Hindawi
Scientific Programming
Volume 2023, Article ID 5500078, 12 pages
https://doi.org/10.1155/2023/5500078

Research Article

A Small Object Detection Network Based on Multiple Feature Enhancement and Feature Fusion

Kun Tan,[1] Shengduo Ding,[1] Shuncheng Wu,[1] Kun Tian,[1] and Jie Ren[2]

[1]*CNPC Research Institute of Safety and Environment Technology, Beijing 100000, China*
[2]*College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China*

Correspondence should be addressed to Shengduo Ding; dingshengduo@cnpc.com.cn

Received 11 August 2022; Revised 15 February 2023; Accepted 28 April 2023; Published 26 May 2023

Academic Editor: Antonio J. Peña

Due to the small size, high resolution, and complex background, small object detection has become a difficult point in computer vision. Making full use of high-resolution features and reducing information loss in the process of information propagation is of great significance to improve small object detection. In this article, to achieve the above two points, this work proposes a small object detection network based on multiple feature enhancement and feature fusion based on RetinaNet (MFEFNet). First, this work designs a densely connected dilated convolutions to adequately extract high-resolution features from C_2. Then, this work utilizes subpixel convolution to avoid the loss of channel information caused by channel dimension reduction in the lateral connection. Finally, this article introduces a bidirectional fusion feature pyramid structure to shorten the propagation path of high-resolution features and reduce the loss of high-resolution features. Experiments show that our proposed MFEFNet achieves stable performance gains in object detection task. Specifically, the improved method improves RetinaNet from 34.4AP to 36.2AP on the challenging MS COCO dataset, and especially achieves excellent results in small object detection with an improvement of 2.9%.

1. Introduction

As a fundamental problem in the field of computer vision, object detection is the basis for many tasks such as image segmentation, object tracking, and image description. With the development of the convolutional neural network [1], many one-stage detectors [2–5] and two-stage detectors [6–9] with remarkable performance have been developed in recent years. Two-stage object detection algorithms are developing rapidly, such as R-CNN [2], Faster R-CNN [4], and Mask R-CNN [5]; the detection accuracy is constantly improving, but the problem of their own architecture limits the detection speed. The one-stage target detection algorithm was proposed later than the two-stage target detection algorithm, due to its relatively simple structure and superior detection speed, it has also attracted the attention of many researchers. The representative algorithms include YOLO and its variants [6–9], SSD and its variants [10–12], RetinaNet [13], and EfficientDet [14].

Although the one-stage object detector is significantly faster than the two-stage object detector, its accuracy has not been comparable to the two-stage object detector. Some one-stage object detection algorithms have improved the detection effect by introducing two-stage object detection algorithms such as Feature Pyramid Network (FPN) [15] and changing the backbone network. FSSD [12] reconstructs the pyramid feature map to fuse features of different scales, which is beneficial to small object detection. EfficientDet [14] uses weighted bidirectional feature pyramid network for feature fusion and scales the model through composite feature pyramid network. Lin et al. believed that the real reason for the low accuracy of the integrated convolutional neural network was the mismatch between the target and background levels in the image, and then proposed RetinaNet [13]. RetinaNet solves the sample imbalance problem by introducing focal loss, which greatly improves the object detection effect. However, the detection effect of small

objects (objects below 32 pixels × 32 pixels [16]) is not competitive with the two-stage target detection algorithm.

We analyzed the RetinaNet [13] and found that the following three points are not conducive to small object detection. Firstly, it does not fully utilize the shallow feature layer C_2. Due to the low resolution and little visual information of small targets, the information of small targets may be lost in the process of network up-sampling, making it difficult for deep feature layers to extract discriminative features. However, the shallow feature layer C_2 has a smaller receptive field, higher spatial resolution, and contains more accurate position information, which are very beneficial for small object detection. Meanwhile, in the FPN-based methods, the network generally uses a simple convolution method to extract the shallow feature C_2, such as [5, 17]. Due to the limitation of the size of the receptive field and the depth of the network, it is difficult to extract shallow features more fully.

Secondly, to reduce the computation, FPN-based methods adopt 1×1 convolutional layers to reduce channel dimensions of the output feature maps C_i from the backbone. C_i generally extract thousands of channels in high-level feature maps. Especially, the high-level features C_4 and C_5 have large channel dimensions, which contain rich semantic information that is beneficial to object detection. The drastic channel dimension reduction (e.g., 2048 to 256) results in the loss of a large amount of channel information, which has a negative impact on small target detection. The existing methods [14, 18] to extract the channel information mainly extract channel-reduced maps by adding additional modules, and act on fewer channel features through more complex network connections to achieve better accuracy. Although [19] makes full use of C_i, it does not fully mine contextual information of the transformed features.

Finally, RetinaNet [13] introduces the top-down feature pyramid structure and performs multiscale feature fusion to improve the detection effect of targets of different scales. It is worth noting that the low-level features are critical for the detection of small objects, which are helpful for more accurate localization. However, due to the limitation of the FPN structure, the path between high-level features and low-level features is long (tens or even hundreds of network layers such as ResNet50 and ResNet101), resulting in less low-level features at the top of the pyramid, which makes the small object detection effect not as good as expected.

Combined with the above analysis and inspired by BFE-Net [20], we believe that improving the utilization of high-resolution features in the network and reduce the loss of features in the process of network propagation is of great significance to improve the effect of small object detection. For one thing, to improve the utilization of high-resolution features, we reuse the shallow features C_2. Inspired by Densenet [21], we designed the multiscale context extraction module to fully extract shallow features. To pursue the balance between accuracy and computational load, this work uses the dense connection mechanism combined with dilated convolution to effectively expand the receptive field and increase the depth of the feature extraction network to some extent, which can extract richer semantic features and location features while effectively realizing feature reuse.

For another thing, to reduce the information loss, this work utilizes subpixel convolution and bidirectional feature pyramid structure. First, inspired by [19, 22], this work designs a subpixel convolution enhancement module to reduce the information loss caused by channel reduction. Specifically, this work uses subpixel convolution to convert low-resolution feature maps into high-resolution feature maps in the horizontal connection of top-down propagation, making full use of channel information and reducing the loss of information during lateral connection. At the same time, the spatial attention mechanism is used for the transformed features to obtain richer contextual information. Second, to reduce the loss of shallow information in the propagation path and inspired by PANet [17], this work introduces a bidirectional fusion feature pyramid structure. We designed a bidirectionally connected feature pyramid structure, which can greatly shorten the propagation path of shallow features to reduce feature loss and better retain shallow feature information. At the same time, the bidirectional feature pyramid network further strengthens the multiscale feature fusion, which greatly enriches the shallow multiscale context information.

