remote sensing

MDPI

Article

Synergistic Attention for Ship Instance Segmentation in SAR Images

Danpei Zhao [1,2,3,*], Chunbo Zhu [1,2,3], Jing Qi [4], Xinhu Qi [5], Zhenhua Su [4] and Zhenwei Shi [1,2,3]

1 Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China; sy1915228@buaa.edu.cn (C.Z.); shizhenwei@buaa.edu.cn (Z.S.)
2 Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China
3 Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technologies, Ministry of Education, Beijing 100191, China
4 DFH Satellite Co., Ltd., Beijing 100094, China; 15151122@buaa.edu.cn (J.Q.); 15151124@buaa.edu.cn (Z.S.)
5 Space Star Technology Co., Ltd., Beijing 100094, China; liumengting@buaa.edu.cn
* Correspondence: zhaodanpei@buaa.edu.cn

Abstract: This paper takes account of the fact that there is a lack of consideration for imaging methods and target characteristics of synthetic aperture radar (SAR) images among existing instance segmentation methods designed for optical images. Thus, we propose a method for SAR ship instance segmentation based on the synergistic attention mechanism which not only improves the performance of ship detection with multi-task branches but also provides pixel-level contours for subsequent applications such as orientation or category determination. The proposed method—SA R-CNN—presents a synergistic attention strategy at the image, semantic, and target level with the following module corresponding to the different stages in the whole process of the instance segmentation framework. The global attention module (GAM), semantic attention module (SAM), and anchor attention module (AAM) were constructed for feature extraction, feature fusion, and target location, respectively, for multi-scale ship targets under complex background conditions. Compared with several state-of-the-art methods, our method reached 68.7 AP in detection and 56.5 AP in segmentation on the HRSID dataset, and showed 91.5 AP in the detection task on the SSDD dataset.

Keywords: synergistic attention; ship instance segmentation; SAR images; feature extraction; feature fusion

Citation: Zhao, D.; Zhu, C.; Qi, J.; Qi, X.; Su, Z.; Shi, Z. Synergistic Attention for Ship Instance Segmentation in SAR Images. *Remote Sens.* **2021**, *13*, 4384. https://doi.org/10.3390/rs13214384

Academic Editor: Dusan Gleich

Received: 17 September 2021
Accepted: 28 October 2021
Published: 30 October 2021

1. Introduction

SAR can eliminate the effects of complex weather, working time limits and flight altitude in on-ground observation. Considering the high-resolution and vast extent characteristics of SAR images, ship detection with SAR images has a unique advantage in marine safety monitoring and marine resources development compared to other methods [1]. Instance segmentation combines semantic segmentation with object detection, capable of locating the edges of instances. This has made the ship instance segmentation with SAR data become a significant research direction in the field of remote sensing applications.

Traditional ship detection algorithms for SAR images are mainly composed of spectral residual (SR) [2], constant false alarm rates (CFAR) [3], as well as the improved algorithms derived from them. However, these methods are highly dependent on the statistical characteristics of the targets and background. In recent years, convolutional neural network (CNN)-based algorithms show high robustness and efficiency. There have been some representative SAR image ship detection datasets such as SSDD [4] as well as object detection methods such as the improved Faster R-CNN [4] and G-CNN [5], which will be introduced in the following section.

As a refined task of object detection, instance segmentation can further segment the specific contour of the target on the basis of determining the position. This is of great

significance for determining detailed information. Before the advent of CNN, the combination of detection and segmentation has always been a challenging task. Especially in SAR images, compared with ordinary detection, the instance segmentation of ship targets is more conducive to obtaining target information and analyzing a target situation in the field of marine monitoring. Therefore, this paper aimed to design a method for the task of ship instance segmentation in SAR images. On the one hand, we planned to provide pixel-level contour information for the detected ship target through the instance segmentation task, which helps researchers further judge the target orientation and even the category—thereby expanding the application field of SAR images. On the other hand, as a multi-task learning method, an instance segmentation framework can use pixel-level supervision information to further improve the performance of ship detection tasks.

The first method proposed for instance segmentation with CNN is Mask R-CNN [6], which adds a segmentation branch to Faster R-CNN [7] to generate the pixel-to-pixel mask. Some instance segmentation detectors have been proposed after Mask R-CNN, such as the Hybrid Task Cascade [8] for accuracy and YOLACT [9] for speed. Based on this paradigm, the success of the instance segmentation method depends on three key aspects: (1) whether the CNN can extract features with a higher degree of representation according to the overall style attributes of the images; (2) whether to adopt a reasonable fusion method using high-level semantic features and low-level detailed features to build a feature pyramid network [10]; and (3) whether the target suspected area extracted for classification and regression is accurate. However, corresponding to the above three aspects, it has been found that existing methods have three major barriers for SAR images, which are described as follows:

- Image-level barrier: Compared with optical images, the color information of the SAR images is lacking, and the target shape is blurred. Furthermore, the speckle noise caused by the imaging mechanism can also interfere with visual effects. The inability to extract features with a sufficiently high degree of characterization is a basic barrier to the subsequent tasks.
- Semantic-level barrier: Low-level detail features of ship targets in SAR images is not as obvious as it is in optical images, which requires more use of high-level semantic features to guide target locating. When fusing the features at different levels, existing methods do not make full use of the guiding significance of semantic features and the result is not sufficiently representative.
- Target-level barrier: Limited by the imaging method and resolution, the targets in the SAR image are relatively small accompanied by complicated sea clutter interference. When traversing the image with dense anchors, it will introduce a large number of negative samples which represent a barrier to the locating process.

To mitigate the adverse effects caused by these issues, we proposed an effective framework for instance segmentation with the attention mechanism, which can be regarded as a resource allocation mechanism to reallocate resources according to the importance of the attention object for the originally evenly allocated resources [11]. Our method consists of three different attention modules to break through the barriers at all three levels discussed above. We proposed the Synergistic Attention R-CNN (SA R-CNN), which integrates three novel components: (1) global attention module (GAM), aimed at the image level, which adds a self-attention [12] module to the basic block of the residual network [13] that can strengthen the effective features; (2) semantic attention module (SAM), aimed at the semantic level, which uses the upper-level semantic feature as the attention mask of the next-level detail feature to obtain a feature pyramid weighted by different semantic information; and (3) anchor attention module (AAM), aimed at the target level, which guides the anchor position according to the distribution of the targets in the feature pyramid and guides the anchor size according to the scale distribution of the targets in the training dataset. Without bells and whistles, our method achieved 2.9 and 1.8 points higher AP in the ship detection and segmentation task, respectively, than several state-of-the-art methods on the HRSID dataset [14].

Our main contributions can be summarized as follows:

- We constructed a synergistic attention mechanism with three new components to propose the SA R-CNN for SAR images, which is an effective instance segmentation framework for ships in diverse scales and complex distribution.
- We proposed the GAM, SAM, and AAM focusing on different stages of the instance segmentation framework to construct attention modules which optimized our method from imaging methods and target characteristics, respectively.
- We tested the proposed method on the HRSID and SSDD [4] datasets for both instance segmentation and ship detection tasks. The results have shown significant improvements over state-of-the-art methods.

2. Related Works

In this section, we will briefly introduce the traditional ship detection methods in SAR images; ship detection methods in SAR images with CNNs; instance segmentation based on deep learning; and the attention mechanism in computer vision. These studies have greatly contributed to our method.

2.1. Instance Segmentation Based on Deep Learning

Instance segmentation is significant to SAR images as it combines semantic segmentation with object detection. Based on locating the ship targets, semantic segmentation divides each pixel of the input image into a semantically interpretable category. As a more complex interpretation method, it has a finer description and perception of the ship targets. As the first attempt of instance segmentation applied on CNN, Mask R-CNN [6] adds a mask branch to predict the segmentation mask for each region of interest (RoI), paralleling the classification and regression branch in Faster R-CNN. Mask Scoring R-CNN [15] uses the product of classification score and IoU score of the mask to define the mask score in order to improve the instance quality. Cascade Mask R-CNN [8] is the hybrid of the Mask R-CNN and Cascade R-CNN. Each cascade structure adds a mask branch to finish the instance segmentation task, which combines the excellent characteristics of the two methods. To improve the detection precision, Hybrid Task Cascade [8] proposes an association between the parallel structure of detection and segmentation, which uses semantic segmentation branches to provide a spatial context for the bounding box.

