IR-Net: An improved RetinaNet-based ship detection detector

Chen Feng

School of Information and Computer Engineering, Northeast Forestry University, Harbin, China.

fc2242895736@163.com

Abstract. The task of ship object detection is of great importance to research. Ship object at sea is key to marine monitoring, real-time rescue, and wartime attack. Whether the ship object can be identified quickly and accurately to support the commander's decision-making is largely related to the success of the maritime mission. However, the dense arrangement and arbitrary direction of ships make it more difficult than the general object detection. Therefore, this paper proposes a rotation detector to deal with these problems. Furthermore, ship detection faces some chanllenges such as the complex background of offshore port images, the large ratio of ship length to width, and the potential of being blocked by clouds. Therefore, this paper proposes a rotatable bounding box based on RetinaNet, named Improved RetinaNet (IR-Net). With the deeper ResNet as its backbone network, this network extracts deeper object features. Furthermore, DCM (deformable convolutional module) is added to the network, so that the sampling points of the network can change with the shape of the object ship, thus extracting the object features more effectively. Finally, in this paper, the Cutout data enhancement strategy in the network is added, which effectively improves its detection accuracy in the case of object occlusion. Ablation experiments show that the IR-Net proposed in this paper has a more accurate detection performance on the HRSC2016 ship dataset.

Keywords: IR-Net, RetinaNet, ship detection detector.

1. Introduction

With modern naval battlefield war as an information war undoubtedly, fully detecting and identifying corresponding ships is the premise of understanding each other's war intention[1]. In addition, ship object detection also plays an imperative role in maritime traffic and maritime detection[2]. With the development of satellite remote sensing technology, ship object detection gets more abundant and accurate information. At the same time, the development of imaging technology and artificial intelligence promotes ship object detection based on the convolutional neural network as a hot research trend. The ship object detector is developed from the general object detector, which is usually divided into two-stage and single-stage networks[3]. In the two-stage network, the first stage network is used for candidate region extraction, and the second stage network classifies and regresses the extracted candidate regions with accurate coordinates, such as R-CNN[4] series. Extracting no candidate regions, the single-stage network is sufficient to complete two tasks of classification and regression, such as YOLO [5] and SSD [6].

However, the horizontal (traditional) bounding box is not suitable for ship object detection. On the one hand, it is difficult to reflect the correct shape of the object ship. On the other hand, it is hard to separate the object object from the background. In addition, when encountering dense object objects, the horizontal bounding box fail to well segment the object objects, while the rotatable bounding box can handle these problems[7]. Specific advantages and disadvantages can be seen in Figure 1:

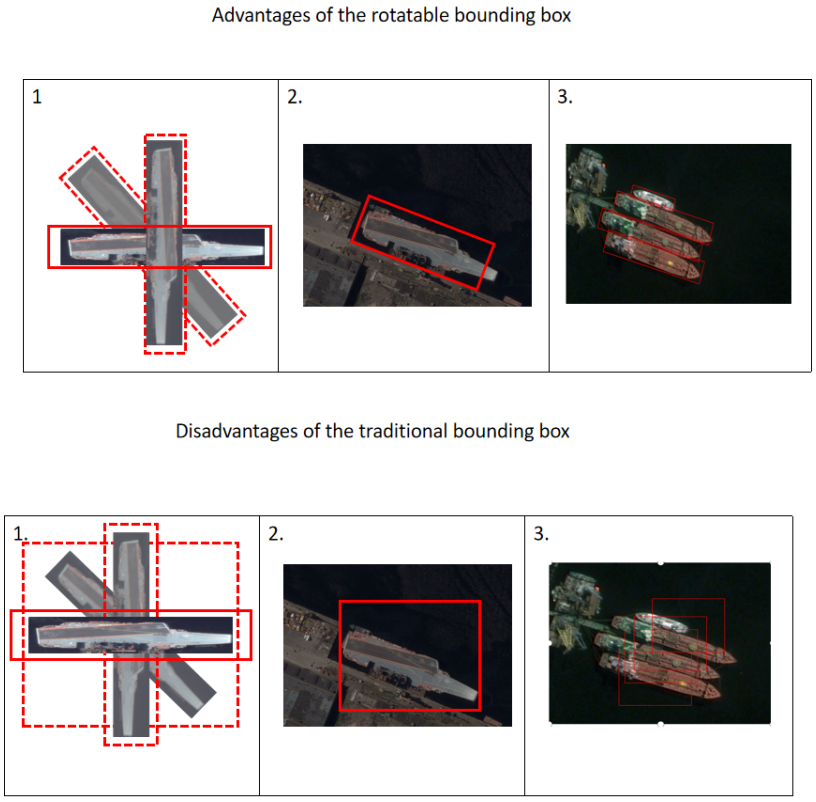

Figure 1. Horizontal v.s. Rotatable Bounding Box

Considering the advantages of the rotatable bounding box in eliminating interference, the Songlin Lei team proposed a ship detection method of rotatable bounding box based on sub-aperture decomposition feature enhancement in SAR images [8].

