Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2021.010065
Tech Science Press

Article

Feature-Enhanced RefineDet: Fast Detection of Small Objects

Lei Zhao[*] and Ming Zhao

School of Computer Science and Engineering, Central South University, Changsha, China
[*]Corresponding Author: Lei Zhao. Email: zlei1995@csu.edu.cn

Received: 30 August 2020; Accepted: 14 January 2021

Abstract: Object detection has been studied for many years. The convolutional neural network has made great progress in the accuracy and speed of object detection. However, due to the low resolution of small objects and the representation of fuzzy features, one of the challenges now is how to effectively detect small objects in images. Existing target detectors for small objects: one is to use high-resolution images as input, the other is to increase the depth of the CNN network, but these two methods will undoubtedly increase the cost of calculation and time-consuming. In this paper, based on the RefineDet network framework, we propose our network structure RF2Det by introducing Receptive Field Block to solve the problem of small object detection, so as to achieve the balance of speed and accuracy. At the same time, we propose a Medium-level Feature Pyramid Networks, which combines appropriate high-level context features with low-level features, so that the network can use the features of both the low-level and the high-level for multi-scale target detection, and the accuracy of the small target detection task based on the low-level features is improved. Extensive experiments on the MS COCO dataset demonstrate that compared to other most advanced methods, our proposed method shows significant performance improvement in the detection of small objects.

Keywords: Small object detection; feature fusion; receptive field block

1 Introduction

Object detection is to determine the location of the object to be searched in the image and to ascertain which category they belong to. With the great achievements in the field of image recognition based on the framework of deep learning in recent years, more and more researchers are attracting research in the field of target detection, and many object detection frameworks based on CNN have been proposed. They have significantly improved the performance of most benchmarks and have made great progress in object detection in terms of accuracy and speed. However, for the existing detectors, detecting small objects is still very challenging due to the low resolution of the pictures and the loss of small object features.

There are two ways to define existing small objects, one is to compare relative scales, and the other is absolute scale. The relative scale is that the length and width of the target size is 0.1 of the input image size, and the absolute size is that the size of the object in the input image is less than 32 pixels × 32 pixels. Small object detection is important in a variety of situations, including autonomous vehicles, image analysis and understanding, intelligent monitoring, and behavioral recognition. There are three reasons for the difficulty of small target detection: the change of target scale, image resolution, and environmental factors. Generally speaking, the feature representation of the small object is insufficient. The existing CNN-based detection methods typically employ several merge/downsample operations in a feed-forward neural network. As the depth of the network increases, the spatial resolution of the target object in the input image is reduced by tens or hundreds of times, which means that small objects that are already blurred in the input image are more difficult to recognize. In the solution of detecting small objects, the existing CNN-based target detection pipeline faces the dilemma: using high-resolution images as input leads to high computational cost, but using low-resolution images as input will lose the characteristic representation of small objects. Therefore, it leads to low precision, so it does not solve the problem well.

In this paper, our main contributions are summarized as follow, based on the RefineDet [1] network framework, we propose our network structure RF2Det by introducing receptive field block (RFB) [2] to solve the problem of small target detection. The RFB is introduced to enhance the semantic information of the low-level feature map while enhancing the distinguishability and robustness of the low-level features. At the same time, we propose a Medium-level Feature Pyramid Networks, which combines high-level context features with low-level features, so that the network can simultaneously use the low-level and high-level features for multi-scale target detection. Extensive experiments on the MS COCO dataset [3] show that the proposed method shows a significant performance improvement in the detection of small objects compared to existing detectors.

The rest of this paper is organized as follows. We introduced the related work in Section 2. Then we describe our RF2Det framework in Section 3. Detailed experiments are presented in Section 4. Finally, we conclude in Section 5.

