Boosting Ship Detection in SAR Images With Complementary Pretraining Techniques

Wei Bao, Meiyu Huang, Yaqin Zhang, Yao Xu, Xuejiao Liu, and Xueshuang Xiang

***Abstract*—Deep learning methods have made significant progress in ship detection in synthetic aperture radar (SAR) images. The pretraining technique is usually adopted to support deep neural networks-based SAR ship detectors due to the scarce labeled SAR images. However, directly leveraging ImageNet pretraining is hard to obtain a good ship detector because of different imaging perspectives and geometry. In this article, to resolve the problem of inconsistent imaging perspectives between ImageNet and earth observations, we propose an optical ship detector (OSD) pretraining technique to transfer the characteristics of ships in earth observations to SAR images from a large-scale aerial image dataset. On the other hand, to handle the problem of different imaging geometry between optical and SAR images, we propose an optical-SAR matching (OSM) pretraining technique, which transfers plentiful texture features from optical images to SAR images by common representation learning on the OSM task. Finally, observing that the OSD pretraining-based SSD has a better recall on sea area while the OSM pretraining-based SSD can reduce false alarms on land area, we combine the predictions of the two detectors through weighted boxes fusion to further improve detection results. Extensive experiments on four SAR ship detection datasets and three representative convolutional neural network-based detection benchmarks are conducted to show the effectiveness and complementarity of the two proposed detectors, and the state-of-the-art performance of the combination of the two detectors. The proposed method won the sixth place of ship detection in SAR images in the 2020 Gaofen challenge.**

***Index Terms*—Common representation learning, optical ship detector (OSD) pretraining, optical-SAR matching (OSM) pretraining, ship detection, weighted boxes fusion (WBF).**

I. Introduction

SYNTHETIC aperture radar (SAR) is an active microwave remote sensing imaging radar with the capability of targeting objects in all-day and all-weather conditions and has been widely applied in many military and civil fields. Ship detection in high-resolution SAR images has drawn considerable attention for its broad application prospects, such as marine surveillance [1], military intelligence acquisition [2], etc. Many traditional SAR ship detection methods have been proposed [3]–[5] to detect multiscale ships in complex surroundings. As one of the most commonly used techniques, the constant false-alarm rate (CFAR) method [3], adaptively adjusts the threshold given a false-alarm rate and leverages the estimated statistical distributions to distinguish objects from the background with the calculated threshold. However, traditional detection methods suffer from tremendous difficulties in accurate detection due to weak feature extraction capabilities.

Recently, benefiting from the rapid development of deep learning, remarkable breakthroughs have been made in deep convolutional neural networks (CNN) [6]-based detection methods. Generally, CNN-based detection methods can be divided into two categories: two-stage detection methods, such as Faster R-CNN [7], Mask R-CNN [8], Cascade R-CNN [9]; and one-stage detection methods, such as SSD [10], RetinaNet [11], YOLO [12]. Specifically, two-stage detection methods adopt the feature maps generated from backbone networks, e.g., residual network (ResNet) [13], to preliminarily extract class-agnostic region proposals of the potential objects with negative locations filtered out, and then further refine these proposals and classify them into different categories. Unlike two-stage detection methods, one-stage detection methods omit the region proposals generation process and consider object detection as a regression problem to directly predict location coordinates and class probabilities for improving the detection speed, but with the precision reduced in general. However, with bags of efficient and effective tricks well used, one-stage detection methods, such as YOLOv4 [14] can also achieve comparable or even better performance to two-stage detection methods.

Because of the powerful feature extraction and representation ability, these CNN-based detection methods have been successfully applied to ship detection in SAR images. Based on two kinds of detection frameworks, effective network structures, training strategies, and tricks are designed to deal with multiscale ship detection, leading to significant performance improvement. As for the two-stage SAR ship detectors (SSDs), Jiao *et al.* [15] fused different feature maps by a densely connected multiscale neural network to solve multiscale ship detection. Lin *et al.* [16] proposed a squeeze and excitation rank architecture to suppress redundant information of feature maps for representative ability improvement based on VGG network [17] pretrained on

Manuscript received March 11, 2021; revised May 22, 2021 and July 30, 2021; accepted August 25, 2021. Date of publication September 1, 2021; date of current version September 16, 2021. This work was supported in part by the Beijing Nova Program of Science and Technology under Grant Z191100001119129 and in part by the National Natural Science Foundation of China under Grant 61702520. *(Corresponding author: Wei Bao.)*

Wei Bao is with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing 100094, China, and also with the School of Information and Electronics Engineering, Beijing Institute of Technology, Beijing 100811, China (e-mail: baowei97@163.com).

Meiyu Huang, Yao Xu, Xuejiao Liu, and Xueshuang Xiang are with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing 100094, China (e-mail: huangmeiyu@qxslab.cn; xuyao@qxslab.cn; liuxuejiao@qxslab.cn; xiangxueshuang@qxslab.cn).

Yaqin Zhang is with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing 100094, China, and also with the School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, China (e-mail: 493834755@qq.com).

Digital Object Identifier 10.1109/JSTARS.2021.3109002

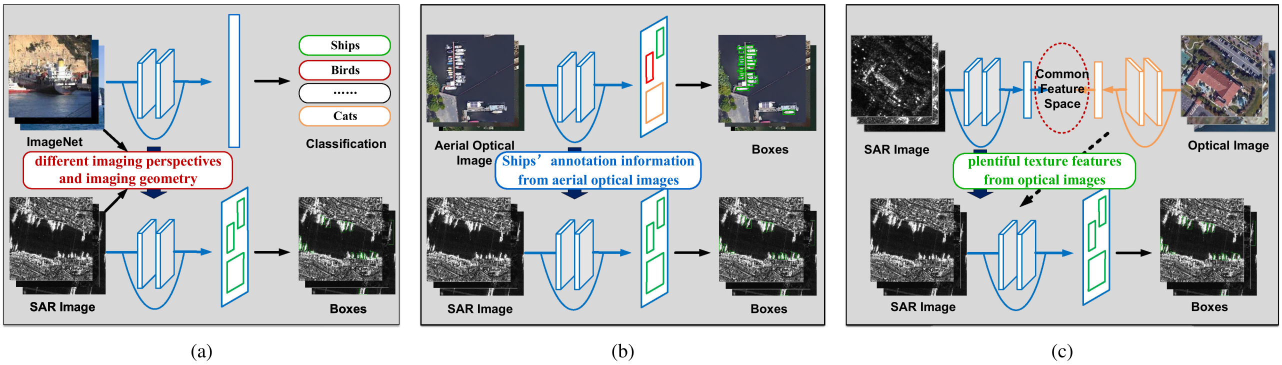

Fig. 1. Illustration of ImageNet pretraining, proposed OSD pretraining and OSM pretraining for SAR ship detection. (a) ImageNet pretraining: hard to obtain a good ship detector because of different imaging perspectives and geometry between ImageNet and SAR images; (b) OSD pretraining: solve the problem of different imaging perspectives by transferring ships' annotation information from aerial optical images to improve the feature learning ability of ships in SAR images; (c) OSM pretraining: solve the problem of different imaging geometry by transferring plentiful texture features from optical images to enhance the feature embedding of SAR ship detection via Optical-SAR common representation learning. (a) ImageNet pretraining. (b) OSD pretraining. (c) OSM pretraining.

ImageNet [18]. Cui *et al.* [19] adopted feature pyramid network (FPN) [20] with convolutional block attention module (CBAM) densely connected to each feature map to integrate resolution and semantic information. Wei *et al.* [21] introduced high-resolution ship detection network (HR-SDNet) to maintain high-resolution features, expecting to improve detection performance. Yan *et al.* [22] proposed attention receptive pyramid network (ARPN) with receptive fields block (RFB) and CBAM combined reasonably. RFB can enhance local features with global dependency, while CBAM can boost useful information and suppress the influence of surroundings. As for the one-stage SSDs, Du *et al.* [23] effectively leverages saliency information to improve the representation capability of original SSD network [10] and enhance target detection results by focusing more on informative regions. Zhang *et al.* [24] proposed a depth-wise separable convolution neural network that leverages multiscale mechanism, concatenation mechanism, and anchor box mechanism to improve ship detection speed dramatically. Inspired by the real-time idea of the YOLO algorithm, Zhang *et al.* [25] introduced a grid CNN equipped with depth-wise separable convolution to speed up the ship detection with little performance loss. Fu *et al.* [26] introduced a feature balancing module for the small-scale ship detection and a feature-refinement module to tackle feature misalignment for better localization accuracy. Instead of horizontal bounding boxes, Chen *et al.* [27] and An *et al.* [28] used the oriented bounding boxes to perform more suitable ship detection for the geospatially arranged objects.

On the one hand, the improvements brought by these detectors attribute to effective network structures. On the other hand, the ImageNet [18] pretraining technique, a common practice, is adopted to support all these SSDs due to the scarcity of labeled SAR images. However, directly using ImageNet pretraining is difficult to obtain a good SSD, which is also a significant issue but with less attention paid. As shown in Fig 1(a), one significant problem is the different imaging perspectives. The ships in ImageNet are taken under natural scenes, while those in SAR images are obtained from earth observations. The inconsistency across ships' viewing perspectives result in annotation information from ImageNet not applicable for SSDs. As depicted in Fig 1(b), we propose an optical ship detector (OSD) pretraining technique to improve the feature learning ability of ships in SAR images based on a large-scale aerial image dataset [29]. The OSD pretraining technique can transfer the characteristics of ships in earth observations to SAR images by fully taking advantage of a large amount of ships' annotation information from aerial images. Another critical problem of directly applying ImageNet pretraining to SAR ship detection in Fig 1(a) is the different imaging geometry between optical and SAR images, making ship detectors unable to obtain powerful SAR feature embedding. Hence, we propose an optical-SAR matching (OSM) pretraining technique to enhance the general feature embedding of SAR images. Specifically, the OSM pretraining technique transfers plentiful texture features from optical images to SAR images by common representation learning via bridge neural network (BNN) [30] on the OSM task. As depicted in Fig 1(c), BNN employs a couple of convolution neural networks (CNN) named left-CNN and right-CNN, which project SAR images and optical images into a common feature space, respectively (see Section II-C for a more detailed discussion). The optical-SAR matching task forces BNN to learn useful fusion features. Then, the left-CNN can be further used as the backbone of the SAR detection framework to perform ship detection.

