Ship Detection Based on RetinaNet-Plus for High-Resolution SAR Imagery

Hao Su, Shunjun Wei*, Mengke Wang, Liming Zhou, Jun Shi, Xiaoling Zhang
School of Communication and Information Engineering, University of Electronic Science and Technology of China,
Chengdu,Sichuan,P.R.China
*Email: weishunjun@uestc.edu.cn

***Abstract*—Ship detection in high-resolution synthetic aperture radar (SAR) imagery is a fundamental and challenging problem due to the complex environments. In this paper, a RetinaNet-Plus method is presented for ship detection in high-resolution SAR imagery based on RetinaNet network modified. In this approach, instead of setting the score for neighboring region proposals to zero as in Non-Maximum Suppression (NMS), Soft-NMS decreases the detection scores as an increasing function of overlap to avoid loss of precision. In addition, focal loss is used to address the class imbalance and to increase the importance of the hard examples during training. The experiments on SAR ship SSDD dataset and TerraSAR-X image from Barcelona port, show that our method is more accurate than the existing algorithms and is effective for ship detection of high-resolution SAR imagery.**

Keywords—Ship detection, focal loss, Soft-NMS, high-resolution SAR imagery, RetinaNet-Plus

I. INTRODUCTION

Synthetic Aperture Radar (SAR) has all-day, all-weather capabilities and it has the advantages of penetrating cloud rain and independent of the light source compared with traditional optical imaging technology. In recent years, with the rapid development of imaging technology in the area of remote sensing, many satellites and aerial sensors provide SAR imagery with high-resolution. These images facilitate a wide range of applications in the field of defense, environmental management, homeland resource exploration, and natural disaster monitoring, where targets are often involved [3].

As a fundamental problem faced for remote sensing image analysis, object detection in remote sensing images plays an important role for both military and civilian applications. However, it's still a challenging problem due to the scale diversity, visual specificity, small target problem, multi-directional problem and background complexity of high-resolution remote sensing images. In addition, images in quantity and quality create an extremely high computational costs, which also increases the difficulties of object detection for near-real-time applications [2]. Recently, Ship detection is an important topic in the field of SAR.

Ship detection is a complex problem, requiring the solution of two main tasks. First, the detector must solve the recognition problem, to distinguish foreground ships from the background and assign them the proper object class labels. Second, the detector must solve the localization problem, to assign accurate bounding boxes to different ships.

In recent years, with the rapid development of deep learning, researchers in the computer vision field have made major breakthroughs in object detection. At present, the mainstream object detection algorithms are mainly based on deep learning models, which can be divided into two categories: (1) two-stage detection algorithm, which divides the detection problem into two stages, first generating region proposals that filters most of the negative samples, then candidate region classification (generally needs to be refined for location). Typical examples of such algorithms are R-CNN algorithms based on region proposal, such as R-CNN [8], Fast R-CNN [13], Faster R-CNN [4], Mask-RCNN [5],etc.; (2) one-stage detection algorithm, which does not require the region proposal stage, directly generates the class probability and position coordinate values of the object, and compares typical algorithms such as YOLO [10] and SSD [11]. Two-stage has the best object detection results, but one-stage is faster. Both of these methods have the problem of class imbalance. The two-stage solves the class imbalance by means of Selective Search [17], EdgeBoxes [18], DeepMask [19, 20] and RPN [4]. Usually, the number of candidate regions is 1-2k, and a fixed foreground-to-background ratio or online hard example mining [21] are performed to maintain the balance between foreground and background in the classification. The one-stage method has 100k different size locations. Due to the majority of the background that is easily distinguished during training, this can cause training inefficiency. Thus, the reference [16] proposes Focal loss and applies it to the one-stage method to solve the extreme category imbalance problem through the loss function.

Objects in large-scale SAR imagery are relatively small in size and appear in densely distributed groups. For ships, the shapes of the same ship are multi-scale due to the influence of various resolutions, and ships with various shapes display differently in the same resolution SAR imagery. Therefore, it is necessary to consider the variance of ship scales. In order to build high-level semantic feature maps at all scales, we apply a feature pyramid structure (FPN) [6] as a backbone network for feature extraction. FPN uses a top-down architecture to fuse the feature of different resolutions from a single-scale input, which improves accuracy with marginal cost.