2 Related Work

Now the proposed object detectors can be roughly divided into two broad categories, one-stage detectors and two-stage detectors. One-stage detectors, such as Retinanet [4], SSD [5] and YOLO 9000 [6], these proposed network models combine extraction and detection into one, which can simultaneously give the object location and classification results. For two-stage detectors, such as fast R-CNN [7], RFCN [8] and FPN [9], they consist of a Regional Proposal Network (RPN) and a classification network. The Fast Detector Single Shot Multi-Box Detector has been greatly improved in speed for eliminating area recommendations and subsequent pixel resampling stage. To improve the accuracy of small objects, the deconvolution single shot detector [10] uses the SSD architecture as the baseline. Since the focus on improving accuracy by using ResNet-101, a lot of speed is inevitably sacrificed.

Small target detection is a problem in the existing deep learning convolutional neural network model. It is even more difficult to achieve balance between accuracy and speed. By increasing the size of the input image [11], the performance of the small object can be improved because the corresponding small target is magnified, but this is treating symptoms and not the root cause. Some people use the image pyramid [12] to solve this problem, but it takes a lot of time to exchange for the improvement of precision, which is not a desirable solution. Other researchers [13–14] have designed a novel network architecture to integrate multiple high-level and low-level feature layers. The early object detection framework (R-CNN, YOLO Series) did not work well for small object detection. In the past two years, the method of using multi-layer feature maps has been proposed, such as feature pyramid, RNN idea, layer-by-layer prediction, which has significantly improved the effect of small object detection.

Menikdiwela et al. [15] proposed a fine-tuned VGG16 network for detecting small objects such as spiders. Christian et al. [16] introduced an improved scheme for generating anchoring proposal and proposed modifications to Faster R-CNN to utilize higher resolution feature mapping for small objects. Cao et al. [17] aim to quickly detect small objects, using the best object detector single shot multi-box detector (SSD) based on accuracy and speed trade-off as the basic structure. They proposed a multi-level feature fusion method that introduces context information in SSD to improve the accuracy of small objects. Wang et al. [18] obtained a conclusion from the study of low-resolution images, a deeper network architecture, more filters or larger filter sizes do not contribute to the classification of small objects. In [19] and [20], Bell et al. considered the object detection of small objects in the context of Fast R-CNN. Wang et al. [21] research on small objects such as company logos, and improve the detection performance of small objects by considering the relationship among the three aspects: the feeling field of features, the size of objects to be searched in the image, and the detection level. Yang et al. [22] solve this problem by creating multi-scale features for small objects by using key techniques such as skip pooling. Wang et al. [23] proposed a cascaded mask generation framework, which takes multi-scale images as input, to solve small target detection. Liang et al. [24] proposed a two-stage detector similar to Faster-RCNN, which can classify more accurately with relatively fewer parameters.

3 RF2Det

3.1 Network Architecture

Our overall network structure is shown in Fig. 1. We use VGG-16 as our backbone network, which is pre-trained on the ILSVRC CLS-LOC dataset [25]. Then we convert the parameters of fc6 and fc7 of VGG-16 into convolutional layers conv_fc6 and conv_fc7 by using sub-sampling parameters.

Before the Hyper Anchor Refinement Module (HARM), the RFB module was introduced to enhance the semantic information of the low-level feature map, while enhancing the resolvability and robustness of the low-level features. Specifically, the features of conv3_3, conv4_3, conv5_3, and conv7_fc are now processed by the RFB module, conv3_3, conv4_3, conv5_3 modules use RFB-s, and conv7_fc modules use RFB. At the same time, we propose a Medium-level Feature Pyramid Networks, which combines opportune high-level context features with low-level features. The network can use the features of both the low-level and the high-level for multi-scale object detection, and the accuracy of small object detection tasks using low-level features is improved.