Based on the two pretraining techniques, we can obtain the OSD pretraining-based SSD (OSD-SSD) and the OSM pretraining-based SSD (OSM-SSD). Furthermore, we propose to combine the two detectors to get a more comprehensively better detector based on the observation of their different advantages. Specifically, since the optical-SAR matching task mainly focuses on land area, the plentiful texture features from optical images can help distinguish the building structures, resulting in fewer false alarms on land area by using OSM-SSD. In contrast, the OSD-SSD can help identify and locate ships on sea area because of more ships' annotation information from the aerial image dataset. Considering these complementary advantages, we employ the weighted boxes fusion (WBF) strategy [31] to fuse the predictions of OSD-SSD and OSM-SSD. The WBF strategy utilizes confidence scores of all predicted bounding boxes to construct the averaged boxes, including confidence

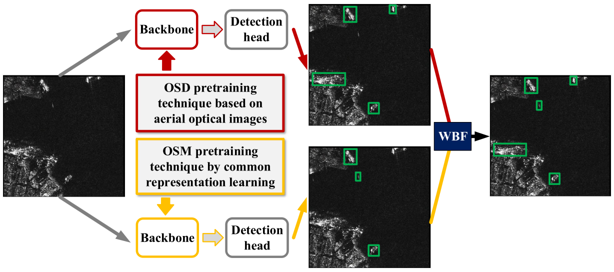
Backbone
Detection head
OSD pretraining technique based on aerial optical images
OSM pretraining technique by common representation learning
Backbone
Detection head
WBF

Fig. 2. Overall process to boost ship detection in SAR images.

scores and coordinate locations, leading to a SSD with better generalization ability.

According to the above analysis, instead of designing sophisticated network structures and using specific tricks, we consider the incompatibility between ImageNet and SAR images and propose two complementary pretraining techniques to boost SAR ship detection. The overall process is depicted in Fig. 2. Specifically, we utilize the OSD pretraining technique and OSM pretraining technique to obtain OSD-SSD and OSM-SSD, respectively. Finally, we leverage the WBF strategy to fuse the predictions of the two complementary detectors. The main contributions of our work can be summarized into the following four aspects:

1) Improve the feature learning ability of ships in SAR images by proposing the OSD pretraining technique, which transfers the characteristics of ships in earth observations to SAR images by fully taking advantage of ships' annotation information from a large-scale aerial image dataset [29].
2) Enhance the general feature embedding of SAR images by proposing the OSM pretraining technique, which transfers plentiful texture features from optical images to SAR images by common representation learning via BNN [30] on the OSM task.
3) Explore the complementary characteristics of the OSD pretraining-based SSD (OSD-SSD) and OSM pretraining-based SSD (OSM-SSD) and thus propose to employ the WBF strategy [31] to fuse the predictions of the two detectors for further improving detection results.
4) Conduct various experiments on four SAR ship detection datasets [32]–[35] and three representative CNN-based detection benchmarks [7], [36] to verify the effectiveness and complementarity of the OSD-SSD and OSM-SSD detectors, and the state-of-the-art performance of the combination of the two detectors.

The rest of this article is organized as follows. Section II introduces our methods in detail. Section III provides the experimental settings and results analysis. Finally, Section IV concludes this article.

II. Methodology

In this section, we first introduce two CNN-based object detection benchmarks: Faster R-CNN [7] and YOLOv3 [36]. Next, we will describe the OSD pretraining technique based on aerial images and the OSM pretraining technique by common representation learning in detail. Finally, we will introduce how the WBF strategy [31] fuse the predictions of OSM-SSD and OSD-SSD.

A. CNN-Based Object Detection Methods

We select two representative methods: Faster R-CNN [7] and YOLOv3 [36], as the benchmarks in our work. Here, we only introduce the critical architecture of these two detection benchmarks, and we refer to their original paper [7], [36] to see a more detailed introduction. It is noted that our OSD pretraining technique, OSM pretraining technique, and the WBF strategy [31] can be easily applied to other state-of-the-art SAR ship detection benchmarks.

Faster R-CNN [7], as the behalf of the two-stage object detection methods, consists of three modules: feature embedding network as the backbone to extract high-level features from the original images, region proposal network (RPN) generating the ship proposals for preliminarily predicting the location of ships and box regression network such as Fast R-CNN [37] finishing the binary classification and bounding box regression. We adopt ResNet50 [13] as the backbone for better feature extraction. Besides, to handle the detection of multiscale ships in multiresolution SAR images, we use the FPN [20] to combine low-resolution, semantic strongly features with high-resolution, semantic weakly features. We initially set three anchor boxes with one scale of size 8 and three aspect ratios of size $\{0.5, 1, 2.0\}$ at each spatial location of each feature map. After RPN generates ship proposals, we adopt the RoIAlign [8] operation to fix the misalignment of feature maps caused by coarse spatial quantization. Furthermore, we select the cross entropy and $smooth$ L1 loss function to optimize the classification and regression task, respectively.

YOLOv3 [36] is a representative one-stage object detection method comprising two modules: feature extraction network and box detection network. The feature extraction network is set as Darknet53 [36] containing 53 convolutional layers and some shortcut connections as used in ResNet [13], making YOLOv3 more powerful and efficient. As similarly used in Faster R-CNN, FPN is adopted in the feature extraction network to enhance multiscale object detection capabilities. The box detection network predicts bounding boxes on three different scales following the feature extraction network. Feature maps with small sizes are used to detect large-size ships, while feature maps with large sizes are utilized to detect small-size ships. The predefined anchor boxes in Faster R-CNN are leveraged because learning the offsets between anchor boxes and the predictions will make the network easier to train. Considering this convenience, YOLOv3 also presets nine anchors with three different sizes on each scale to predict bounding boxes more accurately. The anchor sizes are set according to the dataset. We use the same initial anchor sizes as those in [36]. Moreover, instead of fixing the input image size, we use the multiscale training strategy to resize the input image into one size of $\{320 \times 320, 416 \times 416\}$ randomly in each iteration, forcing the network to perform prediction well across different resolutions.

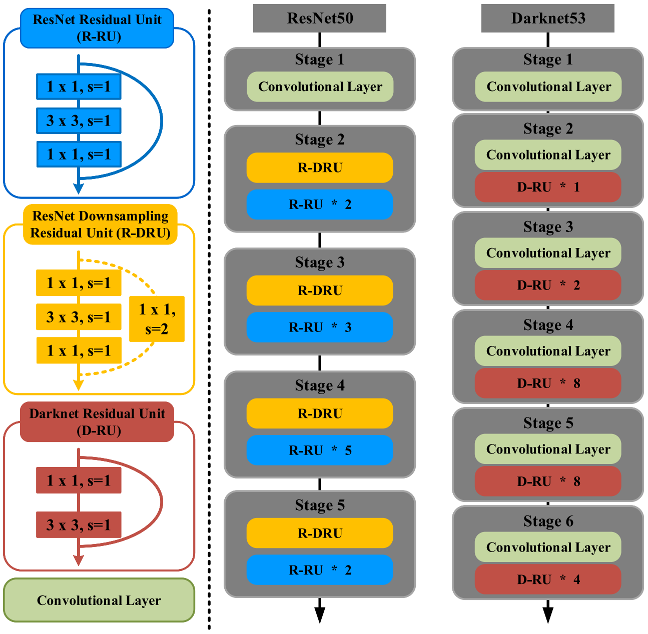

Fig. 3. Illustration of simplified structures of ResNet50 and Darknet53.

Cross entropy and mean square error loss functions are utilized to optimize this network end-to-end. Fig 3 illustrates the simplified structures of ResNet50 and Darknet53.

B. OSD Pretraining Technique Based on Aerial Images

The OSD pretraining technique transfers the characteristics of ships in earth observations to SAR images for improving the feature learning ability of ships in SAR images. Specifically, we first train OSDs in aerial optical images and then leverage the backbone, ResNet50 [13] for Faster R-CNN [7] and Darknet53 [36] for YOLOv3 [36], to further obtain OSD-SSD for fully taking advantage of ships' annotation information. As for object detection methods in aerial images, recent advances have been witnessed because of the construction of many well-annotated datasets [29], [38] and efficient network design for specific problems [39]. Because a large number of high-resolution ships are available, we select DOTA [29], a large-scale dataset for object detection in aerial images, as the basic dataset to train our OSDs. For the ResNet50 backbone, we directly extract it from the trained model provided by the region of interest (RoI) transformer method [39]. RoI transformer applied an RRoI learner on the RoIs to learn spatial transformation from horizontal proposals to oriented bounding box predictions, expecting to solve the common mismatch between horizontal box predictions and oriented objects. This transformation on feature maps enables the ResNet50 backbone more powerful feature extraction capabilities, especially for ships from an overlooking perspective. For extracting the Darknet53 backbone, we adopt the original YOLOv3 method to train the OSD without any architectures modified. In addition to the ResNet50 and Darknet53 backbone, we can also extract FPN [20] from the trained models using RoI Transformer and YOLOv3. We use ResNet50 (Darknet53) to denote ResNet50+FPN (Darknet53+FPN) for convenience.

