

An Improved YOLOv5s Algorithm for Traffic Sign Recognition

xia zhu (✉ snail2024@163.com)

Jinling Institute of Technology <https://orcid.org/0000-0003-0576-0677>

Jing Zhang

Nanjing University of Posts and Telecommunications

Bai Xu

Jinling Institute of Technology

Hongbo Zhu

Nanjing Institute of Posts and Telecommunications: Nanjing University of Posts and Telecommunications

Research

Keywords: traffic sign recognition, YOLOv5s, Depth separable convolution, Bidirectional characteristic pyramid

Posted Date: February 16th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-2200860/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

An Improved YOLOv5s Algorithm for Traffic Sign Recognition

Xia Zhu^{1,2}, Jing Zhang¹, Bai Xu², Hongbo Zhu¹

¹School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China

²Jinling Institute of Technology, Nanjing 211169, China

Corresponding author: Xia Zhu (e-mail: snail2024@163.com).

This work was supported by the National Key Research & Development Program of China under Grant No. 2020YFB1807202, the National Natural Science Foundation Program of China under Grant No. 92067201, the Jiangsu Provincial Key Research and Development Program under Grant No. BE2020084-1, Jiangsu AI Transportation Innovations & Applications Engineering Research Center.

Abstract: Aiming at the problems of excessive parameter quantity, slow detection speed, and low detection accuracy of existing traffic sign recognition models, this paper proposes a traffic sign recognition algorithm with improved YOLOv5s. The algorithm applies depthwise separable convolution to the feature extraction network of YOLOv5s, which significantly reduces the parameters and computation of the backbone network. In the feature fusion stage, the feature maps of different levels obtained by the feature extraction network are input into the weighted bidirectional feature pyramid network structure (CA-BiFPN) for multi-scale feature fusion. In the design process of the loss function, ZIoU is used instead of GIoU as the bounding box regression loss function to solve the problem of overlapping the real box and the prediction box. The Focal loss function replaces the binary cross entropy loss function, which solves the problem of an unbalanced number of positive and negative samples in the detection process. The test results on the TT100K dataset show that the average accuracy of the algorithm reaches 77 %, which is 4.94 % higher than that of YOLOv5s, and the model size is 16MB, which is only 56.7 % of YOLOv5s. The algorithm reduces the amount of calculation and model size to a certain extent and brings the improvement of detection speed and accuracy.

Keywords: traffic sign recognition; YOLOv5s; Depth separable convolution; Bidirectional characteristic pyramid

1. Introduction

As one of the practical applications of computer vision, traffic sign recognition has been widely used in the field of traffic safety, especially in the safety of unmanned driving. It is very important for highway sign and marking maintenance, driver assistance system, and autonomous vehicles. Traffic sign detection and recognition are divided into sign detection and sign classification. The purpose of the detection phase is to locate the region of interest in the image and verify whether the region of interest contains traffic signs; the purpose of the identification stage is to uniquely identify the traffic signs detected in the detection stage. In the face of factors such as weather conditions, light intensity and other environmental changes, the effect of many recognition methods is not ideal, which makes the current application research of traffic sign recognition still very challenging. Accurate and fast traffic sign recognition is the basis for unmanned driving to make the next judgment. Designing a traffic sign recognition method with higher accuracy and faster speed is of great significance to the unmanned driving system.

The traditional research methods of traffic sign recognition systems are mainly based on color

and shape. Cireşan et al. [2] used image preprocessing, color filtering, and color segmentation to detect traffic signs. In the recognition stage, feature extraction and trained neural networks are used to uniquely identify traffic signs. Soetedjo et al.[3] used the RGB color model to classify traffic signs according to color differences, and Dalal et al.[4] proposed a shape detection method using the directional gradient histogram. LIANG et al.[12] proposed a circular traffic sign detection and recognition method, which performs fine-grained classification of traffic sign images through a convolutional neural network with convolution and spatial transformation modules as the main modules. The above methods use the special color and shape of traffic signs to extract features and rely on classifiers for classification. However, the above methods have the problems of slow detection speed and low detection accuracy. It is difficult to achieve the expected goal in practical applications and is greatly affected by light and the environment.

In 2014, Girshick et al. [19] first proposed the RCNN algorithm. As an early deep learning algorithm, RCNN realized the localization and detection of objects. Then, to reduce the repeated operation of candidate regions, Girshick [21] proposed the Fast RCNN algorithm by adjusting the input of the network based on R-CNN, which shortened the inference time of RCNN to some extent. Ren et al. [24] proposed the Faster RCNN algorithm to realize the end-to-end network design by using the region recommendation network instead of the selective search in the candidate region extraction part. The Cascade R-CNN proposed by Cai et al. [5] improves the robustness of the algorithm by adjusting the different detection results by setting the increasing intersection ratio threshold through the cascade structure. Pang et al. [7] proposed Libra RCNN for balanced learning according to the problems of unbalanced positive and negative samples and unbalanced regression and classification loss functions in the detection process. The deep learning method provides a better tool for traffic sign recognition. Compared with traditional methods such as the color model analysis method, it can obtain a better recognition effect and higher recognition rate of traffic signs. Although the two-stage detection network can obtain a better detection result because the two-stage detection network relies on the candidate region extraction network to select the region of interest, the number of model parameters and the amount of calculation is too large, and the inference time is long, which does not meet the real-time requirements of traffic sign recognition.

To improve the detection efficiency, Redmon et al. [26] proposed a network YOLOv1 that uses the whole image as the input of the network and directly returns the position of the detection box and the category of the detection box in the output layer. Compared with the two-stage detection network, the YOLOv1 network has been applied to many fields because of its faster detection speed. However, due to the poor detection effect of YOLO on small targets, weak generalization ability, and large positioning error, the improved network of the YOLO series appeared subsequently, which greatly improved the detection accuracy of small targets. However, the above method has a large amount of calculation, which affects the detection speed, and also ignores the importance of shallow position information to small targets and occluded targets.

