Journal of Artificial Intelligence and Technology, **2021, 1, 68-73**
https://doi.org/10.37965/jait.2020.0025

ISTP
RESEARCH ARTICLE

Detection of Multiscale Center Point Objects Based on Parallel Network

Hao Chen, Hong Zheng, and Xiaolong Li
School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

(Received 12 June 2020; Revised 25 October 2020; Accepted 26 October 2020; Published online 28 December 2020)

Abstract: Anchor-based detectors are widely used in object detection. To improve the accuracy of object detection, multiple anchor boxes are intensively placed on the input image, yet. Most of which are invalid. Although the anchor-free method can reduce the number of useless anchor boxes, the invalid ones still occupy a high proportion. On this basis, this paper proposes a multiscale center point object detection method based on parallel network to further reduce the number of useless anchor boxes. This study adopts the parallel network architecture of hourglass-104 and darknet-53 of which the first one outputs heatmaps to generate the center point for object feature location on the output attribute feature map of darknet-53. Combining feature pyramid and CIoU loss function, this algorithm is trained and tested on MSCOCO dataset, increasing the detection rate of target location and the accuracy rate of small object detection. Though resembling the state-of-the-art two-stage detectors in overall object detection accuracy, this algorithm is superior in speed.

Key words: deep learning; heatmap; feature pyramid networks; object detection; center point

I. INTRODUCTION

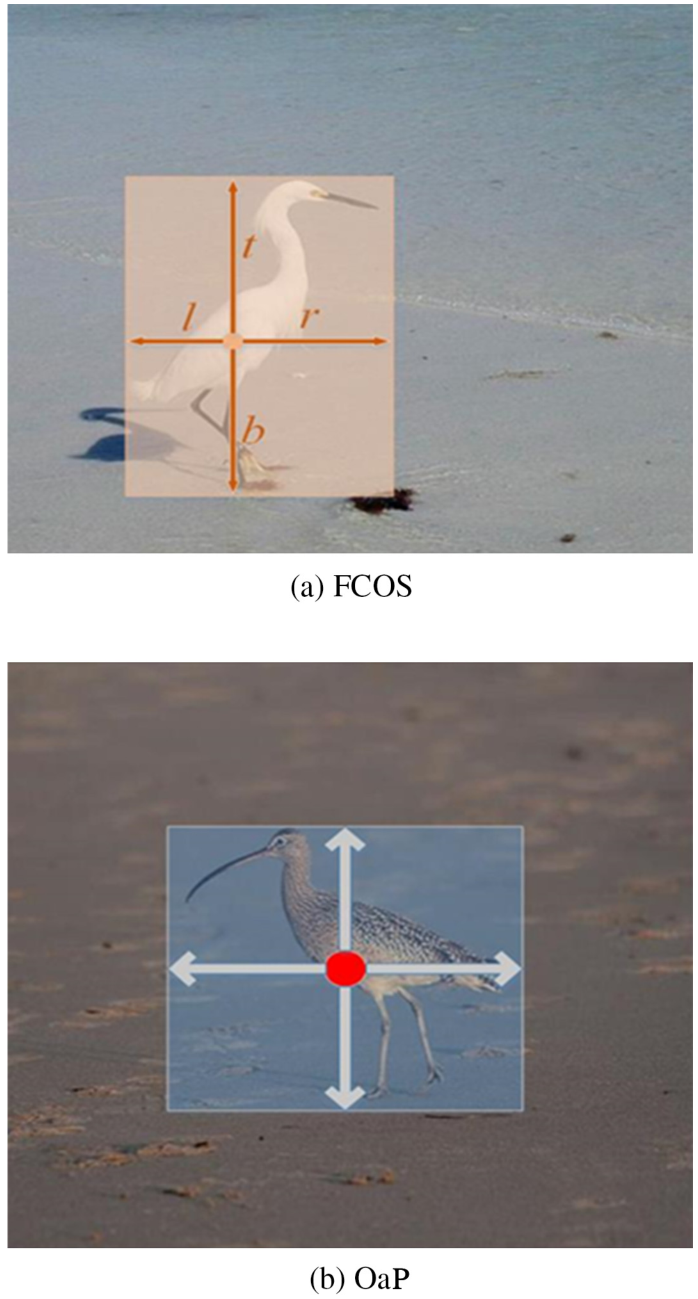
Object detection is a fundamental, and practical, research branch in the field of computer vision, practicing border and category prediction of each instance object in an image by corresponding algorithms. Current mainstream real-time anchor-based detectors, such as Faster R-CNN [1], SSD [2], and YOLOv3 [3], have achieved favorable detection results in VOC dataset [4], but the detection effect in MSCOCO dataset [5] is unsatisfactory. To improve the accuracy of target detection, rectangular bounding boxes with various sizes and aspect ratios are introduced as candidate bounding boxes. Anchor boxes are also widely adopted in one-stage detectors [2], [3], [6], [7], achieving the accuracy of two-stage detectors [1], [8], [9], under better timeliness. One-stage detectors score the anchor boxes distributed densely on the image and generate the final bounding box prediction by improving their coordinates through regression. Although these algorithms are proved successful, the following problems are still noteworthy:

1. To achieve high recall rate, anchor-based detectors are required to densely place anchor boxes on the input image. For example, more than 40K anchor boxes are needed in DSSD [6], 100K in RetinaNet [7], and 180K in feature pyramid networks (FPNs) [10] for images with short side length of 800. Most of these anchor boxes are labeled as negative samples during training; thus, the excessive negative samples aggravate the imbalance between positive and negative samples in training.
2. When computing the intersection over union (IoU) scores between all the anchor boxes and ground-truth boxes during training, the massive number of anchor boxes would also cause the surge of computation and memory consumption, lowering training speed [7].
3. The detection performance is greatly vulnerable to the size, aspect ratio, and number of anchor boxes as these hyperparameters need to be carefully tuned in anchor-based detectors [11]; thus, the training results are affected by human experience factors.

To solve the above problems, some algorithms have been improved. For example, FCOS [11] algorithm directly predicts a 4D vector plus a category at each spatial location of each layer's feature map. As shown in Fig. 1(a), the 4D vector represents the distance from the pixel point to the four borders. Besides, OaP [12] algorithm presents the target by its center point and then regresses some attributes of the target at the center point. As shown in Fig. 1(b), this algorithm turns the object detection problem into a standard key point estimation problem. In the algorithm, a heatmap of an image input into the full convolution network is obtained. The peak point of the heatmap is the center point, and the width and height of the target are predicted by position of the peak point of each feature map. Classified supervised learning is adopted for model training, and reasoning is only a single forward propagation network without postprocessing such as NMS.

The two anchor-free algorithms could effectively address the disadvantages of the anchor boxes algorithm, however, as an object detection algorithm based on semantic segmentation, FCOS contains substantial negative samples, leading to increase of computation, whereas OaP algorithm focusing on center point location ignores, the regression position of border, affecting object detection AP. Therefore, based on the two methods, a multiscale center point object detection method based on parallel network is proposed by this study. This method can make full use of the network and reduce negative samples in training, improving the accuracy of bounding boxes regression.

Corresponding author: Hao Chen (e-mail: chhao626@buaa.edu.cn).



(a) FCOS

(b) OaP

Fig. 1. Detection effect of anchor-free algorithm. The images are from MSCOCOtrain1014 dataset. (a) FCOS algorithm mainly predicts 4D vector, and (b) OaP algorithm mainly predicts the center point.

The main tasks of this paper are

- to analyze the advantages and disadvantages of existing object detection methods and propose feasible solutions to the existing object detection problems,
- to propose a multiscale center point object detection method based on parallel network and detail its theoretical formulas and calculation process, and
- to test and verify the method on the COCO dataset and compare the detection results of this method with those of the other state-of-the-art methods to demonstrate its performance.

