The Visual Computer (2023) 39:2969–2980
https://doi.org/10.1007/s00371-022-02503-4

ORIGINAL ARTICLE

MFANet: Multi-scale feature fusion network with attention mechanism

Gaihua Wang[1,2] · Xin Gan[1] · Qingcheng Cao[1] · Qianyu Zhai[1]

Accepted: 11 April 2022 / Published online: 11 May 2022

Abstract
In order to improve the detection accuracy of the network, it proposes multi-scale feature fusion and attention mechanism net (MFANet) based on deep learning, which integrates pyramid module and channel attention mechanism effectively. Pyramid module is designed for feature fusion in the channel and space dimensions. Channel attention mechanism obtains feature maps in different receptive fields, which divides each feature map into two groups and uses different convolutions to obtain weights. Experimental results show that our strategy boosts state-of-the-arts by 1–2% box AP on object detection benchmarks. Among them, the accuracy of MFANet reaches 34.2% in box AP on COCO dataset. Compared with the current typical algorithms, the proposed method achieves significant performance in detection accuracy.

Keywords Deep learning · Object detection · Multi-feature fusion · Attention mechanism

1 Introduction

Object detection refers to a type of computer vision technology that can classify and locate objects. It is widely used in many fields, such as face recognition [1], gait recognition [2], tracking [3], and crowd counting [4–6]. Traditional object detection [7,8] requires manual feature extraction, which is difficult to obtain robust characteristics and very sensitive to external environmental noise.

With the development of deep learning and the progress of hardware, object detection algorithms based on convolutional neural networks (CNN) develop rapidly. They are mainly divided into two-stage and single-stage algorithms. The two-stage detection algorithm first generates a region proposal, then classifies and calibrates the candidate regions, and obtains the final detection result. Gkioxari [9] proposed RCNN in 2015, which finds the boxes that may contain objects according to the region proposal. Then, the method predicts the bounding box offset and classifies each region. In 2017, Faster-RCNN [10] introduced a Region Proposal Network (RPN) that shares features with the detection network. And it realizes nearly cost-free region proposals. Cai et al. [11] proposed Cascade-RCNN in 2018, which uses different IoU thresholds to divide positive and negative samples, and makes the detector of each stage focus on detecting the proposal of the IoU in a certain range.D2Det [12] introduced a dense local regression that predicts multiple dense box offsets for an object proposal in 2020. Sun et al. [13] proposed Sparse R-CNN in 2021, which uses a fixed number of learnable boxes to replace anchors. These two-stage algorithms have higher detection accuracy. But they have slower detection speed than the single-stage algorithm.

The single-stage detection algorithm directly gives the final detection result without generating candidate boxes. In 2016, YOLO [14] was proposed to frame the object detection as a regression problem. It uses the image as input to directly implement object location regression and classification. SSD [15] was introduced to output a set of default boxes with different aspect ratios at each feature map location. In 2019, Tian et al. [16] proposed FCOS. And the algorithm

G. Wang, X. Gan, Q. Cao and Q. Zhai: These authors contributed equally to this work.

✉ Xin Gan
1989556916@qq.com

Gaihua Wang
20130006@hbut.edu.cn

Qingcheng Cao
2692556245@qq.com

Qianyu Zhai
zhaiqianyu233@163.com

[1] School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China

[2] Hubei Key Laboratory for High-efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China

completely avoids the complex calculations related to the anchors by eliminating the pre-defined anchors. RepPoints [17] learns the offset of deformable convolutions through direct supervision of localization and classification and generates pseudo-boxes by sampling points. LIN et al. [18] designed RetinaNet based on FocalLoss in 2020, which can address the class imbalance. PAA [19] proposes a probabilistic model for assigning labels to anchors in view of the assignment of anchor labels in the current anchor-based model. In 2021, VFNet [20] proposed IoU-aware classification score (IACS) to classify detection, and it combines varifcoal loss, star-shaped bounding box and bounding box refinement to improve detection accuracy. Chen et al. [21] proposed YOLOF, which uses an expansion encoder and unifies matching to narrow the performance gap between SISO and MIMO encoder. The single-stage object detection algorithm does not have a region proposal process. It only needs to be sent to the network once to predict all bounding boxes. The speed is relatively fast, and the number of parameters is small, but the accuracy is lower than the two-stage algorithm.

The ATSS [22] is a one-stage object detection algorithm. The network consists of three parts: backbone, neck and heads. Backbone uses a classification network that removes the fully connected layer to extract image features. Neck is used for feature fusion to achieve multi-scale detection of objects, which adopts the feature pyramid network (FPN) to fuses deep feature maps with low-level feature maps through upsampling to obtain rich semantic information. In order to better calculate the classification and regression loss, heads adopt an adaptive sample selection method to realize the classification and regression of objects.

We believe that FPN only performs feature fusion in the spatial dimension, and this fusion method will lead to the loss of semantic information. Therefore, the paper proposes multi-feature fusion network with attention mechanism (MFANet)-based ATSS. It proposes feature fusion to obtain rich semantic features and adopts a channel attention mechanism to strengthen important features and suppress non-important features. The major contributions of this study can be summarized as follows:

(1) Multi-scale feature fusion uses upsampling and compression operations in the two dimensions of space and channel to fuse feature maps of different sizes. Finally, feature maps of different dimensions are added to obtain rich semantic features.

(2) The attention mechanism obtains feature maps of different receptive fields to get rich contextual information. It divides each feature map into two groups, and realizes channel attention learning of local cross-channel interaction without dimensionality reduction by one-dimensional convolution.

(3) It has achieved remarkable results on the Ms CoCo2017 dataset and PASCAL VOC Datasets.

