RECEPTIVE FIELD-EXPANDED AND CONTEXT-AWARE SINGLE SHOT DETECTOR

Jian ZHANG[12], Yonghui ZHANG[3]*, Hong JIANG[4], Ruonan LIU[5], Jingxuan HE[6]

Single-shot detector (SSD) ignores the context from the proposal boxes and is not accurate enough for small object detection. In this paper, a new object detection method, called RFCSSD, is proposed. The RFCSSD method enhances the context information by extending the receptive field of the SSD target detector multi-scale feature map and fuses the depth semantic abstraction to improve the accuracy of small object detection. The experimental results show that the RFCSSD can significantly improve the weakness of SSD and achieve more accurate detection performance. The RFCSSD has better mAP than the existing algorithms.

Keywords: Receptive Field-extended, Context-aware, Dilate Convolution, SSD

1. Introduction

Multi-scale object detection has always been a key factor affecting the performance of object detector. Using different scale feature maps to predict the targets at different scales (see Fig. 1 (a)) is an inefficient method. The top-level features are used to predict the bounding boxes with different scales and aspect ratios (see Fig. 1 (b)). However, a single top-level feature map does not accommodate the diversity of the target scales in the actual images. Although, the top-level feature map has deep abstract semantic information, which is beneficial for the expression of features, but the top-level features have fixed receptive fields and low resolution that reduce the detection ability of small targets.

[1] School of Information and Communication Engineering, College of Applied Science and Technology, Hainan University, China, e-mail: whealther@hainanu.edu.cn.
[2] School of Information and Communication Engineering, College of Applied Science and Technology, Hainan University, China, e-mail: whealther@hainanu.edu.cn.
[3] Prof., School of Information and Communication Engineering, Hainan University, China, e-mail: yhzhang@hainanu.edu.cn.
[4] Prof., Hainan University, China, e-mail: jhong63908889@sina.com.
[5] Master, School of Information and Communication Engineering, College of Applied Science and Technology, Hainan University, China, e-mail: rnliu@hainanu.edu.cn.
[6] Undergraduate, College of Applied Science and Technology, Hainan University, China, e-mail: hejingxuan@hainanu.edu.cn.
* corresponding author: Yonghui Zhang

The single-shot detector (SSD) [1] uses the bottom-up pyramid structure to detect the targets of different scales and achieves good detection performance (see Fig. 1 (c)). The bottom layer of the pyramid structure is used to detect the small-scale targets, but the feature graph at the bottom layer only contains weak semantic information, which is not conducive to the expression of small targets. In a recent study, Cui et al. [2] attempted to utilize the characteristics of pyramid structures by constructing top-down channels (see Fig. 1 (d)) and improved the accuracy of target detection compared with the standard SSD. However, the details of the small target lost after repeated convolution, which could not be restored even after deconvolution. On the other hand, the lack of contextual information in the underlying features of small receptive field has not been fully supplemented.

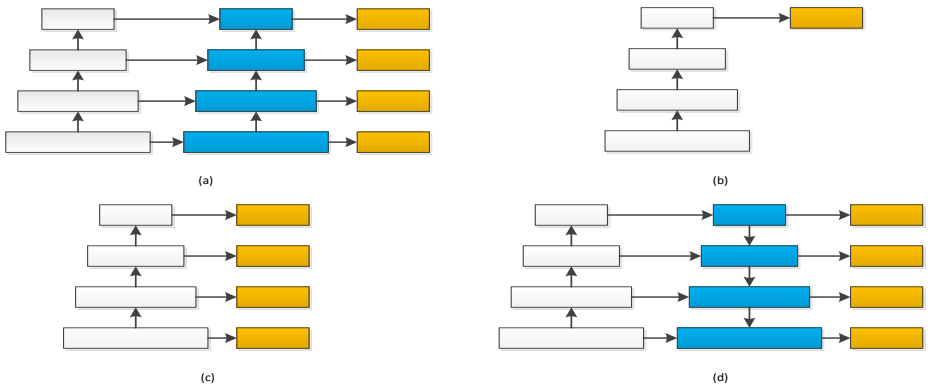

Fig. 1. Multi-scale target detection.

