Semantic Attention-Based Network for Inshore SAR Ship Detection

Wenhao Sun[a, b], Xiayuan Huang[*b]
[a]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China; [b]State Key Lab of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ABSTRACT

The performance of Synthetic Aperture Radar (SAR) ship detector has been significantly improved with the development of convolutional neural network. However, the issue of effective detection of inshore ships is still a challenging problem. In this paper, we propose a novel one-stage SAR ship detector, called Semantic Attention-Based Network (SANet), which can largely improve the accuracy of ship detection in the inshore scenario without compromising the speed. Specifically, we introduce a semantic attention mechanism, which will highlight the features from the ships area and enhance the detector's classification ability. We train the proposed Semantic Attention Module with focal loss, and assign labels for the attention maps by center sampling. Combined with our anchor assign strategy, our SANet achieves state-of-the-art results on the open SAR Ship Detection Dataset (SSDD).

Keywords: Semantic Attention-Based Network (SANet), SAR images, ship detection, inshore scenario

1. INTRODUCTION

As an active microwave imaging sensor, airborne and spaceborne Synthetic Aperture Radar (SAR) is able to detect objects in all-weather and all-day to generate high resolution images. Ship detection in SAR images is an essential step for many real applications, *e.g.,* traffic control and maritime surveillance.

Starting from the pioneering works of Constant False Alarm Rate (CFAR) based SAR ship detectors[1, 2, 3], SAR ship detection has achieved a large number of progress, especially the convolutional neural networks (CNN) based detectors[4, 5, 6]. Cui *et al.*[4] introduced Dense Attention Pyramid Networks for SAR ship detection. Lin *et al.*[5] applied squeeze and excitation mechanism to Faster R-CNN[7] for ship detection in SAR images. Fu *et al.*[6] proposed feature balancing and refinement network (FBR-Net) via anchor-free strategy for multiscale SAR ship detection. However, as shown in Figure 1, since inshore ships suffer a more complex background than offshore ships, the inshore ship detection is still a challenging problem and few works have paid attention to this issue[8, 9]. Zhai *et al.*[8] utilized saliency and context information for inshore SAR ship detection. Liu *et al.*[9] presented a multi-scale full convolutional network and a rotatable bounding box to deal with SAR ship targets in the inshore scenario.

In recent years, a number of methods based on CNN have been proposed for the optical object detection and have achieved satisfactory performance. These methods are mainly composed of two categories: the one-stage methods (*e.g.,* SSD[11]) and the two-stage methods (*e.g.*, Faster R-CNN[7]). The former owns high efficiency while the latter has high accuracy. Recently, Zhang *et al.*[12] proposed an object detector named RefineDet, which consists of two interconnected modules (*anchor refinement module* and *object detection module*). In [12], it has been demonstrated that RefineDet can achieve high accuracy and efficiency in optical object detection tasks.

In this paper, we propose an effective one-stage ship detector called SANet which introduces an attention mechanism to RefineDet for SAR ship detection. The attention algorithm can highlight the features of the ships region, particularly for inshore ships. Therefore, the introduction of the attention algorithm can help improve the discriminative ability of the original RefineDet, and boost its classification capability especially for inshore ships. More specifically, following the similar settings as RefineDet, we adopt feature pyramid network to solve different scales ships using different layers. Then, we introduce an semantic-segmentation-style side branch called Semantic Attention Module (SAM). The classification score indicates how much confident the image region within the receptive field of one location to contain ships with specific scales. The attention maps are then dot-multiplied with the feature maps to highlight the features from

[*] xiayuan.huang@ia.ac.cn

the ships region and make the feature maps more discriminative for classification. Moreover, our anchor setting is designed based on the statistics analysis of the open SAR ship detection dataset SSDD[13]. Our SANet achieves the state-of-the-art performance on SSDD, especially in the inshore scenario.

