Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 2, 2024

Improved traffic sign detection algorithm based on improved YOLOv8s

Xin He, Tianyu Li, and Yiqi Yang

School of Mathematics and Science, Southwest University of Science and Technology, China

Abstract: Aiming at the problem of low accuracy in current traffic sign detection, a new traffic sign detection model YOLOV8-CO with high accuracy was proposed based on YOLOv8 algorithm in this paper. CCMA attention mechanism was introduced and C2f module was replaced by C2fO module. The global average pooling strategy was used to replace the traditional fully connected layer, and the corresponding relationship between feature maps and categories was forced to better conform to the convolutional structure. CIoU and YOLOv8 original loss function were selected to calculate the loss value and train the model. The detection accuracy of traffic signs is improved effectively and overfitting is avoided. TT100K data set and COCO data set were used to verify the validity of the algorithm. The experimental results show that on TT100K, mAP@0.5 increases by 2.2%, mAP@0.7 increases by 2.4%, mAP@0.5:0.95 increases by 1.6%; Improved 0.7% on the COCO dataset, mAP@0.7 improved 0.6%, and mAP@0.5:0.95 improved 0.7%.

Keywords: Traffic sign; Target detection; Attention mechanism; IoU.

1. Introduction

Traffic signs are a kind of static information that vehicles often encounter when driving. They regulate the driving behavior of vehicles and ensure smooth road traffic. Many traffic rules are enforced based on traffic signs, Therefore, it is necessary to obtain traffic sign information completely and quickly for driving.

Research on traffic sign detection began in the 1970s. Due to limitations in hardware levels, it was unable to provide a good computational basis for the algorithm and it was difficult to conduct experimental verification. Therefore, research on traffic sign detection stagnated. However, with the rapid development of science and technology in recent years, the computing power and computing speed of computers have been greatly improved. More and more researchers and automobile manufacturers are committed to the study of traffic sign detection [1]. At this stage, the mainstream target detection algorithms can be roughly divided into two categories: traditional target detection methods and deep learning-based target detection methods [2].

With the in-depth research of domestic and foreign scholars in the field of traffic sign detection, traffic sign detection systems have developed rapidly [3]. The current research on traffic sign detection has gradually developed from simple methods and means such as simple background recognition and color segmentation to the ability to realize traffic sign detection under complex backgrounds and road conditions [4]. Major automobile manufacturers are also trying to integrate traffic sign detection into assisted driving systems, which has high application value. However, due to the problems in the actual driving environment such as complex road conditions, fast driving speeds, defects and occlusions in traffic signs, large changes in light intensity, and difficulty in detecting small targets, it is difficult for the traffic sign detection system to achieve ideal results in actual driving scenarios. [5].

In response to the problem of fast driving speed, this paper selects the YOLOv8n model with the smallest number of parameters and the fastest detection speed in the YOLOv8 algorithm as the basic network structure, and makes improvements on this basis. In order to solve the problem of difficult to distinguish traffic signs with complex road conditions, this paper adopts the CCMA attention mechanism to improve model performance by capturing cross-channel information and also capturing direction-aware and position-sensitive information. In response to the training cost issue, this article also uses the OREPA module to integrate with C2f to compress the training cost.

2. YOLOv8 Overview

As of the time of writing this article, YOLOv8 is the latest one-stage object detection algorithm in the YOLO series. Compared with other mainstream target detection algorithms, it has faster speed, higher accuracy, and better overall performance, reaching the State of the Art (SOTA) level in many tasks. YOLOv8 currently has 5 versions, including YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Taking into account the computing speed, real-time performance, accuracy and the equipment I use, this article selected YOLOv8n as the benchmark and made improvements on this basis.

The structure of YOLOv8 contains four main parts: input, backbone, neck and head, as shown in Figure 1. The input receives an image input of size 640x640. In the backbone part, improvements were made based on YOLOv5, the original 6x6 convolution was changed to 3x3 convolution, and the C3 module was replaced by the C2f module with reference to the idea of YOLOv7 ELAN. In the neck part, the 1x1 convolution sampling layer is removed, and the C2f module is used to replace the C3 module. In the head part, the structure of the decoupled head is used to decouple the classification task and the regression task. During the training process, Mosaic data enhancement is used to preprocess the images. It is worth noting that in the last 10 epochs, the Mosaic data enhancement operation was turned off, which can effectively improve the detection accuracy [6].

Among them, the Conv module is a composite module composed of Conv2d (two-dimensional convolution), BN (batch normalization) and SiLU (Sigmoid-Linear Unit). The specific structure is shown in Figure 2. The convolutional

layer extracts feature information by convolving the input data by applying a set of learnable filters (also called convolution kernels or convolution matrices). These filters have different feature extraction capabilities and can effectively capture the edges, shapes and other features of the input data [7].

