Attention Receptive Pyramid Network for Ship Detection in SAR Images

Yan Zhao, Lingjun Zhao, Boli Xiong, and Gangyao Kuang, *Senior Member, IEEE*

***Abstract*—With the development of deep learning (DL) and synthetic aperture radar (SAR) imaging techniques, SAR automatic target recognition has come to a breakthrough. Numerous algorithms have been proposed and competitive results have been achieved in detecting different targets. However, due to the influence of various sizes and complex background of ships, detecting multiscale ships in SAR images is still challenging. To solve the problems, a novel network, called attention receptive pyramid network (ARPN), is proposed in this article. ARPN is a two-stage detector and designed to improve the performance of detecting multiscale ships in SAR images by enhancing the relationships among nonlocal features and refining information at different feature maps. Specifically, receptive fields block (RFB) and convolutional block attention module (CBAM) are employed and combined reasonably in attention receptive block to build a top-down fine-grained feature pyramid. RFB, composed of several branches of convolutional layers with specifically asymmetric kernel sizes and various dilation rates, is used for grabbing features of ships with large aspect ratios and enhancing local features with their global dependences. CBAM, which consists of channel and spatial attention mechanisms, is utilized to boost significant information and suppress interference caused by surroundings. To evaluate the effectiveness of ARPN, experiments are conducted on SAR Ship Detection Dataset and two large-scene SAR images. The detection results illustrate that competitive performance has been achieved by our method in comparison with several CNN-based algorithms, e.g., Faster-RCNN, RetinaNet, feature pyramid network, YOLOv3, Dense Attention Pyramid Network, Depth-wise Separable Convolutional Neural Network, High-Resolution Ship Detection Network, and Squeeze and Excitation Rank Faster-RCNN.**

***Index Terms*—Attention receptive pyramid network, convolutional block attention module (CBAM), receptive fields block (RFB), synthetic aperture radar (SAR), SAR automatic target recognition (SAR ATR), ship detection.**

I. Introduction

SYNTHETIC aperture radar (SAR) is an active microwave sensor, which could acquire high-resolution data in all weather conditions. Therefore, it has been widely used in military and civil fields such as marine surveillance [1], [2], earth observation [3] and so on. As one of the important applications, synthetic aperture radar automatic target recognition (SAR ATR) aims to figure out locations and class labels of potential targets and has been researched for a long time [4]–[9]. In this field, an important branch is ship detection in SAR images. Although numerous methods have been proposed [10], [11], it is still an enduring hot topic because of several tough problems for detecting multiscale ships in complex surroundings.

In traditional ways, several handcrafted features were explored [12]–[14]. Gray-scale feature, as one of the representative features, has been widely used in this area. Constant false alarm rate (CFAR) based methods, which consider the gray-scale feature, play a remarkable role in ship detection in SAR images. In CFAR-based methods, specific probability distribution models [15], [16], e.g., K and G^0 distributions, are established properly according to the SAR images. Then, binary maps are calculated by CFAR detectors with constant probability of false alarms. Next, specific binary areas are isolated out from these binary maps and discriminated by several well-designed features, such as geometrical features, electromagnetic scattering features, and so on. In this area, several methods were proposed. Leng *et al.* [17] introduced a bilateral CFAR algorithm for ship detection and reduced the influence of SAR ambiguities and sea clutter. Kang *et al.* [18] adopted a CFAR algorithm to detect targets with a threshold determined by the pixels' amplitudes in proposal and background windows. Although these methods detected ships with competitive performance, their generalization capabilities might be weak due to the difficulties in designing proper handcrafted features for various conditions. Besides, these methods might be complex due to numerous parameters predefined for specific conditions.

Recently, convolutional neural network (CNN) based methods have achieved remarkable successes in computer vision tasks, i.e., image classification [19], object detection [20]–[22], and image segmentation [23], [24]. In terms of object detection, methods could be divided into two paradigms, i.e., two-stage detectors and one-stage detectors. In two-stage detectors, e.g., Faster-RCNN [25] and feature pyramid network (FPN) [21], basic features are first extracted by backbone networks, e.g., VGG [26], ResNet [27], Inception [28], and DenseNet [29]. Then, Region of Interest (ROI) proposals are generated by region proposal network (RPN) [25] according to the predefined anchors in the first stage. Then, features of these proposals are resized to fixed sizes and processed by two branches of classification and regression networks. Finally, the detection results

Manuscript received March 20, 2020; revised April 30, 2020; accepted May 20, 2020. Date of publication May 28, 2020; date of current version June 15, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61701508 and Grant 61971426. (*Corresponding author: Lingjun Zhao*.)

The authors are with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, Changsha 410073, China (e-mail: zy34731@qq.com; zhaolingjunkd@126.com; bolixiong@gmail.com; kuangmerg@hotmail.com).

Digital Object Identifier 10.1109/JSTARS.2020.2997081

are acquired by a non-maximum suppression (NMS) operation. NMS is adopted to merge redundant predictions with a fixed threshold. The two-stage detectors can acquire high detection accuracy because of their well-designed modules for feature representation. However, they suffer from low processing speed due to numerous ROI proposals and complex processing schemes. One-stage detectors, e.g., single shot detector (SSD) [30], You Only Look Once (YOLO) [31], [32], and RetinaNet [22], are another paradigm. In this paradigm, networks directly predict locations and class labels of potential objects at several feature maps without cropping and resizing ROI proposals. Thus, the processing schemes are simpler and they can detect objects faster than two-stage detectors. However, one-stage detectors might sacrifice processing accuracy due to their plain strategies for feature extraction and discrimination.

With the surge of high-resolution SAR images, CNN-based methods have been applied to ship detection in SAR images. Kang *et al.* [33] constructed a hybrid detector by combining Faster-RCNN with a CFAR detector. The hybrid detector aimed at alleviating undesired differences caused by multiscale ships and the CFAR detector was adopted for refining object proposals generated by Faster RCNN. To alleviate interference of surroundings, Wang *et al.* [34] adopted SSD with transfer learning to detect ships in SAR images. Chang *et al.* [35] introduced a YOLOv2-reduced network by removing some convolutional layers from original YOLOv2 [36] to reduce computing complexity. The detection results illustrated that the YOLOv2-reduced network could detect ships in SAR images fast without sacrificing much detection accuracy. Besides, Lin *et al.* [37] proposed Squeeze and Excitation Rank (SER) Faster-RCNN for ship detection in SAR images. The method concatenated three levels of feature maps from VGG network to improve representative abilities of network. Besides, an SER mechanism was used after ROI pooling layer to refine significant information. Besides, An *et al.* [38] introduced a one-stage detector named DrBoxv2, to detect ships with various azimuth angles. The detector adopted several anchor sampling strategies with Online Hard Example Mining [39] to detect multiscale ships. Zhang *et al.* [40] proposed a Depth-wise Separable Convolutional Neural Network (DS-CNN) for ship detection in SAR images. In the method, a backbone network, composed of several layers of depth separable convolution [41], was established to extract basic features. The DS-CNN shared similar structure and loss functions with YOLO but further improved the detection speed without sacrificing much performance. Besides, Cui *et al.* [42] proposed a two-stage detector called Dense Attention Pyramid Network (DAPN), for multiscale ship detection in SAR images. Based on the structure of FPN, DAPN densely connected different levels of feature maps and adopted a convolutional block attention module (CBAM) [43] at top-down pathway of the lateral connections to filter out negative objects and suppress interference of surroundings. Wei *et al.* [44] introduced a precise and robust ship detection method called High-Resolution Ship Detection Network (HR-SDNet). In HR-SDNet, a modified High-Resolution Net [45] was established to retain the high-resolution features of ships. Besides, three cascade RCNN networks [46] were adopted for further discrimination.

Network structures, training strategies, and anchor sampling mechanisms carefully designed, the performance of ship detection in SAR images has been obviously improved. However, there still exist some tough problems when detecting multiscale ships. The surroundings of inshore ships are usually more complex than those of offshore ships. Sometimes, backscattering points of interference are even stronger than those of ships and wakes might be similar to ships. It might cause much disturbance for algorithms to locate and discriminate multiscale ships accurately. Therefore, exploring relationships between local features and their global dependences as well as suppressing useless and confusing information is essential for improving the performance of detecting ships in SAR images.

In this article, a novel method named attention receptive pyramid network (ARPN) is proposed. It represents the two key components, i.e., receptive fields block (RFB) [47] and CBAM), and the basic architecture, i.e., a well-designed feature pyramid, of our method. To improve the performance of detecting multiscale ships in complex SAR scenes, RFB, which adopts multibranch convolutions with various asymmetric kernel sizes and dilation rates, is utilized to grab characteristics of multiscale ships with different directions and enhance local features with their global dependences. CBAM, which consists of channel and spatial attention mechanisms, is used to boost the significant features of ships and suppress interference of surroundings, e.g., waves, isles, wakes of ship, and so on, by reweighting feature values intelligently on hierarchical feature maps. By combining RFB and CBAM reasonably, a well-designed lateral connection named attention receptive block (ARB) is introduced in ARPN. When detecting ships, basic feature maps are first established in the bottom-up pathway by using a backbone network, e.g., ResNet-101 in our method. Then, a fine-grained feature pyramid is constructed by the ARB. Next, RPNs are adopted at different levels of feature pyramid to generate ROI proposals. Finally, the final detection results are acquired by using two branches of subnetworks, i.e., classification and regression subnetworks followed by an NMS operation. To evaluate the performance of our method, we conduct several experiments on the SAR Ship Detection Dataset (SSDD) [48] and some large-scene SAR images. The detection results illustrate that our method is efficient for detecting multiscale ships in SAR images with complex scenes, compared with several CNN-based methods, e.g., Faster RCNN, RetinaNet, FPN, YOLOv3, DAPN, DS-CNN, HR-SDNet, and SER Faster RCNN. The major contributions of this article are indicated as follows.

1) A novel detection method called ARPN is proposed for detecting multiscale ships in SAR images.
2) We introduced a well-designed lateral connection called ARB. And two specific modules, i.e., RFB and CBAM, are combined reasonably into ARB to grab features of multiscale ships and refine redundant information when establishing a fine-grained feature pyramid.
3) Experiments on SSDD dataset and large-scene SAR images demonstrate that the proposed method detects multiscale ships with competitive results in comparison with some classical and special CNN-based methods.

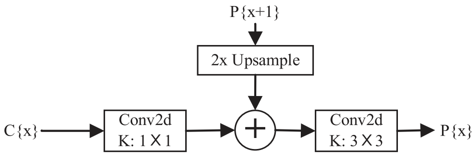


Fig. 1. Lateral connection in FPN.

II. Motivation and Method

In this section, motivations and our method are described in detail sequentially.

A. Idea of the Proposed Method

To detect multiscale objects, a feature pyramid is adopted in FPN. In the bottom-up pathway, basic features are extracted by a backbone network. However, these features at different feature maps might be imbalanced. Furthermore, high-level feature maps contain rich semantic information but are lack of accurate position information. Low-level feature maps contain rich position information but poor semantic information. Therefore, a top-down pathway is introduced by using lateral connections to acquire a fine-grained features pyramid. Fig. 1 is a lateral connection used in FPN. $C\{x\}$ means a basic feature map from the bottom-up pathway. $P\{x+1\}$ and $P\{x\}$ are different levels of fine-grained feature maps from the top-down pathway. To acquire $P\{x\}$, a vanilla convolution with a kernel size of 1×1 and a two times up-sampling operation are conducted on $C\{x\}$ and $P\{x+1\}$, respectively. After an element-wise addition and a convolution with a kernel size of 3×3, the fine-grained feature map $P\{x\}$ is produced.

Although FPN considers multiscale features by establishing a fine-grained feature pyramid, features on identical levels might be naive and local due to the simple convolution with a kernel size of 1×1. The relationships between local and global features at horizontal pathways are still weak. Considering various geometrical shapes of multiscale ships and complex surroundings in SAR images, enhancing relationships of local features with their global dependences and highlighting significant information are essential when establishing the fine-grained feature pyramid. To alleviate the problems, a novel lateral connection called ARB is introduced in our method. It combines RFB and CBAM reasonably. Furthermore, RFB is used to grab features of multiscale ships and enhance local features with their global dependences. It consists of several branches of convolutional layers with asymmetrical kernel sizes and various dilation rates. Dilation rates carefully designed, the receptive fields of RFB are enlarged and appropriate for multiscale ships. However, features from RFB might be redundant and significant features are influenced by useless information. Hence, a powerful attention mechanism called CBAM is employed to highlight significant information. It consists of channel and spatial attention mechanisms to refine features intelligently. By combining the two modules into ARB, not only characteristics of multiscale ships at different feature levels could be acquired, but also significant information is enhanced effectively. In the next section, the whole processing scheme of our method is described in detail.

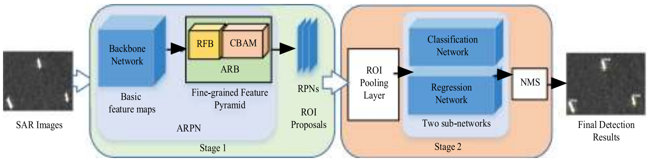


Fig. 2. Overview of the processing scheme.

