Center point prediction using Gaussian elliptic and size component regression using small solution space for object detection

Yuantian Xia[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, lilinlsl@cau.edu.cn]
[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,
China Agricultural University, Beijing, 100193, China
[e-mail: Phil.wang@cau.edu.cn]
[*]Corresponding author: Longhe Wang, Lin Li

Received November 29, 2022; revised March 27, 2023; revised May 25, 2023; accepted July 18, 2023; published August 31, 2023

Abstract

The anchor-free object detector CenterNet regards the object as a center point and predicts it based on the Gaussian circle region. For each object's center point, CenterNet directly regresses the width and height of the objects and finally gets the boundary range of the objects. However, the critical range of the object's center point can not be accurately limited by using the Gaussian circle region to constrain the prediction region, resulting in many low-quality centers' predicted values. In addition, because of the large difference between the width and height of different objects, directly regressing the width and height will make the model difficult to converge and lose the intrinsic relationship between them, thereby reducing the stability and consistency of accuracy. For these problems, we proposed a center point prediction method based on the Gaussian elliptic region and a size component regression method based on the small solution space. First, we constructed a Gaussian ellipse region that can accurately predict the object's center point. Second, we recode the width and height of the objects, which significantly reduces the regression solution space and improves the convergence speed of the model. Finally, we jointly decode the predicted components, enhancing the internal relationship between the size components and improving the accuracy consistency. Experiments show that when using CenterNet as the improved baseline and Hourglass-104 as the backbone, on the MS COCO dataset, our improved model achieved 44.7%, which is 2.6% higher than the baseline.

Keywords: Object detection, Anchor-free, Gaussian elliptic region, Center point prediction, Small solution space regression

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

http://doi.org/10.3837/tiis.2023.08.001 **ISSN : 1976-7277**

1. Introduction

Recently, with the emergence and application of high-performance computers and large-scale public datasets, many models[1-10] with powerful adaptive feature extraction capabilities and high recognition accuracy have been widely used in object detection, promoting the detector's performance. Detection models can be divided into anchor-base and anchor-free, depending on whether or not to use the anchor.

The anchor-base detector counts the sizes of different objects in the dataset through a clustering algorithm to generate a series of prior boxes. These prior boxes will be used as hyperparameters to assist the detector in completing the detection task. The anchor box mechanism effectively improved the detector performance for a specific dataset, but there are also several problems. First, 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. Second, 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[11] on the COCO[12] dataset by 4%. Finally, most of the generated anchors are marked as negative samples, artificially leading to an imbalance between samples. In addition, generating too many anchor boxes will cause a lot of memory and time consumption, aggravate the extra overhead of computing resources, and affect detection efficiency.

In order to solve the negative impact of the anchor mechanism, CenterNet[13] regards the object as a geometric center point and predicts the center point through a heatmap while regressing the object's width and height. As shown in **Fig. 1**, CenterNet predicts each object's center point position and category through the center point Heatmap and corrects the center point coordinates through the Offset feature maps. At the same time, the network regresses the width and height for each center point through the Size Component feature map with two channels, and the two channels respectively predict the width and height corresponding to each center point. Since the whole process does not require defining and calculating any prior boxes, it eliminated the negative impact of the anchor mechanism, making the prediction logic more intuitive and efficient.

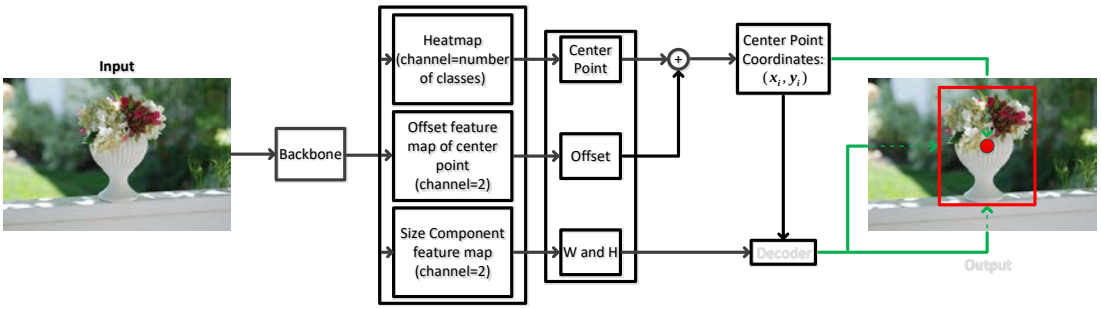


Fig. 1. The network structure of CenterNet.

However, we can see from **Fig. 2** that to satisfy the IOU threshold, CenterNet uses the distance between the upper left corner points of two anchors to approximate the vertical and horizontal distances between them. That is $r \approx d$ is used to construct the Gaussian circle prediction area. Since the object's upper left corner and center point are consistent, they have the same prediction area range. It can be seen that this approximate value method can not get the accurate range of the object center prediction region. In addition, because the same radius

r is used to limit the object's horizontal and vertical directions when the width and height of the objects differ significantly, the object's center point prediction will produce a significant error, thus reducing the prediction accuracy.

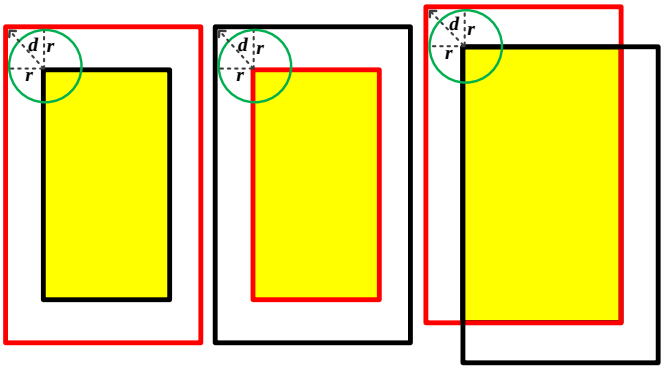


Fig. 2. The Gaussian circle prediction region of CenterNet. The red box is the predicted bounding box, the black box is the ground-true, and the yellow part is the overlapping area. The green circle area is the predicted area of the Gaussian circle, d is the upper left corner distance between the ground-true and the predicted box, and r is the Gaussian circle radius of the predicted area.