This paper is organized as follows: Section II analyzes the advantages and disadvantages of the existing object detection methods and the existing object detection problems. Section III introduces the method presented in this paper. In Section IV, the proposed method is verified by experiments and compared with other methods. Section V constitutes a summary of the problems and methods studied in this paper.

II. RELATED WORK

Object detection is mainly used to predict the border and category of each instance object in the image. Most early algorithms are two-stage detectors using anchor boxes as the main detection method, bearing high detection rate yet poor timeliness. In recent years, one-stage detectors improved based on the two-stage ones become popular sharing both favorable timeliness and detection rate. As anchor boxes and related hyperparameters are removed by some of the improved algorithms, the learning characteristics of the machine are fully exploited.

A. ANCHOR-BASED DETECTORS

The traditional sliding window and proposals-based detectors are inherited by early anchor-based detectors, where anchor boxes are regarded as predefined sliding windows or proposals, and then classified as positive or negative samples. An additional offset regression is also needed to correct the prediction of the frame position. Therefore, anchor boxes in these detectors can be viewed as training samples. Different from the early detectors, such as Fast R-CNN [13], which repeatedly calculates image features for each sliding window, anchor boxes avoid repeated feature calculations and increase detection speed by feature maps of convolutional networks. However, separate proposals are still relied on by the algorithm and cannot be trained end-to-end. Anchor boxes with lower scores are removed by region generation network (RPN) that introduced by Faster R-CNN, and end-to-end training is implemented through joint training of RPN and detection network. In most of the early anchor-based detectors, a set of sparse regions of interest generated is classified by the network, which is what we called two-step method.

To improve the detection speed, the step of region proposals is removed by some researchers to detect the target directly in a single network, which is what we called one-stage method. Anchor boxes are placed densely on the multiscaled feature maps by the SSD algorithm, and each anchor box is classified and refined directly. Though the accuracy and speed of operation are somewhat both ensured, the highest accuracy of the two-stage method still cannot be reached. In YOLOv3, dimension clusters are used as anchor boxes, and multiscale target regression detection is achieved by feature pyramid. The measured accuracy of YOLOv3 is almost the same as two-stage algorithm, but its timeliness is significantly better than that of two-stage method. However, multiple hyperparameters need to be adjusted by anchors, influencing the final accuracy and thereby leaving the detection results of anchor-based detectors vulnerable to artificial presets.

B. ANCHOR-FREE DETECTORS

Anchor-free is not a new concept. As the earliest anchor-free model in the field of target detection, YOLOv1 [14] regards target detection as a spatially separated boundary box and related probabilistic regression problem. The bounding box and classification score can be predicted directly from the entire image. Although this method shares high operation speed, its accuracy is unsatisfactory.

After CornerNet [15] was published in 2018, CornerNet target detection models emerged one after another. Currently, the principle of the main anchor-free detection methods is to replace anchor boxes with key points or intensive predictions. CornerNet, ExtremeNet [16], and OaP are based on key points, and FASF [17], FCOS, and FoveaBox [18] on DenseBox [19].

CornerNet algorithm turns object detection frame into a pair of key points, that is, the upper-left corner and the lower-right corner, to eliminate design of anchor boxes. Corner pooling technology is also adopted in CornerNet, for CNN's better location of corner position. ExtremeNet turns target detection into pure key point estimation issue, in which a target frame is formed by four extreme points and one center point of the target. Similar to the algorithm flow of CornerNet, ExtremeNet generates the target frame only when the response of the five heatmaps predicted by CNN for each target class in the geometric center is large enough. OaP algorithm takes the target as a single point, which is the center point of the bounding box found by key point estimate, to regress other target attributes through the center.