2 Related work

2.1 Multi-scale feature fusion

To solve the problem of predicting objects of different sizes, Lin et al. [23] proposed the famous feature pyramid network (FPN). And the basic idea is to combine the fine-grained spatial information of the shallow feature map and the semantic information of the deep feature map to detect multi-scale objects. On this basis, many researchers have proposed improved FPN structures. Liu et al. [24] proposed PANet, which first uses up-sampling to fuse feature maps of different sizes and then performs down-sampling feature fusion. NAS-FPN [25] is a combination of top-down and bottom-up connections, which can be integrated across a range. AugFPN [26] uses consistent supervision, residual feature augmentation and soft RoI selection modules for FPN defects. BiFPN [27] performs weight fusion of features to learn the importance of different input features. Qiao et al. [28] proposed Recursive-FPN, which inputs the output of traditional FPN to backbone for a second cycle.

These modules only effectively integrate features in the spatial dimension. The information between different channels may be correlated or redundant. Therefore, we propose the multi-dimensional feature pyramid network (MFPN), which adds a branch to fuse feature in the channel dimension. The branch compresses all channel information together and performs semantic fusion and finally obtains rich semantic spatial information.

2.2 Attention mechanism

The attention mechanism originates from the study of human vision. And it was first applied in the field of natural language to realize the efficient allocation of information processing resources. In recent years, the attention mechanism has been rapidly developed in the field of computer vision. In 2018, Hu et al. [29] proposed SENet, which implements the channel attention mechanism through three parts: squeeze, incentive, and scale. In 2018, non-local neural networks were proposed [30] to compute the response at the current area as a weighted sum of the global area. DANet [31] was proposed to use a dual attention network to adaptively integrate local features and global dependencies in 2019. And two types of attention modules are added to the traditional expanded FCN to simulate the semantic interdependence in space and channel dimensions, respectively. In 2020, ASNet [32] introduced a density attention network, and it can provide ASNet with attention masks of different density levels. In 2021,Hou et al. [33] proposed coordinate attention. It captures not only cross-channel information, but also direction-aware and position-sensitive information, which enables the model to more accurately locate and identify the target area.

To show the correlation between different channels, it should strengthen important features and suppress non-important features. This paper proposes multi-receptive field attention mechanism (MFA). It uses 4 parallel branches of different receptive fields. Each branch is divided into two groups, which uses different convolution kernels to obtain channel weights.

3 Our approach

MFANet consists of three parts: backbone, neck and heads. The backbone uses resnet50, which is used to extract the features of the image. Neck is used to connect backbone and heads. And it is used to fuse features of different sizes. Heads are used for object detection to achieve object classification and regression. The loss function is divided into classification loss, regression loss and center loss. The classification loss function adopts FocalLoss, the regression loss adopts GIoULoss, and the center loss adopts CrossEntropyLoss. The network structure is shown in Fig. 1.

3.1 MFPN

The MFPN module is shown in Fig. 2. $[c_3, c_4, c_5], c \in R^{(B,C,H,W)}$ denotes the input feature map. The sizes are $[[B, C_3, H_3, W_3], [B, C_4, H_4, W_4], [B, C_5, H_5, W_5]]$, where B, C, H, W indicate the batch size, channel size, spatial height, and width. The size of C, H, W is expressed by Equation 1.

$$\begin{aligned} C_5 &= 2 * C_4 = 4 * C_3 \\ H_3 &= 2 * H_4 = 4 * H_5 \\ W_3 &= 2 * W_4 = 4 * W_5 \end{aligned} \tag{1}$$

It uses 1*1 convolution to change their channel to the same size C.

The branch1 is to conduct feature fusion in the channel dimension. First, it uses the unfold operation to change the shape of the feature maps. After that, the shape of the feature maps is $[B, C^{'}, L]$. The size of $C^{'}$ is $C^{'} = C * K * K$. And L is expressed by Eq. 2.

$$\begin{aligned} H^{'} &= 1 + \frac{H + 2 * padding - K}{stride} \\ W^{'} &= 1 + \frac{W + 2 * padding - K}{stride} \\ L &= H^{'} * W^{'} \end{aligned} \tag{2}$$

where K is the size of the convolution kernel, and $C^{'}$ represents the size of the sliding window. The padding is the padding size, stride is the step size, and L is the number of sliding windows. Then the output is expressed by Eq. 3.

$$\begin{aligned} c^{'}5 &= F_{UF}(W_a * c_5) \\ c^{'}41 &= F_{UF}(W_a * c_4) \\ c^{'}42 &= F_{UF}(W_a * c_4) \\ c^{'}3 &= F_{UF}(W_a * c_3) \end{aligned} \tag{3}$$

where W_a indicates the $1 * 1$ convolution layer and F_{UF} is an unfold operator. Finally, the output of branch1 is expressed by Eq. 4.

