symmetry

MDPI

Article

FE-RetinaNet: Small Target Detection with Parallel Multi-Scale Feature Enhancement

Hong Liang, Junlong Yang * and Mingwen Shao

School of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; liangh@upc.edu.cn (H.L.); mwshao@upc.edu.cn (M.S.)
* Correspondence: z19070035@s.upc.edu.cn

Abstract: Because small targets have fewer pixels and carry fewer features, most target detection algorithms cannot effectively use the edge information and semantic information of small targets in the feature map, resulting in low detection accuracy, missed detections, and false detections from time to time. To solve the shortcoming of insufficient information features of small targets in the RetinaNet, this work introduces a parallel-assisted multi-scale feature enhancement module MFEM (Multi-scale Feature Enhancement Model), which uses dilated convolution with different expansion rates to avoid multiple down sampling. MFEM avoids information loss caused by multiple down sampling, and at the same time helps to assist shallow extraction of multi-scale context information. Additionally, this work adopts a backbone network improvement plan specifically designed for target detection tasks, which can effectively save small target information in high-level feature maps. The traditional top-down pyramid structure focuses on transferring high-level semantics from the top to the bottom, and the one-way information flow is not conducive to the detection of small targets. In this work, the auxiliary MFEM branch is combined with RetinaNet to construct a model with a bidirectional feature pyramid network, which can effectively integrate the strong semantic information of the high-level network and high-resolution information regarding the low level. The bidirectional feature pyramid network designed in this work is a symmetrical structure, including a top-down branch and a bottom-up branch, performs the transfer and fusion of strong semantic information and strong resolution information. To prove the effectiveness of the algorithm FE-RetinaNet (Feature Enhancement RetinaNet), this work conducts experiments on the MS COCO. Compared with the original RetinaNet, the improved RetinaNet has achieved a 1.8% improvement in the detection accuracy (mAP) on the MS COCO, and the COCO AP is 36.2%; FE-RetinaNet has a good detection effect on small targets, with APs increased by 3.2%.

Keywords: target detection; RetinaNet; multi-scale feature enhancement; bidirectional feature pyramid network

Citation: Liang, H.; Yang, J.; Shao, M. FE-RetinaNet: Small Target Detection with Parallel Multi-Scale Feature Enhancement. *Symmetry* **2021**, *13*, 950. https://doi.org/10.3390/sym13060950

Academic Editors: Whoi-Yul Kim and Moonsoo Ra

Received: 17 April 2021
Accepted: 25 May 2021
Published: 27 May 2021

1. Introduction

The task of object detection has always been one of the main tasks in the field of computer vision. In recent years, thanks to the development of convolutional neural networks, the performance of target detection algorithms has made great progress. Target detection algorithms based on convolutional neural networks can be divided into two categories: one is a two-stage target detection model, such as R-CNN [1], Fast R-CNN [2], Faster R-CNN [3], Mask R-CNN [4], R-FCN [5], etc. The two-stage target detection algorithm distinguishes the foreground and the background in the first stage to generate candidate regions and then extracts features in the second stage to classify and regress the target. The other is a single-stage target detection model, such as SSD [6], YOLOv1 [7], YOLOv2 [8], YOLOv3 [9], YOLOv4 [10], RetinaNet [11], EfficientDet [12], etc. The single-stage target detection algorithm directly generates the class probability and position coordinate value of the object, and the final detection result can be directly obtained after a single detection. In general, compared with the two-stage target detection model, the single-stage

detection model is faster, but the detection accuracy of targets (especially small targets) is slightly lower.

In recent years, researchers have been committed to improving the accuracy of the model as much as possible while ensuring the speed of the single-stage target detection model. As a result, many one-stage target detection algorithms with fast speed and high accuracy have emerged. The paper [13] constructed a top-down feature pyramid network. By fusing high-level strong semantic feature maps and low-level high-resolution feature maps, high-level semantic information is transferred to the lower layers, and the accuracy of the network is improved by fusing multi-scale information. The paper [11] designed the focal loss to solve the problem of a serious imbalance in the proportion of positive and negative samples in single-stage target detection, making RetinaNet more accurate than the most advanced two-stage detection model at the time. Among the existing target detection algorithms, RetinaNet solves the problem of category imbalance by introducing focal loss, which has attracted wide attention from researchers. The standard RetinaNet uses ResNet as the backbone network, uses a top-down feature pyramid network to fuse multi-scale semantic information, and adds two subnets to the back end of the network for target classification and bounding box regression. However, the standard RetinaNet has a good detection effect for large and medium targets, but the detection effect for small targets is not satisfactory.

Through the analysis and research of RetinaNet, we find the two most critical factors that affect the small targets detection performance of RetinaNet. First, the small target information will be lost to varying degrees in the deep and shallow layers of the feature extraction network. On the one hand, the traditional feature extraction networks (such as VGG16 [14] and ResNet [15]) are to repeatedly perform convolution and maximum pooling down sampling operations to extract features. Although a certain degree of semantic information is preserved, it is difficult to extract multi-scale contextual information from the shallow layer to distinguish the background and the target. We use an auxiliary multi-scale feature enhancement module to assist in the extraction of multi-scale shallow features and merge them with the features extracted from the backbone network, which greatly improves the expression ability of small targets. On the other hand, target detection algorithms usually use a backbone network specially designed for image classification, which causes small targets to be ignored in the high-level feature maps of the network. Although the down-sampling feature map with a larger multiple has high-level semantic information, the resolution of the high-level feature map is greatly reduced, which is not conducive to target positioning. This problem also exists in RetinaNet: RetinaNet uses ResNet as the feature extraction network, and adds two additional feature layers, P6 and P7, to improve the detection effect of large targets. This results the resolution of P7 to 1/128 of the original, making smaller targets are not visible in the feature map of this layer. Even if the FPN merges deep and strong semantic information with the shallow layer, the small target is missing in such a deep feature map, the semantic information of the small target is also lost. The task of target detection is different from the task of classification. It not only needs to judge the category of the object but also needs to locate the position of the target in space. This work aims to improve the detection effect of small targets without reducing the detection effect of large targets. Inspired by DetNet [16], the backbone network designed in this work can make the deep feature map retain a larger receptive field without increasing the number of down sampling, which is conducive to retaining the small target information in the deep feature map.