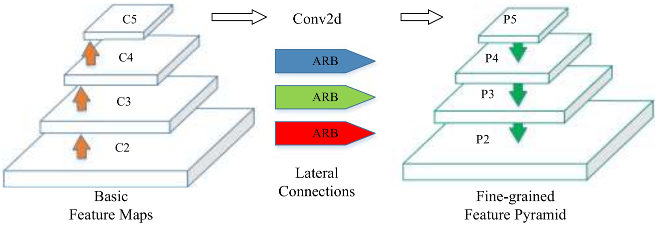


Fig. 3. Attention receptive pyramid network (ARPN).

B. Overview of the Processing Scheme

The processing scheme of our method is shown in Fig. 2. It belongs to two-stage detectors. In the first stage, SAR images are fed into ARPN. It consists of two parts including a backbone network and ARBs. After acquiring the basic feature maps by the backbone network, a fine-grained feature pyramid is established by ARB. RFB and CBAM are used in ARB in sequence when fusing different levels of feature maps iteratively. Next, RPNs are applied on several levels of fine-grained feature maps to acquire two vectors with sizes of $2 \cdot H \cdot W \cdot K$ and $4 \cdot H \cdot W \cdot K$. They encode class labels (background or foreground) and positions (t_x, t_y, t_w, t_h) of ROI proposals, respectively. Here H, W refer to the height and width of feature maps, respectively. K refers to the number of ROIs at each location of feature maps. t_x, t_y encode the normalized central locations of bounding boxes. t_w, t_h encode the normalized widths and heights of bounding boxes, respectively. In the second stage, an ROI warping layer [49] is used to crop and resize ROIs into a fixed size from feature maps. Then, these ROIs are sent to classification and regression subnetworks for discrimination and localization tasks, respectively. The final detection results are acquired by merging and discarding redundant potential objects by using a NMS. Different from FPN, a well-designed lateral connection called ARB is introduced when building a fine-grained feature pyramid. It improves the relationships of different ranges of features and enhance the discrimination abilities of our method for multiscale ships in SAR images.

C. Architecture of ARPN

The structure of ARPN is shown in Fig. 3. Feature maps $C_2, C_3, C_4,$ and C_5 are first extracted by a backbone network in the bottom-up pathway. Then, the fine-grained feature maps $P_2, P_3, P_4,$ and P_5 are built by several lateral connections. In our method, ARB is used at three hierarchical feature maps

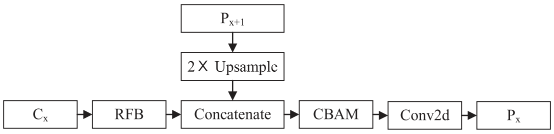


Fig. 4. Attention receptive block (ARB).

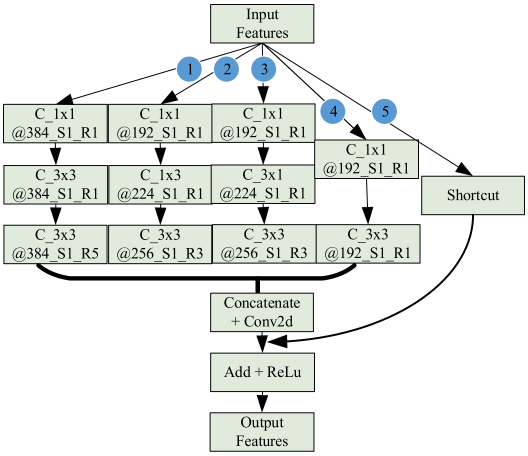


Fig. 5. Structure of receptive field block (RFB).

$C_2, C_3,$ and C_4 and the outputs are $P_2, P_3,$ and P_4, respectively. Considering that C_5 has the smallest sizes and the highest semantic features among other basic feature maps, P_5 is directly constructed by a simple two-dimensional (2-D) convolution. Fig. 4 shows the structure of ARB in detail. C_x refers to a basic feature map. P_{x+1} is two times up-sampled and concatenated with the output of RFB along channel dimension. The fine-grained feature map P_x is acquired after CBAM followed by a simple 2-D convolution.

The following equations formulate ARB:

$$P_5 = \text{Conv}(C_5) \tag{1}$$

$$C_{x\_\text{mid}} = \text{Concat}\{\text{RFB}(C_x), \text{Upsample}(P_{x+1})\} \tag{2}$$

$$P_x = \text{Conv}\{\text{CBAM}(C_{x\_\text{mid}})\}, x = 1, 2, 3\}. \tag{3}$$

D. Receptive Fields Block

The structure and parameters of RFB are shown in Fig. 5 in detail. RFB consists of four branches of convolution and a shortcut connection. We define $C\_AxA@B\_SC\_RD$ to demonstrate convolution of RFB clearly. $A, B, C,$ and D refer to the kernel sizes, the numbers of convolutional filters, the strides, and the dilation factors, respectively. Branch 1 includes three convolutional layers. The last convolution with dilation rate 5 is designed for capturing global features. According to the statistics of bounding boxes' aspect ratios in SSDD, the ratio of length/width is mostly around 0.3. Considering the different directions of ships, we carefully set the convolutional kernel sizes to1×3 and 3×1 in branch 2 and branch 3. The asymmetric convolutions with large dilation rates are suitable for multiscale ships and could grab nonlocal features of them such as edges, profiles. In branch 4, two convolutional layers with kernel sizes of 1×1 and 3×3 are used to fetch local features of ships. After a concatenation followed by a 2-D convolution, the local features with their global dependences of multiscale ships are boosted at identical feature maps. Moreover, a shortcut connection is established in branch 5. This design, which is the same as the residual blocks of ResNet, might retain the original information of input features and avoid gradient vanishment in the training phase. The final output features are acquired after a Rectified Linear Units function used for improving nonlinearity of RFB.

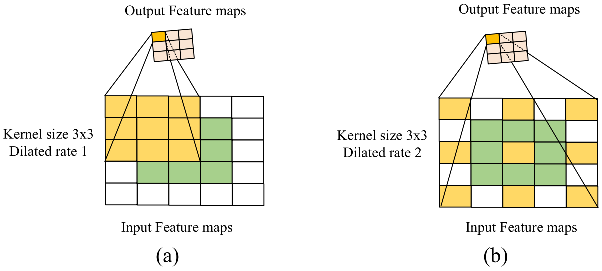


Fig. 6. (a) Classical convolution. (b) Convolution with a dilation rate 2.

Compared with stem modules of inception networks [28], [50], the receptive fields of RFB might be larger with small capacities of network parameters by using convolutions with large dilation rates. Fig. 6(a) and (b) illustrate a classical convolution and a convolution with a dilation rate 2, respectively.

In Fig. 6(a) and (b), yellow rectangles are the real receptive fields. Green rectangles are the original convolutional kernels. Pink rectangles are the output feature maps. In Fig. 6(a), the receptive field is small because of limitation of traditional convolutional kernel, i.e., the dilation rate is 1. However, the receptive field is enlarged by setting the dilation rate to 2 with the kernel size unchanged (the number of yellow rectangles is constant). Thus, convolution kernels with large dilation rates could grab larger ranges of features without increasing much capacities of parameters than the classical convolution. Equations (4) and (5) formulate the links between receptive fields and dilation rates

$$R = K + (K - 1) \cdot (r - 1) \tag{4}$$

$$Y = (X - R + 2P)/S + 1. \tag{5}$$

In (4), R, K, and r refer to sizes of real receptive fields, kernel sizes, and dilation rates, respectively. In (5), Y, X, P, and S refer to sizes of output and input feature maps, paddings, and strides, respectively. Although features from RFB contain rich semantic information, interference caused by complex surroundings, e.g., docks, buildings, isles, wakes of ships might also exist. Thus, it is necessary to refine significant features for discriminating ships clearly. In our method, an attention module called CBAM, is utilized and introduced in the next section.

E. Convolutional Block Attention Module

To refine significant features and improve discrimination of network for multiscale ships, CBAM is employed after RFB. Compared with squeeze-and-excitation (SE) [51], both channel

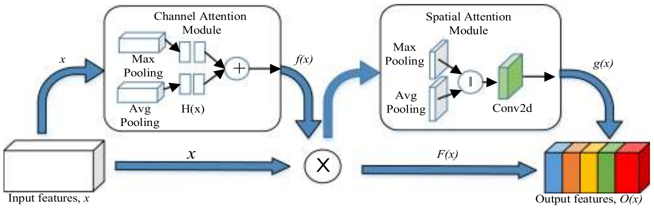


Fig. 7. Convolutional block attention module (CBAM).

and spatial attention mechanisms are introduced in CBAM. Fig. 7 shows its architecture in detail. The input features are multiplied with the outputs of channel attention module and spatial attention module sequentially. In the channel attention module, maximum pooling (Max-Pooling) and average pooling (Avg-Pooling) layers are adopted parallel along the width and the height dimensions of feature maps. Then, a multiple layer perceptron (MLP) $H(x)$ is used to output the weights along the channel dimension of feature maps. Same as the channel attention module, maximum pooling (Max-Pooling) and average pooling (Avg-Pooling) layers are also used in spatial attention module but along channel dimension of the input features. The reweighted vectors $g(x)$ are acquired after a 2-D convolution. The height and width of $g(x)$ are the same as the input features x. Finally, output features $O(x)$ are acquired by multiplying the input features x with $f(x)$ and $g(x)$ in sequence.

CABM is formulated as follows:

$$\mathrm{Att}_{ch} = f(H(\mathrm{Maxpool}(x)) + H(\mathrm{Avgpool}(x))) \tag{6}$$

$$\begin{aligned}\mathrm{Att}_{\mathrm{sp}} = g(&\mathrm{Conv2}d(\mathrm{Maxpool}(\mathrm{Att}_{\mathrm{ch}} \cdot x))\\ &+ \mathrm{Avgpool}(\mathrm{Att}_{\mathrm{ch}} \cdot x))\end{aligned} \tag{7}$$

$$O(x) = x \cdot \mathrm{Att}_{\mathrm{ch}} \cdot \mathrm{Att}_{\mathrm{sp}}. \tag{8}$$

Here f and g refer to sigmoid function. $H(x)$ means multilayer perceptron. $\mathrm{Att}_{\mathrm{ch}}$ and $\mathrm{Att}_{\mathrm{sp}}$ refer to the output vectors from the channel and spatial attention modules, respectively. $O(x)$ refers to the final reweighted features.

F. Loss Functions

Same as other classical two-stage object detection networks, RPN and the final prediction networks are all optimized by using multitask loss, which is given by (9). N_{cls} and N_{reg} refer to numbers of a minibatch samples in the training phase. And $L_{\mathrm{cls}}(p_i, {p_i}^*)$ and $L_{\mathrm{reg}}(t_i, {t_i}^*)$ refer to classification and regression losses, respectively. When training the classification network, Cross Entropy loss (CE) is utilized, which is given by (10). p and y refer to the probabilities of predicted ROIs and the corresponding ground truth labels, respectively. When training the regression network, Smooth L1 loss is used, which is given by (11). x refers to the positional offsets between the normalized ground truth bounding boxes and the predictions of the regression network

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, {p_i}^*) + \lambda \frac{1}{N_{neg}} {p_i}^* L_{reg}(t_i, {t_i}^*) \tag{9}$$

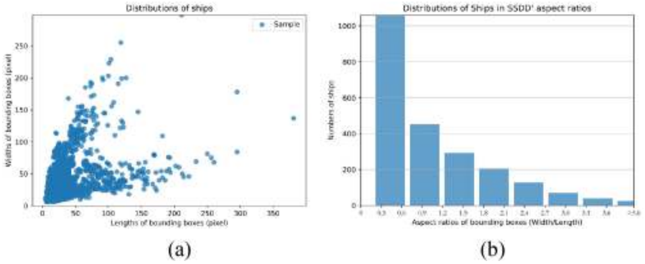


Fig. 8. Distributions of the sizes and aspect ratios of the bounding boxes in SSDD. (a) Distributions of the bounding boxes' sizes in SSDD. (b) Distributions of the bounding boxes' aspect ratios in SSDD.

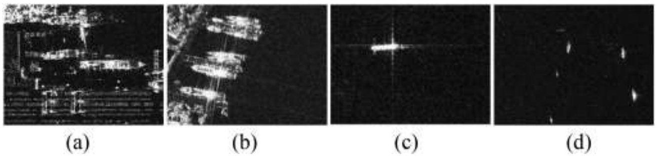


Fig. 9. Samples of inshore and offshore ships in SSDD. (a) and (b) show inshore ships where much interference exists caused by buildings and docks. (c) and (d) show multiscale offshore ships.

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p) & y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases} \tag{10}$$

$$\mathrm{Smooth}_{L1} = \begin{cases} 0.5x^2 & x < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}. \tag{11}$$

III. Dataset and Parameter Settings

In this section, we describe the SSDD dataset, the large-scene SAR images and parameter settings in detail.

A. Dataset Description

To evaluate the performance of our method, an open source dataset, SSDD [48] and two large-scene SAR images are utilized.

SSDD proposed by Li *et al.*, is established by using images from RadarSat-2, TerraSAR-X, and Sentinel-1 satellites. It contains 1160 images and 2456 multiscale ships. The average number of ships per image are 2.12. And resolutions of these images range from 1 to 15 m. The distributions of bounding boxes' sizes and aspect ratios are shown in Fig. 8(a) and (b), respectively.

