ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.060

Research on Small Target Detection Algorithm based on Improved YOLOX

Caixia Meng[1], Hongpeng Chu[1+] ,Jiabao Zhang[1], Kaijie Xi

[1] Xi'an University of Posts & Telecommunication

Abstract. Aiming at the problem of insufficient feature extraction and insufficient accuracy in different stages of feature fusion in the newly launched YOLOX algorithm, an improved YOLOX target detection algorithm is proposed. This method first incorporates the convolutional attention module in the feature extraction stage, which can better capture the original rich information of the feature. Then integrate the attention mechanism in the path fusion module to further improve the feature fusion effect. Finally, the complete intersection ratio loss function is introduced in the bounding box regression process to improve the convergence speed and accuracy of the regression process. In order to test the detection effect of the algorithm, experiments were performed on the MS COCO data set and the PASCAL VOC data set. Compared with the improved YOLOX algorithm, the average accuracy of the proposed algorithm on the two data sets is increased by 1.9% and 4%, respectively, and the effect is improved significantly.

Keywords: deep learning, object detection, ciou, yolox, mff

1. Introduction

In recent years, deep learning has made significant progress in the field of target detection. At present, target detection algorithms based on deep learning are mainly divided into two categories: one-stage detection algorithms and two-stage detection algorithms [1]. The one-stage detection algorithm is a target detection method based on regression, which simultaneously classifies the image and returns the parameters of the candidate frame, eliminating the step of multiple regression; the two-stage detection algorithm is a target detection method based on the candidate area, and the candidate area is selected first. Then classify and regress the candidate area. Compared with the two-stage detection algorithm, the one-stage detection algorithm can directly classify and predict the target without the step of candidateregion classification and regression. Therefore, the one-stage detection algorithm reduces the computational complexity, the time efficiency is significantly improved, and it has greater applicability to real-time target detection and is more widely used. In the one-stage target detection algorithm, YOLOX [2] is an improvement based on the YOLO series network, which inherits the fast characteristics of the previous stage detection algorithm [3], and adopts the improved optimal transmission theory simOTA [4], Which greatly overcomes the problem of category imbalance in the training process. However, there are still areas for improvement in YOLOX, and improvements can still be made in the insufficient feature extraction and feature fusion stages. After the YOLOX model passes through the deep convolutional network, it will output feature maps with inconsistent scales at different stages. Since the downsampling rate corresponding to deep features is usually relatively large, it will cause small targets to have more effective information on the feature map. If it is less, the detection performance of small targets will drop sharply, while the resolution of shallow features is higher, and the details are often learned, which is not conducive to the detection of large targets. Although the feature pyramid PAFPN [5] in YOLOX performs feature fusion through two-way connection, the path fusion feature in PAFPN pays more attention to adjacent layer features and less attention to other layer features.

Therefore, an improved yolox target detection algorithm is proposed in this paper. Aiming at the problem that it is difficult to fully extract and fuse the features in different stages, firstly, the convolution attention CBAM CSP module is introduced into the feature extraction module [6,7]. This method infers the attention map along two independent dimensions (channel [8] and space [9]), and then multiplies the attention map

+ Corresponding author. Tel.: +17719526466.
E-mail address: 2325902720@qq.com.

and the input feature map for adaptive feature optimization, so it can better extract the rich information of different features. Then, a multi-scale feature fusion (MFF) module based on attention map is added behind the feature extraction module. The module includes bottom-up path and top-down path fusion module [10] and feature fusion operation [11], and enhances multi-layer feature fusion through fusion attention mechanism. In view of the inaccurate boundary box regression, starting from the three important geometric factors of the overlapping area, center point distance and aspect ratio of the boundary box, the loss function [12] in yolox algorithm is replaced by the complete intersection union ratio (ciou) loss function [13], which makes the convergence faster and the regression more accurate.

The main points of this paper can be summarized as follows: first, the convolution attention module is introduced into the yolox feature extraction module, which alleviates the problem that the features of different stages can not be fully extracted; Second, a multi-scale feature fusion module is added behind the feature extraction module to further enhance the feature fusion in different stages. Thirdly, ciou loss function is used in the boundary box regression process to make the regression more rapid and accurate.