In this paper, we introduce a RetinaNet-Plus for ship detection in high-resolution SAR imagery. RetinaNet-Plus has three components: a backbone network for feature extraction and two sub-networks (one for classification and the other for box regression). In SAR ship detection dataset (SSDD), there are some dense ship objects. Soft-NMS is introduced into our framework for avoiding loss of precision due to set the score for neighboring region proposals to zero as in NMS, which improves the detection performance of dense ships. Furthermore, focal loss is used to address the class imbalance and to increase the importance of the hard examples during training. Finally, we compare with state-of-the-art deep CNN based methods, demonstrate the effectiveness of the proposed method on SSDD dataset. To better evaluate our proposed method, one TerraSAR-X imagery from Barcelona port are used to test the robustness.

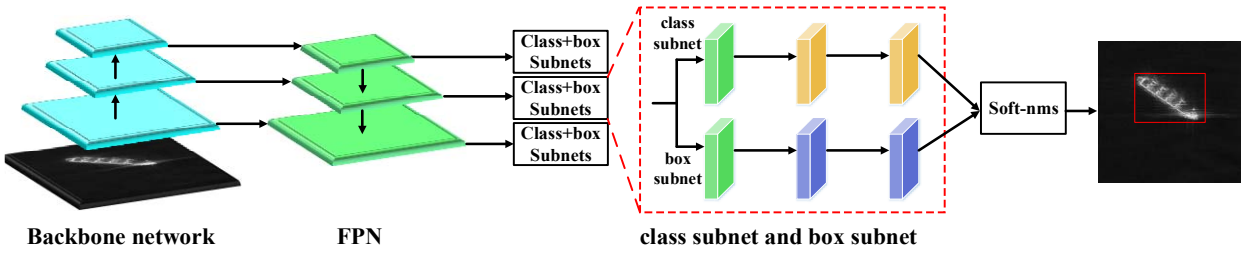

Fig. 1. The architecture of the RetinaNet-Plus method.

II. THE METHOD

A. Focal Loss

RetinaNet belongs to the one-stage object detectors in deep learning. In addition, During the training, the class imbalance and unequal contribution of hard and easy example to the loss have an impact on the detection accuracy in the one-stage object detection scenario. To counter this, the focal loss is proposed in reference [16]. It puts more emphasis on hard example and focuses on the fact that the loss of hard example is higher during training. It is expressed as Equation (1) and has been used to improve the detection accuracy [16].

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

Where α_t and γ are two hypermeters and they function as they function as the role of moderating the weights between easy and hard example.

$$p_t = \begin{cases} p & if\ y=1 \\ 1-p & otherwise \end{cases} \tag{2}$$

Where p is the probability estimated by the model and $y=1$ specifies the ground truth. In general α_t should be decreased slightly as γ is increased, we set $\gamma=2$, $\alpha_t=0.25$ in our experiments.

B. Soft-NMS

Non-Maximum Suppression (NMS) is an essential part of the object detection network to predict final object detections from a set of location candidates, which effectively improve detection performance. In state-of-the-art detectors, these proposals are input to a classification sub-network which assigns them class specific scores. Meanwhile, another parallel regression sub-network refines the region proposals, which could lead to cluttered detections since multiple region proposals often get regressed to the same region of interest (ROI). Therefore, even in state-of-the-art detectors, NMS functions is used to obtain the final set of detections as it significantly reduces the number of false positives [12].

However, the major problem with NMS is that it sets the score for neighboring region proposals to zero. In SAR ship detection dataset (SSDD), there are some dense ship objects. In general, a ship sometimes is surrounded by other ships, so the overlap of nearby ships may be present in that overlap threshold, thus being missed and leading to a drop in average precision. To address this problem, instead of setting the score for neighboring region proposals to zero as in NMS, Soft-NMS decreases the detection scores as an increasing function of overlap, which is denoted as follows:

$$s_i = \begin{cases} s_i, & IoU(M,b_i) < T \\ s_i \times f(IoU(M,b_i)), & IoU(M,b_i) \geq T \end{cases} \tag{3}$$

Where s_i denotes the score of detections, M presents the detection box with the maximum score, and b_i denotes the detection box in the remaining detection boxes, $IoU(M,b_i)$ calculates intersection-over-union between two detection boxes, T denotes IoU threshold.

