IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2023, accepted 24 September 2023, date of publication 4 October 2023, date of current version 11 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321966

RESEARCH ARTICLE

An Anchor-Free Lightweight Object Detection Network

WEINA WANG AND YUNYAN GOU
College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China

Corresponding author: Weina Wang (wangweina@jlict.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 62266046; in part by the Natural Science Foundation of Jilin Province, China, under Grant YDZJ202201ZYTS603; and in part by the Natural Science Foundation of Jilin Provincial Department of Education, China, under Grant JJKH20230281KJ.

ABSTRACT Existing anchor-free object detection methods have achieved some amazing results, but these methods are relatively complex and the inference speed is also slow. In this paper, an anchor-free lightweight object detection network is proposed. The proposed method effectively overcomes the limitation of detection model by anchor-free mechanism, and the lightweight backbone network reduces the computational cost. In addition, the proposed small object enhancement module can enhance the focus on small objects, which improves the detection capability of small objects. Besides, a label assignment strategy is proposed to determine the prominent feature, and a center correction mechanism is introduced to make the predicted bounding box closer to the ground truth to further improve the detection accuracy. Extensive experiments are conducted on MS COCO and Pascal VOC datasets, and the results demonstrate that the proposed method achieves better results than the existing detection methods on detection accuracy by increasing 0.5% on the MS COCO dataset, and has a detection accuracy increase of 0.4% on the Pascal VOC dataset, which proves the superiority of the proposed method.

INDEX TERMS Object detection, anchor-free, lightweight, small object, label assignment strategy.

I. INTRODUCTION

Object detection task is to find out all the objects of interest in the image, determine their class and location, and is one of the core problems in the field of computer vision. In the field of computer vision, there are many important and challenging branches of object detection, such as pedestrian detection [1], [2], medical detection [3], small object detection [4], face detection [5], [6], [7], salient object detectio [8], [9], [10], text detection [11], [12] traffic detection [13], [14]. Due to boom in deep learning, the object detection algorithm has been developed rapidly. Typical methods are Faster R-CNN [15] and fully convolutional networks (FCN) [16]. They have the advantages of conceptual intuitiveness, good flexibility, robustness, and rapidity of training and reasoning. In addition, feature pyramid networks (FPN) [17] use multi-stage image pyramids to obtain feature pyramids, which improved the average accuracy of the detector. Focal loss [18] focuses on sparse hard samples and prevented the large number of negative samples, which can improve the detection performance. However, these anchor-based detectors require setting a large number of anchor points in the detection process, leading to computational cost increase and detection speed degradation.

To overcome the shortcoming of anchor-based detectors, anchor-free detection is proposed. The anchor-free detector utilizes pixel-by-pixel prediction to solve the detection problem. The detector need not require predetermined anchor points, avoiding complex anchor point calculations, and greatly reducing computation time and computational power. Therefore, anchor-free detection has gained great momentum in the object detection field. Cornernet [19] converts the network's focus on the object bounding box into a pair of key points without designing anchor points, which improves the detection accuracy. However, the network only focuses on the edges and corner points, resulting in insufficient internal information acquisition. To overcome the problem, Centernet [20] directly focuses on the centroid

The associate editor coordinating the review of this manuscript and approving it for publication was Badri Narayan Subudhi.

of the object. Although the method effectively improves the accuracy and recall, the detection with occlusion is not effectively solved because of mutual occlusion for the object. Fully convolutional one-stage (FCOS) [21] is a pixel-by-pixel object detection model, which utilizes centrality center-ness to improve the accuracy. Compared with anchor-based methods, FCOS has achieved promising performance in detection accuracy and speed. However, it still has higher computational costs and shortcomings in the detection capability of small objects.

Recently, lightweight networks have shown superior performance in many practical applications. SqueezeNet [22] is a smaller and more intelligent network. The model efficiently reduces the number of parameters, while maintaining good performance. ShuffleNet [23] is specially designed for mobile devices, which have fewer parameters and efficient computation. The traditional approaches to lightening the network are to use simplification, pruning, and compression. MobileNet [24], [25], [26] series discard these approaches and use the efficient network architecture. MobileNets not only have the advantage of less computation, but also can ensure the accuracy of detection. Inspired by the structure of MobileNets, the proposed method introduces the efficient architecture as the backbone network to lighten the network.

Considering all the above-mentioned observations, we propose a lightweight anchor-free detection model, while improving the detection capability of small objects. Firstly, the backbone network of FCOS is optimized to lighten the network. Then, the small object detection enhancement module is constructed to enhance the detection capability of small objects. Finally, a label assignment strategy is proposed, and then the strategy is combined center correction mechanism to improve the limited detection performance due to the lack of accurate label assignment. The main contributions of this paper are as follows:

1) A small object enhancement module is constructed based on the single stage headless face detector. The module can improve the focus on small objects, and enhance the detection capability.

2) A label assignment strategy is proposed to select the optimal anchor point for each object. The top-ranked pixels are selected by self-learning for matching, and the accuracy of the detector is improved.

3) A center correction mechanism is introduced to make the predicted bounding box closer to the ground truth the center of the bounding box. This can avoid the effect of extremely abnormal bad cases, and the predicted index can be further optimized.

4) An anchor-free lightweight object detection network is proposed, which can reduce the computational cost and improve the accuracy of detection. Meanwhile, the proposed network can effectively detect small objects.

5) A comprehensive evaluation of the proposed method on MS COCO and Pascal Voc detection benchmarks. The experimental results not only demonstrate that the proposed detector has better detection accuracy than the other detectors, but also show the parameters of the model are substantially reduced.

The structure of this paper is as follows: Section II introduces three classes of object detection methods and discusses the application of advanced algorithms for object detection. The proposed object detection model is introduced, and the main components and algorithms are described in detail in Section III. In Section IV, a series of experiments and visualizations are carried out to demonstrate the superiority of the proposed model. Our work is summarized, and the research direction of follow-up work is pointed out in Section V.

II. RELATED WORK

A. ANCHOR-BASED OBJECT DETECTION

The workflow of the anchor-based object detector can be summarized as follows: Firstly, a set of predefined anchors is identified, and the image is divided into regions. Then, each candidate box is classified with a classifier to determine whether it contains the target object. Finally, according to the confidence level of the classifier and the position of the bounding box, the final detection result is output. R-CNN [27] first used CNNs for object detection. Although can greatly improve object detection performance, the problem of redundant computation is not solved. Fast R-CNN [28] uses the search method to construct candidate bound, but the speed is still insufficient for real-time requirements. Faster R-CNN [15] utilizes a fully convolutional network as a region suggestion network. The network generates the corresponding candidate windows with associated object scores, which can determine the probability of the appearance of an object. Compared with one-stage networks, the two-stage Faster R-CNN model is more advantageous in objecting high-precision and multi-scale detection. SSD [29] uses several different detection branches to detect multiple scales of objects, thus improving the accuracy of multi-scale object detection. YOLO9000 [30] uses a joint training technique of object classification and detection to expand the network to thousands of detection categories. The network has only one detection branch and lacks the capture of multi-scale contextual information, leading to poor performance for small object detection. Dynamic R-CNN [31] can automatically adjust the labels, and the loss function based on the statistical information is proposed to fit high-quality samples.

However, anchor-based object detection still has the following limitations: 1) The anchor-based detector is very sensitive to changes in anchors. 2) Fixed anchors compromise the versatility of the detector, resulting in the resize for the size and aspect ratio of the anchors for different tasks. 3) The detector needs to generate a large number of anchors to match the ground truth boxes. However, most of the anchors are marked as negative samples, which can cause an extreme sample imbalance. 4) During the training process, the IoU of all anchor boxes with ground truth boxes needs to be calculated, which consumes a lot of memory and time.

To overcome the above drawbacks, the anchor-free detector was proposed.

