CenterNet Based on Diagonal Half-length and Center Angle Regression for Object Detection

Yuantian Xia[1], XuPeng Kou[1], Weie Jia[1], Shuhan Lu[2], Longhe Wang[3*], and Lin Li[1*]
[1] College of Information and Electrical Engineering, China Agricultural University
Beijing, 100083, China
[e-mail: xiayuantian@cau.edu.cn, lilincau@126.com]
[2] School of Information, University of Michigan
Ann Arbor, 48104, USA
[e-mail: shuhanlu@umich.edu]
[3] National Research Facility for Phenotypic and Genotypic Analysis of Model Animals
Beijing, 100193, China
[e-mail: Phil.wang@cau.edu.cn]
[*]Corresponding author: Longhe Wang, Lin Li

Received April 25, 2022; revised March 22, 2023; accepted July 14, 2023; published July 31, 2023

Abstract

CenterNet, a novel object detection algorithm without anchor based on key points, regards the object as a single center point for prediction and directly regresses the object's height and width. However, because the objects have different sizes, directly regressing their height and width will make the model difficult to converge and lose the intrinsic relationship between object's width and height, thereby reducing the stability of the model and the consistency of prediction accuracy. For this problem, we proposed an algorithm based on the regression of the diagonal half-length and the center angle, which significantly compresses the solution space of the regression components and enhances the intrinsic relationship between the decoded components. First, encode the object's width and height into the diagonal half-length and the center angle, where the center angle is the angle between the diagonal and the vertical centreline. Secondly, the predicted diagonal half-length and center angle are decoded into two length components. Finally, the position of the object bounding box can be accurately obtained by combining the corresponding center point coordinates. Experiments show that, when using CenterNet as the improved baseline and resnet50 as the Backbone, the improved model achieved 81.6% and 79.7% mAP on the VOC 2007 and 2012 test sets, respectively. When using Hourglass-104 as the Backbone, the improved model achieved 43.3% mAP on the COCO 2017 test sets. Compared with CenterNet, the improved model has a faster convergence rate and significantly improved the stability and prediction accuracy.

Keywords: Object detection, CenterNet, Prediction stability, Accuracy consistency, Convergence speed

This work was supported by the National Key R&D Program of China(2021ZD0113701).

1. Introduction

As an important research direction in computer vision, object detection has been successfully applied in practical scenarios such as unmanned. Meanwhile, object detection is the foundation of other complex vision tasks, such as image segmentation and object tracking.

Limited by the performance of the computing device, early detectors is based on machine learning. Traditional machine learning methods rely on feature engineering and need to design features manually, resulting in limited feature expression ability. In addition, the method of machine learning also needs to design specific classifiers for different application scenarios, resulting in a serious shortage of generalization ability of the model. With the rapid development of high-performance computers and memory devices, deep learning technology has become the mainstream method to solve computer vision tasks. Neural networks based on deep learning are widely used in object detection models of different frameworks because of their strong adaptive feature extraction and recognition capabilities.

Most of the mainstream detectors are based on the anchor box. The size of different objects is counted through the clustering algorithm first to generate a series of prior boxes for the selected data set. For a specific dataset, these anchors will be used as hyperparameters to assist the model in completing the detection task, which effectively improves the detection accuracy. However, the anchor also brings a series of problems. First, the anchors' size, number, and aspect ratio will seriously impact the detection performance. Some experiments show that adjusting these hyperparameters can increase the AP of Retinanet [1] on the COCO[2] dataset by 4%. Second, these fixed-size anchors significantly impair the universality of the detectors. When the detectors face different tasks and datasets, parameters such as the size of the anchor boxes must be reset. Third, many anchor boxes need to be generated to match the ground-truth box. However, most of them will be marked as negative samples during training, which artificially leads to an imbalance between samples. Finally, during the training process, the IOU between all anchors and ground truths needs to be calculated, resulting in a lot of memory and time consumption.

In order to solve the negative impact of the anchor mechanism, some anchor-free object detectors[3-5] based on the keypoint prediction have been proposed in recent years. These methods regard the object as one or more key points for prediction, and they contain both the location and category information of the object. CenterNet[5] directly regresses the width and height for each predicted key point and decodes it into the bounding box for the corresponding object. Since the prior boxes are not used in the training and inference process, the post-processing process, such as NMS(non-maximum suppression) and other extra computation brought by the anchor mechanism, is eliminated, making the detectors simple and efficient. **Fig. 1** shows the network structure of CenterNet.

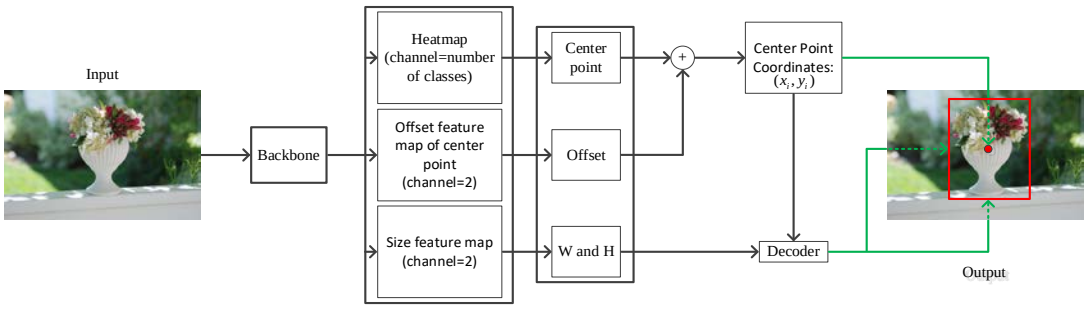


Fig. 1. The structure of CenterNet. Where Heatmap is the center point heatmap.

As shown in **Fig. 1**, CenterNet predicted each object's center point position and category through the center point heatmap and corrected the center point coordinates through the offset feature maps with two channels. The two channels represent the center point in Horizontal and vertical offsets. However, due to the vast difference between the object's width and height, directly regressing them will lead to a vast solution space, which increases the training difficulty of the model, slows the convergence speed, and increases the instability of the prediction results. Moreover, since the width and height are independently predicted in the two channels, the lack of intrinsic correlation between the width and height at the same position will lead to inconsistent prediction accuracy.