Because of the problems existing in the above methods, this paper proposes a traffic sign recognition algorithm MYOLOv5s (Mini YOLOv5s). To reduce the model parameters and calculation amount, the deep separable convolution is introduced based on the backbone network, the two-layer feature information fusion is changed into a three-layer feature fusion in the multi-scale feature fusion part, CA-BiFPN is used instead of the feature pyramid network to better detect targets of different sizes, improve the detection ability of the algorithm, and introduce the focal loss function to solve the problem of imbalance in the number of positive and negative samples in the

dataset so that the model can focus more on the detection of difficult samples. The experimental results show that MYOLOv5s significantly improves the speed and accuracy of traffic sign recognition while reducing model parameters.

2. Related work

2.1 One-stage inspection model

Given the slow reasoning speed of the two-stage network, Liu et al. [18] proposed the SSD algorithm in 2016, which is different from Faster RCNN in candidate region extraction. SSD sets a priori box at each pixel point, uses the feature map with a small receptive field to detect small objects, and uses the feature map with a large receptive field to detect large objects. Finally, the detection results are integrated to obtain the final prediction results. Subsequently, Redmon et al. [25] proposed the YOLOv1 algorithm to treat target detection as a regression problem. YOLOv1 uses a preset number of grid cells for classification and regression. Redmon then introduced batch normalization, dimension clustering, and the K-means algorithm to calculate grid cell size and proportion based on YOLOv1, which improved the speed and accuracy of YOLOv2 [26]. YOLOv3 [27] algorithm replaces the backbone network of YOLOv2 with Darknet-53, which increases the number of network layers. While the number of frames transmitted per second is reduced by half, the average accuracy is increased by 3.1%. Bochkovski et al. [28] added the cross-phase partial connection (CSP), path aggregation network (PAN), Mish activation function, mosaic data amplification, and other methods to YOLOv3 to introduce YOLOv4 algorithm, pushing the precision and speed of YOLO series algorithms to a new height. From YOLOv1 to now, YOLOv5 has developed into YOLOv5. YOLOv5 combines the advantages of previous versions and is superior in detection speed and accuracy. To some extent, YOLOv5 has become the State Of The Art of YOLO series algorithm.

There are four versions of the YOLOv5 network, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5s is the network with the smallest depth and feature map width among the detection networks of this series. Other versions of the network are deepened and widened on this basis. Compared with other series of YOLO detection networks, the weight file of YOLOv5 is relatively small. The YOLOv5s with the smallest weight is only 27MB, and the YOLOv5x with the largest weight is only 170MB. Therefore, based on its small size, YOLOv5s is very suitable for deploying embedded devices to detect traffic lights in traffic scenarios.

Compared with the two-stage network, the one-stage network has fewer model parameters and sizes, and its algorithm reasoning speed is faster. Although the detection accuracy has decreased, it can obtain better detection accuracy in a shorter reasoning time. The one-stage network is used for traffic sign recognition, and its detection speed is consistent with the real-time performance of traffic sign recognition. Based on the comprehensive consideration of weight file size, recognition accuracy, and detection speed, YOLOv5s with fast detection speed and high recognition accuracy are selected for research to improve their network structure.

2.2 MobileNetv3 feature extraction network

Google [9] proposed the MobileNet network in 2017. MobileNetv3 is a lightweight network, which is composed of deeply separable convolution, inverse residual network, and SE attention module. Its core algorithm is to replace standard convolution with deeply separable convolution, which can improve the training speed while ensuring the efficient use of parameters. Depth separable convolution can greatly reduce the computation of the network without affecting the accuracy. The number of channels of the traditional convolution kernel is always consistent with

that of the convolution image. Different from the traditional convolution, the depth separable convolution is to integrate the traditional convolution into two steps, in which each channel of the channel-by-channel convolution is only convolved by one convolution kernel, and one convolution kernel is only responsible for one channel in the feature map to extract the features inside a single channel. The feature map channel obtained after channel-by-channel convolution is consistent with the original feature map channel and then performs point-by-point convolution. The size of each convolution kernel of pointwise convolution is $C \times 1 \times 1$. Fusion the feature information between different channels to obtain a new feature map. The dimension size of the feature graph obtained from the depth separable convolution is the same as that of the standard convolution, but the computation amount and parameter amount of the depth separable convolution are less, which considers the parameter amount and computation amount of the control model while maintaining the performance of the algorithm and improves the efficiency of the algorithm. The structure is shown in Figure 1.

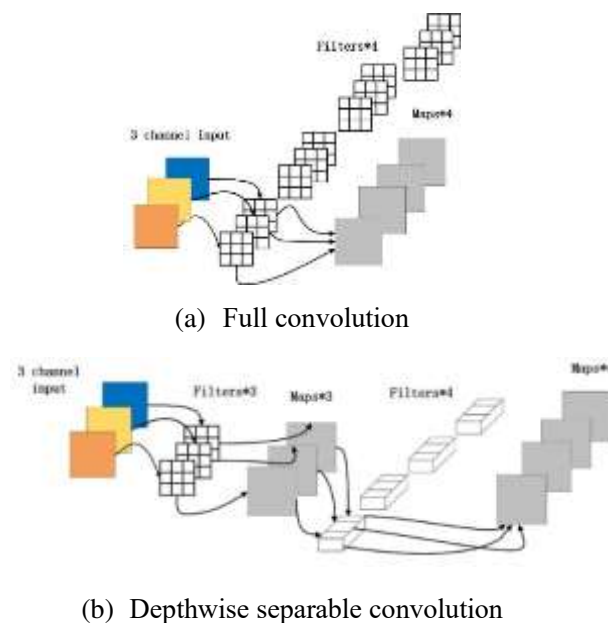


Fig.1 comparison of traditional convolution with depthwise separable convolution

It can be seen from Figure 1 that if the size of the input feature map is $D_H \times D_H \times M$, the size of the convolution kernel is $D_K \times D_K \times M$, the size of the output feature drawing is $D_H \times D_H \times N$, the parameter of the standard convolution layer is $D_K \times D_K \times M \times N$. Depth separable convolution is depth convolution and point-by-point convolution. Let the convolution kernel of depth convolution be $D_K \times D_K \times 1 \times M$, the convolution kernel of pointwise convolution is $1 \times 1 \times M \times N$, the depth convolution parameter is $D_K \times D_K \times 1 \times M$, the point-by-point convolution parameter is $1 \times 1 \times M \times N$. That is, the parameter of depth separable convolution is a traditional convolution $\frac{1}{N} + \frac{1}{D_K^2}$. In general, N is much larger than D_K . By 3×3 Convolution as an example,