B. ANCHOR-FREE OBJECT DETECTION

The workflow of the anchor-free object detector can be summarized as follows: Firstly, the presence of an order is predicted by each pixel on the feature map, and a confidence map is generated. Then, for the pixels on each confidence map, the position of the target bounding box is predicted by a regressor. Finally, based on the confidence map and bounding box location, select the final detection result. Anchor-free detectors do need not predefined anchor boxes, and the detection process is implemented by the following two methods. One method is called the key point method. This method performs object detection by locating multiple predefined or self-learning key points, and then constraining the spatial extent of the object. Cornernet [19] transforms the bounding box detection into the key points detection without designing anchor boxes as priori boxes. However, the network only interests in edges and corner, which not only lacks the internal information but also requires many post-processing mechanisms. Inspired by Cornernet [19], Centernet [20] further improved the accuracy and recall by three points instead of two, which effectively overcome the drawback of too many wrong check boxes and insufficient recognition of intermediate information. ExtremeNet [32] estimated the network detection by standard key points. The object detection is transformed into key points estimation problem, thus avoiding region classification and implicit feature learning. RepPoints [33] represents objects as a collection of sample points. The model constrains the spatial extent of the objects and emphasizes semantically important local regions. YOLO [34] divides the object into some grids, and then predicted the bounding boxes and the corresponding probability. DenseBox [35] uses a circular region at the center to define a positive sample, and then predicts four distances from the circle to the object boundary. FSAF [36] integrates anchor-free branching and online feature selection mechanisms in RetinaNet. The central region of the object is defined as a positive sample and uses the distances from the four edges of the object for localization. FoveaBox [37] assigns different scale objects to different feature layers for direct classification and regression of pixel points. The network determines object locations based on the central concave structure. FCOS [21] defines all positions within the object bounding box as positive samples, and then detected the object by four distance values and centrality scores. Therefore, FCOS [21] has achieved promising performance in detection accuracy and speed.

However, these approaches still have some limitations: 1) The approaches have higher computational complexity and arithmetic power requirements. 2) The detection performance of small objects still needs to be improved. 3) The insufficient recall of boxes leads to accuracy that cannot reach the SOTA of the anchor-based method. To solve the above problems, we propose a lightweight backbone network, a small object enhancement module, and a label assignment strategy to enhance the detection performance.

C. SEMI-ANCHOR-FREE OBJECT DETECTION

The workflow of the semi-anchor-free object detector can be summarized as follows: Firstly, a pretrained convolutional neural network is used to extract features from the input image. Secondly, a proposal frame generation network is used to generate a series of proposal bounding boxes. Then, classification and regression operations are performed on each proposal bounding box. Finally, the detection results are post-processed using a non-maximal suppression algorithm to remove redundant candidate frames. The NMS algorithm filters the proposal bounding boxes based on the degree of overlap between them and retains the ones with the highest confidence level. Semi-anchor-free object detection is an object detection method that is different from the traditional anchor method. Instead of predefining the anchor, the location and scale information of the target is automatically generated by the network. The advantage of this method is that it can better adapt to the size and shape variations of different targets and improve the accuracy of object detection. SAFNet [38] proposes a new enhanced feature pyramid generation paradigm consisting of an adaptive feature fusion module (AFFM) and a self-enhanced module (SEM). The model can obtain a clean and enhanced feature pyramid because the paradigm adaptively integrates multi-scale representations in a nonlinear manner while suppressing redundant semantic information. Second, the adaptive anchor generator (AAG) generates a few suitable anchor boxes for each input image. With this semi-anchor-free approach, the detector overcomes its shortcomings while retaining the points of the anchor-based model. SAFDet [39] solves the problem that two-stage detectors are affected by horizontal recommendation misalignment and complex background interference in accurate object detection. First, the model uses a rotation-anchor-free branch (RAFB) to enhance the foreground feature by accurately regressing the oriented bounding box (OBB). Secondly, the center-prediction module (CPM) is introduced to enhance object localization and suppress background noise. 3SNet [40] object detector uses voxel-based methods to assist in learning point features and achieve anchor-free performance in inference. Then, the model designs a Directional Slice Attention to enhance the discriminability of features. Finally, the framework proposes a region of interest representation based on symmetric feature propagation to alleviate the obstacles caused by incomplete object scanning in autonomous driving scenarios.

III. PROPOSED METHOD

In this section, we propose an anchor-free lightweight detection model, which can effectively detect small objects. Firstly, based on the framework of FCOS, MobileNetV3 is incorporated into FPN to generate relevant features. The detection heads are designed to reduce the parameters of the whole model, while improving the detection speed in the inference phase. Secondly, the small object enhancement module is constructed. The module utilizes single stage headless (SSH) [41] face detector to focus on small objects, and then a scale-invariant network structure is used to further improve the small object detection performance. Thirdly, the label assignment strategy is proposed to design the loss function. The anchor bounding box matching constrain is overcome, and the flexibility of constructing bag-of-center has been enhanced. Finally, the center correction mechanism is introduced in the post-processing stage to tunning the prediction bounding box, and the center is optimally adjusted to achieve higher accuracy. The framework of the proposed model is shown in Fig. 1.

A. LIGHTWEIGHT BACKBONE NETWORK

FCOS avoids a large number of complex anchor calculations, which greatly reduces overhead consumption casts. At the same time, the detection performance is significantly improved over the anchor-based detectors. To further lighten the network, MobileNetV3 is selected as the backbone network of FCOS instead of the original backbone network, i.e. ResNet [42]. By analyzing and comparing the large and small versions of MobileNetV3, both of their structures have the advantage of being more lightweight. From the perspective of inferring efficiency, the choice of MobileNetV3 to replace the original FCOS backbone network is more conducive to the lightweight of the model.

The residual blocks of ResNet are replaced by using the Bneck structure in MobileNetV3. That is, the feature vectors output from the three Bneck structures of the inverse of MobileNetV3 is used as the input vectors of FPN. Different strategies are used for different backbone versions: 1) if the backbone uses the large version, the outputs of the 15th, 16th, and 17th Bneck are used as the inputs of FPN; 2) if the backbone uses the small version, the outputs of the 11th, 12th, and 13th Bneck are used as the input of FPN, as shown in Fig. 2.

Based on the above operations, the lightweight backbone network is obtained. Compared to the original detector, the lightweight backbone network takes into account the loss of information during small object detection, which improves the detector performance for small objects.

B. SMALL OBJECT ENHANCEMENT MODULE

To achieve better performance on small object detection, a small object enhancement module is constructed based on the single stage headless (SSH) face detector. SSH can focus on different-size objects and improve the perceptual field of the model. Therefore, the module embedded with SSH can enhance the performance for small objects.

The execution of the small object enhancement module consists of three steps. Firstly, the different scale features are obtained by the feature pyramid network, and then feature concentration is implemented to form a unified output feature vector. Secondly, the obtained feature vectors input into the small object detection model, as shown in Fig. 3(a). The model includes detection moudle (Fig. 3(b)) and context model (Fig. 3(c)). Detection model consists of a classifier and a regressor for the detection and localization of small objects. Context model uses a larger filter to increase the window size around objects. Different aspect ratios do not have a significant influence on the detection accuracy. Therefore, only anchor boxes with aspect ratio of 1 are retain in process of detection. Finally, the obtained anchors are matched and filtered by the anchor assignment step. In this paper, the proposed label assignment strategy is used to further improve the anchor assignment step, which is described in detail in subsection III-C.

C. LABEL ASSIGNMENT STRATEGY

Inspired by the Freeanchor [43] label assignment, the label assignment strategy is proposed to select the optimal anchor point for each object. Freeanchor uses a self-learning object detection method to match the anchor box. Each object is flexibly matched to the best anchor. The goal of Freeanchor is to discard the hand-designed anchor division while optimizing the following three visual object detection learning indexes. First, to achieve high recall, the detector requires to ensure that at least one prediction of the anchor box for each object is accurate. Second, to achieve high accuracy, the detector requires to classify the anchor box with a large regression error. Third, the prediction of the anchor box should abide by the non maximum suppression program. Otherwise, the predictions with precise positioning but low classification scores may be suppressed. Freeanchor's loss function is defined by:

$$L(\theta) = -\omega_1 \sum_i \log(Mean\text{-}max(X_i)) + \omega_2 FL(P\{a_j \in A_-\}(1 - P_{ij}^{bg}(\theta)) \tag{1}$$

$$X_i = \{P_{ij}^{cls}(\theta) P_{ij}^{loc}(\theta) | a_j \in A_i\} \tag{2}$$

where θ is the network parameter, w_1 and w_2 are the balance factor, and the Mean-max function is used to determine an optimal anchor box for each object from the anchor box set. When training is insufficient, almost all anchor boxes in the anchor box collection are used for training. When the network is fully trained, the confidence level of some anchor boxes increases in training progresses. X_i corresponds to the likelihood set of the anchor bag A_i, FL is the focal loss, A_i represents i-th anchor box set, $A_- \in A$ is the negative sample set, $a_j \in A_-$ represents an anchor box in negative sample set, $P\{a_j \in A_-\}$ is the probability that a_j does not match all objects, bg represents "background", $P_{ij}^{cls}(\theta)$ is the classification confidence, $P_j^{bg}(\theta)$ is the background confidence, $P_{ij}^{loc}(\theta)$ represents the definite position confidence, $P\{a_j \rightarrow b_i\}$ indicates the probability that a_j correctly predicts b_i.