Secondly, RetinaNet adopts the traditional top-down FPN structure, so that feature transfer will be restricted by one-way information flow. Although the pyramid structure used by RetinaNet can transmit strong semantic information from the top layer to the bottom layer, this structure can only inject high-level semantic information into the previous layer and ignores the transmission of shallow high-resolution feature maps. Additionally, because the shallow spatial feature information is gradually faded in the bottom-up transfer process, the small target loses the guidance of the high-resolution feature information after

layer-by-layer convolution and down sampling, resulting in a decrease in model detection performance. Although the papers [12,17] all adopted a two-way feature pyramid network to achieve multi-scale feature fusion, adding additional branches to shorten the process of shallow feature propagation, these works only reuse the shallow features extracted from the backbone network in additional branches, and cannot capture rich contextual information. This work combines the multi-scale feature enhancement module with the bidirectional feature pyramid network, which greatly enriches the shallow multi-scale context information. When detecting objects of different scales, especially small targets, low-level and middle-level high-resolution information and high-level strong semantic information are required [18]. This work implements this point of view and believes that transferring high-level information to the previous layer while transferring low-level information to the next layer is the key to multi-scale target detection, especially small target detection.

Overall, through the research and analysis of the RetinaNet, we propose a one-stage target detection algorithm with the following contributions to improve the detection accuracy of small targets:

1. We propose a simple and effective parallel multi-scale feature enhancement module, which can expand the characteristics of the receptive field without down sampling by using dilated convolution and assist the backbone network to extract shallow features with multi-scale context information.
2. We introduce a method to improve the backbone network specifically for the target detection task, which effectively reduces the gap between the feature extraction network in the detection task and the classification task. This method enables the high-level feature map of the backbone network to preserve the texture information of small targets as much as possible while preserving large receptive fields and strong semantic information.
3. We combine the auxiliary multi-scale feature enhancement module with the original FPN structure to construct a bidirectional feature pyramid network containing multi-scale shallow information. Unlike most bidirectional structures that reuse the backbone network to extract features in additional branches, this article uses the multi-scale feature enhancement module as the input of the additional branches, which brings brand-new feature information to the network.

2. Related Work

The task of small target detection has always been a challenge in the field of computer vision. For most single-stage target detection models, including SSD [6], YOLO [7–10], RetinaNet [11], etc., they perform poorly on small target detection tasks. Reviewing the development of target detection algorithms, the methods to improve the accuracy of target detection are mainly to adopt better feature extraction networks, add more context information, and multi-scale feature fusion.

Better feature extraction networks: Most target detection algorithms use better feature extraction networks to improve the accuracy of the detection algorithm. SSD [6] used VGG-16 as the basic network, and the performance of extracting features is relatively weak, which is one of the reasons for the relatively poor detection effect of SSD. In subsequent improvements, DSSD [19] replaced the VGG with ResNet-101, which improved the performance of the model. YOLOv2 [8] also abandoned the inception structure used by YOLOv1 [7] and constructed a darknet-19 with stronger feature extraction capabilities. YOLOv3 [9] also constructed darknet-53 as the backbone network, which is faster and more effective than ResNet-101. Overall, most advanced target detection algorithms use deeper networks to extract features. Since larger objects are predicted on the deeper feature maps, the requirement for the receptive field corresponding to the original image ratio is also large, so RetinaNet has added two additional layers P6 and P7 based on ResNet to improve the detection effect of large targets. However, on the one hand, the deeper the feature map, the more blurred the edge definition of the object, and the weaker the corresponding

regression. On the other hand, the resolution of the deep feature map is small, and it is difficult to see small objects on the deep feature map. Even if networks such as FPN [13] and RetinaNet [11] add the shallow layer to the deep layer with strong semantics, since the small object target has disappeared in the deep feature map, a large part of the semantic information will still be lost. To solve this problem, DetNet [16] constructed a network dedicated to detection tasks. In the backbone, dilated convolution was used to reduce the number of down sampling and maintain the large receptive field while retaining the texture information of the small target. In the paper [18], to use low and medium-level information to detect small targets, an additional LSN lightweight network was added to extract shallow features. However, training from scratch may encounter convergence problems, and the performance of the auxiliary network is even worse than that of pre-training. The multi-scale feature enhancement module used in this work is different and it can be used as an auxiliary branch in the network for end-to-end training.

More contextual information: There are many works dedicated to increasing the information outside the bounding box, that is, adding more contextual features to improve the accuracy of target detection. DSSD [19] used deconvolution to construct a wide and narrow hourglass structure and improved the detection effect of small targets by introducing large-scale context information in the target detection. The paper [20] mentioned that each pixel has a larger receptive field to avoid missing important information during prediction. Increasing the effective receptive field size is also a way to enrich context information. Target segmentation, image classification, and target detection all require sufficient context information to reduce the possibility of misclassification. Inception series [21–24] designed a series of multi-branch convolution structures to improve the training effect by increasing the width of the neural network. This structure sets up convolution kernels of different sizes in each branch and enriches the context information of the features by extracting features of different receptive field sizes. However, due to the heavy traces of Inception's manual design, ResNext [25] combined the Inception structure with the residual structure, added the idea of multi-group convolution, and constructed a multi-branch model structure. Each branch of the module has the same topology, which makes the model highly modular. RFBNet [26] used dilated convolution to generate a larger receptive field, captured more contextual information in a larger area, and kept the feature map at a higher resolution. The auxiliary multi-scale feature enhancement module is inspired by this kind of structure. It acquires receptive fields of different sizes in multiple branches and adds more context information to the shallow layer.

Multi-scale feature fusion: FPN combined low-resolution, strong semantic feature maps with high-resolution, low-semantic feature maps through a top-down and horizontal connection structure to build high-level semantic feature maps at all scales. This is very important for small target detection. Following this idea, to shorten the transmission path between the bottom and top features, the paper [17] added a bottom-down path enhancement branch based on FPN, which further enhanced the positioning ability of the network. The paper [12] also introduced a weighted two-way feature pyramid network, which is conducive to the simple and fast multi-scale feature fusion of the network. However, most of these two-way feature pyramids are extracted from the reused backbone network. In this work, additional feature enhancement modules are added to form a symmetrical two-way structure to enhance the feature information and to increase the diversity and richness of information.