Based on the online feature selection ability of FPN, FSAF algorithm can dynamically allocate each instance to the most suitable feature layer during training, work together with the module branch with anchors during reasoning, and finally output prediction in parallel. Developed from semantic segmentation, FCOS dispenses with anchor box and region recommendation, avoiding overlapping calculation and performance-sensitive parameter design in model training. FCOS presents a new loss function "Center-ness" to lower the score weight of the bounding boxes far from the center of the object, curtailing low-quality detection boxes without introducing other hyperparameters. Imitating the central fovea of human eyes (i.e., the center of view shares the highest visual acuity), FoveaBox predicts where the object's central area may exist and the bounding box of each valid location. Due to the characteristic representation of feature pyramid, targets of different scales can be detected from multiple feature layers. The core of FoveaBox is to directly learn the probability of the existence of targets and the coordinate position of target box, including prediction of category-related semantic maps and generation of category-irrelevant candidate target boxes, whose size is related to representation of feature pyramid.

Our method is a refined version of the above methods. Taking the parallel structure of Darknet-53 and Hourglass-104 as the backbone, drawing on the idea of FPN, the proposed method detects targets of different sizes by multiple scales. Darknet-53 is used to extract the features of the target and Hourglass-104 to locate the target. This method will increase the expression space of the original features, weaken the correlation between features, and increase the generalization ability of the model.

III. METHOD

A. OVERVIEW

One of the necessary conditions for deep neural network to perform better in detection is that the network can effectively extract the feature information of the detected object. At present, CNN mainly uses one backbone to extract all the feature information of the targets to ensure their correlation, which would affect features' expression space shared by multiple features. As location and attribute information are paramount in object detection, we propose a multiscale center point object detection method based on parallel network to extract object attribute features by Darknet-53 and location features by Hourglass-104. Finally, the two features are combined to detect the object.

B. NETWORK ARCHITECTURE

The algorithm in this paper is refined based on FCOS and OaP methods: (1) The algorithm extracts target attribute features by FCOS method and converts the backbone to Darknet-53 that bears the same accuracy while higher speed (0.47 times faster) that Resnet-101 [3]; (2) Hourglass-104 is used as the parallel backbone to extract target location features and multiscale heatmaps to locate the cells containing the target in the feature map; (3) IoU loss function is replaced by DIoU or CIoU [20]. The structure of the algorithm is shown in Fig. 2.

As shown in Fig. 2, Darknet-53 and Hourglass-104 generate multiscale feature maps and heatmaps in parallel. The feature map of the same scale label corresponds to the heatmap of the same scale label, and the center point generated by heatmaps locates the cell in the feature map. The algorithm uses the cell that contains the center point to complete object detection. Compared with the original FCOS algorithm, this method reduces the number of invalid cells that needs to be calculated and improves the initial state of valid cell. Therefore, Center-ness in the original FCOS method is removed.

C. LOSS FUNCTION

The training loss function is defined as follows:

$$L(\{p_{x,y,c}\},\{t_{x,y}\}) = \frac{-1}{N_{\text{pos}}}\sum_{x,y,c} L_{cls}(p_{x,y,c}, p^*_{x,y,c}) + \frac{1}{N_{\text{pos}}}\sum_{x,y} \mathrm{I}_{x,y} L_{reg}(t_{x,y}, t^*_{x,y}), \tag{1}$$

where L_{cls} removed represents the focal loss [7], $p_{x,y,c}$ removed represents predicted classification scores, $p^*_{x,y,c}$ removed is the class label, N_{pos} removed represents the number of positive samples, L_{reg} removed represents the improved IoU loss function [20], $\mathrm{I}_{x,y}$ removed represents the indicator function (being 1 if there is a center point in the cell and 0 otherwise), $t_{x,y} = [l,t,r,b]$ removed represents the distance from the location to the four sides of bounding boxes, and $t^*_{x,y} = [l^*,t^*,r^*,b^*]$ removed represents the distance from the location to the four sides of the ground-truth bounding boxes. Then, $t^*_{x,y}$ is obtained by the following operation:

$$l^* = x - x_0^{(i)}, \tag{2}$$

$$t^* = y - y_0^{(i)}, \tag{3}$$

$$r^* = x_1^{(i)} - x, \tag{4}$$

$$b^* = y_1^{(i)} - y, \tag{5}$$

where (x,y) is the position of the feature points in the F_{i} map back to the input map, and $(x_0^{(i)},y_0^{(i)})$ and $(x_1^{(i)},y_1^{(i)})$ are the coordinates of the left-top and right-bottom corners of the bounding box.