C. OSM Pretraining Technique by Common Representation Learning

The OSM pretraining technique transfers rich texture features from optical images to SAR images to obtain a specific feature extraction model with better SAR feature embedding capabilities. Specifically, the OSM pretraining technique resorts to a SAR feature embedding operator from common representation learning based on the OSM task. As for common representation learning, several researches [30], [40], [41] were proposed to investigate the relationships between two data sources in different modalities. Among these methods, the BNN proposed in [30] is adopted to perform our common representation learning on the matching problem of optical and SAR images due to its excellent performance. As depicted in Fig 4, BNN acts as a bridge and projects two images from different modalities into a common feature subspace. More specifically, given the OSM task, we should first construct the data pairs of optical images and SAR images. Supposed we have two data sources denoted by $\{X_s, X_o\} \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, where $X_s = \{x_s^i\}_{i=1}^N$ is from SAR images and $X_o = \{x_o^i\}_{i=1}^N$ is from optical images. The ith component $x_s^i \in X_s, x_o^i \in X_o$ are from the same region and match each other, while the ith component $x_s^i \in X_s$ and the jth component $x_o^j \in X_o$ are from different regions and do not match each other. Moreover, we define $D_p = \{x_s^i, x_o^i\}$ as positive samples and $D_n = \{x_s^i, x_o^j\}, i \neq j$ as negative samples. Second, instead of sharing weights, we build the BNN architecture illustrated in Fig. 4, which contains two separate, yet identical CNN: SAR CNN $f_s\{\cdot;\theta_s\}$ and Opt CNN $f_o\{\cdot;\theta_2\}$ with weights (θ_s, θ_o). For a pair of optical image and SAR image (x_s, x_o), the output of SAR CNN and Opt CNN is $f_s(x_s;\theta_s)$ and $f_o(x_o;\theta_o)$, respectively. The CNN can be replaced with ResNet50 [13] and Darknet53 [36] for the Faster R-CNN [7] and YOLOv3 [36] benchmark, respectively. To decrease the feature dimension, we adopt a convolution layer with the filter size of 1×1, a batch normalization layer and a max-pooling layer following the backbone of the detection benchmark to output an m-dimensional feature map. Finally, we add a linear layer followed by the sigmoid activation function to project the m-dimensional feature map into the common feature subspace. The BNN outputs the Euclidean distance of the two outputs of SAR CNN and Opt CNN, which is defined as follows:

$$f\left(x_s, x_o; \theta_s, \theta_o\right) = \frac{1}{\sqrt{n}} \left\|\left(f_s\left(x_s;\theta_s\right) - f_o\left(x_o;\theta_o\right)\right)\right\| \quad (1)$$

where n is the dimension of the common feature. The ultimate goal of BNN is to determine whether the optical and SAR images have a potential relationship, i.e., positive sample or not. Thus, the loss on positive sample set D_p and negative sample set D_n regresses the output $f(x_s, x_o; \theta_s, \theta_o)$ to 0 if $(x_s, x_o) \in D_P$ and to 1 if $(x_s, x_o) \in D_n$, respectively, as follows:

$$l_p\left(D_p;\theta_s,\theta_o\right) = \frac{1}{|D_p|} \sum_{(x_s,x_o)\in D_p} \left(f\left(x_s,x_o;\theta_s,\theta_o\right) - 0\right)^2 \quad (2)$$

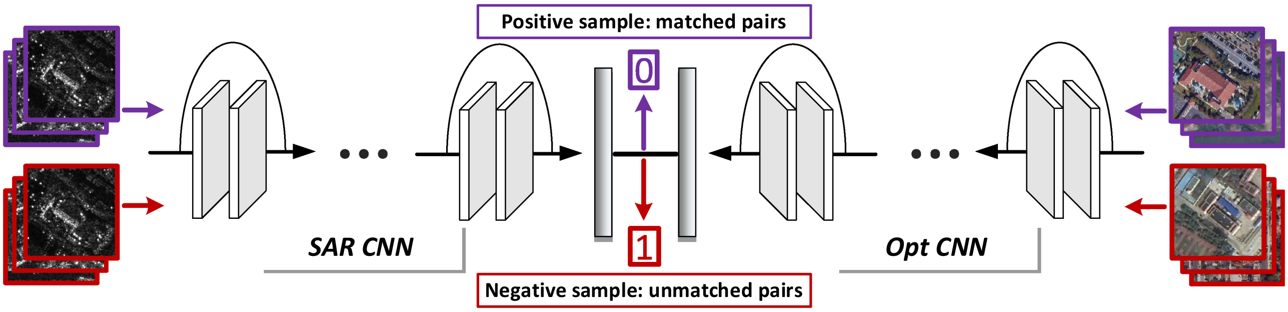
Positive sample: matched pairs
0
1
SAR CNN
Opt CNN
Negative sample: unmatched pairs

Fig. 4. Illustration of common representation learning via BNN [30] on the OSM task, which adopts two CNNs named SAR CNN and Opt CNN to project SAR images and optical images into a common feature space. The Euclidean distance of the two output layers is regressed to 0 or 1 for positive samples or negative samples, respectively.

$$l_n\left(D_n;\theta_s,\theta_o\right) = \frac{1}{|D_n|}\sum_{(x_s,x_o)\in D_n}\left(f\left(x_s,x_o;\theta_s,\theta_o\right)-1\right)^2. \tag{3}$$

Thus, the problem of searching the common feature embeddings of optical and SAR images can be transferred to a binary classification problem, where the overall loss of BNN on D_p and D_n can be designed as follows:

$$L_{\mathrm{bnn}}\left(D_p,D_n;\theta_s,\theta_o\right) = \frac{l_p\left(D_p;\theta_s,\theta_o\right)+\alpha\cdot l_n\left(D_n;\theta_s,\theta_o\right)}{1+\alpha} \tag{4}$$

where α is a hyperparameter to adjust the balance of positive samples and negative samples. Then, the BNN model with the best weights (θ_s^*,θ_o^*) can be find via

$$(\theta_s^*,\theta_o^*) = \mathrm{argmin}_{\theta_s,\theta_o}\, l\left(D_p,D_n;\theta_s,\theta_o\right) \tag{5}$$

In the test phase, BNN uses a predefined threshold parameter γ to decide whether the input data pair is a positive sample or not. We train the BNN model on the QXS-SAROPT [42] dataset. Finally, we can exploit the SAR-CNN as the backbone of our OSM-SSD, expecting that the plentiful texture features from the optical image would enhance the feature embedding ability for SAR images.

D. Weighted Boxes Fusion

We expect to leverage an effective fusion strategy to enhance detection performance based on the observation of the complementary advantages between OSD-SSD and OSM-SSD. A commonly used fusion strategy is nonmaximum suppression (NMS). NMS sorts all the predicted bounding boxes according to classification confidence and removes redundant boxes to keep the one with the highest confidence score. However, NMS pays too much attention to classification confidence without considering localization accuracy. For example, if a predicted box has a high intersection over union (IoU) with ground truth but a low confidence score, it will be removed by another predicted box with high confidence score but a low IoU. Although soft-NMS [43] can alleviate this problem, it will retain largely redundant boxes, increasing many false alarms. To simultaneously take classification confidence and localization accuracy into account, we adopt the WBF strategy [31] to combine the predictions of OSM-SSD and OSD-SSD, expecting to improve the generalization ability of SSDs. The WBF strategy in our method utilizes confidence scores of predicted bounding boxes to construct the averaged boxes, including not only averaged confidence scores, but also averaged localization predictions as depicted in Fig 5. Specifically, a predicted box with higher classification confidence will more proportionally contribute to the final averaged box. Algorithm 1 illustrates the process of WBF strategy.

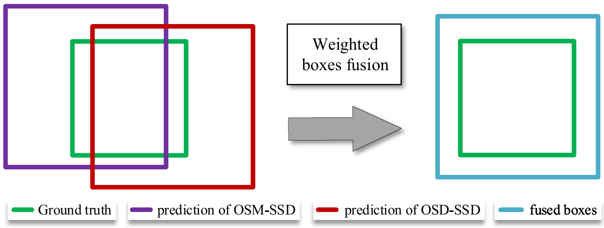

Fig. 5. Illustration of the WBF strategy [31].

III. Experiments

A. Datasets

We conduct experiments on three kinds of datasets: DOTA used for OSD pretraining, QXS-SAROPT used for OSM pretraining, and four SAR ship detection datasets used for OSD-SSD and OSM-SSD. It is noted that the OSD model based on the Faster R-CNN [7] benchmark can be directly obtained from [39], so we only perform OSD pretraining using the YOLOv3 [36] benchmark on DOTA. Table I shows the detailed information of all datasets.