3. Method

This section introduces the overall structure of the detection model, and then introduces three main parts, including the improved Resnet-D, the parallel auxiliary multi-scale feature enhancement model MFEM (Multi-scale Feature Enhancement Model), and the bidirectional feature pyramid network. The bidirectional feature pyramid network proposed in this work combines the features extracted by ResNet-D and the multi-scale context features extracted by the MFEM, which are from the backbone network and MFEM, respectively. It

makes the data flow transfer in a symmetrical way, improves the transmission efficiency of shallow features and improves the network's ability to express small targets.

3.1. Overall Architecture

Figure 1 shows the overall structure of the network, which is mainly composed of three parts: improved ResNet, MFEM, and bidirectional feature pyramid network. Inspired by DetNet, this work improves ResNet and named it ResNet-D. The backbone discards the 64 and 128 times down sampling in the original RetinaNet and uses the dilated residual structure to replace the conventional residual module so that the high-level feature map has a larger receptive field while retaining more small target textures information.

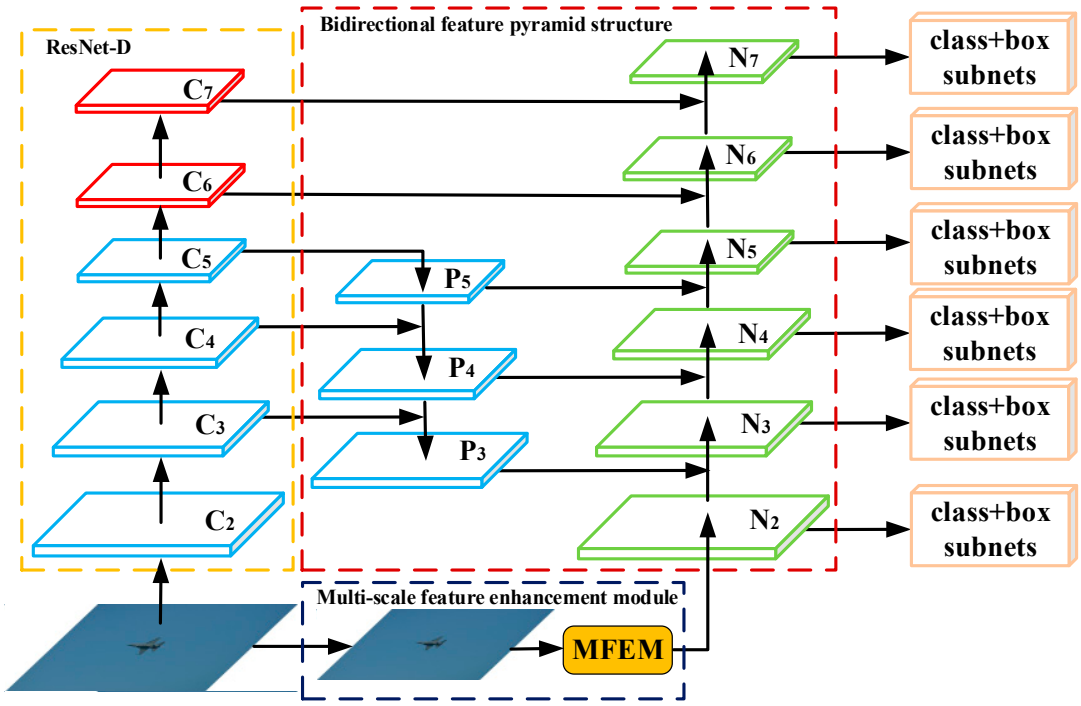

Figure 1. FE-RetinaNet network structure.

Inspired by the Inception structure and the idea of grouping convolution in ResNext, this work designs a multi-scale feature enhancement module, which is a multi-branch structure with different receptive fields and can greatly enrich the shallow context information. The constant receptive field at each prediction layer in RetinaNet captures only a fixed contextual information. Inspired by the Inception structure and the idea of grouping convolution in ResNext, this work designs a multi-scale feature enhancement module (MFEM). MFEM uses 4 dilated convolution branches with different dilation rates to extract feature maps with different receptive fields, that is to extract more features beyond the target to improve the discrimination ability of the standard RetinaNet prediction layer. We start by down sampling an input image with a simple pooling operation to match its size with that of first prediction layer of FE-RetinaNet. Then, the down sampled image is passed through our multi-scale feature enhancement module. In this work, the MFEM is combined with RetinaNet to construct a bidirectional feature pyramid network with a feature enhancement module. On the one hand, this structure can merge the features extracted from the backbone network with high-level semantic information and the shallow high-resolution features from top to bottom. On the other hand, it can use the bottom-down branch fusion of the shallow high-resolution features extracted by the feature enhancement module. Experiments show that this structure can effectively improve the accuracy of small target detection.

3.2. Improved Backbone Network ResNet-D

The commonly used feature extraction strategy in existing detection frameworks is usually to repeatedly stack multiple convolutional layers and maximum pooling layers (such as ResNet-50) to construct a deeper feature extraction network to generate strong semantic information. Semantic information refers the information expressed by images closest to human understanding. High level semantics is the feature information obtained after several times of convolution (feature extraction) and its receptive field is large and the extracted features are more and more abstract. Such a feature extraction strategy is more beneficial to image classification tasks that are more inclined to translation invariance. Different from image classification, target detection also requires accurate object description, and local low/medium level feature (such as texture) information is also the key.

The design of the backbone network usually has two major problems: (1) Keeping the backbone network with a high spatial resolution will consume memory and time greatly; (2) Reducing the number of down sampling will result in a reduction in the receptive field, which is not conducive to large target detection. Inspired by DetNet, this work improves ResNet, aiming to improve the detection effect of small targets without reducing the accuracy of large target detection. A new residual structure is introduced to solve these problems, that is, the 3×3 standard convolution layer in the standard residual module is replaced by the dilated convolution with a convolution core size of 3×3 and expansion coefficient of 2, forming the dilated residual structure, which can effectively expand the receptive field of the feature map, as shown in Figure 2b. Besides, a 1×1 convolutional layer is added to the shortcut branch of the dilated residual structure, so that the seventh stage can obtain new semantic information without down sampling, as shown in Figure 2c.