In Fig. 8(a), heights and widths of bounding boxes range from about 10 to 380 pixels and 10 to 270 pixels, respectively. Aspect ratios of bounding boxes range from about 0.3 to 3.6 and most ships' aspect ratios are about from 0.3 to 0.6 according to Fig. 8(b). Some samples of offshore and inshore ships are shown in Fig. 9.

In our experiment, we randomly divide the SSDD into three parts, i.e., a training set, a validation set and a testing set, with the proportion of 7:1:2. Input images are resized so that the short edges of them are 350 pixels. Data augmentation strategies, e.g., random cropping, flipping, contrast transformation and mirroring, are utilized. Finally, 12 984 training images are collected

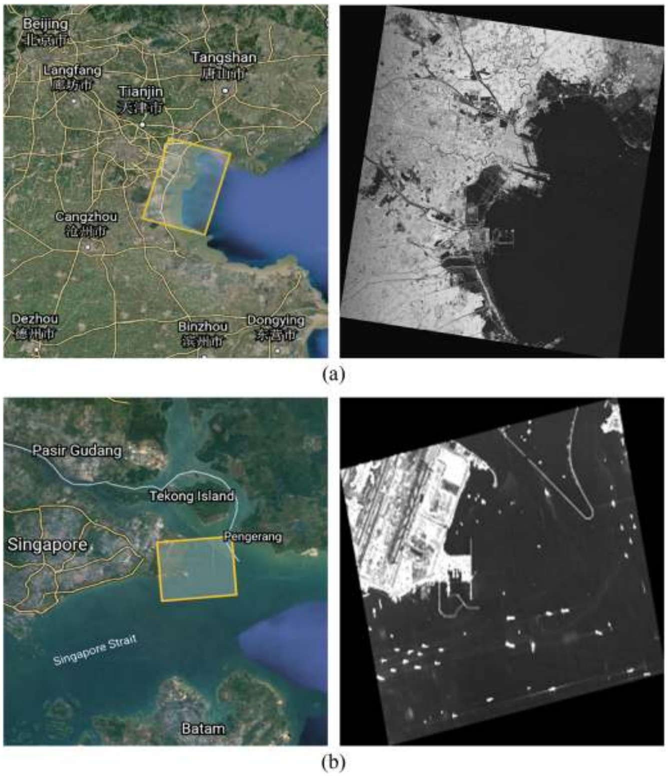


Fig. 10. Locations and thumbnails of two large-scene SAR images for ship detection. Two groups of large-scene SAR images, i.e., (a) and (b). The left images of (a) and (b) are the geographical locations of large scenes SAR images. And the right images of (a) and (b) show the corresponding thumbnails.

TABLE I
INFORMATION OF THE TWO LARGE-SCENE SAR IMAGES

Group Name	(a)	(b)
Product ID	1774400	3375753
Mode	FSII	SL
Polarization	VH	HH
Resolution (m)	10	1
Sizes (Pixels)	12741×13809	30660×26952
Longitude (Degree)	117.656484E	103.98218E
Latitude (Degree)	38.920512N	1.328857N

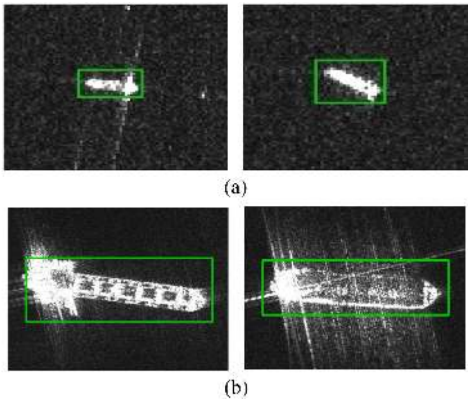


Fig. 11. Ground truth of ships from large SAR images. (a) Ships from the first large-scene SAR image shown in Fig. 10(a). (b) Ships from the second large-scene SAR image shown in Fig. 10(b).

for training our method. Besides, all images are normalized by subtracting the mean values along channel dimension of them to balance the ranges of gray-scale values for stable training.

Furthermore, to further verify the performance of our method for detecting multiscale ships in large SAR images with complex scenes, two large SAR images collected from Chinese GF-3 satellite are also used. The geographical locations and thumbnails of them are shown in the left and the right columns of Fig. 10, respectively. And their imaging information is listed in Table I.

According to Table I, there are some differences between the two images due to their different imaging modes, polarizations, and resolutions. The two images are annotated by experts assisted with LabelImg [52] and GoogleEarth [53] applications. Each ship is marked with two coordinates, i.e., $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$, to define its location. Here $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ refer to the top-left and the bottom-right coordinate values of ships, respectively. All annotation files obey the formats of VOC 2007 dataset [54]. However, due to the complexities of environment around ships and SAR imaging mechanisms, some ships may only appear a few pixels because of small sizes and low image resolutions. It might be hard to discriminate whether they are ships or not. Hence, if lengths and widths of ships are larger than about 10 pixels in SAR images, we would annotate them as ships. To demonstrate our annotations clearly, several ground truth ships from the two large SAR images are shown in Fig. 11(a) and (b). Ships in Fig. 11(a) are a little blurry and from the first large-scene SAR image (product ID 1774400). Ships in Fig. 11(b) are distinct and from the second large-scene SAR image (product ID 3375753).

B. Parameter Settings

In all experiments, ResNet-101 pretrained on ImageNet [19] is adopted as the backbone network in our method. Dilation rates of the first four branches of convolution in RFB are set to 5, 3, 3, 1, respectively. In CBAM, the rates of both MLP in the channel attention module and convolutional layers in the spatial attention module are set to 16. For RPN, sizes and aspect ratios of basic anchors are set to 32, 64, 128, 256, and 512 and 0.4, 0.8, 1.2, 1.6, 2.0, 2.4, 2.8, 3.2, and 3.6, respectively, according to SSDD. Besides, the ratio of foreground and background anchors in a minibatch is set to 1:1 in the training phase. An anchor is assigned to a positive sample if the Intersection of Union between any ground truth bounding box and the anchor is higher than 0.7. Besides, the threshold of NMS, the weight decay rate are set to 0.3, 0.0001, respectively. We train all networks for 50 000 iterations and save checkpoints every 5000 steps. The initial learning rate is 0.001 and decayed 10 times when the steps are at 35 000 and 40 000. All experiments are implemented using TensorFlow framework on Ubuntu with a Nvidia GTX 1080Ti graphics card support.

IV. Experimental Results and Evaluation

The performance of our method is evaluated in this section. First, some evaluation metrics are described. Then, we divide the testing set into two groups, i.e., one with offshore ships and the other with inshore ships, and use them to judge the performance of different methods, respectively. We first exploit the contributions of RFB and CBAM adopted in ARPN. Then, we compare our method with other CNN-based methods by using offshore and inshore ships, respectively. Apart from this, the detection results of our method and some specific CNN-based methods on the two large-scene SAR images are shown.

A. Evaluation Criterions

Since object detection tasks are similar for optical and SAR images, several mature indicators, e.g. Recall rate, Precision rate, F score (F1), and Average Precision (AP), are employed to evaluate the performance of different methods. The following equations formulate these indicators in detail:

$$\text{Precision rate (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{Recall rate (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$\text{F1} = \frac{2PR}{P + R} \tag{14}$$

$$AP = \int_0^0 P(R)dR. \tag{15}$$

Here TP, FP, and FN refer to numbers of True Positives, False Positives, and False Negatives, respectively. Precision rate refers to the proportion of ground truth ships predicted by networks in all predictions. Recall rate refers to the proportion of ground truth ships predicted by networks in all ground truth ships. F1 is a comprehensive indicator used for judging the performance of different networks by combining precision rate with recall rate together. AP describes the area under Precision-Recall (PR) curves and it also illustrates comprehensive performance of different methods. Besides, Frames-Per-Second (FPS) formulated in (16), is used to judge the detection speed of different methods. The higher the FPS is the higher speed a method achieves

$$\text{FPS} = \frac{1}{T_{\text{per-img}}} \tag{16}$$

where $T_{\text{per-img}}$is the inference time cost of a method when processing an image.

B. Contributions of Different Modules

In this section, contributions of RFB and CBAM are exploited on offshore and inshore ships, respectively. Since the proposed method and FPN have the similar structures, FPN is used as a baseline when evaluating RFB and CBAM.

TABLE II
Detection Results of Methods on Offshore Ships

Method	Recall	Precision	F1	AP
FPN	0.964	0.948	0.956	0.978
ARPN - RFB	0.969	0.971	0.970	0.989
ARPN	0.964	0.964	0.982	0.982

TABLE III
Detection Results of Methods on Inshore Ships

Method	Recall	Precision	F1	AP
FPN	0.816	0.627	0.709	0.811
ARPN - RFB	0.767	0.731	0.749	0.816
ARPN	0.854	0.733	0.789	0.841

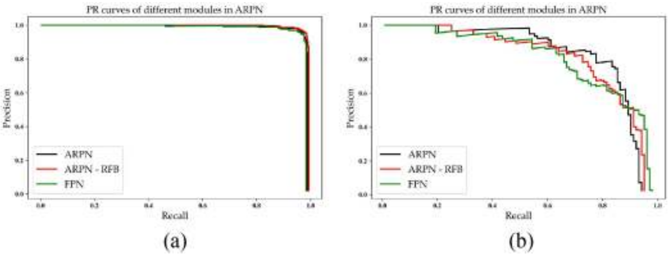


Fig. 12. PR curves of different methods tested on offshore and inshore ships. (a) PR curves of ARPN, ARPN - RFB and FPN on offshore ships. (b) PR curves of ARPN, ARPN - RFB and FPN tested on inshore ships.

1) Contributions of RFB: Tables II and III show the detection results of FPN, ARPN without RFB (ARPN - RFB) and ARPN on offshore and inshore ships, respectively. In Table II, scores of all indicators of ARPN - RFB and ARPN are higher than those of FPN. And ARPN - RFB performs a little better than ARPN. Furthermore, recall rate, precision rate, F1 and AP of ARPN – RFB are 0.5%, 0.7%, 0.6%, and 0.7% higher than those of ARPN, respectively. Because of rich and redundant information acquired by RFB, redundant features exist in the outputs of RFB. These features might cause heavy feature refining load for CBAM. Thus, some false alarms may not be effectively suppressed. Compared with offshore ships, complex surroundings, such as docks, buildings, and vehicles, may cause much interference for detecting inshore ships accurately. Thus, all statistical indicators decrease sharply. For example, recall rate, precision rate, F1 and AP of ARPN tested on inshore ships are 11.0%, 23.1%, 17.5%, and 14.1% lower than those of ARPN tested on offshore ships, respectively. In Table III, recall rate, precision rate, F1 score, and AP of ARPN - RFB are 8.7%, 0.2%, 4.0%, and 2.5% lower than those of ARPN. Especially, a sharp decrease of recall rate appears by using ARPN - RFB. It might be because of the inappropriate feature representation for multiscale ships and the weak relationships of non-local features by using convolution with commonplace kernel sizes, e.g., 3 × 3 and without large dilation rates. Besides, most indicators of ARPN and ARPN - RFB are higher than those of FPN and recall rate and precision rate of FPN is imbalanced.

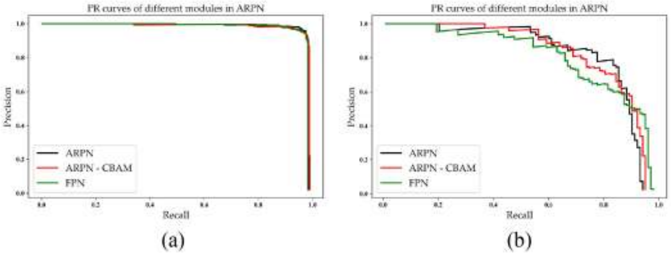
Fig. 12(a) and (b) show the PR curves of FPN, ARPN - RFB, and ARPN tested on offshore and inshore ships, respectively. In Fig. 12(a), there are small differences among ARPN, ARPN - RFB, and FPN. Furthermore, the PR curve of ARPN - RFB [the red curve in Fig. 12(a)] is a little higher than those of ARPN and FPN (the black and the green curves in Fig. 12(a)). In terms of detecting inshore ships, it's distinct that the PR curves of ARPN and ARPN - RFB [the black curve and the red curves in

TABLE IV
DETECTION RESULTS OF METHODS ON OFFSHORE SHIPS

Method	Recall	Precision	F1	AP
FPN	0.964	0.948	0.956	0.978
ARPN - CBAM	0.964	0.952	0.958	0.981
ARPN	0.964	0.964	0.964	0.982

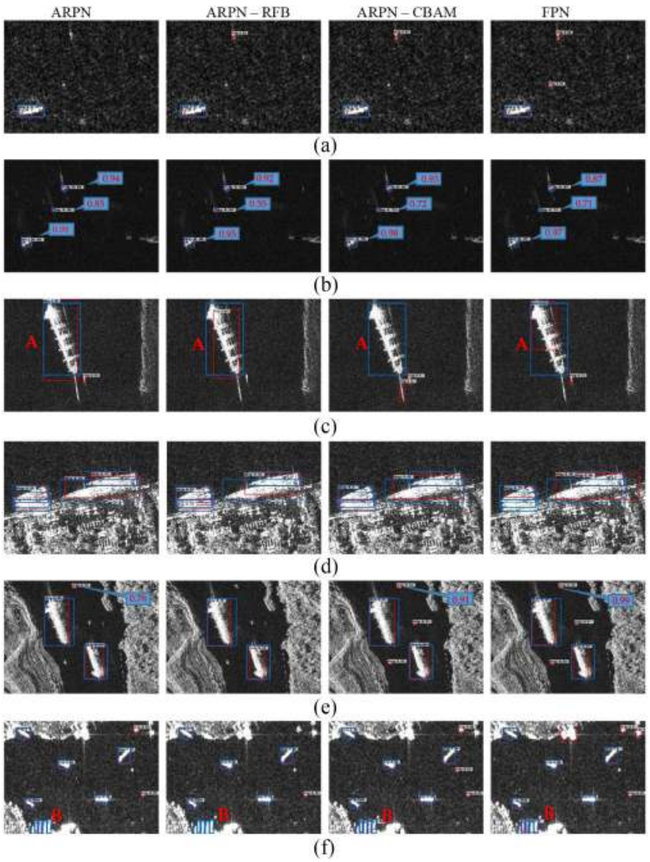
TABLE V
DETECTION RESULTS OF METHODS ON INSHORE SHIPS

Method	Recall	Precision	F1	AP
FPN	0.816	0.627	0.709	0.811
ARPN - CBAM	0.854	0.633	0.727	0.840
ARPN	0.854	0.733	0.789	0.841



(a) (b)

Fig. 13. Precision-Recall (PR) curves of different methods tested on offshore and inshore ships. (a) PR curves of ARPN, ARPN - CBAM, and FPN tested on offshore ships. (b) PR curves of ARPN, ARPN - CBAM, and FPN tested on inshore ships.