In this paper, we proposed the diagonal half-length and center angle regression method to address the above problems. The method dramatically compresses the model's solution space for size information prediction, reduces the training difficulty of the model, and enables the model to converge more quickly. In addition, it also enhanced the intrinsic correlation between the prediction components and improved the consistency of the prediction accuracy.

The following are the main contributions of this paper:

1. In the encoding stage, the object's width and height are encoded into the diagonal half-length and the center angle. Compared with the original width and height, the solution space of the object's size information for prediction will be further compressed.

2. In the decoding stage, the predicted diagonal half-length and the center angle are jointly decoded into two length components in the horizontal and vertical directions where the center point is located, respectively. Since the two length components are obtained by joint decoding of the diagonal half-length and the center angle, the consistency of prediction accuracy is guaranteed.

The rest of the paper is as follows: In section 2, we briefly reviewed the main research work for object detection. In section 3, we propose and introduce the improved method in detail. In section 4, we verify the proposed method's effectiveness and performance by conducting extensive experiments and comparing the improved method with other state-of-the-art detection models. In section 5, we gave the conclusion and summarization.

2. Related Work

With the development of neural network models, many models with high recognition accuracy and feature extraction ability [6-10] have significantly improved the performance of the detectors.

Girshick et al. [11] proposed R-CNN laying the foundation for the subsequent two-stage object detection algorithm. Then Girshick et al. [12] proposed Fast-RCNN base on R-CNN. Fast-RCNN improves the training speed by nine times and the test speed by more than 200 times. However, they all use the selective search [13] algorithm, which is time-consuming. Subsequently, Ren et al. [14] proposed the Faster-RCNN, which used an RPN network to generate candidate regions. They proposed an anchor mechanism for the regression and classification of object location information for the first time, realizing end-to-end training and prediction. He et al. [15] proposed Mask R-CNN, which integrates the dual functions of object detection and instance segmentation and improves the ability of object detection to solve more complex visual tasks. Cai et al. [16] proposed Cascade R-CNN to train multiple Cascade detectors using different IoU thresholds. It can train a higher-quality detection model without reducing the number of samples and improve the detection performance degradation caused by IoU threshold selection in Faster-RCNN. Guan et al. [17] fully studied and improved models such as Faster-RCNN and proposed an image object detection and classification

method based on deep neural network (DNN), improving the quality of object location and classification enhanced the detection performance of the model. Object localization and classification always have object detection's core and key problems. Guan et al. [18] conducted research on this problem and proposed a region-based efficient network that treats object detection as a dual problem of object proposal generation and object classification to detect image objects accurately.

Although the two-stage detectors' performance has been dramatically improved, the detection speed is still inadequate. The YOLOv1 proposed by Redmon et al. [19] uses a single network architecture to simultaneously predict the object's position and category information, significantly improving the detection speed. Liu et al. [20] proposed SSD using multi-scale feature prediction. It introduced the anchor mechanism and used FPN [21] to predict objects on feature maps of different scales. Leng et al. [22] improved the feature fusion method of SSD and proposed ESSD. It adopts bidirectional feature fusion to utilize features of different scales. Fu et al. [23] proposed DSSD, which uses the deconvolution module and adds context information to give the low-level feature map better feature expression ability. Subsequently, the YOLO series of models has been continuously developed. YOLOv2 [24] improved YOLOv1 and used darknet-19 as the backbone, improving detection speed and accuracy. YOLOv3 [25] uses the DarkNet-53 as the backbone. It adds an upsampling-based feature fusion operation based on FPN so that the model can extract the object features more accurately. Songtao et al. [26] summarized and improved the training techniques that have achieved excellent detection performance in recent years and proposed YOLOv4. It adopts CSPNet [10] as the backbone and PAN [27] as the feature fusion network, which improves the model feature extraction and fusion capabilities. Tan et al. [28] proposed EfficientDet by improving the FPN network. It is based on the weighted bidirectional feature pyramid network BiFPN, enabling the model to perform multi-scale feature fusion more conveniently and quickly.

Except for YOLOv1, the above models are all based on the anchor box mechanism. In order to solve the negative impact of the anchor box mechanism on detection performance, some object detection algorithms without anchors have attracted wide attention in recent years. Tian et al. [29] proposed FCOS for pixel-level prediction. First, it reconstructs the detection object in a per-pixel predictive method. Then the method of multi-scale prediction using FPN improves recall and resolves ambiguity caused by overlapping boundaries. Finally, the center-ness branch suppressed the detected low-quality bounding boxes, reducing false-positive boxes and significantly improving detection performance. Law et al. [3] proposed CornerNet based on the keypoint prediction method. It regards the object as a pair of key points and predicts the top-left and bottom-right heatmaps of the object through a single convolutional neural network and the embedding vector for each corner point. Embedding vectors are used to group corners that belong to the same object. To further enhance the accuracy of keypoint-based prediction, Zhou et al. [4] proposed ExtremeNet based on CornerNet, which performs object localization by detecting the four poles of the object. The four poles and the center area of the object are predicted respectively through five heatmaps and combined with the poles of different heatmaps. Although the method of multiple keypoint prediction improves the performance of anchor-free detectors, predicting too many key points increases the difficulty of late matching. Zhou et al. [5] proposed CenterNet, an anchor-free object detector based on center point prediction that only regards the object as a center key point for prediction. First, predict the center area of the object through the center point heatmaps, then adjust the center point through the offset feature maps, and finally, regress the object's width and height. There is no need for complex matching work since there is only one key point for prediction, making

the model simpler and more efficient.

3. Proposed Method

Unlike the anchor-based object detection algorithm, the anchor-free algorithm has a larger and more flexible solution space. Because it does not introduce prior distribution knowledge in advance, avoiding the computational load and post-processing process caused by anchors. Therefore, the anchor-free method improved the accuracy and performance of the detectors.