In addition, since no prior distribution knowledge is introduced in advance, the anchor-free detector has a larger and more flexible solution space. However, the overly flexible and huge solution space dramatically increases the model's training difficulty and generates too many false positives. This problem is particularly prominent for CenterNet. As shown in **Table 1**, since the size range of objects varies greatly, direct regression will lead to too large solution space, slowing down the convergence rate and increasing the instability. Moreover, since width and height are independently predicted in two channels, the lack of intrinsic correlation between width and height at the same position will lead to inconsistent prediction accuracy, thus reducing the model's prediction performance.

Table 1. The width and height range of objects in the MS COCO dataset

Dataset	Width range	Height range
MS COCO	(0,640]	(0,640]

For the problems mentioned above, we proposed an object's center point prediction method based on the Gaussian elliptic region and an object's size component regression method based on a small solution space. The central area of the Gaussian ellipse can accurately limit the prediction area of the object center point and improve the prediction accuracy of different object center points. The size component regression in the small solution space method effectively compresses solution space, increases convergence speed, and improves the consistency of size component accuracy.

The following are the main contributions of this paper::

1. First, we used two independent variables to control the horizontal and vertical distances between the object's ground-truth and predicted bounding box and limit the ratio of these two variables. And then, we calculate the variable critical value that meets the IOU threshold requirements according to the possible positional relationships between the object's ground-truth predicted bounding box. The critical value will be used as the major and minor semi-axes

of the center point prediction region of the Gaussian ellipse to precisely limit the prediction range of the object's center point.

2. Second, we re-encode the object's width and height into a small solution space and directly regress the encoded size components. The encoded size components had a smaller solution space, which can effectively improve the convergence speed and stability.

3. Finally, the predicted size components are jointly decoded into two length components in the horizontal and vertical directions. Since the two length components depend on the joint decoding of the size components, they have a stronger intrinsic correlation and consistency in accuracy.

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

2.1 Two-state anchor-based detectors

The Faster-RCNN proposed by Ren et al.[14] used the anchor mechanism for the first time to classify and localize objects. It adopts the RPN to replace the selective search[15] used in R-CNN[16] and Fast-RCNN[17] to generate candidate regions, which realizes end-to-end training and prediction. He et al.[18] proposed Mask R-CNN based on Faster-RCNN, which integrates the dual functions of object detection and instance segmentation and improves the ability to solve more complex visual tasks. Cai et al.[19] 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. The above are all two-stage algorithms using the anchor mechanism. Although the detection performance has been dramatically improved, there are still significant deficiencies in the detection speed. Therefore, a series of one-stage object detectors have been proposed based on the anchor mechanism.

2.2 One-state anchor-based detectors

Liu et al. [20] proposed an object detection model called SSD based on multi-scale feature prediction. The SSD introduced the anchor mechanism and used FPN[21] to predict objects on different scale feature maps. DSSD[22] used the deconvolution method and added context information so that the low-level feature maps have better feature expression ability. Leng et al.[23] improved SSD and proposed the SSADet. Compared with SSD, SSADet adopts anchor prediction and feature fusion modules to improve the detection accuracy effectively. Subsequently, YOLO series models have been continuously developed. YOLOv2[24] uses the DarkNet-19 network and the anchor mechanism to improve detection accuracy further. YOLOv3[25] improved YOLOv2 by using Darknet-53 as the backbone for feature extraction and added up-sampling feature fusion operation based on FPN so that the model could extract object features more accurately. Hurtik et al.[26] improved YOLOv3 and proposed poly-YOLO. It eliminates the problem of many rewritten labels and inefficient distribution of anchors in Yolov3. Bochkovskiy et al.[27] summarize and improve the training techniques that have achieved excellent detection performance in recent years and proposed YOLOv4. It

adopts CSPNet[5] as the feature extraction network and PAN[28] as the feature fusion network, which improves the model feature extraction and fusion capabilities. Tan et al.[29] 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. Lin et al.[11] proposed the Focal Loss function, which solved the low accuracy of one-stage detectors, thus significantly improving the detection accuracy of one-stage object detectors. Ju et al.[30] proposed an adaptive feature fusion with attention mechanism (AFFAM) method, which further improved the detection performance of YOLOv3 and DSSD.

2.3 Anchor-free detectors

All the above algorithms have used the anchor mechanism. In order to solve the negative impact of the anchor on detection performance, in recent years, some anchor-free detectors have been widely concerned. As the foundation work of the anchor-free detectors, Huang et al.[31] proposed DenseBox based on FCN[32]. DenseBox combines multi-task learning with key-point detection, which enables it to detect the object with severe occlusion accurately and efficiently without generating candidate bounding boxes. Yu et al.[33] proposed UnitBosx based on improving DenseBox. UnitBox adopts the newly designed IOU Loss to process the four coordinate regression values of the object as a whole. As the first version of the YOLO series models, YOLO v1[34] adopts the anchor-free approach to simultaneously predict the object's location and category by using only one neural network, significantly improving the detection speed. Tian et al.[35] proposed a pixel-level prediction object detection algorithm called FCOS. It reconstructs detection objects on a per-pixel basis and uses FPN to improve the recall rate and resolve ambiguity caused by overlapping boundaries. In addition, it proposed the center-ness branch, which reduces the false-positive boxes and dramatically improves the detection performance. Zhu et al.[36] proposed an anchor-free feature selection module FSAF, which solves the limitations of heuristic feature selection and overlap-based anchor sampling in the traditional anchor-based algorithm. Kong et al.[37] proposed an accurate and flexible anchor-free object detection model called FoveaBox. It generates class-agnostic bounding boxes for each location that may contain an object.