Based on the above analysis and strategies, the detection method proposed in this article compared to standard RetinaNet has the following advantages:

(1) To improve the utilization of shallow features, this article designs a multiscale context extraction module (MCEM) consisting of densely connected dilated convolutions, which use convolutional layers with different dilation rates to obtain more effective receptive fields.

(2) To make full use of channel information in the lateral connection and reduce the channel information loss, this article designs a subpixel convolution enhancement module (SCEM), which uses subpixel convolution to convert low-resolution features into high-resolution features to avoid information loss caused by channel dimension reduction in the lateral connection.

(3) To reduce the low-level features loss in propagation process, this article designs a bidirectional fusion feature pyramid structure (BidiFPN), which uses bidirectional feature pyramid structure to shorten the propagation path of shallow features, reducing the shallow feature loss in the propagation process.

2. Related Work

2.1. Object Detectors. At present, there are two types of mainstream deep learning target detection algorithms, two-stage target detection based on region proposal and one-stage target detection based on regression analysis.

2.1.1. Two-stage Detectors. The two-stage target detection algorithm generally uses selective search or region proposal

network to extract candidate frames from the image, and then performs secondary correction on the candidate frame target to obtain the detection result. R-CNN [2] introduces convolutional neural network combined with candidate region proposal to achieve target detection. SPP-Net takes the entire image as input, and realizes feature extraction of any scale area, reducing the amount of computation. Faster R-CNN [4] proposes a region proposal network to extract candidate regions, which improves detection efficiency. Mask R-CNN [5] uses the RoI Align layer to reduce the deviation of the feature map from the original map. Cascade R-CNN [23] introduced multilevel refinement in Faster R-CNN to achieve more accurate target location prediction. The two-stage target detection algorithm is developing rapidly, and the detection accuracy is constantly improving, but the problem of its own architecture limits the detection speed. It cannot meet some downstream tasks with strong real-time performance.

2.1.2. One-stage Detectors. Compared with two-stage detectors, one-stage object detection algorithms do not require classification on candidate regions, and the training process is relatively simple. YOLOv1 [6] is the first one-stage detector in the field of deep learning proposed by Redmon et al., whose biggest advantage is the fast speed. Some scholars have improved on the basis of YOLOv1 [16] and proposed YOLO9000 [7], YOLOv3 [8], and YOLOv4 [9]. SSD [10] proposed in 2015 combines the advantages of YOLO's fast detection speed and Faster R-CNN's accurate positioning. DSSD [11] backbone adopts Resnet-101 and adds deconvolution module to improve the effect of small object detection. FSSD [12] reconstructs the pyramid feature map to fuse features of different scales to enhance the detection effect of small objects. Although the one-stage object detector is significantly faster than the two-stage object detector based on candidate region recommendation, its accuracy has not been comparable to the two-stage object detector. RetinaNet [13] solves the problem of instance sample imbalance by introducing focal loss and realizes a detection framework whose accuracy is comparable to that of two-stage target detectors. However, RetinaNet [13] detection effect on small objects still has room for improvement compared to two-stage target detection algorithms. In addition, EfficientDet [14] uses a weighted bidirectional feature pyramid network for feature fusion. YOLOF [24] designs a dilated encoder and a balanced matching strategy to improve the detection performance.

2.2. Feature Augmentation. As the number of network layers increases, the semantic information and location information of the target are lost layer by layer. Multiscale feature fusion and contextual feature enhancement are effective methods to compensate for information loss.

2.2.1. Multiscale Feature Fusion. To make full use of the features extracted by different feature layers, many researchers optimize the detector architecture to achieve multiscale feature fusion. Most detectors utilize the FPN [15] to detect objects of different sizes, which extracts the features from the bottom to the top, and then performs a top-down feature fusion structure, and finally sends them to the prediction module to output the results. PANet [17] connects the features of the lowest layer of the model with the features of the highest layer, shortens the information path between the top layer and the bottom layer, and further strengthens the connection between the feature maps of each layer. EfficientDet [14] proposes a weighted bidirectional feature pyramid network BiFPN to achieve more efficient multiscale feature fusion. AugFPN [25] utilizes consistency supervision to close the semantic gap before feature fusion and employs residual features to reduce information loss during convolution pooling to better utilize multiscale features. NAS-FPN [26] makes full use of neural network search technology to achieve cross-scale feature fusion through top-down and bottom-up connections. Inspired by [22], Luo et al. used the original channel information for cross-scale output and proposed CE-FPN [19].

2.2.2. Context Feature Enhancement. The detected target has an inseparable relationship with other surrounding objects and the environment. In order to improve detection accuracy by exploring contextual information, CoupleNet [27] improves the detection accuracy by introducing the global and semantic information of the proposal and combining local information and global information. The DetectoRS [28] proposes Recursive Feature Pyramid (RFP) and incorporates additional feedback connections from the feature pyramid network to the bottom-up backbone layers. Lim et al., [29] improved the detection accuracy of small objects by fusing multiscale features and using additional features at different levels as contextual information. Nonlocal [30] proposed a strategy to obtain the dependencies between two locations, solving the problem of limited receptive field obtained by convolution operation at each layer.

3. Methods

This section introduces the small object detection network based on multiple feature enhancement to reduce the loss of high-resolution information and make up for the loss of information during the propagation process and lateral connection. As shown in Figure 1, three components are proposed in MFEFNet: multiscale context extraction module (MCEM), subpixel convolution enhancement module (SCEM), and bidirectional fusion feature pyramid structure (BidiFPN). We have described them in detail as follows.