However, the overly flexible and huge solution space dramatically increases the model's training difficulty, making it difficult to converge during the training process. Moreover, it generates too many false positives, reducing the prediction accuracy. This problem is particularly prominent for CenterNet, which directly regresses the object's width and height. In **Table 1**, We count the width and height information of all objects in the Pascal VOC [30] datasets.

Table 1. The width and height range of objects in the Pascal VOC datasets

Dataset	Wide range	High range
VOC 2007	[8, 499]	[1, 499]
VOC 2012	[4, 499]	[1, 499]

In this section, we proposed a diagonal half-length and center angle regression method to solve the problem mentioned above. The algorithm first re-encodes the object's original width and height information, reduces the regression value range to compress the solution space, and makes the model easier to converge. Then the prediction results are jointly decoded, which enhances the correlation between the decoded components and improves the consistency of prediction accuracy.

3.1 Encoding

In order to compress the huge difference in width and height between objects, we encode them into the diagonal half-length and the center angle corresponding to the object's bounding box. The encoding process is shown in **Fig. 2**.

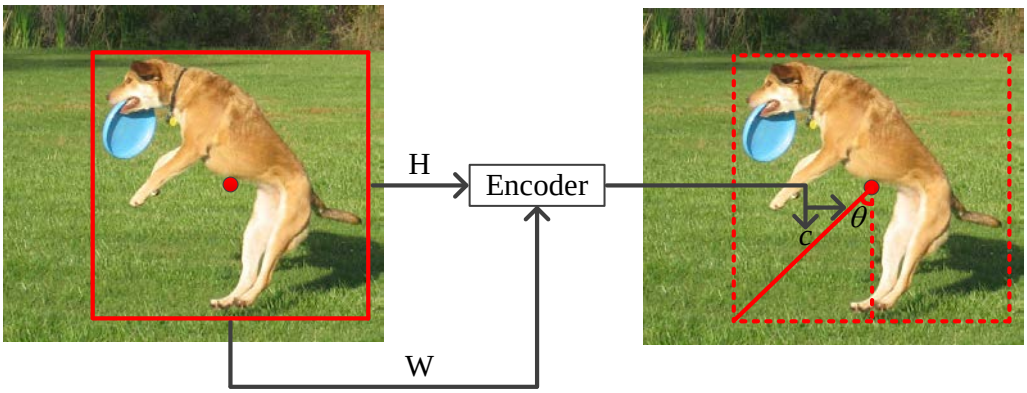


Fig. 2. Diagonal half-length and center angle encoding process. Where H is the height of the object bounding box and W is the width of the object bounding box. c is the half-length of the diagonal after encoding, and θ is the center angle after encoding.

As shown in **Fig. 2**, a right triangle with the center angle θ as the vertex is formed between the diagonal of the object bounding box and the vertical centerline passing through the object's center point. We encode the object's width and height into the hypotenuse length c and vertex angle θ of the triangle. The model will not predict the object's width and height but directly predict c and θ by the size feature map with two channels. According to the Pythagorean theorem, the solution space of the re-encoded hypotenuse length c (that is, the half-length of the diagonal) is smaller than the object's width and height. In addition, no matter how the object's width and height change, the variation range of the center angle θ is constantly kept between (0, π/2). Compared with the vast value range generated by the network's prediction of the object's width and height, the re-encoded prediction scalar significantly compresses the solution space, accelerates the model's convergence speed, and increases the stability of the model prediction accuracy. The calculation process is expressed as follows:

$$c = \frac{\sqrt{w^2 + h^2}}{2} \tag{1}$$

$$\theta = \arctan(\frac{w}{h}) \tag{2}$$

where c is the half-length of the diagonal and θ is the center angle.

For the VOC 2007 dataset, the re-encoded range is reduced from the original [8, 499] and [1, 499] to [$\frac{\sqrt{65}}{2}$, 353] and (0, π/2), respectively. For the VOC 2012 dataset, the re-encoded value range is reduced from the original [4, 499] and [1, 499] to [$\frac{\sqrt{17}}{2}$, 353] and (0, π/2), respectively.

Because the semantic information predicted by the network has been changed, we adjust the loss function responsible for width and height prediction in the original network. We use $(x_1^{(k)}, y_1^{(k)}, w^{(k)}, h^{(k)})$ to denote the bounding box range of the object k, which belongs to the category c_k. Where $x_1^{(k)}$ and $y_1^{(k)}$ are the abscissa and ordinate of the lower-left corner vertex of the object k, respectively. $w^{(k)}$ and $h^{(k)}$ are the width and height of the object k, respectively. The center point of the object k can be denoted as $p_k = (x_1^{(k)} + \frac{w^{(k)}}{2}, y_1^{(k)} - \frac{h^{(k)}}{2})$. For each possible object center point p_k, we use the L1 loss function to regress the encoded component $s_k = (c^k, \theta^k)$, where c^k is the diagonal half-length corresponds to the bounding box of object k, and θ^k is the center angle of object k. The loss function of the prediction component expressed as follows:

$$L_{c\theta} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{p_k} - s_k\right| \tag{3}$$

where N is the number of samples, and $\hat{S}_{p_k}$ is the predicted value of s_k corresponding to the center point p_k. The loss calculation process of the prediction component is shown in **Fig. 3**.

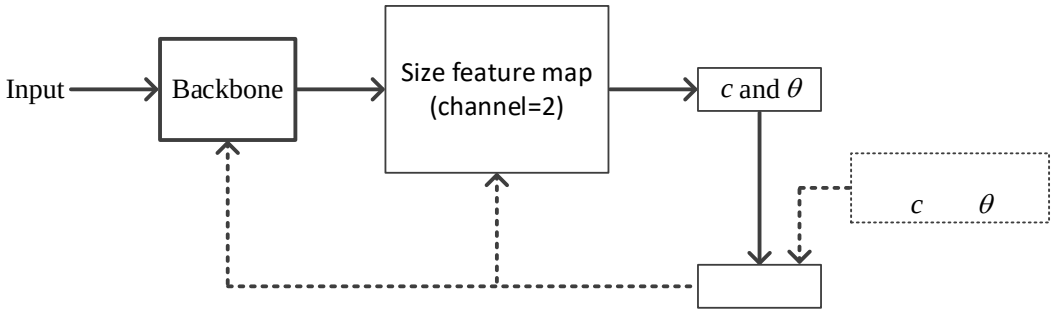


Fig. 3. The loss function calculation process. The size feature map is the predicted component feature map with two channels, and Ground truth is the encoded value of c and θ.