With the continuous improvement of the key point prediction algorithm, more and more detectors have adopted the key point prediction method to complete the object detection task. Law et al.[38] proposed CornerNet based on the corner point prediction. It regards the object as a pair of key points and predicts its top-left and bottom-right heatmaps through a single network and the embedding vector for each corner point. The embedding vectors are used to group corner points that belong to the same object. To further enhance the accuracy of keypoint-based prediction, Zhou et al.[39] 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. For this problem, Zhou et al.[13] 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 heatmaps, then adjust the center point through the offset feature maps, and finally, regress the object's width and height. There is no complex matching work since there is only one key point for prediction, making the model simpler and more efficient.

3. Proposed Method

The detail of our proposed method will introduce in this section. First, we introduced the generation and calculation process of the center point prediction based on the Gaussian elliptic region. Then, we introduce the coding and regression process of the size component based on the small solution space. Finally, we give the joint decoding process of the size components. The object's position can be obtained according to the predicted center point coordinates and the decoded size components.

3.1 Center point prediction based on the Gaussian elliptic region

As shown in **Fig. 3**, for the selected IOU threshold, the positional relationship between the predicted and the ground-truth bounding box is as follows:

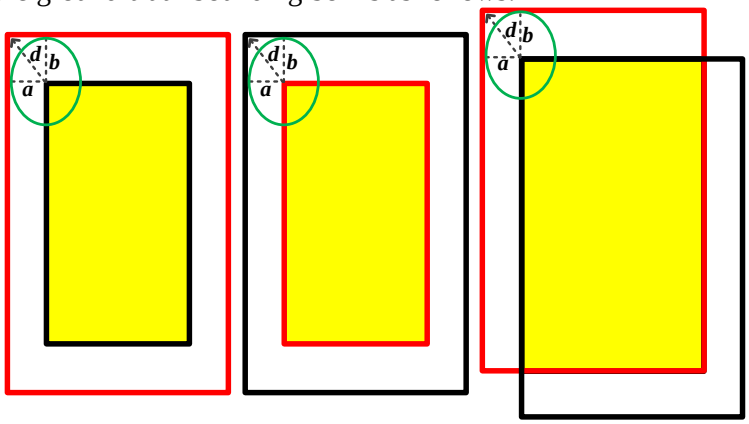


Fig. 3. The Gaussian ellipse prediction region. The red box is the predicted bounding box, the black box is the ground-true bounding box, and the yellow part is the overlapping area. The green ellipse area is the predicted area of the Gaussian ellipse, a is the horizontal distance between the ground-true and predicted bounding box, b is the vertical distance between the ground-true and predicted bounding box, d is the distance between the upper left corner of the predicted and ground-true bounding box.

In order to more accurately describe and limit the possible prediction area range of the object's center point and avoid inaccurate caused by the large gap between the object's size, we control the vertical and horizontal distance between the predicted and ground-true bounding box through a and b, respectively. For each case, the ratio of a and b is specified to be consistent with the aspect ratio of the object's ground-truth bounding box. The relationship between a and b can be described as follows:

$$\frac{a}{b}=\frac{w}{h} \tag{1}$$

where w and h are the width and height of the object's ground-truth bounding box, which are known.

3.1.1 The ground-truth box completely falls within the predicted box

As shown in (a) in **Fig. 3**, for the convenience of observation and calculation, assuming that the center point of the object's ground-true bounding box and the predicted bounding box coincide, the calculation formula of the intersection over union(IOU) is as follows:

$$IOU = \frac{h \times w}{(h+2b)(w+2a)} \tag{2}$$

According to (1) and (2), the following quadratic equation can be obtained:

$$a^2 + hwa + \frac{(IOU-1)hw^2}{4IOU} = 0 \tag{3}$$

where a is the unknown item, w, h and IOU are all known items.

According to the root formula and $a > 0$, the value of a can be obtained:

$$a = \frac{-hw + \sqrt{h(h - \frac{IOU-1}{IOU})}}{2} \tag{4}$$

The value of b can be obtained according to (1):

$$b = \frac{h\left(-hw + \sqrt{h(h - \frac{IOU-1}{IOU})}\right)}{2w} \tag{5}$$

According to the above calculation, in the first case, the distance d_1 between the upper left corner of the predicted bounding box and the ground-true bounding box is:

$$d_1 = \sqrt{a^2 + b^2} \tag{6}$$

3.1.2 The predicted box completely falls within the ground-truth box

As shown in (b) in **Fig. 3**, the calculation formula of the IOU is as follows:

$$IOU = \frac{(h-2b)(w-2a)}{h \times w} \tag{7}$$

According to (1) and (7), the following quadratic equation can be obtained:

$$a^2 - wa + \frac{(1-IOU)w^2}{4} = 0 \tag{8}$$

where a is the unknown item, w, h and IOU are all known items.

According to the root formula, the value of a can be obtained:

$$a = \frac{w(1 \pm \sqrt{IOU})}{2} \tag{9}$$

Since $w-2a$ represents the width of the prediction bounding box in (7), so $w-2a > 0$, and the value of a can be obtained as follows:

$$a = \frac{w(1-\sqrt{IOU})}{2} \tag{10}$$

The value of b can be obtained according to (1):

$$b = \frac{h(1-\sqrt{IOU})}{2} \tag{11}$$

According to the above calculation, in the second case, the distance d_2 between the upper left corner of the predicted bounding box and the ground-true bounding box is:

$$d_2 = \sqrt{a^2 + b^2} \tag{12}$$

3.1.3 The predicted box partially overlaps with the ground-truth box

As shown in (c) in **Fig. 3**, the calculation formula of the IOU is as follows:

$$IOU = \frac{(h-b)(w-a)}{2hw-(h-b)(w-a)} \tag{13}$$

According to (1) and (13), the following quadratic equation can be obtained:

$$a^2 - 2wa + \frac{1-IOU}{1+IOU}w^2 = 0 \tag{14}$$

where a is the unknown item, w, h and IOU are all known items.