The inadequacies of the original yolox algorithm are effectively improved with the aforementioned three changes based on the yolox algorithm. Simultaneously, tests on the MS COCO data set [14] and the Pascal VOC data set [15] confirm the enhanced algorithm's detection performance. The results demonstrate the efficacy of the proposed strategy.

2. Improved YOLOX target detection algorithm

The overall architecture of the improved YOLOX network is shown in Fig.1. First, input a picture, and then extract the image features from the YOLOX backbone network with the CBAM-CSP module. Add the MFF module after the feature multi-scale feature map, and finally perform classification and bounding box regression.

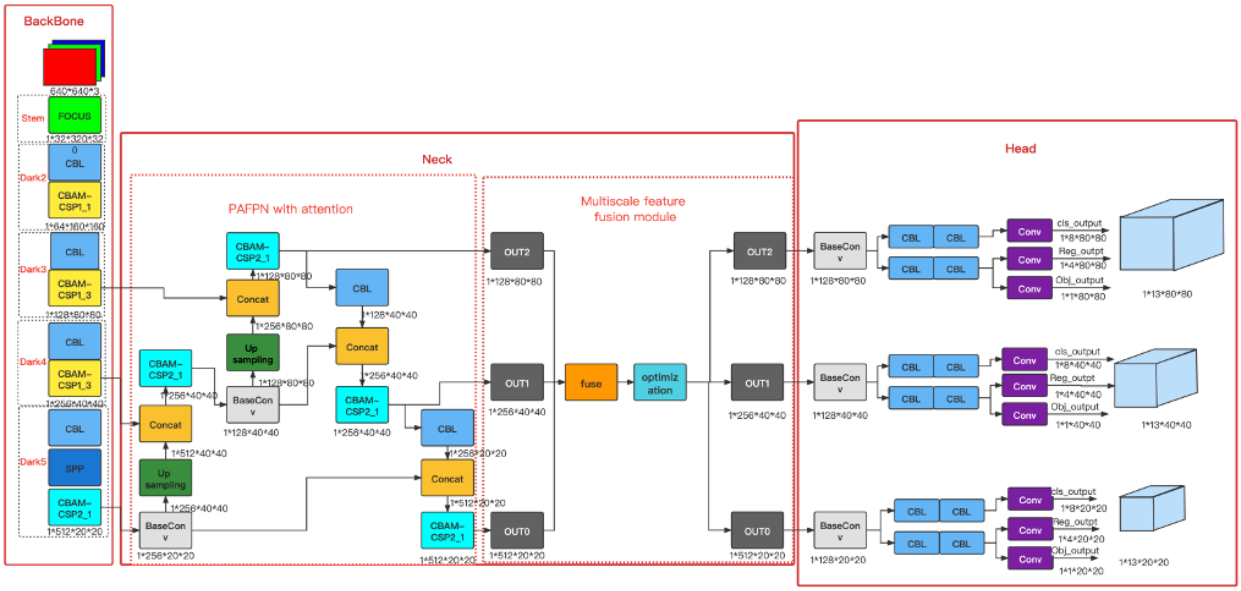

Fig. 1. Improved YOLOX's overall architecture of target detection algorithm

2.1. Feature extraction module based on convolutional attention

The backbone network of YOLOX is based on the feature extraction module of yolov5. FPN uses PAFPN for feature fusion, making it the output of the backbone network and the input of the next stage of feature fusion. Because the YOLOX backbone network has the problem of not fully extracting image feature information, the algorithm in this paper adds a convolutional block attention module (CBAM, convolutional Block Attention Moudle), which is a simple and simple for feedforward convolutional neural network. Effective attention module. Given an intermediate feature map, the CBAM module sequentially infers the attention map along two independent dimensions (channel and space), and then multiplies the attention map with the input feature map for adaptive feature optimization. The attention mechanism is added to the CSP

module of the feature extraction network to form a new CBAM-CSP module, which can effectively help the backbone network to extract rich feature information. Insert the CBAM-CSP module into the Backbone and Neck stages of the network. Take YOLOX-s as an example. Its structure is shown in Table 1. YOLOX-s BackBone has five stages, including stem, dark2, dark3, dark4 and dark5. Each dark module contains a csp module, and an attention module is inserted after each csp module. The specific operation is shown in the figure as shown in 3, a total of 4 attention modules are inserted. The introduced convolutional attention will be described in detail below.

