ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

Object Detection Algorithm based on Improved FCOS

Feng Luo, Huazhang Wang*

Southwest Minzu University college of Electrical Engineering, Chengdu, Sichuan 610041, China

Abstract

Aiming at the use of parallel localization and classification prediction branches in the detection head, due to the different learning mechanisms, there are spatial differences in the learned features, resulting in a certain degree of spatial mismatch. This paper proposes an improved object detection based on FCOS network. algorithm. After the original pyramid feature fusion, a layer of PA feature enhancement network is added, so that the features of the shallow layer can be better integrated into the deep layer, strengthen the feature extraction ability of the network, and improve the detection head to a single-branch detection head, It's used to improve the feature interactivity of classification and localization tasks and reduce the feature space differences in classification and localization tasks. In this paper, the effectiveness of the algorithm is verified on the Pascal VOC2007 and Pascal VOC2012 image data sets. The experimental results show that the improved network has improved accuracy compared with the FCOS network.

Keywords

T-head; PAFPN; ResNet50; FCOS; Object Detction.

1. Introduction

Object detection is to combine object localization and object classification to locate and identify the object of interest in natural pictures, which is usually expressed as a multiple learning problem of object classification and localization. The classification task is designed to identify key and salient features of an object, and the localization task is to precisely locate the entire object and its boundaries. In computer vision, object detection is a basic and challenging task. At present, the mainstream detectors include FCOS[1] for one-stage object detection and Faster-RCNN[2], SSD[3] and two-stage object detection for two-stage object detection. YOLO[4] series. The second-stage anchor frame-based detector has some disadvantages: the detection performance is related to the size, aspect ratio, and number of the predefined anchor frames; the detection effect of small objects with large changes in size is not good, so the aspect ratio needs to be reset to correspond to different objects , which hinders the generalization ability of the detector; in order to obtain a high recall rate, a large number of predefined anchor boxes are used, and most of the anchor boxes are negative samples, which aggravates the imbalance of positive and negative samples; due to a large number of predefined anchor boxes The box will generate complex calculations, the amount of calculation is greatly increased, and the hardware requirements are very high. Recent one-stage detectors attempt to predict the consistent output of two separate tasks with the central part of the object, and they assume that the anchor or frame of the central part of the object can accurately predict classification and localization. For example, FCOS and ATSS[5] use the center branch to enhance the classification scores predicted from localization points near the object center and assign larger weights to the localization loss for the corresponding localization. Due to the differences in the classification and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

localization learning mechanisms, there are spatial differences in the learned features, resulting in a certain degree of spatial mismatch.

In this paper, the FOCS algorithm model is improved, the pyramid feature fusion is improved, and the information of different layers is better enriched by using the shallow features of the network to obtain rich features. At the same time, the single-branch detection detection head is used to provide interactive and special The feature balance between tasks increases the feature interaction between tasks.

2. Related Work

Recent one-stage object detectors attempted to predictconsistent outputs of the two separate tasks, by focusing on the center of an object [1,5,6,7]. In 2019, Tian proposed Fully Convolutional One-Stage Object Detection (FCOS), using anchor-free, which was a single-stage detection network with excellent accuracy and detection speed at that time. In order to solve the long-term problem of low detection accuracy of single-stage detectors due to fuzzy samples and low-quality object frames far from the object center, center-ness is proposed to suppress low-quality object frames far from the object center, and the fuzzy samples are eliminated by feature pyramid, so that single-stage detection has returned to the public view. Its network model is divided into three parts, namely backbone network, feature pyramid and detection head. Its network structure is shown in Figure 1 below.