Wenchao Liu put foward a bounding box for ship detection in remote sensing images based on a convolutional neural network, which is designed to predict the ship bounding box with azimuth information [9].

Tong Wang's team proposed a rotation invariant TSD (RITSD) method, which adopts rotation invariant deformation pooling and angle migration learning when obtaining region of interest (RoI) features. Considering that rotating anchors will increase the computation and training time, they use three trunk sections. First, they use horizontal recommendations to obtain rotation region of interest recommendations. Then, the predicted rotatable bounding boxes are obtained. Finally, a detailed ablation study was carried out on the collected data set and HRSC2016[10].

Given that the ship object is almost invariant in remote sensing images, Linhao Li's team used a new two-branch regression network. Secondly, a shape adaptive pooling method is proposed to overcome the limitation of typical regular ROI pooling in extracting ship features with different aspect ratios. In addition, they advise to fuse multi-level features through adaptive pooling of spatial variation, which is called multi-level adaptive pooling and produces a more compact feature representation more suitable for ship classification and positioning at the same time [11].

However, the detection of rotatable bounding boxed still confronts challenges as follows. Firstly, the background of the offshore port image is complex, the ratio of ship length to width is large, and the potential of being blocked by clouds[12]. In response to these challenges, we have made the following contributions:

First of all, we use a deeper ResNet[13] as the backbone network and increase the number of ResNet layers from 50 to 152. More layers can fully extract the network including not only the low-level features but also the high-level features, which makes the network extracted features more comprehensively. In addition, we also add a deformable convolutional module in the network with an offset that can be adapted to learn. Different from the fixed sampling points of rigid convolution,

the sampling points of deformable convolution can change adaptively according to the shape of the object ship, making the feature extraction more sufficient. Moreover, we also add Cuout data enhancement strategy to randomly generate a 0-mask pixel block on the image, so that the network can learn how to deal with the object detection under occlusion in advance during the training process and select more representative features for extraction.

The structure of this paper is as follows. The second part introduces the proposed network structure and principle, the third part is about experiments, the fourth part is the experimental summary, and the last part is conclusion.

2. Proposed Method

2.1 Overview of Our Proposed Detector

The overall network diagram of this paper is shown in Figure 2, which consists of four parts including the backbone, feature pyramid net, class subnet, and box subnet. The network used in the backbone is ResNet[13] with a better effect when the depth is 152. The deformable convolutional module is also added. In the aspect of data enhancement, we add the Cuout data enhancement strategy.