(a) (b) (c) (d) (e) (f)

Fig. 14. Detection Results of ARPN, ARPN - RFB, ARPN - CBAM, and FPN tested on offshore and inshore ships, respectively. Rectangles with blue color are ground truth ships. Rectangles with red color are predictions. Group (a), (b), and (c) are the detection results on offshore ships; Group (d), (e), and (f) are detection results on inshore ships. Images at different columns are the detection results of different algorithms.

Fig. 12(b)] decrease slowly than that of FPN [the green curve in Fig. 12(b)]. Besides, the PR curve of ARPN - RFB comes to a sharp decrease with an increase of recall rate compared with ARPN. It might be because of the insufficient characteristics extracted by ARPN – RFB, which leads to weak discrimination for ships.

2) Contributions of CBAM: The contributions of CBAM are exploited in this section. Tables IV and V show the statistical results of FPN, ARPN without CBAM (ARPN – CBAM), and ARPN tested on offshore and inshore ships, respectively. In Table IV, small differences exist among the three algorithms in terms of recall rate, precision rate, F1 and AP. Moreover, precision rate, F1 and AP of ARPN are only 1.2%, 0.6%, and 0.1% higher than those of ARPN – CBAM and 1.6%, 0.8%, and 0.4% higher than those of FPN, respectively. It might be because of distinctive features of ships' bodies and clear surroundings of offshore ships in SAR images. Thus, it is easy for these algorithms to learn discriminative features of multiscale ships. In terms of detecting inshore ships, precision rate and F1 of ARPN are 10.0% and 6.2% superior to those of ARPN – CBAM according to Table V. It might be because that redundant features are effectively suppressed by CBAM along channel and spatial dimensions and the network could pay more attention to significant features of ships and discriminate them distinctly.

The PR curves of ARPN, ARPN - CBAM, and FPN tested on inshore and offshore ships are shown in Fig. 13(a) and (b), respectively. In Fig. 13(a), the PR curves of ARPN, ARPN - CBAM, and FPN are very close. It also illustrates that these algorithms perform well when detecting offshore ships. However, obvious differences appear among the PR curves of these algorithms when detecting inshore ships [see Fig. 13(b)]. Especially when the recall rate is about 0.8, the precision rate of ARPN is obviously higher than those of ARPN - CBAM and FPN.

3) Visual Detection Results of ARPN With Different Modules: Visual detection results of ARPN, ARPN - RFB, and ARPN - CBAM are shown in Fig. 14. We totally select six groups of detection results on offshore ships, i.e., groups (a), (b), and (c), and inshore ships, i.e., groups (d), (e), and (f). The detection results of ARPN, ARPN - RFB, ARPN – CBAM, and FPN are shown at each column of Fig. 14, respectively.

According to different groups and columns of Fig. 14, some conclusions are summarized as follows. First, RFB and CBAM enhance network discrimination for features of ships and surroundings. In Fig. 14(a), one or two false alarms exist in the detection results of ARPN – RFB, ARPN – CBAM, and FPN. However, no false alarm exists in the detection results of ARPN. In Fig. 14(b), the three ships are detected with higher probabilities by ARPN than those predicted by other three methods. It might be because that local features with their global dependences are boosted by RFB and significant features are refined by CBAM. Second, RFB mainly pays attention to extract

TABLE VI
DETECTION RESULTS OF CNN-BASED METHODS ON OFFSHORE SHIPS

Method	Recall	Precision	F1	AP	FPS
Faster-RCNN	0.886	0.918	0.902	0.885	16
YOLOv3	0.964	0.968	0.966	0.977	26
RetinaNet	0.945	0.973	0.958	0.934	18
FPN	0.964	0.948	0.956	0.978	12
SER Faster-RCNN	0.913	0.908	0.911	0.893	32
DS-CNN	0.862	0.856	0.859	0.857	77
DAPN	0.983	0.949	0.966	0.987	14
HR-SDNet	0.944	0.951	0.948	0.921	11
Ours	0.964	0.964	0.964	0.982	13

representative features and strengthen links of nonlocal features of multiscale ships. It might improve the accuracy of locating ships and recall rate. In Fig. 14(c), location of ship A predicted by ARPN seems more accurate than those predicted by FPN and ARPN - RFB. In Fig. 14(f), densely arranged ships in area B are detected by ARPN while missed by ARPN - RFB. Third, CBAM concentrates more on feature refinement. In Fig. 14(e) and (f), fewer false alarms exist in the detection results of ARPN than ARPN - CBAM. Besides, the probabilities of false alarms predicted by ARPN are lower than those of ARPN - CBAM and FPN. For example, probability of the false alarm predicted by ARPN is 0.78 while are 0.91 and 0.99 predicted by ARPN - CBAM and FPN, respectively.

In a short, RFB and CBAM are two important modules of our method. On one hand, RFB obtains representative features of multiscale ships and enhances the relationship between nonlocal features. On the other hand, redundant features at different levels of feature pyramid are refined by CBAM. The proposed method, ARPN, could detect multiscale ships in SAR images effectively by combining them reasonably.

C. In Comparison With Other CNN-Based Methods

In this section, offshore and inshore ships of testing sets are used to evaluate the proposed method and other CNN-based methods, e.g., Faster-RCNN, FPN, RetinaNet, YOLOv3, SER Faster-RCNN, DS-CNN, and HR-SDNet, respectively. Precision rate, recall rate, F1, AP, and FPS are also utilized to judge the performance of different methods.

1) Detection Results of Offshore Ships: Table VI shows the detection results of different CNN-based methods tested on offshore ships. The detection results of classical CNN-based methods are shown in the first four rows of Table VI. The fifth to the eighth rows of Table VI are the detection results of CNN-based methods specifically designed for object detection in SAR images. In Table VI, methods with feature pyramids, e.g., FPN, RetinaNet, HR-SDNet, DAPN, and ARPN perform better than those without or with vanilla feature fusion strategies, e.g., Faster-RCNN, SER Faster-RCNN. Although DS-CNN also concatenates multilevel features to acquire fine-grained semantic information, the backbone network constructed by depth-wise and point-wise convolutions, might be weak for extracting significant features. The performance of our method is close to those of FPN, RetinaNet, and YOLOv3 but better than that of Faster-RCNN by a large margin (the first four rows in Table VI). It might be because of the contributions of feature pyramids adopted in FPN, RetinaNet, and YOLOv3 and a weak ability of discriminating features by Faster-RCNN. Besides, YOLOv3 also gets competitive results in terms of F1 and AP because of its well-designed feature fusion and anchor generation strategies. Compared with the CNN-based methods designed for object detection in SAR images, our method performs better than SER Faster-RCNN, DS-CNN, and HR-SDNet. Furthermore, in terms of F1, score of our method is 10.5%, 5.3%, and 1.6% higher than those of DS-CNN, SER Faster RCNN, and HR-SDNet, respectively. AP of our method is 98.2%, which is 12.5%, 7.1%, and 6.1% higher than those of DS-CNN, SER Faster-RCNN, and HR-SDNet, respectively. However, F1 and AP of DAPN are 0.2% and 0.5% higher than those of our method, respectively. Besides, according to FPS shown in the last column of Table VI, DS-CNN runs faster than other methods possibly because of its separable depth-wise and point-wise convolutions used in the backbone network. Besides, because of similar structures, e.g., feature pyramids, and processing schemes among FPN, DAPN, HR-SDNet, and our method, the testing time for offshore ships is very close.

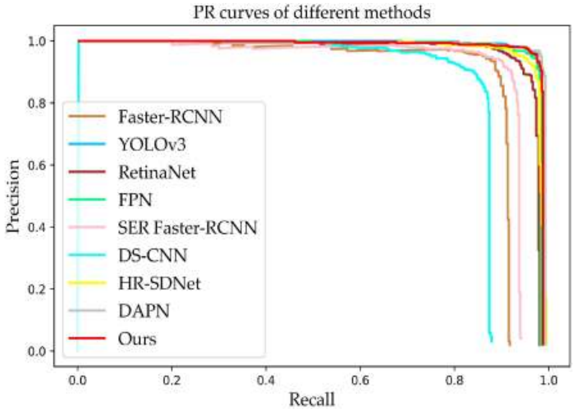


Fig. 15. PR curves of CNN-based methods tested on offshore ships.

Fig. 15 shows the PR curves of different CNN-based methods tested on offshore ships.

According to Fig. 15, the PR curves of these methods are very similar except Faster-RCNN, SER Faster-RCNN, and DS-CNN. It is distinct that the PR curve of DS-CNN is lower than those of other methods. It might be because of the specific convolution operation, which reduces the capacities of parameters but weakens the feature extraction abilities. Besides, the PR curves of Faster-RCNN, SER Faster-RCNN, and DS-CNN decrease sharply when the recall rate is higher than 0.8 possibly due to the lack of fine-grained feature pyramids.

Fig. 16 shows the detection results of different methods at four conditions of offshore ships. At the first condition (the first row to the third row of Fig. 16), ships are densely arranged. More than ten ships are missed by Faster-RCNN, SER Faster-RCNN, and RetinaNet. It might be because of the simple bottom-up pathway of Faster-RCNN, the plain feature fusion strategy of SER Faster-RCNN, and the inappropriate feature levels selected by RetinaNet, respectively. However, fewer ships are missed by FPN, YOLOv3, DAPN, DS-CNN, and our method. It might be

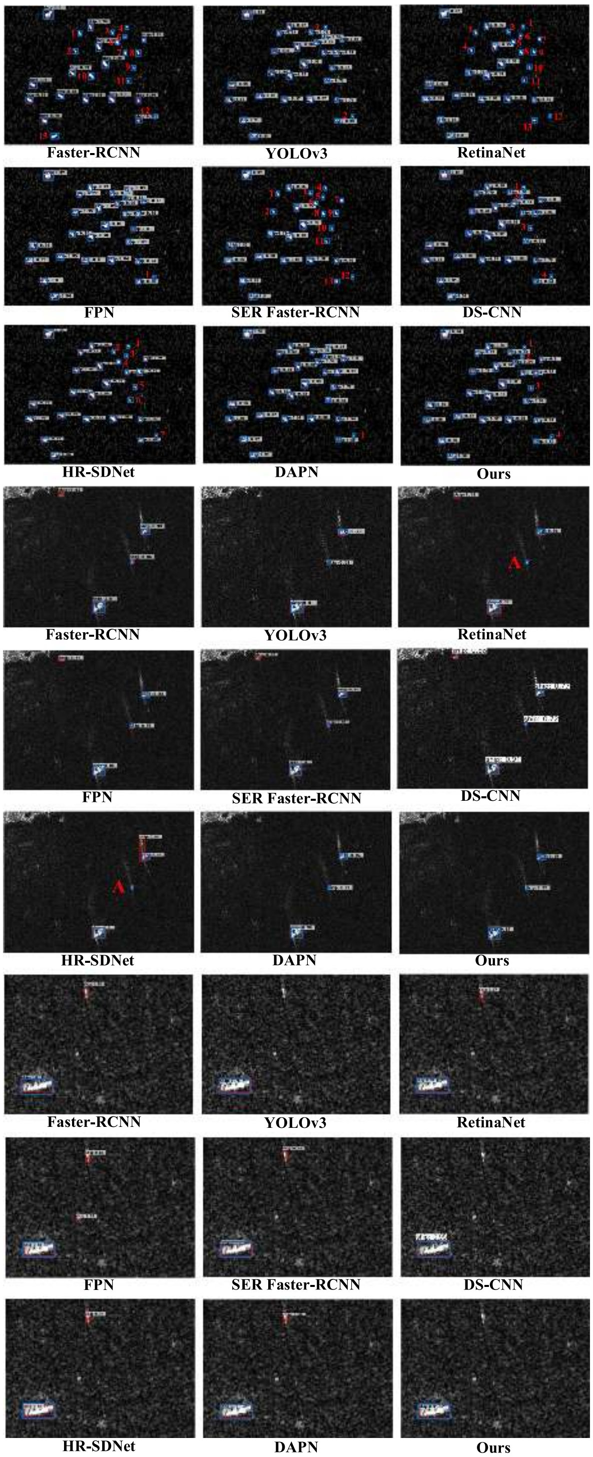


Fig. 16. Visual detection results of CNN-based methods on offshore ships. Four conditions of offshore ships are tested. Rectangles with red color mark the ships predicted by different CNN-based methods. Rectangles with blue color mark the ground truth ships.

