IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 January 2023, accepted 24 January 2023, date of publication 31 January 2023, date of current version 8 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241234

THEORY

SC-YOLO: A Object Detection Model for Small Traffic Signs

YANLI SHI[1], XIANGDONG LI[2], AND MIAOMIAO CHEN[3]
[1]School of Science, Jilin Institute of Chemical Technology, Jilin 132022, China
[2]College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China
[3]Yantai Research Institute, Harbin Engineering University, Harbin 150001, China

Corresponding author: Yanli Shi (syl@jlict.edu.cn)

This work was supported in part by the Natural Science Foundation of Jilin Province under Grant YDZJ202201ZYTS602.

ABSTRACT Automatic traffic sign detection has great potential for intelligent vehicles. In recent years, traffic sign detection has made significant progress with the rise of deep learning. Detecting small traffic signs in real-world scenarios is still a challenging problem due to the complex and variable traffic environment. In this paper, a model with a small number of parameters is proposed to improve the accuracy of small traffic sign detection. Firstly, the cross-stage attention network module is proposed to enhance the feature extraction capability of the network. Secondly, a dense neck structure is proposed to make the detail information and semantic information fully fused. Finally, for the model's loss function, SIOU with direction information is introduced to optimize the model's training process. Tests on the challenging public datasets TT100K, CCTSDB2021, and VOC show that our approach achieves significant performance improvement with the minimum number of parameters compared to existing algorithms.

INDEX TERMS Deep learning, cross-stage attention, small object detection, traffic sign detection.

I. INTRODUCTION

With the rapid development of science and technology, technologies such as driver assistance systems and autonomous driving have gradually emerged. And traffic sign detection system, as a sub-module in intelligent transportation systems, plays an important role in providing current traffic information to drivers and intelligent vehicle control systems to improve driving safety, so the recognition of traffic signs has become a popular topic of research. Many methods have achieved good results on some public transportation sign detection datasets.

In the traditional traffic sign recognition algorithm, the research focuses on feature extraction and feature classification, by segmenting the color space and combining feature extraction methods based on the shape and edges of traffic signs, and then realizing the recognition of traffic signs by completing feature classification through classifiers. Due to the specific shapes and eye-catching colors of traffic signs, many traditional traffic signs detection methods based on manual features were proposed by scholars concerned in the early days based on these characteristics [1], [2]. However, these methods are difficult to be widely applied in practical tasks. First, designing these feature extraction methods requires many human resources. On the other hand, these simple features lack sufficient robustness to cope with complex and changing traffic environments.

With the development of convolutional neural networks, deep learning-based object detection algorithms have gradually replaced traditional object detection algorithms. Traffic sign recognition is a sub-task of object detection, and many general object detection algorithms can be directly applied to traffic sign recognition. However, there is a big difference between the proportion of traffic signs and common objects occupied in the image. The traffic sign seen in a car occupies only a small part of the whole image. As shown in Fig.1, a sign in a 2048*2048 pixels high-resolution image may occupy only 30*30 pixels. Such small objects are still a challenge for object detection due to their low resolution and low information content. In recent years, many scholars have

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

FIGURE 1. Small traffic signs in the image.

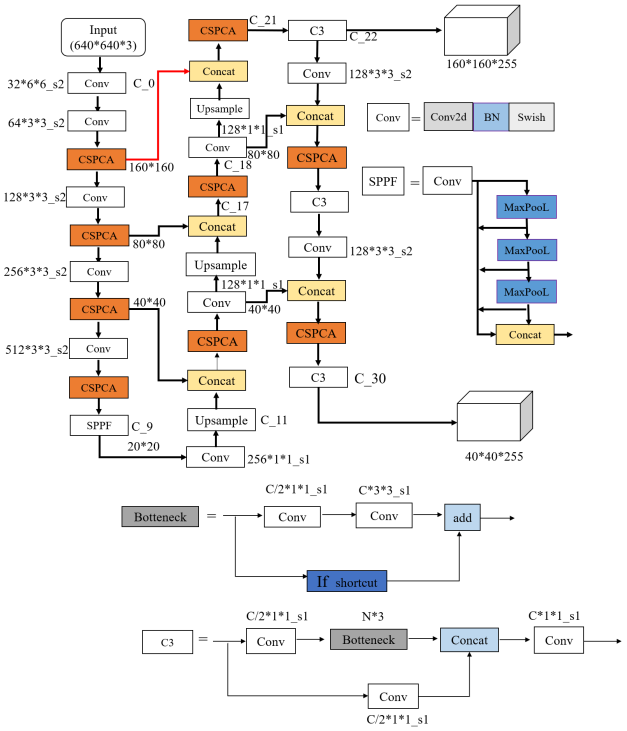


FIGURE 2. The network structure of SC-YOLO. Where 32*6*6 represents the convolution kernel with dimension 32 and size 6*6, s2 represents stride=2 and s1 represents stride=1, C_0 represents the first layer of the network.

proposed some theories and methods to improve the performance of small object detection. The method [3] is to build a high-resolution feature map and make predictions on it. This method obtains fine detail information but loses contextual information. Methods [4], [5] fuse contextual information by building a top-down structure. This method effectively improves detection accuracy by combining low-level details and high-level semantic features at various scales. However, this approach obtains small feature maps by downsampling multiple times and then reconstructing the spatial resolution, which may result in small feature maps retaining little information in small object detection and severely affecting the model's performance. Method [6] use a multiscale strategy to improve small object detection performance, but the shallow feature extraction is insufficient, and the small object detection accuracy improvement is insignificant. These methods are difficult to apply to engineering mobile detection because of their high computational cost and relatively large memory footprint in the training and testing phases.

Inspired by the above methods and combined with the state-of-the-art YOLO series of object detection algorithms, we propose the SC-YOLO network structure shown in Fig.2 to improve the performance of small object detection. To address the loss of deep small object feature map information due to multiple downsampling, we design a cross-stage attention network module (CSPCA), which is used in the backbone network to enable the network to obtain more focused region information and improve the feature extraction capability of the network. In the feature fusion phase of the model, we design a structure with a small number of parameters and a robust feature fusion capability. The previous network performs feature fusion after multiple downsampling, resulting in the loss of much detailed information. We adjust the downsampling multiplier to introduce lower-level detail information and high-level semantic information fusion. The grid of the previous YOLO series algorithm's detection head is too large to detect small objects. We propose a more detailed grid division picture to suppress the background information of a single grid, and reduce the three detection heads of YOLO to two to reduce the model's parameters. Since CIOU calculates the loss values of ground truth and bounding box lacks information on the direction, we introduce the SIOU [7] loss function with directional information to make the model easier to converge. To evaluate the model, we choose the CCTSDB2021 [8] and TT100K [9] datasets containing many small-sized traffic signs in natural scenes. To summarize, the contributions of this paper are as follows.