SPPF (Serial Parallel Pooling Fusion) is a method proposed based on SPP (Spatial Pyramid Pooling), aiming to expand the receptive field. It is implemented by serializing multiple 5*5 max pooling layers, thereby reducing the number of parameters and significantly reducing the amount of calculation. The SPPF module accepts the feature map as input, which is then processed by the ConvBNSiLU module and performs a maximum pooling downsampling operation [8]. Finally, different downsampling results are spliced to form an output feature map. The structure of this module is shown in Figure 3. By using SPPF, problems such as image distortion caused by cropping and scaling operations of the image area can be effectively avoided. This method solves the problem of extracting relevant repetitive features in images by convolutional neural networks, and has a lower amount of parameters and computational load.

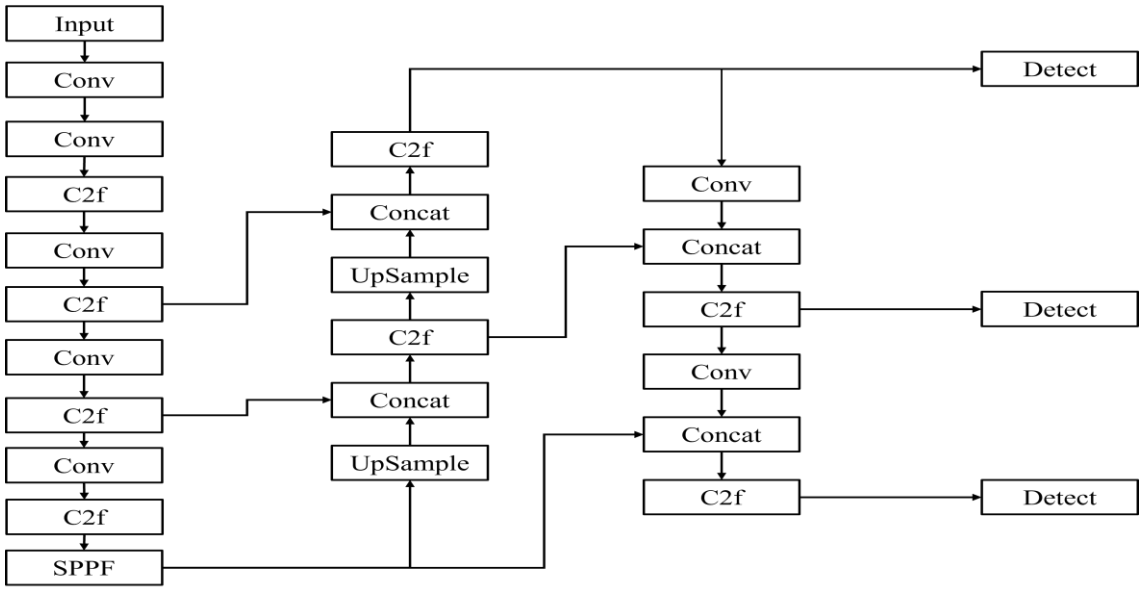

Figure 1. YOLOv8 framework diagram

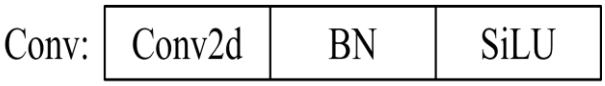

Figure 2. Conv structure diagram

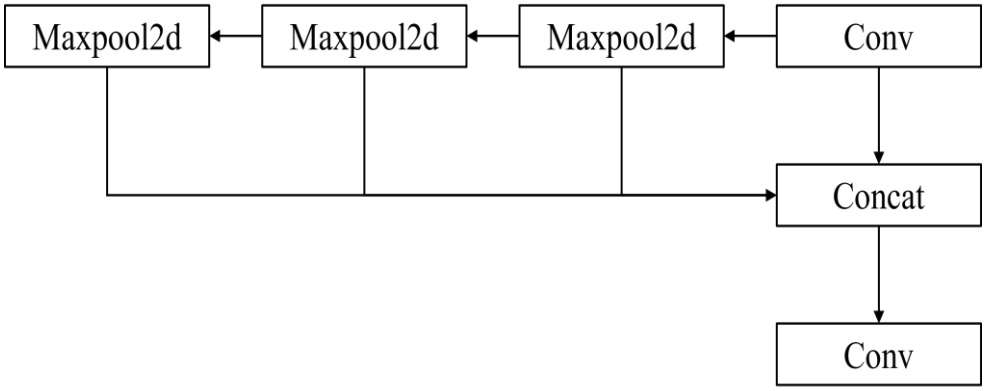

Figure 3. SPPF structure diagram

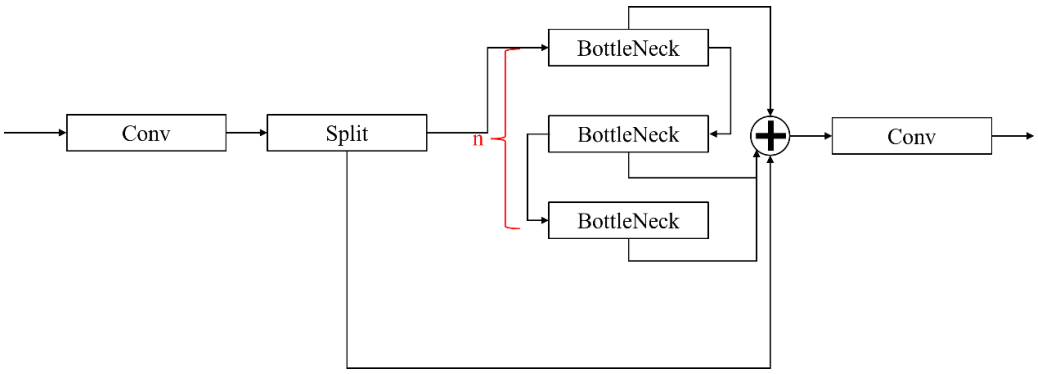

Figure 4. C2f structure diagram

3. The improved algorithm YOLOv8-CO in this article

This paper improves on the model YOLOv8n and proposes a network model suitable for traffic sign detection, namely YOLOv8-CO. First, by introducing the attention mechanism CCMA and replacing the C2f module with the C2fO module, the network's detection accuracy of traffic signs is improved. The network structure of the model is shown in Figure 5.

The attention mechanism is used to enhance the model's attention and weight distribution to the input data. By automatically learning the correlation and importance in the data, the model can process different inputs in a more targeted manner. It enables the model to focus on the most relevant parts of the input by dynamically calculating the weights between elements [9].