because of appropriate feature pyramids constructed by these methods. Therefore, ships with small sizes could be richly represented at low-level feature maps. At the second condition (the fourth row to the sixth row of Fig. 16), wakes of ships and surroundings, e.g., lands and isles, are obvious. Isles are detected as ships by Faster-RCNN, RetinaNet, FPN, SER Faster-RCNN, DS-CNN, and wakes of ships are detected as ships by DPAN. Besides, small ships, e.g., ship A, are missed by RetinaNet and DAPN. It might be because of the weak ability for extracting discriminative features of ships and interference. However, our method could detect these ships without any false alarms, which demonstrates the strong and robust feature representation abilities of our method. In the third condition (the seventh row to the ninth row of Fig. 16), the background is a little rough due to strong waves. In the detection results of different methods, these strong waves are detected as ships by all methods except YOLOv3, DS-CNN, and our method. Because of the effective feature extracting and refining abilities, our method discriminates ship and interference clearly. In the last condition (the last three rows of Fig. 16), several isles exist around ships. Although the three multiscale ships are detected by SER Faster RCNN, DS-CNN, HR-SDNet, and DAPN, isles are also detected as ships. Besides, some small ships are missed in the detection results of Faster-RCNN and RetinaNet, possibly due to the inappropriate feature levels adopted by Faster-RCNN and RetinaNet. However, our method could detect these ships without any false alarms. The reason may be that the powerful feature refinement mechanism, e.g., CBAM, helps our method represent and discriminate ships and false distinctly.

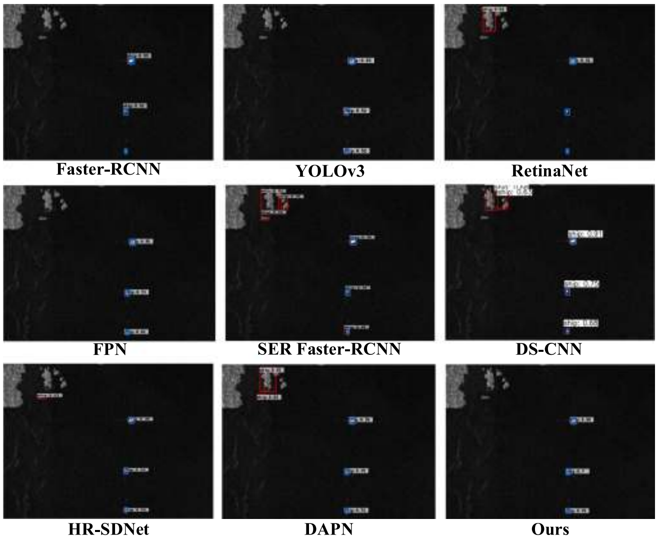


Fig. 16. Continued.

2) Detection Results of Inshore Ships: The detection results of different CNN-based methods tested on inshore ships are shown in Table VII. The first four rows and the fifth to the eighth rows of Table VII show the detection results of classical CNN-based methods and the CNN-based methods designed for object detection in SAR images, respectively.

According to Table VII, due to much interference around inshore ships, precision rate, recall rate, F1, and AP decrease sharply. However, the detection scores, e.g., recall rate, precision rate, F1, and AP, of our method are still higher than those of the classical CNN-based methods (the first four rows of Table VII). Although feature pyramids are adopted in FPN, RetinaNet, and YOLOv3 for extracting multiscale features, F1

TABLE VII
DETECTION RESULTS OF CNN-BASED METHODS ON INSHORE SHIPS

Method	Recall	Precision	F1	AP	FPS
Faster-RCNN	0.816	0.667	0.734	0.746	16
YOLOv3	0.709	0.730	0.719	0.785	26
RetinaNet	0.700	0.727	0.713	0.703	18
FPN	0.816	0.627	0.709	0.811	12
SER Faster-RCNN	0.709	0.403	0.514	0.621	32
DS-CNN	0.786	0.802	0.794	0.808	77
DAPN	0.718	0.685	0.701	0.772	14
HR-SDNet	0.835	0.656	0.734	0.800	11
Ours	0.854	0.733	0.789	0.841	13

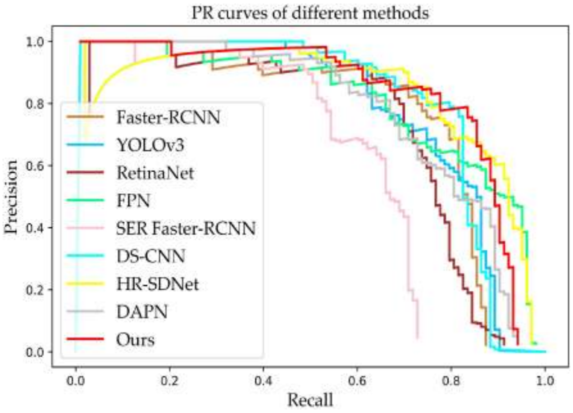


Fig. 17. PR curves of CNN-based methods tested on inshore ships.

of our method is still 8.0%, 7.6%, and 7.0% higher than scores of these methods. Besides, AP of our method is the highest compared with the classical CNN-based methods. Compared with the methods designed for object detection in SAR images, F1 of our method is 8.8% and 5.5% higher than those of DAPN and HR-SDNet but 0.5% lower than DS-CNN, respectively. Besides, AP of our method is 6.9%, 4.1%, and 3.3% higher than those of DAPN, HR-SDNet, and DS-CNN, respectively. It might be because of the well-designed feature extraction, feature fusion, and feature refinement strategies designed in our method. Besides, maybe because of simple feature extraction networks used in Faster-RCNN and DS-CNN, the confusing characteristics in feature maps may not be redundant and cause limited effects on the final detection. Therefore, scores of F1 achieved by Faster-RCNN and DS-CNN are competitive among other methods.

Fig. 17 shows the PR curves of different CNN-based methods tested on inshore ships. In Fig. 17, large differences of the PR curves emerge with an increase of recall rate. Furthermore, the PR curve of SER Faster-RCNN is lower than those of other methods when recall rate is higher than about 0.5. Besides, the trends of other methods' PR curves are not very constant. With an increase of recall rate, jitter might occur. However, the PR curve of ARPN (the red curve in Fig. 17) is generally stable. Although the PR curves of HR-SDNet (the yellow curve in Fig. 17) and FPN (the green curve in Fig. 17) are higher than that of ARPN when recall rate is greater than 0.9, the PR curve of FPN is lower than that of our method when recall rate is lower than 0.9 and the PR curve of HR-SDNet comes to a sharp decrease when recall rate increases from 0 to 0.2. Comprehensively, our method not only acquires high precision and recall rates but also achieves a balance between them. It also proves the effectiveness of our method for detecting inshore ships.

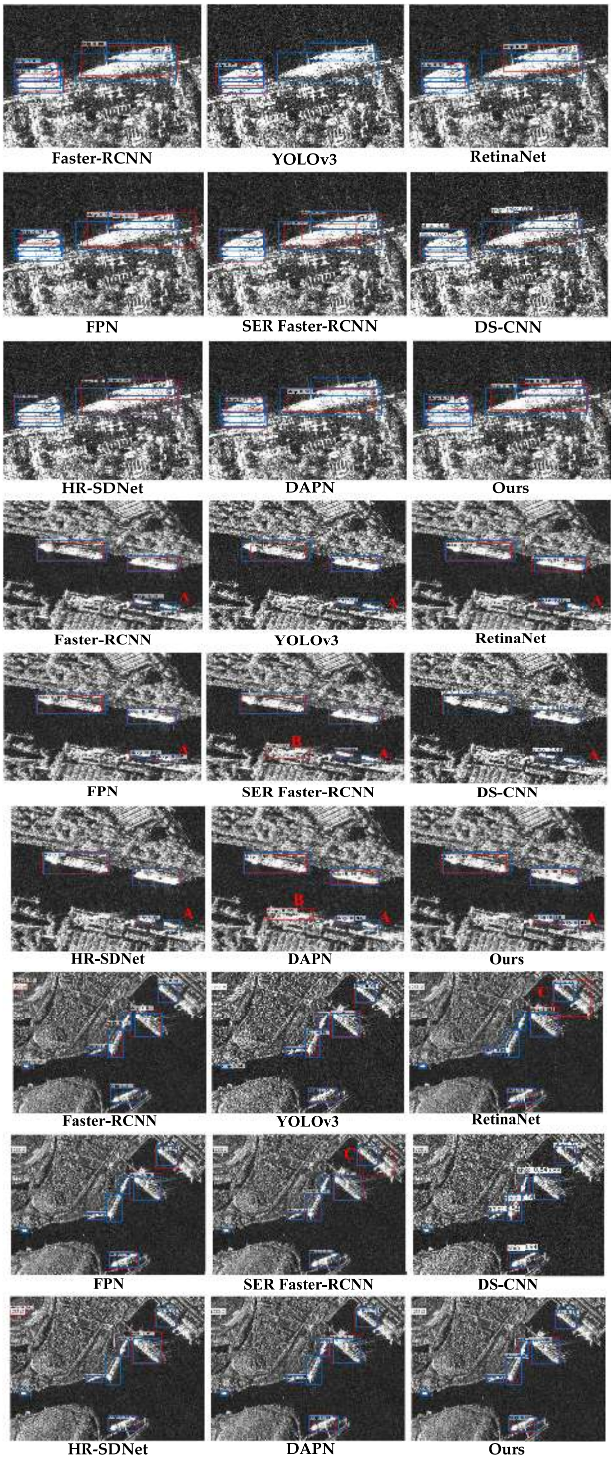


Fig. 18. Visual detection results of CNN-based methods on inshore ships. Four conditions of inshore ships are shown. Rectangles with red color mark the ships predicted by different CNN-based methods. Rectangles with blue color mark the ground truth ships.

Besides, four groups of detection results tested in different conditions of inshore ships are shown in Fig. 18. In the first

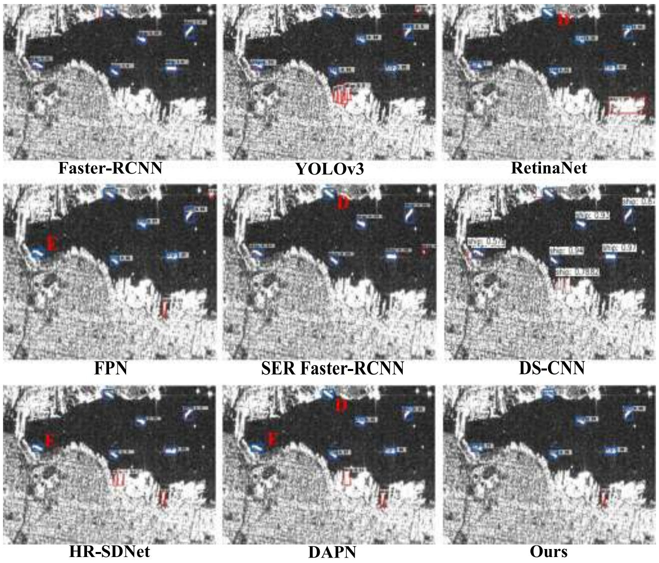


Fig. 18. Continued.

TABLE VIII
DETECTION RESULTS ON TWO LARGE-SCENE SAR IMAGES

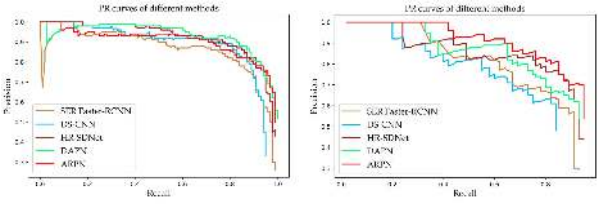
Product ID	Method	AP	Time(s)
1774400	SER Faster-RCNN	0.867	35
	DS-CNN	0.875	14
	HR-SDNet	0.921	97
	DAPN	0.937	52
	ARPN	0.904	65
3375753	SER Faster-RCNN	0.781	17
	DS-CNN	0.701	6
	HR-SDNet	0.829	41
	DAPN	0.820	18
	ARPN	0.873	23

condition (the first three rows of Fig. 18), some densely arranged inshore ships are missed by the compared methods. However, most ships are detected by our method and DS-CNN. In the second condition (the fourth row to the sixth row of Fig. 18), small inshore ships, e.g., ship A, are missed by Faster-RCNN, YOLOv3, RetinaNet, FPN, DS-CNN, SER Faster-RCNN, and HR-SDNet. Besides, false alarms exist in the detection results of SER Faster-RCNN and DAPN, e.g., Region B. However, our method detects these ships correctly. In the third condition (the seventh row to the ninth row of Fig. 18), multiscale ships exist in complex surroundings. And more than one ship is missed by RetinaNet, FPN, HR-SDNet, and DAPN. The locations of ships predicted by RetinaNet and SER Faster-RCNN are inaccurate, e.g., Ship C, compared with the ground truth ships. Besides, the number at the upper left of the image is detected as a ship by Faster-RCNN and HR-SDNet. However, our method could detect these multiscale ships with high probabilities and localization accuracy. In the last condition (the last three rows of Fig. 18), several ships are missed by different methods. Ship D is missed by RetinaNet, SER Faster-RCNN. Ship E is missed by FPN, HR-SDNet. Both ships D and E are missed by DAPN. Besides, more than one false alarm exists in the detection results of HR-SDNet, YOLOv3, and DAPN. Although there exists a false alarm in the detection results of our method, a lower probability is assigned to the false alarm than those predicted by other methods. Besides, our method detects multiscale ships with higher location accuracy than both YOLOv3 and Faster-RCNN.