The reassigned score is associated with the overlap between two boxes. When the IoU is low, these two candidate detections have a high probability to be both true positives. Thus, the $IoU(M,b_i)$ should have some effects on the s_i . Specifically, s_i remains unchanged when the overlap is zero. To address this point, the Gaussian penalty function is used in our framework as follows [12]:

$$f(IoU(M,b_i)) = e^{-\frac{(IoU(M,b_i))^2}{\sigma}} \tag{4}$$

In addition, we set $\sigma=0.5$ in our experiments.

C. RetinaNet-Plus

As shown in Fig.1, our framework is based on RetinaNet [16], which has three components: a backbone network for feature extraction and two sub-networks (one for classification and the other for box regression) [16]. We replace Non-Maximum Suppression (NMS) by Soft-NMS [12], which performs as a post-processing step to obtain the final set of detections.

The ships in high-resolution SAR imagery are various in sizes. In order to build high-level semantic feature maps at all scales, we use feature pyramid network (FPN) [6] as a backbone with ResNet [7] or ResNext [9] of depth 50 or 101 layers to obtain multi-scale features, which are used for object classification and box regression. FPN uses a top-down architecture to fuse the feature of different resolutions

TABLE I. DETECTION PERFORMANCE OF OUR METHOD

Model	Backbone	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet	ResNet-50	NMS	58.5	94.1	65.0	54.2	65.8	52.0
		Soft-NMS	**59.5**	**93.9**	**67.7**	**55.4**	**66.7**	**50.4**
	ResNext-50+32x4d	NMS	58.5	93.7	65.4	54.8	65.1	49.4
		Soft-NMS	**59.3**	**93.6**	**67.4**	**55.8**	**65.9**	**50.1**
	ResNet-101	NMS	58.3	93.8	65.5	54.0	65.5	46.5
		Soft-NMS	**59.5**	**94.2**	**68.6**	**55.4**	**66.8**	**54.9**
	ResNext-101+32x4d	NMS	58.0	92.9	66.3	54.4	64.4	53.7
		Soft-NMS	**59.2**	**93.6**	**68.4**	**55.4**	**65.6**	**52.9**

