Hindawi
Scientific Programming
Volume 2022, Article ID 3583399, 10 pages
https://doi.org/10.1155/2022/3583399

Research Article

Object Detection Algorithm Based on Improved Feature Pyramid

Bai Yu,[1] **Xuhua Pan**,[1] **Xuefeng Li**,[2] **Gaohua Liu**,[2] **and Yunpeng Ma**[1]

[1]*School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China*
[2]*College of Electrical Automation and Information Engineering of Tianjin University, Tianjin 300072, China*

Correspondence should be addressed to Yunpeng Ma; mayunpeng@tjcu.edu.cn

Received 18 February 2022; Revised 21 July 2022; Accepted 24 August 2022; Published 13 September 2022

Academic Editor: Dongpo Xu

Feature pyramid network is widely used in advanced object detection. By simply changing the network connection, the performance of small object detection can be greatly improved without increasing the amount of calculation of the original model. However, the algorithm still has some shortcomings. Therefore, a new attention feature pyramid network (attention PFN) is proposed. Firstly, an improved receptive field module is added, which can make full use of global and local context information. Secondly, the connection mode of the pyramid is further optimized. Deconvolution is used to replace the nearest neighbor interpolation in top-down up-sampling and a channel attention module is added to the horizontal connection to highlight important context information. Finally, adaptive fusion and spatial feature balance are used for each feature pyramid, so that the network can learn the weights of different feature layers. Each pyramid layer contains more discrimination information. Attention PFN is tested on Pascal VOC and MS COCO datasets, respectively. The experiment results revealed that the proposed method has better performance than the original algorithm. Therefore, the attention PFN is an effective algorithm.

1. Introduction

In recent years, object detection algorithms based on deep learning have been widely used in various fields, such as face detection, vehicle-pedestrian detection, dangerous goods detection [1], transmission line defect identification [2], and so on. These object detection algorithms can be divided into two main categories. The first type is a two-stage algorithm based on a region proposal network, including Faster R-CNN [3], R-FCN [4], and others. The other type is a regression-based one-stage algorithm, such as YOLO [5], SSD [6], and others. However, no matter which type of algorithm it is, it faces the difficult problem of poor detection for small targets. In order to solve the problem of small target detection, the Feature Pyramid Network (FPN) was proposed in literature [7], which has been widely used in the two-stage detection algorithm (such as Mask-RCNN [8]) and one-stage detection algorithm (such as RetinaNet [9]).

Based on the structure of ConvNet, the FPN adopts a top-down and horizontal connection method to transfer the top layer semantic information to the low layer. However, the top layer information will be lost. Features of different scales contain information from different abstract levels, so there are large semantic differences in direct addition, and the formed pyramids at all layers are not further fused. Therefore, various variants of the Feature Pyramid Network have been proposed.

In literature [10], all information of a multilevel structure was used to generate a multilevel contextual features pyramid with multiple scales. In literature [11], authors proposed a global information extractor and a local information extractor as a framework for single-stage object detection. Yang et al. [12] used a multilayer feature map stacking method to fuse semantic information and detailed features. PANet [13] added an additional bottom-up path on the basis of FPN to transfer low layer location information to high layer. Libra-RCNN [14] designed a balanced feature pyramid; each layer of the pyramid has the information of all layers. NAS-FPN [15] used neural structure to find irregular characteristic network topology and then repeatedly applied the same block, achieving outstanding results. EfficientDet [16] used EfficientNet as the feature extraction network to design the BiFPN, which has an efficient two-way cross-scale connection and weighted feature fusion with a better precision and efficiency trade-off.

This paper proposes an improved feature pyramid network based on the attention mechanism and receptive field. The improved receptive field module (ARFB) is added to the top of the feature extraction layer to obtain global and local context information. In the next place, we improved the connection method of feature layers with different resolutions. For the top-down up-sampling method, the deconvolution method is used to replace the nearest neighbor-interpolation for reducing information loss. In the horizontal connection, a channel attention mechanism is added to the feature layer of the backbone network before the 1×1 convolution, which can reduce the secondary channel to emphasize important channel information. Finally, the spatial adaptive fusion and balance mechanism are adopted for each layer of the pyramid. The weights of the fused features at all layers can be learned as well as information on each layer of the feature pyramid is fully enhanced.

The principal contributions of the paper are summarized as follows:

(1) The top of the feature extraction layer adopts the ARFB module to obtain ample global and local context information.

(2) Deconvolution is used for up-sampling to reduce the loss of feature information. SE module is added at the horizontal connection of the feature layer of the backbone network to extract effective feature information and reduce the interference of noise information.

(3) The spatial adaptive fusion and balance mechanism is applied to further refine the feature information of each feature layer. It greatly improves the detection ability of the network.

2. Fundamental Knowledge

FPN has two paths: the bottom-up path and the top-down path. The bottom-up path is the feed-forward calculation of networks, such as ResNet [17]. The output of each stage is $\{C_1, C_2, C_3, C_4, C_5\}$, the width and height are halved in turn, and the number of channels is {64, 256, 512, 1024, 2048}.