Due to the lack of label assignment, the FCOS object detector has not fully demonstrated its advantages.

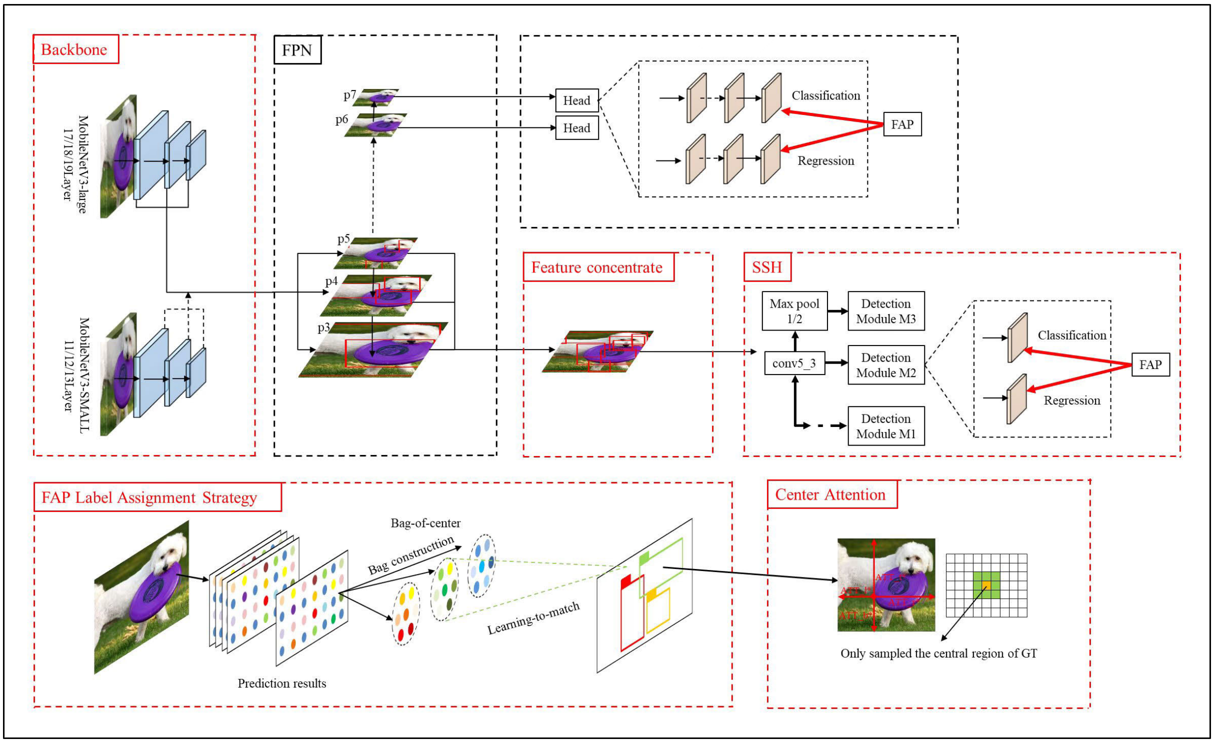


FIGURE 1. The framework of the proposed method.

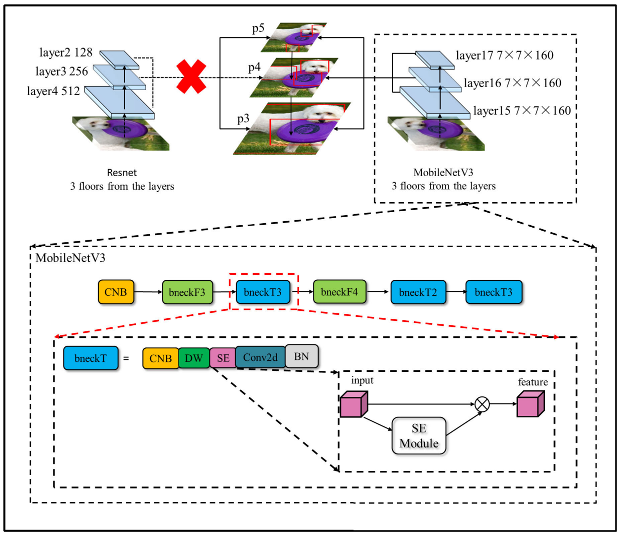


FIGURE 2. Lightweight backbone network.

However, Freeanchor is an anchor-based label assignment, which is not suitable for the anchor-free detector. Therefore, a label assignment strategy is proposed to optimize the Freeanchor label assignment, so that it can be used to improve the detection accuracy by introducing the label assignment to FCOS. The IoU of each anchor point is calculated, and then the optimal anchor point for each object is selected by the Mean-max function. In what follows, the two steps will be explained in detail.

1) P-IoU CALCULATION

As shown in the Fig. 4, box A is the ground truth bounding box, G is the anchor point on the feature map. t_A, b_A, l_A and r_A, represents the distance from point G to the top, bottom, left, and right of ground truth bounding box A. The artificial setting is to treat point G as the center point and assign a fictitious bounding box B around it. The shape of pseudo bounding box B is exactly the same as the truth box. t_B, b_B, l_B, and r_B represent the distance from point G to the left and right of fictitious bounding box B.

Firstly, t_A, b_A, l_A, r_A, t_B, b_B, l_B, r_B, S_A and S_B are calculated as follows:

$$\begin{aligned} t_B &= b_B = (t_A + b_A)/2 \\ l_B &= r_B = (l_A + r_A)/2 \\ S_A &= (t_A + b_A) * (l_A + r_A) \\ S_B &= (t_B + b_B) * (l_B + r_B) \end{aligned} \tag{3}$$

Then, the intersection box quadrangular coordinates t_*, b_*, l_*, r_* are calculated by minimum operation of the corresponding distances. The P-IoU of pseudo bounding box and real boxes is calculated by

$$P - IoU = \frac{|(l_* + r_*) * (t_* + b_*)|}{|S_A + S_B - (l_* + r_*) * (t_* + b_*)|} \tag{4}$$

2) OPTIMAL ANCHOR POINT SELECTION

Each anchor G_i that falls into box A is used to construct the bag-of-center. The P-IoU of the pseudo bounding box B_i formed by each anchor point G_i in the bag-of-center