the depth separable convolution requires about 1/9 of the parameters of the traditional convolution, and the amount of computation is only 1/9 of the original one. It can be seen that the depth separable convolution greatly reduces the number of operational parameters and the amount of computation. Therefore, the depth separable convolution used in this paper helps to better achieve real-time detection.

2.3 CA-BiFPN module

In the original YOLOv5 model, the path aggregation network (PANet) was used for feature fusion. Although the structure constructed a bottom-up and top-down two-way structure, the feature utilization between network levels was low, and too much redundant information appeared, leading to insufficient feature fusion of the model and loss of feature description of shallow edge small targets and area dense targets. So, Tan et al. [13] put forward the multi-scale feature fusion method of BiFPN, added jump connection in the traditional FPN, and replaced the FPN only with top-down feature fusion and bottom-up feature fusion so that the network can fuse more features of the same scale without adding additional computing parameters. Features of different scales can be more fully fused. For nodes with only one input, the contribution of this node to the feature network will be reduced. BiFPN effectively fuses features of different scales and adds information fusion of features of the same scale. Its structural design is shown in Figure 2.

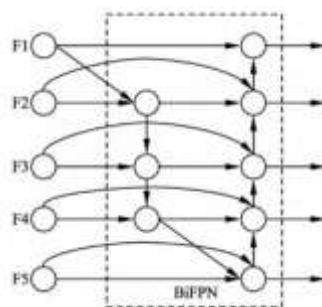


Fig.2 BiFPN structure design

Based on the BiFPN structure, this paper redesigns the channel-weighted bidirectional feature pyramid fusion network, as shown in Figure 3. This structure interlaces the features between different levels. The same level retains the original features, makes full use of shallow details and deep high-level semantic information, and reduces the calculation between nodes. The method of embedding coordinate attention (CA) is used to assign different scores of weights to the multi-scale channels of the BiFPN feature fusion module. Compared with the traditional method of randomly assigning weights, this weighting method can further increase the receptive field of the model, as well as the attention of interested targets and the position sensitivity of small targets.

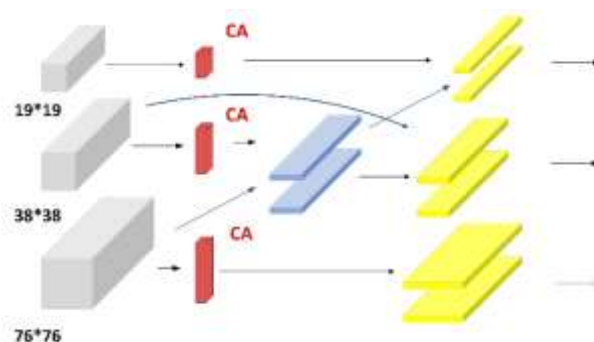


Fig.3 CA-BiFPN weighted feature fusion module

3. MYOLOv5s

MYOLOv5s uses MobileNetv3, which introduces deep separable convolution, as the feature extraction network, and uses CA-BiFPN to fuse the obtained three different scale feature maps for multi-scale feature fusion, regression and prediction of different scales, and detect targets of different sizes. In the loss function part, a ZIoU loss function is selected as the bounding box regression loss, and the focal loss function is selected as the confidence loss and classification loss.

3.1 MYOLOv5s network structure

The network structure of YOLOv5s is shown in Figure 4, in which the backbone network downsamples the input image 5 times to extract the features of the image, and the backbone network contains 4 modules: Focus, CBH, CSP1-x, and SPP. This paper improves the network structure of YOLOv5s from the following two points: (1) MobileNetv3 is applied to the feature extraction backbone network to reduce the number of parameters and calculations of the network; (2) In the feature fusion layer, the feature information in the three multi-scale channels is input into the improved attention-weighted feature pyramid fusion module (CA-BiFPN) to carry out the interactive fusion of shallow and deep detail information and advanced semantic information, and expand attention from same-layer fusion to cross-layer scenarios.

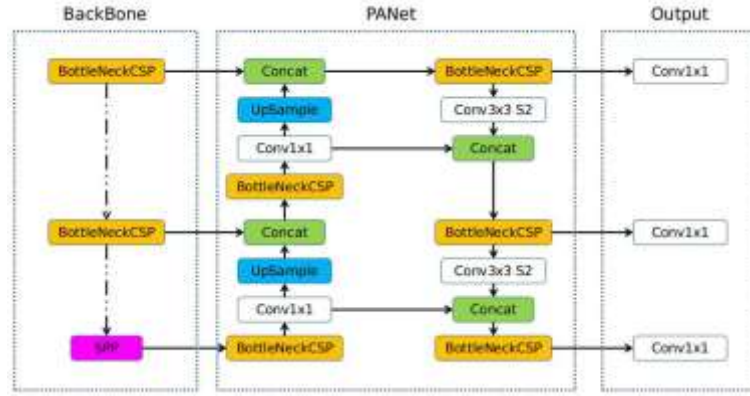


Fig.4 YOLOv5s network structure

The improved MYOLOv5s network structure is shown in Figure 5, which consists of three parts: the feature extraction network, the feature fusion part, and the object detection output. Firstly, three feature channels of different granularity information with scales of 19×19 、 38×38 、 76×76 are generated by the MobileNetv3 feature extraction network, and secondly, the feature information in the three multi-scale channels is input into the improved attention-weighted feature pyramid fusion module (CA-BiFPN) to carry out the interactive fusion of shallow and deep detail information and high-level semantic information, and finally perform object detection and output.