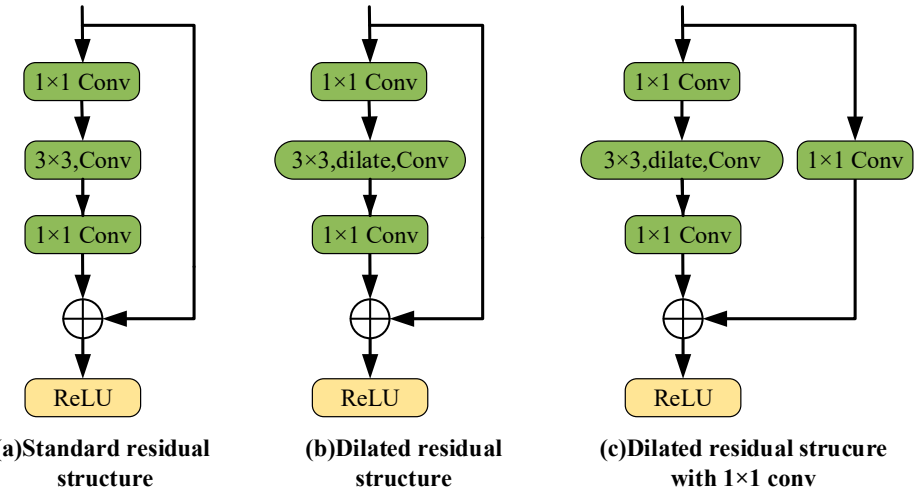

Figure 2. Dilated residual structure. (**a**) Standard residual structure; (**b**) Dilated residual structure; (**c**) Dilated residual structure with 1×1 convolutional layer.

The standard RetinaNet framework uses ResNet to extract features from the backbone network (as shown in Figure 3a), and adds two additional feature maps P6 and P7 (P6 is obtained by convolution operation with convolution kernel of 3×3 and step size of 2 on C5, and P7 is obtained by applying Relu and convolution operation with convolution kernel of 3×3 and step size of 2 on P6). Although it is beneficial to the detection of large targets, high-level feature maps will be ignored due to number of down sampling times being too great (as shown in Figure 3b). To make the high-level feature map have a larger receptive field while retaining more texture information of small targets, this work introduces more stages (namely C6 and C7) on the backbone network. In stages 6–7, we replace the conventional residual module with the dilated residual structure with dilated convolution (as shown in Figure 3c).

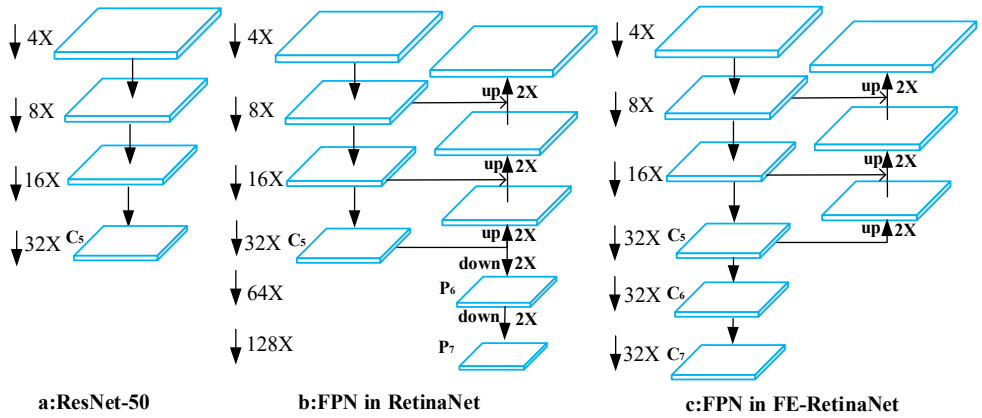

Figure 3. Comparisons of different backbones. (**a**) ResNet-50; (**b**) FPN in RetinaNet; (**c**) FPN in FE-RetinaNet.

The improved ResNet-D is shown in Figure 4. Based on the original 5 stages of ResNet, two additional stages are added, aiming at following the design idea of RetinaNet to improve the detection effect of the network on large targets. ResNet-D maintains that the original residual structure is used in the first five stages, and the dilated residual structure composed of three residual modules is used in the sixth and seventh stages, which can fix the spatial resolution of the two stages to 32 times of down sampling and effectively expand the receptive field of the high-level feature map. The down sampling multiples of each stage of the modified network are (2,4,8,16,32,32,32), respectively, and the feature map with 64 times and 128 times down sampling in RetinaNet is abandoned. There are three expansion bottleneck structures in the sixth and seventh level convolution, which are arranged in the order of (C, B, B). Since the space size is fixed after the fifth stage, in order to introduce a new stage, we use an extended bottleneck and 1×1 convolution projection at the beginning of each stage (Figure 2c). In this stage, the extended convolution is used instead of the traditional convolution structure, which effectively avoids the low sampling operation. Additionally, to reduce the number of network calculations, ResNet-D uses 1×1 convolution after the fourth stage to reduce the dimensionality of the feature map to 512 channels, and the number of feature maps until the seventh stage is 512.

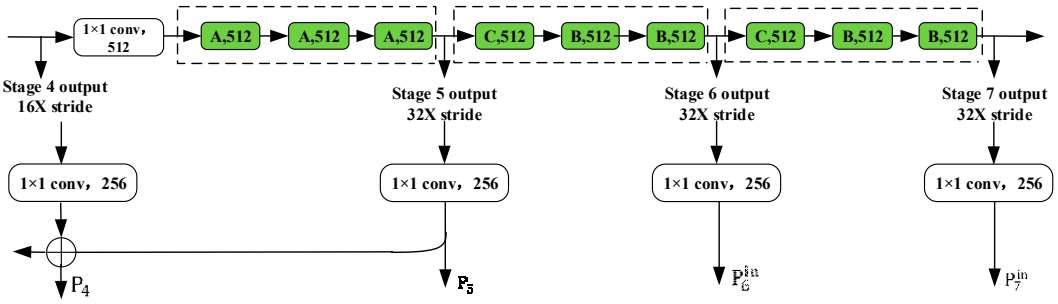

Figure 4. Improved ResNet-D structure.