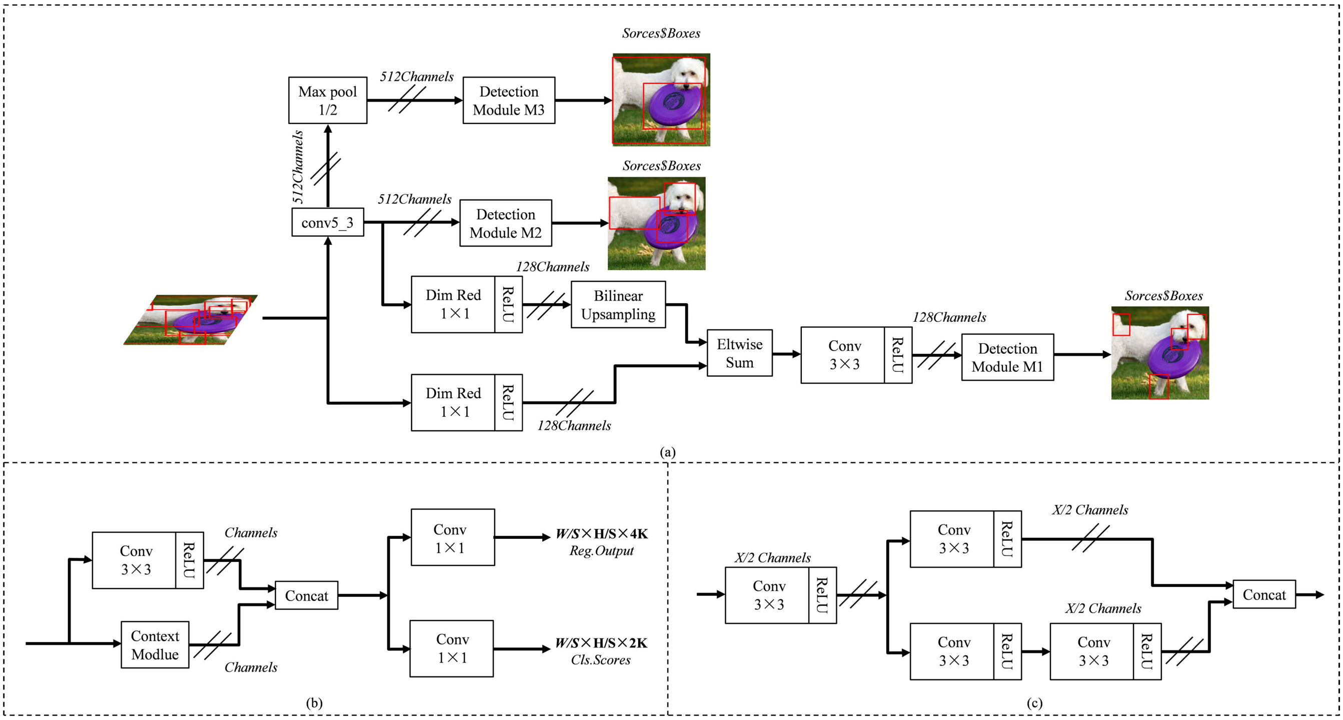


FIGURE 3. (a) Small object detection part model; (b) Detection part; (c) Context part.

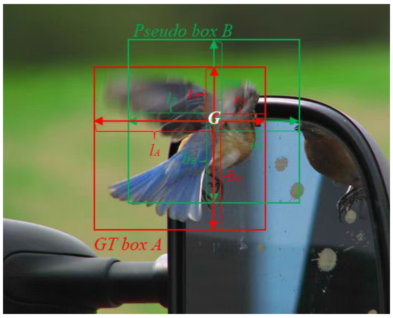


FIGURE 4. The ares in P-IoU loss for box regression.

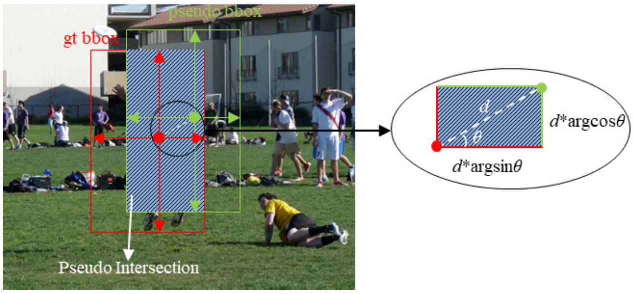


FIGURE 5. Center angle of the ground truth and the pseudo bounding boxes.

set is calculated. Then, the values of P-IoU is sorted in descending order. Finally, the anchor point with the highest confidence is selected as a positive sample, and the rest points are taken as negative samples. It should be noted that the loss term at the angle level is increased to prevent the effect of negative angles on the overall iteration.

The corresponding tilt angles of ground truth center G and I are calculated and are added to the loss function. The loss function of the angle value is calculated as follows:

$$L_{ang} = d(arg\cos(\theta)\omega_1 + arg\sin(\theta)\omega_2) \tag{5}$$

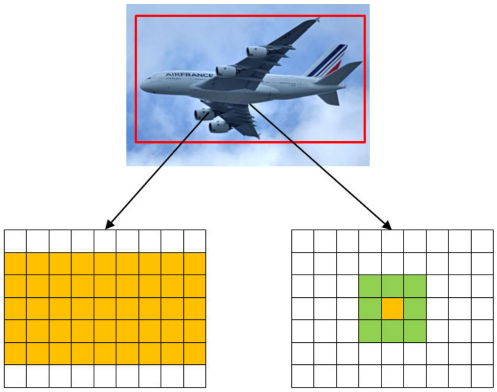
where the distance between the two points G and I is d, the angle between the horizontal direction is lower θ, and ω_1 and ω_2 are superparameters, as shown in the Fig. 5. The optimized loss function is shown as follows:

$$L'(\theta) = L(\theta) + L_{ang} \tag{6}$$

With the above process, the label assignment strategy can be performed in an adaptive manner to select the optimal anchor point for each object.

D. CENTER CORRECTION MECHANISM

The proposed label assignment method may produce some bad cases that can reduce the detection performance of the proposed method. Therefore, a center correction mechanism is proposed by introducing center attention to exclude the influence of extremely abnormal bad cases to a certain extent. Center attention is the tuning operation after the pseudo bounding box is outputted, which is directly used to calibrate the predicted anchor point using the attention parameter. The distance from the anchor of the pseudo bounding box to the four sides of the ground truth bounding box is calculated, and the attention parameter can be obtained by normalizing the



The ground sampling method used by FCOS and the improved sampling method

FIGURE 6. Improved sampling method.

distance. Then, the attention parameter is used for supervised learning.

The center attention can adaptively adjust the position of the center based on the distance between the different centers. The pixel near the center is assigned an initial value, and then the value is updated by the coincidence of the pseudo and ground truth bounding boxes. In order to update and solve this parameter, a sampling method based on pixels has been added, as shown in Fig. 6. When the radius of the center point is 1, the center of the virtual box should fall into the area of these 9 pixels, and HeatMap Loss is used to measure the deviation of the center of the pseudo bounding box. Y_{xyc} indicates the distance weight of *(x,y)* from the positive location of the target center, and the Y_{xyc} closer to 1 means that the *(x,y)* is closer to the positive location of the target center, the better the correction effect. $\hat{Y}_{xyc}$ represents the probability that the *(x,y)* is predicted as the center of the target. *C* indicates the detected target category. Assuming *C=1*, HeatMap Loss can be expressed as:

$$L_D = \frac{-1}{P}\sum_{xyc}\begin{cases}\log(a_{xyc})^{(1-a_{xyc})^{\gamma}} & if\ b_{xyc}=1\\ \log(1-a_{xyc})^{(1-b_{xyc})^{\omega}(a_{xyc})^{\gamma}} & otherwise\end{cases} \tag{7}$$

After obtaining the parameters of the corresponding nine-square grid, *(l, r, t, b)* is multiplied by the corresponding parameters to get the pseudo bounding box, and the purpose of making the center of the pseudo bounding box closer to the center of ground truth bounding box can be achieved.

E. ANCHOR-FREE LIGHTWEIGHT OBJECT DETECTION NETWORK

Anchor-free lightweight object detection network is proposed. It effectively gets rid of the limitation of fixed anchor box size on the detection model ability, and proposes to construct a small object enhancement module while using a lightweight backbone network to reduce the parameter set, so as to improve the detection ability for small objects. In addition, the label assignment strategy is proposed to determine the optimal feature, and the central correction mechanism is introduced to make the prediction more effective, which further improves the detection accuracy. The complete detection algorithm is given in Algorithm 1.