According to the root formula, the value of a can be obtained:

$$a = w(1 \pm \sqrt{\frac{2IOU}{1+IOU}}) \tag{15}$$

Since $w-a>0$ in (13), so the value of a can be obtained:

$$a = w(1 - \sqrt{\frac{2IOU}{1+IOU}}) \tag{16}$$

The value of b can be obtained according to (1):

$$b = h(1 - \sqrt{\frac{2IOU}{1+IOU}}) \tag{17}$$

According to the above calculation, in the second case, the distance d_3 between the upper left corner of the predicted bounding box and the ground-true bounding box is:

$$d_3 = \sqrt{a^2+b^2} \tag{18}$$

3.1.4 The Gaussian ellipse prediction region

For each object, we choose the a and b corresponding to the minimum value of $d = \min(d_1, d_2, d_3)$ in the above three cases as the minor semi-axis and major semi-axis of the generated ellipse. Then the Gaussian function of the ellipse region can be expressed as:

$$f(x,y) = \mathrm{e}^{-\left(\frac{(x-\tilde{p}_x)^2}{2\sigma_a^2}+\frac{(y-\tilde{p}_y)^2}{2\sigma_b^2}\right)} \tag{19}$$

where σ_a and σ_b are the variances corresponding to a and b, respectively. In the Gaussian distribution, the area of the $(\mu-3\sigma, \mu+3\sigma)$ interval accounts for 99.7% of the total area under the Gaussian curve, so the values outside this interval are close to 0 and can be ignored. Therefore, choosing 3σ as the Gaussian radius:

$$\sigma_a = \frac{1}{3}\sigma \tag{20}$$

$$\sigma_b = \frac{1}{3}b \tag{21}$$

Since the coordinates of the upper left corner of the same object are consistent with the center point coordinates, the Gaussian ellipse prediction range of the center point can be obtained by mapping the Gaussian ellipse range of the corner point to the center point of the object. In the Gaussian ellipse predicted range, the center point's value is 1. The further away from the center, the smaller the value. The value near the ellipse boundary gradually

approaches 0, and the value outside the boundary is 0.

3.2 The process of size components encoding

In order to compress the vast difference in the width and height of the objects, we encoded them as the diagonal half-length of the object's boundary box and the cosine of the center Angle. The coding process is shown in **Fig. 4**.

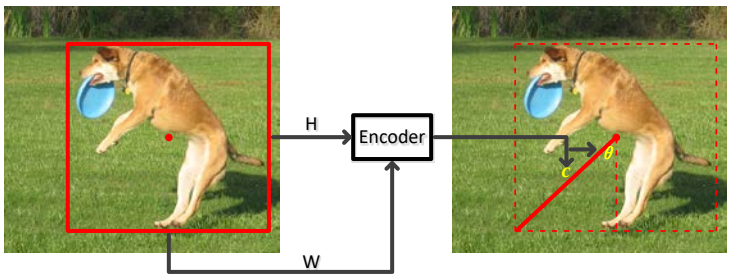


Fig. 4. Encoding process. Where H is the height of the bounding box and W is the width of the bounding box. c is the half-length of the diagonal after encoding, and θ is the central angle after encoding.

As shown in **Fig. 4**, a right triangle with the center angle θ as the vertex is formed between the bounding box's diagonal and the vertical centerline passing through the object's center point. We encoded the width and height as c and $\cos\theta$. The model will not predict the width and height but directly predict c and $\cos\theta$. 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 width and height. In addition, no matter how the width and height change, the variation range of $\cos\theta$ is constantly kept between (0, 1). After coding the object in the COCO dataset, the size components variation range is shown in **Table 2**.

Table 2. The $\cos\theta$ and c range of objects in the MS COCO datasets.

Dataset	$\cos\theta$ **range**	c **range**
MS COCO	(0,1)	[2,450]

Compared with **Table 1**, the re-encoded components significantly compress the solution space. The calculation process is expressed as follows:

$$c = \frac{\sqrt{W^2 + H^2}}{2} \tag{22}$$

$$\cos\theta = \frac{H}{2c} \tag{23}$$

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 $s_k = (c^k, \cos\theta^k)$. The loss function is expressed as follows:

$$Lc\theta = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{p_k} - s_k\right| \tag{24}$$

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. 5**.

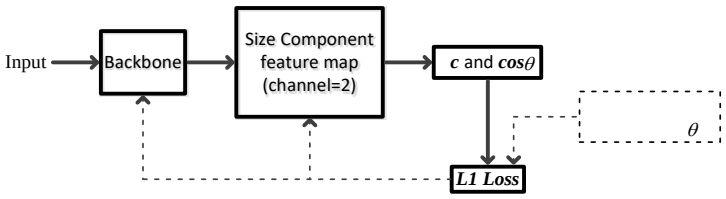


Fig. 5. The loss function calculation process.

As shown in **Fig. 5**, after the input image passes through the backbone, a feature map with two channels is output, and the two channels are the predicted values of c and $\cos\theta$, respectively. By calculating the loss value, the network will gradually obtain better and more accurate values of c and $\cos\theta$.

3.3 The process of size components joint decoding

In CenterNet, the object's width and height are independently predicted in the feature map with two channels, which leads to inconsistency in prediction accuracy, such as the prediction accuracy of one component is high and the other is low. To solve this problem, we jointly decoded the c and $\cos\theta$ obtained in Section 3.2 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 $\cos\theta$ simultaneously, there is a strong internal dependency between the decoded two components, enhancing the consistency of prediction accuracy. The decoding process is shown in **Fig. 6**.