Ren et al. [3] stated that in order to correctly detect the target, the detector should have three basic attributes: 1) the feature graph should have sufficient resolution to represent the fine details of the object; 2) the function that converts the input image into the feature image should be deep enough to abstractly incorporate the appropriate high-level feature of the object into the feature graph; 3) the feature graph should contain appropriate contextual information that can be used to obtain the accurate position of the small target. In the pyramid structure, the shallow feature graph has small receptive field, insufficient contextual information and abstract features, which cannot meet the requirements of 2) and 3). The deep feature graph has large receptive field, sufficient context information and sufficient feature abstraction, but it does not satisfy requirement 1). The low resolution of deep feature graph is not conducive to the recognition of small targets and coordinate regression.

Numerous recent studies have improved the ability of SSD algorithms to identify small targets by transferring the context information contained in the deep feature map to the shallow layer through additional modules [3,4,6,7]. Yu et al. [8] showed that the reduced convolution could expand the receptive field of the

convolution kernel without reducing the resolution and achieved the optimal accuracy in the semantic segmentation of the image. The CSSD [9] improved the SSD algorithm by expanding and convolving the feature maps of different scales to generate the context of different fields of perception. The CSSD method increases the receptive fields of shallow output and does not reduce the resolution of the feature map. However, the shallow feature maps of CSSD still lacks sufficient semantic abstraction information. StairNet [10] starts at its deepest level and incorporates a deep context layer by layer, leading to a deeper semantic abstraction at each level, called progressive semantic aggregation. However, the view of the underlying output in the StairNet has not been substantially expanded.

This paper proposes the RFCSSD framework that expands the receptive field of the feature graph with the expansion convolution keeping the resolution unchanged. In addition, a simple and effective multi-level feature fusion module is designed that transmits powerful semantic information from top to bottom in the network. The proposed design enables the shallow feature mapping to obtain a large sensing field and retain sufficient object details, while the deep strong semantic information is also transferred to the shallow feature graph through the fusion layer by layer. The performance of the proposed framework is better than the most advanced single-level detector.

The main contributions of this paper are as follows:

The RFCSSD framework is proposed. The shallow feature graph obtains large receptive field and maintains high resolution through dilated convolution. Then the shallow feature graph is integrated with the depth feature to significantly improve the detection ability of small target. A multi-level feature fusion module is also designed. The new fusion feature has multi-scale representation and deep semantic information.

A large number of experiments are conducted to provide sufficient options for designing the feature combinations.

The proposed RFCSSD achieves excellent performance on the datasets of PASCL VOC2007 and PASCAL VOC2012, and maintains real-time processing speed.

2. Related works

The target detectors based on deep learning can be divided into two categories: candidate region-based methods [10-13] and regression-based methods [1, 14]. The SSD [1] combines the idea of YOLO regression with the anchor mechanism of Faster R-CNN, generates multiple bounding boxes for each anchor point, and uses the pyramid structure to process the objects of different scales in the prediction stage, achieving higher inference speed and accuracy than YOLO.

The pyramid structure used by the SSD is a bottom-up structure, and its deep feature mapping cannot provide the high resolution required by attribute 1). The shallow feature graph is obtained by the shallow transformation function and cannot meet the requirement of attribute 2). Each feature graph is only responsible for the detection of the object of corresponding size, and the receptive field of the shallow feature graph is extremely small, which cannot meet the requirement of context information in attribute 3). The DSSD [4] adds additional deconvolution structure to improve the recognition ability of small targets by integrating the context information of each prediction layer and corresponding deconvolution layer. However, the infer speed of DSSD is only 11.2 fps.