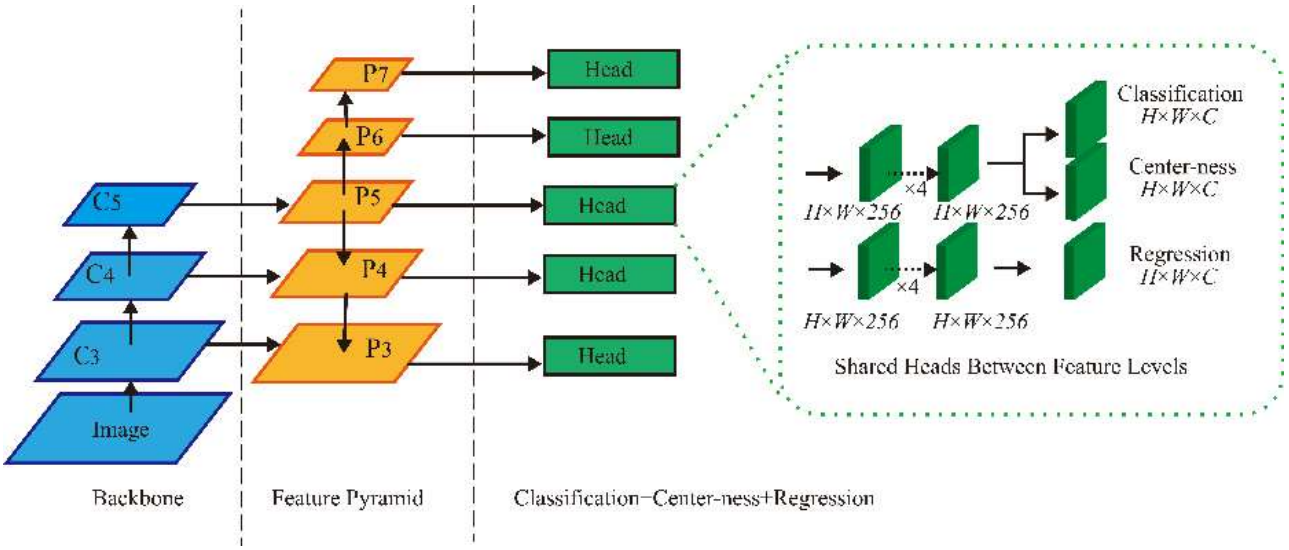


Figure1. FCOS network structure

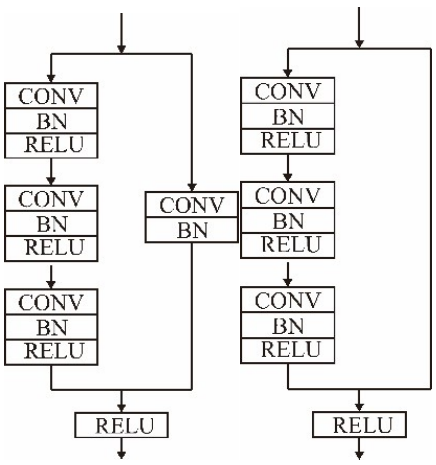


Figure 2. Residual module

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

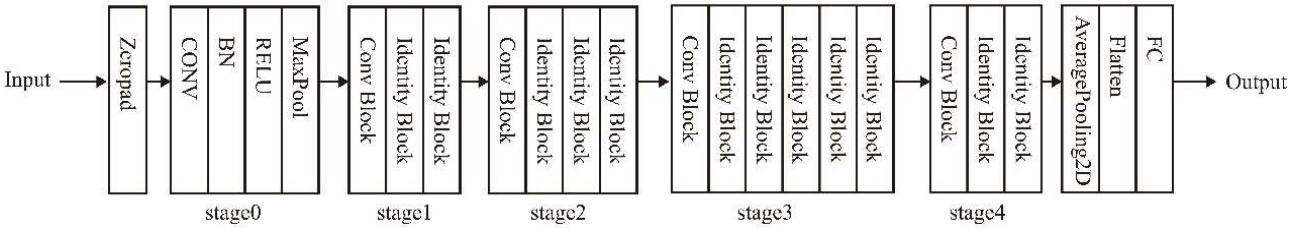


Figure 3. The network structure of the ResNet50 model

Backbone network. The backbone network adopts ResNet50[8] proposed by He Kaiming in 2016, which is a residual network. It uses a short-circuit method to achieve identity mapping, so that the network depth can be deepened while retaining the performance of the network without the problem of a decrease in the recognition rate. The residual module adopted by ResNet50 can be divided into ConvBlock structure and Identity Block according to whether the number of input and output channels is the same, as shown in Figure 2 below. The network structure of the ResNet50 model is shown in Figure 3 below. It consists of 49 convolutional layers and one fully connected layer. It can be divided into five stages. Stage0 is the input image preprocessing, and the remaining four stages include 3, 4, 6 and 3 residual modules, CONV in stage means convolution, BN means Normalization, RELU is ReLU activation function.