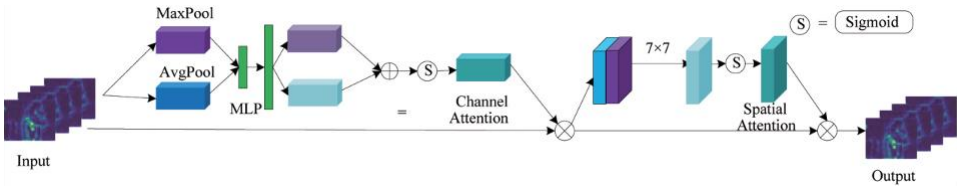

Fig. 2. CBAM attention module

1) Convolutional attention module

Convolutional Block Attention Module (CBAM) represents the attention mechanism module of the convolution module, which is an attention mechanism module that combines spatial and channel. Compared with senet, the attention mechanism that only focuses on channels can achieve better results.

The Fig.2 above shows the overall structure after adding the CBAM module. It can be seen that the output result of the convolutional layer will first pass through a channel attention module, and after the weighted result is obtained, it will pass through a spatial attention module, and finally the result is weighted.

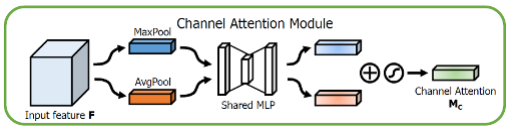

Fig. 3. Channel attention mechanism

The channel attention module is shown in Figure 3. The input feature maps are respectively passed through global max pooling and global average pooling based on width and height, and then passed through MLP respectively. The MLP output features are subjected to an element-wise addition operation, and then a sigmoid activation operation is performed to generate the final channel attention featuremap. Perform an elementwise multiplication operation on the channel attention featuremap and input featuremap to generate the input features required by the Spatial attention module. The above are the steps of the channel attention mechanism.

From another perspective, the Channel Attention Module (Channel Attention Module) compresses the feature map in the spatial dimension to obtain a one-dimensional vector before performing the operation. When compressing in the spatial dimension, not only the average pooling (Average Pooling) but also the maximum pooling (Max Pooling) is considered. Average pooling and maximum pooling can be used to aggregate the spatial information of the feature map, send it to a shared network, compress the spatial dimension of the input feature map, and sum and merge element by element to generate a channel attention map. As far as a picture is concerned, channel attention focuses on which content on the picture is important. Average pooling has feedback for each pixel on the feature map, while maximum pooling is used for gradient backpropagation calculation, only the place with the largest response in the feature map has gradient feedback. The channel attention mechanism can be expressed as:

$$M_c(F) = \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right) \quad (1)$$

$$= \sigma\left(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0(F_{max}^c)\right)\right) \quad (2)$$

The feature map output by the Channel attention module is used as the input feature map of this module. As shown in Fig.4, First do a channel-based global max pooling and global average pooling, and then perform concat operations on these two results based on the channel. Then after a convolution operation, the

dimensionality is reduced to 1 channel. Then generate spatial attention feature through sigmoid. Finally, the feature and the input feature of the module are multiplied to obtain the final generated feature.

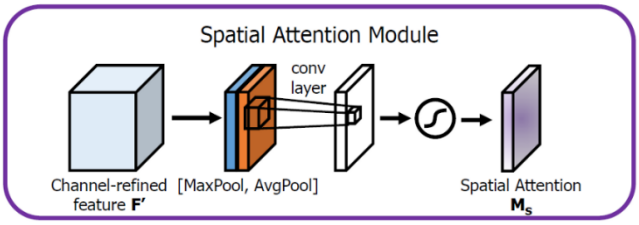

Fig. 4. Spatial attention mechanism

Similarly, the spatial attention mechanism (Spatial Attention Module) compresses the channel, and performs average pooling and maximum pooling in the channel dimensions. The operation of MaxPool is to extract the maximum value on the channel, the number of times of extraction is the height multiplied by the width; the operation of AvgPool is to extract the average value on the channel, the number of times of extraction is also the height of width; then the feature map extracted earlier (The number of channels is 1) Combine to get a 2-channel feature map.