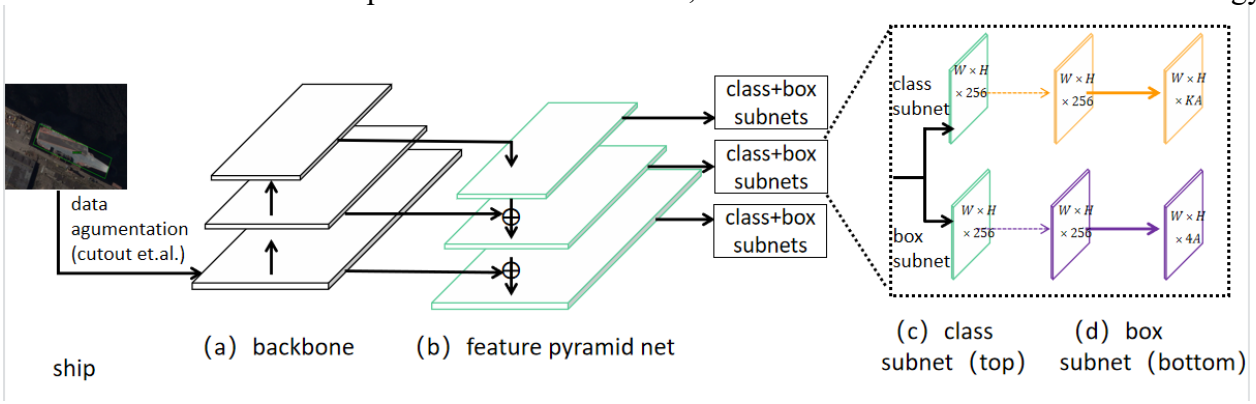

Figure 2. Overall Network Diagram

2.2 ResNet-baesed backbone

In computer vision, the depth of the network is an important factor to achieve good results. However, the phenomenon of gradient explosion and gradient disappearance makes it very difficult to train deep-level networks. Normalization of input data and middle-layer data can only affect the network within dozens of layers. By learning the residual function F (x) = H (x)-x, ResNet enables the accumulation layer to learn new features based on the input features and obtain better results [13].

The structure of the residual block is shown in figure 3 and the expression of the first layer is $\mathcal{F} = W_2\sigma(W_1x)$, where σ represents the nonlinear function ReLU and then $\mathcal{F} = W_2\sigma(W_1x)\sigma y = \mathcal{F}(x,\{W_i\}) + x$ is output after passing through the shortcut and the second ReLU.

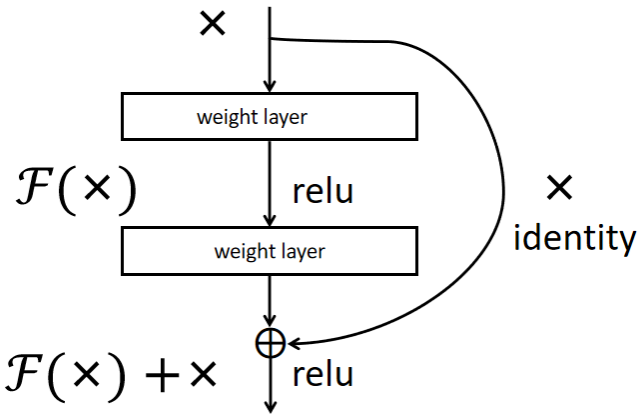

Figure 3. Residual Structure

2.3 (DCM) Deformable Convolution Module

Due to the inherent geometric structure of the CNN module, it is hard to fully realize the modeling of geometric deformation. Therefore, the DCM structure is proposed to improve the modeling of the network to geometric deformation. Figure 4 shows different sampling methods of normal convolution and deformable convolution with a convolutional kernel size of 3x3. Figure (a) shows 9 points sampled under normal convolutional law. (b) (c) (d) is deformable convolution, which adds a displacement to normal sampling coordinates represented by blue arrows. (c) (d) as a special case of (b) shows the special case of deformable convolution as scale transformation, scale transformation, and rotation transformation[14].