3.1. Multiscale Context Extraction Module. Small objects have fewer pixels available than normal-sized objects, and features are difficult to extract. With the deepening of the number of network layers, through continuous downsampling and feature extraction, the feature information and location information of small objects are also lost layer by layer. The shallow target of convolutional neural network contains much small object information due to its small

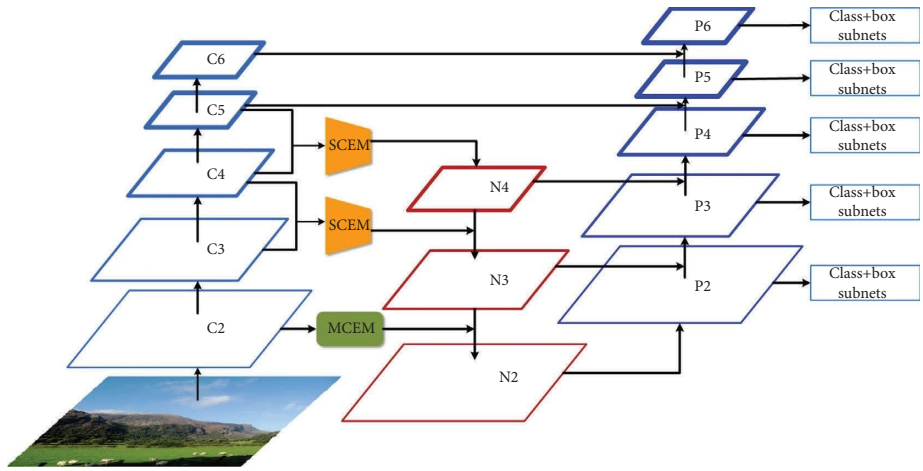


Figure 1: The overall architecture of MFEFNet.

receptive field, high resolution, and rich location information. Therefore, making full use of the shallow feature layer can improve the small object detection effect to a certain extent. RetinaNet [13] does not use the high-resolution pyramid level P_2. We designed the multiscale context extraction module (as shown in Figure 2) to fully extract the features of the high-resolution feature layer C_2 through densely connected dilated convolutions.

Although the shallow feature layer of the convolutional neural network contains rich small object information, its ability to express feature semantic information is weak. Inspired by [21], we perform feature extraction through the dilated convolutional layer with different dilation rates, which enriches semantic information while ensuring rich spatial information, and enhances the high-level semantic information of shallow features.

First, we divide the feature map C_2 into three branches for dilated convolution. Since each dilated convolutional layer has a different dilation rate, three feature maps with different receptive field sizes will be obtained.

$$\begin{cases} F_0 = F_d(C_2, 3), \\ F_1 = F_d(C_2 \oplus F_0, 5), \\ F_2 = F_d(C_2 \oplus F_0 \oplus F_1, 9), \end{cases} \tag{1}$$

where $F_d(\cdot)$ represents the dilated convolution operation function with a convolution kernel of 3×3 and expansion rates of 3, 5, and 9, respectively. The symbol $\oplus$ denotes feature fusion by addition. Then, the three output feature maps containing multiscale context information and C_2 after 1×1 convolution are fused in the concatenate method and then D_2 is obtained through 1×1 convolution layer for channel dimension reduction.

$$D_2 = \mathrm{Conv}_{1\times1}(F_{\mathrm{concat}}(F_0, F_1, F_2, \mathrm{Conv}_{1\times1}(C_2))), \tag{2}$$

where $F_{\mathrm{concat}}(\cdot)$ represents the operation of feature connection in the way of concatenate.

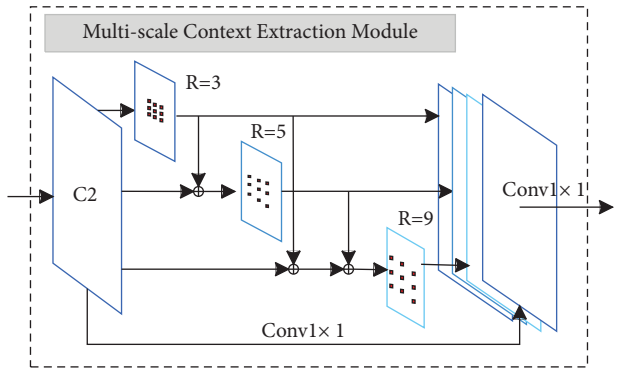


Figure 2: Illustration of multiscale context extraction module (MCEM).

3.2. Subpixel Convolution Enhancement Module. As the number of convolutional layers increases, the network can obtain more effective features. In the RetinaNet [13], with the deepening of the backbone network, feature layers with rich dimensions will be generated in the bottom-up propagation path, especially the high-level features C_4 and C_5, and the feature dimensions are 1024 and 2048, respectively. These high-level features are rich in semantic information. However, in order to reduce the complexity of the network and improve the calculation speed, a 1×1 convolutional layer will be used for dimension reduction in the lateral connection. For example, the dimension of C_5 is reduced from 2048 to 256. The dramatic reduction in dimension will lead to a lot of semantics loss of information.

The loss of semantic information in the top-down propagation process will further affect the detection results, especially the loss of small object features becomes more and more serious. To reduce the loss of semantic information in the lateral connection and make full use of the rich channel information of high-level feature maps, we are inspired by [19] to use subpixel convolution to achieve channel dimension reduction and fully fuse the information

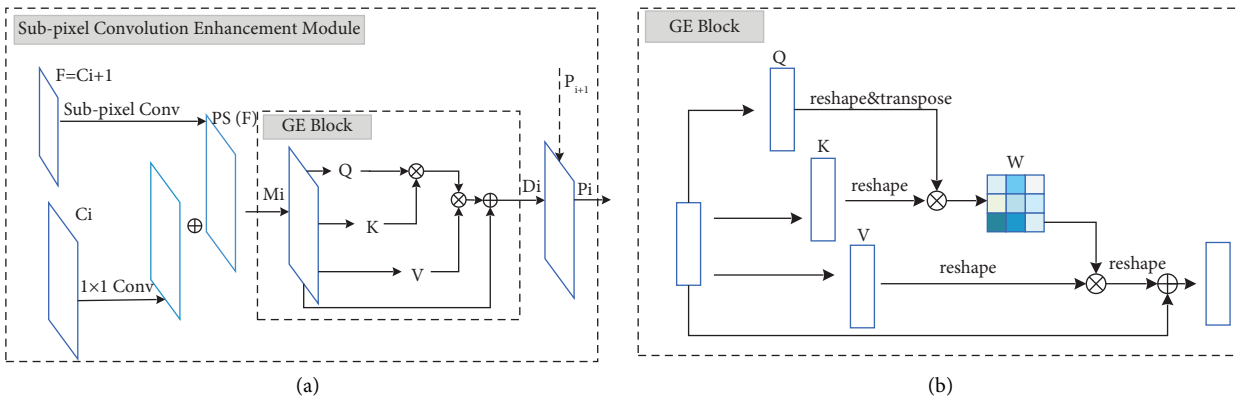


Figure 3: Illustration of subpixel convolution enhancement module (SCEM). (a) The overall of SCEM. (b) The details of GE block.