In the top-down path, the output C5 reduces the number of channels to 256 by a 1×1 convolution, formingP_5. Subsequently, up-sampling is performed by the nearest neighbor interpolation method, and C4 with the same resolution is added to it after a 1×1 convolution becomes P4. Repeating the above operations, a top-down path composed of {P5, P4, P3, and P2} is formed, and they all use a further 3×3 convolution to avoid the aliasing effect caused by up-sampling. In addition, P5 performs max pooling operation to obtain P6. When FPN is combined with Faster R-CNN, a multiscale detection method is adopted. The RPN network searches for the prospect target proposed area in {P2, P3, P4, P5, P6}, and according to the size of the ROI, select the feature map in {P2, P3, P4, P5} to perform the Fast R-CNN operation to obtain the specific target category and more precise location.

However, the FPN has several drawbacks: (1) C5 is located at the top layer. With the deepening of neural networks, the feature extracted by deep convolution has low spatial resolution and position information that can be easily recognized. P5 is formed by channel compression, and then P5 is pooled to form P6, which will cause further loss of characteristic information. (2) The features of different layers contain complex feature information. In the horizontal connection, the direct addition of features at different layers will contain a lot of background noises. It will result in great semantic ambiguity. (3) The formed pyramids at all layers are not further fused effectively.

A novel neural network is proposed in this section called attention FPN, the structure of which is shown in Figure 1. There are three main improvements in the proposed method: (1) An improved receptive field module is added between C5 and P5 to obtain more context information and supplement P5 and P6. (2) The channel attention mechanism is added to the horizontal connection to enhance the effective position information in the feature map in the path and suppress irrelevant background information. At the same time, we also optimize the up-sampling method. (3) The pyramids at all layers have added the adaptive spatial fusion and balance mechanism. The weight of each layer is learned by spatial attention, and the information of all layers of the pyramid makes a more reasonable balance and enhancement.

2.1. Improved Receptive Field Enhancement Module.

Dilated convolution comes from DeepLab [18], which can increase the receptive field without reducing the scale of the feature map and perform well in semantic segmentation. Literature [19] designed a receptive field block RFB in the SSD network, and this structure adopted the multibranch idea of Inception-ResNet [20]. There are different convolution kernels and expansion coefficients on different branches, corresponding to various sizes of targets and receptive fields; their combinations can make full use of context information. As shown in Figure 2, the RFB module can only obtain local context information, so the global context information module GCM is introduced into it as a branch so that global context information and local context information can be well fused; the improved module is named ARFB (augment receptive field block). The specific structure of GCM is shown in Figure 3. First of all, the feature map is globally averaged and pooled to obtain the global feature vector. Then, up-sampling is performed to restore the original size and reduce the number of channels to splice with other branches. In order to prevent information loss, we added residual connections in the input and output feature maps . C5 is connected to P5 through the ARFB module, which can give P5 and P6 rich context information to overcome the deficiency of FPN in large target recognition.

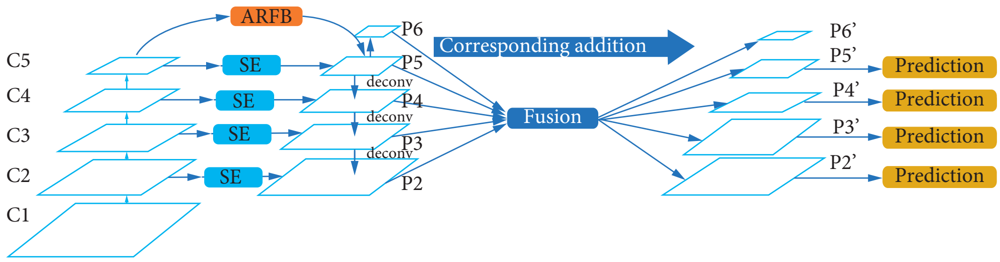


Figure 1: Attention FPN network structure.

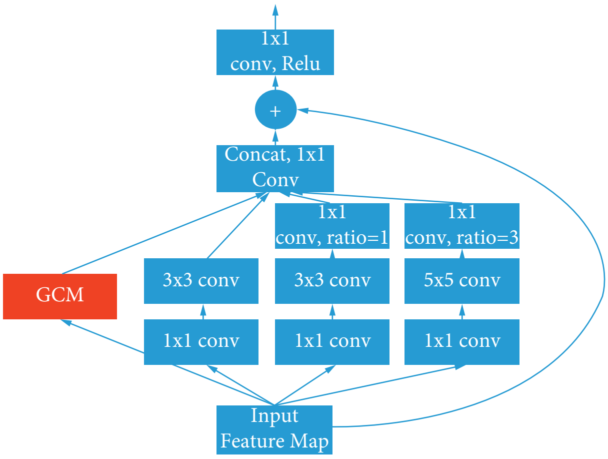


Figure 2: augment receptive field block.

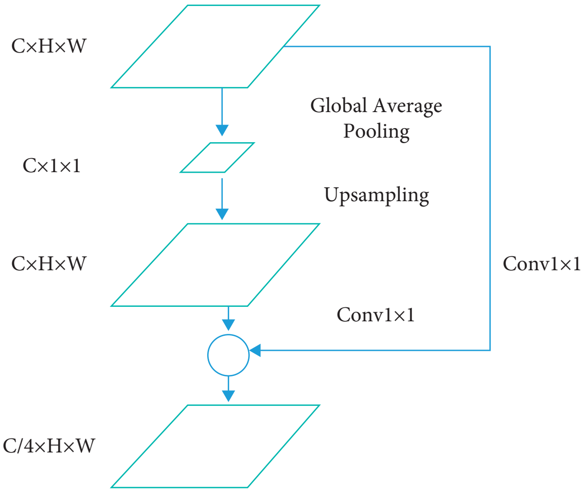


Figure 3: Global context module.

