Electronic Engineering and Informatics
G. Izat Rashed (Ed.)

© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/ATDE240137

Improved Anchor-Free Object Detection Algorithm Based on Task-Aligned One-Stage Object Detection

Zhenyi ZHANG[1], Tianping LI[2]
School of Physics and Electronic Science, Shandong Normal University, Jinan, Shandong, China

Abstract. Numerous fields, including autonomous driving, facial recognition, video monitoring, and medical picture analysis, use object detection. The Task-aligned One-stage Object Detection (TOOD) technique, however, will result in information loss and slow detection speeds when classifying and aligning objects. In order to increase the receptive field of detection and lower the amount of calculations required by the model, a Receptive Field Enhancement Module (RFEM) is added between the backbone network and the neck network of the detector in this article. This effectively addresses the issues of decreasing accuracy and slow speed. Then, in the head network, the original detecting head is switched out for an Enhanced Task Alignment Head (ET-Head) based on Layer Hybrid Attention Module (LHAM), which significantly enhances the detector's performance and feature extraction capability. Additionally, SIOU loss is used in place of regression loss to enhance the training effect. For our experiments, we use the PASCAL VOC dataset. According to experimental findings, the detection accuracy is up 1.1% compared to the TOOD algorithm, and the speed is up 1 frame per second (FPS).

Keywords. TOOD; anchor-free; hole convolution; attention mechanism; object detection

1. Introduction

In the subject of computer vision, object detection is a crucial problem with numerous applications in various facets of society. Deep learning techniques have significantly advanced the field of object recognition in recent years, particularly techniques based on convolutional neural networks like SSD [1], YOLO, Fast R-CNN [2], etc. One-stage object detection algorithms and two-stage object detection algorithms are the two categories into which object identification techniques are separated. One-stage detection algorithms include SSD, YOLO, RetinaNet, CenterNet [3], etc. The two-stage detection algorithms include R-CNN, Fast R-CNN, Faster R-CNN [4], R-FCN [5], etc. The target detection algorithm can also be divided into Anchor-based target detection algorithm and Anchor-Free target detection algorithm. SSD, YOLO, Faster R-CNN, etc. are all Anchor-based target detection algorithms. CenterNet, FCOS [6], FoveaBox, etc. are all Anchor-

[1] Zhenyi ZHANG, School of Physics and Electronic Science, Shandong Normal University, e-mail: zzy0693@163.com
[2] Corresponding author: Tianping LI, School of Physics and Electronic Science, Shandong Normal University, e-mail: [b]sdsdltp@sdnu.edu.cn

Free target detection algorithms.

However, the current two-stage object detection algorithm has the following shortcomings: slow inference speed, complex network structure, and the need for tuning and parameter tuning. Similarly, the Anchor-based target detection algorithm also has the following problems: difficulty in anchor box design, unbalanced training targets, model complexity, and high computational overhead. In order to avoid the above shortcomings, we conduct research based on the one-stage Anchor-Free target detection algorithm TOOD [7]. The core of the TOOD algorithm is to enhance the alignment between classification and localization through a new head structure and an alignment-oriented learning method. In this process, the ability and speed of network access to information will be reduced. Therefore, this paper improves TOOD's Anchor-Free algorithm.

In this paper, we add a receptive field enhancement module (RFEM) to expand the receptive field of detection. Then, the original layer attention module is replaced by a layer hybrid attention module (LHAM) to form an Enhanced Task Alignment Predictor (ETAP), which enhances the ability to acquire feature information. At the same time, replace the regression loss with SIOU loss to improve the detection effect after training. Through the improvement of the network, the detection performance of the network has been significantly improved, surpassing the original TOOD algorithm. The network model is shown in Figure 1.