of adjacent feature layers and designed the subpixel convolution enhancement module (as shown in Figure 3(a)). Subpixel convolution [22] implements the reconstruction process from up-sampling reconstruction from low-resolution images to high-resolution images. This operation is to rearrange the pixels on different channels of the feature map into the same channel space, so as to achieve the purpose of more pixels in the same channel space, mainly by transforming the channel size to increase the width and height. Considering that C_4 and C_5 have 1024 and 2048 channels, respectively, subpixel convolution is performed directly without expanding the channel size. The pixel shuffle operator rearranges the feature of shape $\mathrm{H} \times \mathrm{W} \times \mathrm{C} \cdot \mathrm{r}^2$ to $\mathrm{rH} \times r\mathrm{W} \times \mathrm{C}$, which can be formulated as follows:

$$\mathrm{PS}(F)_{x,y,c} = F_{[x/r],[y/r],C\cdot r\cdot \mathrm{mod}(y,r)+C\cdot \mathrm{mod}(x,r)+c}, \tag{3}$$

where r denotes the up-scaling factor, in this work, $r = 2$. F is the input feature, and F is C_{i+1} in this article as shown in Figure 3(a), and $\mathrm{PS}(F)_{x,y,c}$ denotes the output feature pixel on coordinates x, y, and c. The index x, y, and c start from 0, which represents the coordinates in the high-resolution feature map. M_i is the output obtained by element-wise addition of the low-resolution feature map C_{i+1} and the high-resolution feature map C_i after subpixel convolution.

$$\begin{aligned} M_i &= \mathrm{PS}\left(C_{i+1}\right)_{x,y,c} \bigoplus \mathrm{Conv}_{1\times 1}\left(C_i\right), i = 3, 4, \\ D_i &= \mathrm{GE}\left(M_i\right), i = 3, 4, \end{aligned} \tag{4}$$

where the symbol $\oplus$ denotes feature fusion by addition. $\mathrm{Conv}_{1\times 1}(\cdot)$ represents a 1×1 convolution layer for channel dimension reduction. $\mathrm{GE}(\cdot)$ represents the processing process of GE block.

The standard RetinaNet [13] introduces the feature pyramid network to detect objects of different scales through multiscale representation, enriching the semantic information of shallow features to make it more effective for small objects detection. However, the convolutional neural network can only obtain the local receptive field. Although the receptive field can be expanded through deeper network layers, the global information cannot be obtained. Context information means that in an image, a single pixel or a single target does not exist alone, but has some relationship with the surrounding pixels and targets. Mining and utilizing the contextual information between objects will be beneficial to object detection, especially for small objects that rely heavily on context. Inspired by [30, 31], we design GE Block to model the global context through the self-attention mechanism to effectively capture long-distance feature dependencies. Through the information interaction of the global context, the feature map contains richer semantic information, thereby enhancing the feature response of small objects.

3.2.1. GE Block. To enhance the information fusion between high-resolution feature layer and low-resolution feature layer, we designed a global feature enhancement block (as shown in Figure 3(b)) in SCEM, which utilizes a self-attention mechanism to enhance the representation of features by learning the global dependencies of features. Encode broader contextual information into local features, thereby enhancing its representational power. The processing steps of $\mathrm{GE}(\cdot)$ are as follows.

M_i is redefined as X, and X is used as the input of this model, and $\{Q, K, V\}$ are obtained through three convolutional layers, respectively. Then, perform matrix transpose operation on Q to get Q^T. We performed matrix multiplication of the reshaped K and Q^T to obtain the spatial attention map W. Next, we performed the matrix multiplication operation on the reshaped V and W to weight the spatial information and perform an element-wise addition with M to obtain the final output D as the output of SCEM. In particular, we formulated this procedure as follows.

$$\begin{aligned} q_i, k_j, v_j &= f_q\left(X_i\right), f_k\left(X_j\right), f_v\left(X_j\right), q_i \in Q, k_j \in K, v_j \in V, \\ \widetilde{X}_i &= F_{\mathrm{mul}}\left(F_{\mathrm{nom}}\left(F_{\mathrm{sim}}\left(q_i, k_j\right), v_j\right)\right) + X_i, \end{aligned} \tag{5}$$

where q_i is the i^{th} query; k_j and v_j are the j^{th} key/value pair. $f_q(\cdot)$, $f_k(\cdot)$, and $f_y(\cdot)$ denote the query, key, and value transformer functions [31, 32], respectively. These functions specifically refer to matrix operations using the mapping matrix of q, k, and v and the input features. X_i and X_j are the i^{th} and j^{th} feature positions in $\boldsymbol{X}$, respectively. $F_{\mathrm{sim}}(\cdot)$ is the

similarity function dot product; $F_{\text{nom}}(\cdot)$ is the normalizing function softmax; $F_{\text{mul}}(\cdot)$ is the weight aggregation function matrix multiplication; and D_i is the i^{th} feature position in the output feature map $\widetilde{X}$. $\widetilde{X}$ as the output of $\text{GE}(\cdot)$ is redefined as D_i, and the subscript i is corresponding to the input feature M_i of $\text{GE}(\cdot)$.

3.3. Bidirectional Fusion Feature Pyramid Structure. Multiscale feature fusion integrates low-level features and high-level features through top-down lateral connection and constructs a feature representation with fine-grained features and rich semantic information. The fused features have stronger expressive ability, which is conducive to the detection of small objects. The standard RetinaNet [13] uses a top-down fusion feature pyramid structure, which uses feature pyramid levels P_3 to P_7, where P_3 to P_5 are computed from the output of the corresponding ResNet residual stage (C_3 through C_5) using top-down and lateral connections just as in [15].

Although the feature pyramid structure adopted by the RetinaNet [13] (see Figure 4(a)) can fully integrate multiscale features, the low-level features need to go through hundreds of convolution layers of backbone, resulting in the loss of a large amount of underlying information that is conducive to small object detection during the propagation process. Inspired by PANet [17] (see Figure 4(b)), we designed a bidirectional fusion feature pyramid structure (see Figure 4(c)). The structure adds a bottom-up path enhancement module built with a smaller number of convolutional layers, which ensures that the information of high-level features and low-level features is more fully integrated, while retaining as much low-level information as possible. As in [17], all pyramid levels have $C = 256$ channels.

In the bottom-up backbone network we keep the C_3 through C_6 layers in the standard RetinaNet [13], while making full use of the C_2 which contains rich low-level features.