Several studies have improved the accuracy of detector by adding feature maps of different scales to introduce additional context information [2,6,7,16,17]. Wei et al. [9] improved the pyramid structure of SSD by using deconvolution and expansion convolution. Liu et al. [18] used the multi-branch convolution and the expansion convolution of different cores to enhance the discriminability and the robustness of features. Qin et al. [20] conducted multi-scale feature fusion of the shallow three-layered feature map of pyramid structure using expansion convolution and deconvolution. The MDSSD [2] designed a concise deconvolution fusion module to add the high-level features with semantic information to the low-level features in order to obtain feature mapping with rich information. StairNet [21] introduced a feature composition module to scale up a strong semantic abstraction in a top-down manner in order to address the lack of sufficient semantic information in the shallow layer when detecting small targets. Wu et al. [22] proposed a bidirectional pyramid structure (BPN) to integrate deep and shallow layers. Zhou et al. [23] used the DenseNet as the backbone network and combined it with scale transfer module (STM) to construct the STDN single-stage object detector. The expanded convolution [8] can exponentially expand the receptive field without compromising the resolution and the coverage. The expanded convolution can achieve optimal performance in image semantic segmentation. Using the ResNet as the base network, DetNet [24] redesigned the backbone network specifically for target detection, which used the expansion convolution to expand the receptor field while maintaining the resolution of the feature map without shrinking. Parallel pyramid network [25] improved the identification performance by widening the network width instead of the depth. Wei et al. [9] improved the pyramid structure of SSD by using deconvolution and expansion convolution.

After large step down-sampling, the information of the small target is deteriorated or even disappeared. Since the deep feature graph lacks the information for small target, it is futile to transmit the deep information to the shallow layer. Therefore, dilated convolution is used in the proposed framework to expand the receptive field and maintain the resolution of feature map, so that

the details and the positioning information of small targets will not deteriorate. In addition, a semantic fusion module is designed that fuses the deep semantic information layer by layer into the shallow feature layer. The final model, called RFCSSD, effectively satisfies the three essential attributes of the detector.

3. Method

In this section, the SSD framework is first reviewed and the feature fusion is analyzed. Then the process of expanding the receptive field of the feature graph through dilated convolution is described. Next, the multi-level feature fusion module is introduced. Lastly, the proposed RFCSSD network architecture and training strategy are presented.

3.1 SSD framework

The standard SSD uses the truncated VGG16 [29] as feature extractor, and then adds additional multi-scale prediction structure to predict the targets of different scales. The multi-scale prediction improves the target detection performance and maintains real-time detection speed. However, the shallow feature map lacks sufficient context information and semantic abstraction, and the detection performance of SSD for small targets is poor.

The multi-scale feature graph of SSD can be expressed with a simple mathematical formula as follows:

$$f_n = C_n(f_{n-1}) = C_n(C_{n-1}(\ldots C_1(I))) \tag{1}$$

$$Detection = D(\tau_n(f_n), \ldots, \tau_{n-k}(f_{n-k})), n > k > 0 \tag{2}$$

where I is the input image, f_n is the feature graph of the nth layer, and $C_n(\cdot)$ is the nth nonlinear transformation, including convolution, pooling, ReLU and other operations. $\tau_n(\cdot)$ is the function that converts the feature graph of the nth layer into detection result within a specific size range and D is the final detection output. Obviously, when k is relatively large, the depth of pyramid structure will decrease layer by layer, and the semantic level of feature graph of each layer will also gradually decrease. Thus, the shallowest layer $n \to n-k$ only contains weak semantic information. In addition, the research of Xiang et al. [9] showed that the size of the real receptive field was far smaller than the theoretical one, and the effective receptive field of the Conv4_3 layer used to predict small targets in SSD was only 58.6, accounting for only 1/26 of the original image region. It is easy to see that the layer f_{n-k} used to detect small targets has insufficient contextual information. Obviously, the shallow features defined by formula (2) violate attributes 2) and 3). In this regard, the existing studies have attempted to fuse the features of different scales to improve the performance of the algorithm. The SSD of feature fusion can be described as follows:

$$Detection = \hat{D}(\tau_n(F_n(H), \tau_{n-1}(F_{n-1}(H)),..., \tau_{n-k}(F_{n-k}(H))$$
$$H = \{f_n, f_{n-1},..., f_{n-k},\}, n > k > 0 \quad (3)$$

where H is the set of all feature graphs and $F(\cdot)$ is the function that performs fusion of the feature graph. The study in reference [9] showed that the multi-scale dilated convolution could rapidly expand the TRF size of each prediction layer and ensure large enough region for each feature point. Different from pooling, the dilated convolution will not reduce the resolution of the feature graph. On this basis, if the feature graph f of formula (3) has the following characteristics and can meet the requirements of attributes 1) ~3):

$$f_{n-k}^{'} = \Phi(\varphi_{n-k}(f_{n-k}) + f_{n-k+1})$$
$$\cdots \quad (4)$$
$$f_n^{'} = f_n$$