(1) Propose a cross-stage attention network module structure to enhance the weight of small objects on the feature maps of the backbone and neck networks, suppress the background information of little significance, and reduce the loss of useful information in the deep feature maps.

(2) Propose a fusion structure with a low number of parameters and strong feature fusion capability so that low-level detail information and high-level semantic information can be effectively fused.

(3) Introduce the SIOU loss function to calculate the regression loss. The bounding box regression with direction information is more conducive to model convergence and further improves the recognition accuracy of small objects.

II. RELATED WORK

A. OBJECT DETECTION

Object detection is a technique for locating objects in an image and giving the object class. The current popular object detection algorithms are divided into two categories: Two-Stage object detection algorithms and One-Stage object detection algorithms. The classical mainstream algorithms for Two-Stage object detection are R-CNN [10], SPP-Net [11], Fast R-CNN [12], Faster R-CNN [13], etc. The R-CNN algorithm proposed by Ross Girshick et al. is the first industrial-grade accuracy Two-Stage object detection algorithm. Although the classification-based Two-Stage algorithm has been greatly improved in terms of detection effect, the algorithm's speed still cannot meet the real-time object detection task requirements. With the proposed One-Stage object detection algorithm, the efficiency of object detection has been greatly improved, making it possible to apply it to the real-time sensory detection of objects in autonomous driving systems. The one-Stage object detection algorithm is a new class of detection algorithms based on regression ideas proposed by scholars. The two typical classes of algorithms are the SSD series [14] and the YOLO series. In 2016, Redmon et al. proposed the YOLO algorithm [15], which pioneered the transformation of the detection problem into a regression problem and used convolutional neural networks to directly accomplish the prediction of boundaries and the determination of object classes. The real sense of real-time object detection was achieved, which opened a new era of the One-Stage algorithm for object detection. Later, YOLO was continuously optimized and improved, and YOLO v2, v3, v4, v5, v6, and v7 [16], [17], [18], [19], [20], [21] were proposed. However, the YOLO series algorithm is mainly used to detect general objects, and the detection of small objects like traffic signs needs to be improved. Therefore, this paper proposes SC-YOLO,to improve the detection accuracy of small traffic signs, using the YOLO series algorithm as the basic framework.

B. TRAFFIC SIGN DECTION

As a sub-task of object detection, traffic sign recognition has been continuously proposed by scholars with related theories and solutions. The traditional research methods are mainly based on color and shape for recognition. The literature [22] uses the detection of corner vertices and corner parallels to detect triangular traffic signs. The literature [23] uses color segmentation based on the AdaBoost binary classifier and cyclic Hough transform for traffic sign detection. The literature [24] proposed an Ohta space color probability model for traffic sign detection by drawing color probability maps. The traditional algorithm has weak generalization ability, and the detection effect decreases dramatically when the color fades and the shape changes. With the development of convolutional neural networks, deep learning-based algorithms are widely used to detect traffic signs.

The Tsinghua team [13] produced TT100K traffic sign dataset based on Tencent Street View and proposed a neural network structure to predict and classify. The literature [25] proposed a cascaded R-CNN to obtain multi-scale features of pyramids, weighted multi-scale features by dot product and softmax, and their phases to refine features to highlight traffic sign features and improve the accuracy of traffic sign detection. MR-CNN [6] used a multi-scale inverse fold product structure to combine deep and shallow features. The fused feature mapping can reduce the number of region suggestions to a certain extent and improve the efficiency of traffic sign detection. The above model is based on improving the two-stage object detection algorithm. Although it has made great progress in the traffic sign detection task, it has more computational parameters and a more complex model, which is not conducive to deployment in mobile applications. Therefore, we improve the algorithm on the one-stage object detection algorithm.

C. SMALL OBJECT DETECTION

Small object detection is a challenging task in object detection. On the one hand, small objects have low resolution and little visualization information, making it challenging to extract discriminative features and highly susceptible to interference by environmental factors. On the other hand, small objects occupy a small area in the image, and a single pixel point shift in the prediction bounding box can cause a significant error in the prediction process. Compared with large objects, small objects easily appear in the aggregation phenomenon. When small objects appear in aggregation, the small objects adjacent to the aggregation area cannot be distinguished. When similar small objects appear intensively, the predicted bounding boxes may also be missed due to the post-processing NMS filtering many correctly predicted bounding boxes.

In recent years, several methods have been proposed to improve the accuracy of small object detection. The literature [26] utilizes a multi-scale learning approach with shallow feature maps to detect smaller objects and deeper feature maps to detect larger objects. However, the single multi-scale learning with lower layers does not have enough feature non-linearity to achieve the desired accuracy. The literature [27] uses contextual information, object and scene, and object-object coexistence relationships to improve the performance of small object detection. The method based on context fusion improves the accuracy of object detection to a certain extent, but how to find the contextual information from the global scene that is beneficial to improve small object detection is still a difficult research problem. The literature [28] uses generative adversarial learning by mapping the features of small low-resolution objects into features equivalent to those of high-resolution objects to achieve the same detection performance as that of larger-sized objects. However, generative adversarial networks are difficult to train and do not easily achieve a good balance between generators

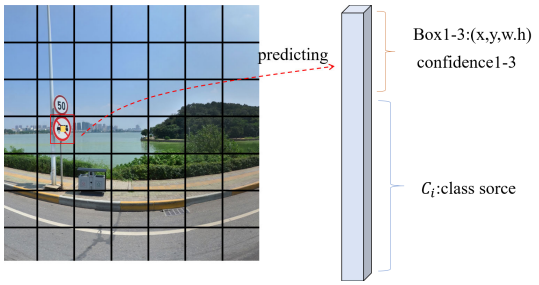


FIGURE 3. The prediction process of YOLO. S*S girds.