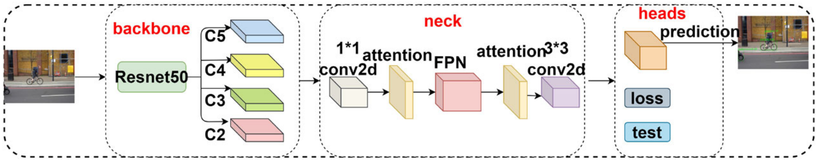

Fig. 1 Illustration of the proposed MFANet

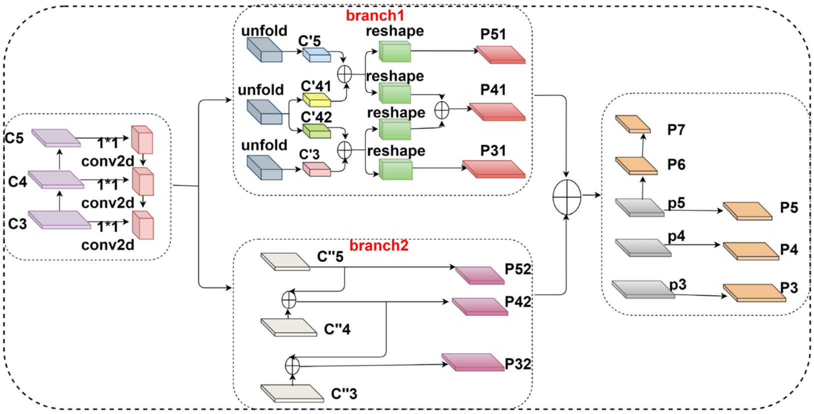

Fig. 2 The proposed MFPN module

$$
\begin{aligned}
p51 &= F_{RS}(c^{'}5 + c^{'}41) \\
p41 &= F_{RS}(c^{'}5 + c^{'}41) + F_{RS}(c^{'}3 + c^{'}42) \\
p31 &= F_{RS}(c^{'}3 + c^{'}42)
\end{aligned}
\tag{4}
$$

where F_{RS} is a reshape operator.

The branch 2 operation is to conduct feature fusion in the spatial dimension of the feature map. $[c^{''}3, c^{''}4, c^{''}5]$ is obtained by $1 * 1$ convolution. The output of branch2 is expressed by Eq. 5.

$$
\begin{aligned}
p52 &= c^{''}5 \\
p42 &= F_{US} * p52 + c^{''}4 \\
p32 &= F_{US} * p42 + c^{''}3
\end{aligned}
\tag{5}
$$

where F_{US} is an upsample operator.

Finally, the feature maps of the two branches are fused to get $[p3, p4, p5]$, and $[P3, P4, P5, P6, P7]$ are obtained after ablation and down-sampling.

3.2 MFA

The MFA is shown in Fig. 3. Let X denote the input feature map, its size is $[B, C, H, W]$, where B, C, H, W indicate the batch size, channel size, spatial height, and width, respectively.

It uses $1 * 1, 3 * 3, 5 * 5, 7 * 7$ convolutions to conduct convolution on X and obtain four tensors $[X_1, X_2, X_3, X_4]$ with different receptive fields. The sizes are all $[B, C, H, W]$, then $[X_1, X_2, X_3, X_4]$ are added to obtain X_5 .

It divides each tensor into two groups in the channel dimension. And the size of each group is $[B, C//2, H, W]$. And it uses two extract modules with different convolution kernel sizes to obtain the channel weights of each group. The convolution kernel sizes are [3,5], respectively. Then it concatenates the two groups in the channel dimension to obtain the weighs of each tensor.

The structure of the extract module is shown in Fig. 4. Let X_{CH} denote the input feature map, its size is $[B, C, H, W]$, where B, C, H, W indicate the batch size, channel size, spatial height, and width, respectively. It obtains $X_a, X_a \in R^{(B,C,1,1)}$ by global average pooling operation. To avoid the model being too complicated, it squeezes and permutes X_a , then obtains $X_s, X_s \in R^{(B,1,C)}$. After that, we use convolution kernel of k*k to realize the local cross-channel interaction to get $X_c, X_c \in R^{(B,1,C)}$. $X_{sg}, X_{sg} \in R^{(B,1,C)}$ is obtained by sigmoid activation function. Finally, it unsqueezes and permutes X_{sg} and then obtains X_{weight}, $X_{weight} \in R^{(B,C,1,1)}$.

The extract module is expressed by Eq. 6.

$$
F_{extractmodulek*k}(X_{CH}) = F_{un} F_{sg} W_{1d} F_s F_a X_{CH} \tag{6}
$$

where F_a is an adaptive avg-pool operator, F_{sg} is a sigmoid operator, $W_{1d}$is a k*k convolution layer, F_s is a compression and swap operator, and F_{un} is a decompression and swap operator. The output of $weigh5$ is expressed by Eq. 7.

$$
\begin{aligned}
X_{51}, X_{52} &= F_{SP} X_5 \\
weight5 &= concat(F_{extractmodulek*k}(X_{51}), \\
&\qquad F_{extractmodulek*k}(X_{52}))
\end{aligned}
\tag{7}
$$

where F_{SP} is a group operator and $concat$ is a splice operator. Finally, we fuse all channel weights and then multiply the $weight$ by X. And it gets the output after channel shuffle. The output of MFA is expressed by Eq. 8.

$$
X_{out} = F_{cs}\left(\sum_{i=1}^{5} weight(i) \odot X\right) \tag{8}
$$

where F_{cs} is channel shuffle operator and $\odot$ is a multiplication operator.

Channel shuffle operator is to integrate channels without increasing the amount of calculation. It is to expand $X, X \in R^{(B,C,H,W)}$ into $X_{cs}, X_{cs} \in R^{(B,G,C//G,H,W)}$

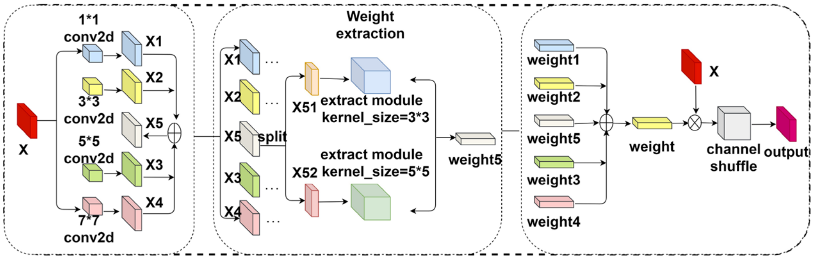

Fig. 3 The proposed MFA module

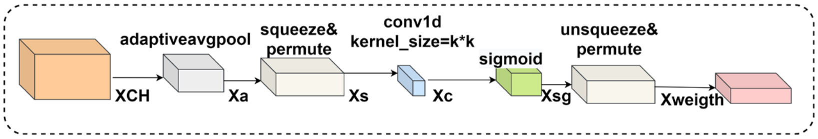

Fig. 4 The structure of extract module

and then reshapes X_{cs} to get X_{sc}, $X_{sc} \in R^{(B,C//G,G,H,W)}$. Finally, it is restored to X, $X \in R^{(B,C,H,W)}$ to achieve global channel information interaction.