Before introducing CCMA attention, we first introduce Coordinate Attention. This article is modified from Coordinate Attention to meet the purpose of this article. Coordinate Attention has the following advantages. First, it not only captures cross-channel information, but also captures direction-aware and position-sensitive information, which helps the model locate and identify objects of interest more accurately. Second, the method is flexible and lightweight. can be easily plugged into the classic building blocks of mobile networks to enhance features by emphasizing information representation. Third, as a pre-trained model, our coordinated attention can bring significant performance improvements to downstream tasks of mobile networks, especially for Those tasks with intensive pre-processing [10]. We have improved on the basis of Coordinate Attention to make this improved attention module more in line with the specific goals of this article.

On the basis of the original, we have made two major improvements:

(1) Convolve again after the last convolution split into X and Y, and split again after activation to obtain X Weight and Y Weight. These two weights are multiplied with X and Y respectively, making the result more refined.

(2) We added an additional branch, first performing global average pooling on the feature map to force the correspondence between the feature map and the category to make it more consistent with the convolutional structure [11], This helps prevent overfitting. Then the global average pooling result is averaged and multiplied by the feature map before splitting into X Weight and Y Weight. This step realizes the fusion of contextual features, and effectively alleviates the information loss caused by the reduction in the number of channels through the fused feature map. This method can better retain multi-scale contextual information, thereby improving detection accuracy.