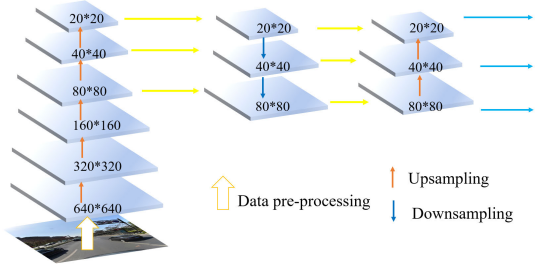


FIGURE 4. YOLO Framework.640*640 represents the size of the feature map.

and discriminators. The literature [29] uses Transformer Prediction Heads (TPH) instead of the original Prediction Heads based on YOLOv5. Although the detection capability of small objects is improved, the Transformer structure is complicated. For the difficulty of traffic small object feature extraction, we use a combination of convolutional attention [30] and contextual information to deal with it.

III. APPROACH

A. YOLO ALGORITHM

The core idea of YOLO is to use the whole image as the network's input and use the CNN network to divide all the input images into S*S grids. As shown in Fig.3, each grid is responsible for detecting the object whose center point falls within the grid and regressing the position and type of the bounding box in the output layer.

After YOLOv3, each grid needs to predict three bounding boxes. Each bounding box needs to predict not only the position coordinates and the confidence value but also the scores of C categories. confidence is the confidence level, which is the probability of the presence of objects in the bounding box; C is related to the category of the dataset. Each bounding box needs to predict five values: x, y, w, h, and confidence. (x, y) denotes the center of the box relative to the boundary of the grid cell; (w, h) denotes the predicted width and height of the box relative to the whole image; confidence denotes the class probability*IOU if there is an object in the bounding box, the class probability is 1. Otherwise, it is 0. Then, when there is an object, the confidence can also be expressed as the IOU between the bounding box and the ground truth.

The general framework of the YOLO family of algorithms after YOLOv3 can be summarized as a four-part composition. As shown in Fig.4, the first part is image pre-processing, which performs Mosaic data enhancement and adaptive image scaling on the input image. The image after pre-processing is richer in image features. The second part is the feature extraction of the image, YOLOv4 and YOLOv5 are the backbone network with CSP network structure, and YOLOV7 is the backbone network with ELAN network structure to realize the feature extraction of the image. The third part is the feature fusion of the image, which consists of the neck network with FPN [31] and PAN [32] structures to fuse the extracted features. The fourth part is the detection layer, which consists of a loss function and a prediction frame screening function to calculate the information loss.

YOLOv5 represents the more widely used and mature YOLO series at this stage. YOLOv5 integrates many of today's state-of-the-art methods, such as mosaic data enhancement, cross-stage partial connection, SPP block [11], PAN structure, and path aggregation module. YOLOv5 is an efficient and powerful object detection model, and after experimental comparison, YOLOv5 performs better in traffic sign detection tasks, so YOLOv5 is used as a baseline in this paper.

B. CROSS-PHASE ATTENTION MODULE

The low resolution of traffic signs makes it difficult to extract features with discriminative power, and they are highly susceptible to interference by environmental factors. We propose the cross-stage attention network module (CSPCA) to enhance the feature extraction capability of small objects of traffic signs.

The overall CSPCA network structure is divided into two branches so that the gradient streams are propagated through two different network paths, the propagated gradient information can have large correlation differences, and the aggregation strategy of the gradient streams is used to prevent different layers from learning duplicate gradient information. As shown in Fig.5, these two branches are called the dense local block and local transport layer, respectively. The dense local layer comprises an ordinary convolution and a CABottleneck, and the local transport layer consists of just an ordinary convolution. The feature map X is mapped into two parts, $X = [x_1, x_2]$. x_1 through the dense local block and x_2 through the local transport layer. The CSPCA network structure not only allows the network depth to be deepened but also to focus on small objects.

The attention mechanism in the CSPCA network structure is inspired by humans. By looking at the global information of an image, humans can object the candidate area of focus under their attention, automatically blocking part of the background and redundant information and quickly locking the focus. The attention mechanism used in CSPCA is designed in the dense local layer. After the feature map X is input, the convolution kernels of $(H, 1)$ and $(1, W)$ are used to encode each channel of the X tensor along the horizontal and vertical

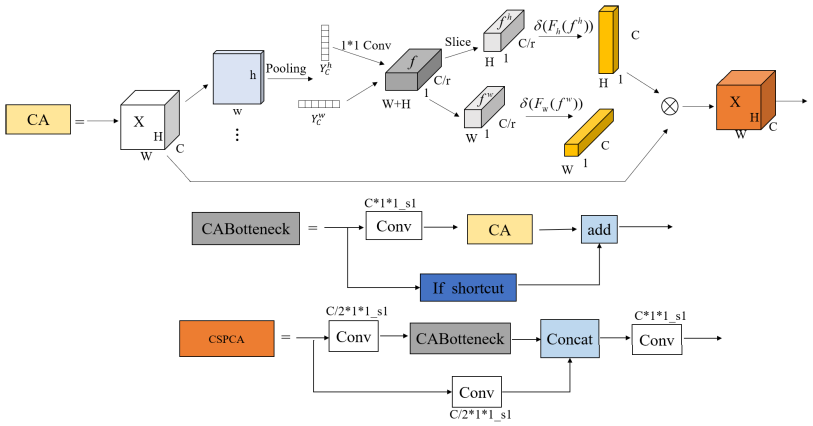


FIGURE 5. CSPCA Structure. Where C*1*1 represents the convolution kernel with dimension C and size 1*1, s1 represents stride=1.

directions, respectively, and H, W are the height and width of the feature map before processing. The output of the cth channel with height h before processing is as follows.

$$Y_c^h(h) = \frac{1}{W}\sum_{i=1}^{W} x_c(h, i) \tag{1}$$

The output of the cth channel of width w before processing is as follows.

$$Y_c^{\mathrm{w}}(w) = \frac{1}{H}\sum_{j=1}^{H} x_c(j, w) \tag{2}$$

Using two one-dimensional convolutional kernels, a global pooling operation is performed on the feature maps, and the input features in horizontal and vertical directions are aggregated into two independent direction-aware feature maps, which are encoded into two attention maps, respectively. Each attention map contains the long-range dependencies of the input feature maps along one spatial direction and preserves the precise location information along the other spatial direction, enabling the CSPCA network to acquire the region of interest more accurately. The horizontally and vertically averaged pooled output tensor is stitched together and then transformed by a shared 1*1 convolution operation as follows.

$$f = \delta(F(Y_c^h, Y_c^w)) \tag{3}$$

The generated $f \in R^{\frac{C}{r}*(H+W)}$ is the intermediate feature map in the horizontal and vertical directions of space, and r denotes the step size of downsampling, which is used to control the size of the attention module.