2.2. Improved Top-Down Path and Horizontal Connection. As shown in Figure 4, for the top-down path, the original interpolation method lacks flexibility and learning ability, resulting in poor performance. Therefore, the deconvolution method is used to replace nearest neighbor interpolation for improving resolution. Moreover, the deconvolution can adjust the training hyperparameter, which can better express the relationship between different layers of features.

In the horizontal connection, the spatial attention SE [21](Squeeze-and-Excitation Networks) module is introduced, which is shown in Figure 5. On the bottom-up path, the SE module is added to enhance the important information in the channel, reducing each feature layer channel to 256. The SE module mainly contains two parts, the input feature map is $U = [u_1, u_2, \ldots, u_c]$, and each channel is $u_c \in R^{H\times W}$. First of all, the global average pooling is used to obtain the global vector Z, and the calculation formula is shown as follows:

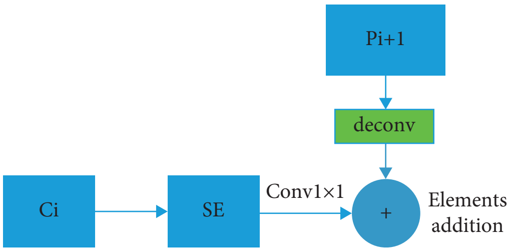


Figure 4: Improved pyramid connection structure.

$$Z_c = \frac{1}{(H\times W)} \sum_{i}^{H} \sum_{j}^{W} u_c(i, j). \tag{1}$$

Then, Z is encoded by using the fully connected network to obtain the dependency between each channel and complete the space compression. The calculation formula is as follows:

$$\widehat{Z}\delta(W_1(W_0(Z))), \tag{2}$$

where $W_0 \in R^{c\times c/2}, W_1 \in R^{c/2\times c}$, δ is the ReLU activation function, $\widehat{Z}_c$ is obtained through the sigmoid activation function, and the value range is adjusted to [0, 1], which is multiplied by the input feature map to obtain the feature after channel excitation:

$$\widehat{U}_{\mathrm{SE}} = \left[\sigma\left(\widehat{Z}_1\right)u_1, \sigma\left(\widehat{Z}_2\right)u_2, \ldots, \sigma\left(\widehat{Z}_k\right)u_k\right], \tag{3}$$

$\sigma(\widehat{Z}_c)$ can be regarded as the importance of each channel. When the network is learning, this module is adaptively tuned well to enhance important channels and suppress irrelevant channels. This module can merge with the feature layer of the top-down path well. Due to the small amount of calculation of the channel attention module, the increased time cost is negligible.

2.3. Adaptive Spatial Feature Fusion and Balance Mechanism. Different feature pyramids have different contributions to various target recognition. Inspired by AugFPN [22] and Libra R-CNN, we propose a spatial adaptive fusion and

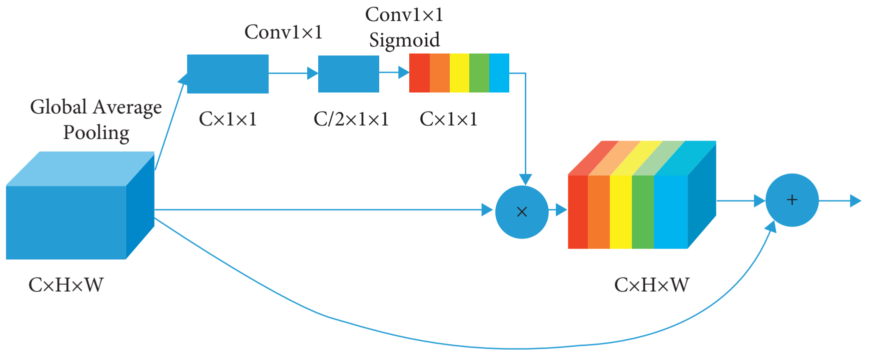


Figure 5: Squeeze-and-Excitation Networks.

balance mechanism. The mechanism is shown in Figure 6. First, {P2, P3, P4, P5, P6} is integrated into the same scale by adopting up-sampling and down-sampling. They were merged into a branch through concatenation operation subsequently. The other branch remains unchanged. The combined features reduce the number of channels by 1 × 1 convolution. It is performed by 3 × 3 convolutions for feature extraction. Then, sigmoid is used to obtain the weight on each channel. The channels perform split operations. The weight of each feature layer is multiplied by the corresponding feature layer to extract effective features. Finally, obtained feature layer performs concatenation operation and reduces the dimension through 1 × 1 convolution. The subsequent steps are the same as Libra R-CNN. The fused features are, respectively, scaled to the original size, and the corresponding original features are added to avoid the loss of some detailed information. The enhanced feature pyramid {P2′, P3′, P4′, P5′, P6’} is formed. The location information and semantic information of each feature layer have been fully enhanced. It is conducive to the detection of multisize targets.

3. Experimental Results and Analysis

The environment configuration of these experiments is presented as follows: CPU model is Intel Sliver 4210; GPU model is NVIDIA GTX 2080Ti and a total of 4 blocks. The operating system is Ubuntu16.04; the acceleration library is CUDA 10.0 and cuDNN 7.6.5. Based on the deep learning framework of PyTorch 1.4 and the implementation network of python 3.7, experimenting them is carried out separately on the PASCAL VOC dataset and MS COCO dataset.