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \tag{3}$$

$$= \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right) \tag{4}$$

Among them, it is the sigmoid operation, 7*7 represents the size of the convolution kernel, and the 7*7 convolution kernel is better than the 3*3 convolution kernel.

2) CBAM-CSP module

Add the CBAM module behind the original CSP1_X and CSP2_X modules in YOLOX to generate CBAM-CSP1_X and CBAM-CSP2_X modules, See Fig.5 and Fig.6 below.

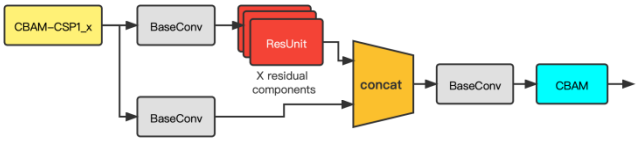

Fig. 5. CBAM-CSP1_x module

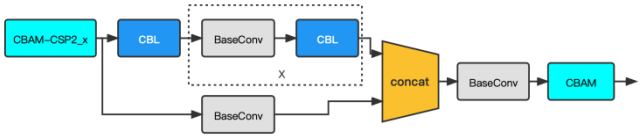

Fig. 6. CBAM-CSP2_x module

The improved CSP module can add an attention mechanism to the feature map output after each dark phase, which can effectively focus on the effective features of the target and enhance the detection ability of the network.

2.2. Multi-scale feature fusion module

The MFF module adds attention mechanism and feature fusion operation to the original YOLOX PAFPN module, including a path fusion module with bottom-up and bottom-up paths and a feature fusion operation.

1）Path Fusion Module

As shown in Fig.1, a path fusion module with bottom-up and bottom-up paths is connected behind the YOLOX feature extraction module. With three different scales output by the dark3, dark4, and dark5 modules, the feature map output by the dark5 module is convolved and upsampled and then spliced directly with the feature map output by the dark4, and then the result is output through the CBAM-CSP2_1 module Through convolution and upsampling, and the feature maps output by the dark3 module are spliced together, and then through the CBAM-CSP2_1 module, the output result is used as the output result of the first scale as out2 on the one hand, and convolution by the CBL module on the other hand The channel is down-

sampled, and then stitched together with the feature map after the previous convolution, and then passed through the CBAM-CSP2_1 module. The output result is used as the result of the second output scale as OUT1, and at the same time as the input of the next scale , And then through the CBL module for convolutional down-sampling, and then spliced together with the feature map obtained by the first convolution, and finally the result of the feature map after a CBAM-CSP2_1 module as the last scale is used as OUT0, and the three Different feature maps are used as the input for the next stage of feature fusion. Down-sampling, down-sampling and splicing through feature maps of different scales constitute a path fusion module with bottom-up and top-down paths. This module adds a convolutional attention mechanism to the original PAFPN, which can focus on the content of different channels in the fusion, and can effectively strengthen the detection ability of the network.

2) Feature fusion operation

The feature fusion operation is mainly divided into two steps of scaling integration and optimization, as shown in Figure 1. First adjust the size of the feature map for fusion, and average the fused features. There are currently three-layer features {OUT2, OUT1, OUT0}. Because the low-level features have high resolution, they often learn detailed features, and the high-level features have low resolution, and they learn semantic features. Therefore, it is necessary to adjust the size of these three-layer features to the size of the middle-level OUT1 feature map for fusion. The operation taken is to downsample OUT0, upsample OUT2, and perform no other operations on OUT1, and then do a simple phase. Add and average operation, as shown in the following formula:

$$N = \frac{1}{L}\sum_{l=l_{min}}^{l_{max}} N_l \tag{5}$$

Among them, L represents the number of feature layers, and represents the l-th feature. Then, the averaged feature maps are further optimized to make the features more discriminative. The optimization operation uses the embedded Gaussian non-local module . The definition of this optimization operation is as follows:

$$M_i = \frac{1}{C(N)}\sum_{\forall j} f\left(N_i, N_j\right) g\left(N_j\right), C(N) = \sum_{\forall j} f\left(N_i, N_j\right) \tag{6}$$

$$f\left(N_i, N_j\right) = e^{\theta(N_i)^T \phi(N_j)} \tag{7}$$

Among them, M and N are feature maps of the same size, i is a pixel position of the feature map, and j is the index of all possible positions. g is a unary input function, generally 1×1 convolution is used, and the purpose is to transform information. f is the pairing calculation function, which calculates the correlation between the i-th position and all other positions. Both θ and ϕ are 1×1 convolution operations, and T is set to 1. C(N) is a normalized function to ensure that the overall information remains unchanged before and after the transformation.