As shown in **Fig. 3**, after Backbone, a feature map with two channels is output, and the two channels are the predicted values of c and θ, respectively. By calculating the loss value, the network will gradually obtain better and more accurate predicted values of c and θ.

Because the computational load of the encoding process is very low, it is all done in the process of loading the dataset before training. Therefore, there will be no impact on model training and inference speed. In addition, since each object's width and height are recorded with limited real-number bits, there will be no data type error or overflow during the encoding conversion process.

3.2 Decoding

In CenterNet, the object's width and height are independently predicted in two channels in the feature map, which leads to inconsistency in prediction accuracy, such as the prediction accuracy of one component is high and the other is low. We jointly decoded the c and θ obtained in Section 3.1 into two length components in the horizontal and vertical directions of the object center point coordinates, respectively. Because the values of the two components depend on the joint decoding of c and θ simultaneously, there is a strong internal dependency between the decoded two components, which enhances the consistency of prediction accuracy. The decoding process is shown in **Fig. 4**.

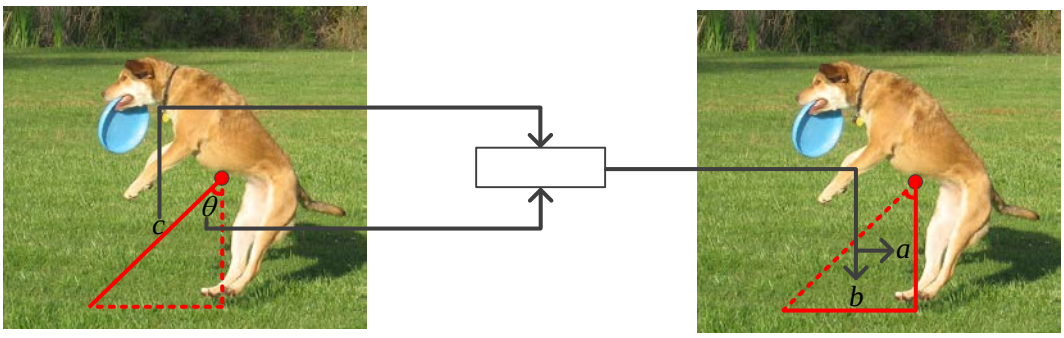


Fig. 4. Decoding process. where a is the decoded vertical length component and b is the horizontal length component.

As shown in **Fig. 4**, c and θ are extracted from the two channels of the feature maps output. They and the object's center point form a right triangle with c as the hypotenuse and

θ as the apex angle. Then use trigonometric functions to decode them into a length component a in the vertical direction of the center point and a length component b in the horizontal direction of the center point. The calculation process is expressed as follows:

$$a = \cos\theta \times c \tag{4}$$

$$b = \sin\theta \times c \tag{5}$$

Because the definition domain of θ is between (0, π/2), the value domain of $\cos\theta$ and $\sin\theta$ is between (0, 1). Therefore, applying the smaller and uniform value domain of $\cos\theta$ and $\sin\theta$ to adjust c can make a and b have better accuracy stability and uniformity, thereby improving the prediction accuracy and performance of the model.

We can obtain the predicted object's bounding box position by calculating the decoded length component and the predicted coordinates of the center point of the corresponding position. The calculation process is expressed as follows:

$$\hat{P}_c = \left\{\left(\hat{x}_i, \hat{y}_i\right)\right\}_{i=1}^{n} \tag{6}$$

$$\hat{P}_{c:left-bottom} = \left(\hat{x}_i - b_i, \hat{y}_i + a_i\right) \tag{7}$$

$$\hat{P}_{c:right-top} = \left(\hat{x}_i + b_i, \hat{y}_i - a_i\right) \tag{8}$$

where $\hat{P}_c$ is the set of n detected center points coordinates of class c. $\hat{x}_i$ and $\hat{y}_i$ are the abscissa and ordinate prediction values of the center point of the i-th object in class c, respectively. a_i and b_i are the decoded length components corresponding to the center point of the i-th object in class c. $\hat{P}_{c:left-bottom}$ and $\hat{P}_{c:right-top}$ are the coordinates of the lower-left and upper-right bounding boxes of the i-th object in class c, respectively. The bounding box generation process is shown in **Fig. 5**.

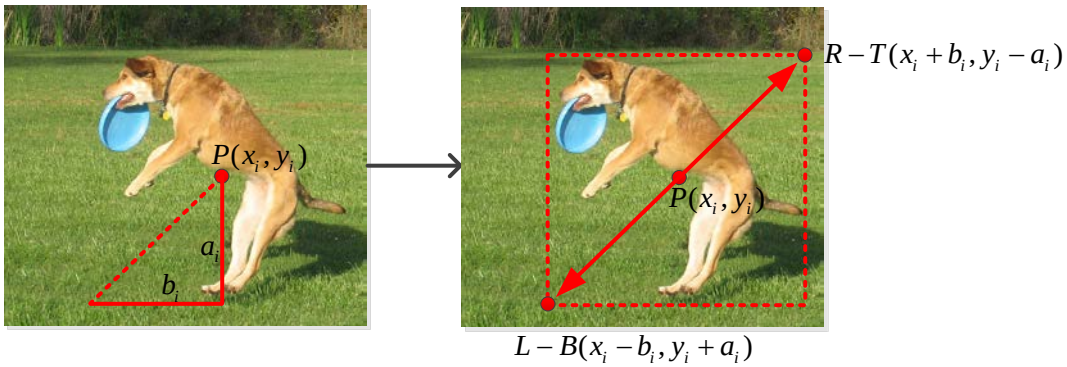


Fig. 5. The bounding box generation process. P is the center point predicted by the network. $L-B$ and $R-T$ are the coordinates of the lower-left corner vertex and the upper-right corner vertex of the object, respectively.