3.3.1. Top-Down Path. The top-down path includes the features of N_2 through N_4. N_4 is the output feature after SCEM with C_4 and C_5 as input features.

$$N_4 = D_4. \tag{6}$$

N_3 is composed of the up-samplingN_4 and the output feature D_3 after the SCEM (Section 3.2) with C_4 and C_5 as the input features. The two parts are fused by the addition method (see Figure 5(b)), which is quite different from [17] (see Figure 5(a)).

$$N_3 = D_3 \bigoplus F_{\text{up}}(N_4), \tag{7}$$

where $\oplus$ is the feature fusion operation, and $F_{up}(\cdot)$ is the up-sampling operation to match the resolution of the feature image to be fused in the lower layer. N_2 is obtained by fusing two parts of features, which are N_3 after up-sampling operation and the output feature D_2 of the MCEM.

$$N_2 = D_2 \bigoplus F_{\text{up}}(N_3). \tag{8}$$

3.3.2. Bottom-Up Enhancement Path. The bottom-up enhancement path includes the features of P_2 through P_6. P_2 through P_4 are generated in the same way just as in PANet [17].

$$P_i = \begin{cases} N_i, & i = 2, \\ \text{Conv}_{1\times1}(N_{i-1}) \bigoplus F_{\text{down}}(P_{i-1}), & i = 3, 4, \end{cases} \tag{9}$$

where $\oplus$ is the feature fusion operation, $\text{Conv}_{1\times1}(\cdot)$ represents a 1×1 conv, and $F_{\text{do wn}}(\cdot)$ is the down-sampling operation to match the resolution of the feature image to be fused in the upper layer. P_5 is obtained by fusing 1×1 conv on C_5 and down-sampling on P_4. P_6 is obtained by fusing 1×1 conv on C_6 and down-sampling on P_5.

$$P_i = \text{Conv}_{1\times1}(C_i) \bigoplus F_{\text{down}}(P_{i-1}), i = 5, 6. \tag{10}$$

4. Experiments

4.1. Dataset and Evaluation Metrics. We perform all experiments on the MS COCO detection dataset with 80 categories, in which objects with scale smaller than 32×32 pixels are considered small objects. MS COCO has a large number of small object objects, and the proportion of small objects accounts for 41.43% [16]. We train models on train2017 and report results of ablation study on val2017. The final results are reported on test-dev. The COCO-style average precision (AP) is chosen as the evaluation metric. AP_{50} and AP_{75} represent the average precision when IoU is set to 0.5 and 0.75, respectively, and AP_{S}, AP_{M}, and AP_{L} represent the average precision of small objects, medium-sized targets, and large-sized targets, respectively.

4.2. Implementation Details. To demonstrate the effectiveness of the MFEFNet proposed in this article, we conducted a series of experiments on the MS COCO dataset for verification. For all experiments in this section, we used SGD optimizer to train our models on a machine, whose CPU is Intel i7-9700k, 32 RAM, × NVIDIA GeForce GTX TITAN X GPUs, the CUDA version is 10.1 and deep learning framework is Pytorch 1.7.1. We initialize the learning rate as 0.01 and decrease it to 0.001 and 0.0001 at 8th-epoch and 11th-epoch. The momentum is set as 0.9 and the weight decay is 0.0001. The classical net-works ResNet-50 and ResNet-101 are adopted as backbones for comparative experiments. Original settings of RetinaNet such as hyper-parameters for anchors and Focal Loss are followed for fairly comparison. For all studies we use an image scale of 500 pixels unless noted for training and testing.

4.3. Main Results. In this section, we evaluated the MFEFNet on the COCO test-dev and compare it with other state-of-the-art one-stage detectors and two-stage detectors. Implementation details and evaluation metrics are set as above. All the results are shown in Table 1.

By analyzing the experimental results in the table, it can be found that when Resnet101 is used as the backbone network, the standard RetinaNet [13] performs better in detecting large

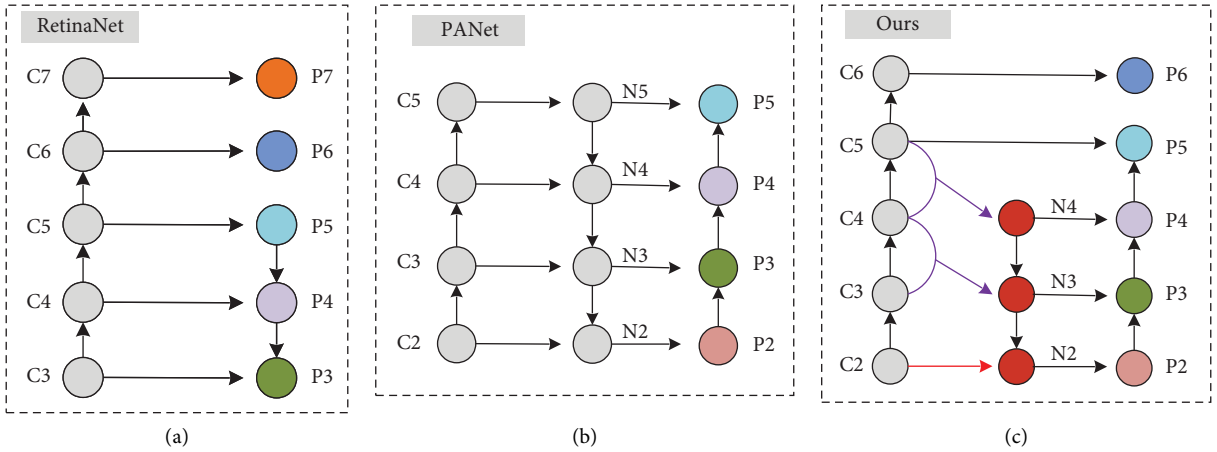


Figure 4: Comparisons of different backbone. (a) RetinaNet; (b) PANet; (c) Ours (MFEFNet), purple line: MCEM, red line: SCEM.

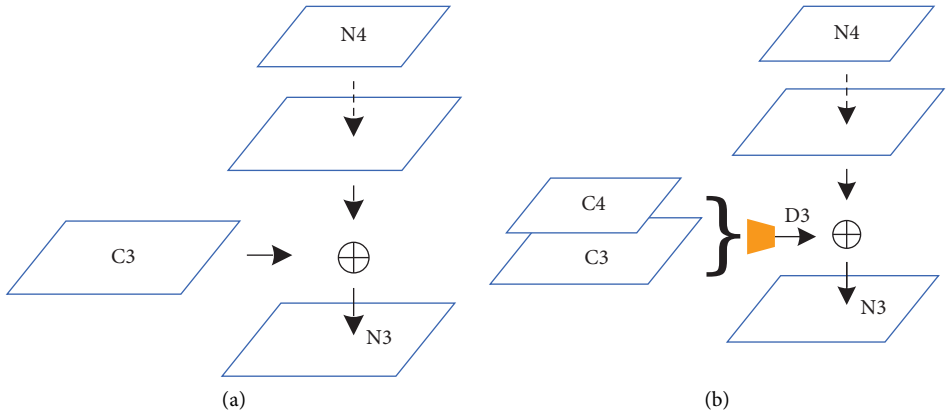


Figure 5: Comparisons of feature layer fusion, the dotted line refers to upsampling and the solid line refers to simple reference/input to the next step. (a) N_3 in PANet; (b) N_3 in MFEFNet (ours), the orange block: MCEM.