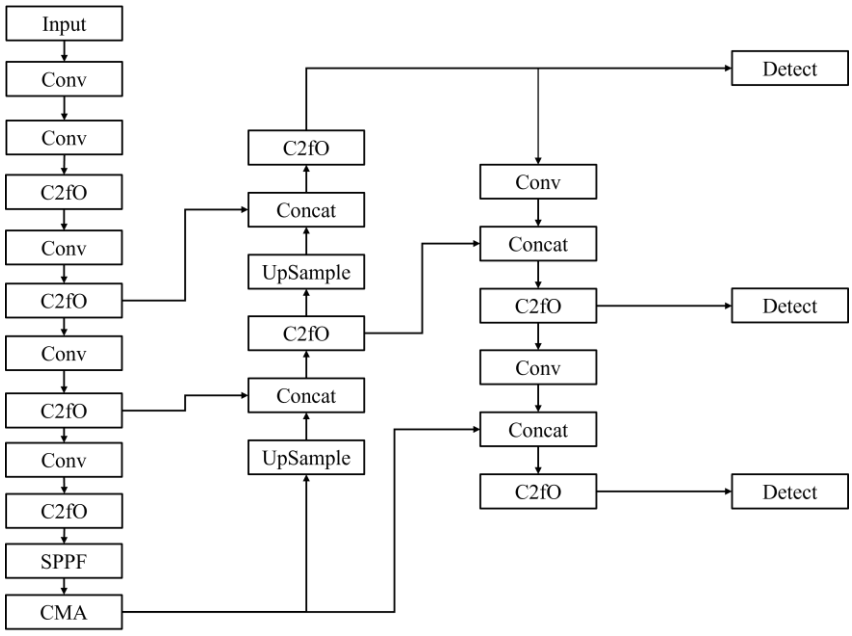

Figure 5. YOLOv8-CO framework

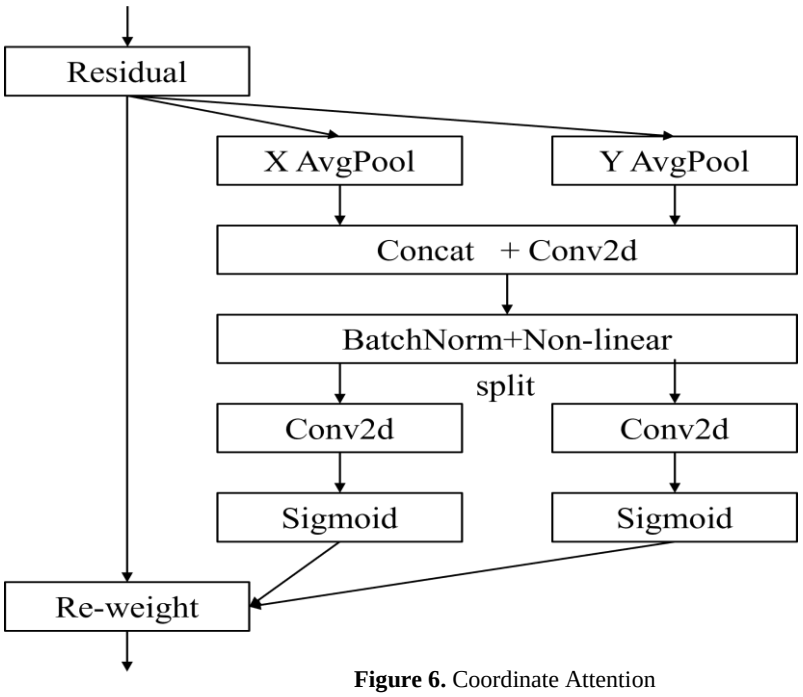

Figure 6. Coordinate Attention

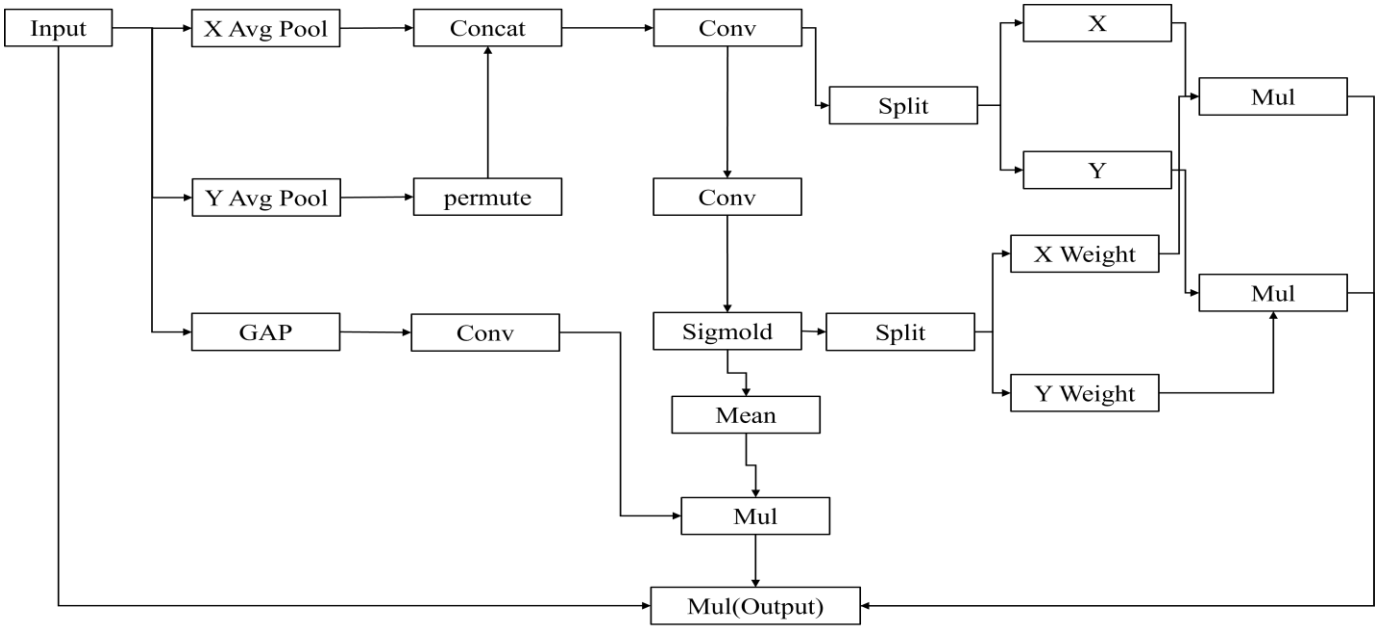

Figure 7. CCMA structure diagram
Note: GAP (Global Average Pooling)

Structural reparameterization has received increasing attention in various computer vision tasks. It aims to improve the performance of deep models without introducing any inference time cost. Although highly efficient during inference, this model relies heavily on complex training time blocks to achieve high accuracy, resulting in substantial additional training costs [11]. In order to effectively solve this problem without much change in accuracy, this article introduces OREPA proposed by Mu Hu et al. in the literature [12] in 2022 to improve the C2f module and combine it into C2fO.

OREPA has a 2-step process. First, during the block linearization process, the OREPA author deleted the nonlinear module of the original module, and added the Scaling module to ensure the same dimensions. During the block compression process, OREPA authors merge blocks into a single convolutional layer (OREPA Conv). Through these steps, the authors experimentally demonstrate that OREPA significantly reduces training costs while maintaining high performance. OREPA result structure is shown in Figure 8.

The C2f module is designed with reference to the ideas of the C3 module and ELAN, allowing YOLOv8 to obtain richer gradient flow information while ensuring lightweight. The current five versions of YOLOv8 are distinguished based on the network depth and width of YOLOv8. The depth and width in the C2f module are mainly determined by changing the number of BottleNecks, so BottleNeck plays an important role in C2f. By modifying C2f (Figure 4) Improve the BottleNeck module in the module, replace the Conv in BottleNeck with OREPA, reduce the training cost of adding

the attention mechanism without reducing the accuracy, so that the advantages of OREPA can be transferred to C2f, further enhancing the performance and practicality of the C2f module, reducing training costs without reducing model accuracy. The improved BottleNeck is shown in Figure 9.