Slice f into two independent tensors, $f^h \in R^{\frac{C}{r}*H}$ and $f^w \in R^{\frac{C}{r}*W}$, along the spatial dimension, after which the feature maps f^h and f^w are transformed to the same number of channels as the X input using two 1*1 convolution F_h and F_{w}, respectively, as follows.

$$g^h = \delta(F_h(f^h)) \tag{4}$$

$$g^h = \delta(F_h(f^h)) \tag{5}$$

where δ represents the sigmoid activation function, reducing the complexity of the model and the computational overhead. The final attention weight matrix is obtained as follows.

$$y_c(i, j) = x_c(i, j) * g_c^h(i) * g_c^w(j) \tag{6}$$

C. SC-YOLO FRAMEWORK

CSPCA is flexible enough to be a backbone network for any off-the-shelf object detector. Considering the trade-off between accuracy and efficiency, we embed it into a one-stage object detection framework, YOLOv5, for demonstration. The performance of YOLOv5 in small object detection needs to be improved. This is because, after multiple downsampling, YOLOv5 extracts little spatial information about small objects.

In this section, we propose SC-YOLO, as shown in Fig.2. The method consists of 3 parts: (1) the backbone part, which uses a cross-stage attention network module (CSPCA) to extract basic features; (2) the neck, which introduces lower-level detail information fused with high-level semantic information and uses a combination of FPN and PAN networks in a model designed for dense PAN; (3) the head, which uses a more fine-grained prediction grid for making predictions. In the following, we describe these three stages in detail.

1) BACKBONE

In this stage, images are adjusted as input, and then feature extraction is performed with a convolutional neural network. YOLOv4 chooses CSPDarkNet53 as the backbone network, and YOLOv5 uses CSP. Although the CSP network can increase the network depth, which somewhat alleviates the problem of network degradation and gradient disappearance, as the network deepens, the feature information of small objects is easily lost. YOLOv7 adopts ELAN network structure as the backbone network, which controls the shortest and longest gradient paths, and the model training convergence time is improved. However, the training requires a large

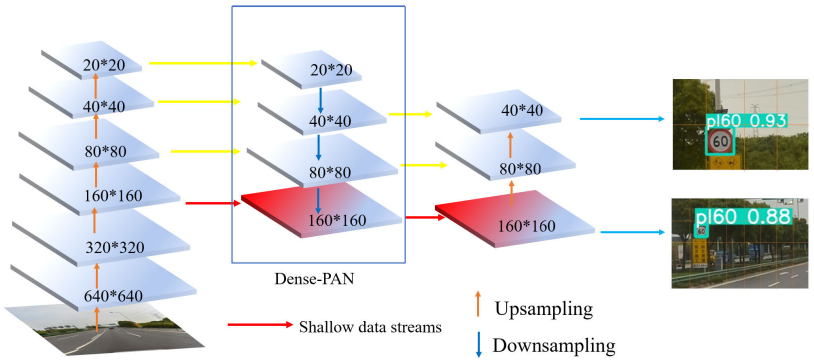


FIGURE 6. Neck&Head structure.

amount of memory, which is more demanding on hardware, and the performance improvement for small objects is not obvious. In this paper, we adopt our designed CSPCA, which is more capable of feature extraction for small objects than the previous backbone network structure.

2) NECK

In this phase, we propose a top-down structure to fuse low and high-level features from different backbone layers for different detection heads. YOLOv4, YOLOv5, and YOLOv7 downsampled the input 8x and 16x for feature fusion of low-level information. With this operation, the low-level feature information is not fused enough, and some feature information of small objects is lost. Therefore, we designed a more dense fusion structure Dense-PAN, as the neck of the network, as shown in Fig.6. The downsampling of 4x, 8x, and 16x for feature fusion improves the neck network's ability to fuse low-level information. For example, for a 640*640 input image, we add the fusion of 160*160 low-level detail features in the neck, and this information can also be used for different detection heads through the top-down network.

Specifically, as shown in Fig.2,the last two convolutional layers of YOLOV5s with 256 and 512 channels are removed, and the total number of parameters is reduced by 1.6M. The convolutional layer C_18 with 128 channels and 16k parameters is added; the convolutional layer C_21 with 128 channels and 84k parameters is added; the convolutional layer C_22 with 128 channels and 74k parameters is added; the convolutional layer C_26 with 128 channels and 74k parameters is added; the convolutional layer C_30 with 256 channels and 296k parameters is added; the total number of parameters is increased by 544K. Compared with YOLOv5s, the total reduction of parameters is 1.05M.

3) HEAD

To reduce the number of parameters in the model, we use only two scales to detect objects on the feature map output by Neck. To locate small traffic signs accurately, we also designed the detection grid for the detection head. For small objects of 640*640 images, we use a 160*160 grid to divide the image instead of an 80*80 grid to predict small objects like YOLOv4, YOLOv5, and YOLOv7. The 160*160 grid divides the image more carefully, twice as much as the previous yolo series, suppresses the interference of background information and is less likely to miss small objects, thus improving the ability to locate small objects. As shown in Fig.6 below, if the input image is divided into a 2*4 grid, the background information will occupy most of it. However, by dividing the image into a 4*8 grid, the background information will occupy less of it. The finer granularity of grid division is beneficial to improve the detection ability of small objects.

D. LOSS FUNCTION

The loss function can calculate how well the model predicts the results and determine whether there is a gap between the model and the actual data, so the loss function is crucial in training the model. The proper loss function is beneficial to get a better model and faster convergence during training.

In the YOLO series of object detection algorithms, the loss function consists of three parts: localization loss, classification loss, and confidence loss. Among them, localization loss is significant for an object detection algorithm. IOU suffers from the problem of scale insensitivity, and the current DIOU and CIOU are improvements on IOU. However, none of the currently proposed and used methods takes into account the direction of the mismatch between the desired real frame and the predicted frame. This deficiency leads to slower and less efficient convergence because the prediction frames may ''wander around'' during the training process and produce worse models. Therefore, we introduce the loss function SIOU, which contains the vector's angle between the true frame and the predicted frame, as shown in Fig.7, and is defined as follows.