4 Experiments

4.1 PASCAL VOC datasets

The PASCAL VOC 2007 and 2012 datasets are divided into four major categories: vehicle, household, animal, and person, and a total of 20 sub-categories (21 categories with background), respectively. PASCAL VOC 2007 object detection consists of 2501 training images, 2510 verification images, 5011 trainval images and 4952 test images. PASCAL VOC 2012 object detection consists of 5717 training images, 5823 verification images, 11540 trainval images and 11540 test images.

4.2 Ms CoCo2017 dataset

The Ms CoCo2017 dataset contains a total of 80 categories for detection. It is a large and rich object detection, segmentation and captioning dataset, which contains four files: annotations, test2017, train2017, and val2017. Among them, train2017 contains 118287 images, val2017 contains 5000 images, and test2017 contains 28660 images. Annotations are a collection of annotation types: object instances , object keypoints and image captions , which are stored in json files.

4.3 Experimental environment

CPU: Intel Xeon E5-2683 V3@2.00GHz; RAM: 32 GB; Graphics card: Nvidia GTX 1080Ti; Hard disk: 500GB.

It built a Python compilation environment with PyTorch1.6.0, torchvision = 0.7.0, CUDA10.0, and CUDNN7.4 as the deep learning framework, and implemented it on the platform mmdetection2.6.

4.4 Experimental strategy

It adjusts the size of all images to 512 × 512 for multi-scale training and uses data enhancement to perform various operations on the image dataset. Limited by experimental equipment, all algorithms use resnet50 as the backbone network. The SGD optimizer is adopted, the learning rate is 0.001, the momentum is 0.9, the weight decay is 0.0001, the learning rate adopts a step adjustment strategy, and the iteration period is 12 epochs.

For PASCAL VOC datasets, the evaluation standard of the experiment adopts mAP. For Ms CoCo2017 dataset, the evaluation standard of the experiment adopts average precision (Average-Precision, AP),AP_{50}, AP_{75}, AP_S, AP_M, AP_L as the main evaluation standards.

4.5 Ablation study

ATSS [22] points out that the essential difference between one-stage anchor-based and center-based anchor-free detectors is actually the definition of positive and negative training samples. However, whether the fusion of image features is sufficient or not directly affects the detection accuracy.

The neck of ATSS [22] adopts the feature pyramid network (FPN), which fuses deep feature maps to low-level feature maps through upsampling to obtain rich semantic features. We believe that the FPN structure is difficult to adequately fuse features in spatial, so it proposes MFPN. In order to reduce redundancy and enhance salient features, it proposes MFA. In this section, ablation experiments will be performed for the proposed method on the PASCAL VOC datasets and Ms CoCo2017 dataset. The 4.5.1 and 4.5.2 test the influence of MFPN and MFA on different networks.

4.5.1 MFPN experiments

In order to verify the effectiveness of the MFPN structure, we conduct ablation comparison experiments on 4 different networks. The experimental results are shown in Tables 1 and 2. Considering our experimental equipment and detec-

Table 1 The influence on Ms CoCo2017 dataset of MFPN on different networks

Model	Neck	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS [17]	FPN	0.291	0.461	0.304	0.104	0.322	0.455
FCOS [17]	MFPN	**0.300**	**0.477**	**0.314**	0.101	**0.336**	**0.468**
VFNet [19]	FPN	0.341	0.497	0.365	0.144	0.376	0.505
VFNet [19]	MFPN	**0.345**	**0.509**	**0.370**	0.140	**0.378**	**0.528**
Foveabox [34]	FPN	0.285	0.464	0.299	0.103	0.322	0.438
Foveabox [34]	MFPN	**0.309**	**0.486**	**0.324**	**0.117**	**0.347**	**0.478**
ATSS [22]	FPN	0.327	0.487	0.348	0.135	0.368	0.498
ATSS [22]	MFPN	**0.340**	**0.507**	**0.363**	0.134	**0.378**	**0.523**

Bold data indicate that the accuracy is improved compared to the original network

Table 2 The influence on PASCAL VOC datasets of MFPN on different

Class	ATSS [22]		Foveabox [34]		FCOS [17]		VFNet [19]	
	FPN	MFPN	FPN	MFPN	FPN	MFPN	FPN	MFPN
Aeroplane	79.8	**80.8**	79.8	**81.1**	77.9	**79.7**	80.0	**83.3**
Bicycle	82.6	**83.6**	81.9	**82.5**	77.1	**80.6**	82.1	**83.2**
Bird	79.6	**80.6**	74.4	**80.0**	77.6	**78.9**	79.3	**79.5**
Boat	68.5	**70.0**	65.4	**69.2**	67.3	**71.2**	70.6	**72.8**
Bottle	66.6	**67.3**	65.3	**66.0**	59.8	**62.6**	65.2	**66.3**
Bus	84.0	**84.8**	80.8	80.3	82.9	**84.8**	83.1	**83.2**
Car	86.1	**86.7**	85.4	**86.1**	83.7	**83.8**	86.4	**86.6**
Cat	88.4	88.3	83.5	**88.2**	86.2	**87.8**	88.5	88.1
Chair	62.4	62.0	60.0	**60.9**	58.5	**59.2**	60.7	**63.1**
Cow	82.9	**83.0**	72.4	**81.5**	79.5	**83.3**	82.8	**84.9**
Diningtable	73.2	73.0	69.7	68.5	63.1	**66.8**	69.1	**71.8**
Dog	87.7	**87.8**	81.5	**87.4**	84.4	**87.8**	87.1	87.0
Horse	85.4	**86.2**	80.0	**83.7**	75.5	**80.2**	85.9	85.8
Motorbike	80.9	**83.9**	80.3	**81.6**	76.5	**76.8**	81.3	**83.0**
Person	82.8	82.7	82.2	82.1	80.4	**80.8**	82.4	**82.6**
Pottedplant	51.3	**53.2**	51.8	**52.4**	53.2	52.5	52.3	**53.5**
Sheep	80.1	**81.7**	75.8	**79.8**	75.0	**78.0**	82.3	**82.6**
Sofa	75.2	**77.4**	72.5	68.5	69.7	**72.7**	76.4	76.1
Train	84.7	**86.5**	82.2	**84.7**	82.1	**84.8**	85.6	85.4
Tvmonitor	77.4	**78.6**	74.3	**77.5**	76.7	**77.4**	78.2	**79.4**
mAP/%	78.0	**78.9**	75.0	**77.1**	74.4	**76.5**	78.0	**78.9**