Corresponding to object detection methods, some one-stage methods for instance segmentation have also appeared in recent years such as YOLACT [9] and SOLO [16]. In addition, there are some methods based on an anchor-free object detection framework, such as PolarMask [17] and BlendMask [18]. These one-stage methods have been widely used in the fields of autonomous driving and face recognition due to their speed advantages. However, in some refined ship detection tasks, the length and contour of ships cannot be measured by detection method only when this information is important for the type of ships. Recently, a dataset of ship instance segmentation in SAR images called HRSID [10] has been proposed, but there is currently no representative work to modify the existing instance segmentation methods for SAR images.

2.2. Attention Mechanism in Computer Vision

The attention mechanism is a method that can strengthen important information and suppress non-important information. In the field of deep learning, the essence of attention is a set of weighted coefficients that are independently learned through the network and use a dynamic weighting method to emphasize the area of interest while suppressing the irrelevant background area. In the field of computer vision, the attention mechanism can be roughly divided into hard attention and soft attention. The former is a kind of random prediction, which emphasizes dynamic changes, but its application is very limited due to its non-differentiable nature. On the contrary, the latter can be obtained by neural network training based on the gradient descent method. At present, the mainstream

attention mechanism can be divided into the three following types: channel attention, spatial attention, and self-attention.

Channel attention aims to assign the weighted coefficients to different channels to strengthen important features through network learning. The representative work in this area is SENet [19], which adaptively adjusts the characteristic response between channels through feature recalibration. In addition, in SKNet, [20] introduces multiple convolution kernel branches to obtain feature map attention at different scales. ECANet [21] uses sparse convolution operations to optimize the fully connected layers and reduce the amount of parameters. Spatial attention aims to improve the feature expression of key regions and generates weight masks for each position and weights the output. Among the outstanding works in this area are CBAM [22] and BAM [23], which are based on the original channel attention and are connected to a spatial attention module. Self-attention is a variant of the attention mechanism. Its purpose is to reduce the dependence on external information and use the inherent information within the feature to interact with attention as much as possible. It first appeared in the transformer architecture [11] proposed by Google. He et al. applied it to the computer vision field and proposed the non-local module [24] which uses the self-attention mechanism to model the global context information, effectively capturing long-distance feature dependence. In addition, CCNet [25] develops two crossed attention modules to equivalently replace attention modeling based on global pixel point pairs to reduce computational complexity; GCNet [12] combines SENet and proposes the use of a simplified spatial attention module to replace the original spatial downsampling process.

Based on the above research, this paper further proposes three attention modules for the ship instance segmentation task in SAR images. The global attention module (GAM) uses self-attention to further explore the correlation between features in the residual network. The semantic attention module (SAM) adopts a construction mode similar to channel attention, generates channel weights according to semantic features and weights the detail features to realize the feature fusion driven by semantic attention. The anchor attention module (AAM), based on the generalized perspective of attention, generates anchors according to the target position and size distribution, which is more efficient than the traditional way of designing it manually. The details are explained below.

2.3. Traditional Ship Detection Methods in SAR Images

Traditional ship detection algorithms for SAR imageries are mainly composed of spectral residual (SR) [2], constant false alarm rates (CFAR) [3], and the improved algorithms derived from them. For specific needs, M.L. Yeremy et al. provided high ship detection rates with polarimetric decomposition methods [26]. Y. Liu et al. designed an automatic ship detection system for both the ship and ship wake detection [27]. Sugimoto et al. proposed two different ship detection methods using different scattering mechanisms between the ship and sea [28]. However, the setting of modules' parameters often comes from experience which lacks some necessary theoretical support. Furthermore, these traditional ship detection methods always rely on complex feature design. In other words, its migration ability is weak, so it is not flexible and intelligent enough which provides great limitations both for speed and accuracy.

2.4. Ship Detection Methods in SAR Images with CNNs

The remarkable characteristic of modern deep learning methods is automatic feature extraction. Recently, object detection based on data-driven and artificial intelligence (AI) methods has been popularized by both two-stage and one-stage detectors. On the one hand, two-stage detectors were first introduced in R-CNN [29]. Gradually derived SPP [30], Fast RCNN [31] and Faster R-CNN [7] further promoted these developments, and proposed region proposal network (RPN) to improve the efficiency of detectors and train detectors end to end. Based on this, researchers also proposed many optimization algorithms from different angles. FPN [10] tackles the scale variance with a feature pyramid. Cascade R-

CNN [32] extends the Faster R-CNN to a multi-stage detector through a powerful cascade architecture. Libra R-CNN [33] proposes three solutions to the three imbalance problems in the object detection algorithm to further improve the accuracy. Based on Cascade R-CNN, Dynamic R-CNN [34] upgrades different thresholds of multiple network heads to adaptive thresholds, which further improves the performance while reducing complexity. On the other hand, some representative one-stage methods, such as YOLO [35–37], Retinanet [38] and SSD [39], do not need to extract the effective region and directly locate the target on the feature map, which makes the detection process simpler and faster.

Considering the powerful feature representation capabilities of deep learning methods, in recent years, many scholars have used these for object detection tasks in SAR images. Similarly to the classification of object detection methods, studies in this research area also have two major directions—those based on two-stage and one-stage detectors. For the former, Kang et al. combined the advantages of CFAR and Faster R-CNN using region proposals as guard windows to detect small targets [40]. SLS-CNN uses the prior results of land–ocean segmentation and the saliency heat map to improve the ship detection performance in SAR images [41]. In addition, the HRFPN structure connects some sub-networks at different resolutions, from high to low, whilst simultaneously realizing precise and robust ship detection in high-resolution SAR images [42]. Based to the above research and the YOLO detection framework, Zhang et al. proposed the grid convolutional neural network (G-CNN) for real-time SAR ship detection [5]. Wang et al. made improvements to Retinanet based on the characteristics of the SAR ship detection task and achieved effective results [43]. Inspired by the research of the lightweight network, Zhang et al. proposed DS-CNN and used depth-wise separable convolution to realize high-speed SAR ship detection [44]. Recently, Guo et al. [45] and Fu et al. [46] applied the latest anchor-free detection method to SAR images and achieved valuable results. Geng et al. [47] made further research on the lightweight nature of the network. Wu et al. [48] and Albuquerque et al. [49] further explored the instance segmentation task of SAR images.

In summary, CNN-based methods achieve considerable results in ship detection in SAR images. However, there are two main areas of work that need to be further resolved. One is how to systematically integrate the advanced achievements in computer vision to fill the gap between the SAR and optical images. The other is mainly aimed at extending ship detection to more applications such as instance segmentation. The two aspects were combined to yield further value from SAR images in the field of remote sensing applications, which also became the topic of this paper.

3. Methods

The overall pipeline of SA R-CNN is shown in Figure 1. The synergistic attention instance segmentation method is mainly based on a two-stage object instance segmentation framework. In the feature extraction phase, the global attention module (GAM) was constructed in the method's backbone. The anchor attention module (AAM) was used to guide the anchor generation of the feature pyramid composed of the semantic attention module (SAM) to select candidate regions. The above parts constitute the synergistic attention mechanism corresponding to different stages of the instance segmentation framework.

Specifically, we used ResNet50 [13] as the feature extraction network, and introduce GAM in its last four stages. Then, the fifth stage has the highest level of semantic information as the deepest feature, and directly serves as the highest level of the feature pyramid. SAM separately merges adjacent stages with semantic attention-guided features, and obtains three feature maps as the remaining three layers of the feature pyramid. Finally, AAM guides the generation of the possible target area to complete the first half of the two-stage framework.

As for the rest of the two-stage framework, the network header part used to complete the instance segmentation task mainly contains the ROI Aligning and Dynamic IoU. The main contribution of the former is to use a bilinear interpolation strategy to optimize the relevant rounding operation. In this way, a more accurate area is obtained to improve

the detection performance for small targets and the method can be applied to segmentation with higher accuracy requirements. Correspondingly, the latter designs a cascade strategy to dynamically adjust the IoU threshold according to the training process. A relatively low IoU threshold should be used to compensate for a few high-quality positive samples in the early stage of the training process and be slowly increased to keep generating proposals with high quality.

