation into SPP structure, the new MDSPP module improves the mAP of the model from 92.4 to 94.1%, and the precision, recall and F1 score are increased by 0.9%, 2.7% and 1.8%, respectively. This indicates that MDSPP can effectively enhance the ability of the network to extract features. Furthermore, when equipped with the proposed SIG module, it achieves 0.1

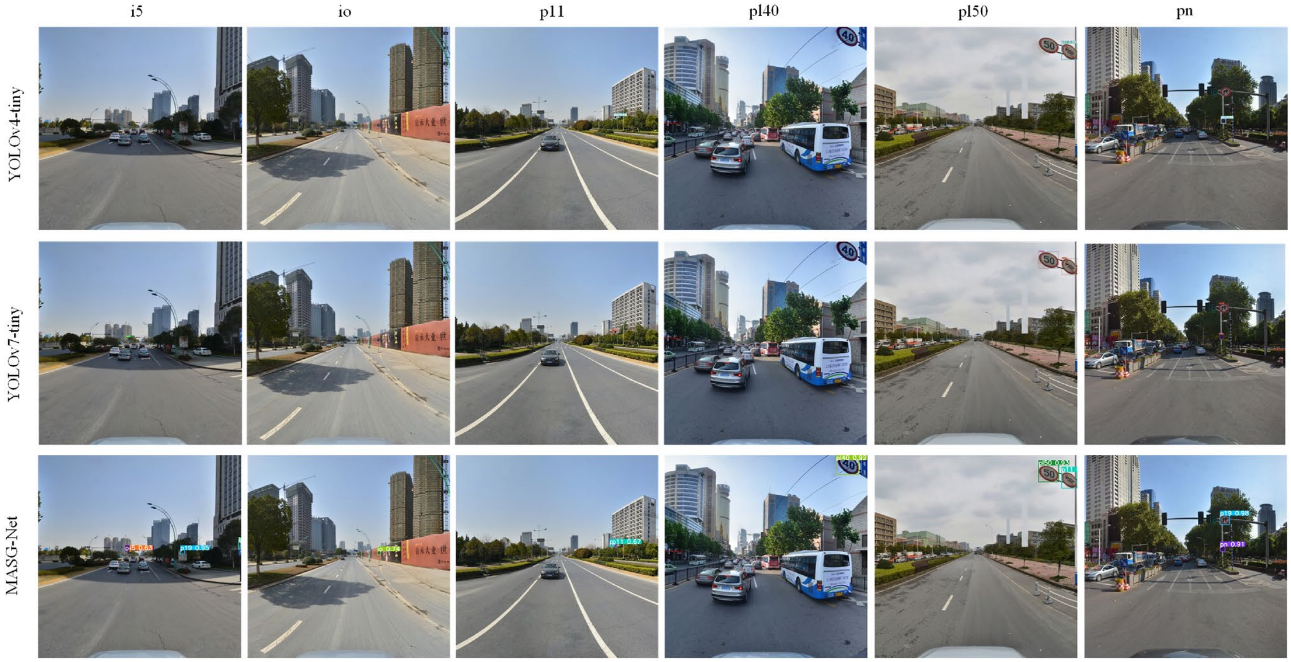


**Fig. 8**. Visualization results of different algorithms on the TT100K dataset.

| Dataset | Network | Precision (%) | Recall (%) | F1-score (%) | mAP (%) |
|---|---|---|---|---|---|
| CCTSDB_s | YOLOv4-tiny | 89.8 | 88.4 | 89.1 | 91.4 |
| | MobileNetV3* | 90.8 | 87.8 | 89.3 | 90.4 |
| | E-mobilenet* | 92.9 | 90.0 | 91.4 | 92.4 |
| | E-mobilenet*+SPP | 91.5 | 90.7 | 91.3 | 92.6 |
| | E-mobilenet*+MDSPP | 93.8 | 92.7 | 93.2 | 94.1 |
| | MASG-Net | 94.5 | 93.4 | 93.9 | 94.2 |
| CCTSDB_l | YOLOv4-tiny | 91.8 | 89.9 | 90.9 | 92.0 |
| | MobileNetV3* | 91.0 | 89.2 | 90.1 | 91.4 |
| | E-mobilenet* | 94.5 | 92.3 | 93.4 | 93.7 |
| | E-mobilenet*+SPP | 94.6 | 93.2 | 93.9 | 94.1 |
| | E-mobilenet*+MDSPP | 95.2 | 94.3 | 94.8 | 95.4 |
| | MASG-Net | 95.8 | 94.7 | 95.2 | 95.6 |
| CCTSDB | YOLOv4-tiny | 91.0 | 89.3 | 90.1 | 91.8 |
| | MobileNetV3* | 91.0 | 88.9 | 90.0 | 91.1 |
| | E-mobilenet* | 93.1 | 92.8 | 93.0 | 93.2 |
| | E-mobilenet*+SPP | 93.7 | 93.2 | 93.5 | 94.0 |
| | E-mobilenet*+MDSPP | 95.0 | 93.5 | 94.3 | 94.5 |
| | MASG-Net | 95.3 | 93.9 | 94.6 | 94.7 |

**Table 6**. Ablation study results on the CCTSDB dataset. *indicates to highlight architectural modifications and their impact on performance metrics.

points gains on mAP, 0.7 points gains on F1 score and the recall increase from 93.2 to 93.9%. This demonstrates the effectiveness of our SIG module in enhancing detection performance of tiny signs.

Compared with the network before the improvement, the accuracy, recall rate, F1 score and mAP of MASG-Net are increased by 4.7%, 5.0%, 4.8% and 2.8%, respectively. It can be seen that MASG-Net effectively solves the problem that YOLOv4-tiny is not strong in feature extraction ability when detecting small targets of traffic signs. Moreover, the MASG-Net is not only suitable for detecting small traffic signs, but also for detecting large traffic signs. Therefore, compared to YOLOv4-tiny, the proposed MASG-Net has improved performance in the recognition of traffic signs of different sizes.