Finally, disperse the optimized features into multi-layer features {OUT0, OUT1, OUT2}, where OUT0 is obtained by up-sampling the optimized features, OUT1 features are directly output, and OUT2 is the optimized features down-sampling of. The above process is the two steps of feature fusion. The path fusion module is used to enhance features and feature fusion operations to fuse multi-layer features, which effectively alleviates the problem of difficulty in fully fusing features at different stages in YOLOX.

2.3. Bounding box regression and classification module

The bounding box regression network and classification network of the algorithm in this paper use the regression and classification network of the original YOLOX algorithm. Among them, YOLOX performed a Decoupled Head operation, commonly known as ‘decoupled head’, and divided the original Yolo Head into three branches: (1) cls_output: mainly predicts the score of the target frame category. Because the COCO data set has a total of 80 categories, and mainly N binary classification judgments, after the Sigmoid activation function is processed, it becomes 20*20*80 in size. (2) obj_output: Mainly determine whether the target frame is the foreground or the background, so it is processed by Sigmoid and becomes 20*20*1 size. (3) reg_output: mainly predict the coordinate information (x, y, w, h) of the target frame, so the size is 20*20*4.

Among them, the regression of the target box uses the intersection ratio IOU_Loss. There are two problems with using IOU_Loss. One is that if the target box and the prediction box do not overlap, the loss

function will not work. The second is that if the sizes of the two pairs of prediction boxes and target boxes are the same, and the intersection values of the two pairs of boxes are also the same, then it is not certain how the two pairs of boxes intersect.

1) Complete intersection ratio loss function

In response to the above problems, the algorithm in this paper replaces the bounding box regression loss function of YOLOX algorithm with CIoU_Loss (Complete-IOU) [14]. CIoU_Loss starts from the three factors of overlap area, center point distance and aspect ratio in bounding box regression. First, directly minimize the normalized distance between the prediction box and the target box to achieve a faster convergence speed; secondly, when the prediction box and the target box do not overlap, or there is overlap or even containment, the regression is more accurate .

First, briefly introduce the definition of Intersection over Union (IoU):

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{8}$$

Among them, is the target frame, which is the prediction frame, x, y, w, and h are the center point coordinates and width and height of the frame, respectively, represent the area of the overlap between the target frame and the prediction frame, and represent the two frames enclosed by the target frame and the prediction frame The total area is shown in Fig.7.

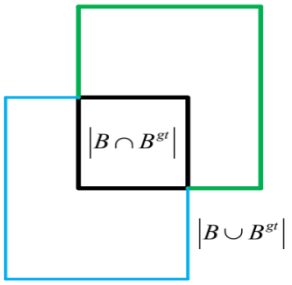

Fig. 7. Schematic diagram of IoU

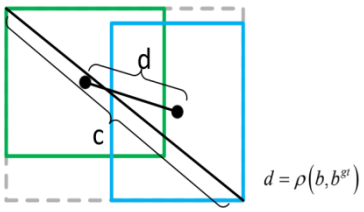

Fig. 8. Schematic diagram of CIoU

Therefore, the CIoU loss function is defined as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{9}$$

Among them, b and represent the center points of the prediction box B and the target box, are the Euclidean distance between the two center points, and c is the diagonal length of the smallest closed box that contains both the prediction box and the target box, as shown in Fig.6. Show. α is the coefficient used to balance the aspect ratio, and v is used to measure the consistency of the aspect ratio between the predicted frame and the target frame. Their definitions are as follows:

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{10}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \tag{11}$$