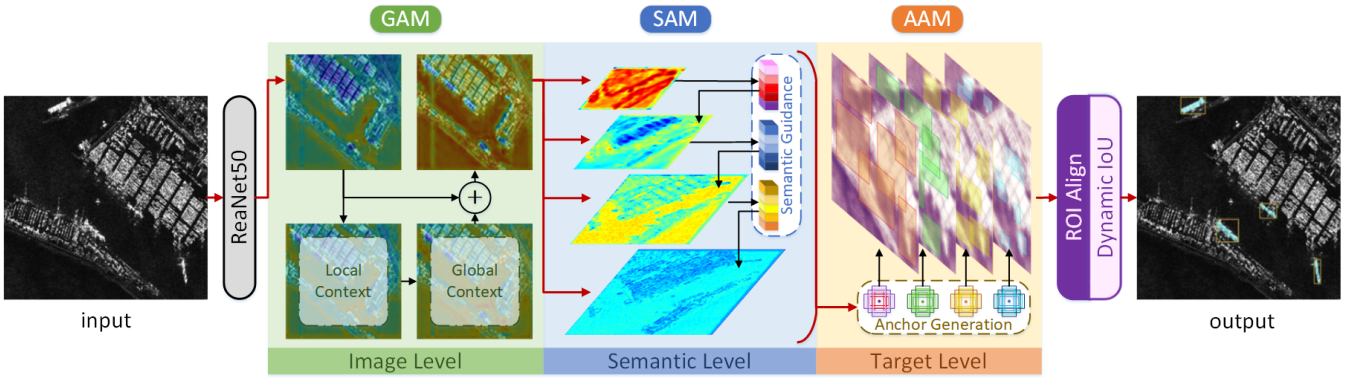

Figure 1. Overview of the proposed SA R-CNN: an overall synergistic attention design for instance segmentation which integrates three novel components: global attention module (GAM); semantic attention module (SAM) and anchor attention module (AAM), respectively, for the barriers in ship instance segmentation in SAR images at the image, semantic, and target levels.

The above is the basic framework of the SA R-CNN. The purpose of our research was to make the instance segmentation method perform better on ship targets in SAR images with the synergistic attention mechanism, thus exploiting the potential of model architectures as much as possible. All components will be detailed in the following sections.

3.1. Global Attention Module (GAM)

At the image level, we hope that the backbone of our method can obtain a more representative feature to guide the subsequent instance segmentation tasks. Image feature can generally be enhanced by two avenues: external and internal. From the external view, SAR images show limited visual features, and the details of the ship target are weak. Therefore, it is difficult to directly transfer the image to obtain more effective features. In contrast, SAR images are more suitable for using the attention mechanism to explore the correlation of features in different dimensions from the internal view of the feature map. Based on the above analysis, we propose the global attention module (GAM). In the residual backbone ResNet50 [13], a global attention block is used to replace the original residual block. The specific difference between them is shown in Figure 2.

The traditional residual block uses convolution (*Conv*) with a kernel size of 1×1 to reduce and expand the feature map dimension so that the number of filters of the convolution with the kernel size of 3×3 is not affected by the input of the previous layer, and its output will not affect the next level. In addition, the residual structure allows the network to achieve better results with a faster convergence speed and a deeper effective number of layers. The variable x in our manuscript means the feature map of the input, and the residual block can be defined as

$$\mathbf{y} = Conv_{1\times1}(Conv_{3\times3}(Conv_{1\times1}(\mathbf{x}))) + \mathbf{x} \quad (1)$$

Compared with the traditional residual block, our global attention block has two components, which can be defined as

$$\mathbf{y} = GCblock(Conv_{1\times1}(DeConv_{3\times3}(Conv_{1\times1}(\mathbf{x})))) + \mathbf{x} \quad (2)$$

In order to take advantage of the short-distance dependence of the features, we replaced the convolution with a kernel size of 3 × 3 in the residual block with deformable convolution (*DeConv*) [50], which provided additional spatial sampling information for the residual block with some offsets to the convolution operation. Based on the above, for the long-distance dependence of features, the feature map with spatial sampling information passes through the global context block (*GCblock*) [12] to be weighted by self-attention. In detail, the global attention pooling in the non-Local block [24] was used here to model the contextual information to obtain spatial dependence. Then, a 1 × 1 convolution and layer normalization was jointly used to form channel dependence with a lower amount of calculation and better generalization. Finally, we used the broadcast element-wise addition to fuse the self-attention and input features. At this point, the feature map in the residual block contains long-distance context information.

As for the ship targets in the SAR images, convolution can only perform context modeling on local areas which leads to limited receptive fields. The global context block actually performs context modeling and covers the receptive field on the entire input feature. Therefore, it is said to be a useful supplement of semantic information. In addition, if the network only uses convolutional stacking to extract features, it can be considered that the same form of function is used to fit the input, resulting in a lack of diversity in the extracted features. The global context block simply increases the diversity of extracted features and compensates for the lack of color and shape details for the SAR images.

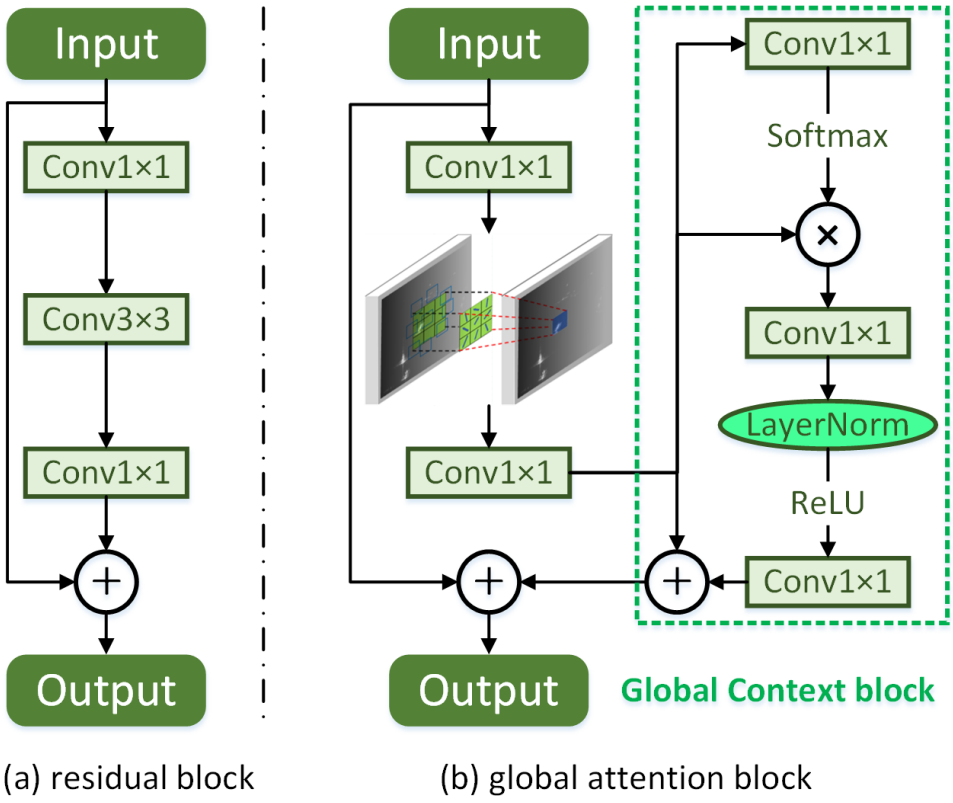

Figure 2. The basic module unit in the traditional residual network and the residual structure with the introduction of the global attention module.

3.2. Semantic Attention Module (SAM)

At the semantic level, SAR images, especially the ship targets, have the following main characteristics compared with optical images. Due to the wide variety of ship targets, there are also large size differences in the same source data. The feature pyramid [10] can fuse the feature maps of different resolutions, which plays an important role in ship detection

tasks. However, after the high-level semantic information is upsampled, the element-wise addition fusion method does not make full use of the guiding significance of the semantic features. Especially in SAR images, the detailed features of the ship are relatively weak in the complex environment of the port which requires supplementary guidance from semantic features. Therefore, we propose the semantic attention module (SAM) to fuse features in different resolutions, which can be defined as

$$\mathbf{y} = \mathbf{x}_{low} \cdot Att(\mathbf{x}_{high}) + Upsample(\mathbf{x}_{high}) \tag{3}$$

Compared with the element-wise addition method in FPN, SAM first processes high-level semantic features ($\mathbf{x}_{high}$) and learns a set of weights through an attention module (Att). It then uses the semantic features to obtain the attention relationship between the different channels of the feature map, and weights the detailed features ($\mathbf{x}_{low}$) according to the corresponding channels. This way, detailed features are weighted by semantic attention before fusion, which contains the semantic information and high resolution at the same time. In particular, the formation method of semantic attention can be defined as

$$Att(\mathbf{x}) = \sigma(W_1(W_0(P_{max}(\mathbf{x}))) + W_1(W_0(P_{avg}(\mathbf{x})))) \tag{4}$$

The specific process is shown in Figure 3. First of all, we use both max-pooling (P_{max}) and average-pooling (P_{avg}), respectively, to aggregate the spatial context of a feature map [22]. Both descriptors were then forwarded to produce the channel attention map with a multi-layer perceptron with one hidden layer, when the weights of the shared network, W_0 and W_1, were shared for both inputs. Finally, we merged the output feature vectors by element-wise addition after the shared network is applied to each descriptor and the feature vectors will become the channel attention weight through the sigmoid function. As a result, the low-level features in high resolution were weighted and semantically guided in the channel domain—before being merged with the upsampling high-level features. Multi-scale information and semantic features cooperate with each other to compensate for the lack of detailed features of SAR images; thus, the feature pyramid formed by SAM will be more representative.