and medium targets, and achieves competitive results compared with the two-stage detectors, respectively, reaching 38.5% and 49.1%. However, when detecting small objects, it is only 14.7%, which is 0.9% and 3.5% lower than the two-stage detectors Faster R-CNN +++ [4] and Faster R-CNN w FPN [15], respectively. Faster R-CNN +++ refers to R-FCN + Resnet-101. In addition, compared with the one-stage detector YOLOv3 [8], it is 3.6% lower, and there is still much room for improvement. It is worth noting that the MFEFNet proposed in this article achieved excellent results in both large and small objects, and the AP_S reached 17.6%, which was improved by 2.9% and 1.0%, respectively, compared with standard RetinaNet [13] and Faster R-CNN+++ [4]. Combining the above analysis and experimental data, it can be found that the model proposed in this article has greatly improved the detection effect of targets of various sizes, especially for small objects.

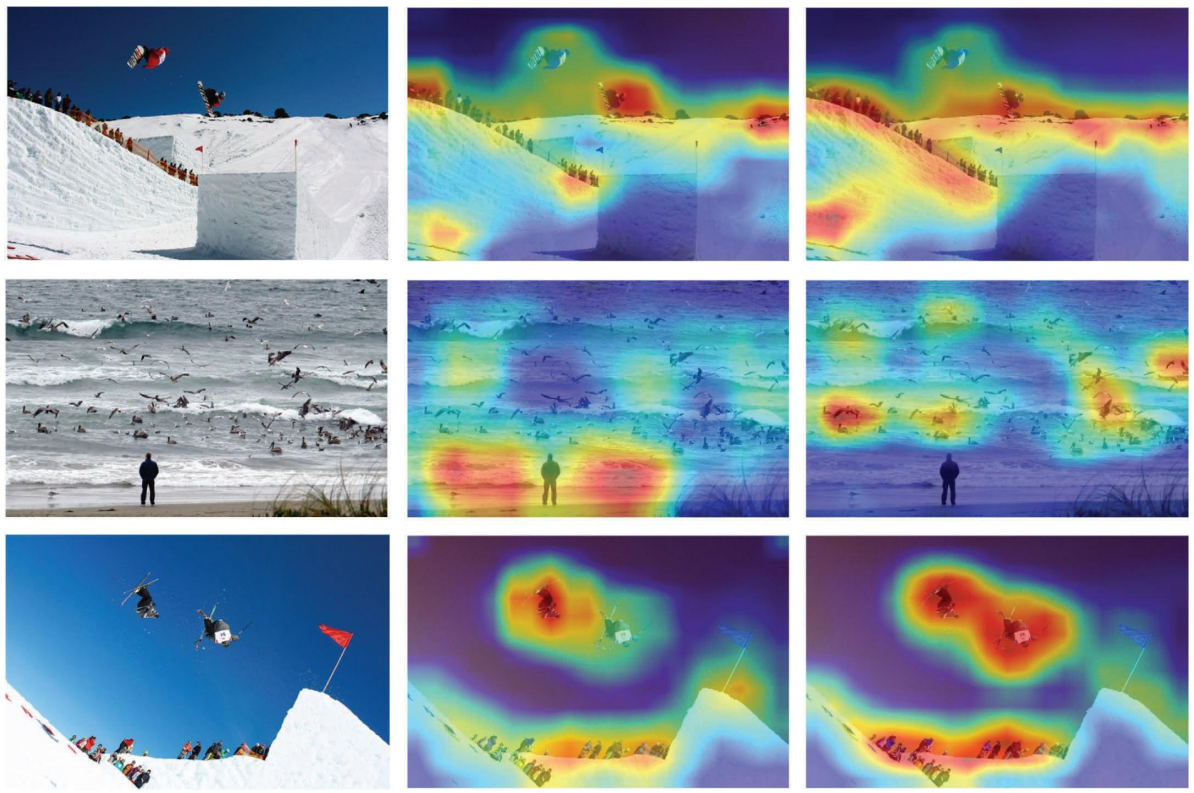
Figure 6 shows the visual comparison of features through convolution layer. Specifically, in this work, we use Grad CAM to calculate and visually display the output of the last convolution layer of the model in combination with the network structure and the weight after training. Column (a) is the original image, and column (b) is the feature visualization result of RetinaNet [13]. It can be found that the heat map does not cover small objects well, which shows that RetinaNet [13] is not sensitive to small objects. The improved network in this article improves the utilization of features and reduces the loss of features. As shown in column (c), it can be found that the feature heatmap of MFEFNet can better cover the boundary of the object, and can pay more attention to more number of small goals. This proves that the improved network can effectively enrich the features of small-scale feature detection, making the network pay more attention to the neglected small objects.

4.4. Ablation Study. In this section, we conducted extensive ablation experiments to analyze the effects of individual components in our proposed method. We also analyze the effect of each proposed component of MFEFNet on COCO val2017. The purpose of this study is as follows.

To analyze the importance of each component in MFEFNet, we gradually applied multiscale context extraction module, subpixel convolution enhancement module, and bidirectional fusion feature pyramid structure to the model to verify the effectiveness. Meanwhile, the

Table 1: MFEFNet vs. other two-stage and one-stage detectors on COCO test-dev.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
DeNet [33]	ResNet-101	33.8	53.4	36.1	12.3	36.1	50.8
CoupleNet [28]	ResNet-101	34.4	54.8	37.2	13.4	38.1	50.8
Faster R-CNN +++ [4]	ResNet-101	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [15]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN [5]	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
Cascade R-CNN [23]	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
One-stage methods							
YOLOv2 [7]	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [10]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
RetinaNet [13]	ResNet-50	32.5	50.9	34.8	13.9	35.8	46.7
YOLOv3 [8]	DarkNet-53	33.0	57.9	34.4	18.3	35.4	51.1
DSSD513 [11]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
EfficientDet-D0 [14]	EfficientNet	34.6	53.0	37.1	—	—	—
RetinaNet [13]	ResNet-101	34.4	53.1	36.8	14.7	38.5	49.1
RefineDet512 [12]	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4
MFEFNet (ours)	ResNet-50	34.8	52.8	37.4	16.8	37.3	47.9
MFEFNet (ours)	ResNet-101	36.2	54.2	38.3	17.6	39.5	50.1



(a) (b) (c)

Figure 6: Comparisons of feature heat maps results. (a) The original images; (b) RetinaNet output features results; and (c) MFEFNet output features results.

improvements brought by the combination of different components are also presented to demonstrate that these components complement each other. The baseline method for all ablation studies is ResNet50. All results are shown in Table 2.