3. Experiment

3.1. Data set and evaluation indicators

The improved algorithm was tested on two public data sets, MS COCO[15] and PASCAL VOC[16]. The MS COCO data set contains 80 categories, 118287 images for training, and 5000 images for verification. ,

20000 pictures used for testing, the algorithm experiment of this paper was carried out on test-dev 2017 and compared with the latest target detection algorithm. The PASCAL VOC data set contains 20 categories, including 22,136 training pictures (trainval 2007+trainval 2012) and 4952 test pictures (test 2007). The experimental results follow the indicators of the VOC data set. The average accuracy indicates that the category’s IoU threshold is The average accuracy of 0.5, using test2007 for ablation experiments. The experimental results all follow the average accuracy (Average Precision, AP) index of the MS COCO standard, where AP means that the IoU starts at 0.5, and the threshold is set every 0.05 until the average accuracy obtained by taking 0.95 is averaged again, indicating that the IoU threshold is 0.5 The average accuracy at time, represents the average accuracy of the small target.

3.2. Model parameter settings

For the COCO data set, first adjust the size of the input image to 640, use Mosaic and MixUp data enhancement and turn it off at the last 15 epochs, and then use Stochastic Gradient Descent (SGD) to optimize all models, and the weight decays to 5e -4, momentum is 0.9, minimum—batch size is 2. The learning rate is initialized to 0.00015625, and a total of 50 epochs are trained. The settings for the VOC data set are the same as the COCO data set. The experiments were carried out on a graphics card model of NVIDIA GTX 965m. The baseline model in this article is YOLOX, and the baseline model on the VOC data set is realized when the other hyperparameter settings are the same. It should be pointed out that the experimental effect of the experimental baseline model YOLOX in this article is 82.1% in the VOC data set.

3.3. Experimental results and analysis

1) Comparative experiment and result visualization

This section evaluates the performance of the improved YOLOX target detection algorithm proposed in this paper on the COCO test-dev 2017 data set and the PASCAL VOC test set. On the COCO data set, in order to make the comparison results more concise and clear, the experiment mainly uses the YoloX-l model, which is to compare the improved YoloX algorithm with Modifified CSP v5 as the backbone network with the latest one-stage and two-stage target detection algorithms. As shown in table 1. It can be seen from Table 1 that the backbone network is Modifified CSP v5's improved YOLOX target detection algorithm AP can reach 51.9%, and the performance has been significantly improved. Compared with other target detection algorithms, the improved YOLOX algorithm achieves the best results.

Table 1: Comparison of Different methods

method	**BackBone**	**AP**	**AP50**	**AP75**	**APs**
Faster R-CNN	ResNet-101	33.9	56.9	-	17.8
Cascade R-CNN	ResNet-50	40.3	59.4	43.7	22.9
EffificientDet-D3	Efficient-B3	45.8	65	49.3	26.6
Sparse R-CNN	ResNet-50	42.3	61.2	45.7	26.7
RetinaNet	ResNet-101	40.5	59.3	43.6	23
FCOS	ResNet-101	41	60.7	44.1	24
PP-YOLOv2	ResNet-50	49.5	68.2	54.4	30.7
YOLOv4-CSP	CSPDarknet-53	47.5	66.2	51.7	28.2
YOLOv5-L	Modified CSP v5	48.2	66.9	-	-
YOLOX-L	Modified CSP v5	50	68.5	54.5	29.8
ours	**Modified CSP v5**	**51.9**	**70.3**	**56.1**	**31.3**

On the VOC test set, the experiment mainly compares the accuracy of each category of the improved YOLOX target detection algorithm (the backbone network is Modifified CSP v5) with the original YOLOX algorithm. The specific results are shown in Table 2. It can be seen from Table that the category average accuracy of the improved YOLOX algorithm is higher than that of the original YOLOX algorithm.