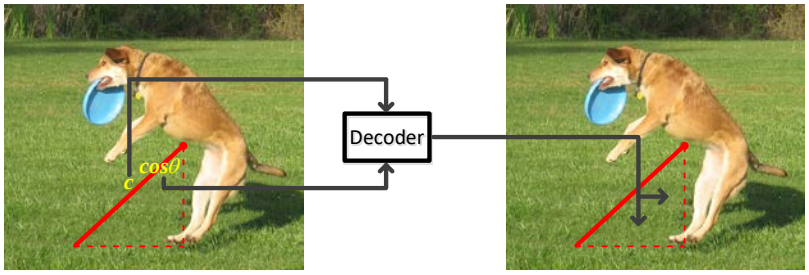


Fig. 6. Decoding process. Where a is the decoded vertical length component and b is the horizontal length component.

As shown in **Fig. 6**, c and $\cos\theta$ can be decoded into the vertical length component a and the horizontal length component b of the center point using the trigonometric function. The calculation process is expressed as follows:

$$a = \cos\theta \times c \tag{25}$$

$$b = \sqrt{c^2 - a^2} \tag{26}$$

Because the definition domain of θ is between $(0, \frac{\pi}{2})$, so the value domain of $\cos\theta$ is between (0, 1). Moreover, c is a positive number, so zero or negative prediction errors will not occur. In addition, applying the smaller value range of $\cos\theta$ to adjust c can make a and b have better accuracy stability and uniformity.

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

$$\hat{P}_c = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{n} \tag{27}$$

$$\hat{P}_{c:left-bottom} = (\hat{x}_i - b_i, \hat{y}_i + a_i) \tag{28}$$

$$\hat{P}_{c:right-top} = (\hat{x}_i + b_i, \hat{y}_i - a_i) \tag{29}$$

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. 7**.

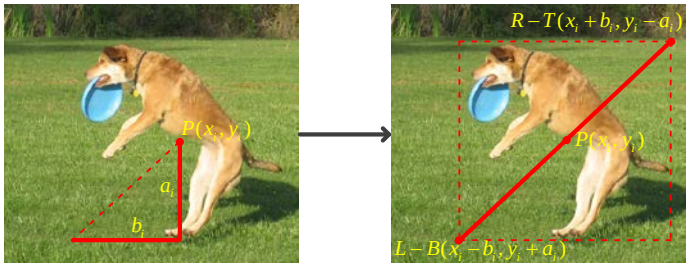


Fig. 7. The bounding box generation process.

As shown in **Fig. 7**, the decoded length components a and b can be used to predict the object 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 of the model.

4. Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed algorithm in this paper, we apply the proposed algorithm to the baseline model CenterNet and conduct extensive experiments on the Microsoft COCO[12] dataset. Microsoft COCO is a large image dataset constructed by Microsoft corporation for object detection, semantic segmentation, key point detection, and other visual tasks. It has been widely used in object detection tasks and recognized by the authority. We selected the most widely used COCO 2017 version for model training and testing. This version consists of a train set with 118,287 images, a validation set with 5,000

images, and a test set with 40,670 images. The total number of images was 163,957 in 80 categories. The COCO dataset images contain natural and common object images in daily life. Compared with other datasets, the background is more complex, the number of objects is larger, and there are many smaller objects, so the visual task on the COCO dataset is more challenging.

4.2 Evaluation metric

We use train 2017 and Val 2017 in the COCO dataset to train and verify the model and evaluate our proposed algorithm in test 2017. Unlike other datasets, 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. The COCO dataset provides AP_{50} and AP_{75} to evaluate the model's performance with IoU thresholds of 50 and 75, respectively. In addition, the COCO dataset also provides AP_S, AP_M, and AP_L to evaluate small objects with pixel areas less than 32×32, medium objects with pixel areas between 32×32 and 96×96, and large objects with pixel areas greater than 96×96, respectively.

4.3 Implement details

Compared with the baseline model CenterNet, the improved model proposed in this paper mainly has the following two changes:

1、The method of the object's center point mapping to the Heatmap.

In Centernet, use the following formula to map all the object's center points in the group truth to the Heatmap:

$$f(x,y)=\mathrm{e}^{-\left(\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2}\right)} \tag{30}$$

By using (30), CenterNet can map the center point coordinates of all objects into a heatmap with a limited range of Gaussian circles when loading data.

In our improved model, (30) needs to be replaced by (19) proposed in Section 3.1.4.

Through (19), the improved model can map the coordinates of the center points of all objects into a heatmap with a limited range of a Gaussian ellipse when loading data.

2、The method of the size components mapping.

In CenterNet, the height and width of all objects in the ground truth are directly loaded as size components during data loading, and the predicted values of height and width are directly output during prediction.

In our improved model, when loading data, the height and width of all objects in the ground truth should be coded according to the method proposed in Section 3.2. The predicted output value should be decoded during the prediction according to the method proposed in Section 3.3.

With the above changes, we can start training and reasoning just like CenterNet. The training and reasoning processes were consistent with the baseline model CenterNet.

4.4 Parameter setting

We selected Hourglass-104 as the backbone and set the input size to 512×512, the same as CenterNet. The model was trained for 100 epochs using the stochastic gradient descent SGD

algorithm, and the batch size was set to 32. The learning rate is dynamically adjusted using warm-up and cosine annealing functions. The number of warm-up steps is set to 5, the initial learning rate is set to 0.00025, and the momentum and weight decay are set to 0.9 and 0.0005, respectively. Image enhancement methods consistent with CenterNet include random flip, random scaling (between 0.6 to 1.3), cropping, and color jittering. The hardware and software environment of the experiment is shown in **Table 3**.

Table 3. The software and hardware environment.

Equipment	Type
CPU:	Intel core i9-9900k
GPU:	NVIDIA GeForce RTX 3090
RAM	64.0 GB
OS:	WIN10 64-bit
Develop software:	Python3.8+Pythorch10.0+cuda11.3+Pycharm

4.5 Loss Comparision

In order to more intuitively demonstrate the improved convergence speed and detection performance of the baseline model by the proposed algorithm, we recorded the central point prediction loss and total loss of each epoch of the baseline model and the improved model during training, respectively. The results are shown in **Fig. 8** to **Fig. 9**.