*Potential limitations*

Although the MDSPP module improves the receptive field and enhances the detection of small targets, the performance may degrade when detecting extremely small traffic signs that occupy only a few pixels in the image. This is due to the inherent limitations of feature extraction at such a small scale. In addition, high-speed motion can cause significant motion blur, which reduces the clarity of traffic signs and makes detection more challenging. While MASG-Net enhances the receptive field and captures global contextual information, extreme motion blur may still lead to missed detections or false positives. As visible, the performance of MASG-Net may still be affected under extreme adverse weather conditions, such as heavy rain, fog, or snow, where the visibility of traffic signs is significantly reduced.

## Conclusion

In this paper, an ultra-lightweight and high-precision network, MASG-Net, is proposed on the basis of YOLOv4-Tiny network for traffic sign detection applications. Firstly, an ultra-lightweight feature extraction network, E-mobilenet, is designed to enhance the feature extraction capability of the network while effectively reducing the number of parameters. Secondly, based on SPP, the MDSPP is proposed, which greatly enhances the receptive field range of the feature map and enables the network to obtain more global information. Finally, we propose a SIG module that utilizes deep feature layers to guide shallow feature layers. By refining the semantics of the shallow feature layer, the influence of complex backgrounds on detection performance is reduced. The combination of the E-mobilenet backbone, MDSPP and SIG modules significantly improves the detection of small, dim, and blurry traffic signs, especially in challenging environments such as low-light conditions. Compared with the network before improvement, the precision, recall rate, F1 score, and mAP of MASG-Net are increased by 4.7%, 5.0%, 4.8% and 2.8%, respectively. It can be seen that MASG-Net effectively solves the problem that YOLOv4-tiny is not strong in feature extraction ability when detecting small targets of traffic signs. Compared to other models, it has better detection accuracy and smaller model complexity. In addition, the feasibility of MASG-Net to detect traffic signs in the real scene is verified by the detection of road environment pictures.

However, there are still some shortcomings in the research work. Vehicles driving at high speed may have an impact on the imaging effect, the pictures captured by the camera may have fuzzy deformation and be difficult to identify, and other vehicles, pedestrians or buildings may also partially block the traffic signs, affecting the detection effect. Subsequent detection algorithms need to be able to accurately identify these situations. To further validate the real-time performance of MASG-Net on devices with limited computational resources (e.g., automotive ECUs or edge devices), we plan to deploy the model on platforms such as NVIDIA Jetson Nano, Raspberry Pi, or similar hardware.

## Data availability

All relevant data are within the paper.

Received: 16 July 2024; Accepted: 14 March 2025
Published online: 24 March 2025

## References

1. Zhao, Z. et al. Dense tiny object detection: A scene context guided approach and a unified benchmark. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–13 (2024).
2. Shi, Y., Zhao, S., Wu, J., Wu, Z. & Yan, H. Fixated object detection based on saliency prior in traffic scenes. *IEEE Trans. Circ. Syst. Video Technol.* **34**, 1413–1426 (2024).
3. Reddy, M. P. et al. A deep learning model for traffic sign detection and recognition using convolution neural network. In *2022 2nd international conference on intelligent technologies (CONIT)*, 1–5 (IEEE, 2022).
4. Njoku, J. N., Nwakanma, C. I., Amaizu, G. C. & Kim, D.-S. Prospects and challenges of metaverse application in data-driven intelligent transportation systems. *IET Intel. Transp. Syst.* **17**, 1–21 (2023).
5. Wu, Z. et al. Enhanced spatial feature learning for weakly supervised object detection. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 961–972 (2024).
6. Yao, H. et al. Occlusion-aware plane-constraints for monocular 3d object detection. *IEEE Trans. Intell. Transp. Syst.* **25**, 4593–4605 (2024).
7. Yang, L. et al. Adadet: An adaptive object detection system based on early-exit neural networks. *IEEE Trans. Cogn. Dev. Syst.* **16**, 332–345 (2024).
8. Yuan, Y. et al. Edge-cloud collaborative UAV object detection: Edge-embedded lightweight algorithm design and task offloading using fuzzy neural network. *IEEE Trans. Cloud Comput.* **12**, 306–318 (2024).
9. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587 (2014).
10. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448 (2015).
11. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inform. Process. Syst.*, **28** (2015).
12. Liang, T., Bao, H., Pan, W. & Pan, F. Traffic sign detection via improved sparse R-CNN for autonomous vehicles. *J. Adv. Transp.*, (2022).
13. Gu, Y. & Si, B. A novel lightweight real-time traffic sign detection integration framework based on YOLOv4. *Entropy* **24** (2022).
14. Hao, C., Zhang, H., Song, W., Liu, F. & Wu, E. Slinet: Slicing-aided learning for small object detection. *IEEE Signal Process. Lett.* **31**, 790–794 (2024).
15. Wang, X. et al. Real-time and efficient multi-scale traffic sign detection method for driverless cars. *Sensors (Basel, Switzerland)* **22** (2022).
16. Liu, W. et al. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37 (Springer, 2016).
17. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).