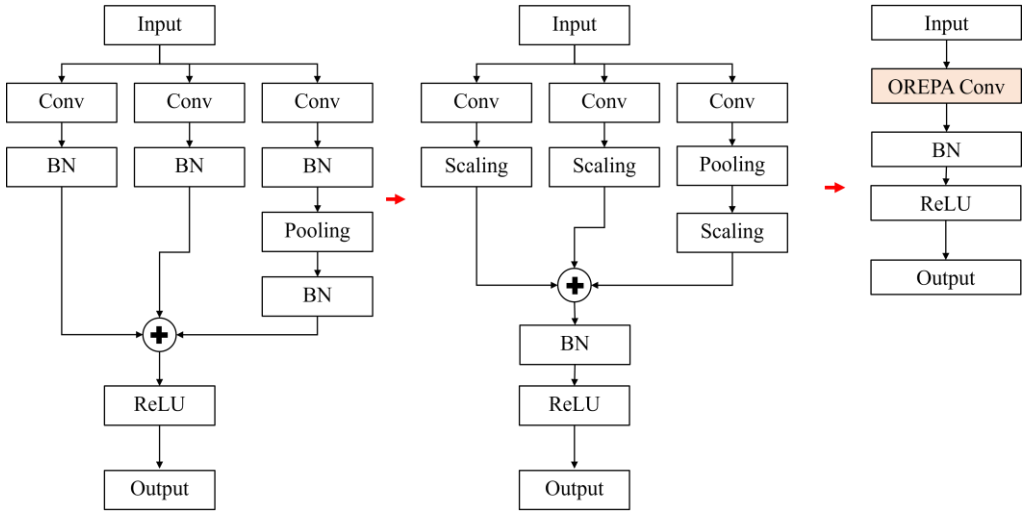

Figure 8. OREPA structure diagram

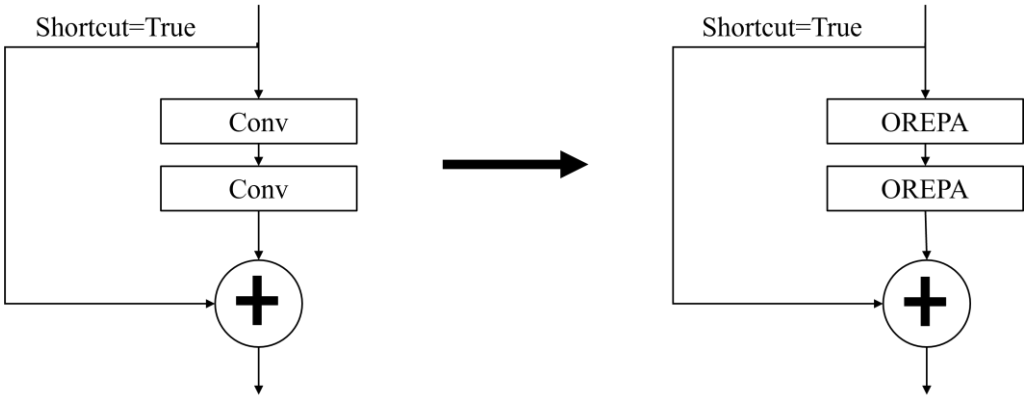

Figure 9. BottleNeck before and after comparison

4. Experimental results and analysis

4.1. Experimental environment

Table 1. The parameter settings of this experiment

name	Environmental parameters
operating system	Windows 10
CPUmodel	lntel(R) Xeon(R) W-2123 CPU @ 3.60GHz
GPUmodel	NVIDIA GeForce RTX 2080 SUPER
Memory size and frequency	64G(2666MHz)
Python version	3.11.5
Pytorch version	2.1.0

The parameter settings of this experiment are as follows: Batchsize is 16, epochs is 150 rounds, the optimizer is SGD, and Mosaic data enhancement is turned off for 10 rounds of training.

4.2. IOU selection

The loss function of YOLOv8 consists of bounding box regression loss and other losses. The bounding box regression loss of the YOLOv8-CO network model constructed in this article is calculated using CIoU and its formula.

The IoU loss function treats the four bounding boxes as a whole, and calculates the intersection and ratio between the predicted box A and the real box B, and then subtracts this value from 1. The final calculated loss value is that A and B do not intersect. The ratio of the part to the union of A and B represents the detection effect of the prediction frame, as shown in the formula below:

$$L_{IoU} = 1 - IoU(A,B) \quad IoU = \frac{A \cap B}{A \cup B}$$

According to the formula above, when the predicted box A and the real box B do not overlap, that is, when IoU(A,B) = 0, the distance between the box A and the box B cannot be measured based on the calculation results. And in the process of training the network, when loss=0, the loss function loses its role and cannot return the gradient, let alone optimize the network parameters. At the same time, relying only on IoU cannot accurately reflect the actual overlap of the two boxes, as shown in Figure 10. It is assumed that the IoU values of these three pictures are the same, but the regression effect displayed is very different. The difference can be clearly seen that the regression effect on the far right is the best and the one on the left is the worst.

Since IoU is a ratio concept, it is insensitive to the scale of the target object. However, in the target detection task, Bounding Box's regression loss optimization and IoU optimization are not completely equivalent. In addition, the Ln norm is sensitive to the scale of the object. The proposal of GIoU alleviates the gradient problem of IoU loss when the detection frames do not overlap. The penalty term IoU loss is calculated, and the GIoU loss function is as follows:

Among them, A represents the real frame, B represents the predicted frame, and C represents the smallest border that can surround frame A and frame B. The specific relationship between the three is shown in Figure 11.