As shown in **Fig. 5**, the decoded length components a_i and b_i can be used to predict the object's bounding box efficiently and accurately. In addition, there is no complex calculation in the decoding and prediction process, which will not affect the inference speed.

4. Experiments

4.1 Datasets

In this section, we apply the diagonal half-length and center angle regression algorithm to CenterNet and conduct extensive experiments using the PASCAL VOC and Microsoft COCO datasets.

As one of the most commonly used benchmark datasets in computer vision, the VOC is widely used to evaluate and validate object detection algorithms. It has two versions, VOC 2007 and 2012, with 20 categories. The VOC 2007 has 5011 images for training and validation and 4952 for testing. The VOC 2012 is an upgraded version of 2007, and it contains a larger number of images and more complex situations than the VOC 2007.

Compared with the VOC dataset, the Microsoft COCO dataset has a more complex background, a higher number of objects, and many smaller objects, so the visual task on the COCO dataset is more challenging. We selected the most widely used COCO 2017 version for model training and testing. The COCO 2017 has 118287 images for training, 5000 images for validation, and 40670 for testing. The total number of images was 163,957 in 80 categories. The COCO dataset images contain natural and common object images in daily life.

4.2 Implement details and evaluation metric

For the result of VOC 2007, we use the trainval set of 2007 and 2012 for training and the 2007 test set for testing. For the result of VOC 2012, we use the trainval set of 2007 and 2012, the test set of 2007 for training, and the 2012 test set for testing. We choose Resnet50 and load its pre-trained parameters as Backbone. We set the input size to 512×512, the initial learning rate to 0.0005, the momentum to 0.9, the weight decay to 0.0005, and the batch size to 32, respectively. We use the stochastic gradient descent SGD algorithm to train the model for 150 epochs. The learning rate is dynamically adjusted using warm-up and cosine annealing functions. Other optimization and data enhancement methods are consistent with their corresponding baseline.

For the COCO dataset, we use train 2017 and Val 2017 to train and verify the model and evaluate our proposed algorithm in test 2017. Unlike the VOC dataset, the COCO dataset uses the new AP metric instead of traditional mAP as the most important metric for detection performance evaluation, which is calculated based on 10 IoU thresholds and the mean of all 80 classes. We choose Hourglass-104 and load its pre-trained parameters as Backbone. We use the stochastic gradient descent SGD algorithm to train the model for 100 epochs and set the initial learning rate to 0.00025. The rest parameter settings were consistent with the training on the VOC. Image enhancement methods consistent with CenterNet.

The results of CenterNet and our improved model are obtained by training with the above parameter settings. The results of other algorithms are from their corresponding papers. The hardware and software environment of the experiment is shown in **Table 2**.

Table 2. The software and hardware environment

Equipment	Type
CPU	Intel core i9-9900k
GPU	NVIDIA GeForce RTX 3090
RAM	32.0 GB
OS	WIN10 64
Develop software	Python3.8+Pytorch1.10.0+cuda11.3+Pycharm

4.3 Comparision of loss value

We recorded the size loss and total loss of each epoch during the training process, respectively, and the results are shown in **Fig. 6** and **Fig. 7**.

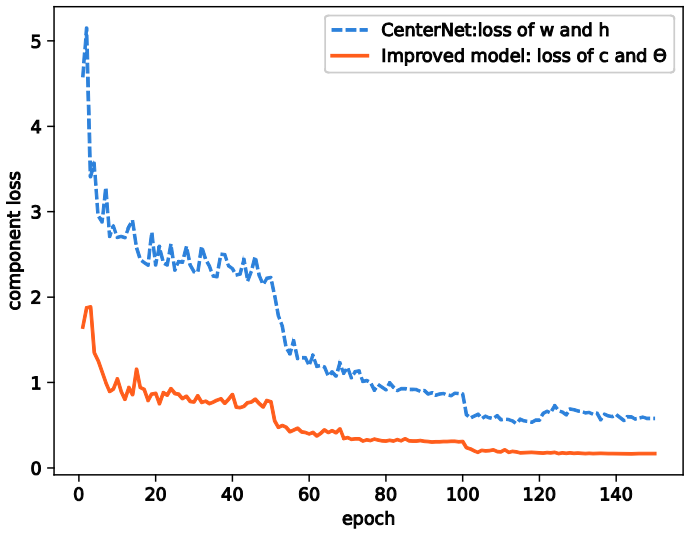


Fig. 6. Comparison curves of size components loss values.

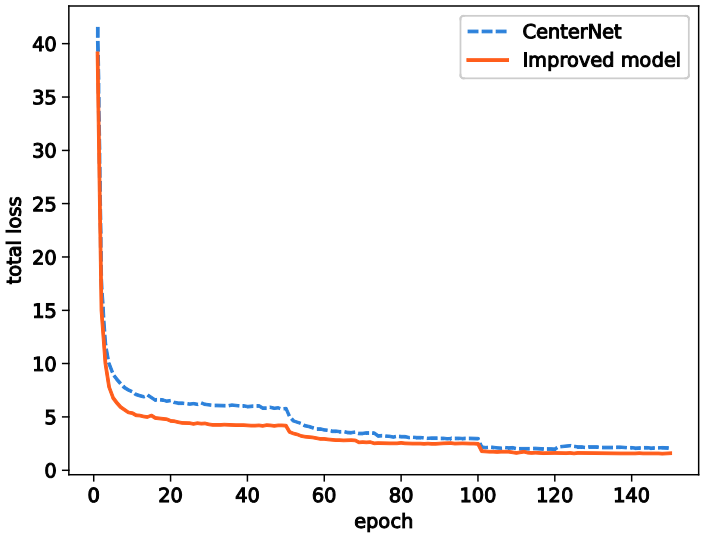


Fig. 7. Comparison curves of total loss values.