3.3. Multi-Scale Feature Enhancement Module

Since most detection algorithms use fixed-size convolution kernels to extract target features, they can only extract local characteristic information, and the size of the receptive field is limited, unable to capture rich contextual information, which is not conducive to detecting complex natural images. Inspired by the idea of acquiring different size receptive fields in Inception and grouping convolution in ResNext, we design a multi-branch structure with different receptive fields. Firstly, the dilated convolution with different expansion rates is used to obtain different scales of information, and then the information with different scales is fused to obtain rich information. Multi-scale Feature Enhancement Module (MFEM) is a simple structure. It is an auxiliary branch directly

connected with ResNet-D to assist in extracting shallow features that carry multi-scale context information. This article introduces the feature extraction strategy used in MFEM and then describes the MFEM architecture.

Multi-scale feature enhancement module MFEM: The standard RetinaNet uses ResNet to extract features, and feature extraction is performed repeatedly by convolution and maximum pooling. Although the down-sampling process retains a certain degree of semantic features, it is still possible to lose low-level features that are helpful for detection. Besides, the current feature extraction network usually sets the receptive field of the same stage to the same size, which will lead to the loss of discriminative and robustness of features. To make up for the loss of shallow features in the down-sampling process and improve the discriminative and robustness of features, the MFEM module in this work provides an auxiliary feature extraction scheme, as shown in Figure 5.

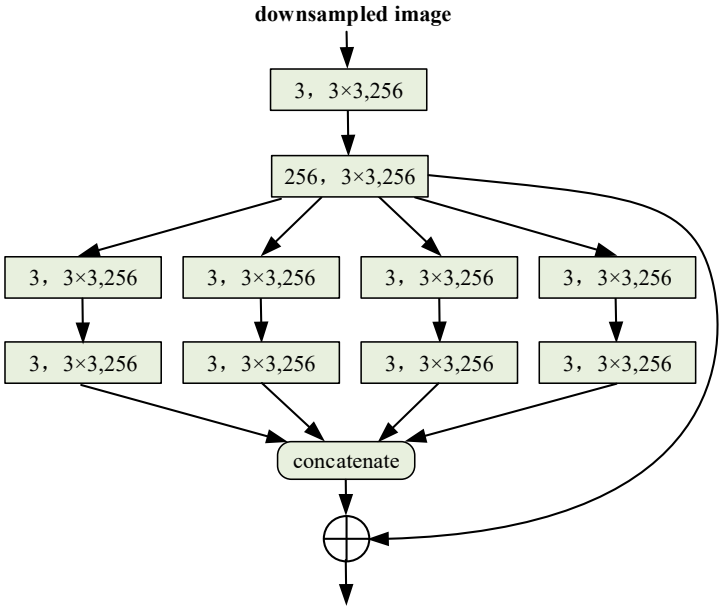

Figure 5. Multi-scale feature enhancement module.

Firstly, the input image I is sampled 4 times down by simple pooling operation to obtain I_t, and the size of I_t matches the second stage feature diagram C_2 of ResNet-D. The down sampled image outputs a feature map containing multi-scale context information through a modular block convolution structure, including segmentation, conversion, and aggregation operations.

The down sampled image I_t first generates the initial feature map $F_{\text{int}(0)}$ through two consecutive convolution layers with the size of 3×3 and 1×1,which is constructed as

$$F_{\text{int}(0)} = \varphi_{\text{int}(0)}(I_t) \tag{1}$$

where $\varphi_{\text{int}(0)}$ represents a serial operation, including a 3×3 convolution and a 1×1 convolution module. Then, the initial feature map $F_{\text{int}(0)}$ is used to generate the intermediate feature set $F_{\text{int}(k)}$ and k represents the number of branches of the multi-scale feature enhancement module:

$$F_{\text{int}(k)} = \omega_{\text{int}(k)}(F_{\text{int}(0)}) \tag{2}$$

where $\omega_{\text{int}(k)}$ represents the kth packet of grouped convolution, including a 1×1 convolutional layer and an expanded convolutional layer with a convolution kernel size of 3×3 and expansion rates of 1, 2, 3, and 4, respectively. Then, four feature maps contain-

ing multi-scale context information are fused in concatenate mode to obtain the fused feature F_{concat}:

$$F_{concat} = \sum_{k=1}^{4} Concat(F_{\text{int}(k)}) \tag{3}$$

where *Concat* represents the merging of information between channels. Then, the initial feature map $F_{\text{int}(0)}$ is identically mapped by the shortcut branch, and the feature F_{concat} is fused by the addition method, and the fused feature is F_{add}:

$$F_{add} = F_{concat} \oplus F_{\text{int}(0)} \tag{4}$$

where $\oplus$ denotes feature fusion by addition. The dimension of F_{add} is upgraded through the 1×1 convolutional layer to obtain the output feature N_2, to prepare these converted features for aggregation through the bottom-up pyramid branch:

$$N_2 = Conv(F_{add}) \tag{5}$$

where *Conv* represents a 1×1 convolution operation, which is used to upgrade the output feature to 256 channels. N_2 represents the output feature after the feature enhancement module, which has the same spatial resolution as C_2.

Figure 6 shows the comparisons of feature heat maps after incorporating the multi-scale feature enhancement module. Specifically, we performed experiments with the improved RetinaNet combined with MFEM on MS COCO and contrasted experiment results with original RetinaNet models. The improved RetinaNet model is trained on Nvidia GeForce GTX TITAN X GPUs. The momentum parameter is set to 0.9, the batch size is set to 32, and the initial learning rate is set to 0.001. For the first 160k iteration, we use a learning rate of 10^{-3}, then 10^{-4} for the 60k iteration, and 10^{-5} for the last 20k. The original RetinaNet model uses the official pre training model. For the fairness of the investigation, the backbone network used the unified ResNet-50 as the primary network structure. For the original image of the MS COCO, the standard RetinaNet output features are shown in column (b), which shows that RetinaNet is not sensitive to small targets. After the auxiliary feature enhancement module is integrated, as shown in column (c), it can be found that the characteristic thermal diagram of this network can better cover the boundary of the object. This proves that the auxiliary multi-scale feature enhancement module can effectively enrich the features of small-scale feature detection, and make the network more concerned with the neglected small targets.