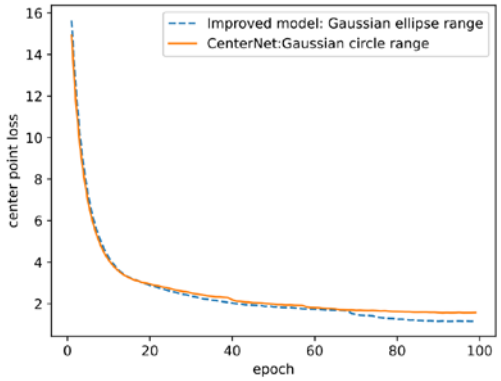


Fig. 8. The predicted loss curve of the central point.

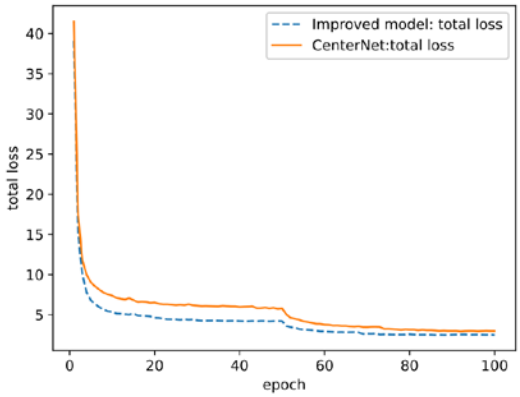


Fig. 9. The total loss curves.

As can be seen from the above comparison figures, compared with the baseline, the improved model has a faster convergence speed, lower loss value, and smoother loss curve, reflecting the improvement of model stability and accuracy to a certain extent.

4.6 Results on MS COCO

To evaluate the effectiveness of the proposed algorithm, we provide a comparison of results between different detectors. **Table 4** shows the comparison results between our proposed algorithm and other state-of-the-art detectors on the MS COCO test 2017, including two-stage, single-stage, anchor-base, and anchor-free object detectors. In order to simplify the expression, we call the improved model proposed in this paper as Enhanced-CenterNet, abbreviated as E-CenterNet. The results of E-CenterNet are obtained by training with the parameters set in Section 4.**4**, and the results of other detectors are from their corresponding papers.

Table 4. Comparison results with other state-of-the-art methods on MS COCO.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
anchor-based detectors							
Mask R-CNN[18]	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
Cascade R-CNN[19]		42.8	62.1	46.3	23.7	45.5	55.2
SSD513[20]		31.2	50.4	33.3	10.2	34.5	49.8
DSSD513[22]		33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet[11]		39.1	59.1	42.3	21.8	42.7	50.2
YOLO v2[24]	Darknet-19	21.6	44.0	19.2	5.0	22.4	35.5
YOLO v3[25]	Darknet-53	33.0	57.9	3434	18.3	35.4	41.9
YOLO v4[27]	CSPDarknet-53	43.5	65.7	47.3	26.7	46.7	53.3
anchor-free detectors							
FCOS[35]	ResNeXt-101	42.1	62.1	45.2	25.6	44.9	52.0
GA-RPN[40]	ResNet-50	39.8	59.2	43.5	21.8	42.6	50.7
FoveaBox[37]	ResNeXt-101	42.1	61.9	45.2	24.9	46.8	55.6
ATSS[41]		43.6	62.1	47.4	26.1	47.0	53.6
Grid R-CNN[42]		43.2	63.0	46.6	25.1	46.5	55.2
FSAF[36]	ResNeXt-64x4d-101	42.9	63.8	46.3	26.6	46.2	52.7
RepDet[44]	ResNet-101-DCN	42.8	65.0	46.3	24.9	46.2	54.7
CornerNet-Lite[43]	Hourglass-54	43.2	-	-	24.4	44.6	57.3
CornerNet[38]	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
ExtremNet[39]		40.1	55.3	43.2	20.3	43.2	53.1
CenterNet[13]		42.1	61.1	45.9	24.1	45.5	52.8
ours							
E-CenterNet	Hourglass-104	**44.7**	**63.4**	**47.8**	**27.1**	**46.1**	**55.5**

As shown in **Table 4**, the AP of E-CenterNet reaches 44.7%, which is the best among all compared models. It is 1.9% higher than the Cascade R-CNN with the highest AP in the two-stage detectors, 1.2% higher than the YOLO v4 with the highest AP in anchor-base detectors, 1.1% higher than ATSS with the highest AP in anchor-free detectors , and 2.6% higher than the baseline CenterNet. The E-CenterNet also achieved excellent accuracy in comparing different IOU threshold indexes, especially in the more stringent AP_{75} index, which reached the best accuracy of all the comparison models. It is 0.4% higher than ATSS, which has the best AP_{75} accuracy in the comparison model, and 1.9% higher than CenterNet, the baseline

model in this paper. It shows that the improved algorithm proposed in this paper can effectively improve the detector's performance under stringent performance requirements. In evaluating detection performance metrics for different size objects, E-CenterNet also achieved the best accuracy. Especially on the most challenging small object detection performance index AP_S, E-CenterNet is 3.0% higher than the baseline CenterNet and 0.4% higher than the YOLO v4 with the highest AP_S in the comparison model, significantly improving the detection accuracy for small objects. It can be seen from the above comparative analysis that the improved algorithm proposed in this paper can significantly improve the detection performance of the model and the detection ability under severe conditions.

4.7 Ablation studies

To further evaluate the effectiveness of the different components of our proposed algorithm and the impact on detection performance, we formed CenterNet+GER based on Gaussian ellipse region prediction and CenterNet+SSS based on small solution space regression. Where GER is the abbreviation of Gaussian elliptic region, SSS is the abbreviation of small analytic space. We tested the models based on the two improved methods on MS COCO dataset, and the results are shown in **Table 5**.