Feature Pyramid. The feature pyramid mainly uses ResNet50 to extract feature maps C3 to C5 of different sizes from bottom to top, and performs 1x1 convolution on C5 to reduce the dimension to obtain P5. In order to obtain a feature map with the same length and width as C4 and C3, which is convenient for subsequent element-by-element addition, P5 is sequentially upsampled to obtain P4 and P3. In order to ensure the diversity of features, P5 is sequentially downsampled to obtain P6 and P7. C3 and C4 are also reduced by 1x1 convolution to make the number of channels fixed to 256. Through horizontal connection, the up-sampled high semantic features and shallow geometric features are fused to obtain new P4 and P3. Objects of different sizes are divided into different feature layers through the feature pyramid network, thereby greatly reducing the number of fuzzy samples.

Detection head. FCOS is a single-stage detector, and for the anchor point (x, y) on the feature map, there is a 4D ground truth vector $t^* = (l^*, t^*, r^*, b^*)$ as the regression object of this position. $x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)}$ are the coordinate values of the upper left corner and the lower right corner of the bounding box, l^*, t^*, r^*, b^* corresponding to the distance from the drawn point to the four sides of the bounding box, and the calculation formula is shown in the following formula Eq.(1):

$$\begin{aligned} l^* &= x - x_0^{(i)}, \quad t^* = y - y_0^{(i)} \\ r^* &= x_1^{(i)} - x, \quad b^* = y_1^{(i)} - y \end{aligned}, \tag{1}$$

In order to suppress many low-quality prediction bounding boxes with high confidence but far from the center of the object generated in the process of object detection, FCOS proposes a center-ness branch parallel to the classification branch without introducing any super-constant. It describes the normalized distance from the location to the center of the object responsible for the location, and the center-ness formula is shown in Eq.(2):

$$\text{centerness}^* = \sqrt{\frac{\min\left(l^*, r^*\right)}{\max\left(l^*, r^*\right)} \times \frac{\min\left(t^*, b^*\right)}{\max\left(t^*, b^*\right)}}, \tag{2}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

The center-ness range is 0 to 1, so binary cross entropy (BCE) loss is used for training, and the final classification score is equal to the original classification score multiplied by the center-ness, so that the objects far from the center of the object have high confidence through the center branch. The bounding box score is reduced and then filtered out by non-maximum suppression (NMS), which greatly improves the detection performance.

3. Improvement of FCOS Algorithm

FCOS is to predict the consistent output of two separate tasks in the central part of the object. It assumes that the anchor point or anchor box in the central part of the object can accurately predict classification and localization, and due to the different learning mechanisms of classification and localization, the learned features If there are spatial differences, there will be a certain degree of spatial mismatch. In order to solve this problem, this paper improves FCOS, and the improved model is shown in Figure 4 below.