Table 2: Comparison of original method and improved method on VOC data

method	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
ours	0.902	0.9	0.877	0.811	0.809	0.903	0.9	0.895	0.778	0.901
YOLOX-L	0.851	0.879	0.834	0.733	0.735	0.881	0.879	0.865	0.721	0.854
diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	MAP
0.829	0.875	0.905	0.898	0.887	0.693	0.879	0.851	0.88	0.854	0.861
0.804	0.828	0.893	0.833	0.821	0.658	0.871	0.814	0.855	0.812	0.821

At the same time, some pictures are randomly selected from the MS COCO data set, and the four pairs of representative test results are selected herein. As shown in Fig.9, the visualization of the YOLOX algorithm, as shown in Fig.9, is shown in Fig.9, is shown in the visual result of the type YOLOX algorithm, it can be seen that the detection result of the improved YOLOX target detection algorithm has higher accuracy. The rate, the detected border is more accurate.

Fig. 9. Comparison of visualization results on coco dataset

2) Ablation experiment

All ablation experiments in this article are performed on the VOC2007 data set. The experimental results are all compared with Baseline, which is the YOLOX algorithm with Modifified CSP v5 as the backbone network. Among them, AP5096 means that IoU starts at 0.5, and every 0.05 is used as a threshold until the average accuracy obtained by taking 0.95 is averaged again. MAP means the average accuracy when the IoU threshold is 0.5, which means the average accuracy of small targets.

Finally, through the pairwise combination and comparative analysis experiment in Table 6, it can be seen that adding a single module to the baseline model YOLOX, or adding two of them, cannot achieve the best performance. The reason is that each module has different functions. For the target detection network as a whole, feature extraction, feature fusion and bounding box regression are all very important parts. Therefore, when improving the target detection algorithm, you should not only focus on part of the network, but analyze the problems of the overall network. Then solve and improve these problems. Therefore, the combination of

these three improvements not only improves the problem of the difficulty in fully extracting and fusing multi-layer features, but also alleviates the inaccuracy of bounding box regression, which verifies the effectiveness of the improved algorithm. As shown in Table 6, the detection accuracy of the improved YOLOX target detection algorithm on the VOC2007 (test) data set is 4% higher than that of the YOLOX algorithm, and the performance is significantly improved.

Table 3: Comparative analysis experiment of cbam-cs module, MFF module and ciou loss function

CBAM-CSP	MFF	CIOU	AP	AP5095
			82.1	63.7
√			83.3	65.5
	√		84.1	65.3
		√	82.9	64.9
√	√		84.2	66.3
√		√	83.9	66.5
	√	√	84.5	67.2
√	√	√	86.1	68.5

4. Concluding remarks

Aiming at the problems of insufficient extraction and fusion of features in YOLOX algorithm and insufficient accuracy of bounding box regression, this paper proposes an improved YOLOX target detection algorithm. Specifically, this paper introduces the convolutional attention module in the feature extraction module and the path fusion module, then combines the path fusion module with the feature fusion operation to form a multi-scale feature fusion module, and finally introduces the CIoU loss in the bounding box regression process function. Through the above improvements, the problems in the YOLOX algorithm have been effectively improved, and its detection performance has been significantly improved on the two public data sets of MS COCO and PASCAL VOC. It is that the improved method proposed at present has not been applied to the two-stage target detection algorithm. The next step will be to carry out follow-up research on the proposed improved method to make it effective in the two-stage target detection algorithm.

5. References

[1] Ying Liu, Luyao Geng, Weidong Zhang, Yanchao Gong, and Zhijie Xu, "Survey of Video Based Small Target Detection," Journal of Image and Graphics, Vol. 9, No. 4, pp. 122-134, December 2021. doi: 10.18178/joig.9.4.122-134

[2] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[3] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[4] Ge Z, Liu S, Li Z, et al. OTA: Optimal Transport Assignment for Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 303-312.

[5] Mei Y, Fan Y, Zhang Y, et al. Pyramid attention networks for image restoration[J]. arXiv preprint arXiv:2004.13824, 2020.

[6] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[7] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[9] Bansal N, Agarwal C, Nguyen A. Sam: The sensitivity of attribution methods to hyperparameters[C]//Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2020: 8673-8683.

[10] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[11] Pang J, Chen K, Shi J, et al. Libra r-cnn: Towards balanced learning for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 821-830.

[12] Zhou D, Fang J, Song X, et al. Iou loss for 2d/3d object detection[C]//2019 International Conference on 3D Vision (3DV). IEEE, 2019: 85-94.

[13] Zheng Z, Wang P, Ren D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021.

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

[15] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015, 111(1): 98-136.