Where $\Phi(\cdot)$ performs dilated convolution on feature graph and $\Phi(\cdot)$ is the nonlinear processing of the feature after fusion. Compared with formula (3), formula (4) is based on multi-scale feature representation of SSD and expands a larger receptive field to each layer of feature graph through dilated convolution in order to obtain broader contextual information and maintains the resolution of the feature graph. These features satisfy attributes 1) and 3). At the same time, the deep strong semantic information can also be transferred to the shallow feature graph layer by layer to satisfy attribute 2).

3.2 Dilation block

The dilated convolution injects holes into standard convolution maps to replace the convolution with a step size greater than 2 or pooling, to generate higher-resolution feature graphs and obtain a wider range of receptive fields to capture the contextual information. This approach has achieved good results in semantic segmentation [26] and target detection [1,27].

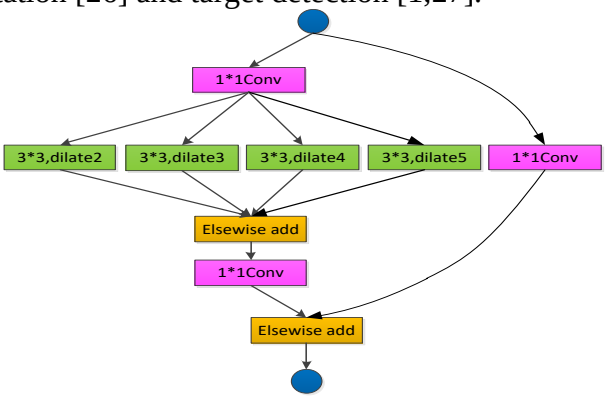

Fig. 2. Dilation block.

In this paper, the expanded convolution block shown in Fig. 2 is introduced to expand the receptive field of the shallow feature map without losing the information in order to ensure each output contains a large range of context information.

3.3 Fusion module

The attribute 2) requires the feature map to have sufficient semantic abstraction. A cascading feature fusion module is designed in this paper to enhance the semantic information of the shallow feature map. The feature fusion module transfers the high-level abstract features to the shallower layer.

In order to combine the information uploaded from the deep layer with the corresponding information from the shallow layer, the Decblock (orange in Fig. 3) is introduced. The Decblock takes the deep fusion feature as the input with a convolution kernel size of 3×3, and scaled by the deconvolution layer with an up-sampling rate of 2. The output of Decblock has 256 channels, and the convolution of 1×1 is used to reduce the dimension and reconstruct the features. Since the features of different layers represent different scale distributions, they are normalized first. Then the features are fused using eltwise add operation (yellow in Fig. 3). The fusion module delivers the deep high-level semantic abstraction to the shallow feature map layer by layer.

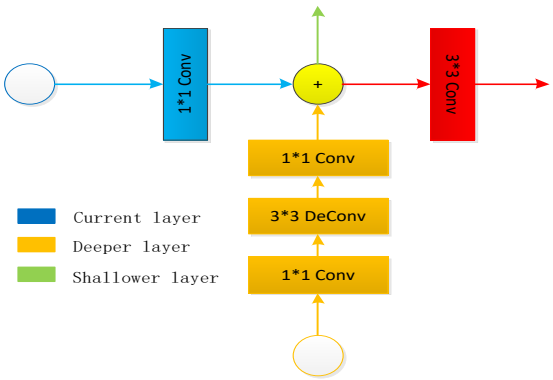

Fig. 3. Fusion module.

The convolution layer of 3×3 (red in Fig. 3) is used to mix the current layer and the deeper information in order to construct the enhanced feature map before the classifier. The enhanced feature map has the same spatial resolution as the original feature map, with larger receptive field and enhanced semantic information.