Bold data represent improved accuracy compared to the original network

tion accuracy, resnet50 is finally used. Resnet101 can better extract features. But it has more complex network and longer training time. And the performance requirements for GPU are also higher.

As Table 1 shows, the AP of ATSS has increased from 32.7% to 34%, AP_{50} and AP_L have even increased by 2%. The AP of FCOS has increased by 0.9% from 29.1%, and its other indicators can also be increased by more than 1%. Vfnet's AP increases by only 0.4% from 34.1%, but AP_L increases from 50.5 to 52.8%. The MFPN has the most obvious improvement in Foveabox. And its AP increases by 2.4%, and AP_L increases from 43.8 to 47.8%. FPN only fuses features of different sizes in space, and MFPN has more feature fusion in the channel dimension. So, the MFPN can obtain richer semantic features. And the accuracy of object detection will be higher.

MFPN has different effects on different networks. It has an AP increase of 0.4% on VFNet, and an AP_L increase of 2.4% on Foveabox. VFNet's original network AP is as high as 34.1%, while Foveabox's AP is only 28.5%. Four different networks use the same backbone, neck and different heads. The detection accuracy of ATSS network is lower than that of VFNet, indicating that the detection accuracy of ATSS heads is lower than that of VFNet heads. MFPN has limited improvement for small object, but it has a significant improvement in the detection of medium-sized object and large object. In the field of object detection, in order to improve the detection accuracy of small objects, a larger size feature map is required.

For PASCAL VOC datasets, as Table 2 shows, ATSS has increased from 78 to 78.9% in the mAP. Foveabox has increased by 2.1% from 75. FCOS has increased from 74.4 to 76.5%. It has mAP increase of 0.9% on VFNet. The MFPN can greatly improve the detection accuracy of most categories on different networks. There are some categories, such as: "cat," "chair," "dog," "horse." Their accuracy has declined. That is because these categories are relatively few in training and are taken as part of pictures rather than as a whole.

4.5.2 MFA experiment

In order to further study the impact of MFA on detection accuracy. We perform MFA ablation comparison experiments on 4 different networks. The experimental results are shown in Tables 3 and 4.

As Table 3 shows, the AP of ATSS increases by 1%, AP_S increases from 13.5 to 14.3%, and AP_L increases by 2.2%. All indicators of FCOS have increased by an average of 1%. VFNet increases by 0.4% AP, but AP_L increases by 2.5%. MFA has the most obvious effect on Foveabox. And its AP

Table 3 The influence on Ms CoCo2017 dataset of MFA on different networks

Model	Attention	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS [17]	×	0.291	0.461	0.304	0.104	0.322	0.455
FCOS [17]	√	**0.300**	**0.479**	**0.311**	**0.106**	**0.338**	**0.468**
VFNet [19]	×	0.341	0.497	0.365	0.144	0.376	0.505
VFNet [19]	√	**0.345**	**0.513**	**0.368**	0.140	**0.381**	**0.530**
Foveabox [34]	×	0.285	0.464	0.299	0.103	0.322	0.438
Foveabox [34]	√	**0.311**	**0.494**	**0.323**	**0.120**	**0.352**	**0.473**
ATSS [22]	×	0.327	0.487	0.348	0.135	0.368	0.498
ATSS [22]	√	**0.337**	**0.508**	**0.360**	**0.143**	**0.377**	**0.520**

× Indicates that there is no attention mechanism. √ indicates that there is a MFA. Bold data indicate improved accuracy compared to comparison networks

Table 4 The influence on PASCAL VOC datasets of MFA on different networks

Class	ATSS [22]		Foveabox [34]		FCOS [17]		VFNet [19]	
	×	√	×	√	×	√	×	√
Aeroplane	79.8	**80.2**	79.8	78.2	77.9	**79.5**	80.0	**82.2**
Bicycle	82.6	**83.1**	81.9	**83.8**	77.1	**81.3**	82.1	**82.3**
Bird	79.6	**80.2**	74.4	**78.5**	77.6	**78.9**	79.3	**80.1**
Boat	68.5	**72.2**	65.4	**70.7**	67.3	**69.4**	70.6	**71.9**
Bottle	66.6	66.3	65.3	**65.8**	59.8	**61.4**	65.2	**66.2**
Bus	84.0	**84.6**	80.8	**82.3**	82.9	**84.8**	83.1	**85.1**
Car	86.1	**86.7**	85.4	85.1	83.7	**84.7**	86.4	86.2
Cat	88.4	88.2	83.5	**88.7**	86.2	**87.9**	88.5	88.4
Chair	62.4	**64.0**	60.0	**61.7**	58.5	**59.1**	60.7	**62.2**
Cow	82.9	82.8	72.4	**81.7**	79.5	**82.5**	82.8	**85.1**
Diningtable	73.2	69.4	69.7	68.1	63.1	**66.5**	69.1	69.0
Dog	87.7	87.0	81.5	**87.6**	84.4	**87.4**	87.1	**87.2**
Horse	85.4	**86.1**	80.0	**84.2**	75.5	**81.3**	85.9	85.3
Motorbike	80.9	**83.1**	80.3	**81.9**	76.5	**80.1**	81.3	**82.3**
Person	82.8	**83.3**	82.2	**82.5**	80.4	**80.7**	82.4	**82.8**
Pottedplant	51.3	**51.8**	51.8	**54.7**	53.2	51.6	52.3	**54.5**
Sheep	80.1	**81.5**	75.8	**79.4**	75.0	**78.7**	82.3	**82.5**
Sofa	75.2	**76.9**	72.5	70.9	69.7	**70.0**	76.4	74.4
Train	84.7	84.3	82.2	82.1	82.1	**85.9**	85.6	**86.3**
Tvmonitor	77.4	**78.1**	74.3	**76.7**	76.7	**77.0**	78.2	**78.8**
mAP/%	78.0	**78.5**	75.0	**77.2**	74.4	**76.4**	78.0	**78.6**