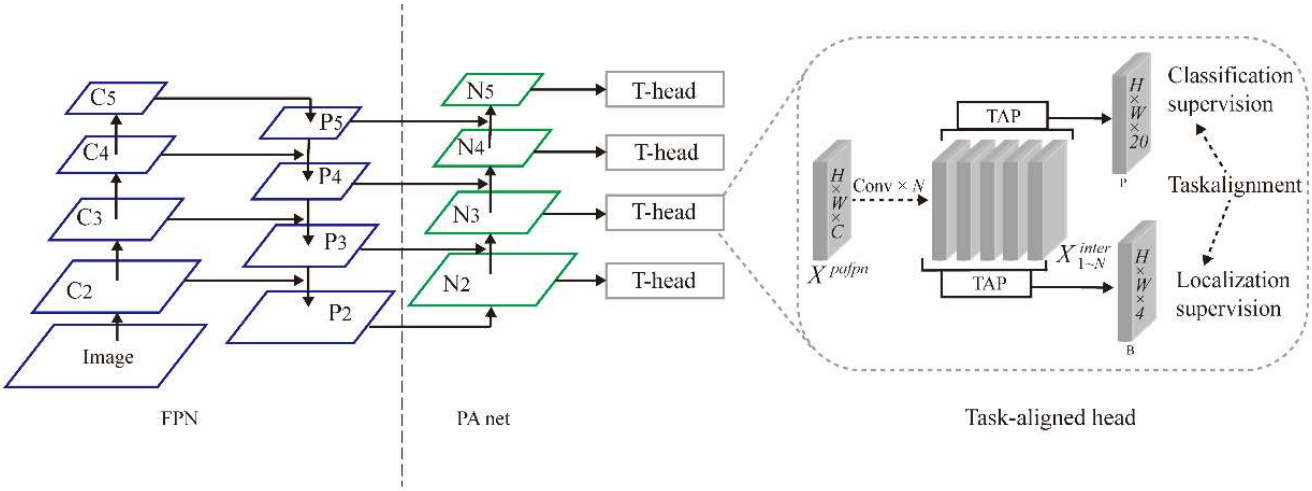


Figure 4. Improved model structure

3.1 Feature Pyramid Improvements

The method of direct feature fusion in FPN[9] still misses a lot of semantic information. The bottom-up path augmentation structure is introduced, and the shallow features of the network are used to better enrich the information of different layers. The PAFPN[10] network structure is shown in Figure 5 below. Use the ResNet50 backbone network to extract five feature layers from C2 to C5. As before, first reduce the dimension of C5 through a 1x1 convolution kernel to generate P5, and then upsample to expand the width and height of the feature map in order to follow up with the corresponding feature layer. Element-by-element addition, different from the previous upsampling attempt P2, in order to better utilize the shallow features of the network, and remove the downsampling P6 and P7 layers. The principle is the same as before. The elements in P2, P3, and P4 are updated through horizontal connection, and then P2 is downsampled to obtain N3, N4, and N5, and then horizontal connection is performed to update the elements to obtain the N2, N3, N4, and N5 feature maps. Used for classification and localization prediction of detection heads. After several passes of the network model, the shallow features are fused into the deep layers through the bottom-up path augmentation structure, which solves the situation that the shallow features cannot be retained in the deep features, so that the deep features also have the characteristics of the shallow features. to more abundant feature information.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

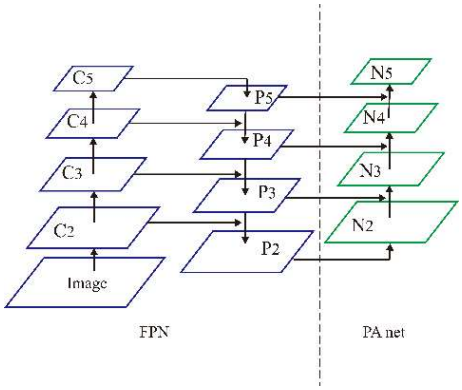


Figure 5. Improved feature pyramid

3.2 Detection Head Improvement

Two parallel prediction branches are used in the original FCOS detection head, and center-ness is added to the classification task to suppress low-quality object boxes far from the object center. However, this enhancement with the center branch is to predict the scores of the anchor points near the center of the object, and assign a larger weight to the corresponding positioning loss, which is the assumption that the anchor points or anchor boxes in the center part of the object can accurately predict the classification and positioning. on a sexual basis. However, due to the differences in the learning mechanism of classification and localization, there will be spatial differences in the learned features. Therefore, in order to enhance the feature interaction between classification and localization tasks, and to close or even unify the optimal anchor boxes in the two tasks, this paper improves the detection head, and adopts the detection head T-head[11] for single-branch detection. Its structure is as follows shown in Figure 6. In the detection head, a feature extractor is first used to form N continuous convolution layers through the activation function of the features obtained from the PAFPN, in which the interactive features of the classification and localization tasks are calculated to improve the interactivity. The calculation formula is as follows Eq.(3):