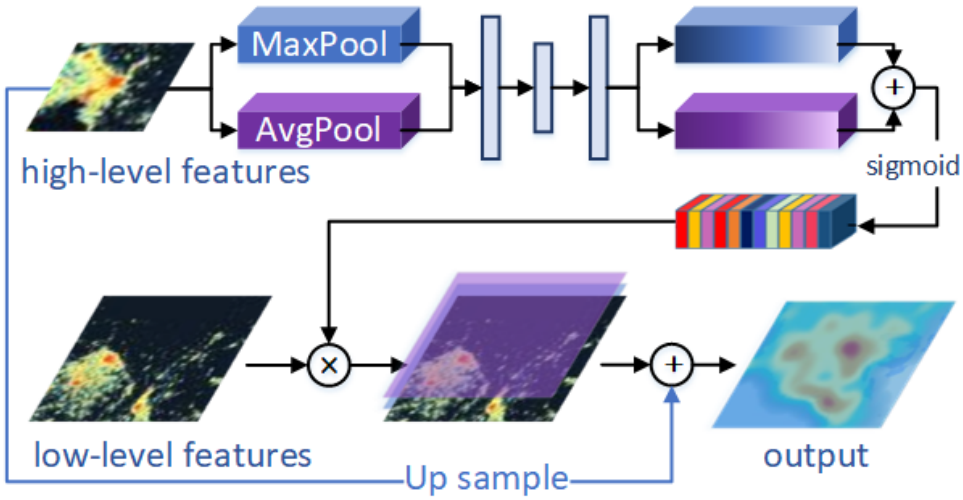

Figure 3. Fusion process of features with different resolution in SAM.

3.3. Anchor Attention Module (AAM)

At the target level, the anchor mechanism is the cornerstone of mainstream deep learning-based object detection methods. However, it has limitations in detection and instance segmentation in SAR images due to the different scales of the ship targets and clutter interference in radar imaging. When traversing the image with dense anchors,

backgrounds such as ports and islands will generate a large number of negative samples. Studies have shown that the anchor mechanism can be more effectively and efficiently implemented [51]. As local attention factors, the semantic features based on the image feature pyramid are used to guide the anchor and constitute the anchor attention module (AAM).

AAM automatically determines the location and shape of the anchor rather than via manually configuring the anchor settings. The positioning of the anchor is to generate a probability map of the same size as the input feature map, where the probability of each feature point at the corresponding position is related to the coordinate position of the target on the original image. First, we used a clustering algorithm to determine the size of the anchor. The K-means [52] algorithm, based on the similarity of the distance between points to calculate the best category attribution, needs to calculate the respective distances between each object and the k-centers, and assign them to the nearest cluster according to the principle of minimum distance. In addition, we used the following distance function to modify the k-means algorithm based on intersection-over-union (IoU), which can be defined as

$$distance(\mathbf{box}, \mathbf{centre}) = 1 - IoU(\mathbf{box}, \mathbf{centre}) \quad (5)$$

This approach makes the prior anchor size more in line with the distribution of the target in the dataset, preventing the adjustment of related hyperparameters. In addition, the position distribution of the anchor on the feature map can also be further optimized. The specific method is to perform a 1×1 convolution after each layer of the feature pyramid to obtain an anchor pyramid with the same resolution as the feature pyramid. We matched the feature map to the ground truth of the target on the anchor pyramid, so as to obtain the response heat map of the target on the feature map. This way, an attention mask based on the ground truth position can be formed to guide the anchor to be more distributed in the position where the target feature response is strong.

Compared with manually setting anchors, AAM has two advantages for object detection. One is that there are more positive proposals, and the other is that there are more proposals with a high IoU. The intuitive difference between them is shown in Figure 4. This way, using the anchor attention module (AAM) to guide the choice of anchor—instead of manually setting them—makes the method more sensitive to the object's area of interest and more robust to background clutter.

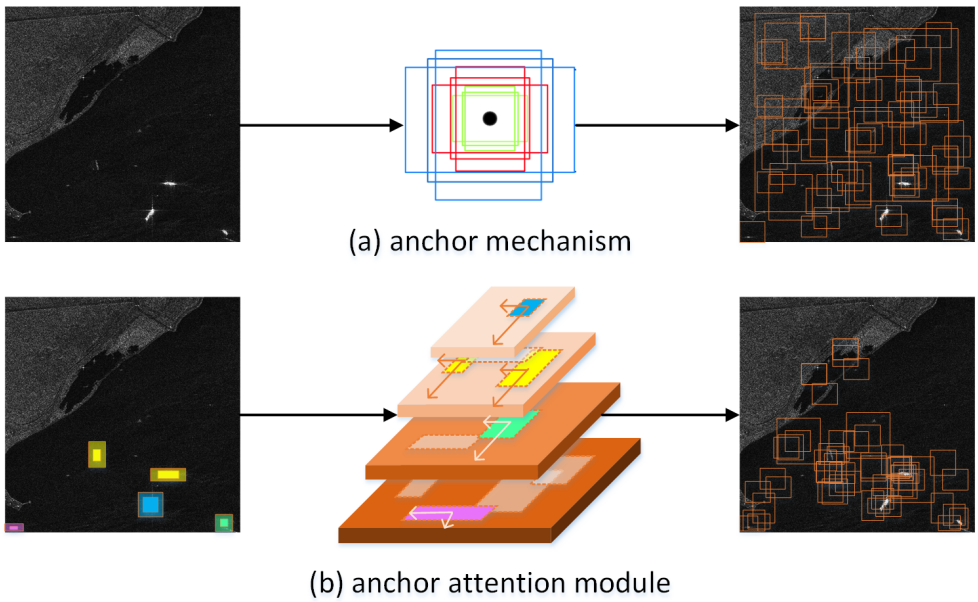

Figure 4. The difference between using artificial anchor settings and feature-based anchor generation with AAM.

4. Experiments

4.1. Dataset and Evaluation Metrics

(1) HRSID [14]: The High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation (HRSID) has 5604 cropped SAR images and 16951 ships from 136 panoramic SAR imageries with a resolution ranging from 1 m to 5 m. The images were cropped to 800 × 800 pixels in the overlapped ratio of 25%. The dataset was divided into a training set (65%) and test set (35%) with the format of MS COCO [53]. In line with reality, it is a challenging dataset with ship targets distributed in large areas of sea and complex coastal areas.

(2) SSDD [4]: The SAR Ship Detection Dataset (SSDD) is a large-scale dataset which consists of 1160 images. Quantitatively, there are 2456 labeled ships with resolutions from 1 m to 15 m which are mainly imaged by RadarSat-2, TerraSAR-X and Sentinel-1 sensors. The diverse scales and resolutions, different berthing areas, sparse distribution, and densely arranged shores make the SSDD a challenging and practical dataset for SAR ship detection task. The dataset is randomly split into a training (70%), validation (10%), and a test dataset (20%) according to the format of PASCAL VOC [54].

We measured the performance by the mean average precision (mAP). The calculation criterion of mAP for the Pascal VOC dataset was based on an IoU threshold of 0.5, while the evaluation metrics in MS COCO are abundant and comprehensive. AP in the MS COCO format is the primary challenge metric with the calculation of the average IoU which has ten IoU thresholds distributed from 0.5 to 0.95. In order to verify the detection performance of the method on multi-scale targets, MS COCO also has evaluation metrics for different scales: APs (area < 32^2 pixels), APm (32^2 pixels < area < 96^2 pixels), and APl (area > 96^2 pixels). Therefore, we characterized the performance of instance segmentation on HRSID according to the format of MS COCO. Correspondingly, we used an AP with an IoU threshold of 0.5 and the recall metric on SSDD according to the format of PASCAL VOC.

4.2. Implementation Details

To quantitatively evaluate the model, we implemented the proposed method with the PyTorch toolbox and mmdetection [55]. All experiments were performed on four NVIDIA 2080Ti GPU for 12 epochs with an initial learning rate of 0.001, which decreased by 0.1 after 8 and 11 epochs, respectively, if not specifically noted. We chose SGD as the optimizer and the momentum and weight decay were set to 0.9 and 0.0001. All other hyperparameters followed the settings in mmdetection if not specifically noted otherwise.