$$L_{\text{cls}} = \begin{cases} (1-p_{x,y,c})^{\alpha}\log(p_{x,y,c}) & \text{if } p^*_{x,y,c} = 1 \\ (1-p^*_{x,y,c})^{\beta}(p_{x,y,c})^{\alpha}\log(1-p_{x,y,c}) & \text{otherwise} \end{cases}, \tag{6}$$

$$L_{\text{reg}} = 1 - \mathrm{IoU} + R(B,B^{gt}), \tag{7}$$

where $R(B,B^{gt})$ denotes the penalty term for predicted box B and target box B^{gt}.

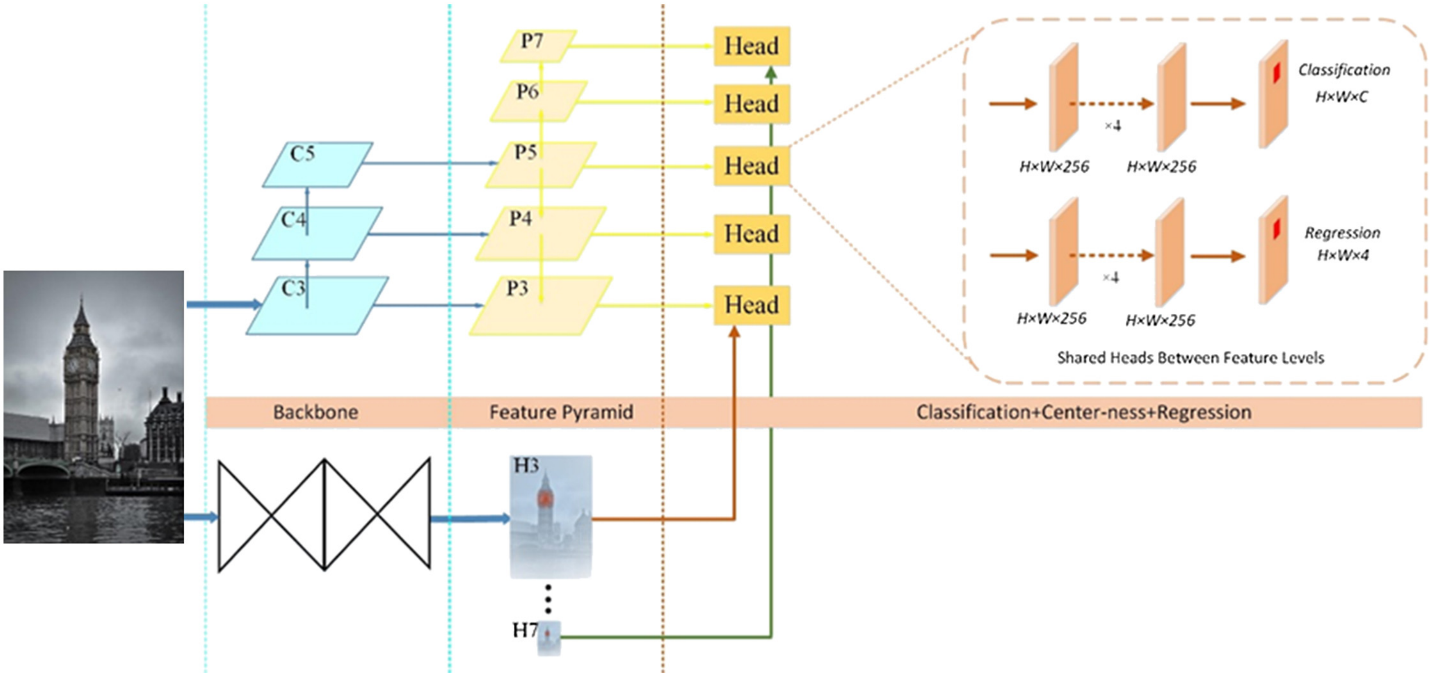


Fig. 2. Network architecture of the proposed model. C3–C5 denote the feature maps of the backbone network. P3–P7 denote the feature levels used for the final prediction. H3 denotes the heatmaps of the Hourglass-104 network, and H3–H7 denote the heatmaps used for the center point. All the numbers are computed with an input of 1024 × 1024.

$$R_{\mathrm{DIoU}}(B,B^{\mathrm{gt}}) = \frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}, \tag{8}$$

where b and b^{gt} stand for the central points of B and B^{gt}, $\rho(\cdot)$ stands for the Euclidean distance, and c stands for the diagonal length of the smallest enclosing box covering the two boxes.

$$R_{\mathrm{CIoU}}(B,B^{gt}) = R_{\mathrm{DIoU}}(B,B^{gt}) + \alpha v, \tag{9}$$

$$\alpha = \frac{v}{1 - \mathrm{IoU} + v}, \tag{10}$$

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

where α is a positive trade-off parameter, and v removed measures the consistency of aspect ratio.

IV. EXPERIMENTS

Conducted on the large-scale detection benchmark COCO [5] according to the common practices [10], [11, [12], our experiments adopted COCO trainval 35K split (115K images) for training and minival split (5K images) for ablation validation. The training data were the COCO trainval 35K split, including all the 80K images from training and a random subset of 35K images from the 40K val split. The detection results of heatmap and multiscale heatmap for attribute feature map location were tested, respectively, along with the results of different IoU loss functions. The method's COCO AP was also compared to that of the state-of-the-art methods on the test-dev split.