3.1. PASCAL VOC Experiment Results. The PASCAL VOC datasets have a total of 20 categories of targets. We use the training validation sets including 16551 pictures of PASCAL VOC 2007 and PASCAL VOC 2012 for training and the testing set of PASCAL VOC 2007 including 4952 pictures. Stochastic gradient descent (SGD) is used for training, momentum is set to 0.9, the input image is re-sized to 1000 × 600, the batch size is set to 16, the weight attenuation coefficient is set to 0.0005, the initial learning rate is set to 0.02, and the maximum number of iterations epoch is 5. At epoch 3, the learning rate is multiplied by 0.1.

The results of our algorithm on the VOC2007 testing set are shown in Table 1 (with ∗ indicating the reimplemented version of PyTorch). We use Faster R-CNN + FPN as the baseline. It can be seen from the result that when ResNet-50 and ResNet-101 are used as feature extraction networks, respectively, we use attention FPN (abbreviated as AFPN in the table) to replace FPN, which not only increases mAP by 2.8% and 1.9%, respectively, but also has a small speed loss. The accuracy of ResNet-50 exceeds that of most algorithms in the table, while the recognition accuracy of ResNet-100 is the highest among all algorithms. Table 2 shows the specific results of 20 categories. Our model obtains the best accuracy in multiple categories. Compared with FPN, attention-FPN has an obvious improvement effect on small targets (such as monitors, water cups, bottles, cows, sheep, and so on).

In order to further study the influence of the added module on the detection effect of the algorithm, we do an ablation experiment. As shown in Table 3, all experiments were carried out on FPN (ResNet-50). The mAP of the original FPN could achieve 79.4%. The RFB module was embedded in the benchmark network for training and testing. The mAP of 80.7% was obtained. The RFB is replaced by the proposed ARFB. The map can reach 80.9%. It proves that ARFB further expands the receptive field and improves detection abilities. The SE module is added to the network based on the embedded RFB. The mAP is improved by 0.6% compared to the embedded RFB only. The SE module can extract effective features and reduce the impact of background information. Next, the spatial adaptive fusion and balance mechanism continue to be added to the network. The mAP can reach 81.8%. The experimental results demonstrate that the spatial adaptive fusion and balance mechanism can significantly intensify the connection between different feature layers and enhance the detection ability. Then, the deconvolution operation is used to replace the nearest neighbor for up-sampling. The improvement effect is not obvious. The possible reason for this phenomenon is that deconvolution will blur the feature of the input. It leads to the gain brought by deconvolution being limited. Lastly, the enhanced receptive field module ARFB is used to replace RFB. The mAP is further increased by 0.3% to 82.2%, which also proves the effectiveness of global context information.

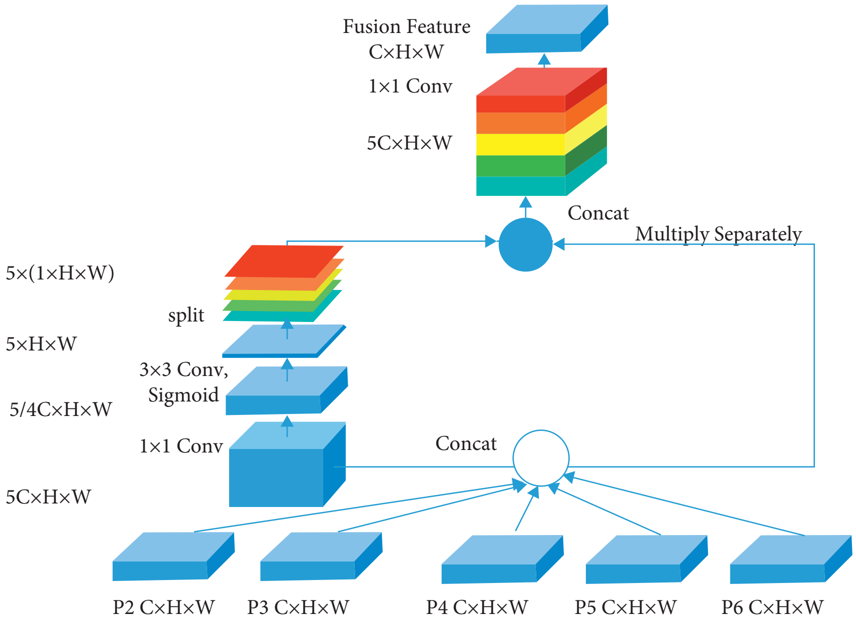


Figure 6: Spatial adaptive feature fusion mechanism.

Table 1: PASCAL VOC 2007 test detection results.