By analyzing the experimental data in the table, it can be found that compared with the standard RetinaNet [13], the three structures proposed in this article have different degrees of improvement in the detection AP of targets of different scales. After adding the BidiFPN to the standard

Table 2: Effect of each component on COCO val-2017.

RetinaNet	BidiFPN	MCEM	SCEM	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√	×	×	×	32.5	50.9	34.8	13.9	35.8	46.7
√	√	×	×	33.3	51.6	35.8	14.9	36.3	47.1
√	√	√	×	34.1	52.3	36.8	16.0	36.9	47.4
√	√	×	√	34.0	52.1	36.6	15.7	36.5	47.5
√	√	√	√	34.8	52.8	37.4	16.8	37.3	47.9

Note: BidiFPN: bidirectional fusion feature pyramid structure; MCEM: multiscale context extraction module; SCEM: subpixel convolution enhancement module.

Table 3: Ablation studies of MCEM on COCO val-2017.

$R+B+S$	Par-dilated	Ser-dilated	Den-dilated	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√	×	×	×	34.0	52.1	36.6	15.4	36.5	47.5
√	√	×	×	34.4	52.6	37.1	16.4	37.2	47.7
√	×	√	×	34.3	52.5	36.9	16.0	37.0	47.7
√	×	×	√	34.8	52.8	37.4	16.8	37.3	47.9

$R+B+S$: RetinaNet + BidiFPN + SCEM.

RetinaNet [13], the AP is increased by 1.2%, and the small object average precision (AP_S) reaches 14.9%, an increase of 1.0%. In addition, after adding MCEM and SCEM, the average precision of small objects (AP_S) is increased by 1.1% and 0.8%, respectively, which also indicates that the shallow features fully extracted by MCEM and channel information at high-level are very helpful for small object detection. In addition to the large improvement in the detection effect of small objects, the detection average precision of large-sized objects and medium-sized objects has also been improved to varying degrees. The improved model improves the AP from 32.5% to 34.8%. Especially, the small object average precision (AP_S) also achieves a very meaningful improvement, from 13.9% to 16.8%, an increase of 2.9%.

To verify the effectiveness of densely connected dilated convolutions with different dilation rates in MCEM, we conducted the following ablation experiments. Feature extraction is performed in the following three ways: Par-dilated means that the three dilated convolutional layers are only connected in parallel to perform feature extraction on the shallow feature C_2. Ser-dilated means that the three dilated convolutional layers are only connected in series for feature extraction, and the convolutional layers are connected in increasing order according to the dilation rate. Den-dilated represents the MCEM used in this article for feature extraction. The experimental results are shown in Table 3. The visual structure diagram of three connection modes is shown in Figure 7.

By analyzing the data in the table, it can be found that the shallow feature extraction in the Den-dilated is more conducive to small object detection, and the average accuracy of small objects reaches 16.8%, an increase of 1.4%. We analyzed when features are extracted by Den-dilated, it fully expands the receptive field and strengthens the information fusion between different feature layers, which can extract more sufficient location information and semantic information. Although the other two methods have different degrees of improvement in the detection results, the effect is weaker than that of Den-dilated. In particular, the detection effect of Par-dilated is better than that of Ser-dilated, especially in small object detection. Par-dilated is 0.4% higher than that of Ser-dilated in small object detection. We believe that the parallel dilated convolution can greatly expand the receptive field, and can more fully extract high-resolution features that are conducive to small object detection.

To verify the effectiveness of GE Block in SCEM, we conducted the following ablation experiments. SCEM can be divided into a channel dimension reduction part based on subpixel convolution and the nonlocal feature extraction part based on GE block. The experimental results are shown in Table 4.

We analyzed the experimental data in the table and found that when only using subpixel convolution for channel dimension reduction, the detection accuracy has been greatly improved, and the average accuracy has increased from 34.1% to 34.6%, an increase of 0.5%. In addition, the small object detection accuracy is improved by 0.6%. However, after adding the GE Block, the detection accuracy of targets of various sizes has been further improved, and the APs has reached 16.8%, an increase of 0.9%. This is due to the fact that GE Block uses the spatial attention mechanism to fully obtain spatial context information, which is very helpful for small objects that rely heavily on context information.

4.5. Visualization of Results. In order to more intuitively demonstrate the effectiveness of the model proposed in this article, we visualized the detection effect of the standard RetinaNet [13] and the MFEFNet proposed in this article on the MS COCO dataset, as shown in Figure 8. The first column in the chart represents the original image, the second column is the detection result of RetinaNet [13], and the last column is the detection result of MFEFNet.

From the detection results, it can be found that compared with the standard RetinaNet [13], MFEFNet can detect more small objects. In the first line of detection results, it can be found that MFEFNet is able to detect people, which are

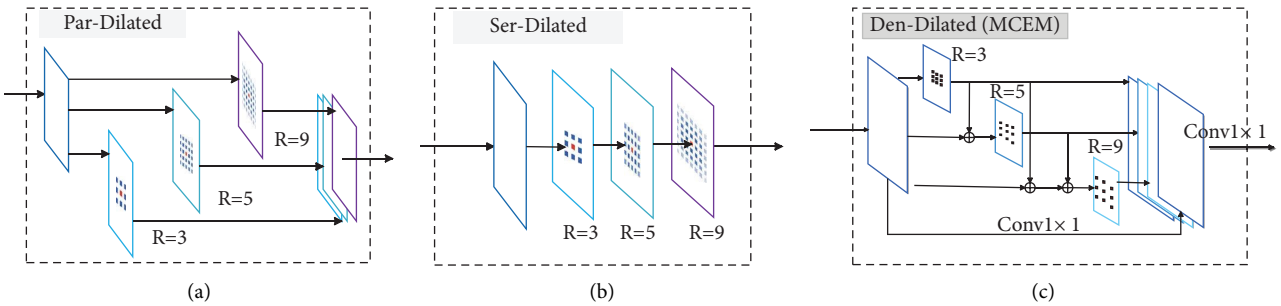


Figure 7: Comparisons of densely connected dilated convolutions. (a) Par-dilated; (b) ser-dilated; and (c) den-dilated (MCEM).