Algorithm 1 Object Detection Algorithm

```
Require: $I$: Input image.
    $B$: A set of ground-truth bounding boxes $B = b_i$.
    $A$: A set of anchors $A = a_j$ in image.
Ensure: $bboxes$, $scores$, $labels$
The last three layers of Backbone output serve as FNP inputs.
The output of the FPN serves as the input to SSH.
for $i$ = 1:MaxIter do
    Forward propagation:
        Predict class $a_j^{cls}$ and location $a_j^{loc}$ for each anchor $a_j \in A$.
    Construct bag-of-center.
        Calculate the IoU between each $a_j$ and $b_i$ by Eqs (3)-(4).
    $A_i \leftarrow$ Select the top anchor $a_j$ according to the IoU of $b_i$.
    for each positive sample do
        Calculate the probability that $a_j$ and $b_i$, taking the maximum value.
        Store the detected bounding box coordinates, confidence scores, and class labels in $bboxes$, $scores$, and $labels$.
    end for
    Calculate the angle-based configuration item through Eq. (5)
    Loss calculation by Eq. (6)
    Calculate the center of the normalization tutorial pseudo and ground truth bounding box by Eq. (7)
    Update $bboxes$, $scores$, $labels$
end for
Return $bboxes$, $scores$, $labels$
```

IV. EXPERIMENTS AND RESULTS

A. DATASETS AND EXPERIMENTAL ENVIRONMENT

1) INTRODUCTION TO DATASETS

The MS COCO [44] dataset is a large-scale dataset that can be used for object detection, semantic segmentation, and image captioning. It has more than 330K images, 1.5 million objects, 80 object categories (pedestrians, cars, elephants, etc.), and 91 material categories. Dataset images are divided into training, validation, and test sets. The representative pictures of the dataset are shown in Fig. 7 (a). The PASCAL Visual Object Classes is a world-class computer vision challenge that includes the following categories: image classification, object detection, object segmentation, action classification, etc. There are 20 main categories of interest in the Pascal VOC [45], [46] dataset. According to the set image,

TABLE 1. Comparisons of state-of-the-art detection methods on the MS COCO dataset.

Method	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN [15]	Resnet-FPN	36.2	59.1	39	18.2	39.0	48.2
Mask RCNN [47]	Resnet-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Libra RCNN [48]	Resnet-FPN	41.1	62.1	44.7	23.4	43.7	52.5
AutoDet [49]	Resnet	39.7	59.8	40.3	21.5	41.8	53.0
YOLOv3 [50]	DarkNet-53	33.0	57.9	34.4	18.3	25.4	41.9
SSD [29]	Resnet	31.2	50.4	33.3	10.2	34.5	49.8
RefineDet [51]	Resnet	36.4	57.5	39.5	16.6	39.9	51.4
RetinaNet [18]	Resnet-FPN	39.1	59.1	42.3	21.8	42.7	50.2
GHM [52]	Resnet-FPN	39.9	60.8	42.5	20.3	43.6	54.1
EMCA [53]	Resnet-FPN	38.1	60.6	50.2	23.6	42.2	48.4
CornerNet [19]	Hourglass	40.6	56.4	43.2	19.1	42.8	54.3
FCOS [21]	Resnet-FPN	41.5	60.7	45.0	24.4	44.8	51.6
ReFPN-FCOS [54]	Resnet-FPN	39.7	58.2	42.7	23.1	43.5	51.8
Pseudo-IoU [55]	Resnet-FPN	41.5	61.0	44.5	24.1	44.6	51.9
ObjectBox [56]	Resnet	46.8	65.9	49.5	26.8	49.5	57.6
Ours	**MobileNetV3-small**	**47.1**	**66.3**	**49.9**	**27.1**	**49.9**	**57.9**
Ours	**MobileNetV3-large**	**47.3**	**66.3**	**50.0**	**27.2**	**49.7**	**58.0**

(a) MS COCO

(b) Pascal VOC

FIGURE 7. The representative pictures of MS COCO and Pascal VOC datasets.

TABLE 2. Comparisons of state-of-the-art detection methods on the Pascal VOC dataset.

Method	VOC07mAP@IoU=0.5	VOC12mAP@IoU=0.5
Faster R-CNN [15]	73.2	70.4
FSNet [57]	80.7	×
SSD [29]	76.8	74.9
YOLO9000 [30]	78.6	73.4
SSD-MSN [58]	82.7	80.8
DF-SSD [59]	78.9	76.5
ObjectBox [56]	83.7	×
Ours-small	**84.1**	**84.5**
Ours-large	**83.9**	**84.5**

it is divided into training set and test set. The representative pictures of the dataset are shown in Fig. 7 (b).

2) EXPERIMENTAL ENVIRONMENT AND HYPERPARAMETER SETTINGS

The experimental environment is the PyTorch deep learning library, of which the Pytorch version is 1.13.1+cu117, the Torchvision version is 0.14.1+cu117, and the Python version is 3.8.15. MobileNetV3 is used as the backbone to build the entire network using a fixed Batch Normalization layer, where class number $C=81$ for MS COCO dataset and $C=21$ for Pascal VOC dataset. The network uses stochastic gradient descent (SGD) to train 256 epochs for MS COCO dataset and 256 epochs for Pascal VOC dataset, and warms up for 500 iterations. The initial learning rate is 0.01, and the weight decay and momentum are set to 0.0001 and 0.9. The input image is resized to 800 on the short side and 1333 on the long side. The proposed center correction mechanism is used as the post-processing mechanism.

B. COMPARISONS WITH STATE-OF-THE-ART METHODS

To verify the effectiveness of the proposed model, the detection results of the proposed model on MS COCO dataset are compared with four multi-stage methods, including Faster R-CNN [15], Mask RCNN [47], Libra RCNN [48], AutoDet [49], six one-stage methods, including YOLOv3 [50], SSD [29], RefineDet [51], RetinaNet [18], GHM [42], EMCA [53] and five anchor-free methods, including CornerNet [19], FCOS [21], ReFPN-FCOS [54], Pseudo-IoU [54], ObjectBox [56] methods, and the comparison results are shown in Table 1.

Table 1 shows that the proposed model achieves the best detection performance in comparison to all other models in terms of AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L metrics. The proposed model for the large version improves the detection of mAP by 0.5%, and the proposed model for the small version improves the detection of mAP by 0.3%. The upper limit of fit of MobileNetV3 is worse than Resnet. Therefore, the results demonstrate that the model enhances the detection performance for small object by embedding the small object detection enhancement module. For different sizes (small/medium/large) objects, the proposed model obtains higher detection accuracy than state-of-the-art methods, indicating that the proposed model not only performs well for the small objects but also achieves good detection performance for medium/large objects.

TABLE 3. Comparisons of the model with and without small object enhancement module.

Method	Small object enhancement module	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MobileNetV-large	×	46.39	65.68	49.29	26.60	49.22	57.29
MobileNetV-large	√	**46.71**	**65.86**	**49.42**	**26.76**	**49.38**	**57.53**
MobileNetV-small	×	46.49	65.56	49.23	26.49	49.28	57.28
MobileNetV-small	√	**46.76**	**65.65**	**49.48**	**26.76**	**49.38**	**57.57**

TABLE 4. Comparisons of the model with and without label assignment strategy.

Method	Label assignment strategy	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MobileNetV-large	×	46.39	65.68	49.29	26.60	49.22	57.29
MobileNetV-large	√	**46.77**	**65.80**	**49.35**	**26.79**	**49.32**	**57.58**
MobileNetV-small	×	46.49	65.56	49.23	26.49	49.28	57.28
MobileNetV-small	√	**46.79**	**65.80**	**49.59**	**26.81**	**49.37**	**57.59**

TABLE 5. Comparisons of the model with and without center correction mechanism.

Method	Center correction mechanism	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MobileNetV-large	×	46.39	65.68	49.29	26.60	49.22	57.29
MobileNetV-large	√	**46.67**	**65.76**	**49.47**	**26.74**	**49.36**	**57.61**
MobileNetV-small	×	46.49	65.56	49.23	26.49	49.28	57.28
MobileNetV-small	√	**46.68**	**65.92**	**49.41**	**26.69**	**49.46**	**47.57**

To further evaluate the performance of the proposed detection model, the experiments on the Pascal Voc dataset are conducted. The comparison with two multi-stage methods, including Faster R-CNN [15], FSNet [57], four one-stage methods, including SSD [29], YOLO9000 [30], SSD-MSN [58], DF-SSD [59] and one anchor-free method ObjectBox [56], is carried out, shown as Table 2.