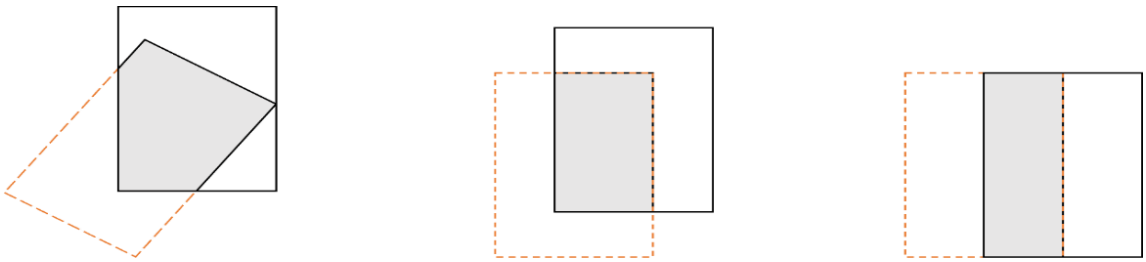

Figure 10. Different regression effects of the same IoU

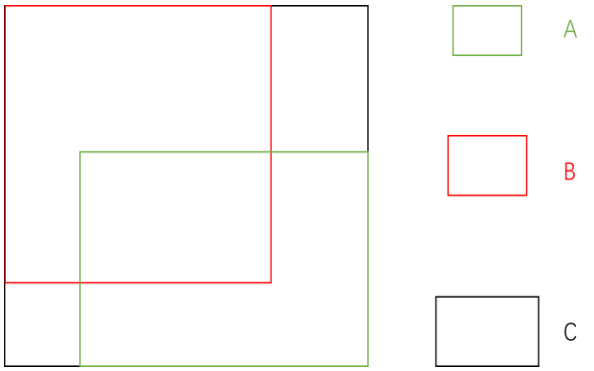

Figure 11. The specific relationship between the three

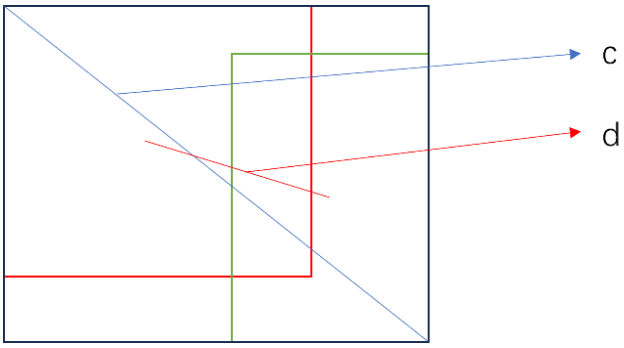

Figure 12. The predicted box and the true box relationship

$$L_{GIoU} = 1 - IoU(A,B) + |C - A \cup B| / |C|$$

GIoU as a loss function will have the following problems. The prediction box is translated and scaled based on the prior box. GIoU first expands the prediction frame so that it can overlap with the real frame, and then calculates IOU(A, B), which will undoubtedly consume a lot of time and reduce the convergence speed of the loss function.

In order to speed up the convergence, DIoU adds a penalty term to the IoU loss, which is designed to minimize the standardized distance between the center points of the two detection frames. The DIoU loss function is as follows:

$$L_{DIoU} = 1 - IoU(A,B) + \frac{\rho^2(b, b^{gt})}{c^2}$$

In order to speed up the convergence speed, DIoU adds a penalty term based on the IoU loss, which is designed to minimize the standardized distance between the center points of the two detection frames. CIoU introduces the aspect ratio based on DIoU. CIoU loss includes three items: overlapping area, center point distance and aspect ratio. The CIoU loss function is as follows:

Among them, b and b^{gt} represent the center points of the predicted box and the real box respectively, and ρ represents the calculation of the Euclidean distance between the two center points. c represents the diagonal distance of the minimum closure area that can contain both the predicted box and the true box [13]. Their relationship is shown in Figure 12.

$$L_{CIoU} = 1 - IoU(A,B) + \frac{\rho^2(b, b^{gt})}{c^2} + a\nu$$

$$\nu = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2$$

$$a = \frac{\nu}{(1 - IoU) + \nu}$$

CIoU introduces the aspect ratio based on DIoU. CIoU loss includes three items: overlapping area, center point distance and aspect ratio. The CIoU loss function is as follows:

When the width and height of the true box and the predicted box are similar, the value of ν is 0, and the penalty term has no obvious impact at this time. Therefore, the function of this penalty term is to control when the width and height difference between the real box and the predicted box is large, so as to prompt the predicted box to adjust to the width and height that matches the real box more quickly. It can be seen that the CIoU loss function can quickly converge even when the intersection ratio is 0. By using the center point of the bounding box and the aspect ratio to perform the regression operation of the prediction box, it helps to achieve better regression effects.

4.3. Evaluation indicators

The Coco metric is an evaluation metric widely used in the field of computer vision to evaluate the performance of object detection and segmentation models. It is based on the calculation method of average average precision (mAP), which can comprehensively evaluate the performance of the

model in different scenarios. This article uses mAP@0.5, mAP@0.7, mAP@0.5 in the COCO indicator: 0.95 standard detection model evaluation indicators. AP is the area enclosed by the PR curve and the coordinate axis. It is based on precision and recall indicators, handles multiple object categories, and uses IoU to define positive predictions. Precision measures the accuracy of the model's positive predictions, while recall measures the ratio of positive examples correctly identified by the model to the actual positive examples. The AP metric provides a balanced evaluation of precision and recall [14]. The calculation method of AP is shown in Figure (13).