3.4 RFCSSD

The proposed RFCSSD detector still uses the multi-scale, single-stage detection framework of the original SSD. Fig. 4 shows the architecture of the

proposed RFCSSD. After the multi-scale feature output, the Dilation block proposed in section 3.2 is embedded to obtain the feature map with high resolution and large receptive field. Different from literature [9], the Dilation block is only added in the shallowest layer. It is believed that since the resolution of the deepest feature map is too small (3×3 and 1×1), it is not suitable to apply the dilated convolution of rate of 5. It is also believed that expanding the receptive field of each layer will introduce excessive background noise, thereby reducing the accuracy, which is also confirmed by the study in Section 4.2. Then, the feature fusion module proposed in section 3.3 is applied to the top-level feature output to transfer the deep semantic information to the shallow feature map layer by layer.

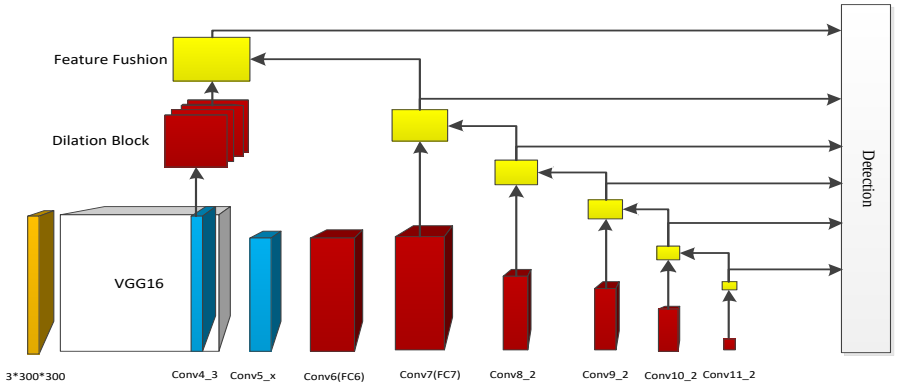

Fig. 4. The architecture of RFCSSD.

Recent studies have shown that using the new backbone network to replace VGG16 improves the feature extraction ability [4,5,21,22], and using the focus loss [28] and combining the two-stage method [17] to solve the problem of category imbalance in training can significantly improve the performance of the detector. However, this paper currently focuses on the framework structure of the original SSD to compare the performance of the proposed approach.

3.5 Training

In this paper, the same training process as the original SSD is followed. The SGD with momentum of 0.9 was used for optimization, weight attenuation was set at 0.0005, and batch was set at 32. Using the same learning strategy as the SSD, 120K iterations were trained, learning rate of 0.001 was used in the first 80K iterations, and then the learning rate was reduced by 10 times per 20K iterations.

4. Experiment

The experiments were conducted on PASCAL VOC2007 and VOC2012 datasets to compare the performance of the proposed method with that of other typical algorithms. The reported performance parameters of the algorithms are used for the comparison. All the experiments in this paper are based on Caffe's framework.

4.1 PASCAL

The proposed model was trained on PASCAL VOC2007 and PASCAL VOC2012 and evaluated using the VOC2007test and VOC2012test test sets. Table 1 shows the test results on the PASCAL2007 test set. The mAP of the RFCSSD reached 79.1, which was 1.6 higher than the original SSD, and better than CSSD [9], DSSD [4], and StairNet [10]. The RFCSSD infer speed reached 57.7 FPS on GTX1080ti*2 platform, which was faster than the DSSD (11.2fps), and slightly lower than the 67fps of the original SSD.