As can be seen from Table 2, the proposed model has significant advantages in all indexes compared to all other methods on Pascal VOC dataset. This demonstrates that the proposed model not only outperforms the traditional multi-stage/single-stage models, but also has better performance than other anchor-free models. The proposed model for the large version improves the mAP index by 0.2%, and the proposed model for the small version improves the mAP index by 0.4%. The superiority of the proposed model is further verified.

In Tables 1 and 2, compared with ObjectBox, the proposed method has only a 0.3%-0.5% improvement. However, compared with other state-of-the-art methods, the proposed method has a significant improvement.

C. ABLATION STUDY

Our method proposes three modules to FCOS, including small enhancement module, label assignment strategy, and center correction mechanism. To evaluate the contributions of the three modules, the ablation experiments for each module are conducted. Meanwhile, the contribution of the lightweight network is verified in terms of network model size and processing time In this experiment, the influence of the three proposed modules is separately verified. To avoid the effect of crossover experiments, the MS COCO dataset is used as the validation dataset, and the remaining modules are excluded when the experiment with a single improvement module.

1) THE PERFORMANCE OF SMALL OBJECT ENHANCEMENT MODULE

The influence of the small object enhancement module is first discussed. To demonstrate the contribution of the small object enhancement module, we compare the performance of the proposed method with and without the small object enhancement module, then show the results in Table 3.

Table 3 shows that the overall metric (AP) of the model using the small object enhancement module is 0.32% higher for the large version than the original framework and 0.27% higher for the small version than the original framework. The improvement of the AP_S metric indicates that the detection ability of the model for small objects is enhanced after embedding the small object detection enhancement module. And meanwhile, the model with the small object enhancement module maintains the detection performance for medium and large objects in terms of AP_M and AP_L.

2) THE PERFORMANCE OF LABEL ASSIGNMENT STRATEGY

The influence of the label assignment strategy is investigated. To demonstrate the contribution of the label assignment strategy, we compare the performance of the proposed method with and without label assignment strategy, then show the results in Table 4.

Table 4 shows that the performance of the model with label assignment strategy has been significantly enhanced in terms of all metrics. There are two reasons for the improved performance: 1) The introduction of the self-learning idea makes each anchor not mechanically assigned. The strategy constructs several sets of bag-of-center before the assignment, and then each set is used to learn the corresponding matching probability to achieve the final matching. 2) The offset probability of anchor and the center is fully considered, and the size of the prediction bounding box is adjusted by introducing a distance factor. The prediction bounding box is dynamically adjusted according to the P-IoU threshold, thus it has higher robustness than the hard-crafted IoU threshold. When dealing with some extremely bad cases, the model can better fit the ground truth bound box.

3) THE PERFORMANCE OF CENTER CORRECTION MECHANISM

The purpose of the center correction mechanism is to tunning the prediction bounding box. The influence of the center correction mechanism is discussed. To demonstrate the contribution of the center correction mechanism, we compare the performance of the proposed method with and without center correction mechanism, then show the results in Table 5.

Table 5 shows that the performance of the model with label assignment strategy has been enhanced to some extent in terms of all metrics. From the perspective of fixed IoU threshold, the model has significant improvement in terms of AP$_5$0. This indicates that label assignment strategy is effective in correcting the large deviation of the center points between the prediction and ground truth bounding boxes, which is in line with the improvement target mentioned above. From the perspective of the detection object size, the improvement is in equilibrium. This indicates that the similar effect can be achieved for different sizes of detection targets without obvious target size tendency.

In Table 3, 4 and 5, although each numeric value has a difference of no more than 0.5% AP when compared in pairs, the overall improvement of the proposed method is significant.

4) THE PERFORMANCE OF LIGHTWEIGHT NETWORK

The purpose of the backbone replacement is to lightweight the network. To demonstrate the contribution of the lightweight network, we compare the network model size and processing time of the network with and without the lightweight network, then show the results in Table 6.

Table 6 shows that the performance of the proposed method with the lightweight network has been enhanced in terms of the network model size and processing time. In comparison with the original backbone, the model size of the large version is reduced to 40% and the inference time is reduced to 62%. The small version of the model size is reduced to 32% and the inference time is reduced to 46%. The proposed model achieves a significant reduction in inference speed while maintaining a small number of parameters.

TABLE 6. Comparisons of the model with and without lightweight backbone.

Backbone	Param(M)	Inference time(ms)
Resnet-50	30.85	105.72
MobileNetV-large	**12.34**	**65.64**
MobileNetV-small	**9.87**	**48.35**

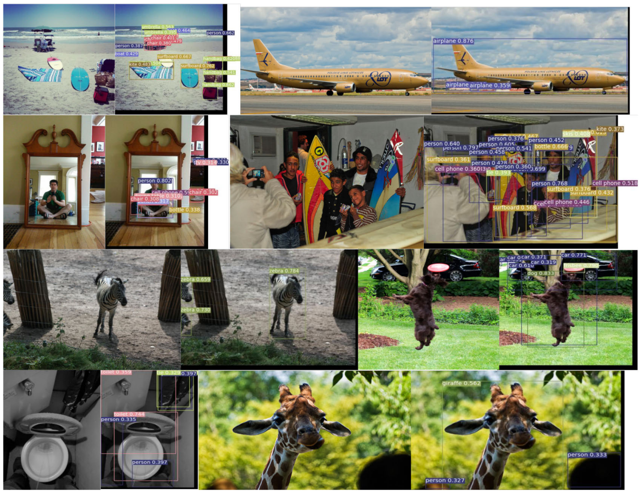


FIGURE 8. Visualization of detection results by the proposed method for MS COCO dataset.

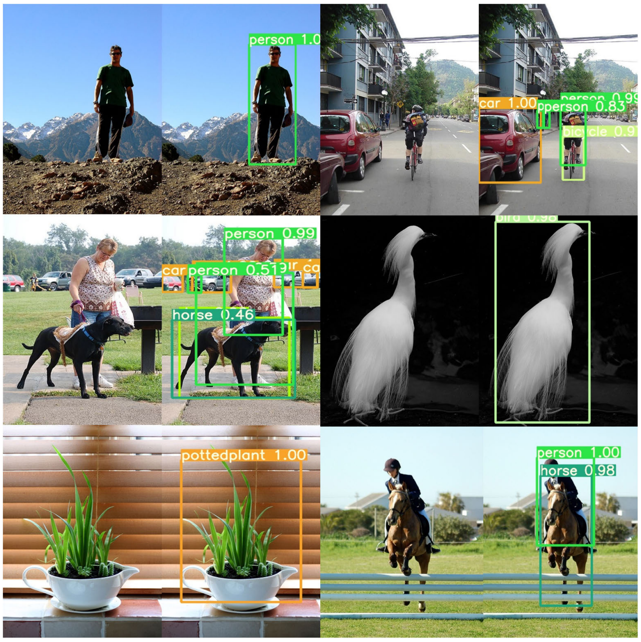


FIGURE 9. Visualization of detection results by the proposed method for Pascal VOC dataset.