× Indicates that there is no attention mechanism. √ Indicates that there is a MFA. Bold data indicate that the accuracy is improved compared to the comparison network

increases from 28.5 to 31.1%, and AP_L increases from 43.8 to 47.3%.

It can be seen from Table 3 that MFA improves AP_{50} more significantly than AP. MFA improves AP_S by an average of nearly 1–2 %, and AP_L can increase by more than 2%. The feature maps of different receptive fields have different effects on object detection of different sizes. The MFA structure integrates the feature maps of 4 different receptive fields, so it can effectively balance object of different sizes. And its extraction of different channel weights can also enhance important features and reduce redundancy, which shows that it is effective for MFA to use feature maps of different receptive fields.

As Table 4 shows, The mAP of ATSS has increased from 78 to 78.5%. Foveabox has increased by 2.2% from 75%. FCOS has increased from 74.4 to 76.4%. It has a mAP increase of 0.6% on VFNet. The MFA can greatly improve the detection accuracy of most categories on different networks. Although there are also some categories that have declined, such as: “cat,” “cow,” “diningtable” and so on. But it is not obvious, it can even be considered as experimental error.

Feature visualization operations are also performed on Ms CoCo2017 dataset. In Fig. 5, Column (a) is the input image. Columns (b) and (d) are the heat maps of the original network and network with MFA, respectively. Columns (c) and (e) are the superimposed effect diagrams of the heat map and the input.

From the column(b) and column(d), it is obvious that without the MFA, the network's attention to pictures is scattered. The feature weights of the objects extracted from the original backbone network are not high. When adding the MFA, the network's attention is focused on the object. The context information in the feature extraction will be aggregated, and important information will be given higher weight (such as the bright spot in Fig. 5). And it is not difficult to find that the attention algorithm can make the framework pay more attention to the area of interest.

4.6 Compare with classic networks

For sub-modules, their outputs are different, but they all perform better than baseline ATSS (Resnet50). From Table 5, it can be seen that the baseline output is only 32.7% AP. There is an increase of 1.0% AP in MFA module and 1.3% AP in MFPN module. After feature fusion, the superposition of the two modules, that is, the output of our model can reach 34.2% AP. Although the AP_L has decreased, the AP_S has increased by up to 2%.

We compare the proposed network with other classic networks. From Table 6, the proposed network is the highest of AP, AP_{50} and AP_S. AP_{75} is 36.6%, second only to 37% in all networks. AP_M is 38.2%, which is only 0.1% lower than the highest 38.3%. Although AP_L is 50.3%, AP_S has improved significantly. And it can effectively balance the detection effect of the network on objects of different sizes.

It compares the detection effect with classic networks. As can be seen from Fig. 6, the loss of SSD information is obvious. In the first image, it does not detect the puppy, and in the fourth image it does not detect the cup. Although Faster-RCNN can detect objects, its false detection is very high. In the first image, the tie is falsely detected many times. In the third image, the front wheel of the motorcycle is falsely detected as a car. In the fourth image, the laptop is falsely detected as tv. The detection effect of ATSS on the second, third and fourth images is ok, but the detection of the first image obviously misses the dog and tie. Although PAA has high detection accuracy, its detection frame redundancy is also high. The proposed method can not only accurately detect objects in images, but also has low missed detection and redundancy rates.

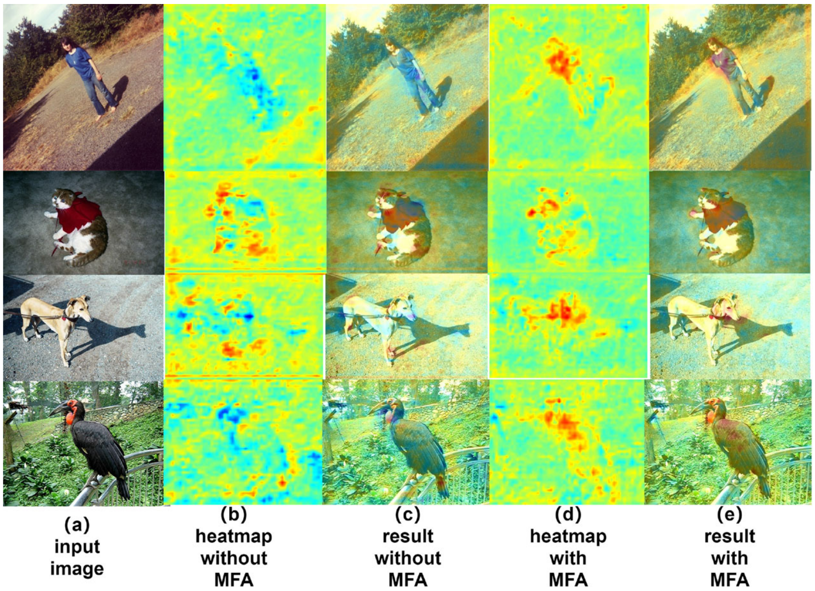

Fig. 5 Visualization on Ms CoCo2017 dataset

Table 5 The effect of different modules on the network

Model	Module	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ATSS [22]	FPN	0.327	0.487	0.348	0.135	0.368	0.498
	FPN+MFA	0.337	**0.508**	0.360	0.143	0.377	0.520
	MFPN	0.340	0.507	0.363	0.134	0.378	**0.523**
	MFPN+MFA	**0.342**	0.506	**0.366**	**0.161**	**0.382**	0.503