Table 1

Detection results on PASCAL VOC 2007 test set. (VOC 07+12:07 trainval+12trainval)

method	Data	mAP (%)	aero	bike	bird	boat	bottle	Bus	car	cat	chair	cow	table	dog	Horse	mbike	persn	plant	sheep	sofa	train	tv
SSD300 [1]	07+12	77.5	79.5	83.9	76	69.6	50.5	87	85.7	88.1	60.3	81.5	77	86.1	87.5	83.9	79.4	52.3	77.9	79.5	87.6	76.8
CSSD [9]	07+12	78.1	82.2	85.4	76.5	69.8	51.1	86.4	86.4	88	61.6	82.7	76.4	86.5	87.9	85.7	78.8	54.2	76.9	77.6	88.9	78.2
DSSD [4]	07+12	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	88.7	86.7	79.7	51.7	78	80.9	87.2	79.4
StairNet [10]	07+12	78.8	81.3	85.4	77.8	72.1	59.2	86.4	86.8	87.5	62.7	85.7	76	84.1	88.4	86.1	78.8	54.8	77.4	79	88.3	79.2
Ours	07+12	**79.1**	81.1	85.0	76.8	73.7	54.6	87.5	87.0	87.3	61.8	86.1	79.4	86.6	88.5	86.1	79.9	53.5	79.0	81.6	87.5	78.2

For VOC2012 task, the training set composed of VOC2007test and VOC2012train was used for training and VOC2012test was used for testing. The results are shown in Table 2, which again validate that the proposed RFCSSD is superior to all other comparison algorithms

Table2

Detection results on PASCAL VOC2012 test set
(All models were trained on 07trainval+07test+12trainval)

Method	Network	mAP(%)
SSD300[1]	VGG16	75.8
StairNet[10]	VGG16	76.4
DSSD[4]	ResNet101	76.3
DSOD[32]	DenseNet	76.3
ours	VGG16	76.6

4.2 Ablation Study on VOC2007

In order to understand the effectiveness of the improvement to the original SSD proposed in this paper, the models were run with different settings on VOC2007, and their recorded evaluations are shown in Table 3.

The CSSD [9] adds expanded convolution structure to each multi-scale feature layer. However, it is believed that: 1) the deepest two layers have enough receptive fields, and their low resolution is not suitable for the use of rate=5 expansion convolution; 2) excessive background noise is introduced while expanding the receptive field. The results in Table 3 confirm the inference. In the end, this paper adopts the method with the highest accuracy, only extending the receptive field to the shallowest conv4_3.

Table3

Effectiveness of dilation numbers on the VOC2007 test set.

Dilation block	0	1	2	3	all
mAP(%)	77.22	77.44	77.39	77.39	77.36

Furthermore, the effects of the Dilation block and the feature fusion on the accuracy of the detector are also examined. Table 4 shows that the Dilation Block+Feature fusion used in this paper improves the accuracy of the original SSD detector by 2%.

Table 4

Effectiveness of various design on the VOC2007 test set.

Component	RFCSSD			SSD
Dilation Block	√		√	
Feature Fusion block	√	√		
mAP(%)	79.1	78.8	77.4	77.2

5. Conclusion

In this paper, an effective improved SSD framework is presented that can obtain more contextual information by expanding the receptive field without reducing the resolution in order to retain more details of small targets and transmit

high-level semantic information layer by layer for accurate target detection. Generally, the traditional detection method at each stage uses the pyramid structure to dispose multi-scale objects, which creates contradictions between shallow and deep feature maps in resolution, contextual information and advanced semantic abstraction. In order to resolve this issue, the dilated convolution module is used to expand the receptive field while keeping the resolution of the feature map unchanged, and the feature fusion module is used to transfer the deep strong semantic information to the shallow feature map layer by layer. The experimental results demonstrate that the accuracy of the proposed method is better than that of existing methods including DSSD, achieving the reasoning speed of 58fps.

Acknowledgment

The research was financially supported by the Hainan Provincial Natural Science Foundation of China "Behavior recognition based on convolution neural network" (618MS027) and the Key development project of Hainan Provincial "Research on Marine biological identification based on edge computing and deep neural network" (ZDYF2019024).

R E F E R E N C E S

[1]. *Liu W, Anguelov D, Erhan D.* Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016, pp. 21-37.

[2]. *Cui L.* MDSSD: Multi-scale Deconvolutional Single Shot Detector for small objects. arXiv preprint arXiv:1805.07009, 2018.

[3]. *Ren J, Chen X, Liu J.* Accurate single stage detector using recurrent rolling convolution[C]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 752-760.