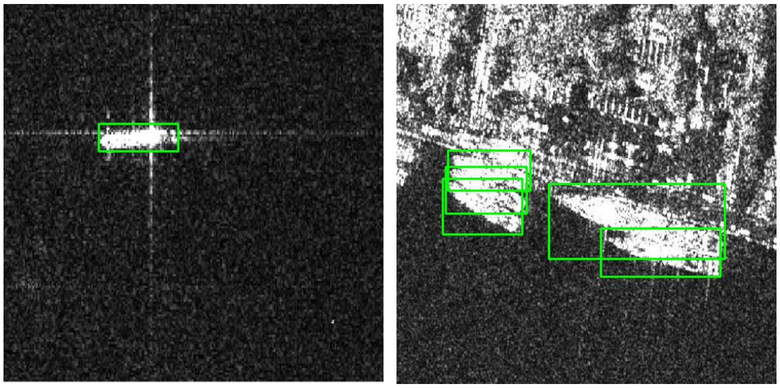


Figure 1. Examples of offshore ship (left) and inshore ship (right) in SAR images.

2. METHODOLOGY

In this section, we first introduce the Semantic Attention Module (SAM), which is the key element of our method, then describe the framework of the network based on SAM for the inshore SAR ship detection.

2.1 Semantic Attention Module

According to recent center-field-based anchor-free object detectors[14, 15], classification branch can learn to predict object categories at every location of feature maps, which can be further used as semantic information. In SAR ship detection settings, there are only two categories to be predicted, *i.e.*, ships and background. The semantic information of ships can be utilized as spatial attention map to help model to focus on ships area and ignore the disturbance of seashores in background. The semantic segmentation technique has been proved useful in face detection task[20, 21].

2.1.1 Semantic Attention Subnet

As shown in Figure 2, according to the classification branch of FCOS[14], we construct the semantic attention subnet consists of two parts: (1) three convolutional modules, in which each consist of one 3×3 convolution layer with 256 output channels, one group normalization layer and a ReLU activation layer, (2) a 3×3 convolution layer with 1 output channel where 1 means the number of foreground object category. For SAR ship detection, we utilize sigmoid activation function at the end of the subnet. All convolution layers in this subnet share parameters across all pyramid levels in our network, except for the first layer since the channel numbers of feature maps are not the same at all levels.

2.1.2 Attention Function

As depicted in Figure 2, let $F_i, G_i \in R^{H\times W\times C}$ be the feature maps at layer i in our backbone. After obtaining the attention maps outputted by the semantic attention subnet, we first deliver them to a sigmoid activation function, then feed them to an exponential operation to rescale the score value from 1 to e, finally dot them with feature maps[20, 21]. The computation process is summarized as

$$SAM\left(F_i, G_i\right) = \exp\left(\sigma\left(H\left(F_i\right)\right)\right) \otimes G_i \tag{1}$$

where σ is the sigmoid function, H is the multilayer convolutional subnet described in Section 2.1.1, $\otimes$ denotes the element-wise multiplication. Some of the attention maps are illustrated in Figure 4. Our attention function can enhance the effect of ships area in feature maps, and weaken that of the seashores background.