DOTA [29] is a large-scale dataset for object detection in aerial images including 2806 aerial images with each image ranging from 800 × 800 to 4000 × 4000 pixels collected from different sensors and platforms. Compared to version 1.0 with 15 categories annotated, the fully annotated DOTA version 1.5 (DOTA-v1.5) contains 16 common object categories, including the ship. We used DOTA-v1.5 and adopted the same dataset split strategy as used in [39]: selecting half of the original images as the training set, 1/6 as the validation set, and 1/3 as the testing set. Each image is cropped into 1024 × 1024 pixels with an overlap

Algorithm 1: Weighted boxes fusion strategy. B^{osd} and B^{osm} are predicted boxes of OSD-SSD and OSM-SSD, respectively. N is the number of models. thr is the IoU threshold. Each predicted box is represented as $(Xmin, Ymin, Xmax, Ymax, C)$.

```
Input: $N, B^{osd}, B^{osm}, thr$.
Output: Fused boxes of OSD-SSD and OSM-SSD.
1: $B^{osd}$ and $B^{osm}$ are added to a list L and sorted in
   decreasing order according to the confidence score $C$
2: Declare an empty list E with each position storing a
   set of boxes (or a single box) for box clusters. Declare
   an empty list F with each corresponding position
   containing only one box to represent the fused box.
3: Loop through all predicted boxes in L, and attempt to
   find a matching box in the list F. If the IoU between a
   box in F and the current box in L is larger than $thr$, we
   define this box as a matching box ($IoU > thr$).
4: If the matching box is not found in step 3, add the
   current box from the list L to the end of lists E and F
   as new entries; proceed to the next box in the list L.
5: If the matching box is found in step 3, add the current
   box from the list L to E and the added position is the
   corresponding position of the matching box in F.
6: Leveraging all $T$ boxes from the same position in E to
   recalculate the box coordinates and confidence score to
   form the fused box in F with the following equation:
       $C = \frac{\sum_{i=1}^{T} C_i}{T}$   (6)
       $Xmin, max = \frac{\sum_{i=1}^{T} C_i \times Xmin_i, max_i}{\sum_{i=1}^{T} C_i}$   (7)
       $Ymin, max = \frac{\sum_{i=1}^{T} C_i \times Ymin_i, max_i}{\sum_{i=1}^{T} C_i}$   (8)
   where a box with larger confidence contributes more
   to the coordinates of the fused box than a box with
   lower confidence.
7: After all boxes in the list L are traversed, adjust the
   confidence scores in F again with the following
   equation:
       $C = C \times \frac{T}{N}$   (9)
```

of 256 pixels. We directly trained the OSDs on the training set and directly report the performance of detectors on the validation set because the annotation information of the testing set is not available.

QXS-SAROPT [42] includes 20000 pairs of corresponding SAR and optical images extracted from GaoFen-3 high-resolution spotlight images and Google Earth remote sensing optical images. The size of each image is 256×256. We select 14000 image pairs as the training set and the remaining 6000 image pairs as the testing set to train our BNN.

SAR Ship Detection Datasets. We conduct SAR ship detection experiments on four datasets: AIR-SARShip-1.0 [32], AIR-SARShip-2.0 [33], which is also the Gaofen competition dataset provided by the "2020 Gaofen Challenge on automated high-resolution earth observation image interpretation," HRSID [34] and SSDD [35]. The AIR-SARShip-1.0 dataset comprises 31 high-resolution large-scale 3000×3000 images collected from the GF-3 satellite. We randomly select 21 images as the training and validation data, and the remaining 10 images as the testing data. The AIR-SARShip-2.0 dataset includes 300 images of size 1000×1000 with spatial resolution ranging from 1 to 5 m collected from the Gaofen-3 satellite. We randomly select 210 images as training and validation data, and the remaining 90 images are used for testing data. Each image of both AIR-SARShip-1.0 and AIR-SARShip-2.0 is cropped into 512×512 pixels with an overlap of 256 pixels for training and testing. The HRSID dataset contains 5604 cropped SAR images with the size of 800×800 and has been divided into a training set and a test set at a ratio of 65 to 35. The SSDD dataset contains 1160 images of resolution from 1 to 15 m in total, where the training set includes 928 images and the testing set includes the remaining 232 images. For the Faster R-CNN benchmark, we directly input each image of AIR-SARShip-1.0, AIR-SARShip-2.0, and HRSID into the network, and resize each image of SSDD into 1000×600 pixels. As for the YOLOv3 benchmark, all the images of the four datasets are resized into 416×416 pixels. We do not apply any data augmentation method for all datasets except for the scaling technique because of the multiscaling training strategy for the YOLOv3 benchmark.

B. Parameter Settings

All the experiments are implemented in the PyTorch 1.7 framework and carried out over an NVIDIA 3070 GPU. The PC operating system is a 64-bit Ubuntu 20.04.

1) OSD Pretraining Technique: For the Faster R-CNN [7] benchmark, we directly extract the ResNet50 [13] backbone from the existing trained model from [39]. As for the YOLOv3 [36] benchmark, we first use MMDetection[1] to train an OSD and then extract the Darknet53 [36] backbone from the trained detector. The OSD is trained with stochastic gradient descent (SGD) for 240 epochs with a total of 12 images per minibatch. The initial learning rate is set as 0.001, which is then divided by a factor of 10 at the 160th and 200th epoch. The weight decay is 0.0005 and the momentum is 0.9. Both the ImageNet-based pretrained ResNet50 and Darknet53 are utilized for a better converge point.

2) OSM Pretraining Technique: Our BNN models based on ResNet50 [13] and Darknet53 [36] are both trained with SGD for 200 epochs with a batch size of 20. The initial learning rate is set as 0.01 and then divided by a factor of 2 at the 30th and 100th epochs. The dimension of the common feature is 50 and m is set as 128. The adjusting factor α is 1 and the threshold γ is set as 0.5. We also adopt ImageNet-based pretrained model to train BNN.

[1] [Online]. Available: https://github.com/open-mmlab/mm-detection

TABLE I
DETAILED INFORMATION OF ALL USED DATASETS

Datasets	Task type	Image numbers		Input image size	
		Training set	Testing set	Faster R-CNN [7]	YOLOv3 [36]
DOTA [29]	Object detection in aerial optical images	11119	3626	1024 × 1024	1024 × 1024
QXS-SAROPT [42]	Optical-SAR matching	14000 pairs	6000 pairs	256 × 256	256 × 256
AIR-SARShip-1.0 [32]	SAR ship detection	637	273	512 × 512	416 × 416
AIR-SARShip-2.0 [33]	SAR ship detection	1248	536	512 × 512	416 × 416
HRSID [34]	SAR ship detection	3642	1962	800 × 800	416 × 416
SSDD [35]	SAR ship detection	928	232	1000 × 600	416 × 416

3) SSDs: We also use MMDetection to implement OSD-SSD and OSM-SSD. For the Faster R-CNN [7] benchmark, all models are trained with SGD for 14 epochs with a total of eight images perminibatch. The initial learning rate is set as 0.02 and then divided by a factor of 10 at the 8th and 12th epochs. The weight decay is 0.0001 and the momentum is 0.9. For the YOLOv3 [36] benchmark, all models are trained with SGD for 240 epochs with a total of 12 images per minibatch. The initial learning rate is set as 0.001, which is then divided by a factor of 10 at the 160th and 200th epochs. The IoU threshold is set as 0.5 when training and testing for rigorous filtering the bounding boxes with low precision. Warm-up [13] is introduced during the initial training stage to avoid gradient explosion and the corresponding number of iterations is 500. We use the same settings for all experiments for a fair comparison.

4) WBF: We fuse predicted boxes from OSD-SSD and OSM-SSD through the WBF strategy [31]. The IoU threshold thr is set as 0.7 verified by many experiments and please see Section III-D4 for a detailed discussion.

C. Evaluation Metrics

Precision, recall, and F1 scores are employed to evaluate the performance of SSDs, and the definition of these evaluation metrics is given as follows:

$$\text{Precision} = \frac{\text{N}_{\text{TP}}}{\text{N}_{\text{TP}} + \text{N}_{\text{FP}}} \tag{10}$$

$$\text{Recall} = \frac{\text{N}_{\text{TP}}}{\text{N}_{\text{TP}} + \text{N}_{\text{FN}}} \tag{11}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

where TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative. N_{TP}, N_{FP}, N_{FN} is the number of TP, FP, and FN, respectively. More specifically, TP indicates the correctly detected ships, FP represents the false alarms and FN denotes the missing ships. A predicted bounding box is considered as a true positive if its IoU with the ground truth is higher than a threshold, such as 0.5. Otherwise, it is regarded as a false positive. Moreover, the predicted bounding box with the highest confidence score is seen as the true positive, if the IoU of several ones with the ground truth are all higher than the threshold. F1 score is a comprehensive evaluation metric for the quantitative performance of different models by simultaneously considering the precision rate and recall rate. To further evaluate the comprehensive quality of SSDs, we also adopt the average precision (AP_i) metrics, which can be defined as follows:

TABLE II
DETECTION RESULTS OF OSDS ON THE DOTA DATASET [29]

Benchmark	Ship			All categories		
	$\text{AP}_{0.5}$	$\text{AP}_{0.75}$	AP	$\text{mAP}_{0.5}$	$\text{mAP}_{0.75}$	mAP
Faster R-CNN [7]	0.807	–	–	0.650	–	–
YOLOv3 [36]	0.605	0.477	0.470	0.509	0.284	0.287

$$\text{AP}_i = \frac{1}{101} \sum_{r \in S} \text{Precision}_i \,|_{\text{Recall}_i = r} \tag{13}$$

where $S = \{0, 0.01, \ldots, 1\}$ representing a set of equally spaced recall rates and i indicating the IoU threshold. $\text{AP}_{0.5}$ denotes AP_i with the IoU threshold being 0.5. We use Precision to represent $\text{Precision}_{0.5}$ for convenience (the same as Recall). To evaluate the localization performance more accurately, we also adopt $\text{AP}_{0.75}$ and AP metrics. AP indicates the averaged AP_i where i is set from 0.50 to 0.95 with the step size set as 0.05, which can be defined as follows:

$$\text{AP} = \frac{1}{10} \sum_{i \in I} \text{AP}_i \tag{14}$$

where $I = \{0.5, 0.55, \ldots, 0.95\}$ representing a set of equally spaced IoU threshold.