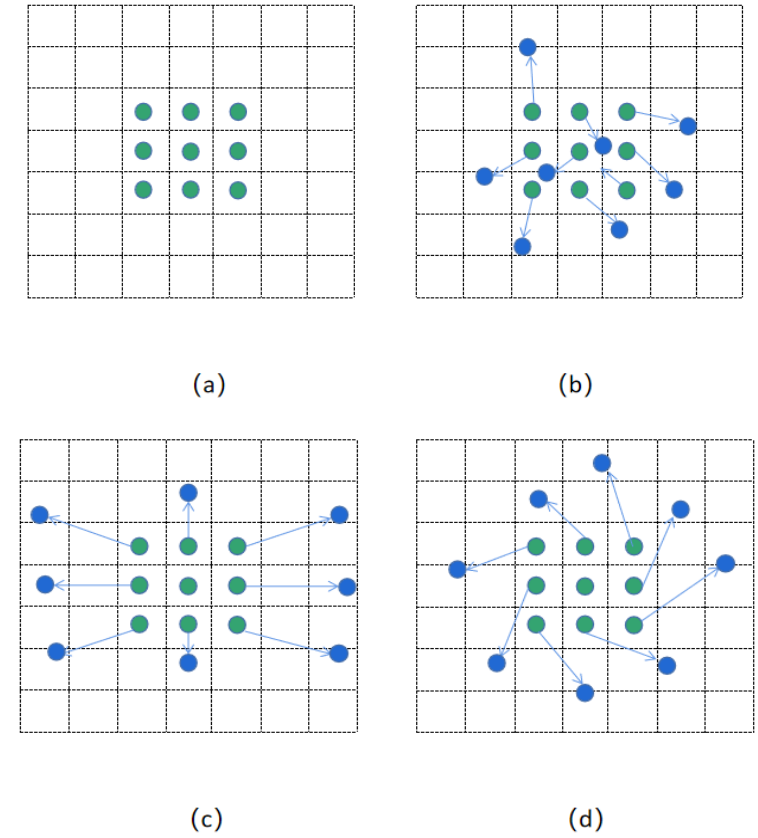

Figure 4. Description of Sampling Position in 3 × 3

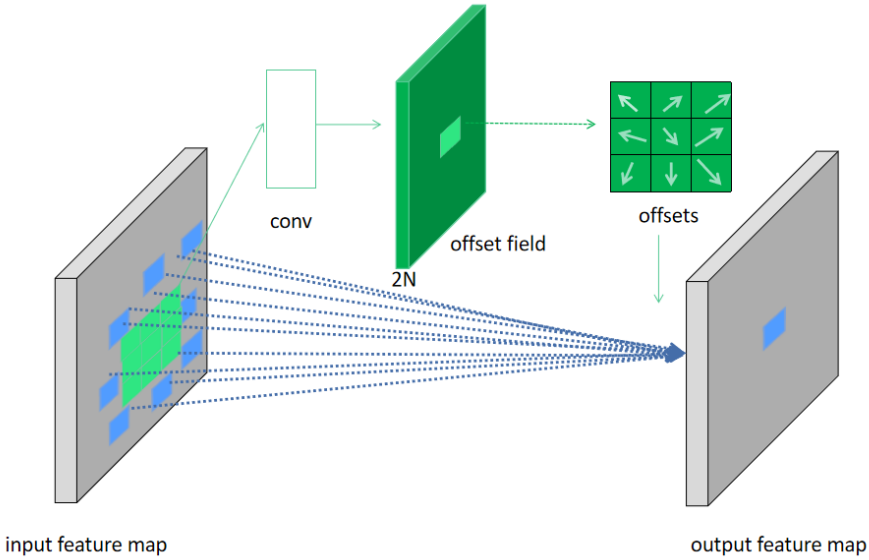

Figure 5. Schematic Diagram of 3*3 Variability Convolution

The following is a schematic diagram of the 3*3 variable convolutional workflow. Assuming that the input feature graph is WH and the variability convolution to be performed is kernel_size=3*3, stride=1, and dialation=1, the offset is first learned using a convolution with the same spatial resolution and expansion rate as the current variability convolutional layer. Convolution then outputs a W×H×2×N offset file. After that, deformable convolutional kernel will perform a convolutional operation according to the offset.[14].

Common convolutional operations are generally divided into two parts:

Sampling on the input feature map using a regular grid $\mathcal{R}$;

The sampling points are weighted, and the size and expansion of the receptive field are defined by $\mathcal{R}$.

$$\mathcal{R} = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$$

For each position p_0 on the feature map of the output, the calculation is performed by the following equation:

$$y(p_0) = \sum_{pn\in\mathcal{R}} w(p_n)\cdot x(p_0 + p_n)$$

Where p_n is the enumeration result of the locations listed in $\mathcal{R}$. However, the operation of variable convolution is different. In the deformable convolutional network, the regular mesh $\mathcal{R}$ is expanded by adding an offset and the output position p_0 becomes:

$$y(p_0) = \sum_{pn\in\mathcal{R}} w(p_n)\cdot x(p_0 + p_n + \Delta p_n)$$

Therefore, the sampling position becomes irregular. For the offset $\Delta p_n$is generally decimal, we need to realize it by bilinear interpolation. The formula is:

$$x(p) = \sum_{q} (q,p)\cdot x(q)$$

2.4 Cuout data augmentation

In the convolutional neural network, Cuout enables CNN to make better use of the global information of the image, instead of relying solely on the characteristics of a certain part. Cutout randomly selects a small square area from an image and sets the pixel value of this area to 0 or other unified values, thus simulating occlusion and improving generalization ability[15]. The renderings are as follows:

Figure 6. Display of Cutout Effect

3. Experiments

3.1 Details of the experiment

The hardware environment of this experiment is 12-core Intel (R) Xeon (R)
Platinum 8255C CPU @ 2.50 GHz and RTX 2080 Ti. The experimental environment is PyTorch 1.6. 0, Python 3.8, and Cuda 10.1 with the learning rate set to 0.001.