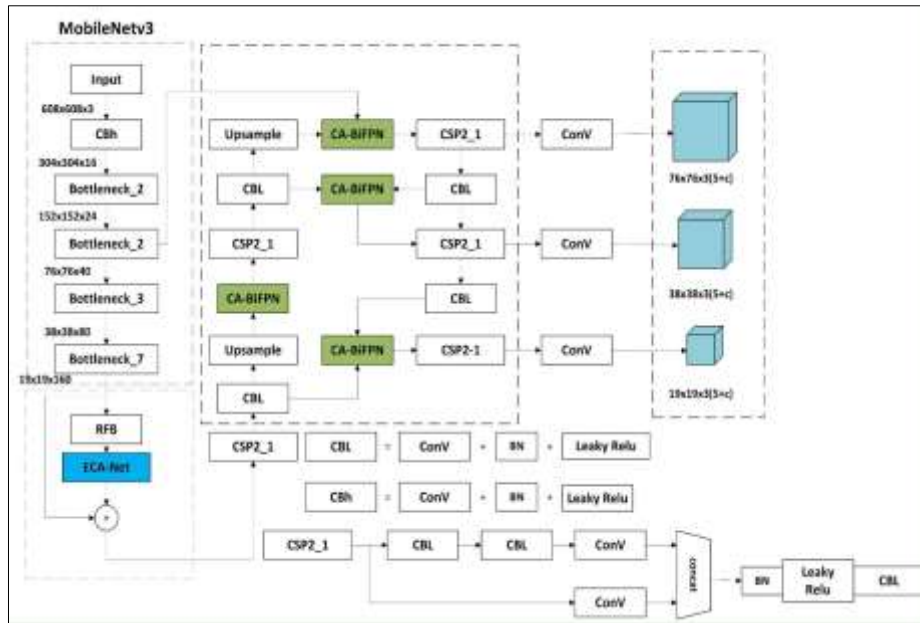


Fig 5. MYOLOv5s network structure

As shown in Figure 5, the feature map of 76×76 is convoluted with a convolution kernel size of 3×3 and a step size of 1 to obtain a feature map of size 38×38 and fuses it with the feature information of the original scale size. Then the feature map of 38×38 after fusion is convoluted in the same way to obtain the feature map of 19×19 and then fused with the feature information of the original scale size. Finally, the total fusion information is upsampled 2x and 4x, and the features of different scales are fused again. The CA-BiFPN module fuses the characteristics of the same layer scene with the cross-layer scene according to attention, including short-hop and long-hop connections, and even performs initial fusion inside it to obtain the high-resolution features of the target and improve the recognition accuracy of the fuzzy target.

3.2 Loss function

The original YOLOv5 loss function is shown in Equation (1), which consists of the bounding box regression loss, confidence loss, and class loss. The confidence loss and category loss are calculated using the binary cross entropy loss function, and the GIoU loss function is selected as the boundary box regression loss.

$$loss_{object} = loss_{loc} + loss_{conf} + loss_{class} \quad (1)$$

$$loss_{loc} = 1 - GIoU \quad (2)$$

GIoU is used in the original YOLOv5 to calculate the boundary box loss:

$$GIoU = IoU - \frac{|C - GT \cup P|}{|C|} = \frac{|P \cap GT|}{|P \cup GT|} - \frac{|C - GT \cup P|}{|C|} \quad (3)$$

GIoU introduces a minimum region, C, which contains exactly two regions, A and B, so that when the two do not intersect, the distance between them can be reflected by the area of C, as shown in Figure 6. Among them, the black rectangle represents the A area, the green rectangle represents the B region, the outermost red rectangle represents the C domain, and the blue part represents the intersection area of A and B.

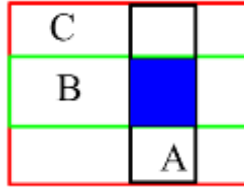


Fig 6 . Schematic diagram of GIoU

Different from the original IoU[31], GIoU not only focuses on the overlapping area between the real box and the prediction box, but also pays attention to other non-overlapping areas, so GIoU can better reflect the overlap between the two than the original IoU, but GIoU always only considers the overlap rate between the real box and the prediction box, and cannot describe the regression problem of the target box well. When the prediction box is inside the real box and the prediction box is the same size, the GIoU degenerates into IoU and cannot distinguish the position relationship between the prediction boxes. As shown in Figure 7, where the black rectangle is the target box (A) and the blue rectangle is the prediction box (B). When A completely includes B, the area of C is equal to the area of A and does not change, and the area of $(A \cup B)$ is also equal to the area of A, so the value of $C \setminus (A \cup B)$ is 0 and the value of GIoU is equal to the value of IoU. However, as can be seen from Figure 7, the middle graph predicts the best results, but the GIoU is 0.35, which does not reflect the advantages and disadvantages in this case.

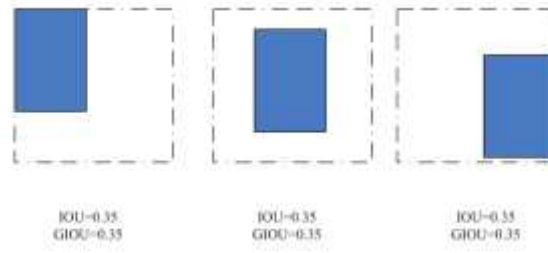


Fig 7. The value of GIoU and IoU when overlapped

Therefore, [29] uses CIoU instead of GIoU as the loss function of the target box regression and the loss function can be defined as

$$\text{CIoU} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (4)$$

$$\alpha = \frac{v}{(1 - \text{IoU}) + v} \quad (5)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (6)$$

Where α is an equilibrium parameter and does not participate in gradient calculation, v is a parameter used to measure the consistency of the aspect ratio. CIoU comprehensively considers the overlap rate, center point distance, and aspect ratio between the real box and the prediction box, so that the target box is more stable in the regression process and the convergence accuracy is higher.