D. Results Analysis

1) OSD-SSD: Table II shows the detection results of our OSDs on DOTA dataset [29] including the mean AP (mAP) for all categories and AP for the ship category. It is noted that the results of the Faster R-CNN [7] benchmark are directly from [39] ("-" means that the original paper did not provide the corresponding results). We can see that $\text{AP}_{0.5}$ achieves 80.7% and 60.5% for the Faster R-CNN and YOLOv3 [36] benchmark, respectively, indicating that ResNet50 [13] and Darknet53 [36] backbone both have excellent ship detection capabilities. As for SSDs, the detection performance is showed in Tables III and IV for the Faster R-CNN and YOLOv3 benchmark, respectively. We can observe that OSD-SSD outperforms the ImageNet pretraining-based SSD (ImageNet-SSD) under different metrics. Take the detection results on AIR-SARShip-2.0 [33] using the YOLOv3 benchmark for an example, the precision rate of the OSD-SSD gains a large improvement of 3.53% and the recall rate achieves 2.12% higher value. Due to these improvements,

TABLE III
OVERALL PERFORMANCE OF DIFFERENT METHODS ON FOUR SAR SHIP DETECTION DATASETS USING THE FASTER R-CNN [7] BENCHMARK

AIR-SARShip-1.0 [32]						
Method	Precision	Recall	F1	$AP_{0.5}$	$AP_{0.75}$	AP
ImageNet-SSD	0.9232	0.8003	0.8574	0.8720	0.5461	0.5129
OSD-SSD	**0.9436**	0.8529	**0.8960**	0.8921	0.6260	0.5605
OSM-SSD	0.9324	0.8248	0.8753	0.8852	0.6153	0.5586
WBF-DM	0.9045	**0.8653**	0.8847	**0.9006**	**0.6452**	**0.5801**
AIR-SARShip-2.0 [33]						
ImageNet-SSD	0.8561	0.7964	0.8252	0.8487	0.5633	0.5190
OSD-SSD	0.8563	0.8232	0.8391	0.8626	0.6028	0.5618
OSM-SSD	**0.8676**	0.8170	0.8415	0.8582	0.6129	0.5528
WBF-DM	0.7944	**0.8482**	**0.8677**	**0.8736**	**0.6451**	**0.5880**
HRSID [34]						
ImageNet-SSD	0.8842	0.8624	0.8736	0.8878	0.7851	0.6703
OSD-SSD	**0.8975**	0.8617	**0.8784**	0.8892	0.7851	0.6743
OSM-SSD	0.8826	0.8658	0.8751	0.8932	0.7890	0.6719
WBF-DM	0.8455	**0.8823**	0.8736	**0.8971**	**0.7988**	**0.6844**
SSDD [35]						
ImageNet-SSD	0.9332	0.9440	0.9385	0.9624	0.7292	0.6231
OSD-SSD	0.9417	0.9558	**0.9487**	0.9704	0.7485	0.6328
OSM-SSD	**0.9429**	0.9485	0.9457	0.9679	0.7427	0.6228
WBF-DM	0.9079	**0.9669**	0.9364	**0.9740**	**0.7534**	**0.6426**

TABLE IV
OVERALL PERFORMANCE OF DIFFERENT METHODS ON FOUR SAR SHIP DETECTION DATASETS USING THE YOLOV3 [36] BENCHMARK

AIR-SARShip-1.0 [32]						
Method	Precision	Recall	F1	$AP_{0.5}$	$AP_{0.75}$	AP
ImageNet-SSD	0.8901	0.8603	0.8744	0.8712	0.5546	0.5024
OSD-SSD	**0.9411**	0.8740	**0.9062**	0.8849	0.6273	0.5398
OSM-SSD	0.9324	0.8797	0.9055	0.8836	0.6146	0.5375
WBF-DM	0.8667	**0.8964**	0.8808	**0.8913**	**0.6529**	**0.5672**
AIR-SARShip-2.0 [33]						
ImageNet-SSD	0.8711	0.7960	0.8323	0.8300	0.4622	0.4649
OSD-SSD	0.9064	0.8172	**0.8593**	0.8449	0.5609	0.5096
OSM-SSD	**0.9091**	0.8150	0.8562	0.8470	0.5509	0.5114
WBF-DM	0.8584	**0.8497**	0.8535	**0.8543**	**0.6156**	**0.5478**
HRSID [34]						
ImageNet-SSD	0.8541	0.8593	0.8567	0.8364	0.5538	0.5229
OSD-SSD	**0.9010**	0.8684	**0.8841**	0.8512	0.6035	0.5496
OSM-SSD	0.8662	0.8658	0.8665	0.8454	0.5731	0.5319
WBF-DM	0.7879	**0.8933**	0.8375	**0.8796**	**0.6581**	**0.5850**
SSDD [35]						
ImageNet-SSD	0.9482	0.9509	0.9493	0.9447	0.6361	0.5840
OSD-SSD	**0.9614**	0.9505	0.9556	0.9474	0.6572	0.5898
OSM-SSD	0.9571	0.9560	**0.9561**	0.9468	0.6759	0.5885
WBF-DM	0.9413	**0.9633**	0.9523	**0.9577**	**0.7017**	**0.6058**

TABLE V
MATCHING RESULTS OF BNN [30] ON THE QXS-SAROPT DATASET

Backbone	Accuracy	Precision	Recall
ResNet50 [13]	0.829	0.748	0.993
Darknet53 [36]	0.828	0.746	0.995

the OSD-SSD finally achieves a 2.7% higher F1 score and 1.70% higher $AP_{0.5}$, reflecting overall performance improvements. Furthermore, when the IoU threshold becomes larger indicating the requirement of localization accuracy gets higher, $AP_{0.75}$ and AP gain a larger improvement of 9.87% and 4.47%, respectively, which means that the predicted bounding boxes are more accurate. Similar phenomena are also presented on other datasets for both the Faster R-CNN and YOLOv3 benchmark, demonstrating the superiority of our OSD-SSD model. In other words, compared to ships in the natural scene from ImageNet, ships' annotation information from earth observations can better improve the feature learning ability of ships in SAR images.

2) OSM-SSD: Table V suggests that our BNN has an outstanding performance on the QXS-SAROPT dataset for both ResNet50 [13] and Darknet53 [36] backbone. Specifically, the pair matching accuracy based on ResNet50 and Darknet53 are 82.9% and 82.8%, respectively, demonstrating that BNN can well predict the relationship of SAR and optical images, and obtain useful common features. As for the performance of OSM-SSD based on the Faster R-CNN [7] and YOLOv3 [36] benchmark, the overall results under different metrics are also illustrated in Tables III and IV, respectively. We can see that our OSM-SSD performs better than ImageNet-SSD under all evaluation metrics for different network architectures and different datasets. Especially on AIR-SARShip-2.0 [33] using the YOLOv3 network, 3.80%, 1.90%, 2.39%, 1.49%, 8.87%, and 4.65% performance improvement can be achieved in terms of precision rate, recall rate, F1 score, $AP_{0.5}$, $AP_{0.75}$, and AP, respectively. All these improved performances can prove that the common features obtained from common representation learning can boost ship detection in SAR images. In other words, BNN can well transfer rich texture features from optical images to SAR images, enhancing the feature extraction capability of SSDs without additional ships' annotation information and any network architecture modified.

3) Comparison Between OSM-SSD and OSD-SSD: For results of AIR-SARShip2.0 based on the Faster R-CNN [7] benchmark in Table III, OSD-SSD and OSM-SSD achieve comparable detection results in different degrees, such as 1.39%, 3.95%, 4.28% improvements verse 0.95%, 4.96%, 3.38% improvements in terms of $AP_{0.5}$, $AP_{0.75}$, and AP, respectively. It is noted that the DOTA dataset [29] on which our OSD pretraining performed has a large amount of annotation information while the QXS-SAROPT dataset on which our OSM pretraining performed only has region matching information without any ship annotation. Moreover, we also conduct experiments under the inshore and offshore scenes, respectively. We can observe from Table VI that OSD-SSD and OSM-SSD have different performances in different scenes. Specifically, compared to OSD-SSD, our OSM-SSD gains 1.50% improvements in precision rate under inshore scenes. As for the offshore ship detection, our OSD-SSD improves the recall rate by 0.25% than OSM-SSD. In addition to quantitative comparisons, we also visualize some detection

TABLE VI
PERFORMANCE OF IMAGENET-SSD, OSD-SSD AND OSM-SSD UNDER TWO DIFFERENT SCENES

Scene	Method	Precision	Recall	F1	$AP_{0.5}$	$AP_{0.75}$	AP
Inshore	ImageNet-SSD	0.8484	0.8139	0.8307	0.8777	0.5336	0.4941
	OSD-SSD	0.8655	**0.8604**	**0.8629**	**0.8973**	**0.5982**	**0.5332**
	OSM-SSD	**0.8805**	0.8197	0.8491	0.8876	0.5789	0.5112
Offshore	ImageNet-SSD	0.9640	0.9919	0.9777	0.9879	0.7861	0.6605
	OSD-SSD	0.9634	**0.9946**	**0.9788**	0.9881	**0.8203**	**0.6705**
	OSM-SSD	**0.9659**	0.9919	0.9787	**0.9885**	0.8074	0.6663

TABLE VII
RESULTS OF WBF-DM UNDER DIFFERENT IOU THRESHOLDS

IoU threshold	$AP_{0.5}$	$AP_{0.75}$	AP
0.4	0.8533	0.6022	0.5400
0.5	**0.8591**	0.6086	0.5461
0.6	0.8581	0.6141	0.5476
0.7	0.8543	0.6156	**0.5478**
0.75	0.8502	**0.6188**	0.5466
0.8	0.8434	0.6136	0.5424

results in Fig. 6 to show an intuitive understanding of their effect. We use red rectangles, green rectangles, orange circles, and yellow circles to figure out ground truth, predicted boxes, false alarms, and missing ships, respectively. In Fig. 6(a), considering the first scene on sea area, two ground-truths are perfectly detected using OSD-SSD while one ship missed in the prediction of OSM-SSD. Similar phenomena that OSD-SSD has a better recall on sea area also occurs in other scenes. We conjecture that this observation is due to the OSD model can better capture ship features as large amounts of ship annotation in the aerial optical dataset. On the contrary, we can see that OSM-SSD has better performance than OSD-SSD on land area from Fig. 6(b). Taking the first row as an example, compared to one false alarm on land area for OSD-SSD, OSM-SSD accurately predicts the ground truth without any false alarms. What causes this phenomenon is that the OSM dataset on which our OSM pretraining performed mainly focuses on land area, leading to OSM-SSD has better feature embedding and fewer false alarms on land area. In conclusion, OSD-SSD and OSM-SSD can improve detection results from different perspectives, proving the effectiveness of our OSD and OSM pretraining techniques. More importantly, these complementary advantages between OSD-SSD and OSM-SSD straightly inspire us to utilize fusion strategies to boost ship detection results, which will be analyzed in the following part.

4) Wbf-Dm: We use WBF-DM to represent the fusion between predictions of OSD-SSD and OSM-SSD. The final results of WBF-DM on four SAR ship detection results are depicted in Tables III and IV. Taking the result on AIR-SARShip 2.0 [33] using the YOLOv3 [36] benchmark as an example, WBF-DM reaches 2.12% and 1.9% higher recall rate but 4.8% and 4.97% lower precision rates compared to OSD-SSD and OSM-SSD, respectively. Is is a natural phenomenon because WBF-DM inevitably produces false alarms when it preserves as many ships as possible to improve the recall rate. Consequently, $AP_{0.5}$, $AP_{0.75}$, and AP gains a large improvement with a slight decrease in F1 score, which demonstrates that WBF-DM can fully take advantage of complementary characteristics between OSD-SSD and OSM-SSD. PR curves of OSD-SSD, OSM-SSD, and WBF-DM compared to ImageNet-SSD under $AP_{0.5}$ and $AP_{0.75}$ metrics are depicted in Figs. 7 and 8, respectively. We can observe that the orange curve corresponding to OSD-SSD is always above the red curve corresponding to ImageNet-SSD. Similarly, the blue PR curve denoting OSM-SSD is also above the red PR curve. Although the green curve corresponding to WBF-DM is slightly lower than the orange or red curve at some points in recall rate, the maximum point in recall rate of the green curve is extremely larger than that of the orange and red curve. Subsequently, the area under the green PR curve increased substantially. This phenomenon also illustrates that WBF is a strategy sacrificing the precision rate for improving the recall rate to achieve a significant increase in AP. Finally, WBF-DM yields a considerable increase in terms of a series of AP metrics compared to ImageNet-SSD, especially 15.34% improvements in $AP_{0.75}$. All of our experiments manifest positive effects of WBF-DM on improving the comprehensive quality of SAR ship detection.

The IoU threshold thr is a significant hyperparameter in WBF-DM to determine the final fusion results to a large extent. In order to analyze the influence of thr, we implement a comparison on AIR-SARShip-2.0 [33] using the YOLOv3 [36] benchmark with all settings remaining identical except for the value of thr in Table VII. We can observe that when thr turns to be 0.5, $AP_{0.5}$ achieves the highest result of 0.8591. However, as thr gradually increases, $AP_{0.5}$ gradually decreases and $AP_{0.75}$ achieves a maximum value of 0.6188 when thr becomes 0.75. It is due to the thr of WBF strategy plays NMS's role and expects to match the requirements of localization accuracy. Hence, we can adaptively change the threshold according to different task requirements. To achieve a comprehensive performance improvement, we set thr as 0.7 in our experiments.

In addition to fusing the predictions of OSD-SSD and OSM-SSD, we also conduct experiments to fuse detection results of different combinations among ImageNet-SSD, OSD-SSD, and OSM-SSD. Table VIII shows the fusion results. We can see that, regardless of any combination, the performance of fusing multiple models is always superior to that of a single model, demonstrating the effectiveness of the WBF strategy. Furthermore, the result of WBF-DM not only outperforms the fusion result between ImageNet-SSD and OSD-SSD, but also

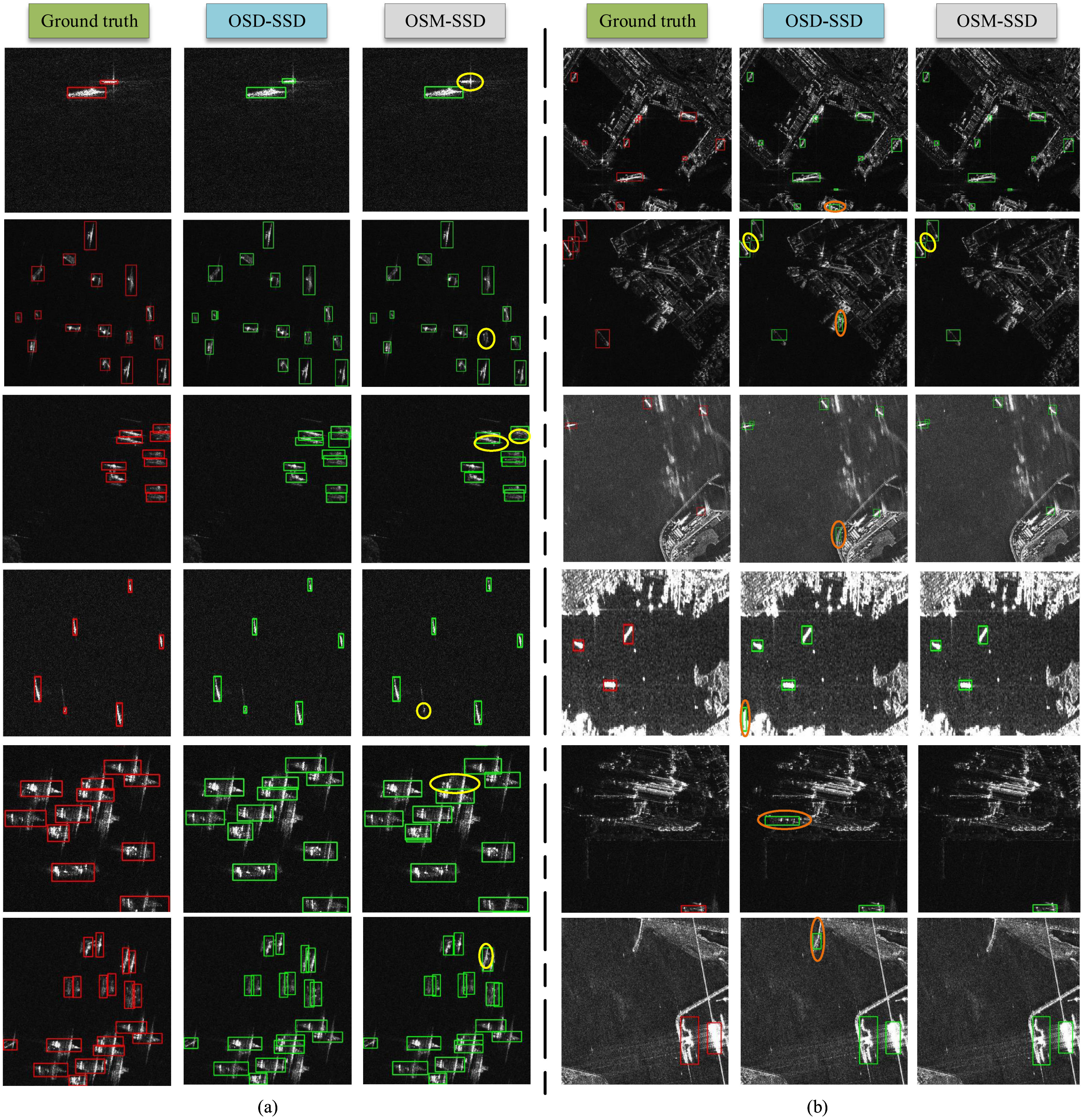

Fig. 6. Comparison results of OSD-SSD and OSM-SSD on four SAR ship detection datasets. The red rectangles are ground truth and the green rectangles represent the detection results. The orange and yellow circles denote the false alarms and missing ships, respectively. Two sets of visualization results demonstrate the complementary characteristics of OSD-SSD and OSM-SSD: (a) OSD-SSD outperforms OSM-SSD on sea area. (b) OSM-SSD outperforms OSD-SSD on land area.

outperforms the fusion result between ImageNet-SSD and OSM-SSD. Compared to ImageNet-SSD fused with OSD-SSD, the better improvements contribute to the fewer false alarms on land area of OSM-SSD. Compared to ImageNet-SSD combined with OSM-SSD, the improved detection results are due to the higher recall on sea area of OSD-SSD. It is the complementary advantage under different scenes that allows the WBF-DM to improve the ship detection performance. Finally, when we combine the predictions of ImageNet-SSD, OSD-SSD, and OSM-SSD, $\mathrm{AP}_{0.5}$ achieves the highest value of 86.05% while $\mathrm{AP}_{0.75}$ and AP become lower. In other words, the improvements vanish as the requirement of localization accuracy gets higher. We conjecture that instead of complementary advantages in different scenes, the increased $\mathrm{AP}_{0.5}$ is because the WBF strategy adjusts not only the confidence of the predicted boxes, but also the position, reaching a better localization accuracy.

NMS is also a commonly used fusion strategy to filter the bounding boxes with low precision, especially for densely

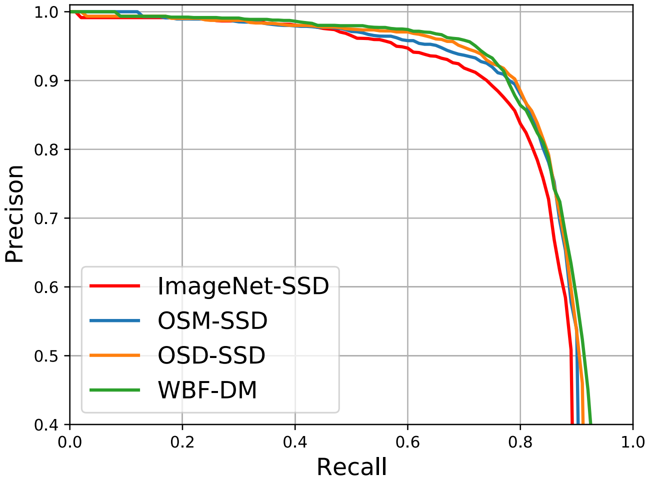

Fig. 7. PR curves of four methods: ImageNet-SSD, OSM-SSD, OSD-SSD and WBF-DM on AIR-SARShip-2.0 [33] using the YOLOv3 [36] benchmark under $\mathrm{AP}_{0.5}$ metric.

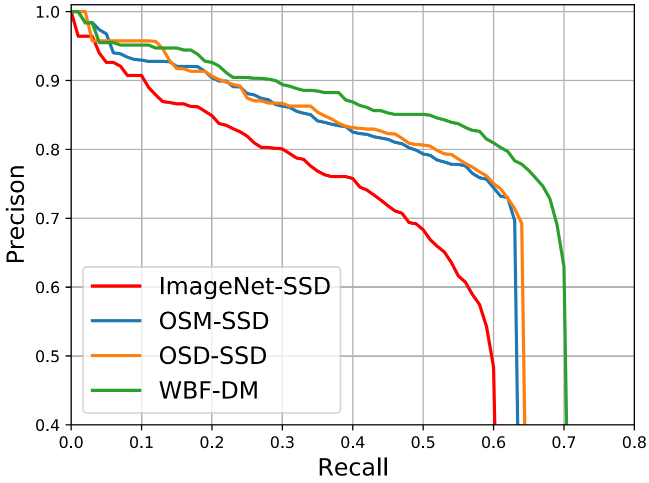

Fig. 8. PR curves of four methods: ImageNet-SSD, OSM-SSD, OSD-SSD and WBF-DM on AIR-SARShip-2.0 [33] using the YOLOv3 [36] benchmark under $\mathrm{AP}_{0.75}$ metric.

TABLE VIII
FUSION RESULTS OF PREDICTIONS OF IMAGENET-SSD, OSM-SSD, AND OSD-SSD UNDER DIFFERENT COMBINATIONS

ImageNet-SSD	OSD-SSD	OSM-SSD	$\mathrm{AP}_{0.5}$	$\mathrm{AP}_{0.75}$	AP
√			0.8300	0.4622	0.4649
	√		0.8449	0.5609	0.5096
		√	0.8470	0.5509	0.5114
√	√		0.8496	0.5811	0.5265
√		√	0.8481	0.5842	0.5283
	√	√	0.8543	**0.6156**	**0.5478**
√	√	√	**0.8605**	0.6087	0.5468

aligned targets. Hence, we perform a comparison between NMS-DM (the fusion between predictions of OSD-SSD and OSM-SSD adopting NMS) and WBF-DM. We conduct experiments on two forms of AIR-SARShip-2.0 [33]: the whole dataset containing 536 images and the part of densely aligned targets containing 84 images (we select 84 images with densely aligned targets from AIR-SARShip-2.0). The IoU thresholds of both NMS-DM and WBF-DM are set as 0.7. Table IX shows the comparison results. For the whole dataset, WBF-DM performs better than NMS-DM under $\mathrm{AP}_{0.75}$ and AP metrics, but is indeed slightly inferior to NMS-DM under $\mathrm{AP}_{0.5}$ metric. For the part of densely aligned targets, WBF-DM is superior to NMS-DM in all metrics. Moreover, when the requirement of localization accuracy gets higher, $\mathrm{AP}_{0.75}$ and AP gain larger improvements. We also visualize a set of comparison results for

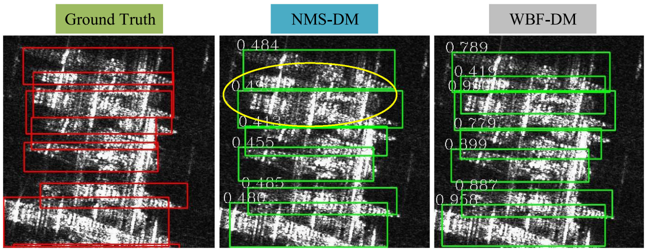

Fig. 9. Visualization of comparison results between NMS-DM and WBF-DM for densely aligned targets.

TABLE IX
COMPARISON BETWEEN NMS-DM AND WBF-DM ON AIR-SARSHIP-2.0 [33]

Datasets (image numbers)	Method	$\mathrm{AP}_{0.5}$	$\mathrm{AP}_{0.75}$	AP
Whole dataset (536)	NMS-DM	**0.8601**	0.5905	0.5302
	WBF-DM	0.8543	**0.6156**	**0.5478**
Densely aligned targets (84)	NMS-DM	0.8636	0.5348	0.5030
	WBF-DM	**0.8654**	**0.5475**	**0.5183**

TABLE X
COMPARISON OF INFERENCE SPEED BETWEEN IMAGENET-SSD AND WBF-DM

Method		Inference Speed (ms)			
		AIR-SAR Ship-1.0 [32]	AIR-SAR Ship-2.0 [33]	HRSID [34]	SSDD [35]
Faster R-CNN [7]	ImageNet -SSD	**26.9**	**26.9**	**49.1**	**45.4**
	WBF-DM	55.7	55.4	98.9	98.7
YOLOv3 [36]	ImageNet -SSD	**17.2**	**17.2**	**17.2**	**17.2**
	WBF-DM	36.2	35.9	35.3	36.4

TABLE XI
OVERALL PERFORMANCE OF DIFFERENT METHODS ON TWO SAR SHIP DETECTION DATASETS USING THE FCOS [44] BENCHMARK

Method	Precision	Recall	F1	$\mathrm{AP}_{0.5}$	$\mathrm{AP}_{0.75}$	AP
AIR-SARShip-1.0 [32]						
ImageNet-SSD	0.8649	0.8301	0.8471	0.8638	0.6280	0.6024
OSD-SSD	**0.9124**	0.8231	**0.8654**	0.8756	0.6673	0.6279
OSM-SSD	0.8714	0.8581	0.8647	0.8883	0.6641	0.6460
WBF-DM	0.8025	**0.8756**	0.8374	**0.8955**	**0.7070**	**0.6657**
SSDD [35]						
ImageNet-SSD	0.9398	0.8521	0.8938	0.9179	0.5946	0.5554
OSD-SSD	**0.9373**	0.8823	**0.9090**	0.9324	0.6485	0.5649
OSM-SSD	0.9341	0.8603	0.8957	0.9248	0.6126	0.5579
WBF-DM	0.8974	**0.9099**	0.9036	**0.9388**	**0.6841**	**0.5864**

TABLE XII
COMPARISON OF DIFFERENT CNN-BASED SHIP DETECTORS ON THE SSDD DATASET [35]

Method		Backbone	Precision	Recall	F1	$AP_{0.5}$	$AP_{0.75}$	AP
One-stage	RetinaNet [11]	ResNet50	0.8310	0.8933	0.8610	0.8891	0.5028	–
	RFA-Det [27]	ResNet50	–	–	–	0.9472	–	–
	DRBox-v2 [28]	VGG16	–	–	0.9149	0.9281	–	–
	FBR-Net [26]	ResNet50	0.9279	0.9401	0.9340	0.9410	0.5906	–
	WBF-DM (YOLOv3)	Darknet53	**0.9413**	**0.9633**	**0.9523**	**0.9577**	**0.7017**	**0.6058**
Two-stage	DCMSNN [15]	ResNet50	0.9049	0.8914	0.8981	0.8943	0.5417	–
	R-DFPN [44]	ResNet50	–	–	0.8529	0.8345	–	–
	DAPN [19]	ResNet50	–	–	–	0.8980	–	–
	HR-SDNet [21]	HRFPN-W40	–	–	–	0.9730	0.7430	0.6370
	WBF-DM (Faster R-CNN)	ResNet50	**0.9079**	**0.9669**	**0.9364**	**0.9740**	**0.7534**	**0.6426**

densely aligned targets in Fig 9. We can observe that NMS-DM misses a ship with yellow circle marked compared to WBF-DM. All these phenomena illustrate that WBF-DM is more effective than NMS-DM for densely aligned targets due to the averaged boxes of WBF strategy.

As for the speed performance, we also conduct comparison experiments between WBF-DM and the two benchmarks. We can observe from Table X that the time consumption of WBF-DM is slightly more than twice that of ImageNet-SSD for both two benchmarks. For YOLOv3, the speed performances of different ImageNet-SSD are identical on four datasets, because the size of input images are the same. The time consumption of WBF-DM on four datasets is slightly different due to the number of boxes to be fused is different. In general, the double-time consumption phenomenon shows that the WBF is a strategy that trades time for accuracy.

5) Generalization Ability: To further demonstrate that the proposed OSM and OSD pretraining techniques can be applied to other SAR ship detection methods, we also conduct experiments on another ship detection method, namely, FCOS [44]. Either Faster R-CNN [7] or YOLOv3 [36] belongs to anchor-based detection methods, where the predefined set of anchors has a great impact on the performance as well as the generalization ability. Furthermore, most anchors are redundant due to the sparsity of ships, which will increase the computation. FCOS is a representative anchor-free detection method and designed to eliminate the predefined set of anchors. Instead of regressing the transformation from predefined anchor boxes to the target bounding box, FCOS explores fully convolutional networks (FCN) [46] to tackle the detection problem in a perpixel prediction manner and directly regresses the target bounding box at each pixel. We perform the FCOS benchmark on AIR-SARShip1.0 and SSDD. As for the experimental parameter settings, the detector is trained with SGD for 36 epochs with a total of eight images per minibatch. The initial learning rate is set as 0.002 and then divided by a factor of 10 at the 24th and 33 rd epochs. The weight decay is 0.0001 and the momentum is 0.9. We adopt ResNet50 as the basic backbone for convenience. Table XI shows the overall performance of different methods on two datasets. We can observe that OSD-SSD and OSM-SSD have higher detection performance than ImageNet-SSD in terms of all evaluation metrics. Moreover, $\mathrm{AP}_{0.5}$, $\mathrm{AP}_{0.75}$, and AP of WBF-DM gain larger improvements due to the complementarity of OSD and OSM pretraining techniques. These phenomena can illustrate the generalization ability of the proposed OSM and OSD pretraining techniques and the effectiveness of the WBF strategy.

6) Comparison With CNN-Based SSDs: We split CNN-based SSDs into two categories: one-stage ship detector and two-stage ship detector for a fair comparison. The overall detection results of different detectors on the SSDD dataset [35] are listed in Table XII. For one-stage detector, we compare the proposed WBF-DM method based on the YOLOv3 [36] benchmark with four CNN-based methods including RetinaNet [11], RFA-Det [27], DRBox-v2 [28], and FBR-Net [26]. Specifically, our method achieves 1.34% and 2.32% higher precision rate and recall rate compared with the FBR-Net method, respectively. In terms of F1 score and $\mathrm{AP}_{0.5}$, our method obtains 1.83% and 1.67% better improvements. Furthermore, our method considerably improves $\mathrm{AP}_{0.75}$ by a large margin of 11.11%, obtaining the predicted boxes with higher localization accuracy. Compared to other methods, similar enhancements can also been achieved. As for two-stage detector, we also compare our WBF-DM based on the Faster R-CNN [7] benchmark with four CNN-based methods including DCMSNN [15], R-DFPN [45], DAPN [19], and HR-SDNet [21]. We can observe that our WBF-DM method can achieve considerable improvements in all degrees compared to other state-of-the-art ship detectors. More importantly, our methods can easily combine with other detectors to boost SAR ship detection. All these improvements verify the effectiveness and complementarity of OSD-SSD and OSM-SSD, and the superiority of the combination of the two detectors.

IV. CONCLUSION

Considering that directly leveraging ImageNet pretraining technique as commonly used is hard to obtain a good SSD, this article introduce a completed framework to boost ship detection in SAR images. To resolve the problem that ships from ImageNet are different from ships from earth observations, we first propose an OSD pretraining technique to improve the feature learning ability of ships in SAR images by fully taking advantage of ships' annotation information from aerial images. Second, to handle the problem of different imaging geometry between optical and SAR images, we propose an OSM pretraining technique to enhance the general feature embedding of SAR images by common representation learning via BNNs. Third, observing the complementary advantages of OSM-SSD and OSD-SSD, this article employ the WBF strategy to combine the predictions of OSD-SSD and OSM-SSD to further improve detection results in SAR images. Finally, various experiments are conducted on four SAR ship detection datasets and three representative CNN-based detection benchmarks to verify the effectiveness and complementarity of OSD-SSD and OSM-SSD, and the state-of-the-art performance of the combination of the two detectors. In the future, we will consider exploring the performance of the proposed method on more complicated networks and more challenging datasets.

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. V. Schie, "Ship surveillance with TerraSAR-X," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1092–1103, Mar. 2011.

[3] B. Hou, X. Chen, and L. Jiao, "Multilayer CFAR detection of ship targets in very high resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 811–815, Apr. 2015.

[4] G. Gao, K. Ouyang, Y. Luo, S. Liang, and S. Zhou, "Scheme of parameter estimation for generalized gamma distribution and its application to ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1812–1832, Mar. 2017.

[5] J. Zhu, X. Qiu, Z. Pan, Y. Zhang, and B. Lei, "Projection shape template-based ship target recognition in TerraSAR-X images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 222–226, Feb. 2016.

[6] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8614–8618.

[7] S. Ren, K. He, R.Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

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

[9] 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.

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

[11] Tsung-Yi 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.

[12] 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.

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

[14] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[15] J. Jiao *et al.*, "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.

[16] 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.

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

[18] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Proc. Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[19] 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.

[20] 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.

[21] S. Wei *et al.*, "Precise and robust ship detection for high-resolution SAR imagery based on HR-SDnet," *Remote Sens.*, vol. 12, no. 1, pp. 167–196, 2020.

[22] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 2, pp. 2738–2756, 2020, doi: 10.1109/JSTARS.2020.2997081.

[23] L. Du, L. Li, D. Wei, and J. Mao, "Saliency-guided single shot multibox detector for target detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3366–3376, May 2020.

[24] 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, pp. 2483–2519, 2019.

[25] T. Zhang and X. Zhang, "High-speed ship detection in SAR images based on a grid convolutional neural network,"*Remote Sens.*, vol. 11, no. 10, pp. 1206–1229, 2019.

[26] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.

[27] S. Chen, J. Zhang, and R. Zhan, "R2FA-Det: Delving into high-quality rotatable boxes for ship detection in SAR images," *Remote Sens.*, vol. 12, no. 12, 2020, Art. no. 2031.

[28] 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.

[29] Gui-Song Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[30] Y. Xu, X. Xiang, and M. Huang, "Task-driven common representation learning via bridge neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 5573–5580.

[31] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, no. 2, pp. 104117, 2021.

[32] S. Xian, W. Zhirui, S. Yuanrui, D. Wenhui, Z. Yue, and F. Kun, "Airsarship-1.0: High resolution SAR ship detection dataset," *J. Radars*, vol. 8, no. 6, pp. 852–862, 2019.

[33] "2020 Gaofen challenge on automated high-resolution earth observation image interpretation," [Online]. Available: http://en.sw.chreos.org

[34] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[35] 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.*, 2017, pp. 1–6.

[36] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[37] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[38] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[39] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.

[40] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4601–4611.

[41] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[42] M. Huang *et al.*, "The QXS-SAROPT dataset for deep learning in SAR-optical data fusion," 2021, *arXiv:2103.08259*.

[43] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS-improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5561–5569.

[44] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[45] X. Yang *et al.*, "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, pp. 132–145, 2018.

[46] 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.

Wei Bao received the B.S. degree in communication engineering from Nanjing Tech University, Nanjing, China, in 2018. He is currently working toward the master's degree with Beijing Institute of Technology, Beijing, China.

He is currently performing cooperation research with researchers at the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. His research interests include computer vision, few-shot learning and remote sensing object detection.

Meiyu Huang received the B.S. degree in computer science and technology from Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree in computer application technology from the University of Chinese Academy of Sciences, Beijing, China, in 2016.

She is currently an Assistant Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. Her research interests include machine learning, ubiquitous computing, human–computer interaction, computer vision, and image processing.

Yaqin Zhang received the B.S. degree in statistics from the School of Science, North University of China, Taiyuan, China, in 2014. She is currently working toward the master's degree with the School of Mathematics and Computational Science, Xiangtan University, Xiangtan, China.

Since 2020, she has been performing cooperation research with researchers with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. Her current research interests include the area of computer vision and computational imaging.

Yao Xu received the B.S. degree in electrical and computer engineering from Shanghai Jiao Tong University, Shanghai, China, in 2013, and the M.S. degree in electrical and computer engineering from the University of California, Irvine, CA, USA, in 2016.

He is currently an Assistant Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. His research interests include deep learning, data Fusion, distributed system, and computer architecture.

Xuejiao Liu received the B.S. degree in computational mathematics from Jilin University, Changchun, China, in 2013, and the Ph.D. degree in computational mathematics from the University of Chinese Academy of Sciences, Beijing, China, in 2018.

She is currently an Assistant Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. Her research interests include deep learning, generative model, and numerical simulation.

Xueshuang Xiang received the B.S. degree in computational mathematics from Wuhan University, Wuhan, China, in 2009, and the Ph.D. degree in computational mathematics from the Academy of Mathematics and Systems Science of Chinese Academy of Sciences, Beijing, China, in 2014.

In 2016, he was a Postdoctoral Researcher with the Department of Mathematics, National University of Singapore, Singapore. He is currently an Associate Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. His research interests include numerical methods for partial differential equations, image processing and deep learning.