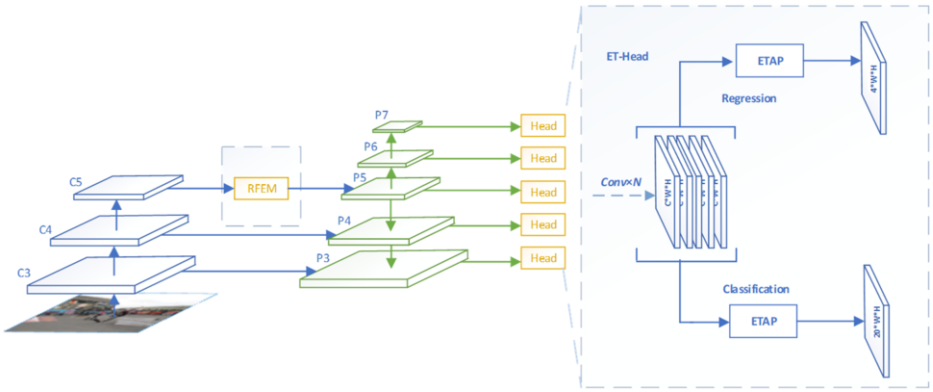

Figure 1. Improved TOOD overall model diagram

2. Proposed Method

2.1 Receptive Field Enhancement Module

In order to solve the problem of decreased accuracy and low speed. The extraction ability and detection rate of feature information at different scales can be enhanced by dilated convolutions [8] of different sizes. We combine dilated convolutional layers with different expansion rates and a spatial pooling pyramid structure, and introduce a residual structure as a receptive field enhancement module (RFEM). Among them, the receptive field enhancement module is connected with the output part of the backbone network at the top layer and the input part of the feature pyramid network.

The receptive field enhancement module is mainly composed of dilated convolutional layers. Hole convolution effectively avoids the use of pooling operations to increase the receptive field while reducing the resolution of the image, making it impossible to recover the problem of loss of detail information. The calculation formula of the dilated convolution receptive field is:

$$r_n = r_{n-1} + (k_n - 1)\prod_{n=1}^{n-1} s_i \quad (1)$$

r_n is the size of the receptive field of the nth layer. k_n is the size of the nth layer convolution kernel. s_i is the step size of the i-th layer. Three consecutive stacked 3×3 hole convolutions, if the expansion rates are 1, 2, and 4, the corresponding receptive fields are 3, 7, and 15. It can be seen that the hole convolution can increase the receptive field.

The structure of the Receptive Field Enhancement Module (RFEM) is shown in Figure 2. We use three expansion rates of 6, 12, and 18 respectively, and the convolution sum is 3×3 hole convolution to extract features to obtain different receptive fields. Then, it is combined with global average pooling and 1×1 convolution in parallel. Among them, global average pooling is to obtain global information, and 1×1 convolution is the original receptive field of the position. After all stacks are stacked in parallel, information of different scales is aggregated through a 1×1 convolution connection. The aggregated output is fused with the residual connection, and finally the fused information is output. The detection receptive field is expanded, and the context information of the feature map can be better obtained at different scales. Therefore, the accuracy and speed are improved, and the problem of occlusion can also be effectively solved.

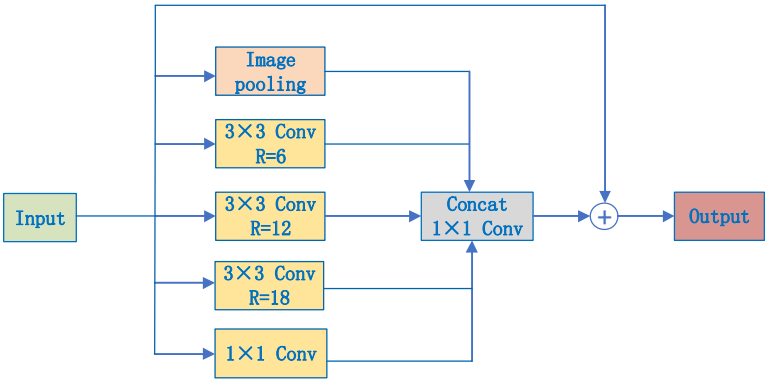

Figure 2. Receptive field enhancement module structure diagram

2.2 Layer Hybrid Attention Module

In order to improve the feature expression ability of the detector, on the basis of the task alignment predictor of the original detection head, the layer attention module is replaced by the layer hybrid attention module (LHAM), which enhances the feature information acquisition ability. Thus, a new enhanced task alignment predictor (ETAP) is formed. The enhanced task alignment predictor network structure diagram is shown in Figure 3.

The layer hybrid attention module is composed of various feature layers of different scales combined with the convolutional layer attention module (CBAM [9]). The layer

mixed attention module weights the feature map with channel dimension and spatial dimension, and selects useful information according to the weight size. It is composed of channel attention module and spatial attention module. Suppose the incoming feature map is $F \in R^{C\times W\times H}$ (C is the number of channels, W is the width, and H is the height.). Through the channel attention module $M_C \in R^{C\times 1\times 1}$, a 1D channel attention map is generated, and through the spatial attention module, a 2D spatial attention map $M_S \in R^{1\times W\times H}$ is generated, which is then transmitted to feature layers of different scales to obtain the output results of the module. The transmission formula of the whole module is:

$$CAM = M_C(F) \otimes F \tag{2}$$

$$SAM = M_C(CAM) \otimes CAM \tag{3}$$

$\otimes$ represents element multiplication, CAM represents the output feature map of the channel attention module, and SAM represents the final output feature map of the spatial attention module. In particular, we added the relu activation function at the end of the main branch of the channel attention module, which saves computation and improves the training effect.