[4]. *Fu C Y, Liu W, Ranga A.* DSSD: Deconvolutional Single Shot Detector. arXiv preprint arXiv:1701.06659, 2017.

[5]. *Jeong J, Park H, Kwak N.* Enhancement of SSD by concatenating feature maps for object detection[J]. arXiv preprint arXiv:1705.09587, 2017.

[6]. *Cao G, Xie X, Yang W.* Feature-fused SSD: fast detection for small objects[C]//Ninth International Conference on Graphic and Image Processing (ICGIP 2017). International Society for Optics and Photonics, 2018, 10615: 106151E.

[7]. *Lee K, Choi J, Jeong J.* Residual features and unified prediction network for single stage detection. arXiv preprint arXiv:1707.05031, 2017.

[8]. *Yu, F., Koltun, V.* Multi-scale context aggregation by dilated convolutions. In: ICLR (2016).

[9]. *Wei Xiang, Zhang D Q, Yu H.* Context-aware single-shot detector[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1784-1793. (CCF-A)

[10]. *Woo S, Hwang S, Kweon I S.* Stairnet: Top-down semantic aggregation for accurate one shot detection[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 1093-1102.

[11]. *Girshick R, Donahue J, Darrell T.* Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, pp. 580-587.

[12]. *He K, Zhang X, Ren S.* Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, **vol. 37**, no. 9, 2015, pp. 1904-1916.
[13]. *Girshick R.* Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[14]. *Ren S, He K, Girshick R.* Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015, pp. 91-99.
[15]. *Redmon J, Divvala S, Girshick R.* You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 779-788.
[16]. *Li Z, Zhou F.* FSSD: Feature Fusion Single Shot Multibox Detector. arXiv preprint arXiv:1712.00960, 2017.
[17]. *Zhang S, Wen L, Bian X.* Single-shot refinement neural network for object detection[C]//IEEE CVPR. 2018.
[18]. *Liu S, Huang D, Wang Y.* Receptive Field Block Net for Accurate and Fast Object Detection[J]. arXiv preprint arXiv:1711.07767, 2017.
[19]. *Han G, Zhang X, Li C.* Single shot object detection with top-down refinement[C]//Image Processing (ICIP), 2017 IEEE International Conference on. IEEE, 2017, pp. 3360-3364.
[20]. *Qin P, Li C, Chen J.* Research on improved algorithm of object detection based on feature pyramid[J]. Multimedia Tools and Applications, 2018, pp. 1-15.
[21]. *Woo S, Hwang S, Kweon I S.* Stairnet: Top-down semantic aggregation for accurate one shot detection[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 1093-1102.
[22]. *Wu X, Zhang D, Zhu J.* Single-Shot Bidirectional Pyramid Networks for High-Quality Object Detection[J]. arXiv preprint arXiv:1803.08208, 2018.
[23]. *Zhou P, Ni B, Geng C.* Scale-Transferrable Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 528-537.
[24]. *Li Z, Peng C, Yu G.* DetNet: Design Backbone for Object Detection[C]//European Conference on Computer Vision. Springer, Cham, 2018, pp. 339-354.
[25]. *Kim S W, Kook H K, Sun J Y.* Parallel Feature Pyramid Network for Object Detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 234-250.
[26]. *Law H, Deng J.* Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 734-750.
[27]. *Kong T, Sun F, Yao A.* Ron: Reverse connection with objectness prior networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 1: 2.
[28]. *Lin T Y, Goyal P, Girshick R.* Focal loss for dense object detection. IEEE transactions on pattern analysis and machine intelligence, 2018.
[29]. *Simonyan K, Zisserman A.* Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[30]. *Chen L C, Papandreou G, Schroff F.* Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.
[31]. *Dai J, Li Y, He K.* R-fcn: Object detection via region-based fully convolutional networks[C]//Advances in neural information processing systems. 2016, pp. 379-387.
[32]. *Shen Z, Liu Z, Li J.* Dsod: Learning deeply supervised object detectors from scratch[C]//The IEEE International Conference on Computer Vision (ICCV). **vol. 3**, no. 6, 2017, pp. 7.