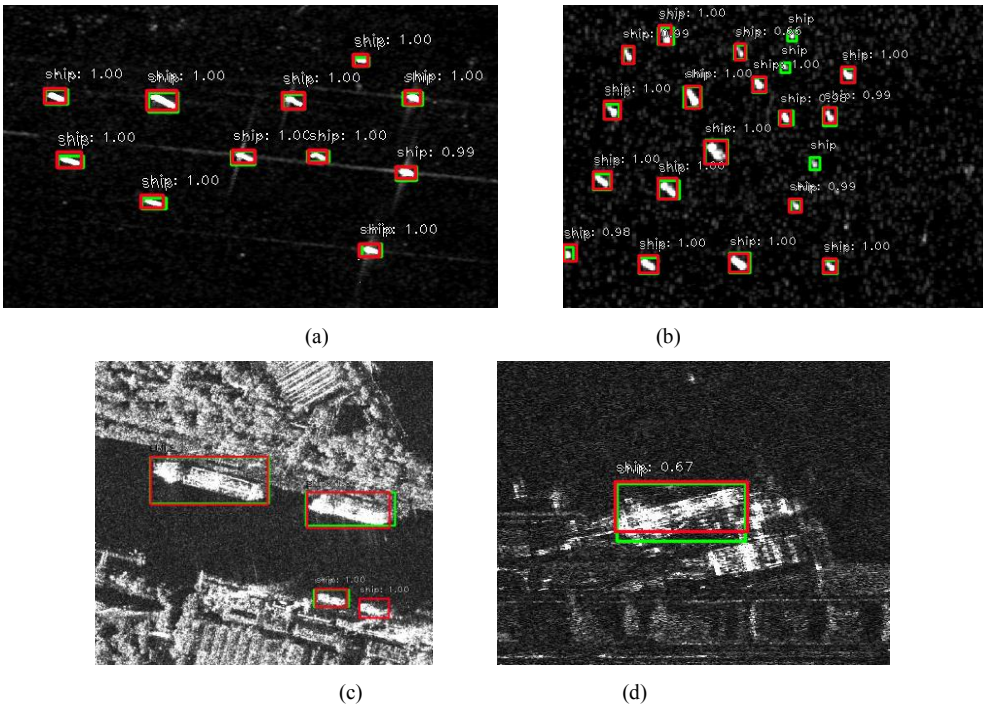

Fig. 2. Results of different scene ship detection in SSDD dataset. (a) and (b) is sea area, (c) and (d) is coast area. (Red boxes denote predicted results; green boxes denote ground truth)

from a single-scale input, which improves accuracy with marginal cost.

RPN is used to generate regions proposals for the class subnet and box subnet. The area of the anchors is best set according to the target size in the statistics set. In this way,

the RPN can handle object of various sizes and aspect ratios. Following the statistical results in SSDD data sets [15] , we assign anchors on different stages depending on the anchor size. Specifically, the area of the anchors are set to $\{8^2, 16^2, 32^2, 64^2, 128^2\}$ pixels on five stages $\{P_3, P_4, P_5, P_6, P_7\}$ respectively. Different aspect ratios $\{1:2, 1:1, 2:1\}$ are also adopted in each stages as in [4].

The classification subnet predicts the probability of object presence at each spatial position for each of the anchors and object classes. This subnet is a small FCN attached to each FPN level; parameters of this subnet are shared across all pyramid levels.

The box regression subnet branch is to provide more accurate bounding boxes for detection, in parallel with the existing branch for classification subnet. In addition, the object classification subnet and the box regression subnet, though sharing a common structure, use separate parameters.

III. EXPERIMENTAL

In this section, we evaluate our method for object detection of high-resolution SAR imagery. We not only compare the object detection performance in terms of average precision (AP) [14], but also show visualized results of our proposed method.

A. Dataset description

SAR ship detection dataset(SSDD) data sets [15] are used in the experiments. SSDD dataset draws on the construction process of PASCAL VOC datasets, including SAR images with different resolutions, polarizations, sea conditions, large sea areas and beaches. This dataset is a benchmark for researchers to evaluate their approaches. In SSDD, there are a totally of 1160 images and 2456 ships. The average number of ships per image is 2.12. In our work, the SSDD data set was randomly divided into 70% for training, and 30% for testing. In order to test our method comprehensively, we expand our dataset by rotating and flipping the image to enhance the image.