$$X_k^{\text{inter}} = \begin{cases} \delta\left(\text{conv}_k\left(X^{pafpn}\right)\right), k=1 \\ \delta\left(\text{conv}_k\left(X_{k-1}^{\text{inter}}\right)\right), k>1 \end{cases}, \forall k \in \{1,2,\ldots,N\}, \tag{3}$$

where $conv_k$ is represented as the k-th convolutional layer, δ is a relu activation function, $X^{pafpn} \in \mathbb{R}^{H\times W\times C}$ is the PAFPN feature, H is the height, W is the width, and C is the number of channels.

After the rich multi-scale features are extracted from the PAFPN features, they are injected into two TAPs to align the classification and localization tasks.

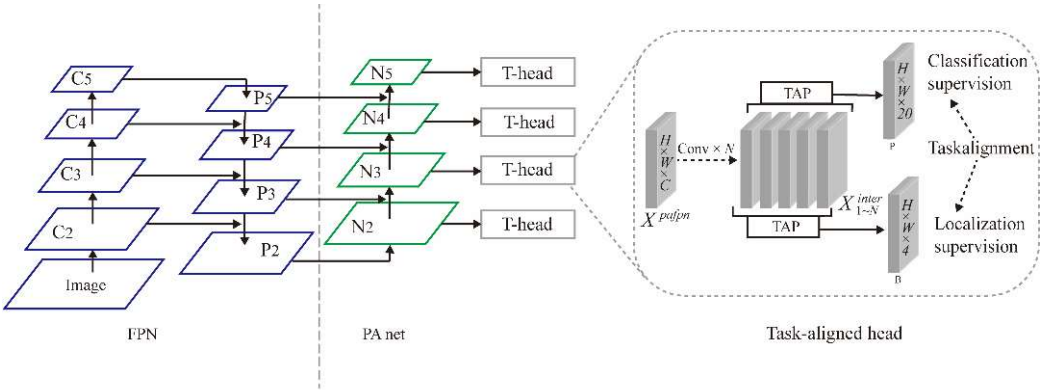


Figure 6. Task-aligned predictor

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

However, due to the use of single-branch detection, the task interaction feature inevitably introduces a certain degree of feature conflict between two different tasks. The task of object classification and localization will have different attention objects because it pays attention to different types of features, and the features of specific tasks are dynamically calculated on the layer. To encourage task decomposition, separate task-specific features for each classification or location. The task-specific features are computed separately for each task ofclassification or localization:

$$X_k^{\text{task}} = \boldsymbol{w}_k \cdot X_k^{\text{inter}}, \forall k \in \{1,2,\ldots,N\}, \tag{4}$$

where w_k is the k-th element in the attention mechanism layer. It can capture the dependencies between convolutional layers:

$$\boldsymbol{w} = \sigma\left(fc_2\left(\delta\left(fc_1\left(\boldsymbol{x}^{\text{inter}}\right)\right)\right)\right), \tag{5}$$

where fc_2 and fc_1 are two fully connected layers, σ is the sigmoid function, and x^{inter} is the connection feature of X_k^{inter} is obtained by global average pooling. The dense classification score $P \in \mathbb{R}^{H\times W\times C}$, or $B \in \mathbb{R}^{H\times W\times C}$ the object bounding box in classification and localization tasks is Z^{task} reduced by $conv_1$ dimensionality reduction through this 1x1 convolutional layer, and then converted into a sigmoid function. the results of classification or localization are predicted from each X^{task}:

$$Z^{\text{task}} = conv_2\left(\delta\left(conv_1\left(X^{\text{task}}\right)\right)\right), \tag{6}$$