As shown in **Fig. 6**, the diagonal half-length and center angle regression algorithm has a faster convergence speed and lower loss value than the direct regression of width and height. In addition, by observing the loss curve after the 100th epoch, it can be found that compared with the width and height loss curve, the loss curve of the improved model is smoother without excessive fluctuations, so it has better stability and accuracy. **Fig. 7** shows the change in the total loss curve, which is equal to the sum of center point loss, offsets loss, and size loss. Since

our proposed method makes the size loss have a faster convergence speed and lower loss value, it speeds up the convergence speed and accuracy of the entire model and improves the stability.

4.4 Results on PASCAL VOC 2007 test set

Table 3 shows the results of our proposed algorithm and other detectors on the VOC 2007 test set. To simplify the expression, we abbreviate the CneterNet based on the diagonal half-length and center angle regression proposed in this paper as DC-CenterNet. It can be observed from **Table 3** that DC-CenterNet achieves 81.6% mAP, which is 1.4% higher than CenterNet with width-height regression. In addition, the categories of boats, cars, and plants often exist smaller or occluded instances. Compared to the baseline, mAP for these challenging categories is improved by 3.9%, 0.1%, and 11.5%, respectively,

The above results fully illustrate that our method improves the model's convergence speed and stability and significantly enhances the detection accuracy and ability to deal with complex scenes, especially the categories with complex object scenes and many small objects. Compared with other detectors, our improved model achieves the best mAP. It achieves 0.1% higher than DSSD with Residual101, 3% higher than YOLOv2 with Darknet19, 4.8% higher than SSD with VGG, 5.2% higher than Faster R-CNN with Residual101, and 8.4% higher than Faster R-CNN with VGG.

Table 3. Experimental results on VOC 2007 test set

Method	Faster R-CNN	Faster R-CNN	SSD512	YOLOv2	DSSD513	CenterNet	DC-CenterNet
Backbone	VGG16	Residual101	VGG	Darknet19	Residual101	ResNet50	ResNet50
mAP	73.2	76.4	76.8	78.6	81.5	80.2	81.6
aero	76.5	79.8	82.4	80.1	86.6	82.1	84.8
bike	79.0	80.7	84.7	81.2	86.2	89.8	90.9
bird	70.9	76.2	78.4	77.1	82.6	80.6	83.5
boat	65.5	68.3	73.8	69.2	74.9	66.7	70.6
bottle	52.1	55.9	53.2	56.3	62.5	58.7	64.9
bus	83.1	85.1	86.2	85.6	89.0	90.5	90.8
car	84.7	85.3	87.5	85.4	88.7	91.2	91.3
cat	86.4	89.8	85.0	88.6	88.8	90.7	91.3
chair	52.0	56.7	57.8	58.2	65.2	65.0	64.9
cow	81.9	87.8	83.1	88.2	87.0	83.8	80.6
table	65.7	69.4	70.2	70.1	78.7	75.1	73.8
dog	84.8	88.3	84.9	88.6	88.2	88.5	87.5
horse	84.6	88.9	85.2	89.2	89.0	89.9	90.6
mbike	77.5	80.9	83.9	81.2	87.5	89.9	90.8
person	76.7	78.4	79.7	791	83.7	86.5	86.4
plant	38.8	41.7	50.3	50.2	51.1	42.0	53.3
sheep	73.6	78.6	77.9	79.2	86.3	82.0	82.9
sofa	73.9	79.8	73.9	80.2	81.6	80.2	79.0
train	83.0	85.3	82.5	85.6	85.7	89.4	87.7
tv	72.6	72.0	75.3	72.2	83.7	83.8	85.6

4.5 Results on PASCAL VOC 2012 test set

Compared with the VOC 2007 dataset, the VOC 2012 dataset has more images and more complex object scenes. **Table 4** shows the results of our proposed algorithm and other detectors on the VOC 2012 test set. It can be observed from **Table 4** that DC-CenterNet with diagonal half-length and center angle regression achieves 79.7% mAP, which is 0.9% higher than CenterNet with width-height regression. In the categories of boats, cars, and plants with smaller objects and more mutual occlusion, the accuracy is improved by 2.2%, 0.7%, and 1.2%, respectively. Because DSSD uses a deeper Residual101 as the feature extraction network, which increases its ability to solve more complex detection tasks, the mAP is 0.3% higher than DC-CenterNet using ResNet50 as the feature extraction network. Compared with other detectors, our improved model achieves the best mAP. It achieves 6.3% higher than YOLOv2 with Darknet19, 4.8% higher than SSD with VGG, 5.9% higher than Faster R-CNN with Residual101, and 9.3% higher than Faster R-CNN with VGG.

Table 4. Experimental results on VOC 2012 test set

Method	Faster R-CNN	Faster R-CNN	SSD512	YOLOv2	DSSD513	CenterNet	DC-CenterNet
Backbone	VGG16	Residual101	VGG	Darknet19	Residual101	ResNet50	ResNet50
mAP	70.4	73.8	74.9	73.4	80.0	78.8	79.7
aero	84.9	86.5	87.4	86.3	92.1	81.6	82.3
bike	79.8	81.6	82.3	82.0	86.6	88.4	87.9
bird	74.3	77.2	75.8	74.8	80.3	81.0	79.2
boat	53.9	58.0	59.0	59.2	68.7	68.9	71.1
bottle	49.8	51.0	52.6	51.8	58.2	62.8	63.5
bus	77.5	78.6	81.7	79.8	84.3	87.7	88.5
car	75.9	76.6	81.5	76.5	85.0	88.6	89.3
cat	88.5	93.2	90.0	90.6	94.6	88.1	90.0
chair	45.6	48.6	55.4	52.1	63.3	62.4	64.2
cow	77.1	80.4	79.0	78.2	85.9	77.3	77.9
table	55.3	59.0	59.8	58.5	65.6	72.5	74.9
dog	86.9	92.1	88.4	89.3	93.0	83.5	84.7
horse	81.7	85.3	84.3	82.5	88.5	88.5	88.4
mbike	80.9	84.8	84.7	83.4	87.8	87.8	88.9
person	79.6	80.7	83.3	81.3	86.4	83.4	84.8
plant	40.1	48.1	50.2	49.1	57.4	50.9	52.1
sheep	72.6	77.3	78.0	77.2	85.2	80.3	78.8
sofa	60.9	66.5	66.3	62.4	73.4	75.7	77.5
train	81.2	84.7	86.3	83.8	87.8	84.5	85.3
tv	61.5	65.6	72.0	68.7	76.8	83.0	82.6