3.4. Bidirectional Feature Pyramid Network

As mentioned earlier, both shallow and deep features are important for the detection of small targets. However, the top-down one-way pyramid structure adopted by RetinaNet focuses on transferring strong semantic information from the top level to the lower level, enhancing the transfer of high-level features in an asymmetrical manner, while ignoring the transfer of low-level features with rich positioning information. In addition, the standard RetinaNet uses ResNet to extract features, and repeated down sampling operations will cause low-level features to be lost. The bi-directional feature fusion used in this work focuses on the downward transmission of high-level features on the one hand. On the other hand, it focuses on the upward transfer of underlying features. This symmetric feature transfer method takes into account the transmission efficiency of shallow features and deep features, and solves the shortcomings of one-way information transfer in the past. In addition, the bi-directional pyramid structure strengthens the transmission of multi-scale shallow information in the bottom-up path to improve the accuracy of the network's detection of small targets.

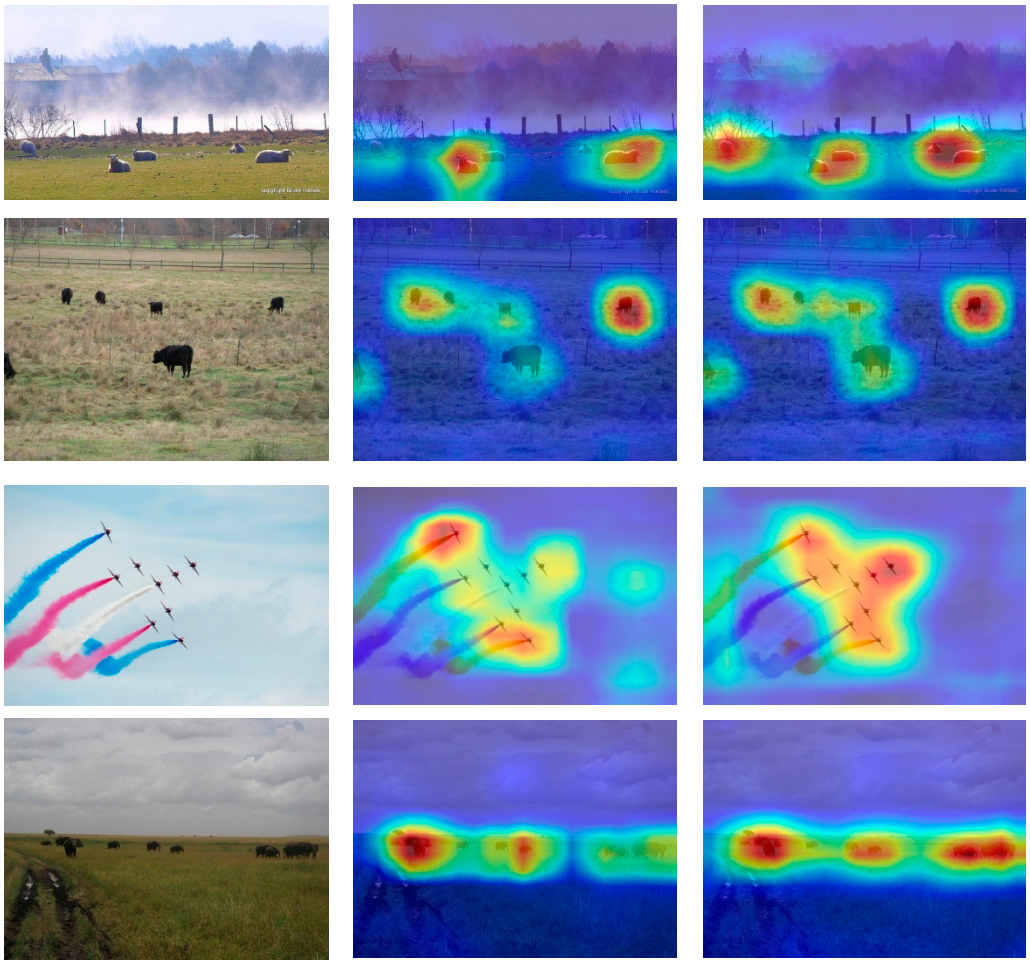

Figure 6. RetinaNet and improved RetinaNet combined with MFEM feature heat maps results. (**a**) Original image; (**b**) Standard RetinaNet output features; (**c**) Improved RetinaNet output features.

Top-down branch: According to the definition of FPN, the feature layers with the same space size are in the same network stage, and each feature level corresponds to a network stage. The improved ResNet-D in Section 3.1 is taken as the basic structure, and $\{C_3, C_4, C_5\}$ is chosen as the feature layer. The top-down path merges the stronger characteristics of high-level semantic information through horizontal connections from top to bottom. Consistent with the standard RetinaNet, each low-resolution feature image is upsampled, and the spatial resolution is expanded to match the size of the next layer of feature maps. The top-down path uses the 3rd to 5th stage features as input, namely

$$\vec{C} = (C_3, C_4, C_5) \tag{6}$$

where C_i represents the output feature map whose resolution is $1/2^i$ of the original image. For example, for an image with an input size of 512 × 512, the resolution size of C_3 is 64 × 64. For each horizontal connection path, a 1 × 1 convolutional layer is used to reduce the dimensionality of each feature map to 256 channels, and obtain $\{P_3^{in}, P_4^{in}, P_5^{in}\}$, corresponding to $\{C_3, C_4, C_5\}$:

$$P_i^{in} = Conv(C_i),\ i = 3, 4, 5 \tag{7}$$

where *Conv* represents a 1 × 1 convolution operation, which is used to reduce the dimensionality of the output features of each stage to 256 channels.

We use a top-down, horizontal connection method to construct a top-down pathway. A top-down path can produce features with higher resolution and rich semantic features, which are enhanced from top to bottom through a horizontally connected path. Each horizontal connection merges feature maps of the same size into one stage. The top-down feature fusion process of traditional FPN can be expressed as:

$$P_i = P_i^{in} \oplus Resize(P_{i+1}),\ i = 3, 4 \tag{8}$$

$$P_5 = P_5^{in} \tag{9}$$

where $\oplus$ is the feature fusion operation, and *Resize* is the upsampling operation to match the resolution of the feature image to be fused in the lower layer.