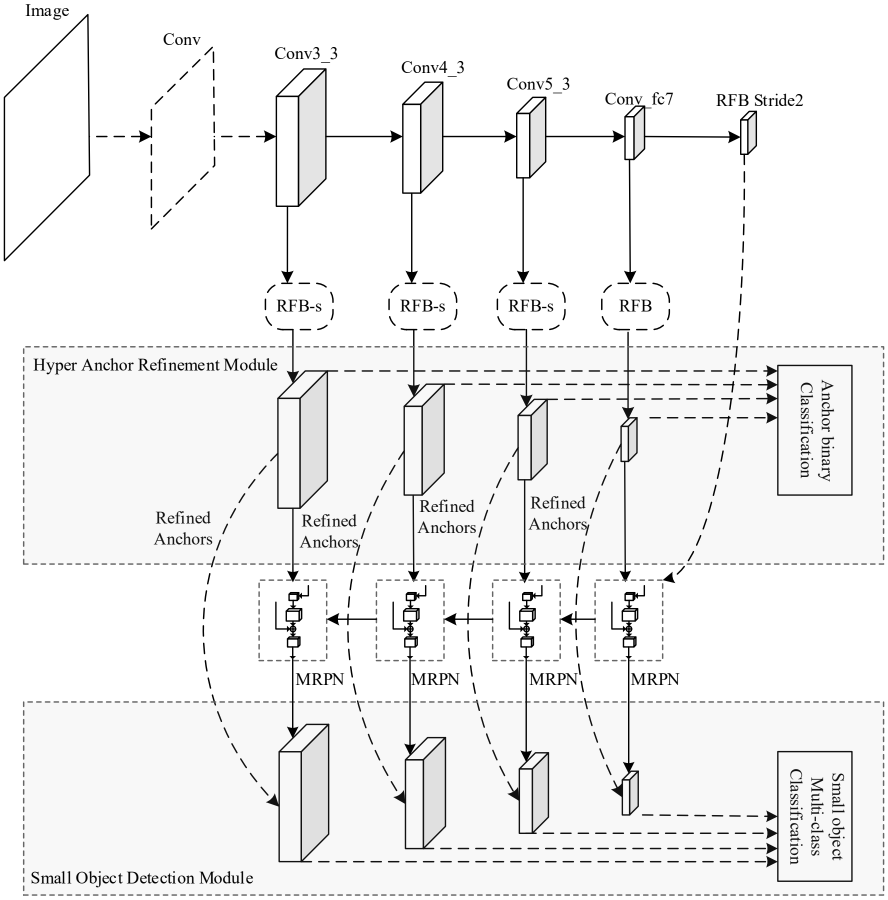

Figure 1: The architecture of RF2Det. The conv3_3, conv4_3, conv5_3 are tailed by RFB-s, and conv7_fc is tailed by RFB. It consists of the baseline VGG-16, Medium-level Feature Pyramid Networks, Hyper Anchor Refinement Module, Small Object Detection Module