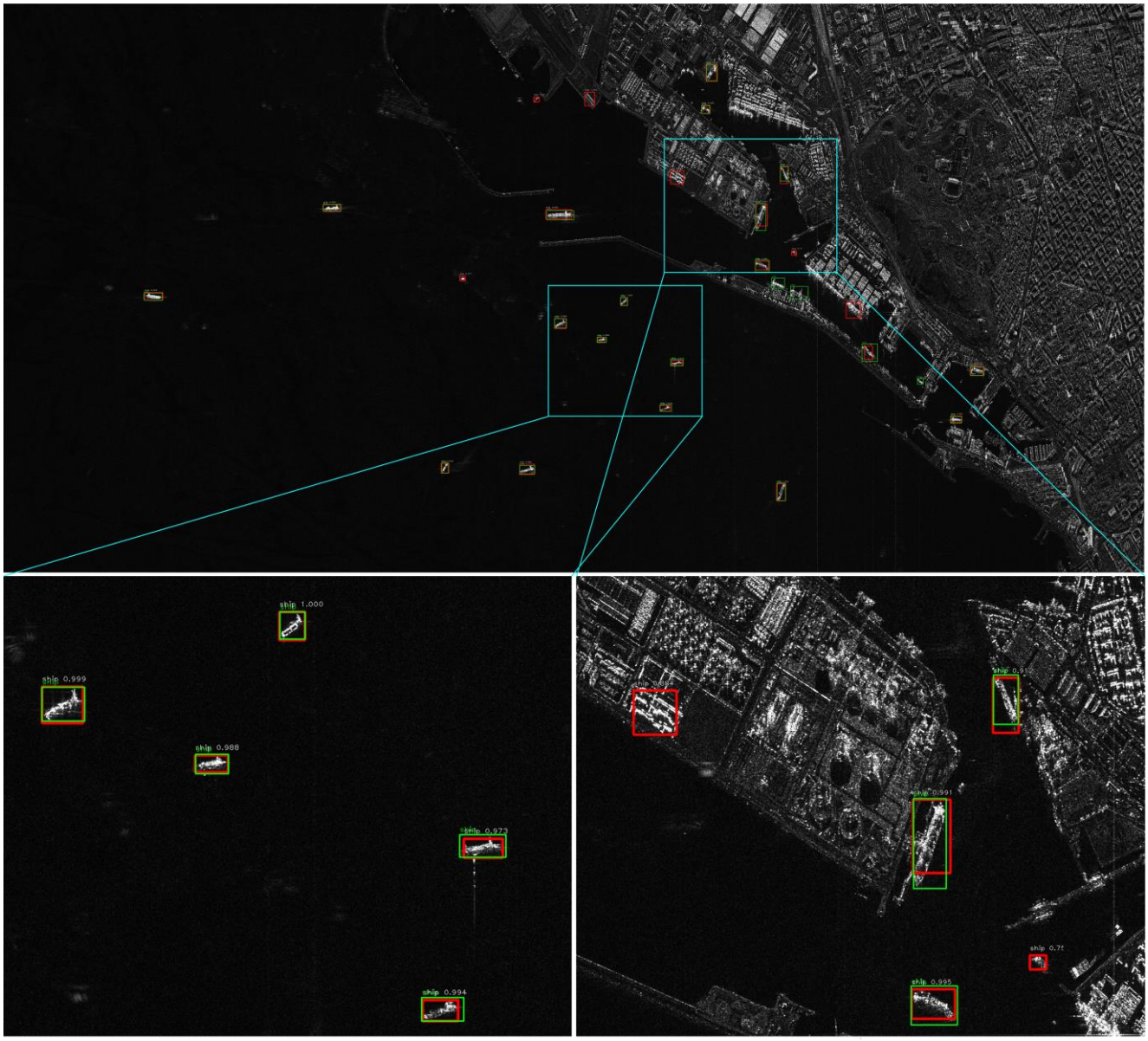

Fig. 3. Ship detection results with the proposed approach on a SAR image. (Red boxes denote predicted results; green boxes denote ground truth)

Furthermore, in order to provide further verification, we evaluated the previous trained detector on a real SAR imagery from Barcelona port. The SAR imagery was acquired from TerraSAR-X sensor, which has a resolution of 2m and a size of 4000×8000 pixels.

B. Experimental Results and Analysis

Fig.2. displays different scene results of ship detection in SSDD dataset, such as sea area and coast area, where the green boxes denote the ground truth, the red boxes represent the predicted results. It shows that our method can accurately detect the ships in a different scene.

TABLE II. THE SHIP DETECTION AP OF TWO MODELS.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv2	50.4	92.9	48.3	52.4	52.5	54.9
Our Method	**59.5**	**94.2**	**68.6**	**55.4**	**66.8**	**54.9**

In TABLE I and TABLE II, we leverage the standard COCO [14] metrics to quantitatively evaluate object detection, including AP, AP_{50}, AP_{75}, AP_S, AP_M, AP_L [14]. Here, AP50 means that the threshold of IoU [14] is set as 0.50. AP_{75} means that the threshold of IoU is set as 0.75. AP represents that the threshold of IoU is set from 0.50 to 0.95, with a step of 0.05. AP_S, AP_M, AP_L are AP for small, medium and large objects. Compared to YOLOv2 [1], the proposed approach has the best performance with a mean AP value of 59.5%. Since the ships are too close and dense, the IoU of their bounding boxes has reached the overlap threshold, thus making one of the ships suppressed. With the help of the Soft-NMS algorithm, our network performs better, which achieves nearly 1% performance gains in terms of mean AP. It can be seen from TABLE I that our framework with Soft-NMS can improve the accuracy, thus making the AP of our method increasing.