3) Detection Results on Large-Scene SAR Images: In this section, two large-scene SAR images are adopted to judge the performance of our method and other CNN-based methods designed for object detection in SAR images, e.g., SER Faster-RCNN, DS-CNN, HR-SDN, and DAPN. In order to acquire the best performance of these methods, we clip the original large-scene SAR images into small chips with fixed overlapped pixels. The first image with a resolution of 10 m (Product ID 1774400, as shown in Table I) is clipped into small chips with a size of 350 × 350 pixels. The second image with a resolution of 1 meter (Product ID 3375753, as shown in Table I), is clipped into small chips with a size of 2000 × 2000 pixels. Then ships are detected on these small chips by different methods and the corresponding results on small chips are merged into the whole images. Finally, the final detection results are acquired after a Global Non-Maximum Suppression with a fixed threshold 0.5. AP and FPS are adopted to evaluate the detection accuracy and speed of different methods, respectively. The detection results are shown in Table VIII.

In terms of detecting the first image (Product ID 1774400), AP of our method is 3.6% and 2.9% higher than those of SER Faster-RCNN and DS-CNN. It might be because of the strong feature representation ability of our method for small ships in the large-scene image with low resolution. Furthermore, finer features could be acquired and carefully refined by RFB and CBAM, respectively, in our method. However, because of the dense vertical connections in top-down pathway of DAPN and high-resolution features maintained at all stages of HR-SDNet, characteristics of offshore ships with small sizes are richly represented. Thus, scores of AP for HR-SDNet and DAPN are 1.7% and 3.3% higher than our method, respectively. In terms of detection speed, DS-CNN runs faster than the other four methods because of its simple processing scheme as well as depth-wise and point-wise convolutional operations. Maybe because of the complex feature fusion strategies and three cascade processing schemes, HS-SDNet costs much time than the other compared methods.

In terms of detecting the second image (Product ID 3375753), due to the influence of SAR image resolution and mode, ships in this image might be different from those in SSDD in several aspects, e.g. geometrical sizes, textures, distributions of backscattering points. Because of the weak feature extraction abilities of the backbone networks and the vanilla feature pyramids adopted by SER Faster-RCNN and DS-CNN, the two methods acquire lower AP than those achieved by the other three methods. Although the high-resolution and rich semantic features of ships are extracted and retained by HR-SDNet and DAPN, the surroundings of ships, e.g., isles, wakes of ships, waves, might be also enhanced in these hierarchical feature maps. It may lead to unsatisfactory performance of detecting



(a) (b)

Fig. 19. PR curves of five methods when detecting ships on the two large-scene SAR images. (a) PR curves on the first large-scene SAR image (Product ID 1774400). (b) PR curves on the second large-scene image (Product ID 3375753).

multi-scale ships by using HR-SDNet and DAPN. Specifically, AP of our method is 17.2%, 9.2%, 5.3%, and 4.4% higher than these methods, respectively. As for the detection speed, DS-CNN still detects ships faster than the other methods results of its simple network structure. Besides, maybe because of the moderate network complexity of our method, it runs faster than HR-SDNet but slower than DAPN.

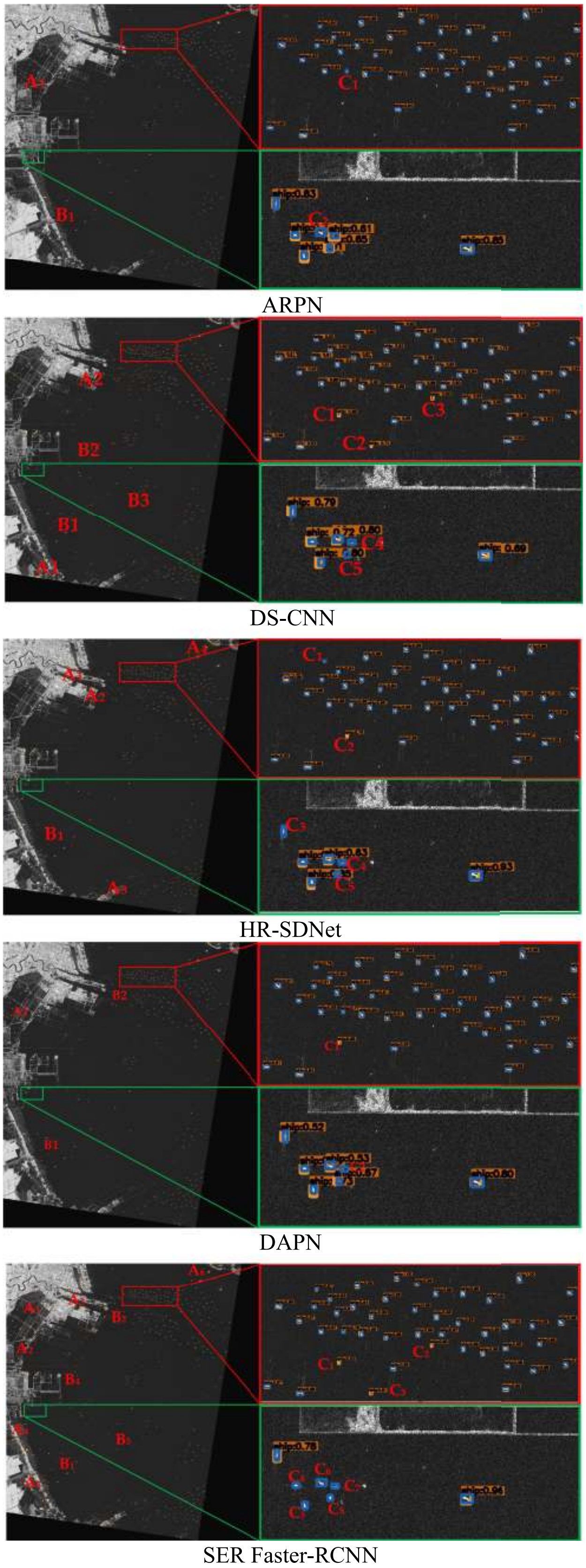
Fig. 19(a) shows the PR curves of different methods tested on the first large-scene SAR image. The PR curve of our method is lower than those of DAPN (the green curve) and HR-SDNet (the brown curve) at most time with an increase of recall rate. Besides, there is a sharp decrease at the PR curve of SER Faster-RCNN when the recall rate increases from about 0 to 0.1. Fig. 19(b) shows the PR curves of different methods tested on the second large-scene SAR image. It is obvious that the PR curve of our method is higher than those of other compared methods by a large margin at all time with an increase of recall rate. It proves that our method has better adaptability to the SAR images from different sources.

Besides, visual detection results on the first large-scene SAR image (Product ID 1774400) are shown in Fig. 20.

In Fig. 20, ships are very small due to low resolution of the image and some conclusions are summarized as follows.

First, among detection results of all the methods shown on the left side of Fig. 20, many false alarms exist on land, e.g., area A_1, areas A_1, A_2, areasA_1 to A_4, area A_1, and areas A_1 to A_6, and in the sea, e.g., area B_1, areas B_1 to B_3, area B_1, areas B_1, B_2, and areas B_1 to B_4, in the detection results of ARPN, DS-CNN, HR-SDNet, DAPN, and SER Faster-RCNN, respectively. It might be because of the similar backscattering characteristics between ships and the surroundings. However, false alarms in the detection results of our method are fewer than those of the compared methods. It might be because of the powerful feature refinement strategies, i.e., CBAM, adopted by our method. Maybe due to the weak feature discrimination, a large number of false alarms exist in the detection results of SER Faster-RCNN than those of the other methods.

Second, most offshore ships are easy to be detected because of their distinct features and clear surroundings. Specifically, although ships in the red rectangular are small, all these methods could detect these ships and only a few ships are missed, e.g., ship C_1 in the detection results of ARPN and HR-SDNet. Besides, some false alarms are detected as ships, e.g., ships C_1 to C_3, ship C_2, ship C_1, and ships C_1 to C_3, in the detection



ARPN

DS-CNN

HR-SDNet

DAPN

SER Faster-RCNN

Fig. 20. Detection results of different methods on the first large-scene SAR image (Product ID 1774400). Rectangles with blue and orange colors refer to the ground truth ships and the predictions, respectively. Two special areas marked with red and green rectangles are enlarged and shown at the right sides.

results of DS-CNN, HR-SDNet, DAPN, and SER Faster-RCNN, respectively.

Third, the performance of our method for detecting densely arranged ships is better than the other methods. In the green rectangular, several inshore ships distribute densely, however, only ship C_2 is missed by our method. Although the number of missed ships in the detection results of DAPN is equal to that of our method, e.g., ship C_2, the detected ships are assigned lower probabilities by DAPN. Besides, many ships are missed, e.g., ships C_4, C_5 and ships C_3 to C_5, in the detection results of DS-CNN and HR-SDNet, respectively. Ships C_4 to C_8 are also missed in the detection results of SER Faster-RCNN, possibly due to the lack of strong feature extraction for these ships.

The detection results on the second large-scene SAR image (Product ID 3375753) are shown in Fig. 21. Images on left side of Fig. 21 are the detection results of different method on the large-scene SAR images. Two specific areas marked with green and red rectangles are enlarged and shown on the right side of Fig. 21.

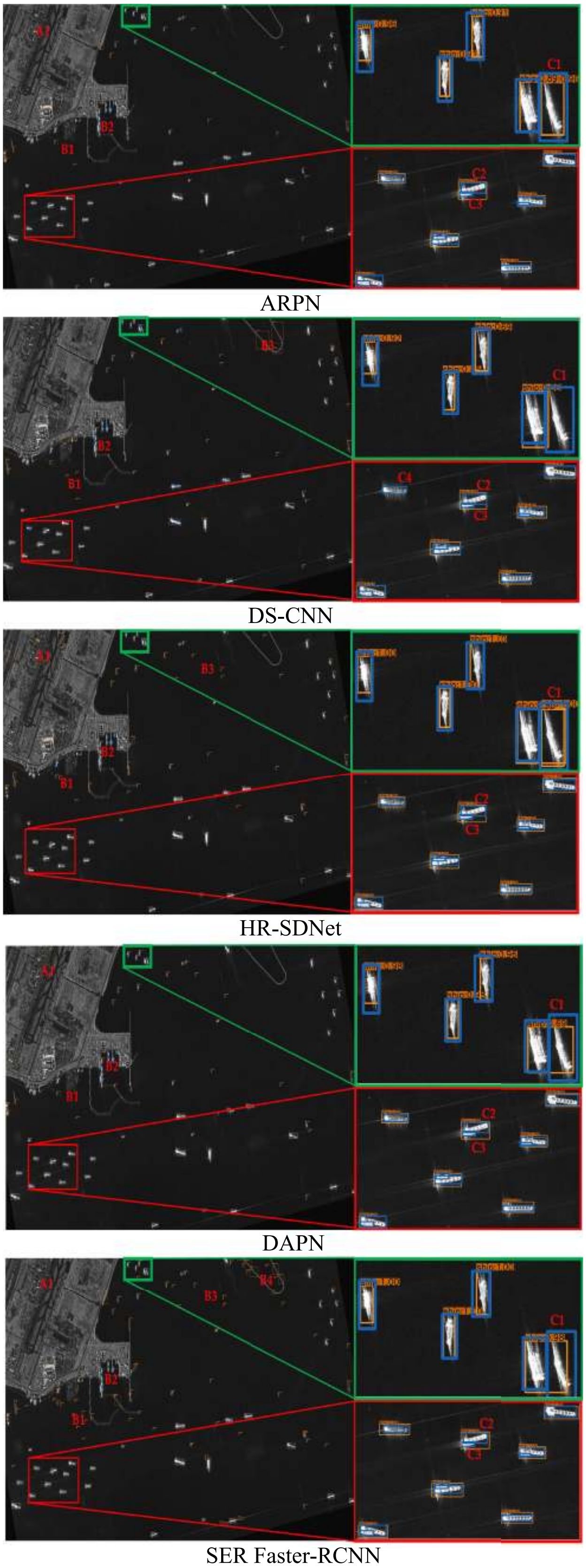
The detection results on the two large-scene SAR images with different resolutions, i.e., images with Product IDs 1774400 and 3375753, demonstrate that our method detects multiscale ships with competitive results and has a high generalization ability compared with the other CNN-based methods designed for object detection in SAR images.