Angle cost:

$$\wedge=1\text{-}2\sin^2(\arcsin(x)-\frac{\pi}{4}) \tag{7}$$

where C_h is the height difference between the center point of the ground truth and the bounding box, σ is the distance between the center point of the ground truth and the bounding

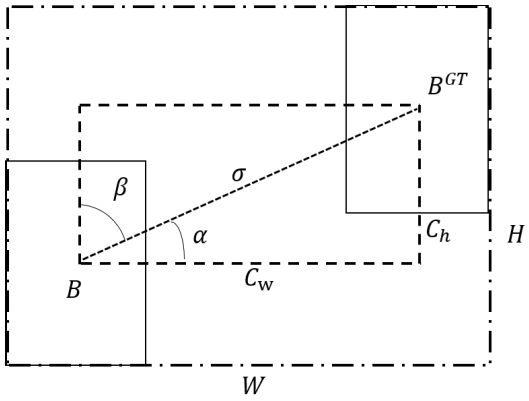


FIGURE 7. SIOU structure.

box, $C_h = \max(b_{c_y}^{g^t}, b_{c_y}) - \min(b_{c_y}^{g^t}, b_{c_y})$, $(b_{c_x}^{g^t}, b_{c_y}^{g^t})$ are the real box center coordinates, (b_{c_x}, b_{c_y}) are the prediction box center coordinates. when α is $\frac{\pi}{2}$ or 0, the angle loss is 0, during the training process if $\alpha < \frac{\pi}{4}$ then minimize α, otherwise minimize β.

Distance cost:

$$\Delta = \sum_{t=x,y} (1 - e^{-\gamma \rho_t}) = 2 - e^{-\gamma \rho_x} - e^{-\gamma \rho_y} \tag{8}$$

where $\rho_x = (\frac{b_{c_x}^{g^t} - b_{c_x}}{W})^2$, $\rho_x = (\frac{b_{c_y}^{g^t} - b_{c_y}}{H})^2$, $\gamma = 2 - \wedge$.

Shape cost:

$$\Omega = (1 - e^{-w_w})^{\theta} + (1 - e^{-w_h})^{\theta} \tag{9}$$

where $w_w = \frac{|w - w^{g^t}|}{\max(w, w^{g^t})}$, $w_h = \frac{|h - h^{g^t}|}{\max(h, h^{g^t})}$. (w, h) and (w^{g^t}, h^{g^t}) are the width and height of the ground truth and the bounding box, respectively, and controls the attention to the shape in order to avoid too much attention to the shape loss and reduce the movement of the prediction frame.

IOU cost:

$$IOU = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|} \tag{10}$$

The final loss function is

$$L = 1 - IOU + \frac{\Omega + \Delta}{2} \tag{11}$$

IV. EXPERIMENTS

A. DATASETS

Dataset 1: CCTSDB2021 [8] traffic sign data set produced by Changsha University of Technology is one of the most recognized data sets of traffic signs in China. The dataset contains three significant categories of traffic signs, namely "directional signs," "prohibition signs," and "warning signs," and includes six kinds of weather conditions, such as night, snow, and rain, which are close to real life. There are 16354 images in the training set and 1500 in the validation set.

Dataset 2: A joint lab of Tsinghua University and Tencent compiled and made public the TT100K [9] dataset, in which they downloaded 100,000 street view images from Tencent's map data center in five different cities in China and later labeled the traffic signs in the images with bounding boxes. The TT100K dataset contains 151 categories, only 45 categories have more than 50 instances, and nearly half of the instances are single-digit categories, which creates a severe data distribution imbalance. Therefore, the dataset was processed, and only the 45 categories with more than 50 instances were retained. The training set has 6107 images, and the validation set has 3073 images.

B. EXPERIMENTAL CONFIGURATION AND EVALUATION INDEXES

The experiments in this paper were conducted under Windows 10 operating system, using the pytorch1.10.0 framework, CUDA version 11.3. Hardware devices: GPU model 3080, 12G graphic memory.

Precision is the proportion of the correct bounding box, and Recall is the proportion of the bounding box among all ground truth. As shown in Eq.(12) and Eq.(13), TP denotes the number of correctly detected objects, FP denotes the number of incorrectly detected objects, and FN denotes the number of unpredicted ground truth. F1 denotes the summed average of Precision and Recall, as defined in Equation Eq.(14). mAP is defined in Eq.(16), mAP is the mean value of AP for all categories, AP is the accuracy rate of a single category, AP is defined in Eq.(15). The higher the value of mAP, the higher the accuracy of the algorithm. The number of parameters(Params) is used to measure the complexity of a model, which is related to the size of the memory resources of the computer occupied by the model, the smaller the Params the fewer the parameters of the model, the smaller the memory occupied. For a certain convolutional layer, its number of Params is shown in equation EquationEq.(17). Where K_h is the height of the convolution kernel, K_w is the width of the convolution kernel, C_{in} is the number of input channels, and C_{out} is the number of output channels.

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{12}$$

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{13}$$

$$F1 = \frac{2PR}{P + R} \tag{14}$$

$$AP = \int_0^1 P(R) \tag{15}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{16}$$

$$Paras = (K_h * K_w * C_{in}) * C_{out} + C_{\text{out}} \tag{17}$$

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) PERFORMANCES ON CCTSDB2021

To demonstrate the superiority of SC-YOLO in traffic sign detection, we conducted experiments on the CCTSDB2021 dataset, and the results are shown in Table1. SC-YOLO is compared with the basic two-stage object detection algo-

TABLE 1. The detection performance comparison of different methods on the CCTSDB2021 dataset.

Method	P	R	F1	mAP	Params(M)
Faster R-CNN [13]	84.4	54.9	66.5	56.5	143.7
Libra R-CNN [33]	83.7	60.0	70.0	61.4	–
Dynamic R-CNN [34]	87.0	58.3	69.8	60.0	–
Sparse R-CNN [35]	94.1	52.6	67.6	59.7	–
SSD [14]	86.5	27.4	42.0	49.2	–
YOLOv3 [17]	84.6	42.7	56.8	50.5	–
YOLOv4 [18]	76.2	52.5	62.2	51.7	–
YOLOv7-tiny [21]	89.8	74.9	81.7	79.9	6.2
YOLOv5s	91.2	73.1	81.2	80.9	7.2
Ours	**93.8**	**76.8**	**84.5**	**84.3**	**6.1**

rithm FasterR-CNN, as well as Dynamic R-CNN and Sparse R-CNN in the last two years. Also, it is compared with the one-stage object detection SSD as well as the algorithms of the YOLO family in recent years.