Bold data indicates the highest precision of the data in the table

Table 6 Comparison of the proposed method with other classic networks

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster-rcnn [10]	0.283	0.450	0.304	0.120	0.310	0.417
FCOS [17]	0.291	0.461	0.304	0.104	0.322	0.455
SSD [15]	0.256	0.440	0.262	0.091	0.290	0.389
Retinanet [16]	0.305	0.473	0.322	0.108	0.352	0.484
Yolov3 [35]	0.246	0.426	0.251	0.082	0.268	0.365
Foveabox [34]	0.285	0.464	0.299	0.103	0.322	0.438
ATSS [22]	0.327	0.487	0.348	0.135	0.368	0.498
PAA [20]	0.342	0.498	**0.370**	0.143	**0.383**	**0.524**
VFNet [19]	0.341	0.497	0.365	0.144	0.376	0.505
Reppoints [18]	0.291	0.472	0.302	0.098	0.372	0.461
The proposed method	**0.342**	**0.506**	0.366	**0.161**	0.382	0.503

Bold data indicates the highest accuracy compared to the comparison network

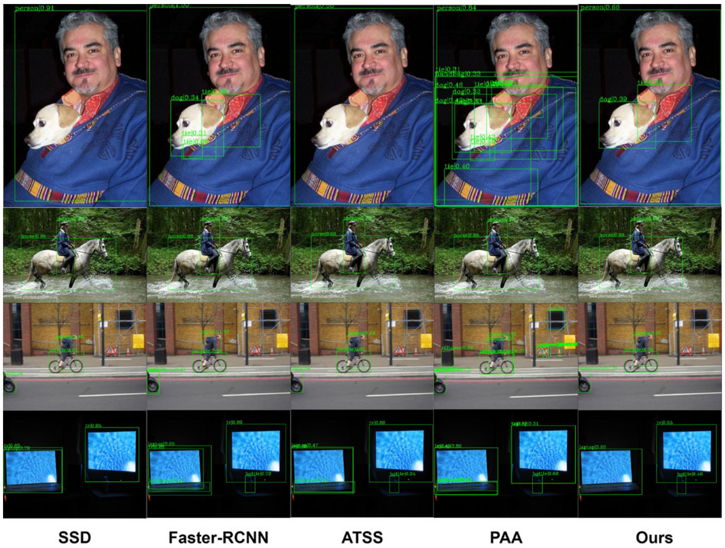

Fig. 6 visual comparison of different networks. All images have the confidence threshold set to 0.3

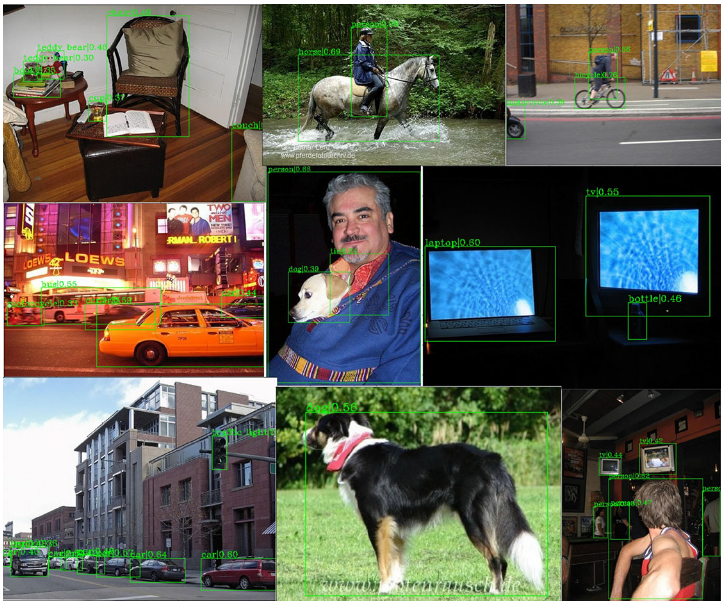

Fig. 7 Qualitative results of MFANet. This model achieves 34.2% in AP. All images have the confidence threshold set to 0.3

It tests the detection effect of the proposed method. As can be seen from Fig. 7, only a dog in the picture, the network can accurately detect the object. When multiple objects in the picture, it can also separate different objects well, such as a person riding a horse. In road traffic scenes, it can detect dense vehicles and traffic lights. In dimly lit scenes, it can also detect cup. In an incomplete picture, it can detect a motorcycle based on a wheel. It is not difficult to see that the proposed method has completed the task of accurate object detection and has an excellent identification effect at the edge.

5 Conclusion

In this paper, we propose the MFANet. The core modules of the network are as follows: multi-scale feature fusion and attention mechanism modules. The feature maps of different sizes are effectively fused in the two dimensions of space and channel. And it realizes channel attention learning of local cross-channel interaction without dimensionality reduction.