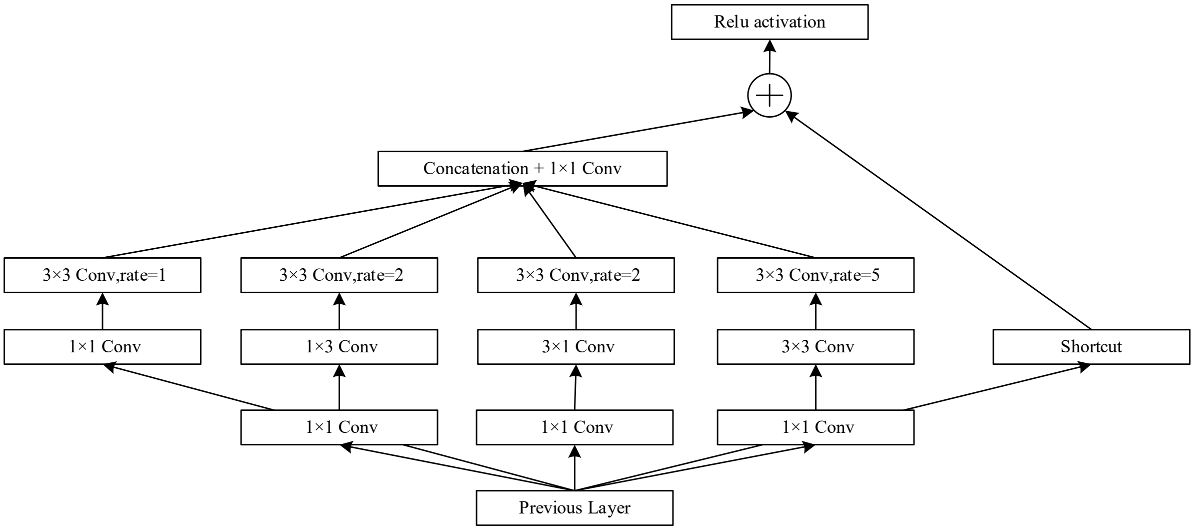

Figure 2: The architecture of RFB-s

3.2 Receptive Field Block

Many recent works have confirmed that low-level feature maps are critical for small object detection. The low-level feature map preserves the rich spatial information of small objects, but because it is located at the lower level of the whole network, the extracted features are not enough to represent the information of the small objects. By introducing the RFB module into the RefineDet, multi-branch pooling with varying kernels is used for the lower-level feature maps. The selection of these kernel parameters is based on the different sizes of RFs to extract the feature information of different receptive fields. In terms of structure, RFB draws on the idea of Inception, mainly adding the divided revolution on the basis of Inception, which is used to simulate eccentricity. Then the features of all the branches are concatenated together to generate a final representation to enhance the semantic information of the low-level feature map, enhancing the resolvability and robustness of the features.

The starting point for the introduction of RFB is because RFB enhances the ability of the network to extract features of images by simulating the receptive field of human vision. In the process of CNN extracting image features, the receptive field is used to represent the size of the mapping area on the original image of the pixel on the feature map output by each layer of the convolutional neural network. RFB highlights the relationship between the size of receptive field and eccentricity. It also improves the importance of features close to the central region and becomes more sensitive to small spatial offsets.

By introducing the RFB module, the effect of dense sampling can be achieved on the original feature map. The advantage of this is that the receptive field is effectively increased, and the extracted features have more global information, which is more conducive to the identification of small objects.

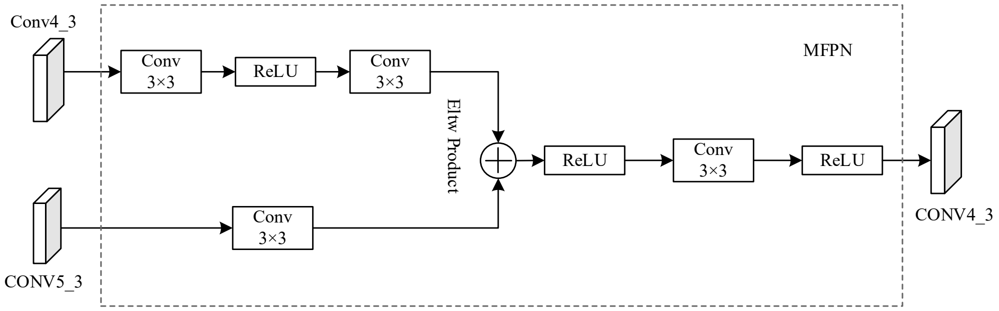

Figure 3: Fusion process of high-level feature map CONV5_3 and low-level feature map conv4_3

3.3 Medium-Level Feature Pyramid Networks

In recent years, some studies have shown the importance of contextual information for object detection, especially for small, blurred objects. In order to improve the performance of the detector processing small objects, low-level features with high resolution play a key role. There are different sizes of objects in the image, and we know that different targets have different characteristics. In order to distinguish small objects, we need to make good use of low-level features. However, if we only use the feature map with low-level and fine-grained features to detect small objects, we will introduce a large number of false positions because of the lack of sufficient discriminant information. In addition, this is not a good solution to solve the problem of fuzzy and low illumination.