Bottom-up branch: This work does not reuse the features extracted from the backbone network in the bottom-up branch-like PANet, but innovatively uses the multi-scale feature enhancement module described in Section 3.3 as the input of the bottom-up branch. Since each branch of the auxiliary feature enhancement module has a different receptive field, the features passing through the structure can be used as input to enrich multi-scale shallow context information.

This paper uses the output feature N_2 of the feature enhancement module as the first extraction layer, which has the same spatial resolution as C_2. The spatial resolutions of the subsequent $\{N_3, N_4, N_5\}$ correspond to $\{P_3, P_4, P_5\}$.inally, C_6 and C_7 are reduced to 256 channels for feature fusion with N_6 and N_7 respectively. There is no down sampling process at this stage. Similar to the top-down branch, the bottom-up branch still adopts a horizontal connection method, using $\{P_3, P_4, P_5, C_6, C_7\}$ as input to obtain the output feature N_i, represented as:

$$P_i^{in} = Conv(C_i),\ i = 6, 7 \tag{10}$$

$$N_i = Resize(N_{i-1}) \oplus P_i,\ i = 3, 4, 5 \tag{11}$$

$$N_i = N_{i-1} \oplus P_i^{in},\ i = 6, 7 \tag{12}$$

where $\oplus$ represents the feature fusion operation; *Conv* represents the 1×1 convolution operation, which is used to reduce the dimensionality of the output features of this stage to 256 channels; *Resize* represents the down sampling operation, and the purpose is to match the resolution of the upper layer feature map.

4. Experiments

To prove the effectiveness of FE-RetinaNet, we conduct experiments on the MS COCO. The experimental conditions in this article are configured as follows: CPU is Intel i7-9700k; memory is 32 G; GPU is NVIDIA GeForce GTX TITAN X; deep learning framework is Pytorch 1.7.1; CUDA version is 10.1.

4.1. Datasets and Evaluation Metrics

The MS COCO is a huge data set composed of 80 different object categories, with a large number of small targets (41% of all target objects), which is especially suitable for evaluating the detection effect of small targets. According to the general classification criteria, this work takes MS COCO trainval35k as the training set (including 80k training and 35k verification images), minival as the verification set (composed of the remaining 5k verification set), and test-dev as the test set to evaluate the model.

In the experiment of the MS COCO, the evaluation metrics used is mean Average Precision (mAP). For IoU, the step size is 0.05 and the setting is from 0.5 to 0.95. For different IoU, we calculate the corresponding mAP, and finally calculate the average of all mAPs. AP_{50} refers to AP at IOU = 0.5, and AP_{75} refers to AP at IOU = 0.75. Objects with a ground truth area smaller than 32×32 are regarded as small targets, objects larger than 32×32 and smaller than 96×96 is regarded as medium targets, and objects larger than 96×96 are regarded as large targets. Finally, the corresponding accuracy rates of objects

of different sizes are counted: small target accuracy rate is regarded as AP_S, the medium target accuracy rate is regarded as AP_M, and large target accuracy rate is regarded as AP_L.

4.2. Experiments on COCO Object Detection

The training of the model in this work uses batch stochastic gradient descent (SGD) to optimize the loss function. The momentum parameter is set to 0.9, the batch size is set to 32, and the initial learning rate is set to 0.001. For the first 160k iteration, we use a learning rate of 10^{-3}, then 10^{-4} for the 60k iteration, and 10^{-5} for the last 20k.

Table 1 shows the comparison between the proposed algorithm and other algorithms on the MS COCO test-dev. Similar to the baseline models, we selected ResNet-50 and ResNet-101 as the backbone.

Table 1. Comparisons of FE-RetinaNet with other methods on MS COCO test-dev set.

Method	Backbone	Input Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods								
Faster R-CNN	VGG-16	~1000 × 600	24.2	45.3	23.5	7.7	26.4	37.1
Faster R-CNN w FPN	ResNet-101	~1000 × 600	36.2	59.1	39.0	18.2	39.0	48.2
Cascade R-CNN	ResNet-101	~1280 × 800	42.8	62.1	46.3	23.7	45.5	55.2
CoupleNet	ResNet-101	~1280 × 800	34.4	54.8	37.2	13.4	38.1	50.8
R-FCN	ResNet-101	~1000 × 600	29.9	51.9	-	10.8	32.8	45.0
Mask R-CNN	ResNet-101	~1280 × 800	38.2	60.3	41.7	20.1	41.1	50.2
One-stage methods								
YOLOv2	DarkNet-19	544 × 544	21.6	44.0	19.2	5.0	22.4	35.5
SSD513	ResNet-101	513 × 513	31.2	50.4	33.3	10.2	34.5	49.8
RetinaNet	ResNet-50	~832 × 500	32.5	50.9	34.8	13.9	35.8	46.7
RetinaNet	ResNet-101	~832 × 500	34.4	53.1	36.8	14.7	38.5	49.1
YOLOv3	DarkNet-53	608 × 608	33.0	57.9	34.4	18.3	35.4	51.1
RefineDet	VGG-16	512 × 512	33.0	54.5	35.5	16.3	36.3	44.3
DSSD513	ResNet-101	513 × 513	33.2	53.3	35.2	13.0	35.4	51.1
RFBNet	VGG-16	512×512	33.8	54.2	35.9	16.2	37.1	47.4
RFBNet-E	VGG-16	512 × 512	34.4	55.7	36.4	17.6	37.0	47.6
EfficientDet-D0	EfficientNet	512 × 512	34.6	53.0	37.1	-	-	-
EFIP	VGG-16	512 × 512	34.6	55.8	36.8	18.3	38.2	47.1
Ours								
FE-RetinaNet	ResNet-50	512 × 512	34.2	52.8	37.1	16.8	37.2	47.6
FE-RetinaNet	ResNet-101	512 × 512	36.2	56.4	39.3	18.0	39.7	49.9

Most two-stage detection algorithms rely on larger input images to improve the performance of the model. In the single-stage detection algorithm, this paper takes the input of ~500 × 500 for comparison. For an input of 832 × 500, the standard RetinaNet obtained an experimental result with an AP of 34.4%. For the detection results (AP_L) of large targets, RetinaNet obtained an excellent result with an AP of 49.1%, but its detection accuracy (AP_S) for small targets was not satisfactory, with an AP_S of 14.7%. The algorithm in this work uses the improved ResNet-101 as the backbone network, surpassing the RetinaNet with ResNet-101 as the backbone network in all indicators, and obtains an AP gain of 1.8%. More importantly, this algorithm has achieved better results in the recognition of small targets, and the AP_S have increased by 3.2%.