4.3. Ship Instance Segmentation Results in SAR Images

We implemented our model and five other representative instance segmentation methods on the HRSID dataset. To control the variables, all methods used Resnet-50 and FPN as the backbone. As shown in Tables 1 and 2, we exhibited the AP in MS COCO format for both ship detection and instance segmentation under the same running environment. Comparatively, our model SA R-CNN beats all other SOTA methods and achieves the highest AP in two tasks, respectively.

Numerically, compared with the Dynamic Mask R-CNN [34] as the baseline, our method surpasses the ship detection results by 2.9 points AP and surpasses the instance segmentation results by 1.8 points AP. In terms of the performance for small objects, our model achieved approximately 70.2 and 58.7 points APs in two tasks, which was a better performance than that for the other methods. Figure 5 shows our detection performance in some complex situations, which contains the inshore ships with large-scale changes. In the situations shown in Figure 5, the proposed method SA R-CNN does not easily miss small targets and effectively reduces the false alarm on ships with complex background inshore. However, due to the uneven distribution of samples with different shapes in the dataset, our method is not optimal on APl. According to the standard of MS COCO, the target size counted by the APl (area > 96^2 pixels) is met by only 2% targets in HRSID, so the gap is in an acceptable range and can perform better on more robust datasets.

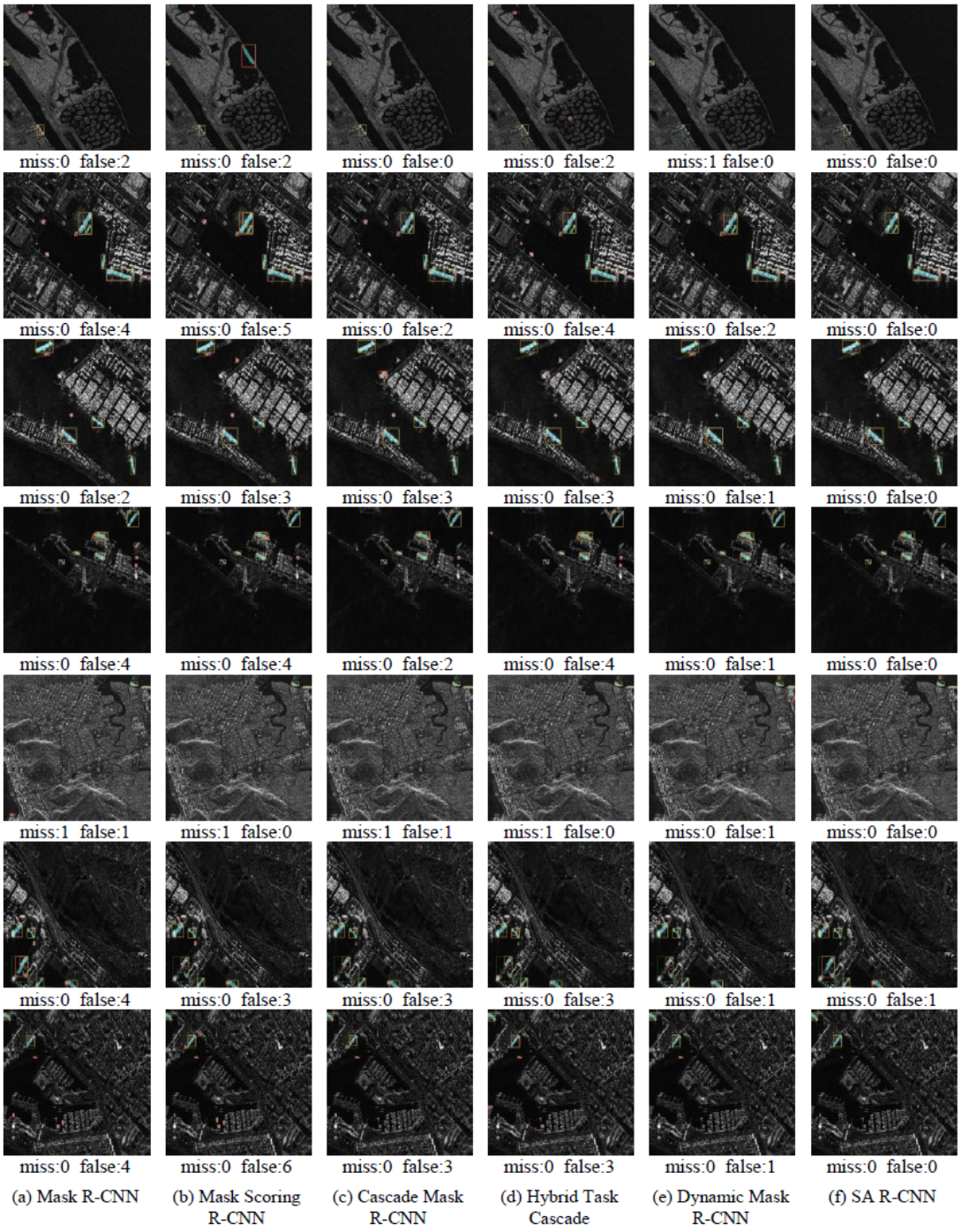

Figure 5. Visible ship instance segmentation results of our SA R-CNN compared with some state-of-the-art methods on complex scenes from the test set of HRSID. The green bounding box denotes the ground truth and the orange bounding box denotes the predicted results. The numbers of failed detections and false alarms are shown below.

Table 1. Detection performance comparison between different instance segmentation methods on HRSID.

Method	AP	AP50	AP75	APs	APm	APl
Mask R-CNN [6]	64.0	87.3	73.2	65.3	65.1	7.7
Mask Scoring R-CNN [15]	64.0	87.3	73.3	65.2	64.9	6.5
Cascade Mask R-CNN [8]	64.6	86.4	74.3	65.9	66.2	8.8
Hybrid Task Cascade [8]	65.1	86.9	74.4	66.1	65.9	9.4
Dynamic Mask R-CNN [34]	65.8	87.0	75.1	67.0	65.1	**13.7**
SA R-CNN (ours)	**68.7**	**90.9**	**78.1**	**70.2**	**67.6**	11.2

Table 2. Segmentation performance comparison between different instance segmentation methods on HRSID.

Method	AP	AP50	AP75	APs	APm	APl
Mask R-CNN [6]	54.2	85.1	64.8	56.3	52.6	4.4
Mask Scoring R-CNN [15]	54.3	84.7	64.6	55.8	**55.7**	2.4
Cascade Mask R-CNN [8]	53.7	84.3	64.2	55.7	53.4	3.7
Hybrid Task Cascade [8]	54.2	85.0	64.8	55.6	54.3	4.9
Dynamic Mask R-CNN [34]	54.7	85.2	65.3	56.4	53.0	**6.8**
SA R-CNN (ours)	**56.5**	**88.4**	**67.9**	**58.7**	54.1	4.2

4.4. Ablation Experiments

4.4.1. Overall Ablation Studies

Taking into account the importance of the synergistic attention mechanism for ship instance segmentation in SAR images, the global attention module (GAM), the semantic attention module (SAM), and the anchor attention module (AAM) were added to Dynamic Mask R-CNN, respectively, in this part. Accuracy with relation to instance segmentation and ship detection in ablation experiments was measured by mask AP and bounding box AP as the format in MS COCO, respectively. Results in Table 3 showed that after introducing the synergistic attention mechanism to the baseline, the segmentation task promotes 1.8 points for the mask AP and the detection task promotes 2.9 points for the bounding box AP. Specifically, the GAM provided a 1.3 point-higher AP for segmentation and 2.1 points-higher AP for detection, while the SAM provided a relatively small increase (0.9 higher AP for segmentation and 1.2 points-higher AP for detection), and the AAM provided the least (0.3 higher AP for segmentation and 0.5 points-higher for detection). Therefore, the three parts of the synergistic attention mechanism have different degrees of efficient improvement in the performance of ship instance segmentation in the SAR images. When they work separately, it is more efficient to use the GAM to improve the feature extraction capabilities of the network or use SAM to fuse the feature maps than to use AAM to generate anchors.

Table 3. Ablation experiment of each attention module proposed in this paper.

Attention	Ship Detection						Instance Segmentation					
	AP	AP50	AP75	APs	APm	APl	AP	AP50	AP75	APs	APm	APl
None	65.8	87.0	75.1	67.0	65.1	13.7	54.7	85.2	65.3	56.4	53.0	6.8
GAM	67.9	90.1	77.5	69.4	**67.9**	7.1	56.0	88.1	66.5	58.1	**54.3**	2.8
SAM	67.0	88.6	76.6	68.1	66.6	13.1	55.6	86.7	67.0	57.6	53.0	5.8
AAM	66.3	87.6	75.7	67.1	67.3	**15.6**	55.0	85.8	65.7	56.9	53.6	**8.6**
All	**68.7**	**90.9**	**78.1**	**70.2**	67.6	11.2	**56.5**	**88.4**	**67.9**	**58.7**	54.1	4.2

4.4.2. Ablation for Different Context Scale in GAM

In this part, we experimentally verify that when the global attention module (GAM) extracts features in ResNet50, both local contexts produced by deformable convolution and global context block play a significant role. In order to contrast these, we performed four experiments in this part, which were: SA R-CNN without GAM; only the local context attention (LCA) was considered; only the global context attention (GCA) was considered; and both of them were used to construct the GAM. As a result, Table 4 illustrates the difference in the effect of the above settings on the task of the ship instance segmentation in SAR images. The experimental results show that, in the ship detection task, LCA and GCA both have 0.8 point improvement in AP, and for instance segmentation, the effects of the two are basically the same (at least 0.6 higher AP) while the effect of GCA is slightly better. This way, our method performs better for targets with a smaller shape. From this perspective, the LCA and GCA modules in GAM start from the local and global contexts, respectively. The evolution of specific features in the backbone ResNet50 is shown in Figure 6.

Table 4. Ablation experiment of context attention in a different scale of GAM.

Context	Ship Detection						Instance Segmentation					
	AP	AP50	AP75	APs	APm	APl	AP	AP50	AP75	APs	APm	APl
None	65.8	87.0	75.1	67.0	65.1	13.7	54.7	85.2	65.3	56.4	53.0	6.8
LCA	66.6	88.5	76.4	67.8	67.4	**13.9**	55.3	86.6	66.1	57.1	**54.4**	**7.2**
GCA	66.6	88.4	76.6	67.9	67.3	11.1	55.5	86.6	66.3	57.6	53.7	4.8
GAM	**67.9**	**90.1**	**77.5**	**69.4**	**67.9**	7.1	**56.0**	**88.1**	**66.5**	**58.1**	54.3	2.8

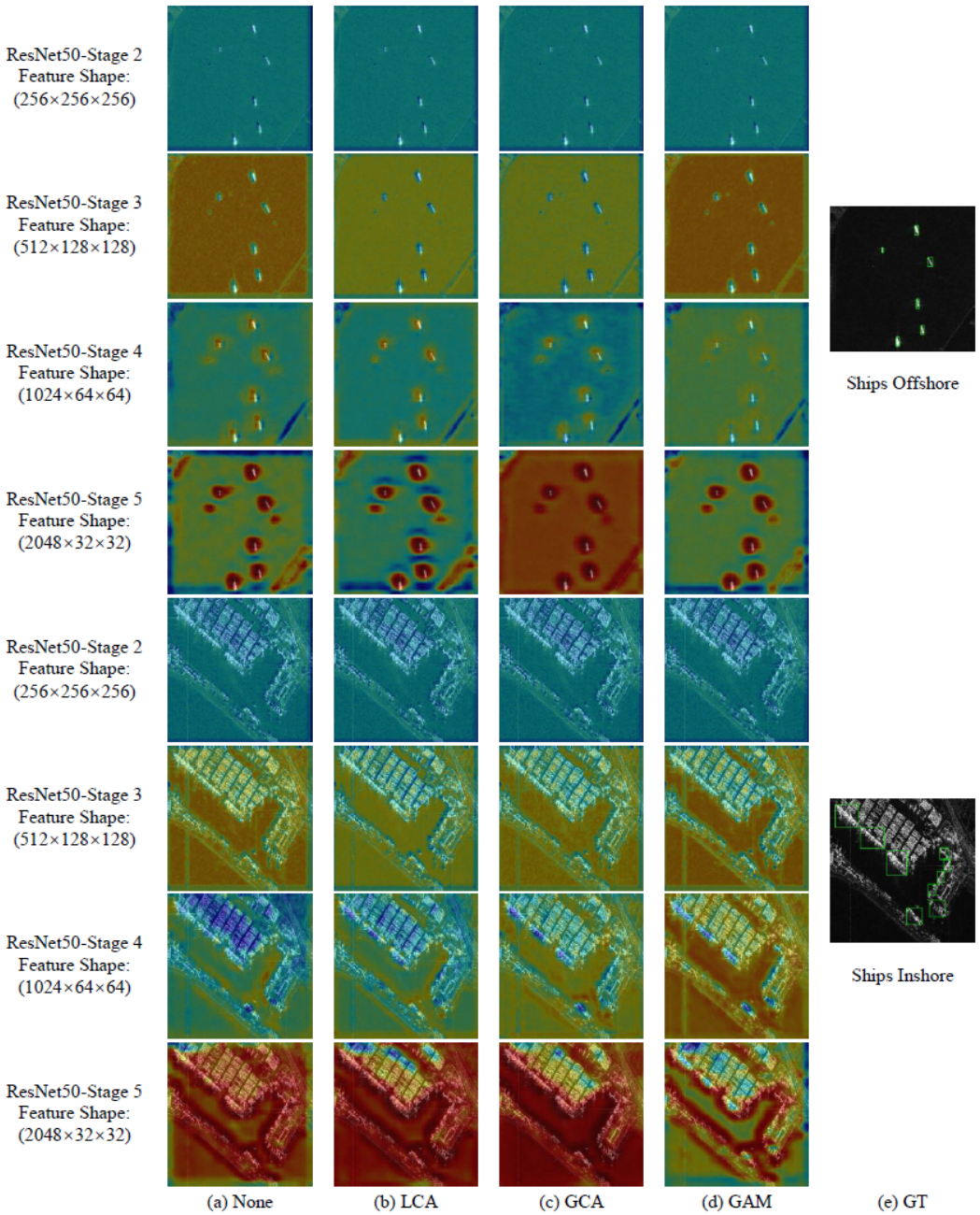

Figure 6. Global attention module (GAM) visualization results for the ships offshore and inshore, respectively. We compared the visualization results of GAM with the baseline ResNet50 and Local Context Attention (LCA) and Global Context Attention (GCA) modules in GAM, respectively. The GAM visualization was calculated for the convolutional outputs from stage 2 to stage 5 in ResNet50 with the feature map in different resolutions, which are used to construct the feature pyramid. The ground-truth (GT) is shown on the right of each input image.

In order to illustrate that GAM has a stronger ability to characterize SAR images at the image level, we provided the visual expression of the characteristics of the two types of ship targets offshore and inshore, respectively, in Figure 6. From the feature extraction process of ResNet50, we extracted the features from stage 2 to stage 5 for visualization which have different resolutions and together form a feature pyramid. The heat map in stage 5 can fully emphasize the location information of the target which is in line with the semantic characteristics of the deep feature. The local context attention will emphasize the local features of the suspected target and suppress the background area, but the global context attention will pay attention to all the positions of the feature map. For the ships offshore, LCA strengthens the characteristic response of the foreground, and GCA highlights the target position from an overall perspective. Similarly, for the ships inshore, GAM can pay more attention to the shore area where the ship is located while ignoring the interference from the content on the shore. The comparison results distinctly illustrate that GAM can effectively improve the network's feature representation ability for SAR images.

4.4.3. Ablation for Mask Branch's Contribution to Ship Detection Task

In order to prove the performance improvement effect of the instance segmentation task on the ship detection task in SAR images, we carried out ablation experiments on Dynamic Mask R-CNN (as the baseline in this paper) and SA R-CNN (our method) with the mask branch removed. The results in Table 5 show that after introducing the mask branch to the baseline Dynamic Mask R-CNN and our SA R-CNN, the ship detection task promotes the 1.2 and 1.1 points bounding box AP, respectively. Especially for APs which evaluate the detection performance of smaller targets, the two methods improved by 1.2 and 1.6 points the bounding box AP, which was also in line with the demonstration in the Mask R-CNN [6] as the instance segmentation task can improve the performance of the detector itself. In comparison with other detection methods, our SA R-CNN also has obvious advantages. This group of ablation experiments shows that there are two meanings for the instance segmentation method of SAR images. In addition to obtaining the outline information for further judging the details of the target, more importantly, a multi-task learning mode can be constructed. On this basis, pixel-level supervision information is introduced which effectively improves the performance of the method for ship target detection in SAR images—especially for densely distributed small targets.

Table 5. Ablation for the mask branch's contribution to the ship detection task.

Method	AP	AP50	AP75	APs	APm	APl
Faster R-CNN [7]	64.2	84.9	73.6	65.3	64.4	11.2
Cascade R-CNN [32]	64.4	85.3	73.5	65.4	63.8	11.0
Libra R-CNN [33]	64.3	85.6	73.7	65.1	65.3	13.6
RetinaNet [38]	63.8	85.1	73.0	65.4	63.7	11.1
SSD [39]	61.3	82.9	71.2	62.4	62.1	7.8
YOLOv3 [37]	63.6	85.7	73.4	65.0	65.5	8.9
Dynamic Mask R-CNN [34] w/o mask	64.6	86.3	74.5	65.8	64.7	13.5
Dynamic Mask R-CNN [34] w mask	65.8	87.0	75.1	67.0	65.1	**13.7**
SA R-CNN (ours) w/o mask	67.6	88.7	77.2	68.6	67.4	11.2
SA R-CNN (ours) w mask	**68.7**	**90.9**	**78.1**	**70.2**	**67.6**	11.2

4.5. Ship Detection Results on SSDD Dataset

Our method was compared with the state-of-the-art detection methods on the SAR ship detection dataset SSDD, and can be better implemented for ship detection. The evaluation index adopts the AP of PASCAL VOC and uses recall reflecting the detection rate of the target in order to facilitate a practical application reference. As shown in Table 6, after adding synergistic attention to the method, our SA R-CNN brings a 3.8 points-higher AP and 2.7 points-higher recall compared with one-stage detection method Retinanet [37]. Further extended to two-stage methods, it achieved 1.3 and 1.2 points-higher AP than

the Faster R-CNN [7]. In addition, we also compared the SA R-CNN with the method specifically designed for the SSDD. The results show that our method surpasses the improved Faster R-CNN by 12.7 points AP and surpasses the scale-transferrable Pyramid Network by 1.1, though the SA R-CNN performed the best recall. Figure 7 shows our detection performance in three types of complex situations for SAR images. In these cases, our method can effectively complete detection. The results show that our method leads to a more powerful feature extractor and a more efficient regression process to make the deep-learning method perform better in SAR images.

Table 6. Detection performance comparison on SSDD.

Method	AP	Recall
Faster R-CNN [7]	90.2	96.4
Cascade R-CNN [32]	90.3	95.3
Libra R-CNN [33]	90.2	96.1
RetinaNet [38]	87.7	94.9
SSD [39]	83.5	94.0
YOLOv3 [37]	86.8	93.4
Improved Faster R-CNN [4]	78.8	/
G-CNN [5]	87.5	89.5
MSARN [56]	83.7	/
Scale-transferrable Pyramid Network [57]	90.4	/
DAPN [58]	83.4	/
B-FPN [59]	88.3	/
SA R-CNN (ours)	**91.5**	**97.6**

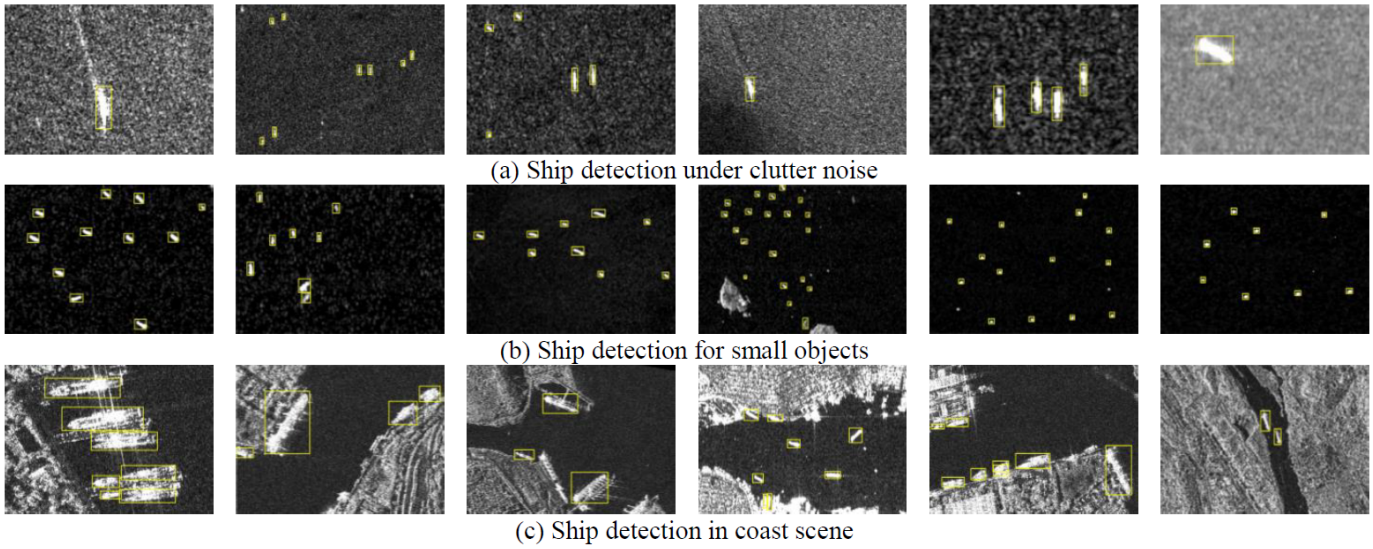

Figure 7. Ship detection performance of our SA R-CNN in three types of complex situations in the SSDD dataset with: (**a**) some noise or clutter; (**b**) a large number of small objects on the sea surface; and (**c**) a complex background around the coast.

5. Discussion

Through the visual analysis of the inference results of the HRSID test set by the SA R-CNN method, we found that hard samples are mainly concentrated in the following two situations. One is that for a target with a larger size, which can easily detect only a part of the target and many inaccurate detection frames will appear as positive samples. Another is that for extremely dense small targets, as some will be missing and there will be false detections. Some specific examples of the above two situations are shown in Figure 8.

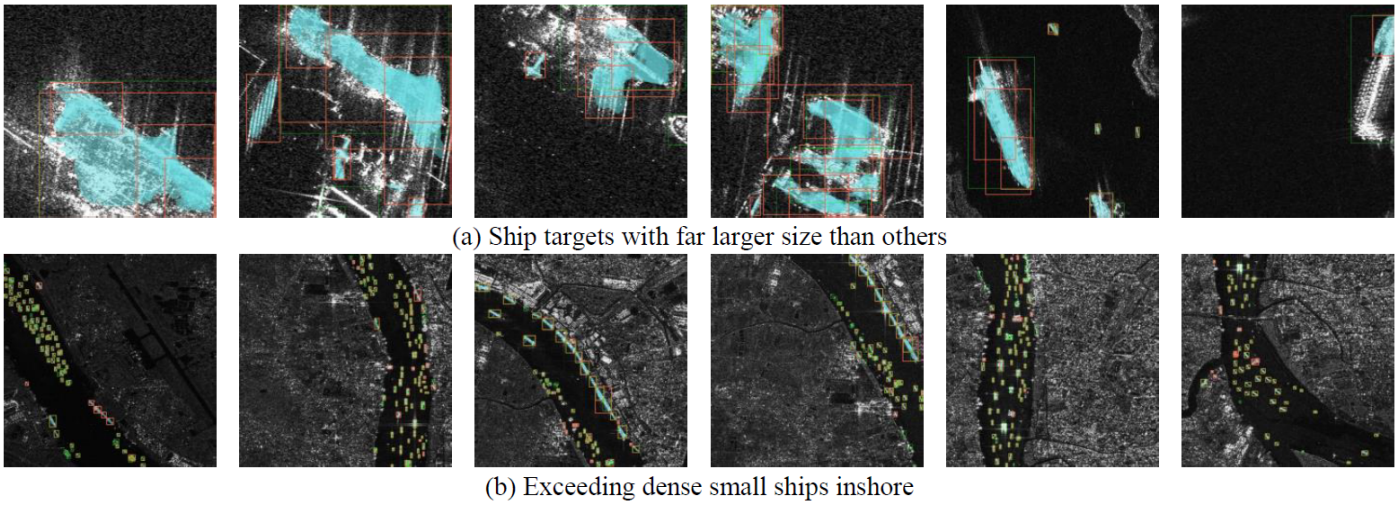

Figure 8. Visible ship instance segmentation results of our SA R-CNN from the test set of HRSID in some difficult situations where the effect was not good enough in these hard samples: (**a**) for some ship targets in a large proportion of the image, only part of the target can be detected but the whole has been cut off; and (**b**) for some excessively dense small inshore ships, our method cannot avoid missing some and making false detections.

On the one hand, for the ship targets with a far larger size than others, the detection bounding box will cut off the ship target and can easily mistake some objects on the shore as targets. The imbalance of the sample results in our method not being sufficiently robust to targets within this scale during the training process, leading to a lower APl in all compared methods. On the other hand, for excessively dense small inshore ships, our method has a poor detection effect. Specifically, for some small docked targets, due to the interference information from the port, the attention mechanism cannot be refined to this degree, resulting in missed detections, which is also the main difficulty in the field of ship detection.

In summary, we can improve the performance of SA R-CNN from the following aspects in future research. Firstly, different datasets can be divided to train the model according to specific task requirements. Then, we will further study the relationship between the attention mechanism and feature maps of different resolutions, and design attention modules for different scales—especially for the targets in extreme shape. In addition, we will also refer to other studies on the detection of dense targets in SAR images for densely distributed ships. In the future, we will strive to improve the performance of SA R-CNN on hard samples through the above methods.

6. Conclusions

Aiming to determine the target's characteristics in SAR images, an instance segmentation method SA R-CNN based on the synergistic attention mechanism is proposed in this paper. Our method improved both the segmentation and detection performance by GAM, SAM, and AAM in terms of the image-level, semantic-level, and target-level aspects. The proposed model alleviates the difficulties caused by the lack of color and texture details in SAR images and the complex distribution of ship targets. The experiments proved that our method has made significant improvements compared with state-of-the-art methods. In particular, the global attention module introduced in the backbone integrated the local and global context information of the feature map. Furthermore, the semantic attention module uses semantic features to guide detailed feature fusion, which is more robust for ship detection around the harbor. In our future work, we will further improve the performance of the method for large samples and densely distributed samples in the dataset. In addition, the method with the synergistic attention mechanism has potential for innovation in other fields of SAR image tasks and other remote sensing applications.

Author Contributions: Conceptualization, D.Z.; methodology, C.Z. and D.Z.; software, C.Z.; validation, C.Z. and D.Z.; formal analysis, X.Q.; resources, J.Q.; writing—original draft preparation, C.Z. and D.Z.; writing—review and editing, C.Z. and D.Z.; visualization, C.Z.; supervision, Z.S. (Zhenhua Su); funding acquisition, Z.S. (Zhenwei Shi) and D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant No. 2019YFC1510905, and in part by the Air Force Equipment pre-research project under grant 303020401 (corresponding author: Danpei Zhao).

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.

Abbreviations

The following abbreviations are used in this manuscript:

SAR	synthetic aperture radar
CNNs	convolutional neural networks
GAM	global attention module
SAM	semantic attention module
AAM	anchor attention module
AP	average precision
LCA	local context attention
GCA	global context attention

References

1. El-Darymli, K.; McGuire, P.F.; Power, D.; Moloney, C. Target detection in synthetic aperture radar imagery: A state-of-the-art survey. *J. Appl. Remote Sens.* **2013**, *7*, 071598. [CrossRef]
2. Wang, H.; Xu, F.; Chen, S. Saliency Detector for SAR Images Based on Pattern Recurrence. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2891–2900. [CrossRef]
3. Robey, F.; Fuhrmann, D.; Kelly, E.J.; Nitzberg, R. A CFAR adaptive matched filter detector. *IEEE Trans. Aerosp. Electron. Syst.* **1992**, *28*, 208–216. [CrossRef]
4. 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; pp. 1–6.
5. Zhang, T.; Zhang, X. High-Speed Ship Detection in SAR Images Based on a Grid Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 1206. [CrossRef]
6. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
7. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef] [PubMed]
8. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid Task Cascade for Instance Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 4969–4978.
9. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT: Real-Time Instance Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 9156–9165.
10. Lin, T.Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
11. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
12. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 1971–1980.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

14. 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]
15. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask Scoring R-CNN. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6402–6411.
16. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. SOLO: Segmenting Objects by Locations. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Online, 23–28 August 2020; pp. 649–665.
17. Xie, E.; Sun, P.; Song, X.; Wang, W.; Liu, X.; Liang, D.; Shen, C.; Luo, P. PolarMask: Single Shot Instance Segmentation With Polar Representation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 12190–12199.
18. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Huang, Y.; Yan, Y. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 8570–8578.
19. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
20. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
21. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
23. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. BAM: Bottleneck Attention Module. In Proceedings of the 2018 British Machine Vision Conference, Newcastle, UK, 3–6 September 2018.
24. Wang, X.; Girshick, R.B.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
25. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Shi, H.; Liu, W. CCNet: Criss-Cross Attention for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 603–612.
26. Yeremy, M.; Geling, G.; Rey, M. Results from the Crusade ship detection trial: Polarimetric SAR. *IEEE Int. Geosci. Remote Sens. Symp.* **2002**, *2*, 711–713.
27. Liu, Y.; Fang, M.; Feng, Q.; Wang, L. An Automatic Ship Detection system using ERS SAR images. In Proceedings of the IGARSS 2003—2003 IEEE International Geoscience and Remote Sensing Symposium, Proceedings (IEEE Cat. No.03CH37477), Toulouse, France, 21–25 July 2003; Volume 4, pp. 2656–2658.
28. Sugimoto, M.; Ouchi, K.; Nakamura, Y. On the novel use of model-based decomposition in SAR polarimetry for target detection on the sea. *Remote Sens. Lett.* **2013**, *4*, 843–852. [CrossRef]
29. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
31. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
32. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
33. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 821–830.
34. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Online, 23–28 August 2020.
35. Redmon, J.; Divvala, S.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
36. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
37. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
38. Lin, T.Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
39. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.Y.; Berg, A. SSD: Single Shot MultiBox Detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.

40. Kang, M.; Leng, X.; Lin, Z.; Ji, K. A modified faster R-CNN based on CFAR algorithm for SAR ship detection. In Proceedings of the 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 19–21 May 2017; pp. 1–4.
41. Liu, Y.; Zhang, M.; Xu, P.; Guo, Z. SAR ship detection using sea-land segmentation-based convolutional neural network. In Proceedings of the 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 19–21 May 2017; pp. 1–4.
42. 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]
43. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. Automatic Ship Detection Based on RetinaNet Using Multi-Resolution Gaofen-3 Imagery. *Remote Sens.* **2019**, *11*, 531. [CrossRef]
44. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. Depthwise Separable Convolution Neural Network for High-Speed SAR Ship Detection. *Remote Sens.* **2019**, *11*, 2483. [CrossRef]
45. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]
46. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1331–1344. [CrossRef]
47. Geng, X.; Shi, L.; Yang, J.; Li, P.; Zhao, L.; Sun, W.; Zhao, J. Ship Detection and Feature Visualization Analysis Based on Lightweight CNN in VH and VV Polarization Images. *Remote Sens.* **2021**, *13*, 1184. [CrossRef]
48. Wu, Z.; Hou, B.; Ren, B.; Ren, Z.; Wang, S.; Jiao, L. A Deep Detection Network Based on Interaction of Instance Segmentation and Object Detection for SAR Images. *Remote Sens.* **2021**, *13*, 2582. [CrossRef]
49. de Albuquerque, A.O.; de Carvalho, O.L.F.; e Silva, C.R.; de Bem, P.P.; Gomes, R.A.T.; Borges, D.L.; Guimarães, R.F.; Pimentel, C.M.M.M.; de Carvalho Júnior, O.A. Instance segmentation of center pivot irrigation systems using multi-temporal SENTINEL-1 SAR images. *Remote Sens. Appl. Soc. Environ.* **2021**, *23*, 100537.
50. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 764–773.
51. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region Proposal by Guided Anchoring. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2960–2969.
52. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.
53. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Miami Beach, FL, USA, 20–26 June 2009.
54. Everingham, M.; Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2009**, *88*, 303–338. [CrossRef]
55. 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.
56. Chen, C.; He, C.; Hu, C.; Pei, H.; Jiao, L. MSARN: A Deep Neural Network Based on an Adaptive Recalibration Mechanism for Multiscale and Arbitrary-Oriented SAR Ship Detection. *IEEE Access* **2019**, *7*, 159262–159283. [CrossRef]
57. Liu, N.; Cui, Z.; Cao, Z.; Pi, Y.; Lan, H. Scale-Transferrable Pyramid Network for Multi-Scale Ship Detection in Sar Images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1–4.
58. Li, Q.; Min, R.; Cui, Z.; Pi, Y.; Xu, Z. Multiscale Ship Detection Based On Dense Attention Pyramid Network in Sar Images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 5–8.
59. Zhang, T.; Zhang, X.; Shi, J.; Wei, S.; Wang, J.; Li, J. Balanced Feature Pyramid Network for Ship Detection in Synthetic Aperture Radar Images. In Proceedings of the 2020 IEEE Radar Conference (RadarConf20), Florence, Italy, 21–25 September 2020; pp. 1–5.