	Two-stage				One-stage		
Method	Backbone	mAP (%)	FPS	Method	Backbone	mAP (%)	FPS
Faster R-CNN	VGG-16	73.2	7	FA-SSD [24]	ResNet50	78.1	30
Faster R-CNN	ResNet-101	76.4	2.4	EFIPNet512 [25]	VGG-16	81.8	28
ION	VGG-16	76.5	1.25	SSD512	VGG-16	79.5	19
MR-CNN	VGG-16	78.2	0.03	DSSD321	ResNet-101	78.6	9.5
Faster R-CNN*	ResNet-50-FPN	79.4	19.1	DSSD512	ResNet-101	81.5	5.5
Faster R-CNN*	ResNet-101-FPN	81.6	16.0	DFPR512 [26]	VGG-16	81.1	34
R-FCN	ResNet-101	80.5	9	RefineDet320 [27]	VGG-16	80.0	40.3
CoupleNet[23]	ResNet-101	82.7	8.2	RefineDet512	VGG-16	81.8	24.1
Ours	**ResNet-50-AFPN**	**82.2**	16.8	RFB Net 300	VGG-16	80.5	83
Ours	**ResNet-101-AFPN**	**83.5**	14.1	RFB Net 512	VGG-16	82.2	38

Figure 7 shows the qualitative results of the comparative experiment. The left side of each picture is the detection result of FPN (ResNet50), and the right side is the detection result of attention FPN (ResNet50). It can be seen that the improved network has fewer redundant frames in terms of large targets because the expanded receptive field makes the ARFB module supplement effective global information for P5 and P6, so the detection of large targets is more accurate. In terms of small targets, ours has a higher recall rate, thanks to the fact that the attention mechanism filters out irrelevant information, giving the feature fusion mechanism more detailed information and semantic information at the bottom.

In order to directly reflect the advantages of the proposed method, the visualization results of some feature layers are displayed in Figure 8. The *a* is the original picture. The *b* is the visualization result of the C5 layer of FPN. The *c* is the visualization result of the C5 layer of FPN which introduces the RFB module. The *d* shows the visualization result of the C5 layer which introduces the SE module and RFB module. The *e* is the visualization result of the P5 layer after spatial adaptive fusion and balance mechanism. The *f* is the visualization result of our proposed methods. We can observe that the proposed methods can lead the networks to pay more attention to object features while ignoring the influence of background noise.

3.2. MS COCO Experiment Results.

MS COCO 2017 dataset is the authoritative dataset currently used to evaluate the performance of target detection algorithms. The data include a total of 80 target objects, of which the training set contains 118287 pictures, the verification set contains 5000 pictures, and the testing set contains 20288 pictures. In COCO, there are more small objects than large objects, of which about 41% are small objects (area <322), 34% are medium objects (322 < area <962), and 24% are large objects (area >962). The measured area is the number of pixels in the segmentation

Table 2: PASCAL VOC 2007 test detection results by categories.

	mAP	Areo	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	M-bike	Person	Plant	Sheep	Sofa	Train	Tv
Faster(Res101)	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	89.8	56.7	87.8	69.4	88.3	80.9	78.4	41.7	78.6	79.8	85.3	72.0
R-FCN	80.5	79.9	87.2	81.5	72	69.8	86.8	88.5	89.8	67	88.1	74.5	**89.8**	79.9	81.2	53.7	81.8	**81.5**	85.9	79.9
RefineDet320	80.0	83.9	85.4	81.4	75.5	60.2	86.4	88.1	89.1	62.7	83.9	77.0	85.4	86.7	82.6	55.3	82.7	78.5	88.1	79.4
RFBNet300	80.5	83.7	87.6	78.9	74.8	59.8	**88.8**	87.5	87.9	65.0	85.0	77.1	86.1	86.6	81.7	58.1	81.5	81.2	**88.4**	80.2
SSD300	77.5	79.5	83.9	76	69.6	50.5	87	85.7	88.1	60.3	81.5	77	86.1	83.9	79.4	52.3	77.9	79.5	87.6	76.8
DSSD321	78.6	81.9	84.9	80.5	68.4	53.9	85.6	86.2	88.9	61.1	83.5	**78.7**	86.7	86.7	79.7	51.7	78	80.9	87.2	79.4
FPN(Res50)*	79.4	82.3	86.9	84.5	68.0	69.7	85.6	87.5	87.9	62.9	85.6	71.9	87.9	82.9	85.1	54.0	81.4	77.1	84.6	76.2
Ours(Res50)	82.2	86.7	87.3	84.8	75.1	72.0	86.4	88.7	89.5	67.9	87.1	77.7	88.7	85.6	86.3	59.0	84.2	78.6	87.4	82.3
Ours(Res101)	**83.5**	**87.7**	**88.5**	**85.3**	**76.9**	**72.7**	88.2	**89.3**	**90.0**	**71.0**	**88.4**	78.1	**89.8**	**86.8**	**87.1**	**60.6**	**87.2**	81.4	87.3	**83.8**

Table 3: Effects on mAP using different modules.

	RFB	ARFB	SE	Deconv	Fusion	mAP
FPN (ResNet-50)						79.4
	√					80.7
		√				80.9
	√		√			81.3
	√		√		√	81.8
	√		√	√	√	81.9
Ours		√	√	√	√	**82.2**