4.6 Results on MS COCO 2017 test set

As shown in **Table 5**, when using CenterNet with Hourglass as the baseline, our improved model achieves 43.3% AP, which is 1.2% higher than the baseline. For different IOU threshold metrics AP_{50} and AP_{75}, our improved model outperforms the baseline by 0.4% and 1.1%, respectively. For evaluating the metrics AP_S, AP_M, and AP_L of different size objects, our improved model outperforms the baseline model by 1.9%, 1.2%, and 0.1%, respectively. Compared with other state-of-the-art detectors, our improved model achieves the best mAP.

Table 5. Experimental results on COCO 2017 test set

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SSD513	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513		33.2	53.3	35.2	13.0	35.4	51.1
YOLO v2	Darknet-19	21.6	44.0	19.2	5.0	22.4	35.5
CornerNet	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
ExtremNet		40.1	55.3	43.2	20.3	43.2	53.1
CenterNet		42.1	61.1	45.9	24.1	45.5	52.8
DC-CenterNet		43.3	61.5	47.0	26.0	46.7	52.9

The results show that our method can significantly improve the detection performance, especially on the more stringent AP_{75} and most challenging small object detection performance metrics AP_S, the accuracy of our improved model has increased substantially.

4.7 Ablation studies

In addition to detection accuracy, detection speed is also an important index to measure object detector performance. In order to better illustrate the effectiveness of our method, we compared detection speeds before and after improvement. **Table 6** shows the comparison results:

Table 6. Comparison of inference speed

Method	Backbone	Input	mAP	FPS
CenterNet	Hourglass-104	512	42.1	28.7
DC-CenterNet			43.3	28.7

Table 6 show that our method will not have any impact on the detection speed of baseline models when the same backbone and input size are used. This is because the coding process is all completed in the process of loading data sets before training, and the computing load on the equipment is very low. It can be seen that our method obtains higher detection accuracy with the same reasoning time.

4.8 Visualization of detection results

We provide a qualitative comparison between the improved model and CenterNet to further illustrate our proposed method's superiority in terms of bounding box prediction accuracy and complex scene object detection. As shown in **Fig. 8**, the accuracy of the bounding box position predicted by CenterNet has obvious inconsistency. Because CenterNet's prediction of width and height is independent, there is no correlation between the prediction components. Therefore, in some pictures, the wide prediction is accurate, but the height is inaccurate, or the high prediction is accurate, but the width is inaccurate. Compared with CenterNet, we generate bounding boxes of objects by jointly decoding the prediction components. Therefore, when the detection accuracy is similar, our improved model has higher accuracy and consistency in predicting the bounding box position.

Fig. 8. Quality comparison of bounding box predictions. The objects with detection scores higher than 0.5 are shown. In each pair, the top is the detection results of CenterNet, and the bottom is our improved model.

Because our proposed algorithm can make more accurate predictions and fine-tune the object bounding box, the model's detection ability in more complex scenarios such as small objects and dense connections is enhanced. As shown in **Fig. 9**, our improved model detects more small objects that CenterNet fails to detect. In addition, our improved model is also more robust to heavily occluded objects, such as occluded cars in the second column of images and birds occluding each other in the last column.

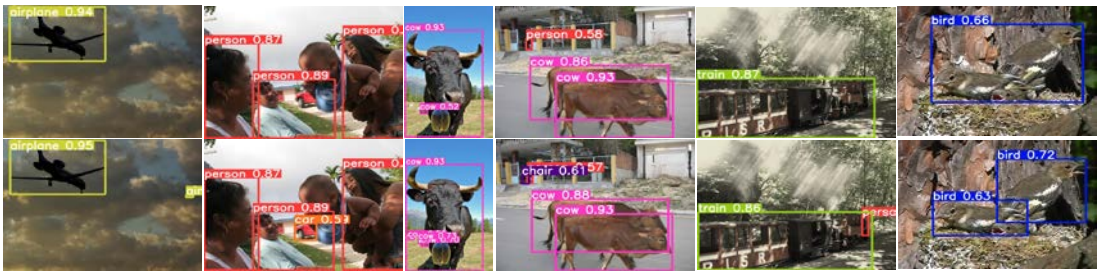


Fig. 9. Comparison of prediction results in complex scenarios. The objects with detection scores higher than 0.5 are shown. In each pair, the top is the detection results of CenterNet, and the bottom is our improved model.

The above qualitative comparison results fully demonstrate that our proposed algorithm can significantly improve the prediction accuracy and quality of object bounding boxes and the detection ability in complex scenes.

5. Conclusion

In this paper, we propose a diagonal half-length and center angle regression method and apply it to CenterNet instead of its direct regression prediction of object width and height. First, the algorithm encodes the object’s width and height into the diagonal half-length and center angle, and the network performs regression prediction on the encoded diagonal half-length and center angle of each object. Compared with the direct regression of width and height, the encoded components have smaller solution space, accelerate the model’s convergence speed, and improve the stability and detection performance. Secondly, the predicted diagonal half-length and the center angle are decoded into the center point's horizontal and vertical length components. Combined with the predicted coordinates of the center point, the predicted bounding box can be accurately decoded. Compared with decoding the bounding box directly by using the width and height, the length component after joint decoding has a stronger

intrinsic correlation and fineness, improving prediction accuracy consistency. Finally, the experimental results show that the diagonal half-length and center angle regression method can significantly improve the convergence speed of the baseline CenterNet, enhancing its stability and prediction accuracy.