Finally, the layer hybrid attention module effectively extracts useful information of different scales through weighting operations without introducing additional parameters and complexity, so that the network can pay more attention to the important features you want, and effectively improve the detection effect of the model.

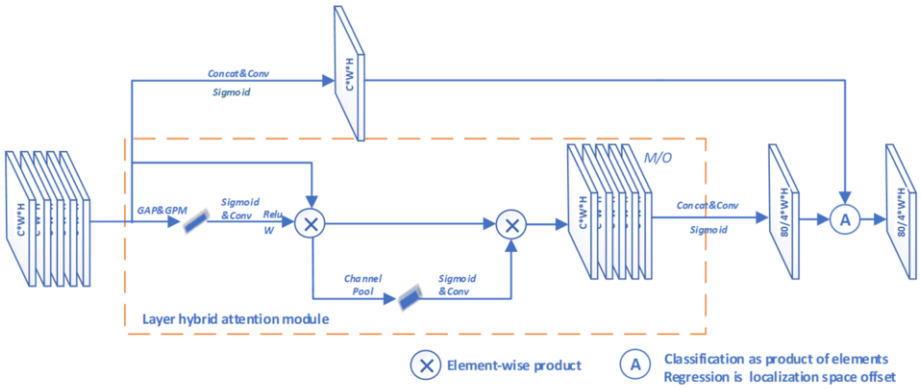

Figure 3. Enhanced Task Alignment Predictor Network Structure Diagram

2.3 Loss function

The loss function is divided into two parts: classification loss L_{cls} and bounding box regression loss L_{reg}. Using Focal Loss as the classification loss function can better deal with the category imbalance problem during training and improve the performance and generalization ability of the model. The formula of Focal Loss here is as follows:

$$L_{focalloss}(p_t) = -\alpha(1-p_t)^{\gamma}\, log(p_t) \tag{4}$$

p_t indicates the predicted probability of the model for the sample. α indicates the weight to control the positive and negative samples, the value range is [0,1], and it is set to 0.25. γ indicates a tunable hyperparameter for adjusting the importance of hard samples. We set γ to 2.0 to make the model pay more attention to difficult samples and weight misclassified samples, thus improving the model's predictive ability for minority classes.

The SIOU [10] loss function is used as a bounding box regression loss function to more accurately evaluate the similarity between object detection boxes at different scales. The formula for SIOU here is as follows:

$$Loss_{siou} = 1 - IOU + \frac{\Delta+\Omega}{2} \tag{5}$$

IOU represents the IOU loss part. Δ represents the distance loss part, i.e. $\Delta = \sum_{t=x,y}(1 - e^{-\gamma p_t})$. Here p_t represents the distance information parameter, which γ is calculated by the related $\wedge$ angle loss. Ω represents the shape loss part, i.e. $\Omega = \sum_{t=w,h}(1 - e^{-w_t})^{\theta}$. Here w_t represents the shape information parameter, the value range is [0,1], and the θ parameter controls the degree of attention to the shape loss.

From the above loss function, the final loss function of the algorithm can be obtained as:

$$L_{total} = w_{cls}L_{cls} + w_{reg}L_{reg} \tag{6}$$

w_{cls} and w_{reg} represent the weight parameters of each loss function. In this algorithm we set $w_{cls} = w_{reg} = 1.0$.

3. Experiment

3.1 Implementation Details

We use the classic PASCAL VOC [11] dataset (VOC 2007+2012) as a benchmark to evaluate our model. The dataset consists of 21503 training images and 4952 testing images. We use the part of the training image that is consistent with the test image as the validation set, and the other part as the training set. The PASCAL VOC dataset has a total of 20 categories of labels. The label categories are shown in Table 1 below.

We set the size of the input image to be 800×1333. The training epoch is set to 15. Batch size and initial learning rate are set to 4 and 0.001, respectively. Among them, the learning rate drops to 0.0001 and 0.00001 at 8 epoch and 11 epoch. Use the Momentum optimizer to optimize the training network model. The momentum of the optimizer is set to 0.9. Finally, we use random ordering and random inversion etc. to improve the robustness of training. All experiments are performed on NVIDIA RTX 3090 GPU.