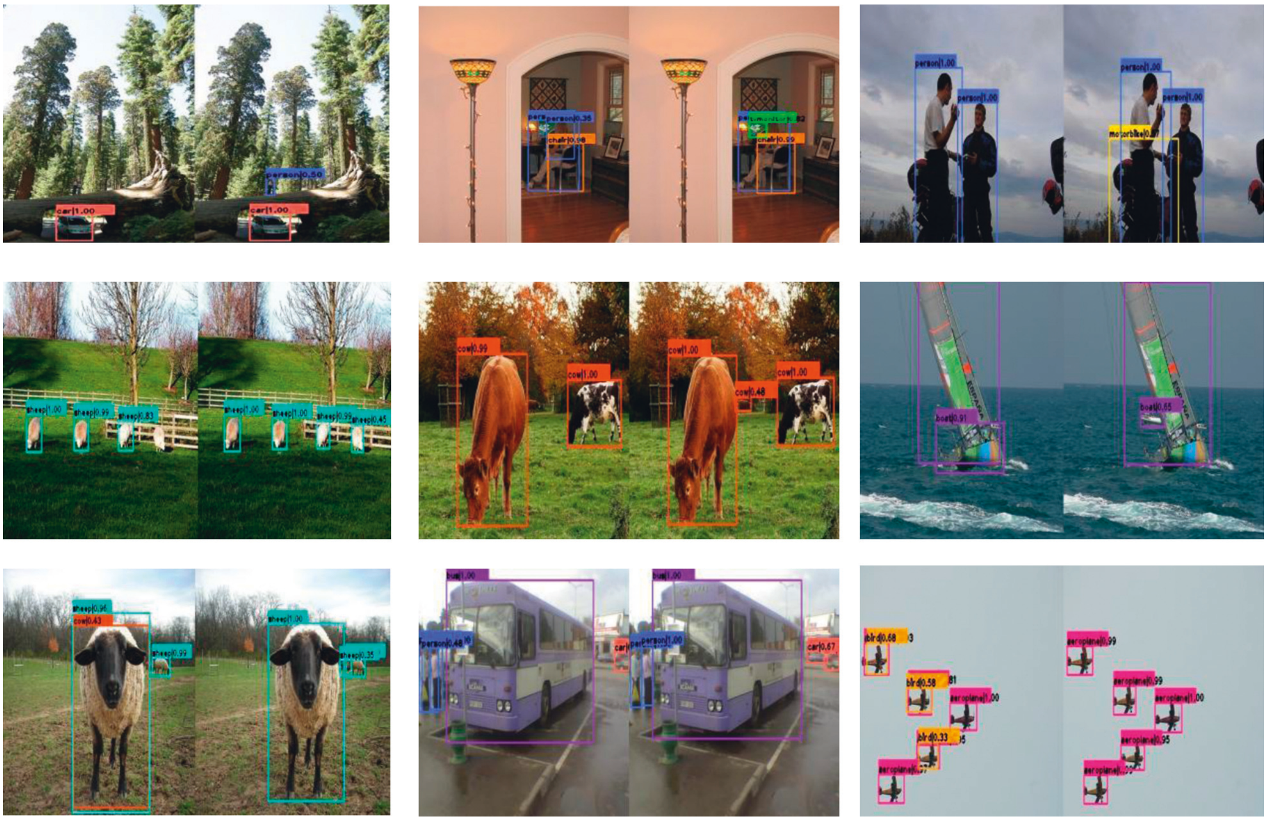
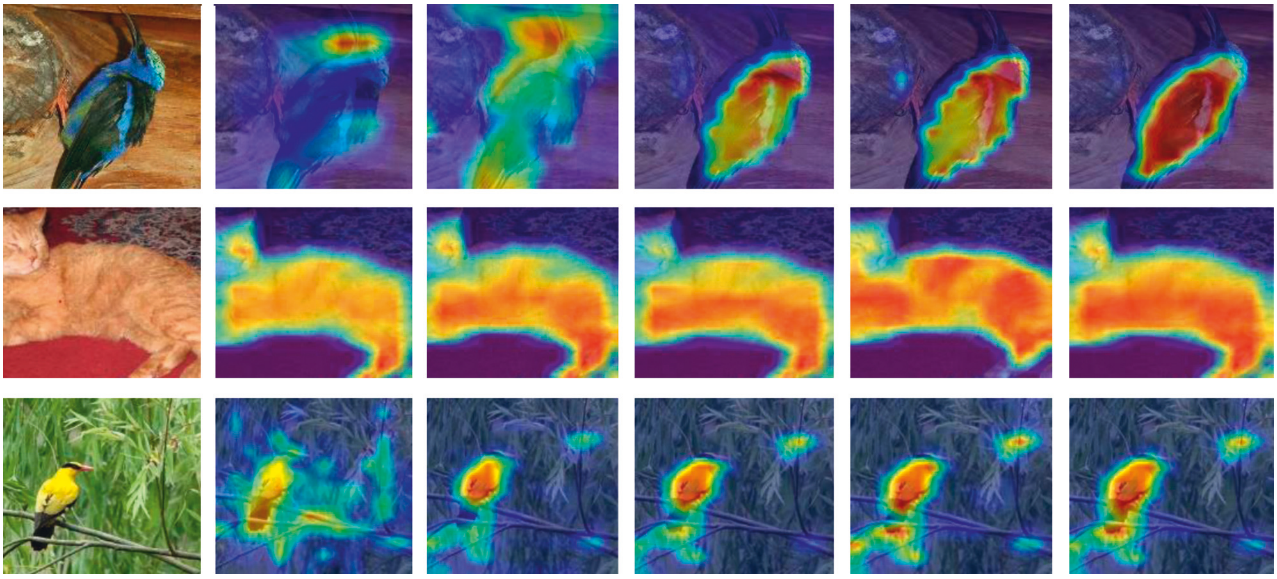


Figure 7: Test result comparison on PASCAL VOC 2007 datasets.

mask. The corresponding evaluation indicators are divided into AP_S, AP_M, and AP_L. $AP_{0.5}$ and $AP_{0.75}$, respectively, represent the average detection accuracy of all categories when the intersection ratio IoU is 0.5 and 0.75. AP represents all 10 IoU thresholds (0.5 to 0.95) and the average of all 80 categories, and it is considered the most important indicator. We train the model on the COCO 2017 training, testing the experimental results on val2017 and test2017-dev, respectively.