In the backbone network of some object detection models, the receptive fields of the top two layers conv7_2 and conv6_2 are 724 and 468, respectively. The input size of our training image is 512, which means that extracting features from conv7_2 and conv6_2 is not suitable for small object detection. Medium-level feature pyramid networks is a top-down fusion starting from the appropriate middle layer. The receptive field is close to half of the input size, so that features that are helpful to each other can be connected together. We use each branch of the feature map of the appropriate middle layer to detect small objects in a specific scale range, so that the detection performance of small objects is improved and the detection speed is also accelerated. The Medium-level feature pyramid networks (MFPN) combines high-level features with low-level features to enhance the semantic information of low-level features, which is more conducive to detecting small objects.

We take the fusion of the high-level feature map CONV5_3 of MFPN and conv4_3 of HARM as an example, and the process is shown in Fig. 3. The convolution kernel has a size of 3 × 3, the number of channels is 256, the size of the deconvolution kernel is 4 × 4, the stride is 2, and the number of channels is also 256. The feature map sizes of the different layers are different, so CONV5_3 is first scaled to the same size as the conv4_3 feature by the deconvolution operation. Then, the fusion is performed by Eltw Product operation to obtain CONV4_3.

4 Experiments

4.1 MS COCO

Tab. 1 shows the comparison of our proposed RF2Det with other advanced methods on the MS COCO dataset. By introducing the RFB module and the MFPN, our proposed RF2Det has a detection accuracy of 17.1% in small objects, which is superior to other methods.

Table 1: Detection results on the MS COCO test-dev set. Bold font indicates the best performance

Method	Backbone	Time	AP_s
Faster R-CNN [23]	VGG-16	147 ms	7.7
R-FCN [24]	ResNet-101	110 ms	10.8
R-FCN w Deformable CNN [30]	ResNet-101	125 ms	14.0
Mask R-CNN [31]	ResNet-101-FPN	210 ms	16.9
RFB Net512 [1]	VGG-16	30 ms	16.2
RefineDet512 [2]	VGG-16	41.5 ms	16.3
YOL Ov2 [22]	Dacknet	25 ms	5.0
SSD321 [29]	ResNet-101	91 ms	6.2
DSSD513 [29]	ResNet-101	182 ms	13.0
RetinaDet500 [20]	ResNet-101-FPN	90 ms	14.7
Ours	VGG-16	52 ms	**17.1**

We set the batch size to 32 and the initial learning speed to 10^{-3}, just like the original SSD, but this makes the training process unstable because the loss fluctuates dramatically. Instead, we use a warm-up strategy that takes the learning rate from 10^{-6} to 4×10^{-3} in the first 5 stages. After the warm-up phase, it returns to the original learning rate plan, dividing by 10 in 150 and 200 periods. The total number of training periods is 250. We use a weight attenuation of 0.0005 and a momentum of 0.9.

Loss Function. The loss function consists of two parts, the loss in the HARM and the loss in the Small Object Detection Module (SODM). For HARM, we assign a binary class label, object or not, to each anchor and return its location and size to get a refined anchor. We then filter the refined anchors whose negative confidence less than the threshold to further predict the object category and the exact object location and size for the SODM. With these definitions, we define the loss function as:

$$\mathrm{L}_{\mathrm{HARM}} = \sum_{\mathrm{i}} L_b(p_i, [l_i^* \geq 1]) + \sum_{\mathrm{i}} [l_i^* \geq 1] L_{\mathrm{r}}(\mathrm{x}_{\mathrm{i}}, \mathrm{g}_{\mathrm{i}}^*) \tag{1}$$

and

$$\mathrm{L}_{\mathrm{SOPM}} = \sum_{\mathrm{i}} L_{\mathrm{m}}(c_i, l_i^*) + \sum_{\mathrm{i}} [l_i^* \geq 1] L_{\mathrm{r}}(\mathrm{x}_{\mathrm{i}}, \mathrm{g}_{\mathrm{i}}^*) \tag{2}$$

where i is the index of the small-batch anchor, l_i* is the ground truth class label of the anchor i, and g_i* is the ground truth position and size of the anchor i. p_i and x_i are the prediction confidence of the anchor as the object and the fine coordinates of the anchor i in the HARM, respectively. c_i and t_i are the coordinates of the prediction object category and the bounding box in the SODM, respectively. The binary classification loss L_b is the cross-entropy loss between two classes, objects and non-objects, and the multi-class classification loss L_m is the softmax loss on multiple types of confidence. Similar to Fast R-CNN, we use the smooth L_1 loss as regression loss L_r.