4.3. Ablation Research

To verify the effectiveness of the strategy proposed in FE-RetinaNet, this paper studies the impact of the proposed improvements (ResNet-D, bidirectional feature pyramid network, and MFEM) on detection performance. ResNet-50 is used as the backbone network. An ablation study was conducted on the COCO minival, and the results are summarized in Table 2.

Table 2. Comparisons of the algorithm effect of incorporating different modules.

RetinaNet	ResNet-D	Bidirectional FPN	MFEM	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√				32.3	50.6	34.5	13.7	35.5	46.3
√	√			32.8	51.3	35.2	14.4	35.9	46.8
√	√	√		33.5	51.9	36.1	15.1	36.1	47.1
√	√	√	√	34.1	52.7	36.9	16.7	37.0	47.5

It can be seen that the backbone improvement strategy and feature enhancement module proposed in this work can effectively improve the detection performance of the algorithm, and the AP increased by 0.5% and 0.6%, respectively. It is worth noting that the improvement proposed by the algorithm in this work has achieved greater gains in the detection of small targets. After using ResNet-D as the backbone network, APs increased by 0.7%; the integration of feature enhancement module MFEM increased APs by 1.6%. Additionally, the overall performance of the algorithm is increased from 32.8% to 33.5% by incorporating the bidirectional feature pyramid network. In this experiment, the top-down branch reuses the backbone network features without adding the MFEM structure.

To further prove the effectiveness of the auxiliary multi-scale feature enhancement module, this paper sequentially adds branches with different expansion rates and performs experiments on the COCO minival data set. The results are shown in Table 3.

Table 3. Comparisons of adding branches with different expansion rates.

FE-RetinaNet	r1 = 1	r2 = 2	r3 = 3	r4 = 4	AP	AP_S	AP_M	AP_L
√	√				33.6	15.3	36.2	47.0
√	√	√			33.8	15.8	36.5	47.2
√	√	√	√		34.1	16.4	36.8	47.3
√	√	√	√	√	34.2	16.8	37.0	47.5

The accuracy of small target detection (AP_S) is improved by 0.5%, 0.6%, and 0.4%, respectively, with different expansion rate branches. When four branches with different expansion rates are added to MFEM at the same time, the optimal result of AP is 34.2%, and the accuracy of AP_S on small targets is 16.8%. This shows that the MFEM structure can improve the detection effect of small targets by acquiring the shallow context features, and the auxiliary multi-scale feature enhancement module can significantly improve the performance of small target detection.

4.4. Small Target Detection Performance Comparision

In order to show the improvement of detection performance more intuitively, FE-RetinaNet and standard RetinaNet are used to detect several selected images in the MS coco dataset, and the results are shown in Figure 7. Specifically, we build the proposed model using deep learning framework PyTorch. Our model is trained on Nvidia GeForce GTX TITAN X GPUs with CUDA 10.1 and CUDNN 7.5. We use the improved model after training to compare with the original model. Comparisons of RetinaNet and FE- RetinaNet small target detection results on MS COCO as follows.

In the comparison diagram, (a) is the original image, (b) is the standard RetinaNet test results, (c) is the improved method test results. In the first row, RetinaNet mis-detects the streetlamp as a boat, and the improved FE-RetinaNet successfully avoids this error. From the first to the fifth rows, it can be seen that the FE-RetinaNet detects small targets more accurately than the original algorithm, and more small targets are detected. Among them, the fifth row of boat targets has more background environment interference. It can be seen that the improved FE-RetinaNet target detection algorithm can more accurately detect and classify small and medium targets in the input image. In complex scenes, FE-RetinaNet can detect more small targets compared with the original RetinaNet.

(**a**) (**b**) (**c**)

Figure 7. Comparisons of RetinaNet and FE- RetinaNet small target detection results on MS COCO. (**a**) Original image; (**b**) Standard RetinaNet test results; (**c**) FE- RetinaNet test results.

5. Conclusions

Aiming at the problem of RetinaNet's poor detection effect on small targets, this work proposes an improved FE-RetinaNet target detection algorithm. First of all, to address the problem of feature loss of small targets in high-level feature maps, this work introduces an improved backbone network. Besides, this work also constructs a parallel feature enhancement module to generate multi-scale context features. Then, the feature enhancement branch is combined with the improved backbone network to form a bidirectional feature pyramid network to enhance the transmission efficiency of shallow features. The algorithm is tested on the MS COCO. The experimental results show that the detection effect of the improved method has been greatly improved, and the mAP on the COCO has increased by 1.8%. The integrated auxiliary multi-scale feature enhancement module effectively improved the detection effect of small targets, and the APs on the COCO increased by 3.3%.

Author Contributions: Conceptualization, H.L.; Funding acquisition, M.S.; Writing—original draft, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research on the Construction of Multi-scale Concept Lattice and Knowledge Discovery Method, grant number 61673396.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in MS COCO at https://doi.org/10.1007/978-3-319-10602-1_48 accessed on 26 May 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
2. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497. [CrossRef] [PubMed]
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
5. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *arXiv* **2016**, arXiv:1605.06409.
6. 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*; Springer: Cham, The Netherlands, 2016; pp. 21–37.
7. 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, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
8. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
9. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
10. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
11. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
12. 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, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
13. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*; Springer: Cham, The Netherlands, 2016; pp. 630–645.
16. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Detnet: A backbone network for object detection. *arXiv* **2018**, arXiv:1804.06215.
17. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
18. Wang, T.; Anwer, R.M.; Cholakkal, H.; Khan, F.S.; Pang, Y.; Shao, L. Learning rich features at high-speed for single-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1971–1980.
19. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
20. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. *arXiv* **2017**, arXiv:1701.04128.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
22. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
23. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
24. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.

25. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
26. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.