And take the average AP of all object categories in the data set as mAP.

$$AP = \int_0^1 P(R)dR$$

Figure 13. The calculation method of AP

4.4. Data set selection

Traffic signs cover a complex and diverse category, and there are significant differences in the traffic signs used in different countries. The most commonly used traffic sign data set in academic research is the German traffic sign data set GTSRB. This data set contains 43 types of traffic signs, of which 12,630 are used as the test set and 39,209 are used as the training set. However, the limitation of this dataset is that each image is clipped, leaving only a traffic sign and 10% of the surrounding background area. This is hugely different from the detection requirements in actual application scenarios, and does not meet the requirements of this article for the data set. Although the China traffic sign data set CCTSDB produced by the Comprehensive Transportation Big Data Intelligent Processing Laboratory of Changsha University of Science and Technology in China contains more than 10,000 data images, it only divides traffic sign annotations into three categories: instruction signs, prohibition signs and warning signs. , it is impossible to conduct fine-grained classification of traffic sign detection, and it cannot meet the needs of traffic sign detection categories in practical applications. TT100K is a traffic sign data set jointly created by Tsinghua University and Tencent. They selected 10 regions in 5 cities in China, covering the urban and suburban areas of each city. In order to meet the demand for traffic sign detection in practical applications, they use professional on-board cameras for image capture. Therefore, this article selects TT100K [15] as the data for the main experiment of this article.

5. Model evaluation

5.1. Detection results on traffic signs (TT100K)

Although TT100K contains a wealth of different types of traffic sign images, the number of traffic signs that are difficult to encounter in daily driving (such as ferries, water crossings, etc.) is very small because the images are shot in the same way as daily driving, and there are many traffic signs (such as no entry, go straight, etc.) that often appear in daily driving. In order to prevent over-fitting, we selected pictures containing the top 40 traffic signs, a total of 9573. In order to make the data distribution more reasonable, we reclassified the data, with 8000 images in the training set and 790 in the test. There are 783 images in the verification set, which is about 8:1:1. The results are shown in the figure below.

Table 2. The results

TT100K	mAP@0.5	mAP@0.7	mAP@0.7: 0.95
YOLOv8n	0.755	0.621	0.530
YOLOv8n-CO	0.777	0.645	0.546

5.2. Detection results on other data sets

In order to check the practicality of the model, we also test it on the COCO dataset.

Table 3. The COCO dataset

COCO	mAP@0.5	mAP@0.7	mAP@0.5: 0.95
YOLOv8n	0.639	0.472	0.443
YOLOv8n-CO	0.646	0.478	0.450

Acknowledgements

Fund Project: Southwest University of Science and Technology College Student Innovation and Entrepreneurship Training Program Project (No. CX23-083)

References

[1] Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3):

[2] Wu Xiaohui, Tian Qichuan. Review of traffic sign recognition methods [J]. Computer Engineering and Applications, 2020, 56(10): 20-26.

[3] Arcos-García Á, Alvarez-Garcia J A, Soria-Morillo L M. Evaluation of deep neural networks for traffic sign detection systems[J]. Neurocomputing, 2018, 316: 332-344.

[4] Li D. Research on Traffic Sign Detection and Recognition Method[J]. World Scientific Research Journal, 2022, 8(3): 462-467.

[5] Temel D, Alshawi T, Chen M H, et al. Challenging environments for traffic sign detection: Reliability assessment under inclement conditions[J]. arXiv preprint arXiv:1902.06857, 2019.

[6] Gui Xiangquan, Liu Shiqing, Li Li, et al. Pedestrian detection algorithm in scenic spots based on improved YOLOv8 [J]. 2023.

[7] Qiu Tianheng, Wang Ling, Wang Peng. Research on target detection algorithm based on improved YOLOv5 [J]. Computer Engineering and Applications, 2022, 58(13): 63-73.

[8] SUN Chuanmeng, WANG Yanping, WANG Chong, et al. Coal-rock interface recognition method integrating improved YOLOv3 and cubic spline interpolation[J]. Journal of Mining andRock Control Engineering,2022,4(01):81-90.

[9] W.H. Li, B. Zhou, B. Hu, Z.H. Zhang. Occluded face detection based on lightweight network[J]. Journal of South -Central Minzu University (Natural Science Edition),2022, 41(03):339-346. (in Chinese)

[10] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[11] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013).

[12] Hu, Mu, et al. "Online convolutional re-parameterization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[13] Cao Chengshuo, and Yuan Jie. "Mask wearing detection method based on YOLO-Mask algorithm." Laser & Optoelectronics Progress 58.8 (2021): 0810019.

[14] WANG Yicheng,ZHANG Guoliang,ZHANG Zijie. Sm-all target detection method based on improved YOLOv5[J]. Computer and Modernization, 2023(05):100-105. (in Chinese)

[15] Zhu, Zhe, et al. "Traffic-sign detection and classification in the wild." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.