Anchors. We use anchor points with three aspect ratios {1:2, 1:1, 2:1}. An anchor of size $\{2^0, 2^{1/3}, 2^{2/3}\}$of the original 3 aspect ratio anchor sets is added at each level. We assign the anchor to the ground truth object box using Intersection over Union threshold of 0.5; if its IoU is at [0, 0.4), it returns the background. If the anchor point is not assigned, there may be an overlap in [0.4, 0.5], the anchor point will be ignored during training. In order to get more low-level anchors, the original SSD is only associated with four default boxes in the location of conv4, conv10, and conv11 feature map and six default anchors in all other layers. As we mentioned above, low-level features are critical for detecting small objects. Therefore, we assume that if more anchor points are added in the low-level feature map, for example, conv3_3, the detection performance for small objects should tend to increase. In the experiment, we placed 9 default anchor points on conv3_3, which further improved our experimental results for our proposed RF2Det.

4.2 Ablation Study

To demonstrate the effectiveness of the different components of the proposed RF2Det, we constructed three different variants and evaluated them at MSCOCO. The results are shown in Tab. 2. Specifically, for fair comparisons, we use the same parameter settings and input sizes in the evaluation.

Table 2: Ablation experiments for RF2Det. We used the trainval35k set for training and test on the test-dev set

Component	RF2Det		
Add RFB ?	✓	✓	✓
Add MRPN ?		✓	✓
Multi-Scale Training?			✓
AP_S(%)	13.9	15.8	17.1

Receptive Field Block. As shown in Tab. 2, by introducing the RFB module, our RF2Det detection results for small objects reached an astonishing 13.9%. The reason is that the introduction of the RFB effectively increases the receptive field, so that the extracted features have more global information and the feature resolvability and robustness are enhanced.

Medium-level Feature Pyramid Networks. The detection results of this module are listed in the second column of Tab. 2, and it can be found that the detection results are increased by 1.9%. The main reason for improving detection accuracy is that RF2Det can inherit the distinguishing features of HARM, and it can use MFPN to combine appropriate high-level context features with low-level features.

Multi-scale Training. For the MS COCO dataset, multi-scale training/testing is still an effective technique for improving performance. By comparing the results of the second and third columns in Tab. 2 (15.8 *vs.* 17.1), it can be seen that multi-scale training improves the detection results by 1.3%.

5 Conclusion

This paper introduces our proposed RF2Det network framework to solve small object detection problems. By introducing RFB, the semantic information of the feature map is enhanced, and the distinguishability and robustness of the features are enhanced. The MFPN combines sufficient high-level context features with low-level features to greatly improve the accuracy of small target detection tasks which takes the low-level features as the main detection basis. Extensive evaluation of the MSCOCO dataset shows that the proposed framework achieves a balance between speed and accuracy, and the results show a significant performance improvement in the detection of small objects. In the future, we will explore the use of RF2Det to detect small faces and introduce focal loss to improve performance.

Funding Statement: There is no fund support for this paper.

Conflicts of Interest: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] S. Bell, C. L. Zitnick, K. Bala and R. Girshick, “Insideoutsidenet: Detecting objects in context with skip poolingand recurrent neural networks,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 2874–2883, 2016.

[2] G. Cao, X. Xie and W. Yang, “Feature-fused SSD: Fast detection for small objects,” in *Ninth Int. Conf. on Graphic and Image Processing*, vol. 1, no. 1, pp. 1–8, 2018.