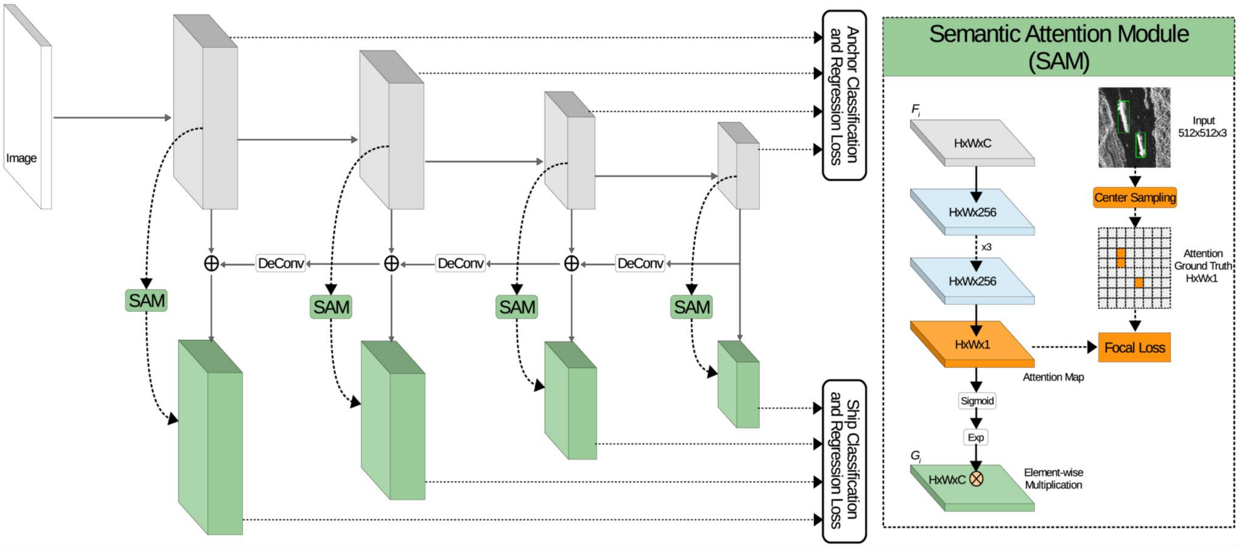


Image
DeConv
DeConv
DeConv
SAM
SAM
SAM
SAM
Anchor Classification and Regression Loss
Ship Classification and Regression Loss
Semantic Attention Module (SAM)
F_i
HxWxC
HxWx256
x3
HxWx256
HxWx1
Attention Map
Sigmoid
Exp
G_i
HxWxC
Element-wise Multiplication
Input 512x512x3
Center Sampling
Attention Ground Truth HxWx1
Focal Loss

Figure 2. The framework of SANet. We only display the layers used for detection. The Semantic Attention Module associated with each detection layer is presented on the right side, which takes center sampling strategy and focal loss in training.

2.1.3 Label Assignment

For feature layer i, we collect all the ground truth bounding boxes B_i, which are matched with preset anchors of layer i. Given an valid ground truth bounding box $(x_1, y_1, x_2, y_2) \in B_i$, the box center (c_x, c_y), width w and height h are

$$c_x = 0.5(x_2 + x_1), c_y = 0.5(y_2 + y_1) \tag{2}$$

$$w = x_2 - x_1, h = y_2 - y_1 \tag{3}$$

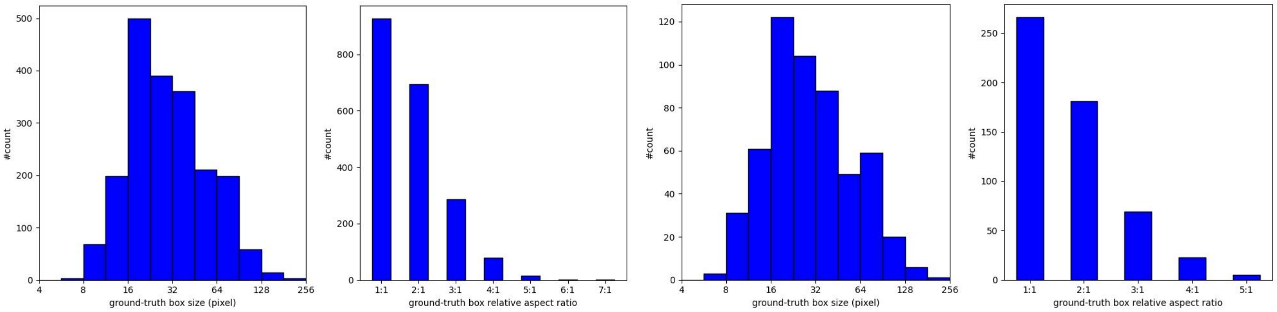
For each location (x, y) on the feature map F_i, it is considered as a positive sample if it falls into the center region of a ground truth bounding box belong to B_i. This strategy is called *center sampling*, which is effective in recent works of object detection[14, 15]. The center region of a ground truth bounding box is defined as

$$(c_x, c_y, \delta w, \delta h) \tag{4}$$

where δ is set to 0.2 in this work. The class label of location (x, y) is set to 1 if it is a positive sample, and 0 otherwise. The assignment process has been described in Figure 2.