GIoU and CIoU are designed to reduce the difference in center point distance and aspect ratio between the two boxes. But when the center points coincide, the two boxes do not necessarily coincide; When the center point coincides with each other, when the aspect ratio is the same, the two frames do not necessarily coincide, so it is not enough to reduce the difference between the center point distance and the aspect ratio, but also to further reduce the distance between the two corners of the upper left corner and the lower right corner of the two frames and the difference between the width and height dimensions of the two frames. Based on the above analysis, if the two boxes are close to each other and the two frames are of the same width, height, and size, it is more helpful to improve the degree of intersection of the two boxes. By studying the characteristics of the traffic sign dataset, this paper finds that most of the traffic signs in the dataset are

Based on CIoU, the distance between the midpoints of the upper and lower boundaries is taken into consideration, and the ratio of the sum of their distances to the square of the diagonal length of the minimum bounding box C is taken as a new penalty term, and a bounding box loss function ZIoU optimized for the traffic sign dataset is proposed as follows:

$$\text{ZIoU} = \text{IoU} - \delta \frac{\rho_1^2(b, b^{gt})}{c^2} - \beta \frac{\rho_2^2(b, b^{gt})}{c^2} - \gamma \frac{|w - w^{gt}|}{w^c} \times \frac{|h - h^{gt}|}{h^c} \quad (7)$$

$\rho_1^2(b, b^{gt}), \rho_2^2(b, b^{gt})$ is the Euclidean distance between the upper left corner and the lower right corner of the anchor box b^{gt} and the dimension box b . w^c, h^c respectively represent the length, width, and height of the diagonal of the minimum closure that can just wrap the anchor box and the label box. δ, β, γ are the weighting coefficients whose value range is $[-1, 1]$ and the initialization is 1.0.

To obtain more accurate classification accuracy, we replace the confidence and the dichotomous cross-entropy loss function of classification with the focal loss function[34]. Focal loss is improved based on cross-entropy loss, aiming at the problem of an unbalanced distribution of positive and negative samples in the classification process. Focal loss reduces the weight of simple background samples, so that the model focuses on the detection of foreground objects, avoiding the tendency of the model to predict a large number of background classes. The focal loss function is expressed as Equation (8).

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t) \quad (8)$$

$$p_t = \begin{cases} p & y = 1 \\ 1 - p & \text{others} \end{cases} \quad (9)$$

Wherein, α_t represents the weight of positive and negative samples, γ represents the weight of difficult and easy samples.

In this paper, we select ZIoU loss as the boundary box regression loss, and focal loss as the prediction confidence and category loss which speed up the prediction box regression speed, reduce the interference of positive and negative sample imbalance on the model, and improve the detection accuracy.

4. Experiment and analysis

4.1 Environment

Using Intel (R) Xeon (R) Platinum8358P processor, A40 graphics card, 56GB CPU memory, and 48G graphics card memory, based on the Darknet deep learning framework, the YOLOv5s network is optimized and trained separately under the Ubuntu 20.04 operating system. At the beginning of training, the initial learning rate is set to 0.01, the attenuation coefficient is set to 0.005, and the OneCycleLR mode is selected to update the learning rate. When the training iteration is completed, the final learning rate is reduced to 0.005 of the initial learning rate.

4.2 Data set

At present, there are several public traffic sign data sets abroad, among which the German traffic sign data set (GTSRB) and Belgian traffic sign data set (BTSD) have been widely used by many scientific research teams for many years. The German traffic sign recognition dataset GTSRB contains 43 categories. The training set contains 39209 traffic sign images, and the test set contains 12630 images. The Belgian traffic sign data set has 7095 images in 62 categories, including 4575 training data sets and 2520 test data sets. There are two publicly available traffic sign data sets in China: one is the Chinese traffic sign detection data set CCTSDB, and the other is the data set TT100K jointly produced by Tsinghua and Tencent. The CCTSDB dataset contains 55 types of traffic signs, and 5920 images in the dataset are selected. The TT100K dataset contains 100K images, including 10592 labeled images and 221 types of traffic signs.

Considering that the research in this paper is applied to the domestic traffic environment, and the German traffic signs are not completely consistent with the Chinese traffic signs, we focus on the domestic traffic sign dataset of the selected dataset. We analyze and process the TT100K dataset itself, and also use data augmentation and data enhancement to equalize the data set, to obtain a large-scale, high-quality small target dataset with Chinese characteristics traffic signs, and then label the equalized data set to obtain the corresponding annotation file of each picture, laying a good foundation for subsequent research.

The TT100K dataset contains 221 types of traffic signs, the training set contains 6107 pictures, and the verification set contains 3073 pictures. Street scenes under different light intensities and weather conditions are selected. The image resolution in TT100K is 2024×2024 , and the input image is cropped to 864×864 in this experiment. Given the long tail effect caused by the imbalance of the number of samples in different categories in the TT100K dataset, this paper expands the data of a small number of categories and uses methods such as adding random Gaussian noise, level flipping, and Mosaic-8 data enhancement to expand the small number of categories. The next experimental dataset in this paper selects images with low similarity and large differences in different scenes in the TT100K dataset for training, and after data enhancement and adding salt and pepper noise, a total of 8000 images are selected, of which 6000 are trained and 2000 are tested.

4.3 Result analysis and algorithm comparison

After MYOLOv5s introduces deep separable convolution into its feature extraction network, the network parameters and model size are significantly reduced compared with the original feature extraction network, and the comparison effect is shown in Table 1. As can be seen from Table 1, after the introduction of deep separable convolution, not only the number of parameters is reduced

to 66%, but also the mAP is improved by 2.1%, and the deep separable convolution effectively reduces the parameter redundancy of the original feature extraction network.