The training details are as follows. Unless specified, Darknet-53 [3] was used as the backbone network, and the used hyperparameters were the same as those of RetinaNet [15]. Specifically, the network was trained with stochastic gradient descent for 90K iterations with the initial learning rate of 0.01 and a minibatch of 16 images. The learning rate was reduced by a factor of 10 at iterations 60K and 80K. The values of weight decay and momentum were set as 0.0001 and 0.9, respectively.

A. ABLATION STUDY

In the experiment, Hourglass-104 outputs single-scale heatmaps and multiscale heatmaps. The head generated by the center point jointly produced by the output attribute feature maps, and Darknet-53 heatmaps was trained, and corresponding test results were compared with those of Yolov3, FCOS, and OaP algorithms, as shown in Table 1.

From Table I, it can be seen that absorbing the strong points of other algorithms, the multiscale center point object detection method based on parallel network has scored remarkable results compared with other combined algorithms. The detection rate of small targets has particularly been improved as the heatmap generated by Hourglass-104 is conducive to location of small targets, intensifying attribute feature network's learning of small targets. Moreover, the multiscale constraints in FCOS method are also conducive to learning of different scale targets by feature pyramid.

FCOS improves detection rate by Center-ness, but IoU or GIoU is used as the loss function in L_{reg}. Yet in this method, the cell is located by the center point, reducing invalid cells that need to be calculated. Therefore, the Center-ness in the original FCOS method is removed.

It is predicted that taking DIoU or CIoU as the loss function may generate a better combination (with the center point) and better object detection effect as three important geometric factors should be valued for good loss of bounding box regression, that is, the overlapping area, the distance between center points, and the aspect ratio. DIoU is superior to GIoU in center point distance and CIoU to DIoU in aspect ratio [20]. Therefore, CIoU can effectively balance the impact of small object in the loss function.