References

[1] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42, no.2, pp. 318-327, 2020. Article(CrossRef Link)

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. of European conference on computer vision*, Zurich, Switzerland, pp.740-755, September, 2014. Article(CrossRef Link)

[3] H. Law, J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. of ECCV 2018*, Munich, Germany, pp. 765-781, September 2018. Article(CrossRef Link)

[4] X. Zhou, J. Zhuo and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," in *Proc. of CVPR 2019*, Long Beach, USA, pp. 850-859, June 2019. Article(CrossRef Link)

[5] X. Zhou, D. Wang, et al, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019. Article(CrossRef Link)

[6] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. Article(CrossRef Link)

[7] C.Szegedy et al, "Going deeper with convolutions," in *Proc. of CVPR 2015*, Long Beach, USA, pp. 1-9, June 2015. Article(CrossRef Link)

[8] M. Lin, Q. Chen, S. Yan, "Network In Network," *arXiv Preprint arXiv: 1312.4400*, 2014. Article(CrossRef Link)

[9] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of CVPR 2016*, Las Vegas, USA, pp. 770-778, June 2016. Article(CrossRef Link)

[10] C. -Y. Wang, H. -Y. Mark Liao, Y. -H. Wu, P. -Y. Chen, J. -W. Hsieh and I. -H. Yeh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," in *Proc. of CVPRW 2020*, Seattle, USA, pp. 1571-1580, June 2020. Article(CrossRef Link)

[11] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proc. of CVPR 2014*, Columbus, USA, pp. 580-587, June 2014. Article(CrossRef Link)

[12] R. Girshick, "Fast R-CNN," in *Proc. of ICCV 2015*, Santiago, Chile, pp. 1440-1448, December 2015. Article(CrossRef Link)

[13] Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T. et al, "Selective Search for Object Recognition," *Int.J. Comput Vis*, vol. 104, pp. 154-171, 2013. Article(CrossRef Link)

[14] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, June 2017. Article(CrossRef Link)

[15] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," in *Proc. of CVPR 2017*, Honolulu, USA, pp. 2980-2988, July 2017. Article(CrossRef Link)

[16] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in *Proc. of CVPR 2018*, Salt Lake City, USA, pp. 6154-6162, June 2018. Article(CrossRef Link)

[17] Y. Guan, M. Aamir, Z. Hu, Z. A. Dayo, Z. Rahman, W. A. Abro, P. Soothar, "An Object Detection Framework Based on Deep Features and High-Quality Object Locations," *Traitement du Signal*, vol.38, no.3, pp.719-730, 2021. Article(CrossRef Link)

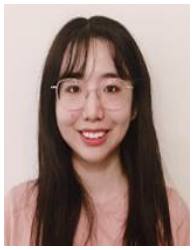
[18] Y. Guan, M. Aamir, Z. Hu, W. A. Abro, Z. Rahman, Z. A. Dayo, S. Akram, "A Region-Based Efficient Network for Accurate Object Detection," *Traitement du Signal*, vol.38, no.2, pp.481-494, 2021. Article(CrossRef Link)

[19] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. of CVPR 2016*, Las Vegas, USA, pp. 779-788, June 2016. Article(CrossRef Link)
[20] W. Liu, et al, "SSD: Single Shot MultiBox Detector," in *Proc. of ECCV 2016*, Amsterdam, Netherlands, pp. 21-37, October 2016. Article(CrossRef Link)
[21] S. W. Kim, H. K. Kook, J. Y. Sun, M. C. Kang, S. J. Ko, "Parallel Feature Pyramid Network for Object Detection," in *Proc. of ECCV 2018*, Munich, Germany, pp. 239-256, September 2018. Article(CrossRef Link)
[22] J. Len, Y. Liu, "An enhanced SSD with feature fusion and visual reasoning for object detection," *Neural Comput & Applic*, vol. 31, pp. 6549-6558, October 2019. Article(CrossRef Link)
[23] C. Y. Fu, W. Liu, A. Ranga, et al, "DSSD: deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017. Article(CrossRef Link)
[24] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proc. of CVPR 2017*, Honolulu, USA, pp. 6517-6525, July 2017. Article(CrossRef Link)
[25] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. Article(CrossRef Link)
[26] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020. Article(CrossRef Link)
[27] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," in *Proc. of CVPR 2018*, Salt Lake City, USA, pp. 8759-8768, June 2018. Article(CrossRef Link)
[28] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *Proc. of CVPR 2020*, Seattle, USA, pp. 10778-10787, June 2020. Article(CrossRef Link)
[29] Z. Tian, C. Shen, H. Chen and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *Proc. of CVPR 2019*, Long Beach, USA, pp. 9626-9635, June 2019. Article(CrossRef Link)
[30] Everingham, M., Van Gool, L., "The PASCAL Visual Object Classes (VOC) Challenge," *Int.J. Comput Vis*, vol. 88, pp. 303-338, 2010. Article(CrossRef Link)

Yuantian Xia is a doctoral candidate at the College of Information and Electrical Engineering, China Agricultural University. His research interests include image processing, deep learning, and computer vision.

Xupeng Kou is a doctoral candidate at the College of Information and Electrical Engineering, China Agricultural University. His research interests include image processing, machine learning, and deep learning.

Weie Jia is a master's student at the College of Information and Electrical Engineering, China Agricultural University. Her research interests include image recognition and image classification.



Shuhan Lu is a master's student at the School of Information, University of Michigan. Her research interests include machine learning, deep learning, and artificial intelligence.

Longhe Wang is a researcher at the National Research Facility for Phenotypic and Genotypic Analysis of Model Animals. His research interests include artificial intelligence, deep learning, and intelligent agriculture.

Lin Li is a professor and doctoral supervisor in the Department of Computer Engineering, College of Information and Electrical Engineering, China Agricultural University. Her research interests include artificial intelligence, deep learning, software and software theory, and big data management and mining.