Table 1 Comparison of introducing depthwise separable convolution

Feature extraction network	SIZE /MB	mAP/%
Full convolution	24	82.8
Depthwise separable convolution	16	84.9

In the multiscale feature fusion part, the multi-scale feature fusion networks such as FPN, PANet, BiFPN are compared with the same feature extraction network and parameter settings, and the results are shown in Table 2. From Table 2, it can be seen that the traditional FPN structure only performs unidirectional feature fusion, and the prediction accuracy is low, while the PANet structure performs top-down and bottom-up bidirectional feature fusion, and introduces additional parameters, BiFPN structure adopts bidirectional feature fusion similar to PANet structure, and accumulates the feature information of different scales according to a certain weight, and achieves the highest accuracy and small number of parameters in the multi-scale feature fusion method.

Table2 comparison of multi-scale feature fusion networks

Multi-scale feature fusion networks	mAP/%	SIZE/MB
FPN	80.1	7.3
PANet	82.4	9.2
BiFPN	85.5	6.9

Table 3 shows the precision comparison under different bounding box regression loss functions. IoU loss comparison smooth L_1 loss can better establish the relationship between the four boundary points. GIoU loss better reflects the coincidence degree of the boundary box. CIOU takes into account the diagonal length of the bounding box center point, the length and width of the bounding box, and the diagonal length of the minimum enclosing rectangle. ZIOU takes the distance between the midpoints of the upper and lower boundaries as a factor based on CIOU, and takes the ratio of the sum of their distances to the square of the diagonal length of the minimum bounding box as a new penalty term.

Table 3 Comparison of different regression loss functions

Loss	AP ₅₀	AP ₇₅
smooth L_1	72.12	62.43
L_{GIoU}	74.70	66.28
L_{CIOU}	75.09	69.71
L_{ZIOU}	77.16	69.93

Finally, the average accuracy of MYOLOv5s network detection of traffic signs reached 0.774, the recall rate reached 0.63, and the accuracy rate reached 0.78. The time to test a single picture was 20 ms, and the average speed of detecting pictures was 33 frames/s. Compared with YOLOv5s, the average accuracy increased by 4 percentage points, the recall rate increased by 2 percentage points, and the accuracy rate increased by 1 percentage point. The time to test a single picture was reduced by 3 ms, and the detection speed increased by 5 frames/s. The overall target detection results under different networks are shown in Table 4.

Table4 Comparison of total object detection results

Algorithm	Time /h	GFLOPS	mAP@.5	mAP@.5:.95
YOLOv5s	0.116	114.6	0.849	0.524
MYOLOv5s	0.049	16.5	0.891	0.457

To better understand the internal network structure of the improved algorithm, and verify the improvement effect of different structural modules (Module 1: Depth Separable Convolution; Module 2: Bidirectional Pyramid Model; Module 3: Loss Function) on the detection performance of the algorithm, a comparative experiment of algorithm ablation was conducted. The first algorithm compared in the experiment is the original YOLOv5s algorithm; Algorithm 2 is the YOLOv5s algorithm that only introduces the MobileNetv3 module; Algorithm 3 is the YOLOv5s algorithm that only adopts CA-BiFPN; Algorithm 4 is the YOLOv5s algorithm that only introduces the ZIOU loss function module; Algorithm 5 is the MYOLOv5s algorithm. The above algorithms are trained on the TT100K dataset and analyzed in comparison, and to fairly compare different experiments, the parameter settings, and training steps are consistent, and the results are shown in Table 5.

Table5 Comparison of ablation test results

Algorithm	Model1	Model 2	Model 3	P/%	R/%
1				84.23	56.13
2	√			84.41	56.44
3		√		84.19	57.71
4			√	85.73	58.98
5	√	√	√	87.23	62.77

It can be seen from Table 5 that compared with Algorithm 5, the P and R values of Algorithm 2~4 have decreased to varying degrees. Algorithm 2 only introduces the MobileNetv3 module, and this algorithm cannot extract the rich detailed feature information of traffic sign targets; Algorithm 3 only introduces CA-BiFPN to replace the characteristic pyramid network. This algorithm can better detect targets of different sizes, improve the detection ability of the algorithm, and all performance indicators are improved compared with YOLOv5; In algorithm 4, the ZIOU loss function module is introduced to learn more accurate location information of traffic sign target anchor box, and its accuracy, recall, and average detection accuracy are improved compared with YOLOv5s algorithm, but each index still has a certain gap compared with algorithm 5. Finally, the P-value and R-value of algorithm 5 of all structural modules introduced in this paper reach 87.23% and 62.77%, respectively, which are significantly improved (all have an increase of 2-4 percentage points) compared with the several algorithms compared in the ablation experiment, which verifies that the proposed algorithm has superior detection performance.

To better prove the superiority of the proposed method, the proposed method is compared with other advanced deep learning-based object detection algorithms. The existing mainstream identification methods are Faster RCNN[24], VGG16[15], SDD[18], RetinaNet[34], etc., the results are shown in Table 6.

Table 6 Results comparison between the proposed method and other methods on the TT100K dataset

Method	TT100K	CCTSDB	GTSRB
Faster RCNN[24]	68.1%	70.2%	72.6%
SDD[18]	71.2%	76.3%	80.7%
VGG16[15]	76.4%	80.9%	85.8%
RetinaNet[34]	75.9%	81.8%	85.4%
YOLOv5s	78.7%	82.8%	85.5%
MYOLOv5s	83.5%	85.4%	86.4%

The accuracy of target recognition based on Faster RCNN and SDD methods has made a

breakthrough improvement in the field of high-resolution large target object detection in the public dataset Pascal VOC, which has been proven to be conducive to the extraction of complex image features. However, this series of detection models are not ideal for the detection of objects with small pixel areas, so it is not suitable for direct application to the detection task of small target objects. With the deepening of the number of neural network model layers, the recognition ability of the model to the data set is better, and the recognition rate of VGG16 and RetinaNet is more accurate, it can be seen that when the dataset used is the TT100K dataset, MYOLOv5s has a higher recognition rate than other methods, up to 83.5%, which is the highest recognition accuracy method in the current data set. At that time, when using the CCTSDB dataset and GTSRB dataset, we found that the accuracy of each method was improved by about 2-4 points, because the number of classifications in the CCTSDB dataset and GTSRB dataset was much less than that of TT100K, and there was not much interference.