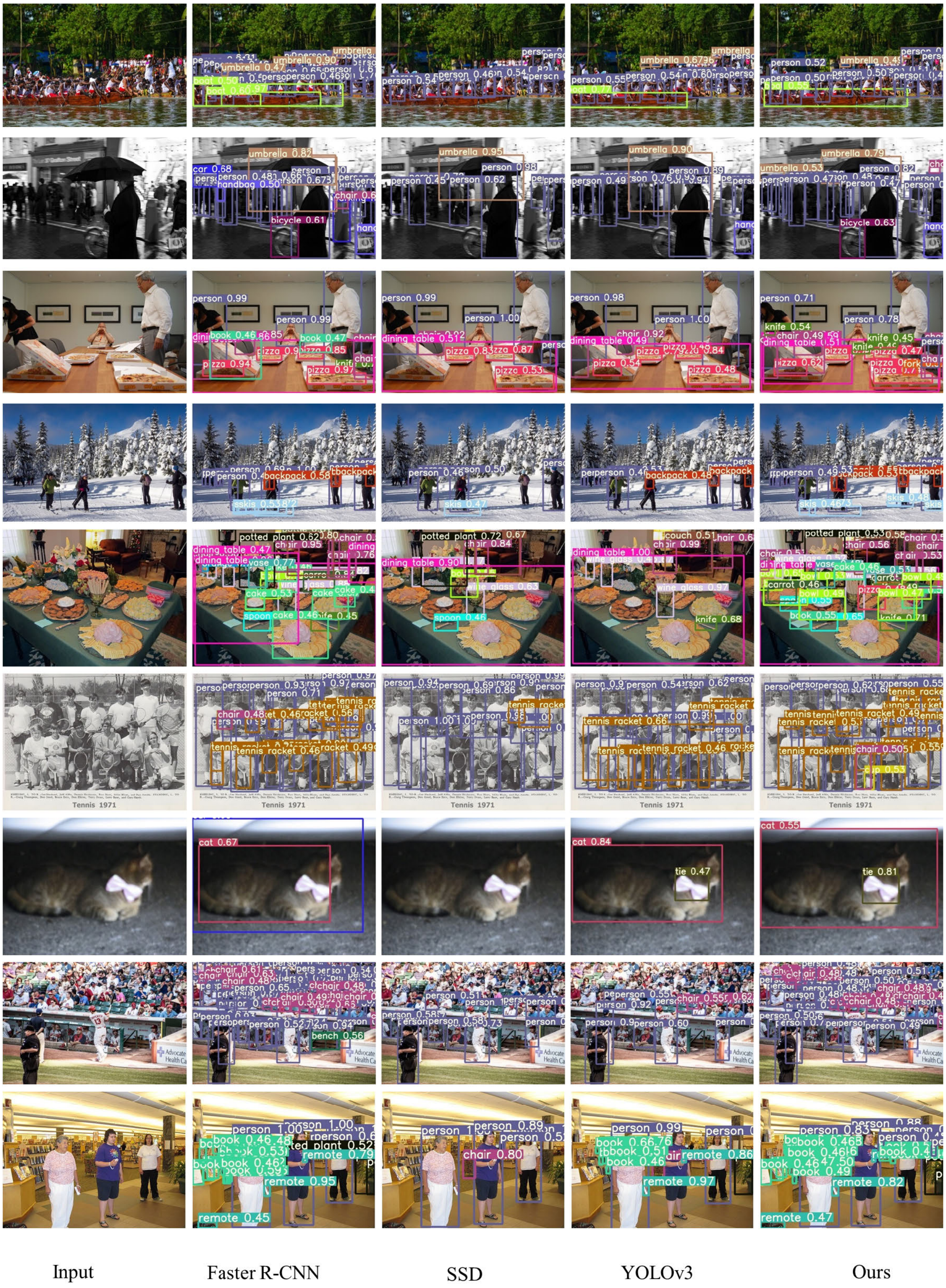


FIGURE 10. Qualitative comparisons with some representative methods for MS COCO datasets.

D. QUALITATIVE RESULTS

1) QUALITATIVE EVALUATION

The qualitative results on MS COCO and Pascal VOC datasets are shown in Figs. 8 and 9. Figs. 8 and 9 include different object detection scenarios, such as crowded, occluded, highly overlapping, and small object. The results demonstrate that the proposed model achieves better performance for different object detection scenarios.

2) QUALITATIVE COMPARISON

To further demonstrate the qualitative performance of the proposed model, the qualitative results on the MS COCO

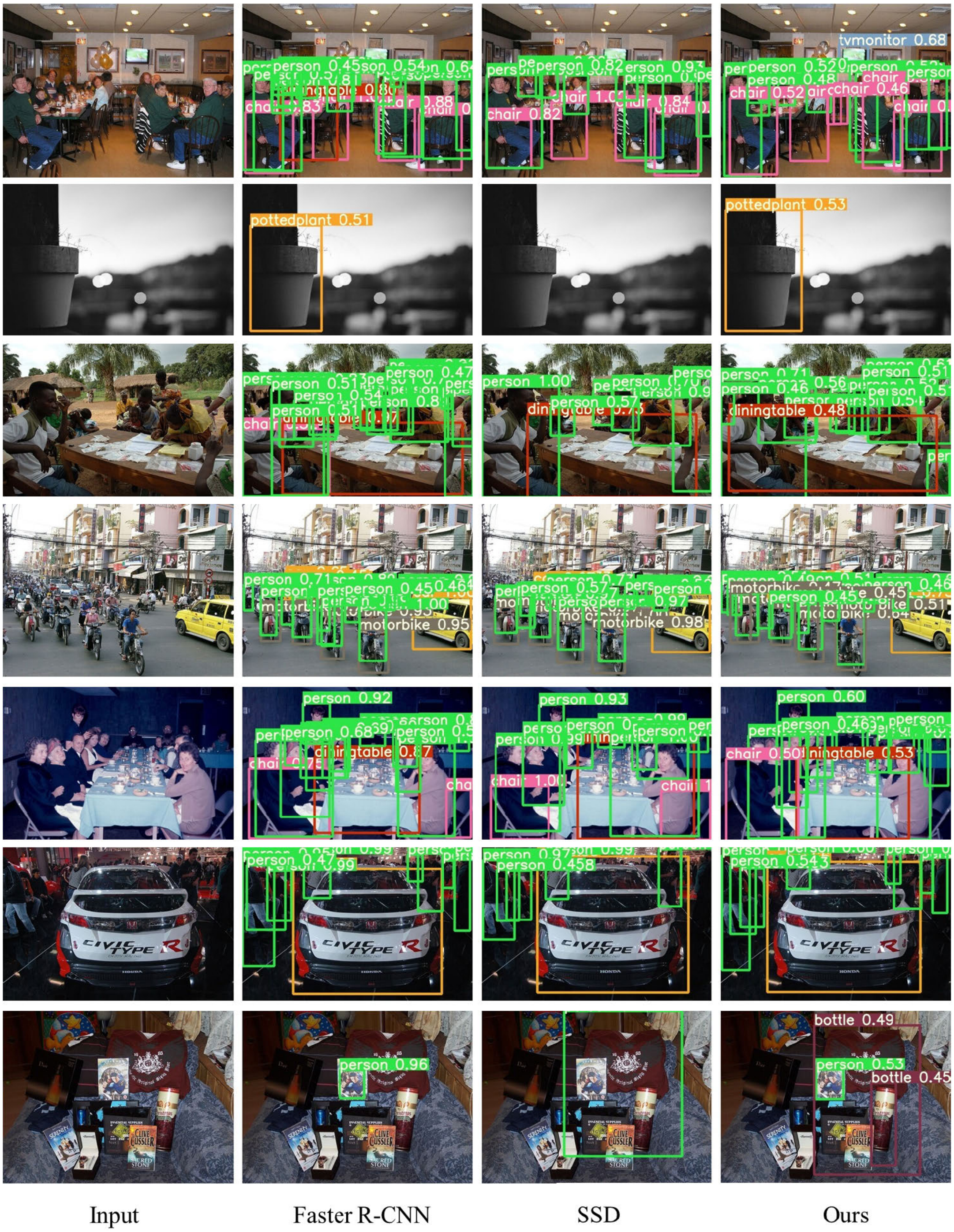


FIGURE 11. Qualitative comparisons with some representative methods for Pascal VOC datasets.

dataset are compared with Faster R-CNN, SSD, YOLOV3, and the results are shown in Fig. 10. The qualitative results of the proposed model on Pascal VOC dataset are compared with Faster R-CNN, SSD methods, and the results are shown in Fig. 11. From Figs. 10 and 11 can see that the proposed model achieves superior detection performance for different object detection scenarios. (1) More small target objects can be detected; (2) The accuracy of detection is improved; (3) Compared with other detectors, the possibility of missing or misdetection is smaller, and there are more types of detection.

