Journal of
Marine Science and Engineering

MDPI

Article

An Effective Multi-Layer Attention Network for SAR Ship Detection

Zhiling Suo, Yongbo Zhao * and Yili Hu

National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China
* Correspondence: ybzhao@xidian.edu.cn

Abstract: The use of deep learning-based techniques has improved the performance of synthetic aperture radar (SAR) image-based applications, such as ship detection. However, all existing methods have limited object detection performance under the conditions of varying ship sizes and complex background noise, to the best of our knowledge. In this paper, to solve both the multi-scale problem and the noisy background issues, we propose a multi-layer attention approach based on the thorough analysis of both location and semantic information. The solution works by exploring the richness of spatial information of the low-level feature maps generated by a backbone and the richness of semantic information of the high-level feature maps created by the same method. Additionally, we integrate an attention mechanism into the network to exclusively extract useful features from the input maps. Tests involving multiple SAR datasets show that our proposed solution enables significant improvements to the accuracy of ship detection regardless of vessel size and background complexity. Particularly for the widely-adopted High-Resolution SAR Images Dataset (HRSID), the new method provides a 1.3% improvement in the average precision for detection. The proposed new method can be potentially used in other feature-extraction-based classification, detection, and segmentation.

Keywords: multi-layer attention module (MAM); ship detection; synthetic aperture radar (SAR); multi-scale feature maps

Citation: Suo, Z.; Zhao, Y.; Hu, Y. An Effective Multi-Layer Attention Network for SAR Ship Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 906. https://doi.org/10.3390/jmse11050906

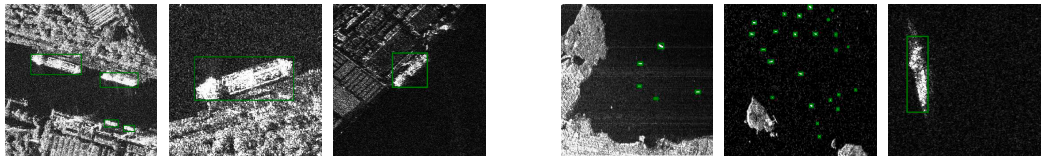
Academic Editors: Merv Fingas and Lev Shemer

Received: 2 February 2023
Revised: 31 March 2023
Accepted: 18 April 2023
Published: 23 April 2023

1. Introduction

Synthetic aperture radar (SAR) is widely employed in marine exploration (itself an important field for multiple areas, ranging from biological to chemical research [1]) for being able to overcome interference caused by adverse weather or insufficient light during observations [2,3]. Not only has SAR become more widely used, but the volume of data produced through SAR usage has also increased steadily due to the proliferation of high-resolution satellite imaging. Among the interesting applications, SAR-based ship detection has garnered special attention in international research in the field of maritime exploration due to its applications in trajectory prediction, congestion avoidance, and activity monitoring [4,5].

The conventional methods of ship detection in SAR images have a low level of generalization, which is an issue as SAR is used in very different scenarios, such as varying sunlight and changing weather conditions, and for different types of ship [6,7]. Another problem, illustrated in Figure 1a, is that the images obtained are often full of unnecessary and interfering information that pertains to the environment, such as the bank edges and piers, but does not include the ship, which is the object that needs to be detected. Finally, as shown in Figure 1b, ships come in very different sizes, so an efficient detection solution needs to be able to handle objects of various scales and still properly detect them, which is an issue since flexibility and adequate response is essential [8]. There are many works in the literature on SAR-based ship detection, as will be elaborated in Section 2, but none handles all the problems listed here.



(a) (b)

Figure 1. Images generated through SAR containing ships [9]. In (**a**), inshore ships with interference such as the shipyards and piers in the background are shown. In (**b**), multi-scale ship objects in SAR images are shown. Green rectangles are the ships to be detected.

In this paper, we will propose a solution that addresses all of these issues. Our proposal is based on the construction of a precise feature map that identifies key regions in the figure where the ship probably is. This is performed by careful analysis of the semantic and location information of the image through a multi-layer attention module (MAM) in our novel multi-layer attention network (MANet). The processing of both semantic and spatial data is performed on feature maps generated on the original image from ResNet50, a widely adopted standard neural network, which trains deep neural networks and extracts features at different resolutions. This richness in information not only allows our proposal to exclude noise of unimportant parts of the image more efficiently but also allows us to lower the chance of ship misidentification due to the scale of the vessel. The resulting enhanced feature map produced by our MANet can then be used as an input for the Feature Pyramid Network (FPN) [10], a standard deep learning solution for multi-scale object detection, for improved performance at ship detection from SAR images.

The main contributions of this paper are as follows. (i) We propose a cross-level fusion of multi-layer feature maps that can integrate semantic and location data, thus providing a better solution to the multi-scale problem. (ii) We design an attention mechanism that ensures that the feature maps of each layer are efficiently built so that the interference of inshore objects is reduced and more focus is given to the ship object. (iii) We carefully evaluate the performance of our proposed approach and show that our method is significantly better in detection precision than conventional solutions.

The remainder of this paper is organized into the following sections. Section 2 discusses in depth the state-of-the-art methods regarding ship detection. Section 3 explains what our proposed approach is. Section 4 contains the performance evaluation on our proposed solution through images obtained by SAR. Section 5 concludes the paper.

2. Related Works

Ship detection in SAR images can be divided into traditional SAR image processing and SAR image processing based on deep learning.

2.1. Traditional SAR Ship Detection

The traditional solutions for ship detection are based on a statistical clutter modeling called constant false alarm rate (CFAR) [6,7,11–15]. The foundation of such methods is the exploring of the statistical distribution of different marine clutters and using adaptive thresholding techniques. For example, Ai et al. [11] utilized the strong correlation in gray intensity and 2D joint log-normal distribution of pixels in their surrounding area to identify ships. Wang et al. [6] used a quick block detector to remove sea clutter and then identified ships based on kernel density estimate and aspect ratio of the remaining pixels. Leng et al. [7] focused on minimizing the impact of ambiguities and sea clutter in SAR images. Wang et al. [13] introduced an intensity space domain-based detector that takes advantage of the intensity and correlation between pixels. Pappas et al. [14] used super-pixels to define the guard bands and backdrops used in CFAR. Similarly, Li et al. [15] differentiated objects from clutter through weighted information entropy, and used a two-stage CFAR detection system that offered global and local detection. However, in the above CFAR-based solutions, key parameters related to multiple scenario conditions have

to be manually pre-determined, significantly decreasing the capacity of generalization in traditional solutions and worsening the overall performance for unexpected situations. They are also particularly sensitive to objects in the background. This deteriorates the performance of the detector when the backdrop objects have complex features.

2.2. Deep Learning-Based SAR Ship Detection

With the popularization of deep learning [16–19], the application of convolutional neural network (CNN) in SAR-based ship detection is becoming a research hotspot [20]. For example, Li et al. [20] evaluated hard negative mining, transfer learning, feature fusion, and other implementation techniques, alongside a novel dataset and CNNs to achieve faster detection [16]. Moreover, Zhou et al. [21] evaluated the lightweight algorithm of You Only Look Once (YOLO) v4 [19] applied to ship detection to reduce the number of needed features by depth-wise separable convolution and MobileNet v2 as the network architecture. Chang et al. [22] simplified the classic YOLOv2 architecture for CNN to reduce calculation time when detection ships in SAR images. Zhang et al. [23,24] proposed a variant of YOLO-FA to incorporate the frequency-domain information. Sun et al. [25] used fully convolutional one-stage object identification as the foundation of their detector to improve the network's position regression branch features, and bounding box regression detection. Similarly, Mao et al. [26] proposed an anchor-free SAR-based ship detection system for bounding box regression, and a different network for score map regression. Zhou et al. [27] added to this by simplifying the CNN operation through a two-way dense connection module. Bai et al. [28] adapted the anchor-free detector and established a boundary frame detection strategy to detect objects. However, these works fail to address the substantial interference caused by near-shore background and the vast diversity of scale (see Figure 1 for an example). Because of that, the accuracy of detecting multi-scale ships using SAR still needs improvement, even when using deep learning.

Existing research shows that feature pyramid networks (FPN) [10] are one of the solutions that can address multi-scale object detection and significantly improve small object identification. Additionally, an attention mechanism object detection can detect and focus on specific areas of interest in a big image [29,30], which is a possibility with useful application in SAR-based ship detection [31–33]. Cui et al. [31] introduced a dense attention pyramid network that used a convolutional block attention module (CBAM) densely connected to each concatenated feature map from top to bottom. Zhao et al. [32] also used CBAM with an attention-receptive pyramid scheme. Yang et al. [33] used a coordinate attention module and a receptive field expanded module that works better with varying sizes of vessels. Bai et al. [34] proposed a feature enhanced pyramid and shallow feature reconstruction network to improve the detection for small and medium-sized ships. However, these studies did not fully integrate the feature maps of low and high levels, resulting in limited identification performance.

Nonetheless, none of the mentioned solutions integrate multi-layer features, which is essential for the success of deep learning-based object recognition. For example, for identifying ships, high-level feature maps are better due to their wider receptive field, more semantic information, and less spatial information. However, low-level feature maps are more suited for recognizing smaller ships due to their narrower receptive fields, greater emphasis on spatial information, and less reliance on semantic information. Multi-scale ship detection relies heavily on this kind of seamless combination of semantic and geographical data. Such a technique allows the accuracy to be improved by extracting relevant data from bigger feature maps. Attention-based approaches can even be used to draw focus to relevant details and ignore backdrop noise to further increase ship detection accuracy. Our proposed MAM is based on these ideas, integrating semantic and spatial information to highlight the salient features of the desired object and using an attention-based technique alongside a multi-layer feature map to highlight features of the object. These characteristics make the feature map output by our proposal more useful for the underlying network.

3. Methods

In this section, we will explain the details of our proposed MANet-based architecture for ship detection in SAR images.

3.1. Overall Design

Figure 2 shows how the CNN-based solution for SAR-based ship detection works and where we propose to integrate our novel MANet, composed of a Faster R-CNN [16] and a new multi-layer attention module (MAM), into such a system. In Figure 3, we show, in detail, how our proposed MAM is designed. We divide it into two sub-modules, multi-layer channel attention module (MCAM) and multi-layer spatial attention module (MSAM).

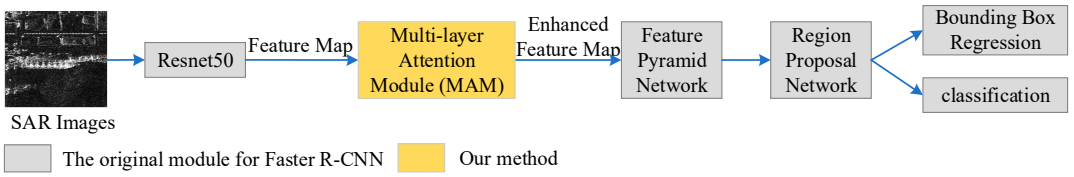


Figure 2. The multi-layer attention network's overall structure. Note that R-CNN is the region convolutional neural network [16].

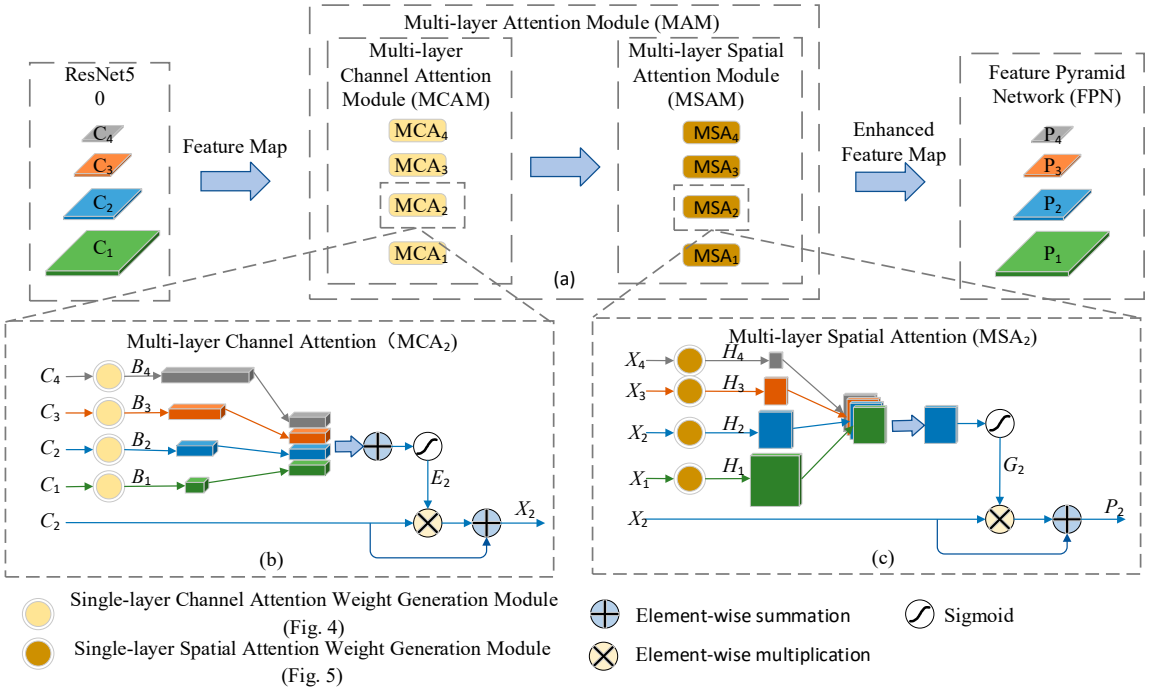


Figure 3. Overview of the multi-layer attention module (MAM). In (**a**), the MAM module's general structure is outlined. In (**b**,**c**), the flow charts of the multi-layer channel attention module (MCAM) and the multi-layer spatial attention module (MSAM), respectively, are shown. *C* represents the feature map, *P* is the input to the FPN, *B* is the channel attention, *E* and *G* are the outputs of the sigmoid functions, *H* is the spatial attention map, and the subscripts 1~4 represent different channels.

Our MAM works as follows. First, the MCAM performs channel attention processing on each layer's feature maps generated by the backbone network, i.e., ResNet50 in Figure 3, so that the significant features on the channel axis are identified. Each channel's focus is then combined and applied to the initial feature map to enable the adaptive acquisition of a feature map with significant scale information. Then, the MSAM will perform spatial attention processing on the outputs of each layer of the MCAM. These will be fused to obtain weights for each position. The final optimum feature map is obtained by broadcasting

these weights to MCAM's feature maps for each layer's output. Through this process, the prominent aspects of the multi-scale feature map can be employed by the MAM when searching for ships. While FPN is capable of fusing feature maps from top to bottom, our proposed MAM uses an attention-based approach to reduce the amount of interference from the environment, filter out undesirable noise, highlight the needed information, and produce more relevant feature maps through a complete integration of semantic and geographic information contained in the data of each layer. The use of deep learning is paramount in identifying the relevant patterns in the data [35].

3.2. MAM

As mentioned earlier, the MAM is divided into two sub-modules, MCAM and MSAM, that are, themselves, responsible for the functions of multi-layer channel attention (MCA) enhancement and multi-layer spatial attention enhancement, respectively. The ResNet50 structure is chosen as the feature extraction network because it produces four layers of feature maps at varying resolutions. The set of generated features is denoted by $\{C_1, C_2, C_3, C_4\}$. Accordingly, the MCAM is composed of the processing unit set $\{\mathrm{MCA}_1, \mathrm{MCA}_2, \mathrm{MCA}_3, \mathrm{MCA}_4\}$. Each processing unit implements a channel attention enhancement focused on its corresponding feature map, and the resulting enhanced feature maps are defined as $\{X_1, X_2, X_3, X_4\}$. Similarly, the MSAM consists of the processing unit set $\{\mathrm{MSA}_1, \mathrm{MSA}_2, \mathrm{MSA}_3, \mathrm{MSA}_4\}$, which implements spatial attention enhancement for $\{X_1, X_2, X_3, X_4\}$, respectively. The final output feature graph, which will be the input of the FPN network, is denoted as $\{P_1, P_2, P_3, P_4\}$. Therefore, the whole MAM process can be summarized as follows:

$$X_i = \mathrm{MCA}_i(C_i) \quad i = 1, 2, 3, 4; \tag{1}$$

$$P_i = \mathrm{MSA}_i(X_i) \quad i = 1, 2, 3, 4. \tag{2}$$

3.3. MCAM

The MCAM is made of four MCA processing units denominated as $\{\mathrm{MCA}_1, \mathrm{MCA}_2, \mathrm{MCA}_3, \mathrm{MCA}_4\}$, as is illustrated in Figure 3a. Because the process is identical for each unit, the following explanations will consider a single MCA only. The specific operation steps of MCA are divided into the generation of the single-layer channel attention (SCAM), the generation of the map weights, and the generation of the feature map. The three steps are thoroughly explained as follows.

3.3.1. Generation of Single-Layer Channel Attention

To capture rich semantic and location information, all four feature maps of the backbone network module, $\{C_1, C_2, C_3, C_4\}$, are used as input of the MCA processing unit. First, each feature map $\{C_i, i = 1, 2, 3, 4\}$ goes through the SCAM (see Figure 4 for an illustration). For convenience of description, the subscript of C_i is removed to obtain $C \in \mathrm{R}^{\mathrm{c}\times\mathrm{h}\times\mathrm{w}}$, where c, h, and w represent the number of channels, height, and width of the feature map, respectively. The detailed procedures that make up SCAM's operation are presenting in the following.

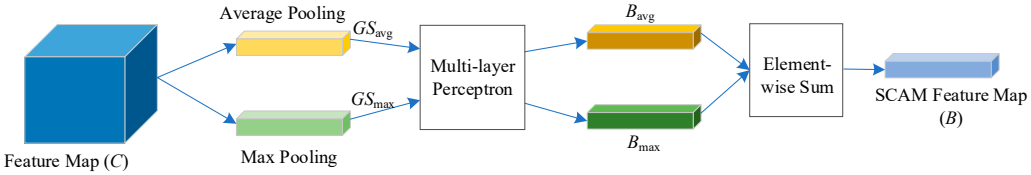


Figure 4. The single-layer channel attention module (SCAM). GS_{avg} and GS_{max} are the average and maximum of the global spatial pooling, respectively. B_{avg} and B_{max} are the average and maximum of the channel attention measures, respectively.

Global spatial average pooling (GSAP). The operation of GSAP for feature map $C \in \mathrm{R}^{\mathrm{c}\times\mathrm{h}\times\mathrm{w}}$ is represented as follows

$$GS_{\mathrm{avg}}^{k} = \frac{1}{\mathrm{h}\times\mathrm{w}} \sum_{\substack{m=1,2,3,\ldots,\mathrm{h} \\ n=1,2,3,\ldots,\mathrm{w}}} C_{m,n,k} \tag{3}$$

where $C_{m,n,k}$ denotes the feature map of the m-th row and n-th column on the k-th channel, and GS_{avg}^{k} represents the feature map of the k-th channel after GSAP. In other words, the set of all GS_{avg}^{k} make up the bigger feature map $GS_{\mathrm{avg}} = \left\{ GS_{\mathrm{avg}}^{k} | k = 1,2,3,\ldots,c \right\} \in \mathrm{R}^{c\times1\times1}$ after GSAP is concluded.

Global spatial max pooling (GSMP). The operation of GSMP for feature map $C \in \mathrm{R}^{\mathrm{c}\times\mathrm{h}\times\mathrm{w}}$ is represented as follows

$$\begin{aligned} GS_{\mathrm{max}}^{k} = \max(C_{m,n,k}) \quad & m = 1,2,3,\ldots,\mathrm{h} \\ & n = 1,2,3,\ldots,\mathrm{w}, \end{aligned} \tag{4}$$

where max indicates an operator that obtains the maximum value in the channel, and GS_{max}^{k} denotes the feature map of the k-th channel after GSMP. In other words, the set of all GS_{max}^{k} make up the bigger feature map $GS_{\mathrm{max}} = \left\{ GS_{\mathrm{max}}^{k} | k = 1,2,3,\ldots,c \right\} \in \mathrm{R}^{c\times1\times1}$ after GSMP is concluded.

Then, the channel attention measures B_{avg} and B_{max} are generated by using a multi-layer perceptron (MLP) that receives, respectively, GS_{avg} and GS_{max} as input. Then, an element-wise summation is performed to obtain the single-layer channel attention B, i.e., channel attention B is computed as:

$$\begin{aligned} B &= B_{\mathrm{avg}} + B_{\mathrm{max}} \\ &= \mathrm{MLP}\left(GS_{\mathrm{avg}}\right) + \mathrm{MLP}(GS_{\mathrm{max}}). \end{aligned} \tag{5}$$

To minimize the number of parameters in the MLP model, the hidden activation size is changed to $\mathrm{R}^{\frac{\mathrm{C}}{\mathrm{r}}\times1\times1}$, where r represents the reduction ratio.

Using this method, $B_1 \in \mathrm{R}^{\mathrm{c}_1\times1\times1}$, $B_2 \in \mathrm{R}^{\mathrm{c}_2\times1\times1}$, $B_3 \in \mathrm{R}^{\mathrm{c}_3\times1\times1}$, and $B_4 \in \mathrm{R}^{\mathrm{c}_4\times1\times1}$ are obtained through SCAM, where c_i $(i = 1,2,3,4)$ is the channel index, and, without loss of generality, the condition $\mathrm{c}_1 < \mathrm{c}_2 < \mathrm{c}_3 < \mathrm{c}_4$ is assumed.

3.3.2. Generation of MCA

Considering that the dimensionality of the attention maps B_i $(i = 1,2,3,4)$ are not necessarily the same, it is necessary to first convert them into the same dimension. Taking the MCA_2 processing module as an example, B_1, B_3, and B_4 should be converted to the exact dimensions as B_2. The conversion process is indicated as follows:

$$\begin{aligned} D_i &= FC(B_i) \quad i = 1,3,4, \\ D_2 &= B_2, \end{aligned} \tag{6}$$

where FC represents the convolution operation.

After the conversion, the outputs will satisfy $D_1 \in \mathrm{R}^{\mathrm{c}_2\times1\times1}$, $D_3 \in \mathrm{R}^{\mathrm{c}_2\times1\times1}$, and $D_4 \in \mathrm{R}^{\mathrm{c}_2\times1\times1}$. By doing this, all four layers of the channel attention map can be added through an element-wise summation. The sigmoid procedure is then used to determine the MCA map's weight. This process is represented by

$$E_2 = \sigma(D_1 + D_2 + D_3 + D_4), \tag{7}$$

where σ denotes the sigmoid function.

Similarly, $E_1 \in R^{c_1 \times 1 \times 1}$, $E_3 \in R^{c_3 \times 1 \times 1}$, and $E_4 \in R^{c_4 \times 1 \times 1}$ can be obtained following the same process.

3.3.3. Generation of the MCA Feature Map

Finally, the MCA feature map of each layer is obtained through Equation (8):

$$X_i = C_i \times (1 + E_i) \quad i = 1, 2, 3, 4. \tag{8}$$

3.4. MSAM

The MSAM, which is illustrated in Figure 3a, is made up of four MSA processing units represented by $\{MSA_1, MSA_2, MSA_3, MSA_4\}$. The multi-channel attention feature $\{X_1, X_2, X_3, X_4\}$ obtained according to Equation (8) is used as input for the MSA units. As with the MCAM, the four MSA processing units have the same flow, thus our description will focus on a single module. Additionally, similar to MCA, the specific operation steps of the MSA are divided into the generation of the single-layer spatial attention, generation of the multi-layer spatial attention, and generation of the multi-layer spatial attention feature map.

3.4.1. Generation of Single-Layer Spatial Attention

First, each feature map $\{X_i, i = 1, 2, 3, 4\}$ goes through the single-layer spatial attention module (shown in Figure 5). To simplify the description, the subscript of the output characteristics of the MCA is removed to obtain $X \in R^{c \times h \times w}$.

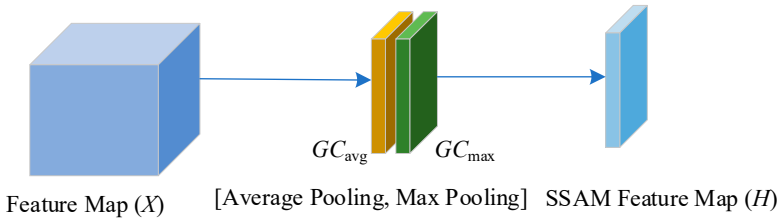


Figure 5. The single-layer spatial attention module (SSAM). GC_{avg} and GC_{max} are the average and maximum of the global channel pooling, respectively.

Global channel average pooling (GCAP). The operation GCAP is completed for each feature map $X \in R^{c \times h \times w}$, with process being computed as follows

$$GC_{m,n,avg} = \frac{1}{c} \sum_{k=1}^{c} X_{m,n,k}, \tag{9}$$

where $X_{m,n,k}$ denotes the feature mapping value at the position of row m and column n on channel k, and $GC_{m,n,avg}$ represents the feature map of the corresponding spatial position after GCAP. The set of all $GC_{m,n,avg}$ makes up the feature map $GC_{avg} = \{GC_{m,n,avg} | m = 1, 2, 3, \ldots, h, n = 1, 2, 3, \ldots, w\}$ after the GCAP operation is completed, where $GC_{avg} \in R^{1 \times h \times w}$.

Global channel max pooling (GCMP). The operation of GCMP is completed for each feature map $X \in R^{c \times h \times w}$, with the operation being summarized by Equation (10).

$$GC_{m,n,max} = Max(X_{m,n,k}), \tag{10}$$

where Max obtains the maximum value in row m and column n among all channels. Moreover, $GC_{m,n,max}$ indicates the feature map of the corresponding spatial position after GCMP is completed. The set of all $GC_{m,n,max}$ forms the feature map $GC_{max} = \{GC_{m,n,max} | m = 1, 2, 3, \ldots, h, n = 1, 2, 3, \ldots, w\}$ after GCMP is completed, where $GC_{max} \in R^{1 \times h \times w}$.

GC_{avg} and GC_{max} are then concatenated, and a convolution with a standard layer is performed, culminating in the creation of the spatial attention map. The operation is summarized as

$$H = FC\left(Concat\left(GC_{avg}, GC_{max}\right)\right). \tag{11}$$

where Concat and FC stand for concatenation and convolution with a 7 × 7 filter size, respectively. Next, we obtain single-layer spatial attention feature map $H \in R^{1\times h\times w}$.

Following the described process, a four-layer spatial attention map is obtained composed of $H_1 \in R^{1\times h_1\times w_1}$, $H_2 \in R^{1\times h_2\times w_2}$, $H_3 \in R^{1\times h_3\times w_3}$, and $H_4 \in R^{1\times h_4\times w_4}$, where h_1, w_1, h_2, w_2, h_3, w_3, h_4, and w_4 are the height and width of the four-layer feature map and, without loss of generality, we assume that $h_1 > h_2 > h_3 > h_4$ and $w_1 > w_2 > w_3 > w_4$.

3.4.2. Generation of Multi-Layer Spatial Attention

Given that the dimensions of H_1, H_2, H_3, and H_4 are different, we need to modify them so that they have same dimensionality. For example, considering the MSA_2 processing module, H_1, H_2, and H_4 should be converted into the same dimension as H_2. The conversion process is calculated by Equation (12).

$$I_i = Sample(H_i) \quad i = 1, 3, 4,$$

$$I_2 = H_2, \tag{12}$$

where Sample denotes an up-sampling/down-sampling transformation.

After transformation, $I_1 \in R^{1\times h_2\times w_2}$, $I_3 \in R^{1\times h_2\times w_2}$, and $I_4 \in R^{1\times h_2\times w_2}$ are obtained, and then concatenated, convolved, and activated by a sigmoid function. The whole process is as follows

$$G_2 = \sigma(FC\ (Concat\ (I_1,\ I_2,\ I_3,\ I_4))) \tag{13}$$

where $G_2 \in R^{1\times h_2\times w_2}$.

Similarly, $G_1 \in R^{1\times h_1\times w_1}$, $G_3 \in R^{1\times h_3\times w_3}$, and $G_4 \in R^{1\times h_4\times w_4}$ can be obtained, giving us the global spatial attention feature map weights of each layer.

3.4.3. Generation of the Multi-Layer Spatial Attention Feature Map

The final refined output can be obtained by Equation (14).

$$P_i = X_i \times (1 + G_i) \quad i = 1, 2, 3, 4. \tag{14}$$

4. Experiments

In this section, we will detail the experiments conducted to evaluate how well the suggested approach detects the objects of interest. First, the dataset and parameters used in the experiments are presented. Then, the influence of our proposed MANet is shown by comparing its performance with other techniques in the existing datasets.

4.1. Datasets

Three standard SAR-based ship detection datasets were used for the experiments in this paper. The SAR Ship Detection Dataset (SSDD) [9] generated by the RadarSat-2, TerraSAR-X, and Sentinel-1 [36] satellites has 1160 pictures of ships in various situations, which cover places including Yantai, China, and Visakhapatnam, India. Following the original paper, the dataset was divided in a ratio of 8:2 between training and testing. The SAR-Ship-Dataset [37] was created using 102 Chinese Gaofen-3 pictures and 108 Sentinel-1 pictures. These images cover several near-shore areas in China, South Korea, Japan, etc. A sliding window of 256 × 256 pixels with a sliding step of 128 pixels was used on these pictures, resulting in 43,819 images that include 59,535 ships. Out of these images, a ratio of 7:2:1 was used to randomly select samples for training, verification, and testing. Finally, the high-resolution SAR images dataset (HRSID) [38] has 5604 cropped images and 16,951 ships placed at Houston, Sao Paulo, Aswan Dam, Shanghai, etc. These images come

from 136 panoramic SAR pictures, with a 25% overlap rate. The original pictures have a relatively high resolution of greater than 800 × 800 pixels. Overall, 65% of the images from the HRSID were used for training, while 35% were used for testing.

4.2. Parameters and Metrics

The MMDetection framework [39] and machines powered by AMD Ryzen 7 3700X processors and NVIDIA GeForce GTX 1080Ti GPU cards manufactured by AMD and NVIDIA respectively were used for all tests. The machine was running Ubuntu 18.04.

The parameters in the experiment are set in Table 1. Such values were obtained after careful tests to find the best hyperparameters for our desired application [40]. Other hyperparameters are the same as the default values in MMDetection.

Table 1. Test settings.

Batch Size	Initial Leaning Rate	Momentum Decay	Weight Decay	Number of Epochs
1	1.25×10^{-4}	1×10^{-4}	0.9	30

Four assessment indicators, precision, recall, F1 score, and AP (short for "average precision", a quality assessment index used in the literature), as given below [41], were used to gauge how well our proposed strategy performed.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where false-positive (FP) represents the number of false detections if the object is not present, false-negative (FN) represents the number of no-detections when the object exists, true-negative (TN) is the number of the correct no-detections, and true-positive (TP) represents the number of correct detections.

AP is defined as the integral of the precision as a function of the recall, as given by

$$AP = \int_0^1 Precision(Recall) \mathrm{d} Recall$$

The most often used indices for precision and recall are the ratios of properly predicted ships in all forecasts, for precision, and in all ground truth ships, for recall. Precision and recall are then combined to provide the full assessment measure known as the F1 score. Furthermore, AP is another widely used metric that also considers precision and recall. AP_{50}, AP_{75}, and AP_s are the metrics calculated here, with calculations following the same process seen in the literature [42]. Specifically, AP_{50} is the value of AP when the ratio of the intersection area of the detection bounding box and the real bounding box to the union area of the two boxes is greater than 50%. Similarly, AP_{75} is the value of AP when this ratio is greater than 75%. AP_s is the AP when dealing with small-sized ship objects with the area of the bounding box below 32 × 32 pixels.

4.3. Results and Analysis

Tables 2–4 show the performance of our proposed solution and other existing CNN methods on the datasets SSDD [9], SAR-Ship-Dataset [36], and HRSID [38], respectively.

Table 2. Comparison of different methods on SSDD.

Method	Precision	Recall	F1	AP_{50}
RetinaNet [43]	0.910	0.915	0.912	0.896
Cascade R-CNN [44]	0.940	0.899	0.919	0.893
SSD512 [18]	0.929	0.896	0.912	0.937
CenterNet [45]	0.943	0.945	0.944	0.935
CenterNet++ [46]	0.932	0.945	0.938	0.927
Baseline [16]	0.946	0.935	0.940	0.950
Proposal	0.953	0.949	0.951	0.957

Table 3. Comparison of different methods on SAR-Ship-Dataset.

Method	Precision	Recall	F1	AP_{50}
RetinaNet [43]	0.881	0.938	0.909	0.938
Cascade R-CNN [44]	0.910	0.926	0.918	0.920
SSD512 [18]	0.901	0.905	0.903	0.942
CenterNet [45]	0.927	0.935	0.931	0.950
CenterNet++ [46]	0.925	0.934	0.929	0.949
Baseline [16]	0.941	0.949	0.945	0.948
Proposal	0.952	0.950	0.951	0.954

Table 4. Comparison of different methods on HRSID.

Method	Precision	Recall	F1	AP_{50}
RetinaNet [43]	0.701	0.838	0.763	0.826
Cascade R-CNN [44]	0.899	0.794	0.843	0.793
SSD512 [18]	0.909	0.855	0.881	0.888
CenterNet [45]	0.818	0.874	0.845	0.863
CenterNet++ [46]	0.822	0.873	0.847	0.863
Baseline [16]	0.926	0.891	0.908	0.923
Proposal	0.931	0.902	0.916	0.936

SSDD. Our proposal can achieve the best results in all metrics, with a precision of 0.953, recall of 0.949, F1 of 0.951, and AP of 0.957. Compared to the single-stage approach RetinaNet [43], our proposal shows an improvement of approximately 6.1%. Even when compared with two-stage algorithms, such as Cascade R-CNN [44], the proposed method is approximately 6.4% better.

SAR-Ship-DataSet. As was the case with SSDD, our proposal shows the best performance, being, for example, approximately 3.4% better than Cascade R-CNN.

HRSID. Compared to the other datasets, HRSID has more complex backdrops and many pictures with tiny boats. These characteristics significantly degrade the performance of the other methods. However, the proposed solution is still showing the best performance, with an improvement of up to 14.3% compared with RetinaNet and 1.3% compared with Faster R-CNN, the second-best solution.

In summary, the mechanisms in our proposal to analyze semantic and spatial data gives it an edge that increases its accuracy and precision. The index fluctuation is also extremely small, highlighting the strong generalization of the proposal. The results in all datasets confirm that our method is highly competitive and consistent. The results on HRSID, in particular, show that the proposal is good at both multi-scale situations and the suppression of undesired background data.

To further understand what is behind the efficacy of our proposal, ablation experiments were performed where portions of our framework were removed to measure their impact. First, we compared the performance of our full proposal (MANet) and versions of it with just the MCAM and with just the MSAM. Baseline performance in the form of pure Faster R-CNN is also provided for reference. Results are shown in Table 5. First, it is

noteworthy that full MANet is better than its counterparts with just MCAM and with just MSAM. For example, on SSDD, our full proposal sees improvements of greater than approximately 0.7%. Comparing the baseline and MCAM, the AP_{50} improves from 0.950 to 0.953, showing that incorporating MCAM and producing more accurate feature maps provides a small advantage. Comparing the baseline and MSAM, the AP_{50} increased to 0.956 and the AP_s increased from 0.683 to 0.693. This indicates that the module can improve the identification of the object location, reduce the influence of complex backgrounds on the detection performance, and improve the detection performance of small ships.

Table 5. MCAM and MSAM ablation experiments on SSDD.

Method	MCAM	MSAM	AP_{50}	AP_{75}	AP_S
Baseline			0.950	0.823	0.683
MCAM	√		0.953	0.830	0.684
MSAM		√	0.956	0.848	0.693
Ours (MANet)	√	√	0.957	0.852	0.697

Note: MCAM represents multi-layer channel attention module. MSAM represents multi-layer spatial attention module. MANet stands for multi-layer attention network.

Next, tests were performed on different architectures of the backbone network, with results being shown with and without the use of the MAM. The performance is shown in Figure 6 and proves that adding MAM increases the method's overall recognition rate, regardless of the underlying network.

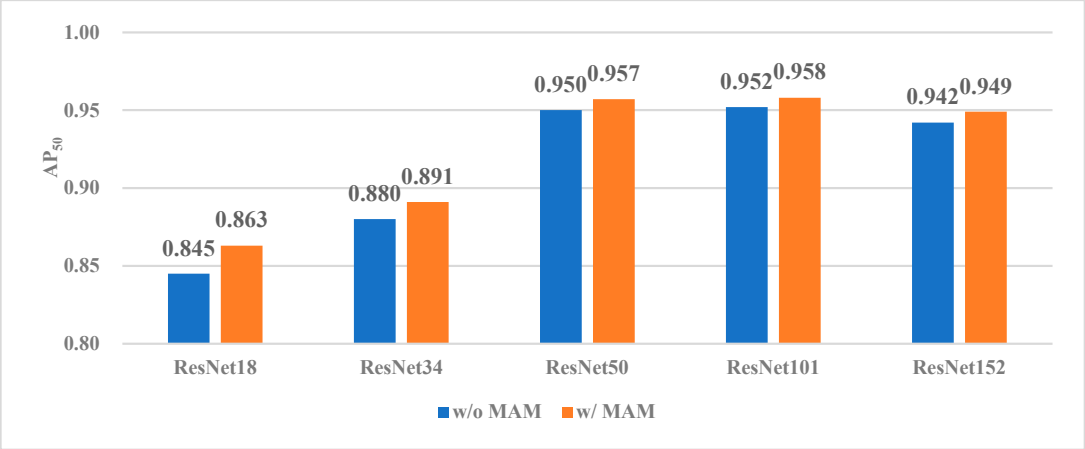
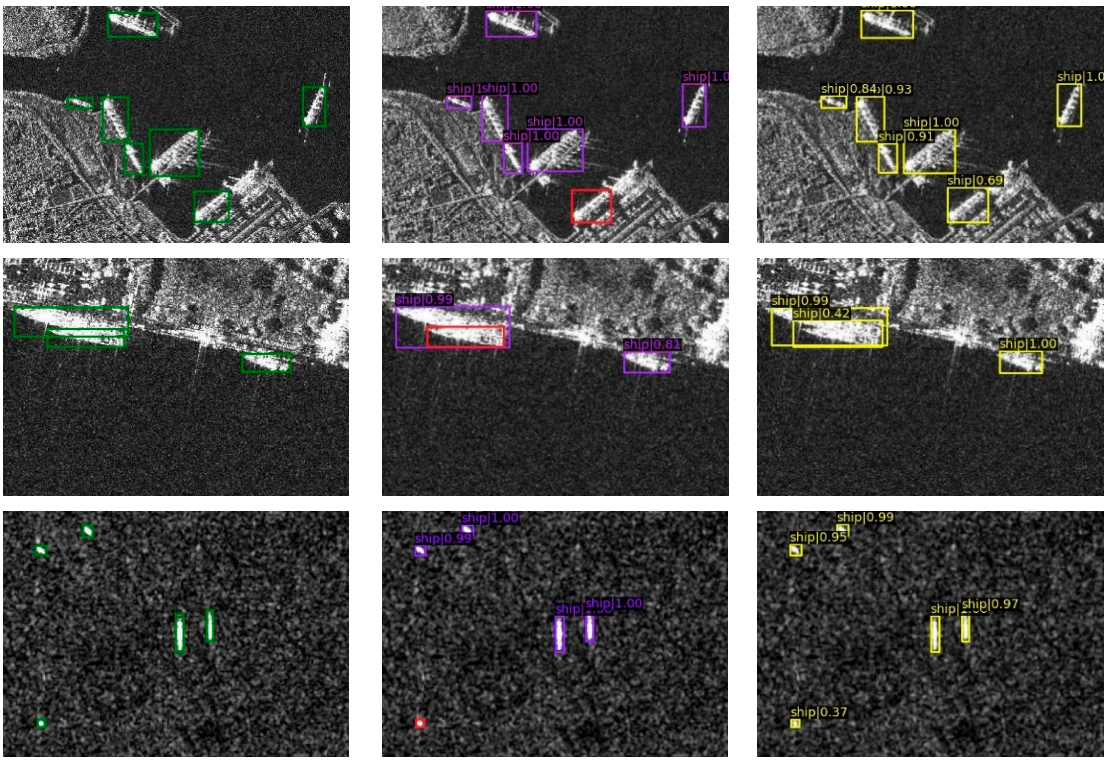


Figure 6. Comparison of different backbone methods.

Figure 7 provides a visual illustration of the detection effects between the baseline (pure Faster R-CNN) and our proposed MANet. The real ship position (ground truth) is given by Figure 7a, highlighted by the green boxes. The detection outcome of the baseline is seen in Figure 7b. The ships highlighted in purple were successfully detected while the ships highlighted in red were missed. In Figure 7c, the ships identified by our proposal, MANet, are highlighted in yellow. This works as an example of how our proposal is capable of detecting small vessels and ships among complex backgrounds even when the conventional methods encounter difficulty and fail.



(**a**) (**b**) (**c**)

Figure 7. Results of the suggested procedure compared to the baseline. The green, purple, yellow, and red boxes indicate the ground truths, the test results of baseline, the test results of our method, and the missed ship object, respectively. The proposed MANet generates the most accurate detection results, missing no ship objects. (**a**) Ground truth. (**b**) Baseline. (**c**) MANet.

4.4. Limitations and Possible Improvements

As for the densely arranged and parallel-placed ships, we have found that the new method has some limitation in identifying the vessels. The detection of densely side-by-side ships in the nearshore scene is the focus of future research. It is also the focus of future research to detect the difference between ships and objects, such as small islands. The result is not included to save space. We conject the reason for this limitation is that the horizontal and vertical bounding boxes we use may not be close enough to the ships' edges, especially when the ships do not appear horizontal or vertical in the image.

In addition, the vertical bounding box we use in this work would limit the detection performance, especially for the inshore scenario, which may contain buildings and other types of ships. Using an oriented bounding box can be a possible improvement of the proposed method. Along with the position information of the detected ships, the oriented bounding box can also provide the aspect ratio of the ship and the heading angle information, which greatly help the trajectory prediction and attitude determination.

Although our proposed method works well on the datasets, it is still an open problem to improve the detection precision of the small ships. The small ships in the images have fewer features than the large objects by nature. The multi-layer attention model we use in the work to extract the features of the object to be detected is one of the possible methods that can be adopted for small object detection. A modified multi-layer attention model or more advanced models will possibly appear and become feasible to enhance the detection performance for small ships in the future.

From the perspective of real-time application of the proposed method, we still have room for improvement. As of now, our main focus is to improve the detection precision. To realize real-time detection, we need to reduce the complexity of the network and/or

implement the method on more powerful computational platforms. Considering the size, weight, and power constraint systems, such as miniature drones, small satellites, and unmanned aerial vehicles, developing low-complexity methods and small-scale models seem to be a more feasible way to realize real-time object detection.

5. Conclusions

In this paper, we proposed a multi-layer attention network (MANet) method as a solution to both the multi-scale and complex inshore background issues in SAR-based ship detection. The proposed new approach can simultaneously explore the rich semantic information of high-level features and the accurate location data of low-level features. Different from the conventional methods, we integrate an attention mechanism into the feature map of each layer to adaptively improve the weight of essential details depending on the importance of different scales and positions for that particular image. Finally, we conduct extensive performance evaluation based on multiple widely-accepted SAR datasets. Result shows that the proposed multi-layer attention mechanism network significantly increases the accuracy of ship detection compared to the existing conventional solutions.

In addition to ship detection from SAR images discussed in the paper, the proposed method can be utilized in many other applications. Based on the structures and procedures of the new method, we can extend its application to any network based on feature extraction for classification, detection, and segmentation. In the future, we will investigate the detection of densely side-by-side ships in the nearshore scene, explore these applications of our new method and conduct more experiments with various datasets. Other focuses of future research include detecting the difference between ships and objects, such as small islands, and improving the computational efficiency.

Author Contributions: Conceptualization, Z.S.; methodology, Z.S.; software, Z.S. and Y.H.; validation, Z.S.; formal analysis, Z.S.; investigation, Z.S. and Y.H.; resources, Z.S. and Y.Z.; data curation, Z.S. and Y.Z.; writing—original draft preparation, Z.S.; writing—review and editing, Y.Z., and Y.H.; visualization, Z.S. and Y.H.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 62271379.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Karthäuser, J.F.; Koc, J.; Schönemann, E.; Wanka, R.; Aldred, N.; Clare, A.S.; Rosenhahn, A.; Laschewsky, A. Optimizing Fouling Resistance of Poly(Sulfabetaine)s through Backbone and Charge Separation. *Adv. Mater. Interfaces* **2022**, *9*, 2200677. [CrossRef]
2. Crisp, D.J. *The State-of-The-Art in Ship Detection in Synthetic Aperture Radar Imagery*; Defence Science and Technology Group: Canberra, Australia, 2004.
3. Pelich, R.; Longépé, N.; Mercier, G.; Hajduch, G.; Garello, R. AIS-Based Evaluation of Target Detectors and SAR Sensors Characteristics for Maritime Surveillance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3892–3901. [CrossRef]
4. Son, J.; Kim, D.-H.; Yun, S.-W.; Kim, H.-J.; Kim, S. The Development of Regional Vessel Traffic Congestion Forecasts Using Hybrid Data from an Automatic Identification System and a Port Management Information System. *J. Mar. Sci. Eng.* **2022**, *10*, 1956. [CrossRef]
5. Sun, C.; Xue, M.; Zhao, N.; Zeng, Y.; Yuan, J.; Zhang, J. A Deep Learning Method for NLOS Error Mitigation in Coastal Scenes. *J. Mar. Sci. Eng.* **2022**, *10*, 1952. [CrossRef]
6. Wang, C.; Jiang, S.; Zhang, H.; Wu, F.; Zhang, B. Ship Detection for High-Resolution SAR Images Based on Feature Analysis. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 119–123. [CrossRef]
7. Leng, X.; Ji, K.; Yang, K.; Zou, H. A Bilateral CFAR Algorithm for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1536–1540. [CrossRef]

8. Zhang, X.; Crisci, R.; Finlay, J.A.; Cai, H.; Clare, A.S.; Chen, Z.; Silberstein, M.N. Enabling Tunable Water-Responsive Surface Adaptation of PDMS via Metal–Ligand Coordinated Dynamic Networks. *Adv. Mater. Interfaces* **2022**, *9*, 2200430. [CrossRef]
9. Zhang, T.W.; Zhang, X.L.; Li, J.W.; Xiao, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
10. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22 July 2017.
11. Ai, J.; Qi, X.; Yu, W.; Deng, Y.; Liu, F.; Shi, L. A New CFAR Ship Detection Algorithm Based on 2-D Joint Log-Normal Distribution in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 806–810. [CrossRef]
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June 2016.
13. Wang, C.; Bi, F.; Zhang, W.; Chen, L. An Intensity-Space Domain CFAR Method for Ship Detection in HR SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 529–533. [CrossRef]
14. Pappas, O.; Achim, A.; Bull, D. Superpixel-Level CFAR Detectors for Ship Detection in SAR Imagery. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1397–1401. [CrossRef]
15. Li, T.; Liu, Z.; Xie, R.; Ran, L. An Improved Superpixel-Level CFAR Detection Method for Ship Targets in High-Resolution SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 184–194. [CrossRef]
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June 2016.
18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11 October 2016. [CrossRef]
19. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2004**, arXiv:2004.10934.
20. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017. [CrossRef]
21. Zhou, L.-Q.; Piao, J.-C. A Lightweight YOLOv4 Based SAR Image Ship Detection. In Proceedings of the 2021 IEEE 4th International Conference on Computer and Communication Engineering Technology (CCET), Beijing, China, 13 August 2021. [CrossRef]
22. Chang, Y.-L.; Anagaw, A.; Chang, L.; Wang, Y.C.; Hsiao, C.-Y.; Lee, W.-H. Ship detection based on YOLOv2 for SAR imagery. *Remote Sens.* **2019**, *11*, 786. [CrossRef]
23. Zhang, L.; Liu, Y.; Zhao, W.; Wang, X.; Li, G.; He, Y. Frequency-Adaptive Learning for SAR Ship Detection in Clutter Scenes. *IEEE Trans. Geosci. Remote Sens.* 2023; *early access*. [CrossRef]
24. Wei, S.; Su, H.; Ming, J.; Wang, C.; Yan, M.; Kumar, D.; Shi, J.; Zhang, X. Precise and Robust Ship Detection for High-Resolution SAR Imagery Based on HR-SDNet. *Remote Sens.* **2020**, *12*, 167. [CrossRef]
25. Sun, Z.; Dai, M.; Leng, X.; Lei, Y.; Xiong, B.; Ji, K.; Kuang, G. An Anchor-free Detection Method for Ship Targets in High-Resolution SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7799–7816. [CrossRef]
26. Mao, Y.; Yang, Y.; Ma, Z.; Li, M.; Su, H.; Zhang, J. Efficient Low-Cost Ship Detection for SAR Imagery Based on Simplified U-Net. *IEEE Access* **2020**, *8*, 69742–69753. [CrossRef]
27. Zhou, L.; Wei, S.; Cui, Z.; Fang, J.; Yand, X.; Ding, W. Lira-YOLO: A Lightweight Model for Ship Detection in Radar Images. *J. Syst. Eng. Electron.* **2020**, *31*, 950–956. [CrossRef]
28. Bai, L.; Yao, C.; Ye, Z.; Xue, D.; Lin, X.; Hui, M. A Novel Anchor-Free Detector Using Global Context-Guide Feature Balance Pyramid and United Attention for SAR Ship Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4003005. [CrossRef]
29. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2033. [CrossRef]
30. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 14 September 2018.
31. Cui, Z.; Li, Q.; Cao, Z.; Liu, N. Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8983–8997. [CrossRef]
32. Zhao, Y.; Zhao, L.; Xiong, B.; Kuang, G. Attention Receptive Pyramid Network for Ship Detection in SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2738–2756. [CrossRef]
33. Yang, X.; Zhang, X.; Wang, N.; Gao, X. A Robust One-Stage Detector for Multiscale Ship Detection with Complex Background in Massive SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5217712. [CrossRef]
34. Bai, L.; Yao, C.; Ye, Z.; Xue, D.; Lin, X.; Hui, M. Feature Enhancement Pyramid and Shallow Feature Reconstruction Network for SAR Ship Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1042–1056. [CrossRef]
35. Guo, Q.; Tang, F.; Rodrigues, T.K.; Kato, N. Five Disruptive Technologies in 6G to Support Digital Twin Networks. *IEEE Wirel. Commun.* 2023; *early access*. [CrossRef]
36. Geudtner, D.; Gebert, N.; Tossaint, M.; Davidson, M.; Heliere, F.; Traver, I.N.; Furnell, R.; Torres, R. Copernicus and ESA SAR missions. In Proceedings of the 2021 IEEE Radar Conference (RadarConf21), Atlanta, GA, USA, 7–14 May 2021; pp. 1–6.

37. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds. *Remote Sens.* **2019**, *11*, 765.
38. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
39. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
40. Rodrigues, T.K.; Kato, N. Deep Q Networks with Centralized Learning Over LEO Satellite Networks in a 6G Cloud Environment. In Proceedings of the 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 4–8 December 2022. [CrossRef]
41. Li, J.; Chen, J.; Cheng, P.; Yu, Z.; Yu, L.; Chi, C. A Survey on Deep-Learning-Based Real-Time SAR Ship Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3218–3247. [CrossRef]
42. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22 October 2017.
43. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6 September 2014. [CrossRef]
44. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Salt Lake City, UT, USA, 19 July 2018.
45. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
46. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]