During prediction, the computed task interactivity features are used to align the two predictions, and the two tasks are aligned by adjusting the spatial distribution of the two prediction tasks. The two prediction tasks perform different homogeneous methods. we use a spatial probability map $M \in \mathbb{R}^{H\times W\times 1}$ to adjust the classification prediction:

$$P^{\text{align}} = \sqrt{P \times M} \tag{7}$$

Eq.(9) is calculated by the interaction feature, which can learn the degree of consistency between the two tasks in the spatial position, and the formula is shown in Eq.(8). The positioning prediction uses the spatial offset map $O \in \mathbb{R}^{H\times W\times 20}$ learned from the interaction features to adjust the predicted bounding box of each position to find the best boundary prediction for the best alignment anchor. The formula is as follows, and the calculation formula is as follows Eq.(10).

$$B^{\text{align}}(i,j,c) = B(i + O(i,j,2\times c), j + O(i,j,2\times c+1), c), \tag{8}$$

$$M = \sigma\left(conv_2\left(\delta\left(conv_1\left(X^{\text{inter}}\right)\right)\right)\right), \tag{9}$$

$$O = conv_4\left(\delta\left(conv_3\left(X^{\text{inter}}\right)\right)\right), \tag{10}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

where an index (i, j, c) denotes the (i, j)-th spatial location at the c-th channel in a tensor, which is calculated by bilinear interpolation.

3.3 Loss Function Improvements

The classification loss function in FCOS adopts the Focal loss[12] function, and the positioning adopts the IOU loss function. The loss function in the improved classification task is shown in Equation (11).

$$L_{cls} = \sum_{i=1}^{N_{Pos}} \left| \hat{t}_i - s_i \right|^{\gamma} BCE\left(s_i, \hat{t}_i\right) + \sum_{j=1}^{N_{\text{neg}}} s_j^{\gamma} BCE\left(s_j, 0\right) \tag{11}$$

where N_{pos} is the number of positive anchor points, BCE is the binary cross entropy, t is the maximum IoU in each instance, i is the ith in the total number of positive points in each instance, and j is among the total number of negative points in each instance The j-th negative anchor point, is the focus parameter. The localization loss function is shown in Equation (12).

$$L_{reg} = \sum_{i=1}^{N_{pos}} \hat{t}_i L_{GIoU}\left(b_i, \overline{b}_i\right) \tag{12}$$

where b is the predicted bounding box, $\overline{b}_i$ is the corresponding background ground-truth box of b, and LGIoU is the GIoU loss function[13].

4. Experimental and Results

Dataset and evaluation protocol. All experiments are implemented on the large-scale detection benchmark VOC 2007 and 2012[14]. The Pascal VOC 2007 and 2012 datasets contain 20 different objects, with 9963 images in VOC2007 and 11530 images in VOC2012. Objects include 20 categories including people, animals (such as cats, dogs, islands, etc.), vehicles (such as cars, boats, planes, etc.), furniture (such as chairs, tables, sofas, etc.), and each image has an average of 2.4 object.

Implementation details. In this paper, various types of mean average precision (mAP) and center offset distance are used as the main evaluation indicators, and the convergence speed and object detection speed are used as auxiliary evaluation indicators. Object detection models for comparison.

The model training and experimental results testing in this paper are all run on the server. The code programming language is implemented in python. The experiment uses 5 GPUs for distributed parallel computing, which effectively shortens the training time of the model compared to single GPU training.

Algorithm comparison. To visually describe the effectiveness of the network model, the model is trained using the VOC2007 and 2012 datasets. During the experiment, by observing the loss function image, set the initial learning rate to 0.001 to speed up the convergence, set the batch to 12, adjust the learning rate to 0.0001 when training to the 12th batch, and set the number of iterations to 588.