Tabel 1. PASCAL VOC dataset label category

serial number	category	serial number	category
1	aeroplane	11	diningtable
2	bicycle	12	dog
3	bird	13	horse

4	boat	14	motorbike
5	bottle	15	person
6	bus	16	pottedplant
7	car	17	sheep
8	cat	18	sofa
9	chair	19	train
10	cow	20	tvmonitor

3.2 Results and Analysis

We test the performance of our improved method and some popular object detection algorithms on the PASCAL VOC dataset. As shown in Table 2, mAP represents the average precision when the IoU threshold of this category is 0.5, and FPS represents the frame rate. Our method achieves the highest mAP of 81.7% and a speed of 27FPS, which is close to the level of real-time detectors. Finally, a comparative analysis of experimental data demonstrates that our method achieves significant advantages in both accuracy and speed.

In addition, we also visualized the training results of our network model. The test results are shown in Figure 4. When detecting blurred objects, object occlusions, and small objects in images, our method can effectively classify and localize objects. Therefore, our improved network can also be well resolved for complex detection environments and small target problems.

Table 2. Detection performance results of different methods on the PASCAL VOC dataset

Methods	Backbone	Input Resolution	FPS	mAP (%)
Fast R-CNN [2]	VGG-16	600×1000	0.5	70.4
Faster R-CNN [4]	VGG-16	600×1000	2	72.2
R-FCN [5]	ResNet-101	600×1000	7	80.5
SSD300 [1]	VGG-16	300×300	-	74.3
SSD512 [1]	VGG-16	512×512	19	76.8
YOLOv2	DarkNet-19	544×544	40	76.6
YOLOv3	DarkNet-53	512×512	43	79.3
CenterNet [3]	ResNet-50	512×512	23	75.6
FCOS [6]	ResNet-50	800×1333	16	76.4
FoveaBox	ResNet-50	800×1333	-	78.4
TOOD [7]	ResNet-50	800×1333	26	80.6
Ours	ResNet-50	800×1333	27	81.7

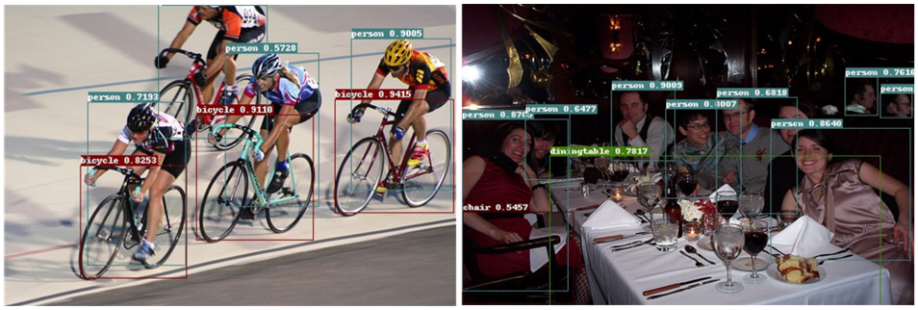

Figure 4. Pictures of test results

4. Conclusion

To summarize, we use the one-stage anchor-free detector TOOD as a benchmark to avoid the drawbacks of two-stage object detectors and anchor-based object detectors. Information loss and slow detection speed are drawbacks for the TOOD object detector. To enhance the detector's performance in terms of detection, we enhance the TOOD algorithm. In order to increase the receptive field of detection and decrease the quantity of calculation, the receptive field enhancement module of dilated convolution is first introduced. Second, the layer combines the attention module, enhancing the feature's capacity for expressiveness. According to experimental findings, our technique is faster and more accurate than a few well-known algorithms. fresh approaches and techniques for further investigation are provided.

Acknowledgments

We would especially like to thank Mark Everingham, Andrew Zisserman, and all of the other contributors who made data sets available for this article as well as the scientists who submitted experimental data.

References

[1] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). https://www.springer.com/cn

[2] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448). https://ieeexplore.ieee.org/document/466933

[3] Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*. https://arxiv.org/abs/1904.07850

[4] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*. https://ieeexplore.ieee.org/document/7485869/

[5] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, *29*. https://dl.acm.org/doi/10.5555/3157096.3157139

[6] Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627-9636). https://www.computer.org/csdl/proceedings/iccv/2019/1hQqfuoOyHu

[7] Feng, C., Zhong, Y., Gao, Y., Scott, M. R., & Huang, W. (2021, October). Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3490-3499). IEEE Computer Society. https://www.computer.org/csdl/proceedings/iccv/2021/1BmEezmpGrm

[8] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122. https://arxiv.org/abs/1511.07122

[9] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19). In Proceedings of the European conference on computer vision2018. https://link.springer.com/book/10.1007/978-3-030-01246-5

[10] Gevorgyan, Z. (2022). SIoU loss: More powerful learning for bounding box regression. *arXiv preprint arXiv:2205.12740*. https://arxiv.org/abs/2205.12740

[11] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*, 303-338. https://link.springer.com/article/10.1007/s11263-009-0275-4