Table 5. The effect of different improved components on detection accuracy of baseline.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
CenterNet	Hourglass-104	42.1	61.1	45.9	24.1	45.5	52.8
CenterNet+GER	Hourglass-104	43.7	62.6	47.0	26.5	46.3	54.8
CenterNet+SSS	Hourglass-104	43.1	61.7	46.4	26.1	45.6	54.0

As can be seen from **Table 5**, the AP accuracy of CenterNet+GRE using the Gaussian ellipse region prediction method reaches 43.7%, and the accuracy is improved by 1.6% compared with CenterNet using the Gaussian circle region prediction method. The AP of CenterNet+SSS with small solution space regression is 43.1%, which is 1.0% higher than that of CenterNet with direct regression width and height. In the AP_{50} index, the acureacy of CenterNet+GER and CenterNet+SSS are 1.5% and 0.6% higher than the baseline. Under the more stringent IOU threshold of AP_{75}, the AP_{75} of CenterNet+GER and CenterNet+SSS are 1.1% and 0.5% higher than the baseline, respectively. The accuracy of the improved model based on different components is also greatly improved on the indexes used to evaluate objects of different sizes. Especially on the most difficult small object indicator AP_S, the accuracy of CenterNet+GER and CenterNet+SSS are 2.4% and 2.0% higher than the baseline, respectively. It can be seen from the above analysis that both the center point prediction method based on the Gaussian ellipse region and the size component regression method based on the small solution space proposed in this paper can significantly improve the detection performance. Because the center point prediction method based on the Gaussian elliptic region can more reasonably and accurately define the potential region of the object's center point, it can predict the target center point more accurately. In addition, the size component regression method based on the small solution space can effectively compress the regression space of the object size, so the prediction of the object size is more stable and accurate, further improving the model's detection performance.

4.8 Qualitative evaluation

We provide a qualitative comparison between the improved model and CenterNet to illustrate further our proposed method's superiority in complex scene object detection and bounding box prediction accuracy.

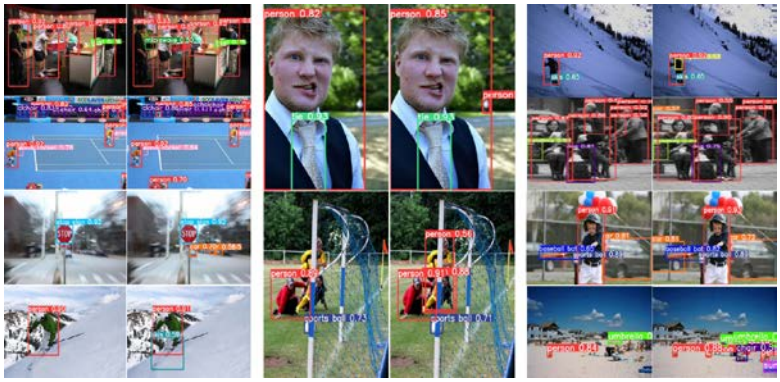


Fig. 10. Comparison of prediction results in complex scenarios. The objects with detection scores higher than 0.5 are shown. In each pair, the left side is the detection results of CenterNet. The right side is our improved model.

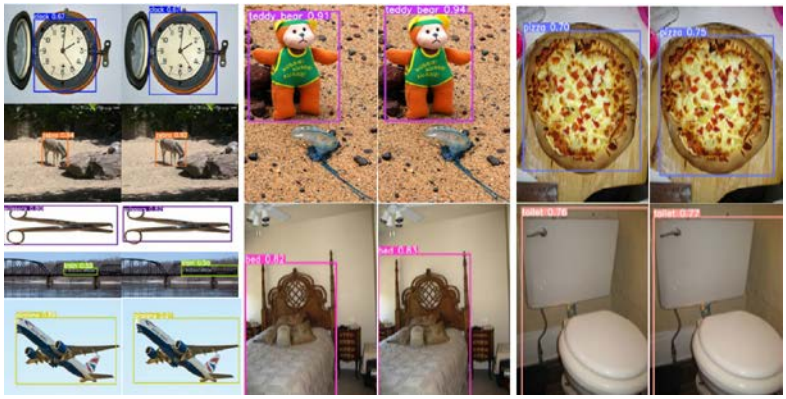


Fig. 11. Quality comparison of bounding box predictions. The objects with detection scores higher than 0.5 are shown. In each pair, the left side is the detection results of CenterNet. The right side is our improved model.

Fig. 10 shows the detection performance comparison between the improved and baseline models in complex scenarios. In complex scenes, there are usually many complex situations such as mutual occlusion, uneven illumination, and dense connections between objects, which pose a higher challenge to the performance of the detectors. As shown in **Fig. 10**, the improved model proposed in this paper successfully detects most of the objects that the CenterNet does not detect, which are complicated by small size and poor visual features. It fully shows that the algorithm proposed in this paper can significantly improve the object detection ability of the model in complex scenes. **Fig. 11** compared the improved and baseline models for the prediction quality of the object bounding box. As shown in **Fig. 11**, 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 width prediction is accurate, but the height is inaccurate, or the high prediction is accurate, but the width is inaccurate. Compared

with CenterNet, we compressed the object size component's regression solution space by recoding the object's width and height and generating the object size by joint decoding. Therefore, when the detection accuracy is similar, our improved model has higher accuracy and consistency in predicting the bounding box position. The above qualitative comparison results fully show that the improved algorithm proposed in this paper can significantly improve the detection accuracy of the model in complex scenes and the prediction quality of the object bounding box, thus significantly improving the model's detection performance.