Based on the same configuration and platform, it verifies the excellent performance of the proposed algorithm. Under the premise of the same configuration, our algorithm improves 1.5% AP, 2.9% AP_{50}, 1.8% AP_{75}, 2.6% AP_S, 1.4% AP_M and 0.5% AP_L, respectively. In future work, we will investigate how feature fusion differs in channel dimension and spatial dimension. We will also explore their respective effects on the detection accuracy of objects of different sizes.

Author Contributions Gaihua Wang, Xin Gan, Qing Caocheng and Qianyu Zhai conceived the experiments. Xin Gan and Qingcheng Cao conducted the experiments. All authors reviewed the manuscript.

Funding This work is supported in part by the National Key R&D Program of China under Grant 2017YFB1302400.

Declarations

Conflict of interest This article has no conflict of interest with any individual or organization.

Code or data availability Code and data are available.

Ethics approval The experiments in this article are all realized through program operation, which will not cause harm to humans and animals and will not cause moral and ethical problems.

Consent to participate Welcome readers to communicate.

Consent for publication Completed at Hubei University of Technology on December 14, 2021.

References

1. Sugiura, M., Miyauchi, C. M., Kotozaki, Y.: Neural mechanism for mirrored self-face recognition. Cereb. Cortex **25**(9), 2806–14 (2015)
2. Boulgourisa, N.V., Plataniotis, K., Hatzinakos, D.: Gait recognition using linear time normalization. Pattern Recogn. **39**(5), 969–979 (2006)
3. Mei, J., Zhou, D., Cao, J., et al.: HDINet: hierarchical dual-sensor interaction Network for RGBT tracking. IEEE Sens. J. **21**(15), 16915–16926 (2021). https://doi.org/10.1109/JSEN.2021.3078455
4. Chaudhry, H., Rahim, M. S. M., Saba, T.: Crowd detection and counting using a static and dynamic platform: state of the art. Int. J. Comput. Vis. Robot. **9**(3), 228–59 (2009)
5. Cerezo, E., Pérez, F., Pueyo, X.: A survey on participating media rendering techniques. Vis. Comput. **21**(5), 303–328 (2005)
6. Wang, G., Zhai, Q.: Feature fusion network based on strip pooling. Sci. Rep. **11**(1), 1–8 (2021)
7. Verschae, R., Ruiz-del-Solar, J.: Object detection: current and future directions. Front. Robot. AI **2**, 29 (2005)
8. Xiao, Y., Tian, Z., Yu, J.: A review of object detection based on deep learning. Multimed. Tools Appl. **79**(33/34), 23729–91 (2020)
9. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r* cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1080–1088 (2015)
10. Ren, S., He, K., Girshick, R.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–49 (2017)
11. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
12. Cao, J., Cholakkal, H., Anwer, R.M., Khan, F.S., Pang, Y., Shao, L.: D2det: towards high quality object detection and instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11485–11494 (2020)
13. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Luo, P.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14454–14463 (2021)
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham (2016)
16. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636 (2019)
17. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: Point set representation for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9657–9666 (2019)
18. Lin, T.-Y., Goyal, P., Girshick, R.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–27 (2020)
19. Kim, K., Lee, H.S.: Probabilistic anchor assignment with IOU prediction for object detection. In: European Conference on Computer Vision, pp. 355–371. Springer, Cham (2020)

20. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: An IOU-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8514–8523 (2021)
21. Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., Sun, J.: You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13039–13048 (2021)
22. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9759–9768 (2020)
23. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125(2017)
24. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
25. Ghiasi, G., Lin, T.Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7036–7045(2019)
26. Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C.: Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12595–12604 (2020)
27. Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)
28. Qiao, S., Chen, L.C., Yuille, A.: Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 10213–10224 (2021)
29. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
30. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
31. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
32. Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Pang, Y.: Attention scaling for crowd counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4706–4715 (2020)
33. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)
34. Kong, T., Sun, F., Liu, H.: FoveaBox: beyound anchor-based object detection. IEEE Trans. Image Process. **29**, 7389–98 (2020)
35. Li, D., Huang, C., Liu, Y.: YOLOv3 target detection algorithm based on channel attention mechanism. In: 2021 3rd International Conference on Natural Language Processing (ICNLP), pp. 179–183. IEEE (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Gaihua Wang is an associate professor in the School of Electrical and Electronic Engineering, Hubei University of Technology. She is responsible for his work and has a strong sense of innovation. She has made some achievements in teaching and scientific research. She has presided over one project of National Natural Science Foundation of China, two projects of Hubei Provincial Education Department (all of which have been concluded), and participated in a number of national projects. She has published more than 10 papers in EI and above journals, including 2 SCI articles. In the teaching work, he has won many awards such as Outstanding Instructor, Class Tutor Model, and Outstanding Annual Assessment. She guides graduate students to participate in provincial competitions and undergraduate students to carry out undergraduate innovation and entrepreneurship projects.

Xin Gan received the B.E. degree in JiMei University, Fujian, China, in 2019. He is currently pursuing the M.S. degree with Hubei University of Technology, Wuhan, China. His current research interests include deep learning and object detection. The author's honors include the second prize of the 16th Graduate Electronic Design Competition in Central China Business Competition and the third prize of the Technology Competition.

Qingcheng Cao received the B.E. degree in Qingdao University of Technology, Shandong, China, in 2020. He is currently pursuing the M.S. degree with Hubei University of Technology, Wuhan, China. His current research interests include deep learning and object detection. The author's honors include the second prize of the 16th Graduate Electronic Design Competition in Central China Business Competition and the third prize of the Technology Competition.

Qianyu Zhai received the B.E. degree in Hubei University of Technology Engineering and Technology College, Wuhan, China, in 2020. He is currently pursuing the M.S. degree with Hubei University of Technology, Wuhan, China. His current research interests include deep learning and image segmentation. The author's honors include the second prize of the 16th Graduate Electronic Design Competition in Central China Business Competition and the third prize of the Technology Competition.