4.4 Qualitative analysis

To compare the detection performance of the verification algorithm more intuitively, the detection results of MYOLOv5s and YOLOv5s on the TT100K dataset are qualitatively compared and analyzed, as shown in Figure 8. According to Figure 8(a), YOLOv5s has the problem of false detection and missed detection of traffic signs in the detection results, and MYOLOv5s solves the above problems and shows good detection performance



(a) Detection effect of the YOLOv5s



(b) Detection effect of the MYOLOv5s

Fig.8 Results comparison between the YOLOv5 and the proposed method

As shown in Figure 8, the picture in Figure 8(a) missed detection, and the traffic signs in No. 113, No.109, and No.107 were missed. In Figure 8(b), MYOLOv5s implemented the correct positioning and classification of all signs in the data sample, including the partially obscured sign on the right in picture No. 109, proving that the ZIoU loss function can optimize the accuracy of the algorithm for the anchor box positioning of traffic signs. To sum up, according to the comparative analysis of 16 groups of test results under different image enhancement processing, it is verified that MYOLOv5s can use its own network's good feature extraction ability and target box positioning accuracy to solve the problems of false detection and missing detection in the traditional detection network, and it is verified that the algorithm in this paper has good generalization and robustness and can better complete the detection task of road traffic signs.

The detection effect in the actual environment is shown in Figure 9, from which it can be seen that the original network has many traffic signs that have not been detected (line 2), while the improved network (line 1) has a better detection effect and higher recognition accuracy for traffic signs. It can be seen that the improved network is more suitable for the detection and recognition of traffic signs.



Fig.9 Comparison of detection effect in the actual environment

5. Conclusion

In this paper, an improved traffic sign recognition method MYOLOv5s is proposed to solve the problems of too large model parameters, slow recognition speed, and low recognition accuracy in the existing traffic sign recognition methods. Aiming at the problem that the YOLOv5s network model is relatively large, the MobileNetv3 backbone network is applied to reduce the model parameters and accelerate the recognition speed. The CA-BiFPN structure is used in the multi-scale feature fusion part to strengthen the deep and shallow feature information fusion and obtain more same-level feature information. The focal loss is used as the confidence and classification loss to improve the detection ability and accuracy of the model for difficult samples. Through the verification and analysis of the experimental results, it can be seen that the model of MYOLOv5s is only 56.7 % of YOLOv5s, and the mAP is increased by 4.94 percentage points. This method is superior to the YOLOv5s in model size, detection accuracy, and detection speed and can meet the speed requirements of traffic sign detection. Moreover, the model is lightweight and easier to deploy in driving vehicles, which has practical application value for the rapid development of autonomous driving.

In this paper, an improved traffic sign recognition method MYOLOv5s is proposed, which aims to solve the problems of excessive model parameters, slow recognition speed and low recognition accuracy in existing traffic sign recognition methods. MYOLOv5s aims at the relatively large problem of YOLOv5s network model, applies MobileNetv3 backbone network to reduce model parameters, speed up recognition, adopts CA-BiFPN structure in the multi-scale feature fusion part, strengthens the fusion of deep and shallow feature information and obtains more feature information at the same level, takes the focal loss as the confidence and classification loss, and

improves the detection ability and accuracy of the model for difficult samples. Through the verification and analysis of experimental results, it can be seen that the model size of MYOLOv5s is only 56.7% of YOLOv5s, the mAP increased by 4.94 percentage, the method is better than YOLOv5s in model size, detection accuracy and detection speed, can meet the rate requirements of traffic sign detection, and the model is lightweight, more convenient to deploy in driving vehicles, and has practical application value for the rapid development of automatic driving.

Abbreviations

- 1.) YOLOv5s : you only look once version 5 (model is small)
- 2.) CA-BiFN: coordinate attention bidirectional feature pyramid network
- 3.) TT100k: Tsinghua-Tencent 100K Tutorial
- 4.) RCNN: Rich Convolutional Neural Network
- 5.) MYOLOv5s: Mini YOLOv5s
- 6.) SSD: Single shot multibox
- 7.) CSP: Cross-phase partial connection
- 8.) PAN:Path aggregation network
- 9.) FPN:Feature Pyramid Network
- 10.) IoU:Intersection over Union
- 11.) GIoU: Generalized Intersection Over Union
- 12.) DIoU: Distance Intersecton Over Union
- 13.) GTSRB: German traffic sign data set
- 14.) BTSD: Belgian traffic sign data set
- 15.) CCTSDB: Chinese traffic sign detection data set

Availability of data and material

The datasets used for testing in this article are all open source datasets

Competing interests

Not applicable

Acknowledgements

I extend my sincere thanks to Dr. Xinyun Yan for her valuable suggestions and to improve the quality of the paper.

Author contributions

Contribution of the first author: designed, developed and Python implementation of the three modules they are (i) new CA-BiFPN Module (ii) ZIoU algorithm . Contribution of the second author: (i) discussed the literature survey and its outcomes . Contribution of the third author: He helped me test the data and compare different scenarios. Contribution of the forth author: Discuss and correct the test scenario.Both authors read and approved the fnal manuscript.

Funding

This work was supported by the National Key Research & Develoment Program of China uder Grant No. 2020YFB1807202, the National Natural Science Foundation Program of China under Grant No. 92067201,the Jiangsu Provincial Key Research and Development Program under Grant No. BE2020084-1.Jiangsu AI Transportation Innovations & Applications Engineering Research Center.