V. Discussion

Because of the contributions of RFB and CBAM, the proposed method could detect multiscale ships in complex environments. However, there also exist other strategies for extracting semantic features of multiscale objects, e.g., Stem [28], [50], Atrous Spatial Pooling Pyramid (ASPP) [55]–[57], and for feature refinement, e.g., SE [51] block. Therefore, it might be necessary to further exploit the effectiveness of RFB and CBAM. In this section, different multiscale feature extraction strategies and attention mechanisms are discussed by replacing RFB and CBAM with other modules, respectively. Indicators such as recall rate, precision rate, F1, AP, and FPS are also utilized to evaluate the performance of different algorithms. Besides, basic and fine-grained feature maps at bottom-up and top-down pathways of our method are shown to demonstrate inner feature processing mechanisms of RFB and CBAM, respectively.

A. Multiscale Feature Extraction

There are several useful strategies, e.g., Stem and ASPP, for detecting multiscale objects. Fig. 22(a) and (b) show the structures of a Stem and an ASPP, respectively. A Stem consists of several branches of convolutional layers with various kernel sizes to capture fine-grained and rich semantic features. However, compared with RFB, the feature representation abilities of Stem might be still weak by using plain convolutions with a kernel size of 3 × 3. Besides, there are many parameters in a Stem as a result of stacking several convolutional layers. ASPP, which consists of several parallel convolutional layers, is introduced for image segmentation at first. It not only enlarges receptive



ARPN

DS-CNN

HR-SDNet

DAPN

SER Faster-RCNN

Fig. 21. Detection results of different methods on the second large-scene SAR image (Product ID 3375753). Rectangles with blue and orange colors refer to the ground truth ships and the predictions, respectively. Two special areas marked with red and green rectangles are enlarged and shown at the right sides.

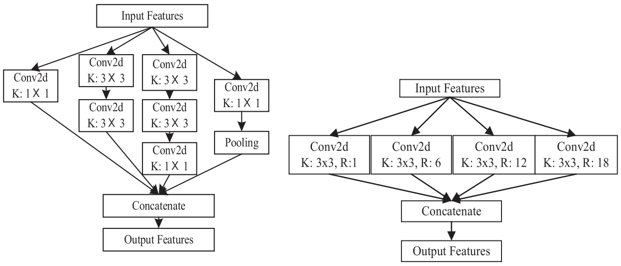


Fig. 22. Multiscale feature extraction strategies. (a) Stem. (b) ASPP.

TABLE IX
DETECTION RESULTS OF ARPN WITH DIFFERENT MULTISCALE MODULES

Type	Method	Recall	Precision	F1	AP	FPS
Offshore	*(RFB)	0.964	0.964	0.964	0.982	12
	* (Stem)	0.966	0.950	0.958	0.982	13
	* (ASPP)	0.947	0.947	0.947	0.970	14
Inshore	* (RFB)	0.854	0.733	0.802	0.841	12
	* (Stem)	0.767	0.669	0.714	0.792	13
	* (ASPP)	0.573	0.702	0.631	0.630	14

* refers to ARPN.

fields of convolutional kernels but also reduces capacities of parameters by using convolutions with constant kernel sizes but various dilation rates at different branches. Compared with ASPP, RFB also consists of multibranch convolutional layers with different dilation rates but a little more complex because of several specific convolutional operations with asymmetric kernel sizes and a shortcut connection involved.

To exploit the effectiveness of specific structure adopted by RFB, we test the modified networks, i.e., ARPN(Stem) and ARPN(ASPP), on offshore and inshore ships by replacing RFB with a Stem and an ASPP, respectively. Table IX shows the detection results in detail.

In Table IX, the first and the last three rows of Table IX show the detection results of modified algorithms tested on offshore and inshore ships, respectively. In terms of detecting offshore ships, recall rate, precision rate, F1 and AP all come to slight decreases by using ARPN(Stem) and ARPN(ASPP). Specifically, F1 of our method is 0.6% and 1.7% higher than those of the two modified networks. AP of ARPN(RFB) and ARPN(Stem) are the same while 1.2% higher than that of ARPN(ASPP). However, because of the simplest structure of ASPP among RFB, Stem, and ASPP, ARPN(ASPP) runs with the highest FPS among ARPN(RFB), ARPN(Stem), and ARPN(ASPP). In terms of detecting inshore ships, there are great differences among original ARPN and the two modified ARPN algorithms. For example, recall rate of ARPN(RFB) is 8.7% and 28.1% higher than those of ARPN(Stem) and ARPN(ASPP), respectively. Precision rate of our method is 3.1% and 3.4% higher than those of ARPN(ASPP) and ARPN(Stem), respectively. Besides, F1 of ARPN(RFB) is 8.8% and 17.1% higher than those of ARPN(Stem) and ARPN(ASPP), respectively. As for AP, the score of ARPN(RFB) is 4.9% and 21.1% higher than those of ARPN(Stem) and ARPN(ASPP), respectively. Because of the inappropriate settings of dilation rates in ASPP, the receptive fields of ASPP might be large but with some holes in feature maps. Hence, much significant information is missed, which leads to serious gridding effects [58]. The performance of ARPN(RFB) and ARPN(Stem) is similar. However, ARPN(RFB) performs better than ARPN(ASPP) by a large margin. The reason may be that the Stem has more similar receptive fields with RFB than ASPP.

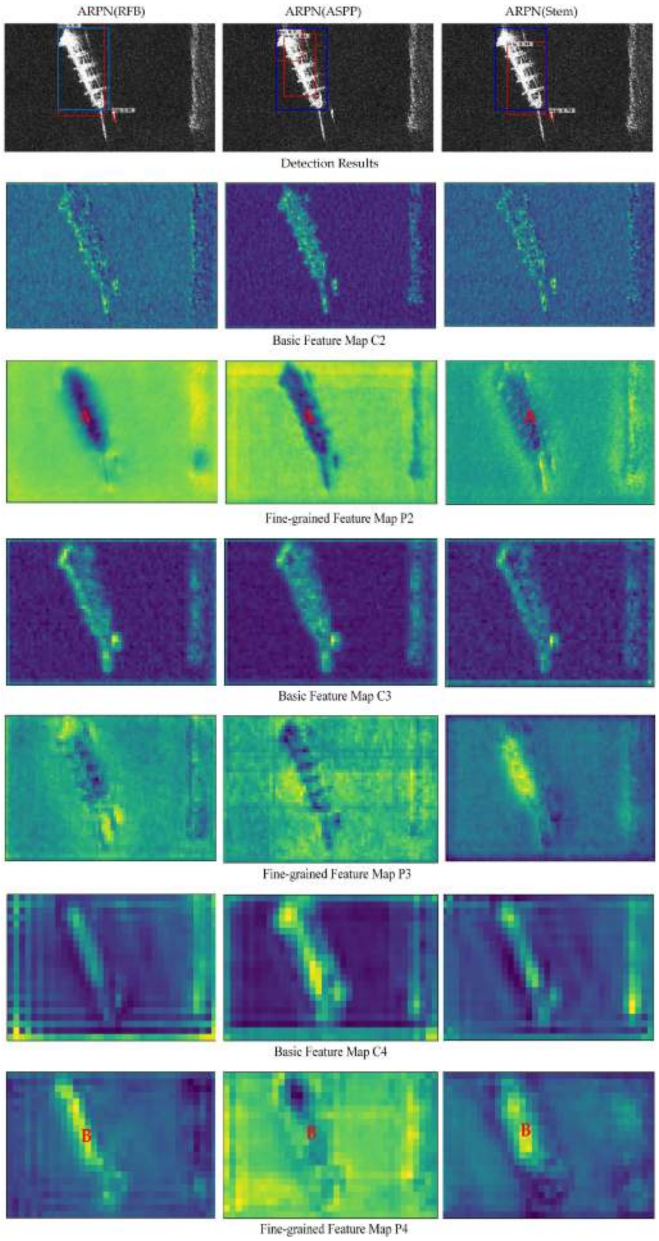


Fig. 23. Detection results and feature maps of ARPN(RFB), ARPN(ASPP), and ARPN(Stem) tested on offshore ships.

Besides, basic and fine-grained feature maps, e.g., C2, C3, C4, P2, P3, and P4, tested on offshore and inshore ships are shown in Figs. 23 and 24, respectively.

Fig. 23 shows the detection results, the basic feature maps, C2, C3, and C4, and the fine-grained feature maps, P2, P3, and P4, of ARPN(RFB), ARPN(ASPP), and ARPN(Stem) tested on

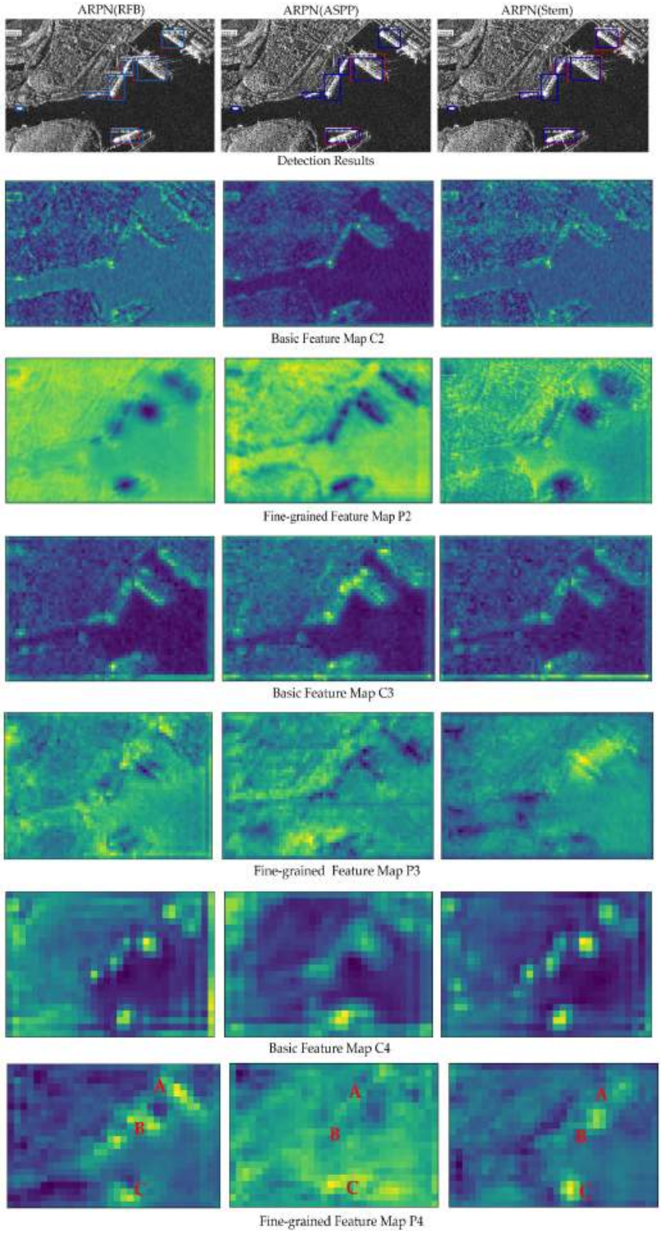


Fig. 24. Detection results and feature maps of ARPN(RFB), ARPN(ASPP), and ARPN(Stem) tested on inshore ships.

TABLE X
DETECTION RESULTS OF ARPN WITH DIFFERENT ATTENTION MECHANISMS

Type	Method	Recall	Precision	F1	AP	FPS
Offshore	* (CBAM)	0.964	0.964	0.964	0.982	12
	* (SE)	0.959	0.934	0.950	0.970	13
Inshore	* (CBAM)	0.854	0.733	0.802	0.841	12
	* (SE)	0.718	0.611	0.663	0.752	13

* refers to ARPN.

offshore ships. In the detection results (the first row of Fig. 23), rectangles with blue and red colors refer to the ground truth ships and the predictions by the methods, respectively. Compared with the two modified ARPN networks, ARPN(RFB) performs better. Moreover, the locations of ships predicted by our method are more accurate than those predicted by ARPN(Stem) and ARPN(ASPP). Besides, our method assigns lower probabilities for the false alarms than those assigned by the other two methods. According to the feature maps (the second to the last rows of Fig. 23), active areas in the fine-grained feature maps, P2, P3, and P4, constructed by our method distribute centrally. The highlighted areas mostly exist at main body of the ships, e.g., area A in P2 and area B in P4. However, active areas predicted by ARPN(ASPP) and ARPN(Stem) are blurry and disperse a little at the same areas. Specifically, active area B in P4 of ARPN(ASPP) is chaotic, which might lead to serious offsets between the ground truth ships and the predictions.