The stochastic gradient descent (SGD) method is used for training, the momentum of which is set to 0.9, the weight attenuation coefficient is set to 0.0005, the input image is resized to 1333 × 800, each GPU is assigned two images, and the batch size is set to 8. The maximum number of iterations epoch is 12, and the initial learning rate is set to 0.01, which is multiplied by 0.1 when the epoch is 8 and 11.

As shown in Table 4, on val2017, we choose three backbones: ResNet-50, ResNet-101, and the more powerful ResNeXt-101 [28](32 × 4d), which are compared with the baseline. As we all know the same backbone, Faster R-CNN combined with FPN only has advantages in small target recognition compared to the original Faster R-CNN. The detection ability of medium and large targets will be reduced a lot. However, the detection accuracy of attention FPN on objects of different sizes is all improved. As seen from Table 5, on test2017-dev, the test results of the proposed algorithm are also better than the baseline. The AP values of attention FPN of the three feature extraction networks increased by 1.5%, 0.7%, and 0.2% respectively, reaching 37.7%, 39.5%, and 40.6%. It is superior to algorithms such as Mask R-CNN and CoupleNet which also use ResNet-101.



Original picture (a) FPN (b) FPN+RFB (c) FPN+RFB+SE (d) FPN+RFB+SE+Fusion (e) our (f)

Figure 8: The visualization results in different feature layers. (a) Original picture. (b) FPN. (c) FPN + RFB. (d) FPN + RFB + SE. (e) FPN + RFB + SE + Fusion four.

Table 4: COCO val 2017 detection results.

	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_S	AP_M	AP_L
Faster R-CNN*	ResNet-50	36.6	58.5	39.2	20.7	40.5	47.9
Faster R-CNN*	ResNet-101	38.8	60.5	42.3	23.3	43.1	50.3
Faster R-CNN*	ResNet-50-FPN	36.4	58.4	39.1	21.5	40.0	46.6
Faster R-CNN*	ResNet-101-FPN	38.5	60.3	41.6	22.3	43.0	49.8
Faster R-CNN*	ResNeXt-101(32 × 4d)-FPN	40.1	62.0	43.8	23.4	44.6	51.7
Ours	**ResNet-50-AFPN**	**37.5**	59.9	40.5	22.3	41.5	48.1
Ours	**ResNet-101-AFPN**	**39.4**	61.6	42.8	23.4	43.7	50.9
Ours	**ResNeXt-101(32 × 4d)-AFPN**	**40.3**	63.4	43.7	24.1	44.8	52.6

Table 5: COCO test 2017-dev detection results.

	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_S	AP_M	$AP_{\ L}$
Faster R-CNN	ResNet-101	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN*	ResNet-50-FPN	36.2	58.5	38.9	21.0	38.9	45.3
Faster R-CNN*	ResNet-101-FPN	38.8	60.9	42.1	22.6	42.4	48.5
Faster R-CNN*	ResNeXt-101(32 × 4d)-FPN	40.4	62.2	43.6	24.0	43.8	50.3
Mask R-CNN	ResNet-101-FPN	29.9	60.3	41.7	20.1	41.1	50.2
R-FCN	ResNet-101	33.1	51.9	—	10.8	32.8	45.0
CoupleNet	ResNet-101	**37.7**	53.5	35.4	11.6	36.3	50.1
YOLO v2	Darknet-19	21.6	44.0	19.0	5.2	22.4	35.5
YOLO v3	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
SSD512	VGG-16	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RFB Net512	VGG-16	33.8	54.2	35.9	16.2	37.1	47.4
M2Det [29]	VGG-16	33.5	53.4	35.6	14.4	37.6	47.6
IBi-FPN [11]	VGG-16	32	50.9	33.7	12	37	47.1
M2Det [29]	ResNet-101	38.8	59.4	41.7	20.5	43.9	53.4
NETNet [30]	ResNet-101	38.5	58.6	41.3	19.0	42.3	53.9
MREFP-Net [10]	ResNet-101	39.3	60.1	43.1	20.6	43.9	52.0
MREFP-Net [10]	VGG-16	38.5	59.1	41.1	18.6	42.9	49.4
Ours	**ResNet-50-AFPN**	**39.5**	60.3	40.5	22.2	40.8	46.5
Ours	**ResNet-101-AFPN**	**40.6**	62.1	42.9	23.1	43.0	49.5
Ours	**ResNeXt-101(32 × 4d)-AFPN**	**41.8**	64.1	44.1	24.1	44.2	51.4

4. Conclusion

This paper proposes a target detection algorithm based on the improved feature pyramid network, called attention FPN. In the attention FPN, the improved receptive field module is used to obtain the global and local context information; the channel attention module is added to the transverse connection to enhance the characteristics that contribute greatly to the key information. We also used deconvolution instead of nearest neighbor interpolation to reduce information loss in the up-down up-sampling method. Finally, the spatial attention style is used to weigh the integration of the various characteristic layers. The mAP of our improved algorithm on the PASCAL VOC 2007 test set reached 83.5% and the AP on COCO test 2017-dev reached 40.6%; the results show that our algorithm is better than the original algorithm and some mainstream algorithms with a considerable speed.