[3] E. Christian, B. Anton, W. Stephan, Z. Dan and L. A. Rainer, “Closer look: Small object detection in Faster R-CNN,” in *IEEE Int. Conf. on Multimedia and Expo*, vol. 1, no. 1, pp. 421–426, 2017.

[4] J. Dai, Y. Li, K. He and J. Sun, “R-FCN: object detection via region-based fully convolutional networks,” *Advances in Neural Information Processing Systems*, vol. 29, no. 1, pp. 379–387, 2016.

[5] C. Eggert, A. Winschel, D. Zecha and R. Lienhart, “Saliency-guided selective magnification for companylogo detection,” in *IEEE Int. Conf. on Pattern Recognition*, vol. 1, no. 1, pp. 651–656, 2016.

[6] M. Everingham, L. van Gool, C. K. I. Williams, J. Winn and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[7] C. Fu, W. Liu and A. Ranga, “DSSD: Deconvolutional single shot detector,” *Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 1–11, 2017.

[8] R.Girshick, J. Donahue, T. Darrell and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 580–587, 2014.

[9] R. Hong, W. H. Cheng and T. Yamasaki, “Small object detection using deep feature pyramid networks,” *Advances in Multimedia Information Processing*, vol. 3, no. 11166, pp. 554–564, 2018.

[10] T. Lin, P. Goyal, R. B. Girshick, K. He and P. Dollár, “Focal loss for dense object detection,” in *IEEE Int. Conf. on Computer Vision*, vol. 1, no. 1, pp. 2999–3007, 2017.

[11] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona *et al.,* “Microsoft COCO: common objects in context,” in *European Conf. on Computer Vision*, vol. 1, no. 1, pp. 740–755, 2014.

[12] S. Liu, D. Huang and Y. Wang, “Receptive field block net for accurate and fast object detection,” *European Conf. on Computer Vision*, vol. 1, no. 1, pp. 1–16, 2017.

[13] W. Liu, B. Leibe and J. Matas, “SSD: single shot multibox detector,” in *European Conf. on Computer Vision*, vol. 1, no. 9905, pp. 21–37, 2016.

[14] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng and S. Yan, “Perceptual generative adversarial networks for small object detection,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 1951–1959, 2017.

[15] M. Menikdiwela, C. Nguyen, H. Li and M. Shaw, “CNN-based small object detection and visualization with feature activation mapping,” in *Int. Conf. on Image and Vision Computing New Zealand*, vol. 1, no. 1, pp. 1–5, 2017.

[16] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis, “SSH: Single stage headless face detector,” in *IEEE Int. Conf. on Computer Vision*, vol. 1, no. 1, pp. 4885–4894, 2017.

[17] J. Redmon, A. Farhadi, “YOLO9000: better, faster, stronger,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 6517–6525, 2017.

[18] S. Ren, K. He, R. B. Girshick and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligenc*e, vol. 39, no. 6, pp. 91–99, 2015.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh *et al,* “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[20] G. Wang, Z. Xiong and D. Liu, “Cascade mask generation framework for fast small object detection,” in *IEEE Int. Conf. on Multimedia and Expo*, vol. 1, no. 1, pp.1–6, 2018.

[21] J. Wang, Y. Yuan and G. Yu, “Face attention network: an effective face detector for the occluded faces,” *Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 1–10, 2017.

[22] S. Yang, Y. Xiong, C. C. Loy and X. Tang, “Face detection through scale-friendly deep convolutional networks,” *Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 1–12, 2017.

[23] Z. Wang, S. Chang, Y. Yang, D. Liu and T. S. Huang, “Studying very low resolution recognition using deep networks,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 4792–4800, 2016.

[24] S. Zhang, L. Wen and X. Bian, “Single-shot refinement neural network for object detection,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 1, pp. 4203–4212, 2018.

[25] S. Zhang, X. Zhu, X. Lei, H. Shi, X. Wang and S. Z. Li, “S3FD: Single shot scale-invariant face detector,” in *EEE Int. Conf. on Computer Vision*, vol. 1, no. 1, pp. 192–201, 2017.