In Figure 7, the abscissa represents the number of iterations, and the ordinate represents the loss value. It can be clearly seen from the figure that the loss value continues to decrease during the model training process, indicating that the neural network is in the process of training. The difference between the predicted value and the actual value The difference is decreasing. Comparing the last few trainings of the network, the PAFPN_T-head loss value fluctuates between 0.106 and 0.111, but the overall level is still flat, which shows that the training of the neural network has been completed.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

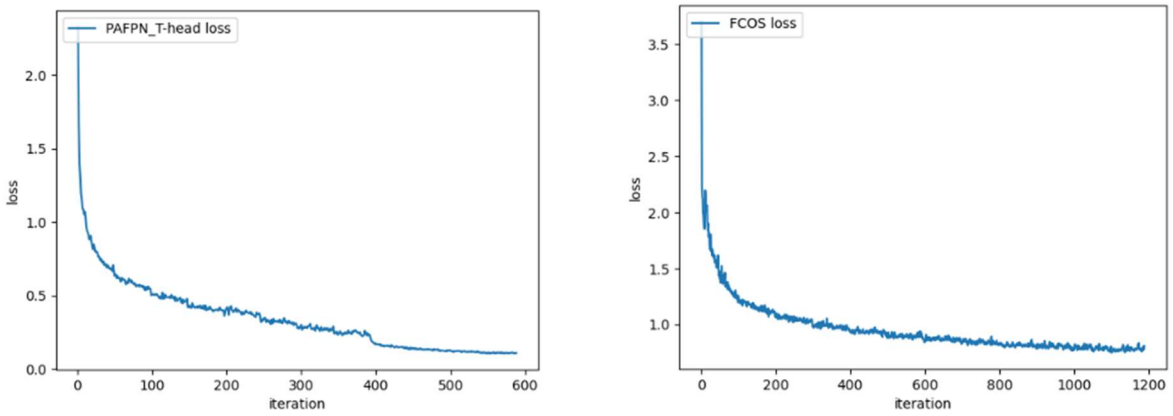


Figure 7. Loss function image

In the process of training the FCOS algorithm model, due to its small memory allocation, the number of iterations reached 1188. As shown in Figure 7 above, the Fcos loss value fluctuated between 0.779 and 0.807 to complete the convergence. Comparing the loss value, it can be seen that, The improved model in this paper has a stronger ability to fit the data samples. In Figure 8 below, the abscissa represents the number of training batches, and the ordinate represents the average accuracy of various types of model detection. It can be seen from the figure that the improved algorithm in this paper achieves the highest detection accuracy after the ninth batch of training, and the same is true for the FCOS algorithm. After the ninth training, the highest detection accuracy is achieved, and the accuracy of the subsequent batches has declined. It can be clearly seen from the figure that the improved model accuracy of the FCOS algorithm in this paper is higher than the FCOS algorithm from the beginning. itself.

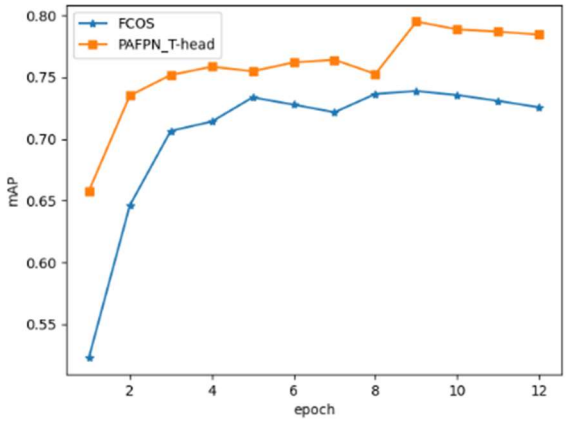
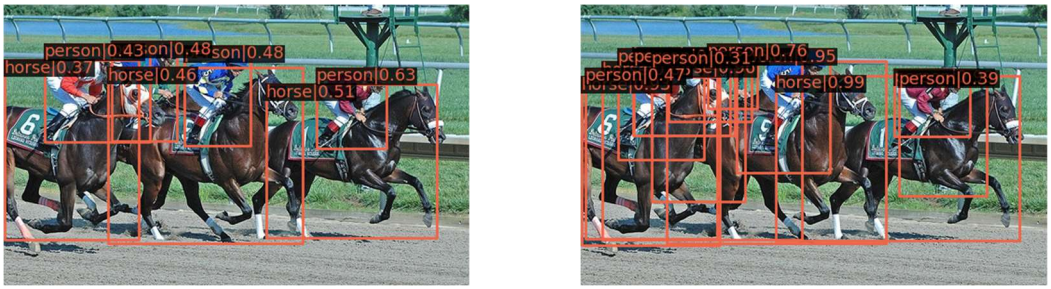