2.2 Semantic Attention-Based Network (SANet)

To highlight the features of ships region, particularly for inshore ships, the proposed SANet integrates the above SAM and our baseline RefineDet[12]. As illustrated in Figure 1, the architecture of SANet uses the same feature pyramid backbone as RefineDet, and the extended two phases prediction (*anchor refining phase* and *ship detecting phase*). The feature pyramid backbone can generate fused multiscale feature maps using deconvolution for better classification and regression of multiscale ships. The two phases prediction can filter out easy negative examples to reduce search space for the classifier, and regress bounding boxes using refined anchors. We utilize the same data augmentation and hard negative mining strategies as RefineDet.



(a) Distribution of training set

(a) Distribution of test set

Figure 3. Distribution of box size and relative aspect ratio (longer side : shorter side) in SSDD training and test set. (a) Left: almost 86% box samples in training set have a object size between 16 and 128 pixels. Right: almost 81% box samples in training set have a relative aspect ratio of 1:1 or 2:1. (b) Left: almost 82% box samples in test set have a object size between 16 and 128 pixels. Right: almost 83% box samples in test set have a relative aspect ratio of 1:1 or 2:1.

2.2.1 Anchor Assign Strategy

In SANet, we assign different scales to anchors of different detection layers. We analyze the statistic distribution of ground truth box size and aspect ratio for the SSDD training set. As shown in Figure 3(a), almost 86% ships have an object scale from 16 to 128 pixels, and almost 81% ships have an relative aspect ratio (longer side : shorter side) of 1:1 or 2:1. Considering ships will be enlarged after data augmentation, we set our anchor scales as [32, 64, 128, 256], and anchor aspect ratios as [0.5, 1, 2] for better positive samples selecting and easier boxes regressing. During training, we first match each ship to the anchor box with the best jaccard overlap score, then match the anchor boxes to any ships with jaccard overlap higher than 0.5.

2.2.2 Loss Function

We adopt a multi-task loss to jointly optimize SANet's parameters:

$$\begin{aligned} L = {} & \lambda_1 \frac{1}{N_{arp}} \sum_k \left(\sum_{i \in A_k} L_b\left(n_i, l_i^*\right) + \sum_{i \in A_k} l_i^* L_r\left(d_i, g_i^*\right) \right) \\ & + \lambda_2 \frac{1}{N_{sdp}} \sum_k \left(\sum_{i \in A_k} L_b\left(o_i, l_i^*\right) + \sum_{i \in A_k} l_i^* L_r\left(b_i, g_i^*\right) \right) \\ & + \lambda_3 \sum_k L_a\left(s_k, s_k^*\right) \end{aligned} \tag{5}$$

where k is the index of an feature level, A_k is the set of anchors assigned in feature level k, i is the index of anchor in A_k, l_i^* is the ground truth class of anchor i (1 for ship, 0 for background), g_i^* is the ground truth position and size of anchor i. n_i and d_i are the predicted confidence of anchor i for two classes and refined coordinates of anchor i in the *anchor refining phase*. o_i and b_i are the predicted probability for two classes and coordinates of bounding box in the *ship detection phase*. N_{arp} and N_{sdp} are the number of positive samples in the above two phases. The binary classification loss L_b is the cross-entropy loss over two classes (ships *vs.* background). The regression loss L_r is the smooth L_1 loss.

We use pixel-wise sigmoid Focal Loss[16] as our attention loss L_a. s_k and s_k^* are the attention map generated at level k and the ground truth map described in Section 2.1.3, respectively. λ_1, λ_2 and λ_3 are employed to balance three loss terms, and we set $\lambda_1 = \lambda_2 = \lambda_3$ in this work.

3. EXPERIMENTS

In this section, we firstly introduce the dataset and the detailed experimental settings, then analyze our model in an ablative way, finally compare our model performance with the state-of-the-art methods and introduce its run time performance.