The experiment is shown in the following figure.

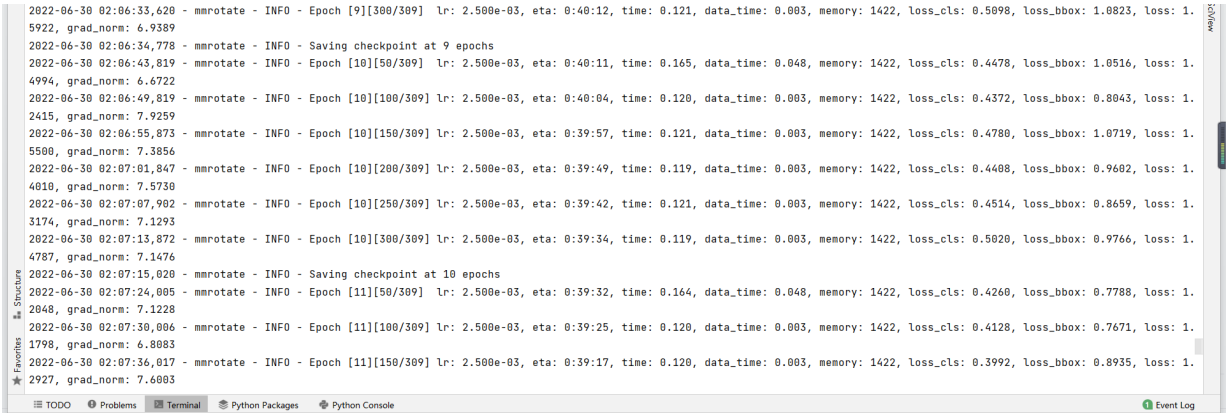

Figure 7. Experiment Demonstration

3.2 Ablation study

In this experiment, the backbone network of RetinaNet is Resnet 50.

3.2.1 ResNet Deep Effectiveness Experiment

Table 1. ResNet Deep Effectiveness Experiment

Settings	MAP	AP50	AP75	AP95
ResNet 50	0.521	0.841	0.592	0.011
ResNet 101	0.522	0.847	0.598	0.03
ResNet 152	0.547	0.84	0.605	0.009

Through this experiment, it can be found that the value of mAP increases gradually with the increase of the depth of ResNet. When the depth is 152, the value of mAP is 0.547, which is 0.026 higher than that when the depth is 50. It can be seen that increasing the depth effectively improves the accuracy of the network.

3.2.2 Validity Experiment of Deformable convolution

Table 2. Validity Experiment of Deformable convolution

Settings	MAP	AP50	AP75	AP95
ResNet 152	0.547	0.84	0.605	0.009
ResNet 152 + DCM	0.548	0.845	0.609	0.006

It can be found that when the deformable convolutional module is added, the mAP of the network increases obviously with the mAP value as 0.548, which is 0.001 higher than that before DCM is added. It shows that the object detection effect under this method is better.

3.2.3 Effectiveness Test of Cutout

Table 3. Effectiveness Test of Cutout

Settings	MAP	AP50	AP75	AP95
ResNet 152 + DCM	0.548	0.845	0.609	0.006
ResNet 152 + DCM + Cuout	0.558	0.852	0.625	0.015

We can find that after adding the Cuout data enhancement strategy, the value of mAP is 0.558, 0.01 higher than that without adding the Cuout data enhancement strategy. Therefore, after adding the Cuout module the accuracy of the network has achieved better results than before.

Some experimental visualizations of ResNet 152 + dcm + Cuout are shown in the following figure.