Reference

- [1] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012. 2
- [2] D. Cireşan, U. Meier, J. Masci and J. Schmidhuber. A committee of neural networks for traffic sign classification. The 2011 International Joint Conference on Neural Networks, 2011, pp. 1918-1921, doi: 10.1109/IJCNN.2011.6033458.
- [3] Soetedjo, Aryunto and Koichi Yamada. An Efficient Algorithm for Traffic Sign Detection. J. Adv. Comput. Intell. Informatics 10 (2006): 409-418.
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886-893 vol. 1.
- [5] Z. Cai and N. Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154-6162.
- [6] Y. Liu, H. Liu, J. Fan, Y. Gong, Y. Li, F. Wang, and J. Lu. A Review of Research and Application of Small Target Detection Based on Deep Learning. Chinese Journal of Electronics, vol. 48, no. 03, pp. 590–601, 2020.
- [7] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang and D. Lin. Libra R-CNN: Towards Balanced Learning for Object Detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 821-830.
- [8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014.
- [9] Howard, Andrew G. et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv abs/1704.04861 (2017): n. pag.
- [10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. Proc. IEEE Conf. computer. Vis. Pattern Recognition., pp. 248-255, 2009.
- [11] Bin Yan, Pan Fan, Xiaoyan Lei, Zhijie Liu, and Fuzeng Yang. A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. Remote Sensing 2021, Vol. 13, Page 1619, 13(9):1619, apr 2021.
- [12] LIANG Min jian et al. Traffic sign recognition methods based on HOG-Gabor feature fusion and Softmax classifier [J]. Journal of Traffic and Transportation Engineering, 2017, 17(3):151-158.
- [13] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 10778–10787, nov 2019.
- [14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.
- [15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al. Going deeper with convolutions. arXiv preprint arXiv: 1409, pp. 4842, 2014.
- [16] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside outside net: Detecting objects in context with skip pooling and recurrent neural networks. In CVPR, 2016.

- [17] J.Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In NIPS, 2016.
- [18] W.Liu, D. Anguelov, D. Erhan, C. Szegedy et al. SSD: Single Shot Multibox Detector. European Conference on Computer Vision. Springer, 2016, pp. 21–37.
- [19] R.Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
- [20] R.Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In CVPR, 2015.
- [21] R. Girshick. Fast R-CNN. In ICCV, 2015.
- [22] K.He, X.Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV. 2014.
- [23] K. He, X.Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
- [24] S.Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
- [25] K.He, G.Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386–397, 2020.
- [26] J.Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
- [27] J.Redmon, A. Farhadi. YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6517–6525.
- [28] J.Redmon, A. Farhadi. YOLOv3: An Incremental Improvement. CoRR, vol. abs/1804.02767, 2018.
- [29] Alexey Bochkovskiy, Chien-Yao Wang, and Hong Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934, 2020.
- [30] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. AAAI, 2020, pp. 12993–13000.
- [31] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid and S. Savarese. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 658-666.
- [32] Howard, Andrew G. et al. Searching for MobileNetV3. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 1314-1324.
- [33] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. CoRR, abs/1312.2249, 2013.
- [34] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In CVPR, 2017.
- [35] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In CVPR, 2017.
- [36] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318-327, 1 Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.
- [37] ZHANG Hongyi et al. Mixup: beyond empirical risk minimization [EB/OL]. (2018-04-

- 27).<http://arxiv.org/abs/1710.09412>.
- [38] RAJENDRAN S P ,SHINE L et al. Real-time traffic sign recognition using yolov3 based detector[C]//IEEE.10th International Conference on Computing ,Communication and Networking Technologies. New York: IEEE, 2019:1-7.
 - [39] STALLKAMP J et al. The German traffic sign recognition benchmark: A multi-class classification competition[C]// IEEE.2011 International Joint Conference on Neural Networks. New York: IEEE, 2011:1453-1460.
 - [40] Xu Yan et al. Traffic Sign Recognition Algorithm Combining Weighted ELM and AdaBoost [J].Small Microcomputer Systems, 2017, 38(09):2028-2032.
 - [41] ZHANG Huibing et al. Real-time detection method for small traffic signs based on YOLOv3 [J]. IEEE Access, 2020, 8:64145-64156.
 - [42] WU Yi-qiang et al. Real-time traffic sign detection and classification towards real traffic scenes [J].Multimedia Tools and Applications, 2020, 79(25/26):18201-18219.
 - [43] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez,H. Gomez-Moreno, and F. Lopez-Ferreras. Road-sign detection and recognition based on support vector machines. Intelligent Transportation Systems, IEEE Transactions on, 8(2):264–278, June 2007.
 - [44] Li, Yongjun et al. YOLO-ACN: Focusing on small target and occluded object detection. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3046515.
 - [45] Guo Lei, Wang Qiulong, Xue Wei, Guo Ji. Small object detection algorithm based on improved YOLOv5 [J]. Journal of University of Electronic Science and Technology of China, 2020, 51(02):251-258.
 - [46] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 658–666.
 - [47] GAO Tao et al. Traffic sign detection algorithm based on pyramid multi-scale fusion [J].Journal of Traffic and Transportation Engineering, 2022, 22(3):210-224.
 - [48] J. Jin, K. Fu, and C. Zhang. Traffic sign recognition with hinge loss trained convolutional neural networks. IEEE Transactions on Intelligent Transportation Systems, 15(5):1991–2000, 2014.
 - [49] Luis Perez, Jason Wang.The Effectiveness of Data Augmentation in Image Classification using Deep Learning. 2017.
 - [50] P. Li, H. Wang and X. Huang.Wheat straw target detection algorithm based on improved YOLOv5.2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS), 2022, pp. 380-385.
 - [51] Guanhao Yang, Wei Feng, Jintao Jin, Qujiang Lei, Xiuhao Li, Guangchao Gui, and Weijun Wang. Face Mask Recognition System with YOLOV5 Based on Image Recognition. 2020 IEEE 6th International Conference on Computer and Communications, ICC3 2020, pages 1398–1404, dec 2020.
 - [52] Fangbo Zhou, Huailin Zhao, and Zhen Nie. Safety Helmet Detection Based on YOLOv5. Proceedings of 2021 IEEE International Conference on Power Electronics, Computer Applications, ICPECA 2021, pages 6–11, jan 2021.