3.1 Dataset and Settings

Our experiments are conducted on the most commonly used SAR ship detection dataset SSDD, which contains 1,160 SAR images with 500×500 average image size. We follow the data split appealed in [10] for fair comparison, and use the 8:2 ratio as training and test split. As shown in Figure 3(b), the distribution of the test set is close to that of the training set, which implies the reasonableness of the split. We adopt the same data augmentation as [12] in our settings, resulting in an 512×512 input size.

SANet takes VGG16 as the primary backbone and uses batch size 16 in training. We set the learning rate to 0.002 for the first 200 epochs, and decay it to 0.0002 and 0.00002 for training another 66 and 34 epochs, respectively. All models are trained from scratch. We implement SANet in Pytorch[17].

The evaluation metric AP of the largest object detection benchmarks MS COCO[18] is adopted in our experiments. To obtain AP, we first calculate a series of average precision by taking a prediction as true positive when its IOU with a ground truth is large than [0.5: 0.05: 0.95] respectively, then take the average of them. AP is a strict metric and can reflect not only the classification ability under strict conditions, but also the regression capability of models.

3.2 The effect of Semantic Attention Module

In this paper, we propose Semantic Attention Module (SAM) to enhance the ship detection performance in SAR image especially in the inshore scenario. We evaluate this component by comparing with the baseline RefineDet, the experiments are carried out on the same settings.

According to Table 1, some promising conclusions can be summarized. After adding SAM to the baseline, we get higher AP in all scenarios, especially 1.1% higher for inshore ship detection, which indicates that the spatial attention provided by SAM can help in promoting the performance of detector especially in the inshore case. As shown in Figure 4, at different feature levels, the attention maps produced by SAM can be activated by objects with different scales. Our SAM is trained to highlight the ships area in the input images. After the attention operation, the feature maps are enhanced and become more discriminative for the classification task. Therefore, our SANet is able to suppress both false positives and false negatives and achieve a better performance.

Table 1. Effectiveness of semantic attention module. Bold fonts indicates the best AP.

Methods	AP (%)		
	All scenarios	Inshore	Offshore
Baseline	60.9	51.7	65.0
SANet (Baseline + SAM)	**61.4 (+0.5)**	**52.8 (+1.1)**	**65.2 (+0.2)**