5. Conclusion

In this paper, we proposed a center point prediction algorithm based on the Gaussian elliptic region to more accurately predict the object's center point with a large difference in size and improve the accuracy of object position prediction. And then, we proposed the size component regression algorithm based on a small solution space to speed up the convergence and improve the stability and consistency. The experiment results fully show that our method can significantly enhance detectors' accuracy and detection ability. However, the method based on central point prediction also has the following shortcomings, which are also the focus of our follow-up research and solution. In the training process, if some objects of the same class are too close, the downsampling operation may cause the center points of these objects to overlap and eventually cause the model to train these different objects as the same object. Similarly, in the predictive reasoning process, if the center points of some objects of the same class also overlap after downsampling, then the model can only predict one object.

References

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

[2] C. Szegedy, W. Liu, Y.Jia, P.Sermanet, A.Rabinovich, “Going deeper with convolutions,” in *Proc. of CVPR 2015*, Long Beach, USA, pp.1-9, June 2015. Article(CrossRef Link).

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

[4] 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).

[5] 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).

[6] Y. Gao, O. Beijbom, N. Zhang and T. Darrell, "Compact Bilinear Pooling," in *Proc. of CVPR 2016*, Las Vegas, USA, pp.317-326, June 2016. Article(CrossRef Link).

[7] D. Erhan, C. Szegedy, A. Toshev and D. Anguelov, "Scalable Object Detection Using Deep Neural Networks," in *Proc. of CVPR 2014*, Columbus, USA, pp.580-587, June 2014. Article(CrossRef Link).

[8] P. O. Pinheiro, R. Collobert, P. Dollar, “Learning to segment object candidates,” *arXiv preprint arXiv:1506.06204*, 2015. Article(CrossRef Link).

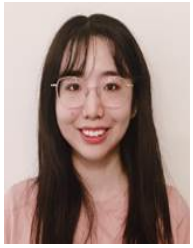
[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classication with deep convolutional neural networks,” in *Proc. of NIPS 2012*, Lake Tahoe Nevada, USA, pp.1097-1105, December 2012. Article(CrossRef Link).

[10] H. Zhang, X. Cao, J. K. L. Ho and T. W. S. Chow, "Object-Level Video Advertising: An Optimization Framework," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp.520-531, April 2017. Article(CrossRef Link).

[11] 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, 1 February 2020. Article(CrossRef Link).
[12] 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 ECCV 2014*, Zurich, Switzerland, pp.740-755, September 2014. Article(CrossRef Link).
[13] X. Zhou, D. Wang, et al, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019. 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] 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).
[16] 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).
[17] R. Girshick, "Fast R-CNN," in *Proc. of ICCV 2015*, Santiago, Chile, pp.1440-1448, December 2015. Article(CrossRef Link).
[18] 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).
[19] 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).
[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] C. Y. Fu, W. Liu, A. Ranga, et al, "DSSD: deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017. Article(CrossRef Link).
[23] J. Leng, Y. Liu, "Single-shot augmentation detector for object detection," *Neural Comput&Applic*, vol. 33, pp.3583-3596, July 2021. 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] P. Hurtik, V. Molek, J. Hula, et al, "Poly-YOLO: higher speed, more precise detection and instance segmentation for YOLOv3," *Neural Comput&Applic*, vol. 34, pp.8275-8290, February 2022. Article(CrossRef Link).
[27] 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).
[28] 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).
[29] 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).
[30] M. Ju, J. Luo, Z. Wang, et al, "Adaptive feature fusion with attention mechanism for multi-scale target detection," *Neural Comput&Applic*, vol. 33, pp.2769-2781, July 2021. Article(CrossRef Link).
[31] L. Huang, Y. Yang, Y. Deng, "DenseBox: Unifying Landmark Localization with End to End Object Detection," *arXiv preprint arXiv:1509.04874*, 2015. Article(CrossRef Link).
[32] J. Long, E. Shel-hamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pat-tern Analysis & Machine Intelligence*, vol.39, pp.640-651, 2017. Article(CrossRef Link).

[33] J. Yu, Y. Jiang et al, "UnitBox: An Advanced Object Detection Network," in *Proc. of the 24th ACM international conference on Multimedia*, Netherlands, Amsterdam, pp.516-520, October 2016. Article(CrossRef Link).
[34] 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).
[35] Z. Tian, C. Shen, H. Chen and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *Proc. of ICCV 2019*, Long Beach, USA, pp.9626-9635, June 2019. Article(CrossRef Link).
[36] C. Zhu, Y. He, M. Savvides, "Feature Selective Anchor-Free Module for Single-Shot Object Detection," in *Proc. of CVPR 2019*, Long Beach, USA, pp.15-20, June 2019. Article(CrossRef Link).
[37] T. Kong, F. Sun, H. Liu et al, "FoveaBox: Beyound anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp.7389-7398, june 2020. Article(CrossRef Link).
[38] 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).
[39] 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).
[40] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin, "Region proposal by guided anchoring," in *Proc. of CVPR 2019*, Long Beach, USA, pp.2960-2969, June 2019. Article(CrossRef Link).
[41] S. Zhang, C. Chi, Y. Yao, Z. Lei and S. Z. Li, "Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection," in *Proc. of CVPR 2020*, Seattle, USA, pp.9756-9765, June 2020. Article(CrossRef Link).
[42] X. Lu, B. Li, Y. Yue, Q. Li and J. Yan, "Grid R-CNN," in *Proc. of CVPR 2019*, Long Beach, USA, pp.7355-7364, June 2019. Article(CrossRef Link).
[43] Hei Law, Yun Teng, Olga Russakovsky, and Jia Deng, "Cornernet-lite: Efficient keypoint based object detection," *arXiv preprint arXiv:1904.08900*, 2019. Article(CrossRef Link).
[44] Z. Yang, S. Liu, H. Hu, L. Wang and S. Lin, "RepPoints: Point Set Representation for Object Detection," in *Proc. of ICCV 2019*, Seoul, Korea, pp.9656-9665, 2019. 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.



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.