It can be observed from Table II that DIoU and CIoU achieve better detection results as their combination with the center point is indeed better than that of the original IoU loss function. The difference emerges because the increased constraints in the improved loss functions and the parameters in the center point can be better integrated, balancing the loss of small objects. The improved loss functions overlap the effect of Center-ness in part, but both are conducive to improving small object training results.

B. COMPARISON WITH THE STATE-OF-THE-ART DETECTORS

The final test results of the proposed method are compared with those of the state-of-the-art algorithms, as shown in Table III.

It can observe from Table III that due to application of anchor-free methods, current one-stage detectors can almost reach the accuracy of two-stage detectors after years of development, let alone the advantage of speed. The reason why anchor boxes can achieve high accuracy is that preset anchors that can cover almost all the targets are introduced for every point on the feature map. The more presets entail, the more computation and the higher accuracy. However, in the actual scenarios, the utilization rate of these preset

Table I. Comparison table of related algorithm characteristics on COCO test-dev (single-scale/multiscale test for OaP and our method)

Method	Backbone	AP	AP50	AP75	APS	APM	APL
YOLOv3	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
FCOS	ResNet-101-FPN	41.0	60.7	44.1	24.0	44.1	51.0
OaP	Hourglass-104	42.1/45.1	61.1/63.9	45.9/49.3	24.1/26.6	45.5/47.1	52.8/57.7
Ours	Darknet-Hourglass	42.6/45.5	61.7/64.3	46.2/49.6	25.2/28.2	45.7/47.9	52.9/57.8

Table II. IoU loss function comparison table. DIoU and CIoU are better than traditional IoU loss function. CIoU loss function is more conducive to network learning of small objects, yet the overall object detection rate records limited improvement

Method	Backbone	AP	AP50	AP75	APS	APM	APL
FCOS	ResNet-101-FPN	41.0	60.7	44.1	24.0	44.1	51.0
Ours_GIoU	Darknet-Hourglass	45.5	64.3	49.6	28.2	47.9	57.8
Ours_DIoU	Darknet-Hourglass	45.7	66.2	50.7	29.3	48.2	57.9
Ours_CIoU	Darknet-Hourglass	46.0	67.1	51.2	30.1	48.5	57.9

Table III. Ours_CIoU vs. the other state-of-the-art comparison on COCO test-dev. Top: two-stage detectors; bottom: one-stage detectors. We show multiscale testing for most one-stage detectors

Method	Backbone	AP	AP50	AP75	APS	APM	APL	Time (ms)
Two-stage methods:								
Faster R-CNN w/FPN	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2	172
MaskRCNN	ResNeXt-101	39.8	62.3	43.4	22.1	43.2	51.2	91
SNIPER [21]	DPN-98	46.1	67.0	51.6	29.6	48.9	58.1	400
TridentNet [22]	ResNet-101-DCN	48.4	69.7	53.5	31.8	51.3	60.3	1429
One-stage methods:								
SSD513	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8	125
YOLOv3	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9	51
RetinaNet	ResNeXt-101-FPN	40.8	61.1	44.1	24.1	44.2	51.2	185
CornerNet(multi)	Hourglass-104	42.1	57.8	45.3	20.8	44.8	56.7	244
ExtremeNet(multi)	Hourglass-104	43.7	60.5	47.0	24.1	46.9	57.6	323
FASF(multi)	ResNeXt-101	44.6	65.2	48.6	29.7	47.1	54.6	370
FCOS	ResNet-101-FPN	41.0	60.7	44.1	24.0	44.1	51.0	74
OaP(multi)	Hourglass-104	45.1	63.9	49.3	26.6	47.1	57.7	128
Ours_CIoU	Darknet-Hourglass	46.0	67.1	51.2	30.1	48.5	57.9	182