Overall, the two-stage object detection algorithm is more accurate than the previous one-stage, but YOLOv5 and this year's YOLOv7 achieved good detection results. The Precision, Recall, and mAP of the classical FasterR-CNN were only 84.4%, 54.9%, and 56.5%. The improved algorithms after FasterR-CNN, Dynamic R-CNN, and Sparse R-CNN made some progress. Dynamic R-CNN utilizes a Dynamic label Assignment strategy to adaptively fit the variation of the distribution of the regression labels, which is 3.3% and 3.5% higher than the F1 and mAP of FasterR-CNN. Sparse R-CNN utilizes a purely sparse image target detection method with many-to-one label assignment; the F1 and mAP are 1.1% and 3.2% higher than those of the Faster R-CNN. Among them, Precision is more prominent in Sparse R-CNN with a P value of 94.1%, and SC-YOLO's Precision is 93.8%, which is 0.3% lower than it, but the combined index F1 of Precision and Recall is 16.9% higher than it.

One of the first-stage object detection algorithms, SSD, has the worst performance, with F1 and mAP values of only 42% and 49.2%, indicating a single multiscale learning with low model nonlinearity. yolov7-tiny and yolov5 perform contextual information fusion and have a better overall performance. yolov7-tiny has the highest F1 among the YOLO family of algorithms, but SC -YOLO has 2.7% and 4.4% higher F1 and mAP than YOLOv7-tiny. Compared with YOLOv5s, SC-YOLO has 2.6%, 3.7%, 3.4%, and 3.4% higher P, R, F1, and mAP, respectively, with 15% fewer model parameters. Fig.8 and Fig.9 show the comparison of the training process between the SC-YOLO model and the YOLOv5s model, from which it can be seen that the SC-YOLO model converges faster and the training process is smoother than the YOLOv5s. The above comparison shows that our proposed method has good performance.

2) PERFORMANCES ON TT100K

To further demonstrate the superiority of SC-YOLO in traffic sign detection, SC-YOLO is compared with Faster R-CNN, YOLOv5s, and YOLOv7-tiny object detection

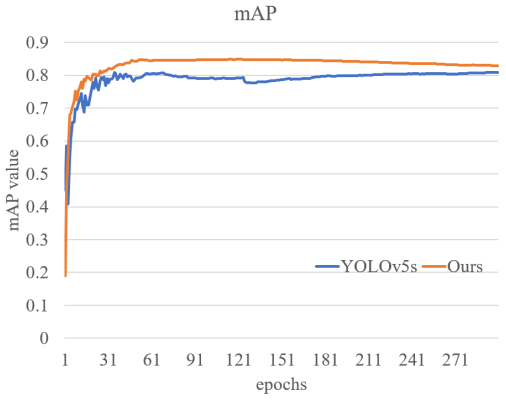


FIGURE 8. Comparison chart of mAP of YOLOv5s and our method.

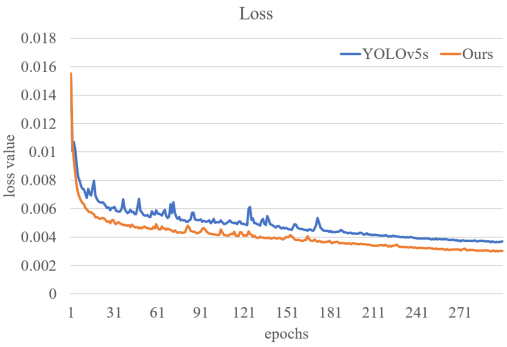


FIGURE 9. Comparison of training loss between YOLOv5s and our method.

algorithms on TT100K data set, and also with zhu, DR-CNN, MSA_YOLOV3, and IFA-FPN some traffic sign detection algorithms were compared. Tsinghua zhu's team created the TT100K dataset and obtained the most advanced results on the TT100K dataset at that time. DR-CNN, MSA_YOLOV3, and IFA-FPN are the new advancement of the TT100K dataset.

The detection performance of our proposed method and other methods is shown in Table2. The results show that SC-YOLO performs best on TT100K with the minimum number of parameters. Faster R-CNN, as a representative of a two-stage target detector, has an average performance of 64.8% and 73.4% of F1 and mAP for traffic signs, respectively, due to multiple downsampling, which leads to the loss of small target information. DR-CNNN uses a two-stage adaptive loss function, which is 1.4% and 0.2% higher than the Precision and Recall of TT100K creator zhu. IFA-FPN introduced Integrated Operation (IO) to solve the imbalance problem of Region-of-Interests (ROIs) in pyramid levels, which is higher than TT100K creator zhu's Precision, Recall, and mAP by 3.3%, 1.2%, and 5.6%. At the input image of 1280*1280, Precision, Recall, and mAP of SC-YOLO are 92.3%, 92.6%, and 95.2%, which are 0.5%, 1.7%, and

TABLE 2. The detection performance comparison of different methods on the TT100K dataset.

Method	P	R	F1	mAP	Params(M)
Faster R-CNN [13]	56.9	77.2	64.8	73.4	143.7
Zhu et al [9]	87.7	91.0	89.3	88.0	81.2
DR-CNN [37]	89.1	91.2	90.0	–	147.9
IFA-FPN [40]	91.0	92.2	91.6	93.6	–
YOLOv3 [17]	73.0	79.6	76.2	81.5	–
MSA_YOLOv3 [36]	79.6	84.4	81.9	86.3	–
YOLOv7-tiny [21]	85.3	78.9	81.9	83.7	6.2
YOLOv5s$_{(640)}$	85.3	76.8	80.8	83.6	7.2
Ours$_{(640)}$	89.6	85.0	87.2	90.4	6.1
YOLOv5s$_{(1280)}$	91.8	90.9	91.3	94.0	7.2
Ours$_{(1280)}$	**92.3**	**92.6**	**92.5**	**95.2**	**6.1**

FIGURE 10. Comparison chart of YOLOv5s and our method. (a)Results for YOLOv5s (b)Results for ours.

1.2% higher than Precision, Recall, and mAP of YOLOv5s, respectively, and higher than F1 of DR-CNN with the largest parameters 1.4%, and 1.6% higher than the mAP of IFA-FPN.

In addition, the model performance improvement varies with the resolution size of the input image. When the input image is 640*640, SC-YOLO has 6.4% and 6.8% higher F1 and mAP than YOLOv5s, respectively. At the input image of 1280*1280, SC-YOLO has 1.2% higher F1 and mAP than YOLOv5s. SC-YOLO significantly improves at low resolution, indicating that general target detection algorithms, such as YOLOv5, have weaker feature extraction ability and lose more information on low-resolution images. At the same time, our method can better enhance the model's feature extraction ability at low-resolution images feature extraction ability and reduce the information loss of small targets.

The detection results of YOLOv5s and SC-YOLO on the TT100K dataset are visualized as shown in Fig.10. When the traffic signs are very small, YOLOv5s has problems with missed detection, false detection, or low confidence. YOLOv5s only detects the near traffic signs, while SC-YOLO detects both distant and near traffic signs. YOLOv5s incorrectly identifies traffic signs with a speed limit of 80 as 60, while SC-YOLO accurately identifies traffic signs with a speed limit of 80. YOLOv5s has a confidence level of only 0.51 for prohibited signs, while SC-YOLO has a confidence level of 0.91 for prohibited

TABLE 3. Speed performance comparison of different methods on TT100K dataset.

Method	mAP	Speed	FPS	GPU
Faster R-CNN [13]	73.4	0.33s	3.0	GTX980
Zhu et al [9]	88.0	5.83s	–	GTX980
Lu et al [41]	87.0	0.26s	3.8	GTX980
MSA_YOLOv3 [36]	86.3	42.0ms	23.8	Tesla P100
YOLOv5s	83.6	28.3ms	35.3	RTX3080
Ours	90.4	29.7ms	33.7	RTX3080

TABLE 4. The detection performance comparison of different methods on the VOC dataset.

Method	Backbone	Params(M)	mAP
Faster R-CNN [13]	ResNet-101	–	76.4
Ganster R-CNN [38]	ResNet-101	–	80.7
YOLOv4	MobileNet-v2	46.3	81.5
YOLOv4	MobileNet-v3	47.3	78.9
YOLOv4	EEEA-Net-C2 [41]	31.2	81.8
YOLOv7-tiny [21]	ELAN	6.2	81.0
YOLOv5s	CSPNet	7.2	82.6
Ours	CSPCA	**6.1**	**83.0**

3) PERFORMANCES ON SPEED

To verify the speed performance of the model, we tested it on the public dataset TT100K, and the results are shown in Table3, we can see that SC-YOLO achieves the highest map, 6.8% higher than the mAP of YOLOv5s, although it is 1.7 FPS slower than YOLOV5S. SC-YOLO achieves 33.7 FPS, reaching the speed of processing 30 images a second.

4) PERFORMANCES ON VOC

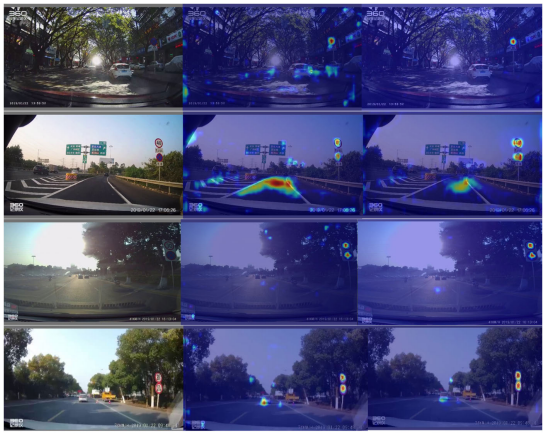
To illustrate the generalizability of SC-YOLO, we perform model performance validation on the dataset VOC. As can be seen in Table4, our method is optimal, with 6.6% higher mAP than Faster R-CNN, 0.4% higher mAP than YOLOv5s, and 2% higher mAP than YOLOv7-tiny. SC-YOLO achieves better performance even when compared with algorithms improved in recent years. It is 2.3% and 1.2% higher than the mAP of Ganster R-CNN and EEEA-Net-C2, respectively. This indicates that our proposed method can also improve the detection capability of the model for general objects.

5) ANALYSIS OF ABLATION EXPERIMENTS

To further analyze the effectiveness of our proposed critical method, we performed ablation experiments at CCSTDB2021. Since the classical object detection YOLOv5s performs best synthetically, we used YOLOv5s as a baseline to confirm our method. "CSPCA" represents the feature extraction network we proposed, "Neck&head" represents the neck and head structure we proposed for small objects, and "SIOU" represents the introduction of

TABLE 5. Ablation experiments on the components in the proposed method.

Method	CSPCA	Neck&head	SIOU	F1	mAP
YOLOv5s				81.2	80.7
A	✓			82.2	81.6
B	✓	✓		84.0	84.1
C	✓	✓	✓	84.5	84.3



(a) (b) (c)

FIGURE 11. Comparison of yolov5s and CSPCA network's grad-cam. (a) Original image (b) Results of YOLOv5s (c) Results of experiment ours.

the loss function with orientation information. Experiment A indicates that only the CSPCA was used, and experiment B indicates that only the CSPCA and the neck improvement were used. Experiment C indicates that all methods were used together. Meanwhile, we visualize the heat map of CSPCA and the result is shown in Fig.11.

As can be seen from Table5, each innovation point has a role in the model performance. Our proposed feature extraction network improves F1 and mAP by 1% and 0.9%, respectively, over YOLOv5s, which indicates that CSPCA has a better small-object feature extraction ability and the attention mechanism of CSPCA can reduce the loss of small-object information due to network deepening. "Neck&head" improves the F1 and mAP of the model by 1.8% and 2.3%, respectively, which indicates that our designed neck and head networks have a better ability to fuse the contextual information and retain the low-level detail information better. When SIOU was used as the loss function, the F1 and mAP of the model were improved by 0.5% and 0.2%, respectively, which indicated that the loss function with directional information helped the model optimization.

V. CONCLUSION

This paper is devoted to improving the accuracy of small traffic sign recognition with a model of small number of parameters. Small traffic sign detection has been a challenge for object detection. Although previous methods have achieved good results in this direction, the complexity and accuracy of the model still need to reach a rational level. This paper proposes a high-performance object detection model, SC-YOLO, for small-scale traffic sign detection. In the feature extraction phase, we propose the cross-stage attention network module to make the model more accurate in obtaining the region of interest. In the feature fusion stage, we propose a fusion of lower-level detail information and higher-level semantic information of the neck, which is more conducive to detecting small objects. In the detection phase, we propose a more detailed grid to detect small objects, suppressing the interference of background information. Finally, in the training stage, we introduce the loss function SIOU with direction information, and the model converges faster and smoother when training. SC-YOLO is evaluated on the public transportation sign datasets TT100K and CCTSDB2021, and the results show the feasibility and effectiveness of the model. Meanwhile, we validate the generalizability of our algorithm on the VOC dataset. In future work, we plan to study real-time small traffic sign recognition on mobile systems with limited memory and computational power. In addition, we intend to handle special weather conditions in traffic, such as rain and snow, in our future work.

REFERENCES

[1] A. Ruta, Y. Li, and X. Liu, "Real-time traffic sign recognition from video by class-specific discriminative features," *Pattern Recognit.*, vol. 43, no. 1, pp. 416–430, 2010.

[2] J. M. Lillo-Castellano, I. Mora-Jiménez, C. Figuera-Pozuelo, and J. L. Rojo-Álvarez, "Traffic sign segmentation and classification using statistical learning methods," *Neurocomputing*, vol. 153, pp. 286–299, Apr. 2015.

[3] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 21–30.

[4] L. Cui, P. Lv, X. Jiang, Z. Gao, B. Zhou, L. Zhang, L. Shao, and M. Xu, "Context-aware block net for small object detection," *IEEE Trans. Cybern.*, vol. 52, no. 4, pp. 2300–2313, Apr. 2020.

[5] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.

[6] Z. Liu, J. Du, F. Tian, and J. Wen, "MR-CNN: A multi-scale region-based convolutional neural network for small traffic sign recognition," *IEEE Access*, vol. 7, pp. 57120–57128, 2019.

[7] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," 2022, *arXiv:2205.12740*.

[8] J. Zhang, X. Zou, L.-D. Kuang, J. Wang, and R. S. Sherratt, "CCTSDB 2021: A more comprehensive traffic sign detection benchmark," *Hum.-Centric Comput. Inf. Sci.*, vol. 12, pp. 1–21, May 2022.

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

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

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jul. 2015.

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

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[16] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[17] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[18] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[19] *Ultralytics/YOLOV5: V6.0*. Accessed: May 2020. [Online]. Available: https://github.com/ultralytics/yolov5
[20] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
[21] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[22] R. Belaroussi and J.-P. Tarel, "Angle vertex and bisector geometric model for triangular road sign detection," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Dec. 2009, pp. 1–7.
[23] H. Fleyeh, R. Biswas, and E. Davami, "Traffic sign detection based on AdaBoost color segmentation and SVM classification," in *Proc. Eurocon*, 2013, pp. 2005–2010.
[24] Y. Yang and F. Wu, "Real-time traffic sign detection via color probability model and integral channel features," in *Proc. Chin. Conf. Pattern Recognit.*, Heidelberg, 2014, pp. 545–554.
[25] J. Zhang, Z. Xie, J. Sun, X. Zou, and J. Wang, "A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection," *IEEE Access*, vol. 8, pp. 29742–29754, 2020.
[26] L. Cui, R. Ma, P. Lv, X. Jiang, Z. Gao, B. Zhou, and M. Xu, "MDSSD: Multi-scale deconvolutional single shot detector for small objects," 2018, *arXiv:1805.07009*.
[27] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 385–400.
[28] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1222–1230.
[29] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Jul. 2021, pp. 2778–2788.
[30] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2021, pp. 13713–13722.
[31] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 2117–2125.
[32] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Jun. 2019, pp. 8440–8449.
[33] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830.
[34] H. Zhang, H. Chang, M. Bingpeng, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2020, pp. 260–275.
[35] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14454–14463.
[36] H. Zhang, H. Zhang, L. Qin, J. Li, Y. Guo, Y. Zhou, and J. Zhang, "Real-time detection method for small traffic signs based on YOLOv3," *IEEE Access*, vol. 8, pp. 64145–64156, 2020.
[37] Z. Liu, D. Li, S. S. Ge, and F. Tian, "Small traffic sign detection from large image," *Appl. Intell.*, vol. 50, no. 1, pp. 1–13, Jan. 2020.
[38] K. Sun, Q. Wen, and H. Zhou, "Ganster R-CNN: Occluded object detection network based on generative adversarial nets and faster R-CNN," *IEEE Access*, vol. 10, pp. 105022–105030, 2022.
[39] C. Termritthikun, Y. Jamtsho, J. Ieamsaard, P. Muneesawang, and I. Lee, "EEEA-net: An early exit evolutionary neural architecture search," *Eng. Appl. Artif. Intell.*, vol. 104, Sep. 2021, Art. no. 104397.
[40] Q. Tang, G. Cao, and K.-H. Jo, "Integrated feature pyramid network with feature aggregation for traffic sign detection," *IEEE Access*, vol. 9, pp. 117784–117794, 2021.
[41] Y. Lu, J. Lu, S. Zhang, and P. Hall, "Traffic signal detection and classification in street views using an attention model," *Comput. Vis. Media*, vol. 4, no. 3, pp. 253–266, Sep. 2018.

YANLI SHI received the M.S. degree in basic mathematics from Beihua University, Jilin, China, in 2005, and the Ph.D. degree in basic mathematics from the Dalian University of Technology, Dalian, China, in 2017. Since 2005, she has been with the Jilin Institute of Chemical Technology, where she became an Associate Professor with the School of Science, in 2014. Her research interests include data mining and deep learning.

XIANGDONG LI received the bachelor's degree in automation from Henan Polytechnic University, Henan, China, in 2020. He is currently pursuing the M.S. degree with the Jilin Institute of Chemical Technology. His main research interests include deep learning and object detection.

MIAOMIAO CHEN received the bachelor's degree in automation from Henan Polytechnic University, Henan, China, in 2020. She is currently pursuing the M.S. degree with Harbin Engineering University. Her research interests include data mining and deep learning.