Nevertheless, there are certain limitations in our proposed algorithm. Compared to the original algorithm, the number of parameters of our proposed algorithm has increased by 10%. In addition, a small amount of background noise still exists, which affects the detection efficiency. In the future, we will reduce the number of parameters of the network and study a more effective mechanism to identify target features. Furthermore, we will combine the proposed algorithm with practical application.

Data Availability

The data that support the findings of this study are openly included within the articles [Lin *T* Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[J]. European Conference on Computer Vision, 2014.] at https://doi.org/10.1007/978-3-319-10602-1_48 and [Koskela M, Laaksonen J, Viitaniemi V. The 2005 PASCAL Visual Object Classes Challenge[J]. 2007.] at https://doi.org/10.1007/s11263-014-0733-5, separately.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of Tianjin (Grant No. 20JCQNJC00430), the Science and Technology Plan Project of Tianjin (Grant No. 19JCTPJC53400), and the Special Fund Project of Tianjin Technology Innovation Guidance (Grant No. 21YDTPJC00370).

References

[1] C. Y. Gao, W. H. Zhao, and M. L. Zhang, "A vehicle bottom dangerous object detection algorithm based on YOLOv3," *Journal of Tianjin University*, no. 4, p. 4, 2020.

[2] X. Li, H. Su, and G. Liu, "Insulator defect recognition based on global detection and local segmentation," *IEEE Access*, vol. 8, pp. 59934–59946, 2020.

[3] S. Q. Ren, K. M. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[4] J. Dai, Y. Li, K. M. He, and J. Sun, "R-fcn:Object Detection via Region-Based Fully Convolutional networks," in *Proceedings of the Neural Information Processing Systems(NIPS)*, pp. 379–387, Barcelona, Spain, December 2016.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look once: Unified, Real-Time Object detection," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[6] 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, Spring Verlag, Amsterdam, Netherlands, September 2016.

[7] T. Y. Lin, D. Piotr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object detection," in *Proceedings of the IEEE Computer Vision and Pattern Recognition(CVPR)*, pp. 2117–2125, IEEE, Hawaii, USA, July 2017.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, December 2017, Article ID 29612969.

[9] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2017.

[10] B. Laa, Md. Sah Bin Haji Salam Fc a, and C. Sa, "Multi-level Refinement Enriched Feature Pyramid Network for Object detection," *Image and Vision Computing*, vol. 115, Article ID 104287, 2021.

[11] T. N. Quang, S. Lee, and B. C. Song, "Object detection using improved bi-directional feature pyramid network," *Electronics*, vol. 10, no. 6, p. 746, 2021.

[12] A. P. Yang, L. Y. Lu, and Z. Ji, "Multi-feature concatenation network for object detection," *Journal of Tianjin University*, no. 6, p. 13, 2020.

[13] S. Liu, L. Qi, H. F. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759–8768, IEEE, Salt Lake City, Utah, October 2018.

[14] J. M. Pang, K. Chen, J. P. Shi, and R. Libra, "Towards balanced learning for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 821–830, IEEE, Long Beach, CA, February 2019.

[15] G. Ghiasi, T. Y. Lin, and Q. Le, "NAS-FPN: learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7036–7045, IEEE, Long Beach, CA, October 2019.

[16] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection[J/OL]," 2019, https://arxiv.org/abs/1911.09070.

[17] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep Residual Learning for Image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, IEEE, Las Vegas, NV, USA, May 2016.

[18] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J/OL]," 2016, https://arxiv.org/abs/1606.00915.

[19] S. Liu and D. Huang, "Receptive Field Block Net for Accurate and Fast Object detection," in *Proceedings of the European Conference on Computer Vision*, pp. 385–400, IEEE, Munich, Piscataway, September 2018.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-Resnet and the Impact of Residual Connections

on learning," in *Proceeding of the AAAI:National Conference on Artificial Intelligence*, pp. 1425–1438, IEEE, Phoenix, Arizona, USA, August 2016.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, IEEE, Salt Lake City, June 2018.

[22] C. X. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "AugFPN: Improving Multi-Scale Feature Learning for Object Detection [J/OL]," 2019, https://arxiv.org/abs/1912.05384.

[23] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu, "CoupleNet: Coupling Global Structure with Local Parts for Object Detection," in *Proceeding of the International Conference on Computer Vision*, pp. 4146–4154, IEEE, Venice, Italy, August 2018.

[24] J. S. Lim, M. Astrid, H. J. Yoon, and S. I. Lee, "Small Object Detection Using Context and Attention," in *Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Jeju Island, Korea (South), April 2021.

[25] Y. Pang, T. Wang, R. M. Anwer, F. S. Khan, and L. Shao, "Effificient featurized image pyramid network for single shot detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7336–7344, Long Beach, CA, USA, June 2019.

[26] T. Kong, F. Sun, C. Tan, H. Liu, and W. Huang, "Deep feature pyramid reconfiguration for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 169–185, 2018.

[27] S. F. Zhang, L. Y. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4203–4212, IEEE, Salt Lake City, Utah, November 2018.

[28] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, IEEE, Honolulu, Hawaii,USA, 2017.

[29] Q. Zhao, T. Sheng, Y. Wang et al., "M2det: a single-shot object detector based on multi-level feature pyramid network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9259–9266, 2019.

[30] Y. Li, Y. Pang, J. Shen, J. Cao, and L. Shao, "NETNet: neighbor erasing and transferring network for better single shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13349–13358, May 2020.