Table 4: Ablation studies of SCEM on COCO val-2017.

$R+B+M$	SubD	GE-B	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√	×	×	34.1	52.1	36.7	15.9	36.9	47.4
√	√	×	34.6	52.5	37.2	16.5	37.2	47.6
√	√	√	34.8	52.8	37.4	16.8	37.3	47.9

Note: SubD: sub-pixel convolutional dimension reduction and GE-B: GE block. $R+B+M$: retinaNet + BidiFPN + MCEM.

Figure 8: Comparisons of small object detection results. (a) The original images; (b) RetinaNet test results; (c) MFEFNet test results; and (d) the ground truth.

targets not detected by RetinaNet [13]. In the second row, RetinaNet [13] detected false objects and missed some objects. The white tent was mistakenly identified as sheep, the grass was mistakenly identified as cow, and the distant cow was not detected, which were successfully avoided in MFEFNet. From the experimental results in the third and fourth rows, it can be found that MFEFNet can also accurately identify a larger number of small objects such as cows. These experimental results show that the improved model in this article can further enhance the representation ability of the model and can greatly improve the missed detection and false detection of small objects.

5. Conclusions

This article deeply analyzes the key factors affecting small object detection and points out the shortcomings of the excellent single-stage object detector RetinaNet in small object detection. This work proposes a small object detection network based on multiple feature enhancement (MFEFNet) starting from improving high-resolution utilization and reducing information loss during propagation. First, it uses densely connected dilated convolutions to adequately extract shallow layer C_2, improving the utilization of high-resolution features. Second, this work introduces a bi-directional feature pyramid structure to shorten the shallow feature propagation path. Finally, this work makes full use of channel features containing rich semantic information through subpixel convolution to avoid channel information loss caused by channel dimension reduction in lateral connections. This article conducts sufficient experiments and stable detection improvements on the challenging MS COCO dataset, and the experimental results show that the detection effect of the improved method is greatly improved, and the AP is improved by 2.3%. The AP_S is increased by 2.9%, which effectively improves the detection effect of small objects. This article demonstrates the effectiveness of the model through sufficient experiments, and we believe this work can help future object detection research. [34].

Data Availability

The data presented in this study are openly available in MS COCO at https://doi.org/10.1007/978-3-319-10602-1_48.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[2] R. Girshick, J. Donahue, T. Darrell, and M. Jitendra, "Rich feature hierarchies for accurate object detection and semantic segmenta-tion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[4] S. Ren, K. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[5] K. He, G. Gkioxari, and P. Dollár, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, G. Ross, and F. Ali, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEE, Las Vegas, NV, USA, June 2016.

[7] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[8] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[9] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[10] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, pp. 21–37, Springer, Cham, Amsterdam, Netherlands, October 2016.

[11] Z. Li and F. Zhou, "FSSD: feature fusion single shot multibox detector," 2017, https://arxiv.org/abs/1712.00960.

[12] S. Zhang, L. Wen, X. Bian, L. Zhen, and Z. L. Stan, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE con-ference on computer vision and pattern recognition*, pp. 4203–4212, Salt Lake City, UT, USA, June 2018.

[13] T. Y. Lin, P. Goyal, R. Girshick, H. Kaiming, and D. Piotr, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, Venice, Italy, October 2017.

[14] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790, Seattle, WA, USA, June 2020.

[15] T. Y. Lin, P. Dollar, and R. Girshick, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Honolulu, HI, USA, July 2017.

[16] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Proceedings of the European conference on computer vision*, pp. 740–755, Springer, Cham, Zurich, Switzerland, September 2014.

[17] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Salt Lake City, UT, USA, June 2018.

[18] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: towards balanced learning for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 821–830, Long Beach, CA, USA, June 2019.

[19] Y. Luo, X. Cao, and J. Zhang, "CE-FPN: enhancing channel information for object detection," 2021, https://arxiv.org/abs/2103.10643.

[20] Q. Zhang, J. Ren, H. Liang, Y. Yang, and L. Chen, "BFE-net: bidirectional multi-scale feature enhancement for small object detection," *Applied Sciences*, vol. 12, no. 7, p. 3587, 2022.
[21] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: implementing efficient convnet descriptor pyramids," 2014, https://arxiv.org/abs/1404.1869.
[22] W. Shi, J. Caballero, F. Huszár, T. Johannes, P. A. Andrew, and B. Rob, "Real-time single image and video super-resolution using an efficient sub-pixel convolu-tional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, Las Vegas, NV, USA, June 2016.
[23] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," 2017, https://arxiv.org/abs/1712.00726.
[24] Q. Chen, Y. Wang, T. Yang, Z. Xiangyu, C. Jian, and S. Jian, "You only look one-level feature, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition," pp. 13039–13048, 2021, https://arxiv.org/abs/2103.09460.
[25] C. Guo, B. Fan, Q. Zhang, and X. Shiming, "AugFPN: improving multi-scale feature learning for object detection," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, June 2020.
[26] G. Ghiasi, T. Y. Lin, and Q. V. Le, "NAS-FPN: learning scalable feature pyramid architecture for object detection," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, CA, USA, June 2019.
[27] Y. Zhu, C. Zhao, J. Wang, and Z. Xu, "Couplenet: coupling global structure with local parts for object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 4126–4134, Seattle, WA, USA, October 2017.
[28] S. Qiao, L. C. Chen, and A. Yuille, "Detectors: detecting objects with recursive feature pyramid and switchable atrous convolu-tion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10213–10224, Long Beach, CA, USA, June 2021.
[29] J. S. Lim, M. Astrid, H. J. Yoon, and I. L. Seung, "Small object detection using context and attention," in *Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 181–186, IEEE, Las Vegas, NV, USA, April 2021.
[30] X. Wang, R. Girshick, A. Gupta, and H. Kaiming, "Non-local neural networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Salt Lake City, UT, USA, December 2018.
[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.
[32] H. Zhang, H. Zhang, C. Wang, and X. Junyuan, "Co-occurrent features in semantic segmentation," in *Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition*, pp. 548–557, Seattle, WA, USA, June 2019.
[33] J. Huang, V. Rathod, C. Sun, Z. Menglong, K. Anoop, and F. Alireza, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7310-7311, Honolulu, HI, USA, July 2017.
[34] N. Carion, F. Massa, G. Synnaeve, U. Nicolas, K. Alexander, and Z. Sergey, "End-to-end object detection with transformers," in *Proceedings of the European conference on computer vision*, pp. 213–229, Springer, Cham, Heidelberg, Germany, November 2020.