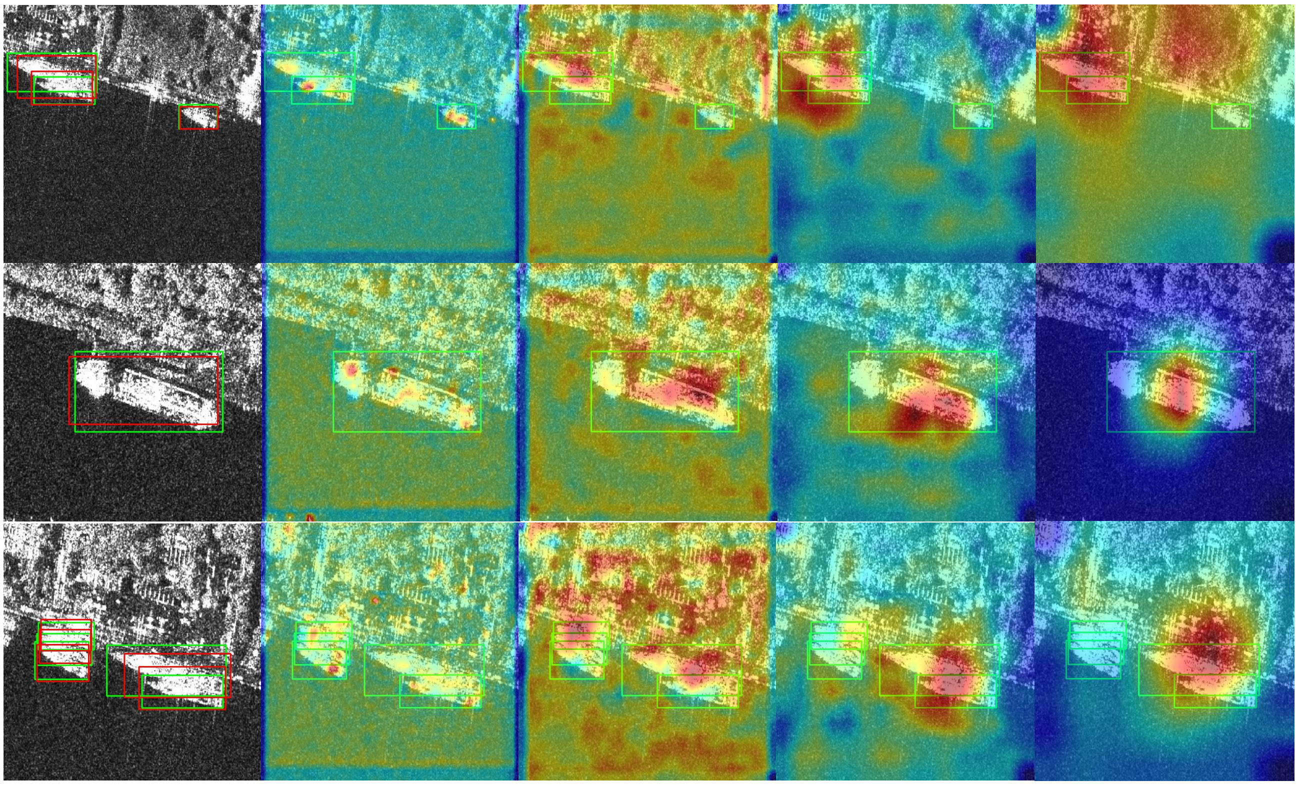


Figure 4. The attention maps of SAM. First column: input images, green boxes means ground truths, red boxes means predictions. Second to last column: attention maps from low detection layer to high layer, respectively.

3.3 Comparison with the state-of-the-art methods

We also report the results of some state-of-the-art methods. The experiments of Faster R-CNN[7], Cascade R-CNN[19], SSD[11] and RefineDet on SSDD are implemented by ourselves.

As shown in Table 2, when compared with both two-stage and one-stage methods, SANet can achieve the highest AP in all scenarios, which indicates that the proposed Semantic Attention Module can provide SANet with better classification ability. Moreover, considering the strict IOU thresholds taken by the AP metric, we can conclude that the regression capability of our model has also been boosted with the help of our attention mechanism. This is an additional gain, but is not the focus of this work.

Table 2. Detection results on SSDD test set. Bold fonts indicates the best AP.

Methods	AP (%)			FPS
	All scenarios	Inshore	Offshore	
two-stage:				
Faster R-CNN	32.0	20.7	37.2	5.2
Cascade R-CNN	59.8	49.8	64.3	14.8
one-stage:				
SSD	56.6	44.5	62.4	28.4
RefineDet	60.9	51.7	65.0	52.3
SADet (Ours)	**61.4**	**52.8**	**65.2**	**45.0**

3.4 Run time efficiency

We present the run time performance at the last column of Table 2. The speed is evaluated with NVIDIA Titan X (Pascal), CUDA 10.1 and cuDNN v7. As shown in Table 2, two stage methods are much more time-consuming than one stage methods. Cascade R-CNN has a higher AP and faster processing speed described by FPS (Frames Per Second) than Faster R-CNN, but the FPS of Cascade R-CNN is only 14.8. SSD has 28.4 FPS while RefineDet achieves the largest 52.3 FPS. Our SANet processes an image in 45.0 FPS, which still keeps relatively high efficiency.

4. CONCLUSION

In this paper, we focus on the problem of SAR ship detection in the inshore scenario. We propose an one-stage semantic attention-based detector, which integrates our semantic attention module and our specifically adapted baseline for ship detection. The semantic attention mechanism can highlight the features from the ships region and enhance the classification ability of our detector. We carry out several experiments on the popular SAR ship detection benchmark SSDD and its challenging inshore subset to demonstrate that SANet achieves the state-of-the-art accuracy and keeps high efficiency.