18. Redmon, J. & Farhadi, A. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271 (2017).
19. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018).
20. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020).
21. Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021).
22. Li, C. et al. Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022).
23. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475 (2023).
24. Munir, A. & Siddiqui, A. J. Vision-based UAV detection in complex backgrounds and rainy conditions. arXiv preprint arXiv:2305.16450 (2023).
25. Zhu, Y. & Yan, W. Q. Traffic sign recognition based on deep learning. *Multimedia Tools Appl.* **81**, 17779–17791 (2022).
26. Qu, S., Yang, X., Zhou, H. & Xie, Y. Improved yolov5-based for small traffic sign detection under complex weather. *Sci. Rep.* **13**, 16219 (2023).
27. Dey, A., Biswas, S. & Abualigah, L. Efficient violence recognition in video streams using ResDLCNN-GRU attention network. *ECTI Trans. Comput. Inform. Technol.* **18**, 329–341 (2024).
28. Wang, G. et al. Fighting against terrorism: A real-time CCTV autonomous weapons detection based on improved YOLO v4. *Digital Signal Process.* **132**, 103790 (2023).
29. Parisae, V. & Bhavanam, S. N. Multi scale encoder-decoder network with time frequency attention and s-tcn for single channel speech enhancement. *J. Intell. Fuzzy Syst.* **46**(4), 10907–10907 (2024).
30. Vanambathina, S. D. et al. Speech enhancement using u-net-based progressive learning with squeeze-tcn. In *Proceedings of the international conference on advances in distributed computing and machine learning*, 419–432 (2024).
31. Parisae, V., Bhavanam, S. N. & Devi, M. V. Progressive learning framework for speech enhancement using multi-scale convolution and s-tcn. In *Proceedings of the international conference on inventive systems and control*, 83–89 (2024).
32. Gogineni, R., Ramakrishna, Y., Veeraswamy, P. & Chaitanya, J. Pansharpening of multispectral images through the inverse problem model with non-convex sparse regularization. In *Proceedings of the international conference on robotics, control, automation and artificial intelligence*, 513–525 (2022).
33. Sermanet, P. et al. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013).
34. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1904–1916 (2015).
35. Huang, Z. et al. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based yolo for object detection. *Inf. Sci.* **522**, 241–258 (2020).
36. Bayramoglu, N., Tiulpin, A., Hirvasniemi, J., Nieminen, M. T. & Saarakkala, S. Adaptive segmentation of knee radiographs for selecting the optimal ROI in texture analysis. *Osteoarthritis Cartilage* **28**, 941–952 (2020).
37. Xu, X. et al. Crack detection and comparison study based on faster R-CNN and mask R-CNN. *Sensors* **22**, 1215 (2022).
38. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969 (2017).
39. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440 (2015).
40. Zhang, D., Zhang, W., Li, F., Liang, K. & Yang, Y. Pnanet: Probabilistic two-stage detector using pyramid non-local attention. *Sensors (Basel, Switzerland)* **23** (2023).
41. Hao, C., Hou, C. & Orlando, D. Performance analysis of an enhanced two-stage detector. In *2015 IEEE China summit and international conference on signal and information processing (ChinaSIP)* 292–295 (2015).
42. Bandiera, F., Besson, O., Orlando, D. & Ricci, G. A two-stage detector with improved acceptance/rejection capabilities. In *2008 IEEE international conference on acoustics, speech and signal processing* 2301–2304 (2008).
43. Li, Z. et al. Light-head r-cnn: In *Defense of two-stage object detector*. ArXiv **abs/1711.07264** (2017).
44. Zhang, S., Wen, L., Bian, X., Lei, Z. & Li, S. Z. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4203–4212 (2018).
45. Gevorgyan, Z. Siou loss: More powerful learning for bounding box regression. arXiv preprint arXiv:2205.12740 (2022).
46. Munir, A. & Siddiqui, A. J. Vision-based UAV detection in complex backgrounds and rainy conditions. arXiv preprint arXiv:2305.16450 (2023).
47. Guan, L., Jia, L., Xie, Z. & Yin, C. A lightweight framework for obstacle detection in the railway image based on fast region proposal and improved yolo-tiny network. *IEEE Trans. Instrum. Meas.* **71**, 1–16 (2022).
48. Oltean, G., Florea, C., Orghidan, R. & Oltean, V. Towards real time vehicle counting using yolo-tiny and fast motion estimation. In *2019 IEEE 25th international symposium for design and technology in electronic packaging (SIITME)* 240–243 (2019).
49. Padala, A. & Malathi, P. An optimized object detection system using you only look once algorithm and compare with tiny-yolo algorithm with increased accuracy. In *2022 2nd international conference on innovative practices in technology and management (ICIPTM)* **2**, 606–610 (2022).
50. Sumit, S. S., Rambli, D. R. A., Mirjalili, S. H., Ejaz, M. & Miah, M. S. U. Restinet: On improving the performance of tiny-yolo-based cnn architecture for applications in human detection. Applied Sciences (2022).
51. Adarsh, P., Rathi, P. & Kumar, M. Yolo v3-tiny: Object detection and recognition using one stage improved model. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, 687–694 (IEEE, 2020).
52. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. Scaled-YOLOv4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 13029–13038 (2021).
53. Wang, Y., Zhang, J., Kan, M., Shan, S. & Chen, X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* 12272–12281 (2020).
54. Li, Y., Zeng, J., Shan, S. & Chen, X. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. Image Process.* **28**, 2439–2450 (2019).
55. McClenny, L. D. & Braga-Neto, U. M. Self-adaptive physics-informed neural networks using a soft attention mechanism. ArXiv **abs/2009.04544** (2020).
56. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).
57. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).
58. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13713–13722 (2021).
59. Wang, Q. et al. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542 (2020).
60. Zhang, Q. & Yang, Y. Sa-net: Shuffle attention for deep convolutional neural networks. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2235–2239 (2021).
61. Xu, W. & Yi, W. Ela: Efficient local attention for deep convolutional neural networks. arXiv preprint arXiv:2403.01123 (2024).