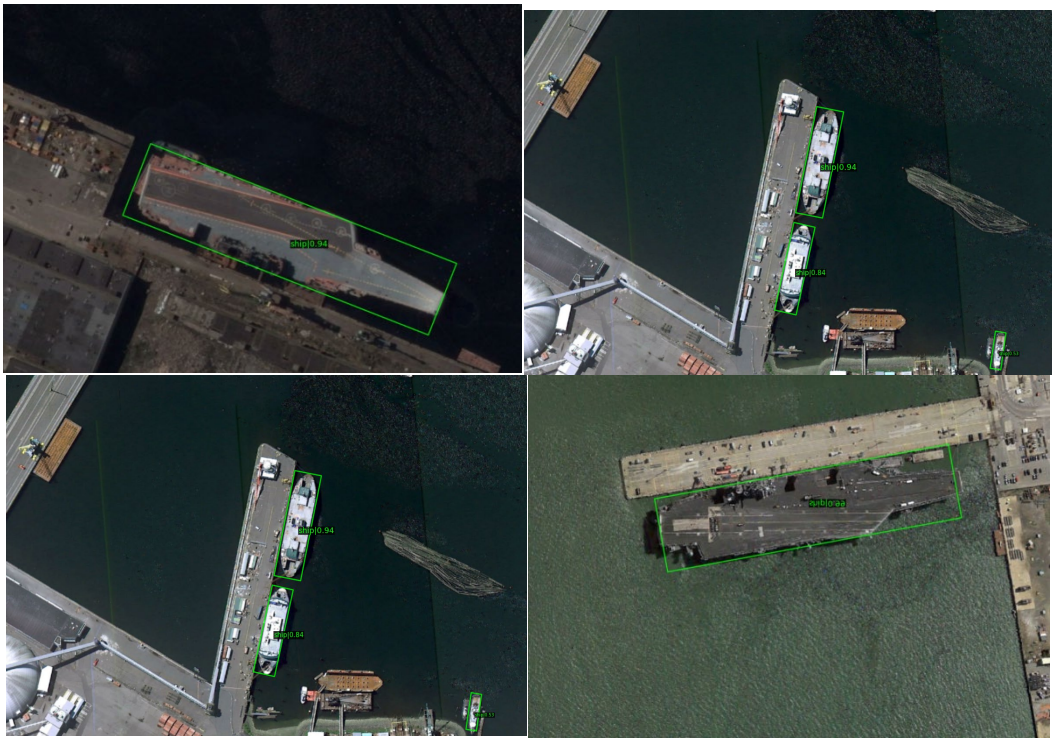

Figure 8. Experimental Visualization Results

4. Conclusion

In this paper, a ship detection network IR-Net based on rotatable RetinaNet is proposed to increase the depth of the network, so that the network can more fully extract object features and increase the deformable convolutional module. In this way, sampling points can change adaptively with the object shapes and features can be more fully extracted. Then the Cuout data enhancement strategy is added so that the network can learn the coping scheme under occlusion and improve the accuracy of object detection. Ablation experiments show that the above improvement measures have achieved excellent detection results.

References

[1] An, Y. (2014). Research on ship object detection and recognition sea battlefield. (Doctoral Dissertation). Harbin Engineering University, Harbin, Heilongjiang. DOI:10.7666/d.D749653.

[2] Liu, G., Zhang, X. & Meng, J. (2019). A small ship object detection method based on polarimetric SAR. Remote Sensing, 11(24), 2938.

[3] Zhao, Z. Q., Zheng, P., Xu, S. et al. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11): 3212-3232.

[4] Girshick, R., Donahue, J., Darrell, T. et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 580-587.

[5] Redmon, J., Divvala, S., Girshick, R. et al. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 779-788.

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

[7] Liu, L., Pan, Z., Lei, B. (2017). Learning a rotation invariant detector with rotatable bounding box. arXiv preprint arXiv:1711.09405.

[8] Lei, S., Qiu, X., Ding, C. et al. (2021). A feature enhancement method based on the sub-aperture decomposition for rotating frame ship detection in SAR images. IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, 3573-3576.

[9] Liu, W., Ma, L. & Chen, H. (2018). Arbitrary-oriented ship detection framework in optical remote-sensing images. IEEE Geoscience and Remote Sensing Letters, 15(6), 937-941.

[10] Wang, T. & Li, Y. (2021). Rotation-Invariant Task-Aware Spatial Disentanglement in Rotated Ship Detection Based on the Three-Stage Method. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-12.

[11] Li, L., Zhou, Z., Wang, B. et al. (2020). A novel CNN-based method for accurate ship detection in HR optical remote sensing images via rotated bounding box. IEEE Transactions on Geoscience and Remote Sensing, 59(1), 686-699.

[12] Liu, Z., Hu, J., Weng, L. et al. (2017). Rotated region based CNN for ship detection. 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 900-904.

[13] He, K., Zhang, X., Ren, S. et al. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

[14] Dai, J., Qi, H., Xiong, Y., et al. (2017). Deformable convolutional networks. Proceedings of the IEEE international conference on computer vision, 764-773.

[15] DeVries, T. & Taylor, G. W. Improved regularization of convolutional neural networks with Cuout. arXiv preprint arXiv, 1708.04552, 2017.