V. CONCLUSION

This paper proposes an anchor-free lightweight detection method that more focuses on small object detection. Firstly, the lightweight network MobileNetV3 is used as the backbone, which significantly reduces the parameters of the whole model and also significantly improves the detection speed in the inference stage. Secondly, the proposed model embeds a small object enhancement module, and the detection efficiency for small objects is further improved by adding a scale-invariant network structure. Unlike the two-stage proposal/classification approach, it detects small objects in a one-stage, and locates and detects them simultaneously from the early convolutional layers of the classification network. In addition, SSH is introduced to be scale-invariant when detecting different small object scales in a single forward channel of the network, instead of processing the input pyramid thus significantly reducing the detection time. Finally, the label assignment strategy and center correction mechanism are proposed to further enhance overall accuracy by selecting the most suitable center for the object in a self-learning manner. The experimental results show that: 1) The improvement in the APS index indicates that the performance for small objects has been enhanced. 2) The robustness of the method with label assignment strategy is higher than the other methods with the hard-crafted IoU threshold. 3) The proposed method has well performance for different sizes of detection objects.

Although the proposed method has the advantages of being lightweight, focusing on small targets, and high detection accuracy, the label allocation strategy in the method still has room for improvement. It is necessary to combine the output of the center correction mechanism to do the tuning operation after the prediction of the bounding box, and use the ground truth center attention parameter to calibrate the obtained predicted center. Therefore, we will focus on several promising research directions including: 1) optimize the allocation strategy of anchor-free detection, and further studying a more concise allocation strategy to improve the detection performance of the model; 2) propose a lighter network to improve the efficiency; 3) modularize the method to achieve a plug-and-play effect.

REFERENCES

[1] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, “From handcrafted to deep features for pedestrian detection: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4913–4934, Sep. 2022.

[2] Z. Shao, G. Cheng, J. Ma, Z. Wang, J. Wang, and D. Li, “Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic,” *IEEE Trans. Multimedia*, vol. 24, pp. 2069–2083, 2022.

[3] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, and H. Zhang, “SANet: A slice-aware network for pulmonary nodule detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4374–4387, Aug. 2022.

[4] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, “Extended feature pyramid network for small object detection,” *IEEE Trans. Multimedia*, vol. 24, pp. 1968–1979, 2022.

[5] X. Ming, F. Wei, T. Zhang, D. Chen, N. Zheng, and F. Wen, “Group sampling for scale invariant face detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 985–1001, Feb. 2022.

[6] W. Wang, X. Wang, W. Yang, and J. Liu, “Unsupervised face detection in the dark,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1250–1266, Jan. 2023.

[7] S. Zhang, C. Chi, Z. Lei, and S. Z. Li, “RefineFace: Refinement neural network for high performance face detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4008–4020, Nov. 2021.

[8] Z. Yao and L. Wang, “Boundary information progressive guidance network for salient object detection,” *IEEE Trans. Multimedia*, vol. 24, pp. 4236–4249, 2022.

[9] N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, “Employing bilinear fusion and saliency prior information for RGB-D salient object detection,” *IEEE Trans. Multimedia*, vol. 24, pp. 1651–1664, 2022.

[10] Q. Ren, S. Lu, J. Zhang, and R. Hu, “Salient object detection by fusing local and global contexts,” *IEEE Trans. Multimedia*, vol. 23, pp. 1442–1453, 2021.

[11] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, “Real-time scene text detection with differentiable binarization and adaptive scale fusion,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 919–931, Jan. 2023.

[12] B. A. Plummer, K. J. Shih, Y. Li, K. Xu, S. Lazebnik, S. Sclaroff, and K. Saenko, “Revisiting image-language networks for open-ended phrase detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2155–2167, Apr. 2022.

[13] T. Liang, H. Bao, W. Pan, X. Fan, and H. Li, “DetectFormer: Category-assisted transformer for traffic scene object detection,” *Sensors*, vol. 22, no. 13, p. 4833, Jun. 2022.

[14] T. Liang, H. Bao, W. Pan, and F. Pan, “Traffic sign detection via improved sparse R-CNN for autonomous vehicles,” *J. Adv. Transp.*, vol. 2022, pp. 1–16, Mar. 2022.

[15] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[16] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[19] H. Law and J. Deng, “CornerNet: Detecting objects as paired keypoints,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 765–781.

[20] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint triplets for object detection,” 2019, *arXiv:1904.08189*.

[21] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.

[22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size,” 2016, *arXiv:1602.07360*.

[23] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” 2017, *arXiv:1707.01083*.

[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, *arXiv:1704.04861*.

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” 2018, *arXiv:1801.04381*.

[26] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, “Searching for MobileNetV3,” 2019, *arXiv:1905.02244*.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[28] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, “SSD: Single shot MultiBox detector,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 21–37.

[30] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[31] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, “Dynamic R-CNN: Towards high quality object detection via dynamic training,” in *Proc. ECCV*, Aug. 2020, pp. 260–275.

[32] X. Zhou, J. Zhuo, and P. Krähenbühl, “Bottom-up object detection by grouping extreme and center points,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 850–859.

[33] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," 2019, *arXiv:1904.11490*.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[35] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "DenseBox: Unifying landmark localization with end to end object detection," 2015, *arXiv:1509.04874*.

[36] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849.

[37] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[38] Z. Jin, B. Liu, Q. Chu, and N. Yu, "SAFNet: A semi-anchor-free network with enhanced feature pyramid for object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 9445–9457, 2020.

[39] Z. Fang, J. Ren, H. Sun, S. Marshall, J. Han, and H. Zhao, "SAFDet: A semi-anchor-free detector for effective detection of oriented objects in aerial images," *Remote Sens.*, vol. 12, no. 19, p. 3225, Oct. 2020.

[40] H. Peng, G. Tong, and Y. Shao, "3SNet: Semi-anchor-free 3D object detector with slice attention and symmetric features propagation," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 18, 2023, doi: 10.1109/TITS.2023.3292945.

[41] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis, "SSH: Single stage headless face detector," 2017, *arXiv:1708.03979*.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] X. Zhang, F. Wan, C. Liu, R. Ji, and Q. Ye, "FreeAnchor: Learning to match anchors for visual object detection," 2019, *arXiv:1909.02466*.

[44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

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

[47] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[48] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830.

[49] Z. Li, T. Xi, G. Zhang, J. Liu, and R. He, "AutoDet: Pyramid network architecture search for object detection," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1087–1105, Jan. 2021.

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

[51] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.

[52] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proc. AAAI Conf. Artif. Intell.*, Sep. 2019, vol. 33, no. 1, pp. 8577–8584.

[53] E. M. Bakr, A. El-Sallab, and M. Rashwan, "EMCA: Efficient multiscale channel attention module," *IEEE Access*, vol. 10, pp. 103447–103461, 2022.

[54] J. Zeng, J. Xiong, X. Fu, and L. Leng, "ReFPN-FCOS: One-stage object detection for feature learning and accurate localization," *IEEE Access*, vol. 8, pp. 225052–225063, 2020.

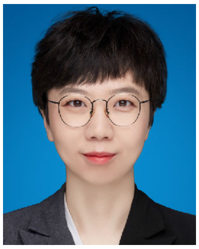
[55] J. Li, B. Cheng, R. Feris, J. Xiong, T. S. Huang, W.-M. Hwu, and H. Shi, "Pseudo-IoU: Improving label assignment in anchor-free object detection," 2021, *arXiv:2104.14082*.

[56] M. Zand, A. Etemad, and M. Greenspan, "ObjectBox: From centers to boxes for anchor-free object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2022, pp. 390–406.

[57] J. Jiang, H. Xu, S. Zhang, Y. Fang, and L. Kang, "FSNet: A target detection algorithm based on a fusion shared network," *IEEE Access*, vol. 7, pp. 169417–169425, 2019.

[58] Z. Chen, K. Wu, Y. Li, M. Wang, and W. Li, "SSD-MSN: An improved multi-scale object detection network based on SSD," *IEEE Access*, vol. 7, pp. 80622–80632, 2019.

[59] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE Access*, vol. 8, pp. 24344–24357, 2020.



WEINA WANG received the M.Sc. degree in applied mathematics from Dalian Maritime University, China, in 2007, and the Ph.D. degree in control theory and control engineering from the Dalian University of Technology, China, in 2016. She is currently a Professor with the School of Sciences, Jilin Institute of Chemical Technology. Her research interests include object detection, data mining, and time series analysis.

YUNYAN GOU received the bachelor's degree in engineering from Lanzhou Jiaotong University, China, in 2020. She is currently pursuing the master's degree in electronic information with the Jilin Institute of Chemical Technology. Her main research interest includes object detection.