62. Kim, J., Nang, J. & Choe, J. Lmlt: Low-to-high multi-level vision transformer for image super-resolution. arXiv preprint arXiv:2409.03516 (2024).
63. Howard, A. G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017).
64. Yu, X., Lyu, W., Zhou, D., Wang, C. & Xu, W. Es-net: Efficient scale-aware network for tiny defect detection. *IEEE Trans. Instrum. Meas.* **71**, 1–14 (2022).
65. Qin, Z., Zhang, P., Wu, F. & Li, X. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 783–792 (2021).
66. Zhang, J., Huang, M., Jin, X. & Li, X. A real-time Chinese traffic sign detection algorithm based on modified YOLOV2. *Algorithms* **10**, 127 (2017).
67. Stallkamp, J., Schlipsing, M., Salmen, J. & Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Netw.* **32**, 323–332 (2012).
68. Zhu, Z. et al. Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2110–2118 (2016).
69. Ren, K., Huang, L., Fan, C., Han, H. & Deng, H. Real-time traffic sign detection network using DS-DetNet and lite fusion FPN. *J. Real-Time Image Proc.* **18**, 2181–2191 (2021).
70. Tang, Q., Cao, G. & Jo, K.-H. Integrated feature pyramid network with feature aggregation for traffic sign detection. *IEEE Access* **9**, 117784–117794 (2021).
71. Zhang, J., Xie, Z., Sun, J., Zou, X. & Wang, J. A cascaded r-cnn with multiscale attention and imbalanced samples for traffic sign detection. *IEEE Access* **8**, 29742–29754 (2020).
72. Yao, Y., Han, L., Du, C., Xu, X. & Jiang, X. Traffic sign detection algorithm based on improved YOLOv4-tiny. *Signal Process.: Image Commun.* **107**, 116783 (2022).
73. Varghese, R. & Sambath, M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *Proceedings of the IEEE international conference on advances in data engineering and intelligent computing systems (ADICS)*, 1–6 (2024).

## Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61671192, and in part by the National Science Foundation for Post-Doctoral Scientists of China under Grant 2017M114, and in part by the Top-Ranking Discipline a Class of Electronics Science and Technology in Zhejiang Province, China.

## Author contributions

Chenjie Du was responsible for manuscript editing, data collection, and statistical analysis. Siyu Su and Chenwei Lin verified the entire article and participated in experimental design and data analysis. Yingbiao Yao and Xinghua Hong did the supervision and provided guidance on the topic. Ran Jin was responsible for writing the literature review and discussion sections and made contributions to the execution of the experiments. All authors jointly reviewed and approved the final version of the article.

## Declarations

### Competing interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

## Additional information

**Correspondence** and requests for materials should be addressed to R.J.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.