5. ACKNOWLEDGEMENT

This work is supported in part by the National Natural Science Foundation of China under Grant 61802408, Grant 91648205, Grant 61627808, Grant 91948303 and Grant 61806202.

REFERENCES

[1] Hou, B., Chen, X., and Jiao, L., "Multilayer CFAR Detection of Ship Targets in Very High Resolution SAR Images," IEEE Geoscience and Remote Sensing Letters, 12, 811-815(2015).

[2] An, W., Xie, C., and Yuan, X., "An Improved Iterative Censoring Scheme for CFAR Ship Detection With SAR Imagery," IEEE Transactions on Geoscience and Remote Sensing, 52, 4585-4595(2014).

[3] Leng, X., Ji, K., Yang, K., and Zou, H., "A Bilateral CFAR Algorithm for Ship Detection in SAR Images," IEEE Geoscience and Remote Sensing Letters, 12, 1536-1540(2015).

[4] Cui, Z., Li, Q., Cao, Z., and Liu, N., "Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images," IEEE Transactions on Geoscience and Remote Sensing, 57, 8983-8997(2019).

[5] Lin, Z., Ji, K., Leng, X., and Kuang, G., "Squeeze and Excitation Rank Faster R-CNN for Ship Detection in SAR Images," IEEE Geoscience and Remote Sensing Letters, 16, 751-755(2019).

[6] Fu, J., Sun, X., Wang, Z., and Fu, K., "An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images," IEEE Transactions on Geoscience and Remote Sensing, 59, 1331-1344(2021).

[7] Ren, S., He, K., Girshick, R.B., and Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 1137-1149(2015).

[8] Zhai, L., Li, Y., and Su, Y., "Inshore Ship Detection via Saliency and Context Information in High-Resolution SAR Images," IEEE Geoscience and Remote Sensing Letters, 13, 1870-1874(2016).

[9] Liu, L., Chen, G., Pan, Z., Lei, B., and An, Q., "Inshore Ship Detection in Sar Images Based on Deep Neural Networks," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 25-28(2018).

[10] Zhang, T., Zhang, X., Shi, J., Wei, S., Wang, J., Li, J., Su, H.,and Zhou, Y., “Balance scene learning mechanism for offshore and inshore ship detection in sar images,” IEEE Geoscience and Remote Sensing Letters, 1–5(2020).

[11] Liu, W., Anguelov, D., Erhan, D., et al., "Ssd: Single shot multibox detector," European conference on computer vision. Springer, Cham, 21-37(2016).

[12] Zhang, S., Wen, L., Bian, X., Lei, Z., and Li, S., "Single-Shot Refinement Neural Network for Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4203-4212(2018).

[13] Li, J., Qu, C., and Shao, J., "Ship detection in SAR images based on an improved faster R-CNN," 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), 1-6(2017).

[14] Tian, Z., Shen, C., Chen, H., and He, T., "FCOS: Fully Convolutional One-Stage Object Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 9626-9635(2019).

[15] Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., and Shi, J., "FoveaBox: Beyound Anchor-Based Object Detection," IEEE Transactions on Image Processing, 29, 7389-7398(2020).

[16] Lin, T., Goyal, P., Girshick, R.B., He, K., and Dollár, P., "Focal Loss for Dense Object Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 42, 318-327(2020).

[17] Paszke, Adam, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library." Advances in Neural Information Processing Systems, 32, 8026-8037(2019).

[18] Lin, T. Y., Maire, M., Belongie, S., et al., "Microsoft coco: Common objects in context," European conference on computer vision. Springer, Cham, 740-755(2014).

[19] Cai, Z., and Vasconcelos, N., "Cascade R-CNN: Delving Into High Quality Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6154-6162(2018).

[20] Wang, J., Yuan, Y., and Yu, G., "Face Attention Network: An Effective Face Detector for the Occluded Faces," ArXiv, abs/1711.07246(2017).

[21] Zhuang, C., Zhang, S., Zhu, X., Lei, Z., and Li, S., "Single Shot Attention-Based Face Detector," CCBR(2018).