Figure 8. Comparison of model accuracy

As a single-stage detector, this paper compares its detection effect with the current mainstream two-stage detector Faster RCNN. It can be seen from the comparison charts of model detection effects in Figure 9(b) that the two-stage detector Faster RCNN although The detection accuracy of a single object is high, but because it will generate a large number of redundant anchor boxes, the detection accuracy of the coincident multi-object detection is greatly reduced. The PAFPN_T-head used in this paper is based on an anchor point. The stage detector can accurately select the detected object without

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

generating a large number of redundant anchor boxes. Table 1 shows the test results of the method in this paper and the existing mainstream two-stage object detection algorithm Faster RCNN and one-stage object detection algorithm FCOS in Pascal VOC2007/2012 trainval sum. The experimental results show that the average accuracy of the improved FCOS algorithm in this paper is increased by 5.5%, which greatly improves the detection accuracy, which shows the effectiveness of the improved algorithm in this paper. The experiment also shows that the method in this paper is better than Fster RCNN in general problems.



(a) (b)

Figure 9. Comparison diagram of detection effect

Table 1. Performance comparison of different algorithms

	Backbone	Epoch	mAP/(VOC2007/2012trainval)
Faster RCNN[15]	ResNet50	12	69.7
Faster RCNN+FPN[15]	ResNet50	12	70.5
FCOS	ResNet50	12	73.9
PAFPN_T-head	ResNet50	12	79.4

5. Conclusion

In order to effectively improve the positioning accuracy of object detection, this paper improves the FCOS object detection model. Different from the features extracted by FPN pyramid feature fusion in FCOS and input to the classification and bounding box regression network at the same time, this paper first adjusts the structure of FPN, and adds a PA structure to it to enhance the extraction of shallow geometric information, allowing The deep features are better fused with the shallow layers to get better features, and then the detection head is improved into a single-branch detection head, and the feature extractor in the single-branch detection head is used to obtain multi-scale rich feature information, allowing classification and The feature generation of localization improves the interactivity, thereby suppressing the spatial mismatch to a certain extent caused by the spatial difference between the two tasks of classification and localization due to the features extracted by different learning mechanisms, and effectively improving the detection accuracy.

References

[1] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 9627–9636, 2019.

[2] Shaoqing Ren and Kaiming He and Ross B. Girshick and Jian Sun.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1137-1149, 2015.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In Proceedingsof the European Conference on Computer Vision, pages 21–37, 2016.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0064

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. Y ou only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788, 2016.

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qing ming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 6569–6578, 2019.

[6] Tao Kong, Fuchun Sun, Huaping Liu, Y uning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing, 29:7389–7398, 2020.

[7] Shifeng Zhang, Cheng Chi, Y ongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9759–9768, 2020.

[8] Kaiming He and X. Zhang and Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770-778, 2016.

[9] Lin Tsung-Yi, Dollar Piotr, Girshick Ross, et al. Feature Pyramid Networks for Object Detection. IEEE Conference on Computer Vision and Pattern Recognition. pages 936-944, 2017.

[10] Liu Shu, Qi Lu, Qin Haifang, et al. Path Aggregation Network for Instance Segmentation. IEEE Conference on Computer Vision and Pattern Recognition, pages 8759-8768, 2018.

[11] Chengjian Feng and Yujie Zhong and Yu Gao and Matthew R. Scott and Weilin Huang. TOOD: Task-aligned One-stage Object Detection. arXiv preprint arXiv: 2108.07755, 2021.

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on ComputerVision, pages 2980–2988, 2017.

[13] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 658–666, 2019.

[14] Everingham Mark, Van Gool Luc, Williams Christopher K I, Winn John, Zisserman Andrew. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, pages 303-338, 2010.

[15] Wang Xianbao, Zhu Xiaoyong, Yao Minghai. Target detection method based on improved Faster RCNN. High-tech communication, pages 489-499, 2021.