Fig. 24 shows the detection results and the basic feature maps, C2, C3, and C4, and the fine-grained feature maps, P2, P3, and P4, of ARPN(RFB), ARPN(ASPP), and ARPN(Stem) tested on inshore ships. Compared with ARPN(RFB) and ARPN(Stem), several ships are missed by ARPN(ASPP). Additionally, the active areas predicted by ARPN(RFB) are integral and distinct at P2. However, the same active areas predicted by ARPN(Stem) are more dispersed. Besides, textured features of the background areas in P2 outputted by ARPN(RFB), are more uniform than those predict by ARPN(Stem) and ARPN(ASPP). Although P4 is 16 times downsampled of the input images, the active areas extracted by ARPN(RFB) are still discriminative, e.g., areas A, B, and C in the last row of Fig. 24. Because of the appropriate receptive fields constructed by RFB, the primary areas of ships might be completely represented at several fine-grained feature maps. Moreover, the asymmetrical convolutional kernels, i.e., kernel sizes of 1×3 and 3×1, might be more effective for grabbing features of ships with various aspect ratios. However, the simple structures and unmatched receptive fields of ASPP might miss significant features of ships. It might lead the features between ships and surroundings at high feature levels might be chaotic and blurry.

B. Attention Mechanisms

In this section, CBAM is substituted by a SE module to verify the effectiveness of channel and spatial attention mechanisms. We name the original ARPN and ARPN with CBAM as ARPN(CBAM) and ARPN(SE), respectively. Table X shows the detection results of the two methods in detail.

The first and the last two rows of Table X are the detection results of different algorithms tested on offshore and inshore ships, respectively. Generally, most indicators decrease a little when using ARPN(SE). Scores of recall rate, precision rate, F1, and AP achieved by ARPN(SE) are 0.5%, 3.0%, 1.4%, and 1.2% lower than those of ARPN(CBAM) on offshore ships. Because of the simple and clear surroundings, there are slight differences between ARPN(SE) and ARPN(CBAM). However, when detecting inshore ships, scores of these indicators achieved by ARPN(SE) are 13.6%, 12.2%, 13.9, and 8.9% lower than those of ARPN(CBAM). Because of the unique spatial attention mechanism, the interference caused by surroundings might be suppressed effectively at spatial dimension and the networks

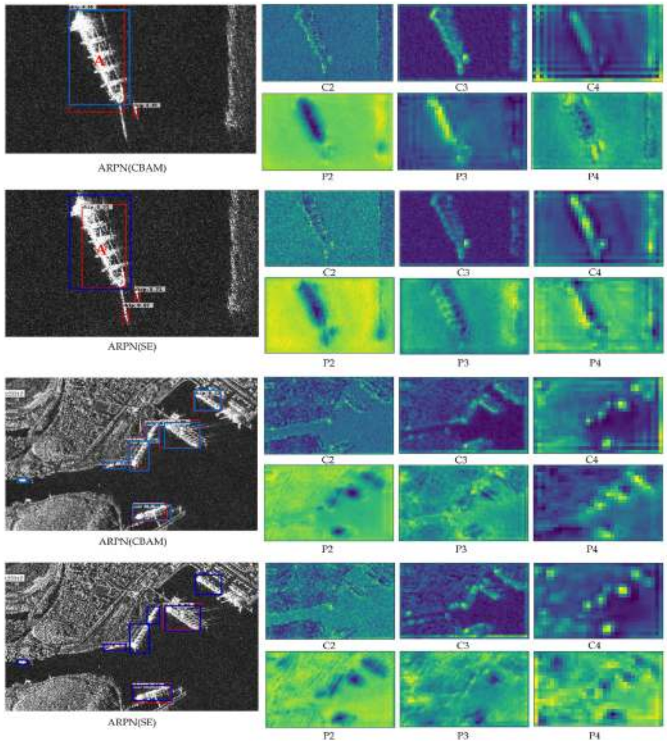


Fig. 25. Detection results and feature maps of ARPN(CBAM) and ARPN(SE) on offshore and inshore ships.

could pay more attention to significant features of ships than ARPN(SE). Additionally, the testing speeds between the two methods are almost the same. It might be because that only slight extra computing cost is introduced by the spatial attention mechanism of CBAM.

Detection results and different levels of feature maps acquired by ARPN(CBAM) and ARPN(SE) are shown in Fig. 25. In each row of Fig. 25, the final detection result is placed on the left and the different feature maps, C2, C3, C4, P2, P3, and P4, are placed on the right. Besides, the first two rows of Fig. 25 are detection results on offshore ships and the last two rows of Fig. 25 are the detection results on inshore ships.

In terms of detecting offshore ships, two false alarms exist in the final results of ARPN(SE). Besides, ship A is not completely enclosed by a bounding box predicted by ARPN(SE). However, locations of ship A predicted by our method are precise. In terms of different feature maps, areas of wakes at P2 of ARPN(SE) are strongly activated. However, areas of ships at P2 of ARPN(CBAM) are stronger than those of ARPN(SE). In terms of detecting inshore ships, the C2, C3, and C4 predicted by ARPN(CBAM) and ARPN(SE) are similar. However, several differences exist on the P2, P3 and P4 between ARPN(CBAM) and ARPN(SE). Furthermore, active areas of ships in these feature maps of ARPN(CBAM) are clearer and more distinct than those of ARPN(SE). Although P4 has the smallest sizes among these fine-grained feature maps, active areas of P4 extracted by ARPN(CBAM) are discriminative and their locations are also accurate. Because of contributions of spatial attention mechanism adopted in ARPN(CBAM), the significant features at specific areas could be boosted effectively. However, active areas of P4 predicted by ARPN(SE) are blurry.

In summary, RFB and CBAM are two essential modules for multiscale feature representation and refinement. RFB is more efficient than Stems and ASPP for enhancing the relationships among different ranges of features. CBAM is more important for refining redundant features than SE. Moreover, RFB and CBAM are complementary and the performance of our method is competitive by combining them reasonably.

VI. Conclusion

In this article, a two-stage detector called ARPN was proposed for detecting multiscale ships in SAR images. We carefully design a well-designed lateral connection named ARB to extract representative features of multiscale ships and suppress interference of surroundings by combining RFB with CBAM reasonably. Specifically, RFB, which consists of multibranch convolutional layers with specific asymmetric kernel sizes and dilation rates, was utilized to extract information of multiscale ships with various directions as well as enhancing relationships of nonlocal features. CBAM was adopted to make the network focus on significant features for detecting ships by reweighting feature maps using channel and spatial attention modules in sequence. In the experimental part, we proved superiorities of our method by exploiting the contributions of RFB and CBAM separately as well as evaluating the performance of our method with some CNN-based methods on the SSDD and two large-scene SAR images. Besides, we further compared RFB and CBAM with other multiscale feature extraction and refinement strategies, respectively. The inner processing mechanisms reflected by visualized feature maps illustrate that our method could detect multiscale ships with preferable performance and competitive generalization.

In the future, we will concentrate on combining backscattering properties of ships in SAR images with convolutional design of networks and introducing a strong restriction, e.g., mask, to further improve the detection accuracy as well as detection speed.

References

[1] D. Cerutti-Maori, J. Klare, A. R. Brenner, and J. H. Ender, “Wide-area traffic monitoring with the SAR/GMTI system PAMIR,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 3019–3030, Oct. 2008.

[2] S. Brusch, S. Lehner, T. Fritz, M. Soccorsi, A. Soloviev, and B. van Schie, “Ship surveillance with TerraSAR-X,” *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1092–1103, Mar. 2011.

[3] A. S. Solberg, G. Storvik, R. Solberg, and E. Volden, “Automatic detection of oil spills in ERS SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 4, pp. 1916–1924, Jul. 1999.

[4] S. Quan, B. Xiong, D. Xiang, and G. Kuang, “Derivation of the orientation parameters in built-up areas: With application to model-based decomposition,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4714–4730, Aug. 2018.

[5] X. Leng, K. Ji, S. Zhou, X. Xing, and H. Zou, “Discriminating ship from radio frequency interference based on noncircularity and non-Gaussianity in sentinel-1 SAR imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 352–363, Jan. 2018.

[6] S. Quan, B. Xiong, S. Zhang, M. Yu, and G. Kuang, "Adaptive and fast prescreening for SAR ATR via change detection technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1691–1695, Nov. 2016.

[7] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[8] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[9] C. He, M. Tu, D. Xiong, F. Tu, and M. Liao, "A component-based multi-layer parallel network for airplane detection in SAR imagery," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1016.

[10] C. C. Wackerman, K. S. Friedman, W. G. Pichel, P. Clemente-Colón, and X. Li, "Automatic detection of ships in RADARSAT-1 SAR imagery," *Can. J. Remote Sens.*, vol. 27, no. 5, pp. 568–577, 2001.

[11] F. Xu and J.-H. Liu, "Ship detection and extraction using visual saliency and histogram of oriented gradient," *Optoelectron. Lett.*, vol. 12, no. 6, pp. 473–477, 2016.

[12] A. C. Copeland, G. Ravichandran, and M. M. Trivedi, "Localized Radon transform-based detection of ship wakes in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 1, pp. 35–45, Jan. 1995.

[13] H. Lang, J. Zhang, X. Zhang, and J. Meng, "Ship classification in SAR image by joint feature and classifier selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 212–216, Feb. 2015.

[14] C. Wang, S. Jiang, H. Zhang, F. Wu, and B. Zhang, "Ship detection for high-resolution SAR images based on feature analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 119–123, Jan. 2014.

[15] S. Watts, "Radar detection prediction in sea clutter using the compound K-distribution model," *IEE Proc. F, Commun., Radar Signal Process.*, vol. 132, no. 7, pp. 613–620, 1985.

[16] Z. He, X. Zhou, J. Lu, and G.-Y. Kuang, "A fast CFAR detection algorithm based on the G0 distribution for SAR images," *J. Nat. Univ. Defense Technol.*, vol. 31, no. 1, pp. 47–51, 2009.

[17] X. Leng, K. Ji, K. Yang, and H. Zou, "A bilateral CFAR algorithm for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1536–1540, Jul. 2015.

[18] M. Kang, X. Leng, Z. Lin, and K. Ji, "A modified faster R-CNN based on CFAR algorithm for SAR ship detection," in *Proc. Int. Workshop Remote Sens. With Intell. Process.*, 2017, pp. 1–4.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[20] L. Liu *et al.*, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, pp. 261–381, 2019.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 91–99, Jun. 2015.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, 2017, pp. 2261–2269.

[30] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[32] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[33] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 860.

[34] Y. Wang, C. Wang, H. Zhang, C. Zhang, and Q. Fu, "Combing single shot multibox detector with transfer learning for ship detection using Chinese GAOFEN-3 images," in *Proc. Prog. Electro. Res. Symp.-Fall (PIERS-FALL)*, 2017, pp. 712–716.

[35] Y.-L. Chang, A. Anagaw, L. Chang, Y. C. Wang, C.-Y. Hsiao, and W.-H. Lee, "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, vol. 11, no. 7, p. 786, 2019.

[36] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[37] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019.

[38] Q. An, Z. Pan, L. Liu, and H. You, "DRBox-v2: An improved detector with rotatable boxes for target detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8333–8349, Nov. 2019.

[39] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.

[40] T. Zhang, X. Zhang, J. Shi, and S. Wei, "Depthwise separable convolution neural network for high-speed SAR ship detection," *Remote Sens.*, vol. 11, no. 21, 2019, Art. no. 2483.

[41] F. Chollet, "XCEPTION: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[42] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[43] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[44] S. Wei *et al.*, "Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 167.

[45] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[46] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[47] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.

[48] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era, Models, Methods Appl. (BIGSARDATA)*, 2017, pp. 1–6.

[49] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3150–3158.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[52] "tzutalin/labelImg," GitHub, 2020. [Online]. Available: https://github.com/tzutalin/labelImg, Accessed on: Mar. 3, 2020.

[53] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, 2017.

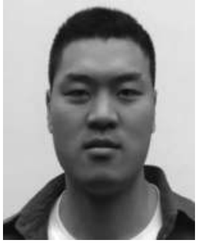
[54] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[55] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[56] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[57] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFS," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[58] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.



Yan Zhao received the B.S. degree in electronic engineering from Beijing Institute of Technology, Beijing, China, in 2018. He is currently working toward the M.S. degree in the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, Changsha, China.

His current research focuses on the application of machine learning and deep learning algorithms to SAR automatic target recognition.

Lingjun Zhao received the B.S., M.S., and Ph.D. degrees in information and communication engineering from National University of Defense Technology, Changsha, China, in 2003, 2004, and 2009, respectively.

She is currently an Associate Professor with the School of Electronic Science, National University of Defense Technology. Her current research interests include remote sensing information processing and SAR automatic target recognition.

Boli Xiong received the B.S. degree in electronic engineering, the M.S. degree in photogrammetry and remote sensing, and the Ph.D. degree in communication engineering from the National University of Defense Technology, Changsha, China, in 2004, 2006, and 2012, respectively.

He is currently an Associate Professor with the School of Electronic Science, National University of Defense Technology. His current research interests include SAR/PolSAR image registration and change detection, SAR automatic target recognition, pattern recognition, and machine learning.

Gangyao Kuang (Senior Member, IEEE) received the B.S. and M.S. degrees in geophysics from the Central South University of Technology, Changsha, China, in 1988 and 1991, respectively, and the Ph.D. degree in communication and information from the National University of Defense Technology, Changsha, in 1995.

He is currently a Professor with the School of Electronic Science, National University of Defense Technology. His current research interests include remote sensing, SAR image processing, change detection, SAR ground moving target indication, and classification with polarimetric SAR images.