Fig.3 indicates the qualitative result on the TerraSAR-X test image from Barcelona port, where the green boxes denote the ground truth of ship target, the red boxes represent the predicted results of ship detection. To see the detection results more clearly, we enlarged two small regions denoted by cyan rectangles. From this figure, we can conclude briefly that (1) Whether in the sea area or coast area, most of the

ships has been successfully detected, which shows that our method is effective and useful. It is noteworthy that our method takes a large-scale and high-resolution SAR imagery as input, and outputs the ship detection results directly. As shown in this figure, there are nearly no false alarms on land area. Therefore, it has great potential for wide field application. (2) For the coast areas, ships are small in size and dense with complex environments, our method still achieved satisfying detection performance, which demonstrates that our method is effective for dense and small size ships.

IV. CONCLUSIONS

In this paper, we proposed a RetinaNet-Plus based method for ship detection in high-resolution SAR images. Soft-NMS is introduced into our framework for avoiding loss of precision due to set the score for neighboring region proposals to zero as in NMS, which improves the detection performance of dense ships. Furthermore, focal loss crack the class imbalance and unequal contribution of hard and easy examples to the loss. The approach is evaluated on SSDD dataset and one TerraSAR-X image from Barcelona port to demonstrate its capability of detecting ships. The results show that our framework is more robust to high-resolution SAR images than RetinaNet and YOLOv2. Meanwhile, our method achieved better performance for detecting ships with large-scale and high-resolution SAR imagery. Therefore, it has great potential for wide field application. To improve the accuracy and reduce the effort to design the anchor boxes, we will propose a general approach to optimize anchor boxes for ship detection in the future.

ACKNOWLEDGMENT

The authors would like to thank Jianwei Li, who generously provided SSDD data set. This work was supported by the National Natural Science Foundation of China (61501098) and the High Resolution Earth Observation Youth Foundation (GFZX04061502).

REFERENCES

[1] Chang Y L, Anagaw A, Chang L, et al. Ship Detection Based on YOLOv2 for SAR Imagery[J]. Remote Sensing, 2019, 11(7): 786.

[2] Deng Z, Sun H, Zhou S, et al. Multi-scale object detection in remote sensing imagery with convolutional neural networks[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 145: 3-22.

[3] Shahzad M, Maurer M, Fraundorfer F, et al. Buildings Detection in VHR SAR Images Using Fully Convolution Neural Networks[J]. IEEE transactions on geoscience and remote sensing, 2019, 57(2): 1100-1116.

[4] Ren, Shaoqing, et al. "Faster R-CNN: towards real-time object detection with region proposal networks." International Conference on Neural Information Processing Systems , pp. 91-99, 2015.

[5] He, Kaiming, et al. "Mask R-CNN." IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 99, pp. 1-1, 2017.

[6] Lin, Tsung-Yi, et al. "Feature Pyramid Networks for Object Detection." CVPR. Vol. 1. No. 2. 2017.

[7] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[8] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[9] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[10] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016.

[11] Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision. Springer, Cham, pp. 21-37, 2016.

[12] Bodla N, Singh B, Chellappa R, et al. Soft-NMS--Improving Object Detection With One Line of Code. Proceedings of the IEEE International Conference on Computer Vision. 2017: 5561-5569.

[13] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision, pp. 1440-1448, 2015.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. pp. 740–755, 2014.

[15] Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14; pp. 1–6, 2017.

[16] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[17] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W.Smeulders. Selective search for object recognition. IJCV, 2013. 2, 4.

[18] C. L. Zitnick and P. Doll´ar. Edge boxes: Locating object proposals from edges. In ECCV, 2014. 2.

[19] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In NIPS, 2015. 2, 4.

[20] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Doll´ar. Learning to refine object segments. In ECCV, 2016. 2.

[21] A. Shrivastava, A. Gupta, and R. Girshick. Training regionbased object detectors with online hard example mining. In CVPR, 2016. 2, 3, 6, 7.