anchor boxes is not high; thus, many operations are actually invalid, to which the anchor-free method serves as a solution by reducing the calculation caused by massive of invalid anchor boxes. Anchor-free method does not explicitly preset the size and scale of various anchor boxes in each location, but the location information is still reserved. It can be equivalently considered that anchor-free transforms all kinds of anchor boxes in each position into one anchor. As a result, the number of anchor boxes is linearly reduced, yet most of them are still useless. Combining two anchor-free methods, this study further reduces the number of useless anchor boxes by introducing the central point of the heatmap scoring better detection results than most of the current object detectors. However, as the heatmap generation process of Hourglass-104 is time consuming, the speed of the algorithm has not been explicitly improved, requiring further improvement in the future.

V. CONCLUSION

This study adopts parallel backbone architecture in the deep learning model, in which Hourglass-104 outputs the heatmap and Darknet-53 the attribute feature map of the original map. The center point generated by heatmap can be used to locate the target feature on the attribute feature map, further reducing the number of anchor boxes and optimizing the initial state of anchor regression on the basis of anchor free. Two parallel backbones are simultaneously employed to separately extract attribute features and position features of the target to lower the correlation between different features and enlarge their expression space, thereby enhancing the learning of target features by the network. In this method, multiscale and CIoU loss function are jointly adopted to improve the detection accuracy of small targets. However, due to the time-consuming heatmap generation process of Hourglass-104, the proposed algorithm's speed is affected though better than that of two-stage detectors with the same accuracy. Further speed improvement is needed for real-time detection.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster RCNN: Towards real-time object detection with region proposal networks," in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in Proc. Eur. Conf. Comp. Vis., Springer, 2016, pp. 21–37.

[3] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[4] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2014.

[5] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in Proc. Eur. Conf. Comp. Vis., Springer, 2014, pp. 740–755.

[6] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," arXiv preprint arXiv:1701.06659, 2017.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2017, pp. 2980–2988.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2014, pp. 580–587.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. Comput. Vis., 2017.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2017, pp. 2117–2125.

[11] Z. Tian, C. Shen, H. Chen, et al., "FCOS: Fully convolutional one-stage object detection," in Proc. IEEE Int. Conf. Comput. Vis., 2019.

[12] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv: 1904.07850, 2019.

[13] R. Girshick, "Fast R-CNN," arXiv preprint arXiv:1504.08083, 2015.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Uni_ed, real-time object detection," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016, pp. 779–788.

[15] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proc. Eur. Conf. Comput. Vision, 2018, pp. 734–750.

[16] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2019.

[17] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2019.

[18] T. Kong, F. Sun, H. Liu, et al., "FoveaBox: Beyond anchor-based object detector," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2019.

[19] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., arXiv preprint arXiv:1509.04874, 2015.

[20] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, Distance-IoU loss: Faster and better learning for bounding box regression. Presented at the AAAI Conf. Artificial Intelligence (AAAI), 2020.

[21] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient multi-scale training," in Proc. NIPS, 2018.

[22] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," arXiv preprint arXiv:1901.01892, 2019.

ABBREVIATIONS

The following abbreviations are used in this manuscript:

COCO	Common Objects in COntext
VOC	The pascal visual object classes challenge: A retrospective
AP	Average Precision
R-CNN	Rich feature hierarchies for accurate object detection and semantic segmentation
Faster R-CNN	Faster RCNN: Towards Real-time Object Detection with Region Proposal Networks
SSD	Single Shot Multibox Detector
YOLOv1	You only look once: Unified, real-time object detection
YOLOv3	Yolov3: An Incremental Improvement
DSSD	Deconvolutional Single Shot Detector
FCOS	Fully Convolutional One-Stage Object Detection
OaP	Objects as Points
MS COCO	Microsoft Common Objects in COntext
FASF	Feature Selective Anchor-Free Module for Single-Shot Object Detection
CNN	Convolutional Neural Networks
GIoU	Generalized Intersection over Union
DIoU	Distance-IoU
CIoU	Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression