remote sensing

MDPI

Article

Ship Instance Segmentation Based on Rotated Bounding Boxes for SAR Images

Xinpeng Yang [1], Qiang Zhang [1,*], Qiulei Dong [2,3], Zhen Han [4], Xiliang Luo [1] and Dongdong Wei [4]

[1] Remote Sensing Image Processing and Fusion Group, School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] Hangzhou Institute of Technology, Xidian University, Hangzhou 311200, China
* Correspondence: zhangqiang@xidian.edu.cn

Abstract: Ship instance segmentation in synthetic aperture radar (SAR) images is a hard and challenging task, which not only locates ships but also obtains their shapes with pixel-level masks. However, in ocean SAR images, because of the consistent reflective intensities of ships, the appearances of different ships are similar, thus making it far too difficult to distinguish ships when they are in densely packed groups. Especially when ships have incline directions and large aspect ratios, the horizontal bounding boxes (HB-Boxes) used by all the instance-segmentation networks that we know so far inevitably contain redundant backgrounds, docks, and even other ships, which mislead the following segmentation. To solve this problem, a novel ship instance-segmentation network, called SRNet, is proposed with rotated bounding boxes (RB-Boxes), which are taken as the foundation of segmentation. Along the directions of ships, the RB-Boxes can surround the ships tightly, but a minor deviation will corrupt the integrity of the ships' masks. To improve the performance of the RB-Boxes, a dual feature alignment module (DAM) was designed to obtain the representative features with the direction and shape information of ships. On account of the difference between the classification task and regression task, two different sampling location calculation strategies were used in two convolutional kernels of the DAM, making these locations distributed dynamically on the ships' bodies and along the ships' boundaries. Moreover, to improve the effectiveness of training, a new adaptive Intersection-over-Union threshold (AIoU) was proposed based on the aspect-ratio information of ships to raise positive samples. To obtain the masks in the RB-Boxes, a new Mask-segmentation Head (MaskHead) with the twice sampling processes was explored. In experiments to evaluate the RB-Boxes, the accuracy of the RB-Boxes output from the Detection Head (DetHead) of SRNet outperformed eight rotated object-detection networks. In experiments to evaluate the final segmentation masks, compared with several classic and state-of-the-art instance-segmentation networks, our proposed SRNet achieved more accurate ship instance masks in SAR images. The ablation studies demonstrated the effectiveness of the DAM in the SRNet and the AIoU for our network training.

Keywords: synthetic aperture radar (SAR) image; ship instance segmentation; rotated bounding box; dual feature alignment; adaptive IoU threshold (AIoU)

Citation: Yang, X.; Zhang, Q.; Dong, Q.; Han, Z.; Luo, X.; Wei, D. Ship Instance Segmentation Based on Rotated Bounding Boxes for SAR Images. *Remote Sens.* **2023**, *15*, 1324. https://doi.org/10.3390/rs15051324

Academic Editor: Antonio Iodice

Received: 23 December 2022
Revised: 20 February 2023
Accepted: 24 February 2023
Published: 27 February 2023

1. Introduction

Synthetic aperture radar (SAR) is a kind of active microwave imaging sensor that is able to work at all times in severe weather conditions; therefore, SAR is always used for ocean observation [1]. With the development of SAR-related technology, the quantity of SAR ocean data increases rapidly, and automatic ship-information-extraction has become a crucial technology for the understanding of various ocean scenes. Ship instance segmentation is an all-around ship-information-extraction technology. Although it is harder

than ordinary detection [2], it is able to obtain the location and mask of each ship and thus is used widely in various fields. In the military field, ship instance segmentation is helpful to analyze the marine situation and improve marine defense capacity. In the civil field, ship instance segmentation is helpful for the monitoring and management of marine transportation [3]. Therefore, the instance segmentation for ships in SAR images has drawn many researchers' attention.

At present, many instance-segmentation networks are proposed for optical images and remote-sensing images, such as Mask RCNN [4], Mask Scoring RCNN [5], Cascade Mask RCNN [6], Hybrid Task Cascade [7], and FL CSE ROIE [8]. These networks obtain the masks of the objects based on horizontal bounding boxes (HB-Boxes), which get some achievement but have unavoidable problems for ships in SAR images, as shown in Figure 1. Different from the RGB information in the optical images, which can clearly reflect each ship in detail, the reflection intensities of ships and even docks in SAR images which are obtained via the SAR sensors based on the electromagnetic wave are similar. Therefore, densely packed ships are confused easily, and the docks also can be segmented as ships. The first row and the second row of Figure 1 list three source SAR images and their ground-truth images. From (a) and (c) of the first row, we can see that the parallel ships have a similar appearance, and (b) of the first row shows us that the dock where a ship is moored looks like a part of the ship. For the ship instance-segmentation networks mentioned above, because of the inclining directions and large aspect ratios of ships, their HB-Boxes contain other information which cannot be distinguished by the following segmentation process. When objects with similar reflection characteristics are contained in one HB-Box, the segmentation network cannot determine which parts belong to the ship that really needs to be segmented. As shown in the third row of Figure 1, which lists three instance masks based on HB-Boxes, (a) illustrates that the densely packed ships are segmented repeatedly, (b) illustrates that the dock is segmented to a part of the ship, and (c) illustrates that the parallel ships are predicted to one ship.

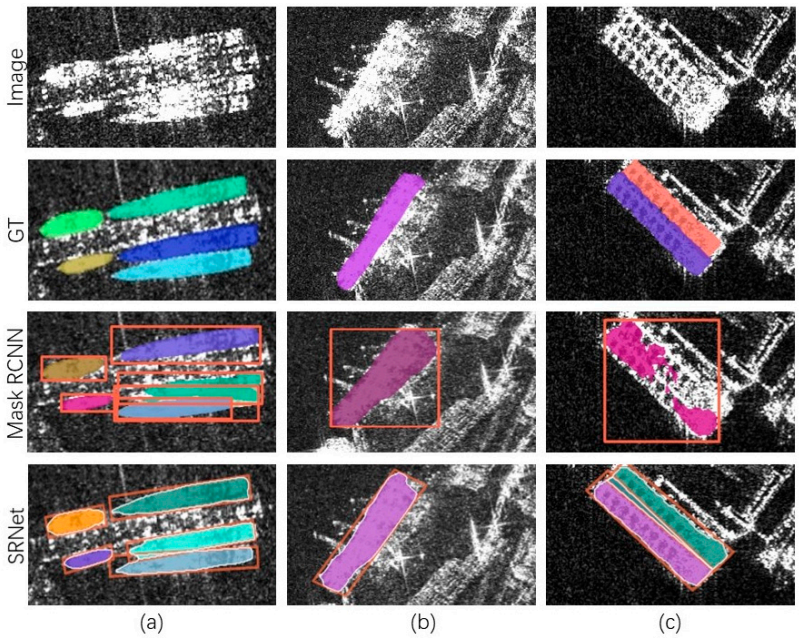


Figure 1. Visual comparison between the instance masks from the HB-Box-based Mask RCNN and from our RB-Box-based SRNet. (**a**) The scene of the densely packed ships. (**b**) The scene of the ship berthing in a dock. (**c**) The scene of parallel ships.

Different from the HB-Boxes, the RB-Boxes can grasp the directions of ships and surround ships tightly. Therefore, the following segmentation process extracts masks easily in the RB-Boxes containing fewer backgrounds, thus making it possible to avoid missed

segmentation, repeated segmentation, and error segmentation. Based on this idea, a novel instance-segmentation framework based on RB-Boxes is formed. The fourth row of Figure 1 lists the instance masks output from our SRNet based on this framework, which is able to eliminate the problems mentioned above to correctly segment the ships in different scenes. It should be noted that the detection of RB-Boxes is the core of the framework, which directly influences the segmentation performance. Therefore, the accuracy of the RB-Box needs to be improved further.

Recently, many rotated object detection networks have been proposed to obtain the RB-Boxes for remote-sensing images, such as faster RCNN [9], RetinaNet [10], Reppoints [11], ROI Trans [12], R3Det [13], ReDet [14], Oriented RCNN [15], and S2ANet [16]. In these networks, before producing the regression and classification of RB-Boxes, the feature sampling locations are unable to fit ships completely for effective feature extraction, and the same feature maps, which reflect the single property of ships, are used to produce both the RB-Boxes and their classification scores. In their training processes, the positive samples of the ships with large aspect ratios are distributed on their center areas; thus, the samples at the bow and stern are not effectively trained.

For the RB-Boxes of ships, the segmentation strategy for HB-Boxes cannot be used, and how to obtain the instance masks from the RB-Boxes is another obstacle. Because the RB-Boxes are misaligned with the grids of feature maps and the original image, it is necessary to design sampling strategies to set up the corresponding relationships between the sampling locations in the RB-Boxes and the grids.

Based on the instance-segmentation framework based on RB-Boxes mentioned above, a ship instance-segmentation network, called SRNet, is proposed in this paper. To raise the accuracy of RB-Boxes, a new dual feature alignment module (DAM) is designed to refine the feature maps from the feature pyramid network (FPN [17]), and a new adaptive IoU threshold (AIoU) calculation method is presented to raise the positive samples in the network training process. In the DAM, on account of the properties of the regression task and classification task, two alignment convolutions extract the feature maps for regression and classification, respectively, based on two different convolutional kernels. In these kernels, the sampling locations change with the directions and shapes of ships. In the network training, the aspect ratio of each ship is considered in the calculation of the Intersection-over-Union (IoU) threshold [18], which is used to rationally allocate the samples at the bow and stern to positive samples. With the RB-Boxes output from the DetHead of the SRNet, a MaskHead is designed to extract the features in the RB-Boxes by feature sampling and then predict the masks based on these features and transfer them onto the original image.

In sum, the main contributions of our paper are as follows:

1. The new instance-segmentation framework based on RB-Boxes is proposed for ships in SAR images, which reduces missed segmentation, repeated segmentation, and error segmentation.
2. The new dual feature alignment module (DAM) is designed to extract the feature maps for the regression task and classification task. In this module, two different sampling strategies for the two convolutional kernels are used, which ensure the two alignment convolutions enable to focus on the areas beneficial to their corresponding tasks.
3. The adaptive IoU threshold (AIoU) calculation method for the sample allocation is proposed. On account of the aspect ratio of each ship, the AIoU raises the number of positive samples, and the features at the bow and stern are effectively trained.
4. The MaskHead is designed to obtain masks in RB-Boxes and transfer them onto the original image.

2. Related Works

2.1. Instance Segmentation

Instance segmentation obtains the locations and masks of objects, which is more accurate than ordinary detection [2]. With the progress of deep learning, many instance-

segmentation networks are proposed for optical images and remote-sensing images [19]. These networks are the HB-Box-based segmentation networks, which first locate the objects with HB-Boxes and then predict instance mask in each HB-Box. Mask RCNN [4] achieves instance segmentation by adding the mask branch in the RCNN of Faster RCNN [9] and improves the detection performance by the ROI Align [4]. Mask Scoring RCNN [5] combines the classification scores and IoU scores to calculate the mask scores to improve the instance masks' qualities. Cascade Mask RCNN [6] cascades RCNN three times based on Mask RCNN to improve the accuracy of location and segmentation. Hybrid Task Cascade [7] proposes a new cascade structure and attaches a semantic branch to supply semantic information. FL CSE ROIE [8] achieves instance segmentation by combining context ROI extractor [20], attention mechanisms [21–23], and CARAFEB [24] so as to improve the location performance. In general, these networks have made a great effort to improve the performance of the HB-Box-based location and the quality of segmentation in HB-Boxes. However, due to a lack of consideration for the imaging method of SAR sensors and the complex ocean scene in SAR images, these HB-Box-based networks cannot achieve a good performance for ship instance segmentation in SAR images.

2.2. Object Detection

Object detection in SAR images is an important part of SAR-image understanding, and many accurate detection methods for different application requirements are proposed. Armando Marino et al. researched the ship detection method based on spectral analysis [25]. T. Zhang et al. proposed a two-stage ship detection method for PolSAR images [26]. X. G. Leng et al. designed a ship detection method based on complex generalized Gaussian distribution fast shape parameter estimation [27]. For instance-segmentation networks, object detection is also the primary step. In the beginning, object detection networks based on HB-Boxes for optical images are proposed [28], such as Faster RCNN [9] and RetinaNet [10]. Faster RCNN [9] is a classic detection network with a two-stage structure, which improves HB-Box-based location precision. This network has been used in many object detection tasks, such as ship detection for SAR images. RetinaNet [10] proposes the focal loss function to improve the quality of HB-Box-based location by balancing the samples. Although great efforts have been made to improve the location precision in these networks based on HB-Boxes, the HB-Boxes of ships in SAR images always contain too much redundant information, due to the ships sailing and berthing in arbitrary directions. Especially the densely packed ships, which always have similar reflection characteristics, may be contained in one HB-Box.

For remote-sensing images, many object detection networks based on RB-Boxes are proposed [29] in order to obtain more accurate location information. Rotated RetinaNet [10] attaches extra angle prediction to the source RetinaNet for rotated object detection. However, this network does not align the features with the shapes of the ships, leading to bad rotated detection precision. R3Det [13] and S2ANet [16] propose different alignment methods to optimize features. R3Det fuses features of five key points to obtain aligned features and S2ANet extracts aligned features by an alignment convolution, but the two networks use the same feature maps for both the classification task and regression task. The difference between classification and regression for ships is not considered in them. ROI Trans [12] and ReDet [14] extract rotation invariant features to improve the location precision with RROI extractors and E2CNN [30], respectively, and ReDet [14] aligns features with the angles of RB-Boxes to improve the location precision further. Oriented RCNN [15] represents RB-Boxes by the distance from vertexes of RB-Boxes to the vertexes of RB-Boxes' bounding boxes, which proposed a new oriented RPN to balance the accuracy and speed. However, these networks are trained with the sample allocation strategy with a fixed IoU threshold, which leads to undersampling or oversampling for ship detection due to various aspect ratios of ships. ATSS [31] is a frequently used sample allocation strategy, which calculates the different IoU thresholds for different ground-truth boxes. However, ATSS [31] only considers the information of the samples in the neighborhood of the center of the

ground truth, leading to a bad performance on the large aspect ratio ships. In a word, from the feature extraction to the sample allocation for network training, the properties of ships are not considered thoroughly by these networks. Thus, improvement of the RB-Box-based detection for ships in SAR images is crucial work in our proposed SRNet.

3. Methodology

In this section, based on RB-Boxes, the ship instance-segmentation network SRNet is described in detail. Because the RB-Boxes surround ships tightly and contain fewer backgrounds, a slight deviation of the RB-Boxes will corrupt the completeness of segmentation. To improve the location accuracy of the RB-Boxes, the DAM is presented, in which the two alignment convolutions are designed for the regression task and classification task. In addition, an AIoU is adopted for network training to increase the positive samples at the bow and stern, which trained the features comprehensively. To predict the instance masks in the RB-Boxes, the MaskHead is presented with a feature-sampling strategy and a mask-sampling strategy.

In the following subsections, the overall architecture of the SRNet is introduced first. Then the three crucial modules, DAM, DetHead, and MaskHead, are described. Finally, the AIoU and the loss functions are presented.

3.1. Architecture

The overview architecture of our SRNet is shown in Figure 2. We can see that the SRNet consists of a feature-extraction procedure, a detection procedure, and a segmentation procedure. The feature-extraction procedure consists of a backbone and FPN, which outputs the feature maps of five scales. With these feature maps, the DAM in the detection procedure produces two kinds of aligned feature maps for each scale, and then the DetHead outputs the RB-Boxes of ships based on the results of the DAM and determines the scale that each ship is in. Finally, the MaskHead segments the ships in the RB-Boxes by using the feature maps of FPN from the corresponding scales.

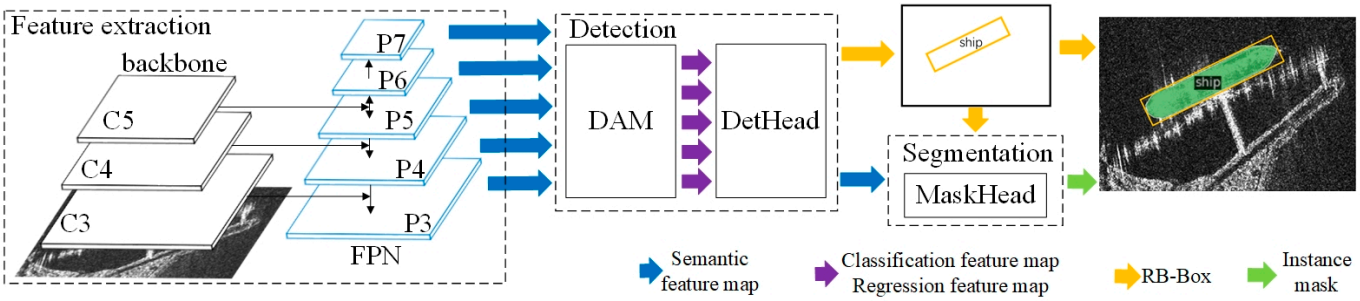


Figure 2. Overview of the proposed SRNet, consisting of a feature extraction procedure, a detection procedure, and a segmentation procedure. In the feature extraction procedure, the FPN takes feature maps C3, C4, and C5 extracted by the backbone as inputs and then outputs semantic feature maps P3–P7. In the detection procedure, these semantic feature maps are first transmitted to the DAM to produce the classification feature maps and regression feature maps. Then the DetHead takes the outputs of DAM as inputs to output the RB-Boxes and their classification scores. The segmentation procedure takes the RB-Boxes and the semantic feature maps as inputs and then outputs the final instance masks.

3.2. Dual Feature Alignment Module

In SAR images, because ships always have arbitrary directions and various aspect ratios, the standard convolution based on the regular grid kernel makes it difficult to obtain the features corresponding to the shapes of ships, thus causing the misalignments between the standard convolutions and the rotated ships [32]. Moreover, the classification and regression tasks need to extract the features with different properties [33–36]. The classification task concentrates on the areas of the object itself to predict a correct classification score [36].

On the other hand, the regression task concentrates on the areas of the object itself and the areas around the object to predict an accurate object position [36]. Thus, the proposed DAM is designed to extract two different feature maps with different characters, using two alignment convolutions, which are based on two kinds of specially designed sampling locations. The two alignment convolutions are the classification alignment convolution (CAConv) and the regression alignment convolution (RAConv).

As shown in Figure 3, the sampling locations of RAConv (a) are uniformly distributed on the ship area and its bounding areas, and the sampling locations of CAConv (b) are intensively distributed inside the ship area. Therefore, the RAConv based on the sampling locations (a) can easily extract the feature map with the properties needed by the regression task, and the CAConv based on the sampling locations (b) can easily extract the feature map with the properties needed by the classification task. To do this, some parameters (the center coordinates, the width and height, and the rotation angle of the convolutional kernel) are needed to calculate these two kinds of sampling locations. Especially due to the fact that the sampling locations of CAConv are not on the regular center points, nine offsets are needed. Therefore, in DAM, two lightweight convolutional branches are used to produce these parameters.

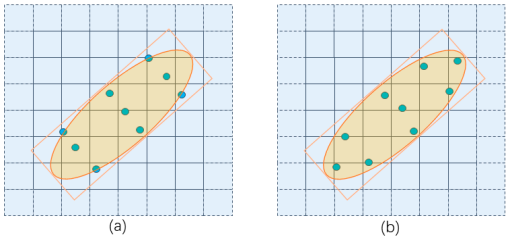


Figure 3. The two sets of sampling locations (blue points). The sampling locations of the RAConv (**a**). The sampling locations of the CAConv (**b**). The orange RB-Boxes indicate the kernels. The yellow ellipses circle the ship areas.

As shown in Figure 4, with the feature map (F^S) of each scale (S) as input, the R-Anchor branch uses a three-layer convolutional network to predict the coordinates (C^K) of the kernel based on a horizontal square anchor [10], which includes the center coordinate, (X, Y); the width and height, (W, H); and the rotation angle, θ. The C-Offset branch also uses a three-layer convolutional network to estimate the 9 offsets of the sampling locations of the kernel, represented by $O = \{(O_{xi}, O_{yi})|i = 1, \cdots, 9\}$.

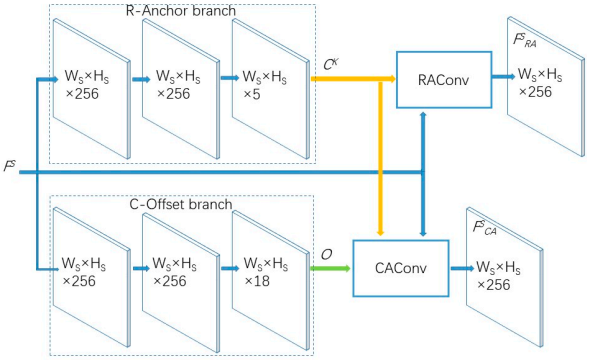


Figure 4. Overview of the DAM, which consists of two convolutional branches (R-Anchor branch and C-Offset branch) and two alignment convolutions (RAConv and CAConv). With the feature map (F^S) of scale (S) as input, the R-Anchor branch produces the kernel coordinates (C^K), and the C-Offset branch produces the sampling offsets (O). The $W_S \times H_S \times Channel$ represents a feature map with the width of W_s, the height of H_s, and the channel of $Channel$ extracted by convolution. With the kernel coordinates (C^K) and the sampling offsets (O), the RAConv and CAConv produce the regression feature map (F^S_{RA}) and classification feature map (F^S_{CA}), respectively.

With these parameters calculated by the R-Anchor branch and the C-Offset branch, the initial sampling locations of standard convolutions can be moved to the sampling locations of RAConv and CAConv, respectively, by the deformation step, rotation step, and translation step. Figure 5a,b correspond to the three steps.

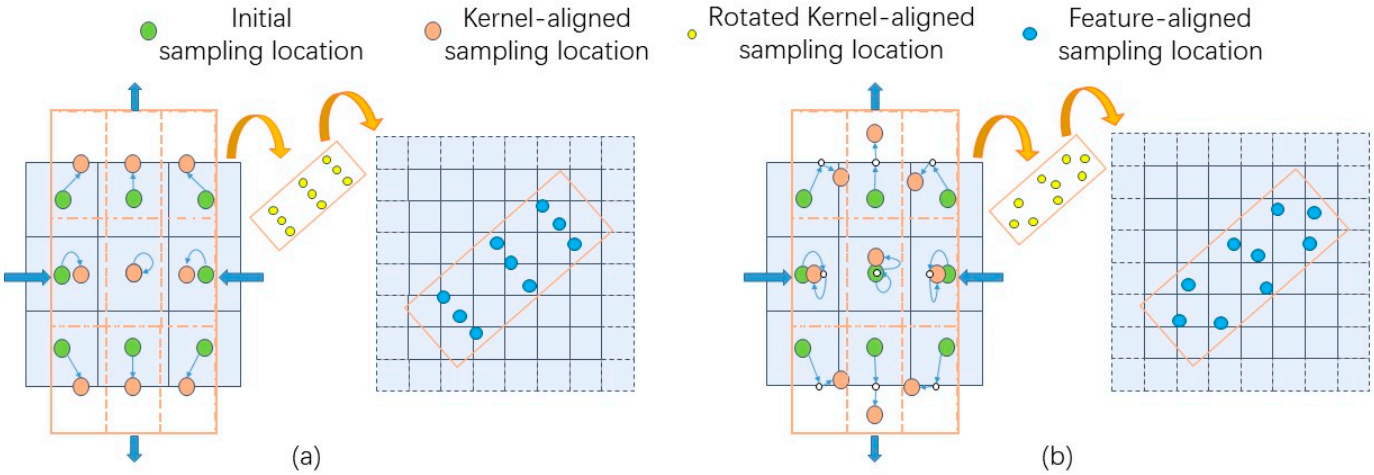


Figure 5. The calculation pipeline of the sampling locations for the RAConv (**a**) and the CAConv (**b**). The left figures in (**a**,**b**) demonstrate the process of the initial sampling locations moving along the blue arrows to the kernel-aligned sampling locations. With the rotation (middle figure in (**a**,**b**)) and translation (right figure in (**a**,**b**)) the feature-aligned sampling locations on the feature map coordinate system are obtained.

Firstly, the initial sampling locations perform the deformation step to obtain the kernel-aligned sampling locations. As shown in the left figure of (a), for RAConv, the initial sampling locations of a standard 3×3 kernel are deformed, corresponding to the width and height (W, H) of the kernel predicted by the R-Anchor branch to obtain the kernel-aligned sampling locations of RAConv, P^K_{RAC}. In the left figure of (b), for CAConv, the initial sampling locations of a standard 3×3 kernel are added, along with the 9 sampling offsets (O) estimated by the CAConv, and then reshaped according to the width and height (W, H) of the kernel; the kernel-aligned sampling locations of CAConv, P^K_{CAC}, can be obtained after this deformation step. The calculation processes of two kernel-aligned sampling locations, P^K_{RAC} and P^K_{CAC}, are described by Equation (1).

$$\begin{cases} P^K_{RAC} = P * \begin{pmatrix} \Delta W & \Delta H \end{pmatrix} \\ P^K_{CAC} = (P + O) * \begin{pmatrix} \Delta W & \Delta H \end{pmatrix} \end{cases} \tag{1}$$

where the P represents the 9 initial sampling locations of the standard kernel, defined as $\{(x, y)|x = -1, 0, 1 \; and \; y = -1, 0, 1\}$; O includes 9 sampling offsets, defined as $\{(O_{xi}, O_{yi})|i = 1, \cdots, 9\}$; ΔW and ΔH are $W/3$ and $H/3$; and * indicates the element-wise product.

Secondly, the kernel-aligned sampling locations perform the rotation step and translation step to obtain the feature-aligned sampling locations. As shown in the middle and right figures of (a) and (b), the two kinds of kernel-aligned sampling locations are firstly rotated by the rotation angle (θ) predicted by the R-Anchor branch and then transferred to the feature map coordinate system. The two kinds of feature-aligned sampling locations of RAConv and CAConv, P^F_{RAC} and P^F_{CAC}, can be obtained after these two steps. The calculation processes can be described by Equation (2):

$$\begin{cases} P^F_{RAC} = (X, Y) + P^K_{RAC} R_\theta \\ P^F_{CAC} = (X, Y) + P^K_{CAC} R_\theta \end{cases} \tag{2}$$

where $R_\theta = (\cos(\theta), -\sin(\theta); \sin(\theta), \cos(\theta))$ is a rotation matrix.

Thirdly, with the two kinds of sampling locations calculated by the three steps, the convolutional kernels are aligned with the ships. The feature map with the properties needed by the regression task is extracted by the RAConv based on the feature-aligned sampling locations, P^F_{RAC}. The feature map with the properties needed by the classification task is extracted by the CAConv based on the feature-aligned sampling locations, P^F_{CAC}. These feature map extraction processes are achieved by the deformable convolution, which is the weighted sum of the feature values at the 9 sampling locations, which are described by Equation (3):

$$\begin{cases} F^S_{RA} = \sum_{i=1}^{9} W^S_{RA}(i) F^S\left(P^F_{RAC}(i)\right) \\ F^S_{CA} = \sum_{i=1}^{9} W^S_{CA}(i) F^S\left(P^F_{CAC}(i)\right) \end{cases} \tag{3}$$

where W^S_{RA} and W^S_{CA} represent convolutional weights of the RAConv and the CAConv, and the optimal values of these parameters are calculated gradually during the network training. $F^S(p)$ denotes the features at the location p in F^S.

Due to the results of the R-Anchor branch including the center coordinate, shape, and rotation angle of ships, the DetHead predicts the final RB-Boxes more easily based on them rather than restarts. Thus, besides the regression feature map and the classification feature map, the C^K values are also inputted into the DetHead as the initial anchors to produce the final RB-Boxes of ships.

3.3. DetHead and MaskHead

The DetHead and MaskHead are two output heads, as shown in Figure 6, that are both designed on the lightweight convolutional framework commonly used in various object-detection networks and instance-segmentation networks [10,13,16,28,31,32,36]. In (a), the DetHead contains two branches, the F^S_{RA} and F^S_{CA}, which are extracted by the DAM and taken as inputs to produce RB-Boxes based on C^K and their classification scores. Then, based on the classification scores, the rotated NMS algorithm, in which the rotated IoU calculation algorithm is used, filters the RB-Boxes to remove the overlapping and redundant RB-Boxes and produces the final RB-Boxes and their corresponding classification scores. In (b), the MaskHead chooses a single-scale feature map, F^{S_l}, from FPN according to the area of each RB-Box [4] and extracts the texture feature map, $F^{S_l}_T$, with a fully convolutional network. Then, with $F^{S_l}_T$, the instance masks in the final RB-Boxes are predicted by the RB-Box instance mask prediction as follows.

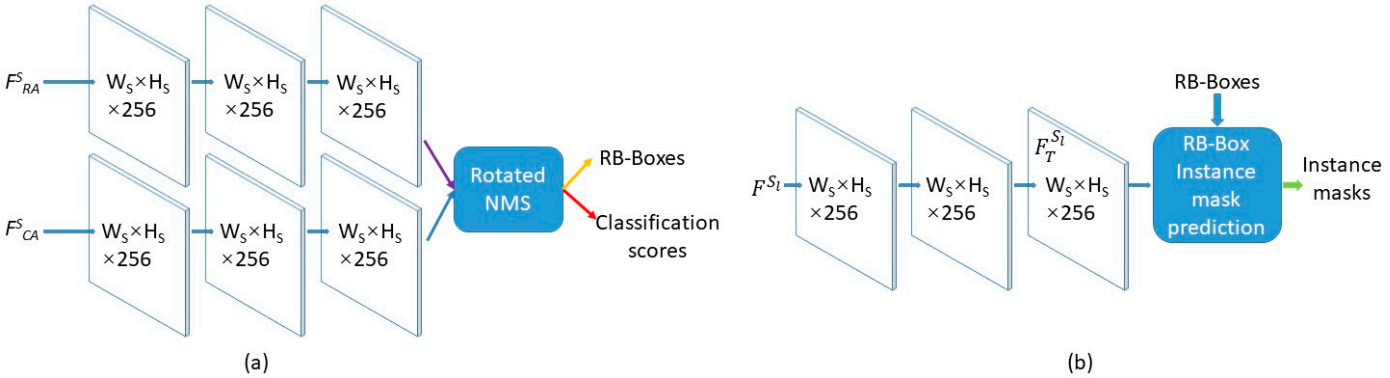


Figure 6. Overviews of the DetHead (**a**) and the MaskHead (**b**). In (**a**), F^S_{RA} and F^S_{CA} are input into two different three-layer convolutional networks to predict the RB-Boxes and their classification scores. Then the rotated NMS algorithm produces the final results. In (**b**), with the F^{S_l} as input, the feature map, $F^{S_l}_T$, output from a three-layer convolutional network and the RB-Boxes is used to produce the final instance masks by the RB-Box instance mask prediction.

Due to the RB-Boxes having arbitrary directions and the ROI Align [4] used in many HB-Box-based instance-segmentation networks obtaining the feature sampling locations in a horizontal area, the ROI Align [4] cannot be used in our MaskHead directly. Therefore, as shown in Figure 7, we sample the features in the RB-Boxes first, then predict the instance masks in the RB-Boxes by a 1×1 convolution, and finally transfer the masks in the RB-Boxes onto the origin image.

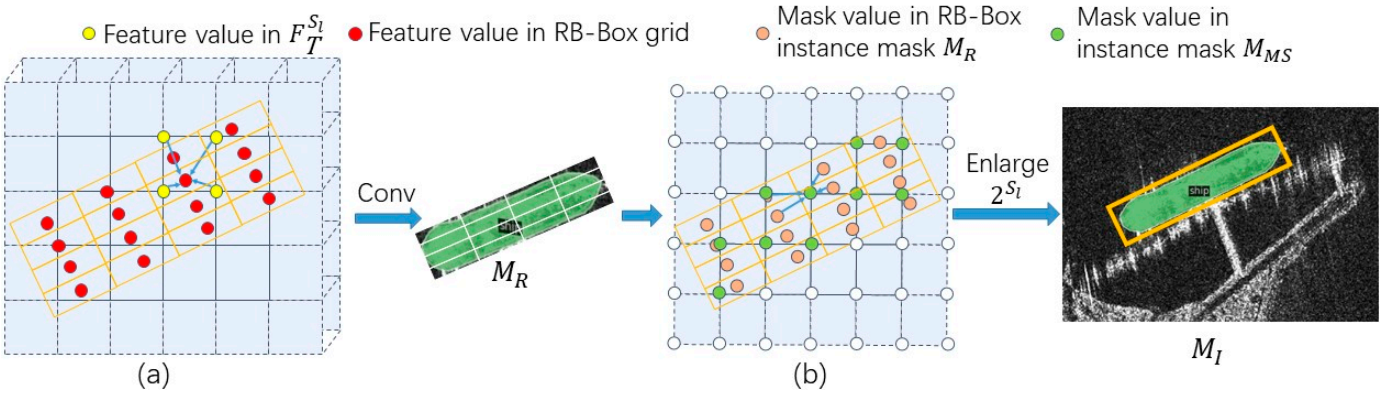


Figure 7. The pipeline of the RB-Box instance mask prediction: (**a**) indicates the feature sampling process, and (**b**) indicates the mask sampling process. In (**a**), the feature values in the RB-Box are computed with their nearest feature values in $F_T^{S_l}$. Then a standard 1×1 convolution, Conv, is used to produce the instance mask, M_R, in the RB-Box. In (**b**), the instance mask values in instance mask M_{MS} are computed with their nearest mask values in the RB-Box instance mask, M_R. Then, the M_{MS} is enlarged 2^{S_l} times to obtain the instance mask M_I in the original SAR image.

As shown in Figure 7a, we divide the RB-Box uniformly into a 28×28 grid, and the bins in it have the same size. The center points in these bins are selected as the feature sampling locations, P_{FS}. The bilinear interpolation algorithm is used to calculate the features F_{FS} at the P_{FS} with the four nearest feature points in $F_T^{S_l}$. The feature sampling process can be described by Equation (4):

$$F_{FS} = Bilinear\left(F_T^{S_l}\left(Near\left(P_{FS}\middle|F_T^{S_l}\right)\right)\right) \tag{4}$$

where $Near(p|F)$ represents the coordinates of the four nearest feature points of the coordinate p in F, and $F_T^{S_l}(p)$ represents the features at the coordinates p in $F_T^{S_l}$. *Bilinear* represents the bilinear algorithm.

With F_{FS} as input, the corresponding instance mask, M_R, in the RB-Box can be predicted by a 1×1 convolution. Then, as shown in Figure 7b, we choose the feature points in the RB-Box as the mask sampling locations, P_{MS}. The same as the feature sampling, the instance mask values at the mask sampling locations are aligned using their four nearest instance mask values in M_R to produce the instance mask M_{MS} in the coordinate system of the feature map $F_T^{S_l}$. The mask sampling process can be described by Equation (5):

$$M_{MS} = Bilinear(M_R(Near(P_{MS}|M_R))) \tag{5}$$

where $M_R(p)$ represents the mask values at the coordinate p in M_R. With the scale S_l of the input F^{S_l}, the instance mask M_{MS} is enlarged 2^{S_l} times to output the final instance mask M_I in the original SAR image.

Because of the dense sampling points in the RB-Boxes, the bilinear interpolation algorithm computes more accurate values for each sampling location, especially the instance mask values on the boundary of ships.

3.4. Network Training

3.4.1. Adaptive IoU Threshold

As mentioned in Section 3.2, the outputs of the R-Anchor branch directly influence the performance of the two alignment convolutions in the DAM and the final RB-Boxes predicted by the DetHead, which can be improved by reasonably raising the number of positive samples in network training. The frequently used sampling allocation strategy, ATSS [31], calculates an IoU threshold to gather positive samples, which considers the mean and standard deviation of the IoUs of the samples nearby the center of the RB-Box ground truth. Because this IoU threshold reflects the statistic value of the IoUs around the center of the ground truth box, it is more suitable for the object whose aspect ratio approximates 1. However, the ships in SAR images always have large aspect ratios. The IoU thresholds, which are calculated based on the center neighborhood samples only, will suppress the positive samples at the bow and stern, and thus the features at the bow and stern cannot be effectively trained. As shown in Figure 8a, the positive samples allocated by ATSS are distributed near the center of ground truth, but there are almost no positive samples at the bow and stern.

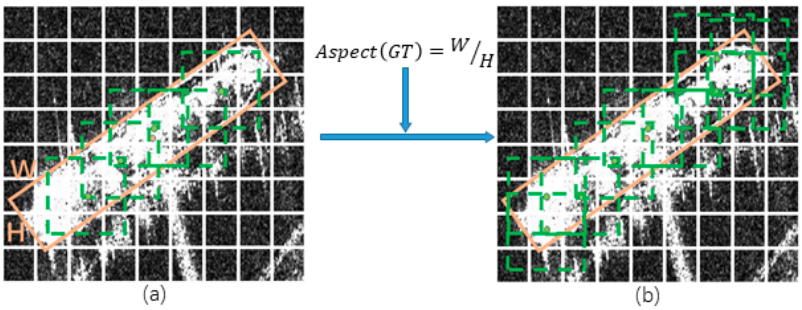


Figure 8. Comparison illustration of two IoU thresholds. (**a**) The positive sample allocation results of the ATSS. (**b**) The positive sample allocation results of the new adaptive IoU threshold with the aspect ratio. The green square boxes represent positive samples, and the green points represent their center points.

Therefore, on account of the aspect ratios of ships, a novel adaptive IoU threshold, called AIoU, is used in our network training to raise positive samples, as shown in Figure 8b. For each ship, we calculate the angle between its length and width and normalize it with $2/\pi$ to limit it to 0 to 1. From Figure 9 we can see that the larger the aspect ratio, the larger the normalized angle becomes. Then we invert the normalized angle by Equation (6) to obtain the factor F:

$$F = 1 - {}^{2}/_{\pi} \arctan(\alpha \cdot Aspect(GT)) \tag{6}$$

where $Aspect(GT)$ represents the aspect ratio of a ship, α is a factor which is used to adjust the change rate of F, and arctan represents the arctangent function.

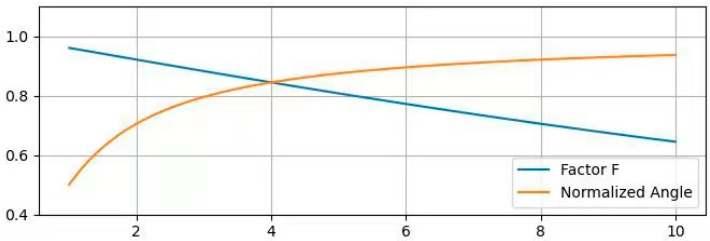


Figure 9. Illustration of the factor F and the normalized angle with $\alpha = 1$. The horizontal axis represents the aspect ratio. The vertical axis represents the value of the factor F and the normalized angle.

Moreover, the AIoU can be obtained by Equation (7):

$$T_{AIoU} = T_{ATSS} \cdot F \tag{7}$$

where the T_{ATSS} is the IoU threshold calculated by ATSS. As shown in Figure 9, F decreases with the increase of the aspect ratio, which reasonably adjusts the IoU threshold to raise the positive samples.

To evaluate the effectiveness of our AIoU, we test the number of positive samples for a RB-Box with the rotated angle of 60, which is used as the ground truth to allocate samples. We fix the width of the RB-Box and stretch its length to obtain ground truths with different aspect ratios. The IoU thresholds calculated by the state-of-the-art ATSS and our proposed AIoU are used to allocate samples. Figure 10 shows the statistics of the positive samples. It can be seen that the number of positive samples obtained by the AIoU is more than the number of positive samples obtained by the ATSS. Especially for the large aspect ratios, the difference in quantity between positive samples obtained by ATSS and AIoU increases obviously.

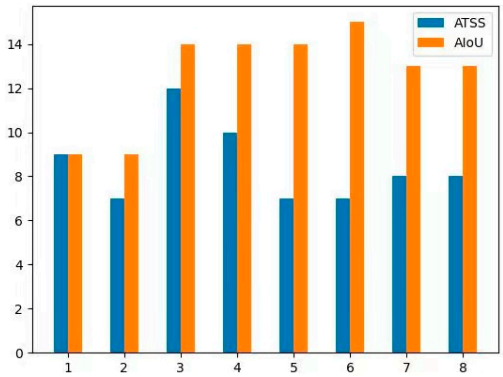


Figure 10. Illustration of the number of positive samples allocated by the ATSS and the AIoU. The horizontal axis represents the aspect ratio. The vertical axis represents the number of positive samples.

3.4.2. Loss Function

Our network SRNet is trained with a multi-task loss function, L, which consists of a classification loss, a regression loss, and a segmentation loss. Besides the RB-Boxes output from the DetHead, the C^K output from the R-Anchor branch in the DAM is also optimized in the regression loss. The multi-task loss function, L, can be described by Equation (8):

$$L = L_{cls} + L_{reg} + L_{seg} \tag{8}$$

where the classification loss, L_{cls}, is the focal loss [10] to optimize the classification scores to improve the performance of the classification. The regression loss L_{reg} consists of two smooth L1 losses [37] and is represented by Equation (9).

$$L_{reg} = L_{R-Anchor} + L_{RB-Box} \tag{9}$$

where $L_{R-Anchor}$ is used to optimize the C^K to improve the inputs of the DetHead, and L_{RB-Box} is used to optimize the final RB-Boxes predicted by DetHead to improve the location precision.

The L_{seg} is the binary cross-entropy loss [4], which optimizes the instance masks, M_I, predicted by MaskHead to improve the performance of mask value prediction at each location in M_I.

4. Experiments

In experiments, the SAR Ship Detection Dataset (SSDD) [38,39] and the Rotated Ship Detection Dataset in SAR images (RSDD) [40] were used to evaluate the performance of our

SRNet. Because the precision of RB-Box is the key factor for the instance segmentation in our network, the ship-detection results are also provided to demonstrate the accuracy of the RB-Boxes output from the DetHead. Moreover, we labeled 3555 SAR images from RSDD manually at the pixel level to produce a new dataset, Instance-RSDD, for experiments of ship instance segmentation. Eight rotated-object networks and five instance-segmentation networks were used as comparison networks. The ablation studies were conducted to verify the effectiveness of the components proposed in our network.

4.1. Dataset and Evaluation Metrics

4.1.1. SSDD [38,39]

The SAR Ship Detection Dataset (SSDD) is the first objects-detection dataset for ships in ocean SAR images. Since it was developed, the SSDD dataset has been expanded with the instance-segmentation labels and the RB-Box labels [39], so it is now widely used in rotated object detection and instance segmentation for ocean SAR images. The SSDD contains 1160 SAR images, with resolutions ranging from 1 m to 15 m, which are collected by RadatSat-2, TerraSAR-X, and Sentinel-1 sensors. These SAR images contain 2456 ship instances in total, and their contents involve a large number of water regions and dock regions, in which various scenes and densely packed ships are the challenges for rotated object detection and instance segmentation. In experiments, we divided the SSDD according to [39]: 928 images were used for network training and 232 images for evaluating the rotated-object-detection and instance-segmentation performance.

4.1.2. RSDD [40] and Instance-RSDD

The Rotated Ship Detection Dataset (RSDD) is the latest rotated-objects-detection dataset for ships in ocean SAR images. RSDD contains 7000 SAR images with resolutions ranging from 2 m to 20 m, which are collected by TerraSAR-X and Gaofen-3 sensors. These SAR images contain 10,263 ship instances in total and involve abundant scenes. Due to the various imaging styles and polarization modes, the visual difference between the images is obvious, making it hard to extract representation feature maps, which is a challenge for rotated object detection and instance segmentation. In experiments, we devided the RSDD according to [40]: 5000 images were used for training, and 2000 were used images for evaluating the rotated object detection performance.

Due to there being no instance-segmentation labels in RSDD, we expanded it to produce a new dataset Instance-RSDD for instance-segmentation experiments. In the dataset, 3555 clear SAR images are picked up and manually labeled at the pixel level to ensure that each ship has an accuracy instance mask. The Instance-RSDD can be downloaded at GitHub [41]. In experiments, 2840 images in the dataset are used for network training, and 715 images are used for evaluating the instance-segmentation performance.

4.1.3. Evaluation Metrics

In rotated-object-detection experiments, we use two metrics to evaluate the RB-Boxes output from the rotated-object-detection networks. One metric is Recall, which is the proportion of the true predicted positive samples in all ground truth boxes. Besides considering the proportion of the true predicted positive samples in all ground truth boxes, the metric mean average precision (mAP) also takes the proportion of the true predicted positive samples in all predicted samples into account. Compared with Recall, mAP is able to evaluate the networks more comprehensively [42].

In instance-segmentation experiments, we use several metrics to evaluate the instance masks output from the instance-segmentation networks. These metrics are the APs under the conditions of different instance mask IoUs [43], which are in MS COCO format [43] and can evaluate instance-segmentation performance comprehensively [19]. To evaluate the performance of instance segmentation under different overlap ratios, three APs are used in this paper. AP is the most important metric which is under the condition of the IoU thresholds in the range of 0.5 to 0.95. AP_{50} and AP_{75} are under the conditions of the IoU

thresholds of 0.5 and 0.75, respectively. It is worth noting that the higher the IoU threshold is, the more strictly the predicted instance mask is required to match the ground truth mask. To evaluate the performance of instance masks for the objects with different scales, respectively, AP_S, AP_M, and AP_L are used, which are calculated under the condition of IoU threshold equal to 0.5. AP_S calculates the AP value for the object with a small scale (area of object < 32^2 pixels), AP_M calculates the AP value for the object with a medium scale (32^2 pixels < area of object < 96^2 pixels), and AP_S calculates the AP value for the object with a large scale (96^2 pixels < area of object) [43]. In experiments, these evaluation metrics can comprehensively evaluate the instance-segmentation performance of networks.

4.2. Implementation Details

We implement our network and all comparison networks by using the mmrotate [44] and mmdetection [45] toolkits based on the PyTorch deep-learning framework. The resnet50 [46] is taken as the backbone for all networks. In our SRNet, the α is set to 1/16. During network training and testing, the image size is set up to 608 × 608 for the SSDD dataset, and the image size is set up to 512 × 512 for the RSDD dataset. The SGD optimization algorithm with a momentum of 0.9 and decay weight of 0.0001 is used to update the parameters of the networks, and the initial learning rate is 0.0025. We train the networks for 36 epochs and adjust the learning rate to 0.00025 and 0.000025 at the 24th epoch and 33rd epoch, respectively. For both training and testing, a PC with an Ubuntu operating system and an NVIDIA 3090GPU is used.

4.3. Rotated-Object-Detection Results for Ships in SAR Images

To verify the rotated-object-detection performance of our SRNet, eight rotated-object-detection networks designed for remote-sensing images are used for comparison, namely Rotated Faster RCNN [9], Rotated RetinaNet [10], Rotated Reppoint [11], ROI Trans [12], R3Det [13], ReDet [14], Oriented RCNN [15], and S2ANet [16]. Tables 1 and 2 report the rotated-object-detection performance of the different networks. In Table 1, the SRNet achieves the best performance compared with all the networks on the SSDD dataset. Specifically, it obtains an 0.8% higher in Recall and 0.1% higher in mAP than the best in the comparison networks. In Table 2, on the RSDD dataset, although the SRNet achieves a 0.2% lower Recall than the best in the comparison networks, the SRNet achieves the best mAP and a 0.1% higher than the best in the comparison networks. These two tables illustrate that most of the ships can be detected by the SRNet and that the RB-Boxes output from SRNet can correctly and tightly surround the ships in SAR images. Moreover, from the instance-segmentation results of our SRNet mentioned later, it can be concluded that the performance of the RB-Boxes output from the Dethead is good enough to meet the foundation of the following segmentation.

Table 1. Rotated object detection performance of different networks on the SSDD dataset.

Networks	Recall	mAP
Rotated Faster RCNN	92.7	89.6
Rotated RetinaNet	93.0	88.6
Rotated Reppoint	91.0	87.5
ROI Trans	92.9	90.1
R3Det	93.4	90.0
ReDet	94.5	90.3
Oriented RCNN	94.5	90.2
S2ANet	93.8	90.0
SRNet	**95.3**	**90.4**

Table 2. Rotated object detection performance of different networks on the RSDD dataset.

Networks	Recall	mAP
Rotated Faster RCNN	88.2	79.2
Rotated RetinaNet	88.3	78.8
Rotated Reppoint	88.5	78.5
ROI Trans	90.4	88.1
R3Det	89.4	80.8
ReDet	**92.5**	89.3
Oriented RCNN	90.8	88.0
S2ANet	91.8	89.0
SRNet	92.3	**89.4**

Figures 11 and 12 show four visualization examples of all the comparison networks on the SSDD dataset and RSDD dataset, respectively. In Figure 11, (a) is the scene of a single ship berthing in a dock, (b) is the scene of parallel ships berthing in a dock, (c) is the scene of multiple parallel ships berthing in a dock, and (d) is the scene of multiple ships sailing in the water region. In Figure 12, (a) is the scene of the ships berthing on one side of a dock, (b) is the scene of the ships berthing on both sides of a dock, and (c) and (d) are the scene of ships sailing on the water region. From the two figures, we can see that the parallel berthing arrangement and the non-ship objects in the dock have a bad influence on the output RB-Boxes. For example, in Figure 11a–c and in Figure 12a,b, there are a large number of missing alarms and false alarms in the results of comparison networks. How to distinguish the different parallel ships is especially a hard barrier to all comparison networks, which predict the parallel ships to one single ship. Because of the interference of textures and shapes of isles or other ship-like objects, it is difficult for comparison networks to detect ships in the water region. In Figures 11d and 12c,d, there are also a large number of missing alarms and false alarms in the results of comparison networks. However, in the results of our SRNet, for the scene of multiple parallel ships berthing in a dock, almost all the parallel ships can be distinguished by the SRNet, and the RB-Boxes almost have the accurate shapes and location, that can be seen in Figures 11a–c and 12a,b. This is mainly because the DAM provides the representative features for regression so that the RB-Boxes can grasp the shapes and locations of ships. Moreover, allocating samples by the AIoU threshold makes the features at the bow and stern trained effectively, which helps distinguish the parallel ships. For the scene of multiple ships sailing in the water region, there are fewer missing alarms or false alarms in the results of the SRNet, and almost all RB-Boxes have the rational classification scores, which can be seen in Figures 11d and 12c,d. This is mainly because the DAM in the SRNet can provide the representative features to produce accurate classification scores so that the ships can be accurately detected, and the docks can be distinguished.

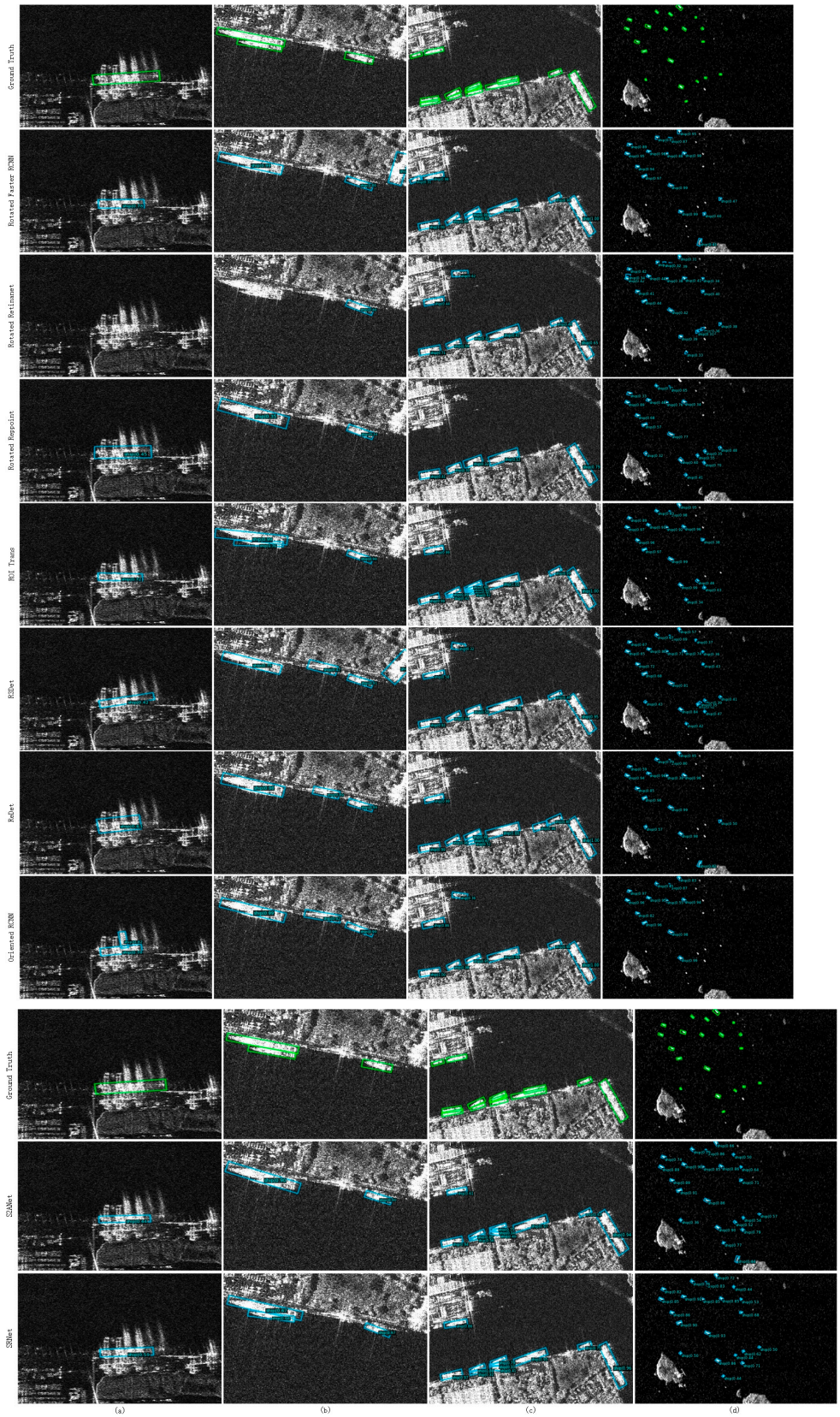


Figure 11. Four visualization illustration images of different comparison networks for rotated object detection on the SSDD dataset. (**a**) The scene of a single ship berthing in a dock. (**b**) The scene of parallel ships berthing in a dock. (**c**) The scene of multiple parallel ships berthing in a dock. (**d**) The scene of multiple ships sailing in the water region.

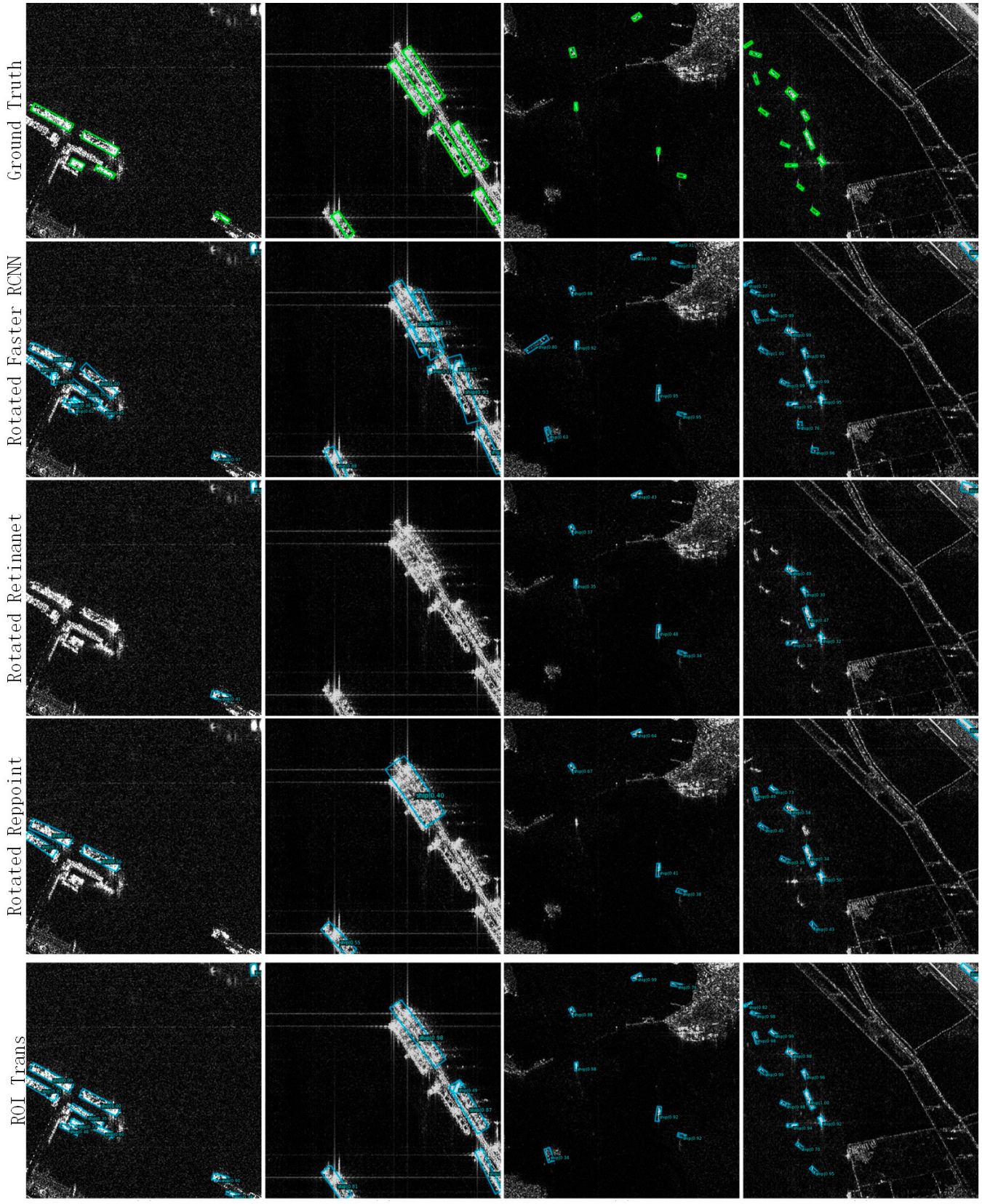


Figure 12. *Cont*.

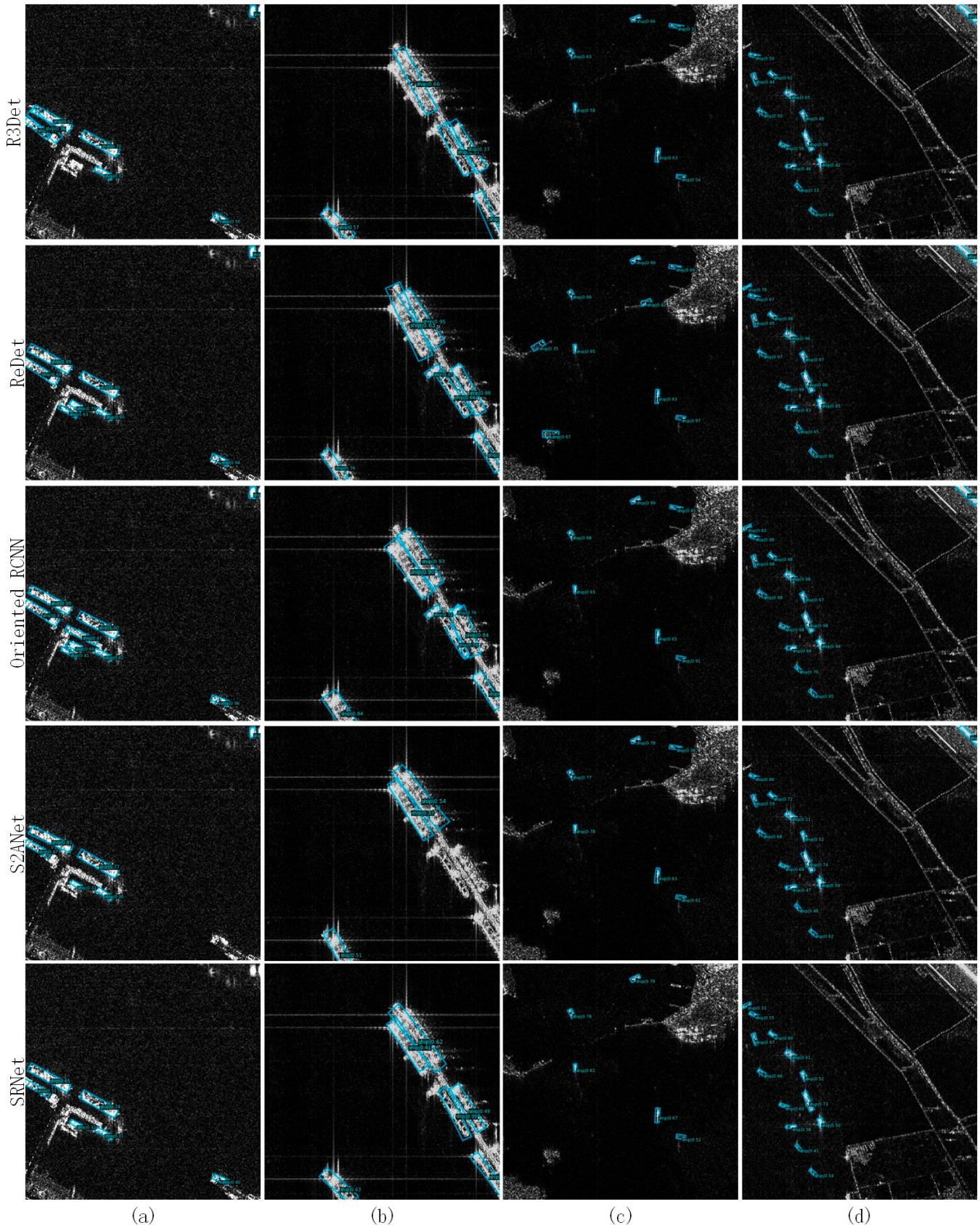


Figure 12. Four visualization illustration images of different comparison networks for rotated object detection on the RSDD dataset. (**a**) The scene of the ships berthing on one side of a dock. (**b**) The scene of the ships berthing on both sides of a dock. (**c**,**d**) The scene of ships sailing in the water region.

4.4. Ship Instance Segmentation in SAR Image

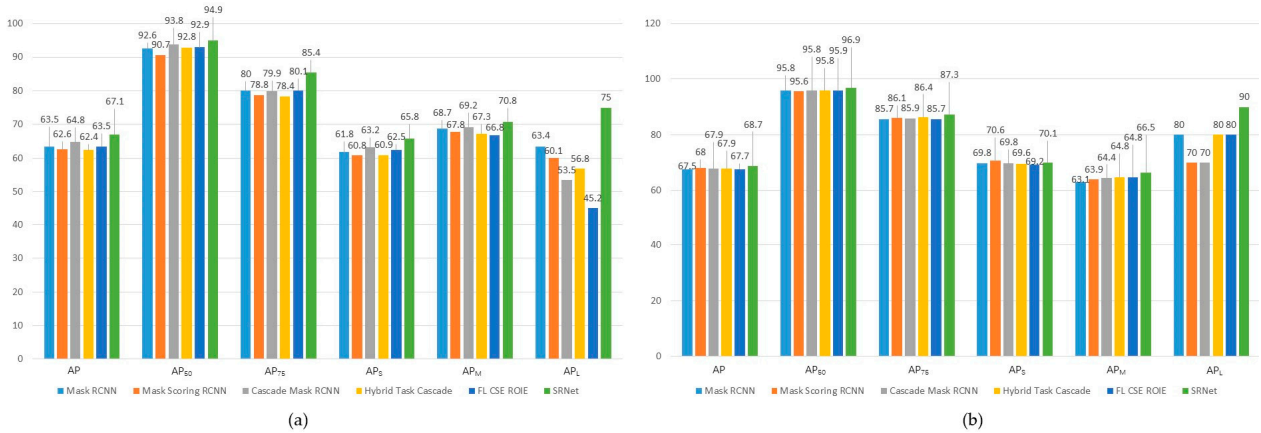
To verify the instance-segmentation performance of our SRNet, five instance-segmentation networks were used for comparison, namely Mask RCNN [4], Mask Scoring RCNN [5], Cascade Mask RCNN [6], Hybrid Task Cascade [7], and FL CSE ROIE [8]. The first four networks were originally designed for optical images, which are frequently used for SAR images. The last network was designed for SAR images. Tables 3 and 4 report the instance-segmentation performance of the different networks on the SSDD and Instance-RSDD, respectively. Our network SRNet achieved the highest AP on the two datasets and outperformed all other state-of-the-art networks in almost all metrics. To enhance the visualization of the two tables, their corresponding histograms are shown in Figure 13. The AP_{75} is hard to improve, which requires a higher overlap ratio with the instance mask ground truth. Meanwhile, the AP_{75} columns of SRNet in the two histograms are both higher than the others. This illustrates that the instance-segmentation results of SRNet are closer to the instance mask ground truth. In addition, it can be seen that the AP_M and AP_L columns of SRNet in the two histograms are much higher than others. This illustrates that the ships with middle scale and large scale can be segmented better by our network. This is mainly because the AIoU threshold provides enough positive samples for the network training, and it is obvious for ships with middle scales and large scales.

Table 3. Instance-segmentation performance of different networks on the SSDD dataset.

Networks	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask RCNN	63.5	92.6	80.0	61.8	68.7	63.4
Mask Scoring RCNN	62.6	90.7	78.8	60.8	67.8	60.1
Cascade Mask RCNN	64.8	93.8	79.9	63.2	69.2	53.5
Hybrid Task Cascade	62.4	92.8	78.4	60.9	67.3	56.8
FL CSE ROIE	63.5	92.9	80.1	62.5	66.8	45.2
SRNet	**67.1**	**94.9**	**85.4**	**65.8**	**70.8**	**75.0**

Table 4. Instance-segmentation performance of different networks on the Instance-RSDD dataset.

Networks	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask RCNN	67.5	95.8	85.7	69.8	63.1	80.0
Mask Scoring RCNN	68.0	95.6	86.1	70.6	63.9	70.0
Cascade Mask RCNN	67.9	95.8	85.9	69.8	64.4	70.0
Hybrid Task Cascade	67.9	95.8	86.4	69.6	64.8	80.0
FL CSE ROIE	67.7	95.9	85.7	69.2	64.8	80.0
SRNet	**68.7**	**96.9**	**87.3**	70.1	**66.5**	**90.0**



(a) (b)

Figure 13. The histograms of instance-segmentation performance on the SSDD dataset (**a**) and the Instance-RSDD dataset (**b**).

Figures 14 and 15 show four visualization instance-segmentation examples of all the comparison networks on the SSDD dataset and RSDD dataset, respectively. In Figure 14, (a) is the scene of parallel ships berthing at a dock, (b) and (c) are the scenes of multiple parallel ships berthing on both sides of a dock, and (d) is the scene of the ships sailing in the water region. In Figure 15, all the images show scenes of the ships berthing in the docks, and there are some non-ship objects with similar visual appearance to ships in the images. In the several segmented images, besides the final instance masks, the HB-Boxes of the comparison networks and the RB-Boxes of our SRNet are marked on the segmented images. From these figures, we can see that there are inherent shortcomings in the results of HB-Box-based instance-segmentation networks. For example, in Figures 14c,d and 15a,b,d, there are a large number of repeated HB-Boxes which are predicted for the same ship, and there are a large number of missing alarms and false alarms in the results of comparison networks. There are two situations of missing alarms, the first one is that the multiple ships are predicted to be one ship by one HB-Box, and the second one is that the ships with low classification scores are wrongly filtered. In Figures 14a–c and 15a–c, there are many phenomena that the instance masks of the parallel ships are connected in one HB-Box, and non-ship objects are predicted to be part of ships in the results of the comparison networks. This illustrates that the multiple ships or non-ship objects being easily contained in one HB-Box is another barrier for the HB-Box-based instance-segmentation networks. However, in the results of our SRNet, for the scenes of multiple parallel ships berthing in a dock, in Figures 14a–c and 15a–c, one RB-Box always contain one ship, the instance masks of ships are separated from each other, and there are almost no phenomena in which the instance masks of the parallel ships are connected. In addition, the docks are almost excluded from the RB-Boxes and not segmented into part of instance masks of ships. This is mainly because the RB-Boxes have the advantage that can distinguish parallel ships and non-ship objects. With lower amounts of other information, instance segmentation based on the RB-Boxes, which is the core idea of our SRNet, can almost avoid the inherent shortcomings in HB-Box-based instance-segmentation networks; thus, the SRNet achieves a high instance-segmentation performance. For the scenes of a single ship berthing in a dock or multiple ships sailing on the water region, in Figures 14d and 15d, there are almost no missing alarms or false alarms, thus illustrating that instance segmentation based on the RB-Boxes, our SRNet, is suitable for most of the complex scenes. Another important point is that, in all scenes, the instance masks are clear and complete; this is mainly because the MaskHead in SRNet can accurately predict the instance masks in the RB-Boxes, causing the ships to be correctly segmented.

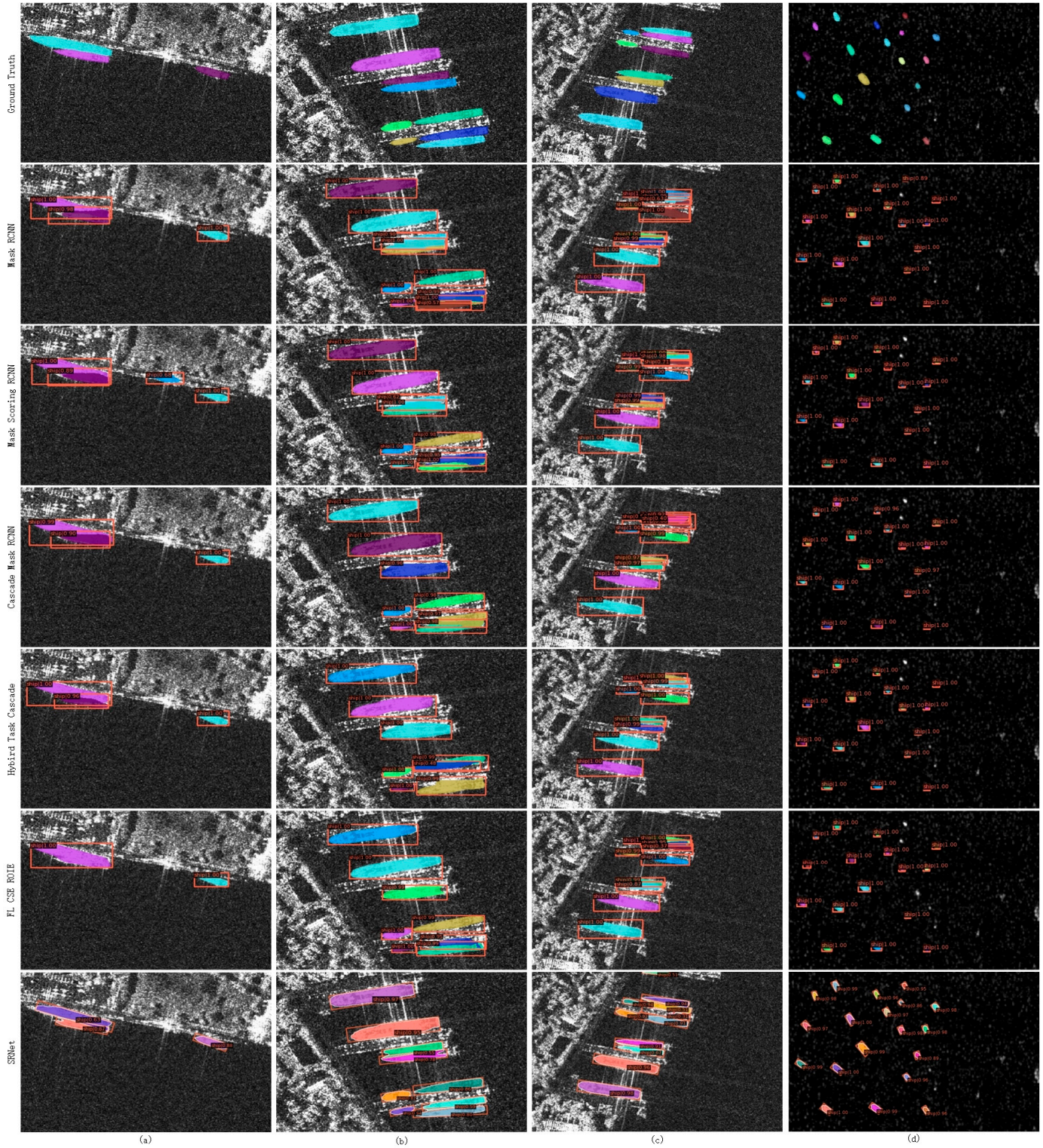


Figure 14. Four visualization illustration images of different comparison networks for instance segmentation on the SSDD dataset. (**a**) The scene of parallel ships berthing at a dock. (**b**,**c**) The scenes of multiple parallel ships berthing on both sides of a dock. (**d**) The scene of multiple ships sailing in the water region.

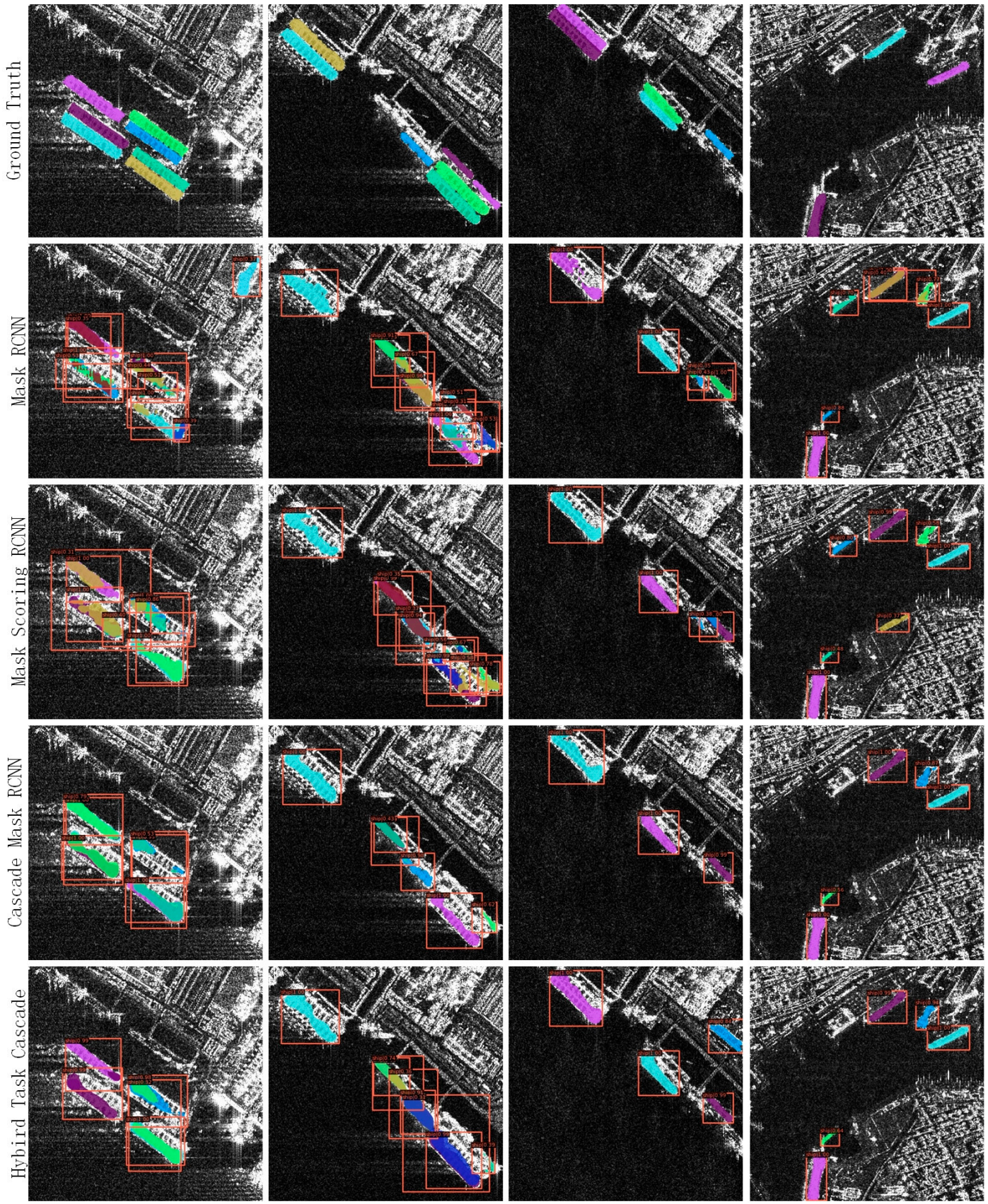


Figure 15. *Cont*.

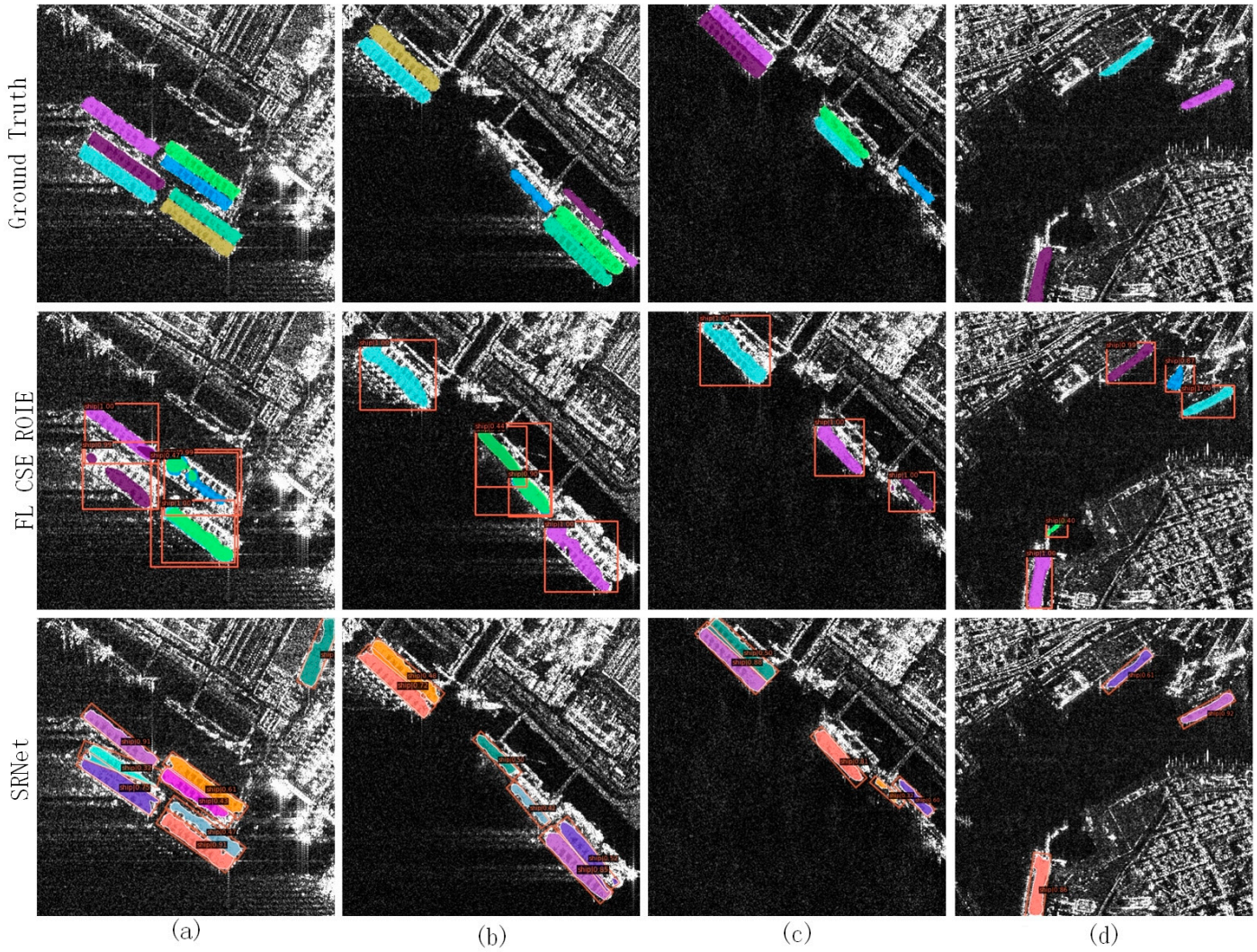


Figure 15. Four visualization illustration images of different comparison networks for instance segmentation on the Instance-RSDD dataset. (**a**–**d**) The scenes of the ships berthing in the docks.

4.5. Ablation Experiments

4.5.1. Ablation for the Components Proposed in SRNet

In order to verify the performance of the proposed DAM and the AIoU calculation method for network training, three ablation studies with different settings were conducted on the SSDD dataset and Instance-RSDD dataset, respectively. To verify the effectiveness of the DAM, based on the instance-segmentation framework based on the RB-Boxes which is proposed in this paper, the single feature map was taken as the input of the DetHead to compare with two feature maps with different properties output from the DAM. Moreover, the single feature map was extracted by a standard convolution and a deformable convolution. To verify the effectiveness of the AIoU, the ATSS was used for network training for comparison. Thus, in Tables 5 and 6, Conv and DeformConv denote that the standard convolution and the deformable convolution are used to provide a single feature map as the input of the DetHead, and DAM denotes that the DAM is used to provide two feature maps as inputs of the DetHead. ATSS denotes using the ATSS to calculate the IoU threshold for network training.

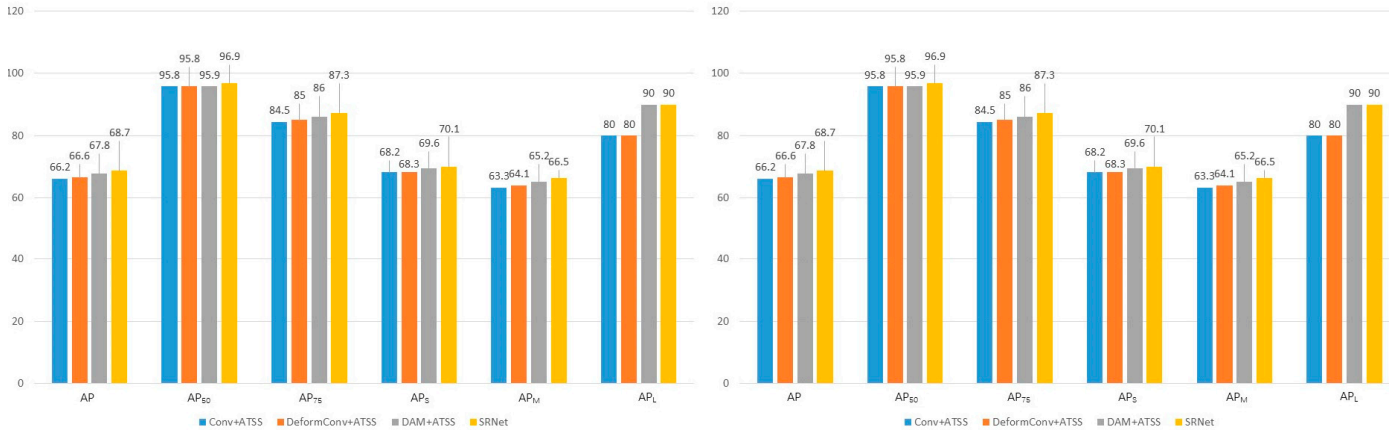
Table 5. Ablation experiment results of our network framework on the SSDD dataset.

Networks	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Conv+ATSS	64.8	94.9	81.5	64.5	66.3	72.5
DeformConv+ATSS	65.6	94.9	81.9	64.4	69.9	67.6
DAM+ATSS	66.9	94.9	84.2	65.7	71.1	65.0
SRNet	**67.1**	**94.9**	**85.4**	**65.8**	70.8	**75.0**

Table 6. Ablation experiment results of our network framework on the Instance-RSDD dataset.

Networks	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Conv+ATSS	66.2	95.8	84.5	68.2	63.3	80.0
DeformConv+ATSS	66.6	95.8	85.0	68.3	64.1	80.0
DAM+ATSS	67.8	95.9	86.0	69.6	65.2	90.0
SRNet	**68.7**	**96.9**	**87.3**	**70.1**	**66.5**	**90.0**

Tables 5 and 6 show the results of the ablation studies. From the whole two tables, we can see that the SRNet, which is equipped with the DAM and trained with the AIoU calculation method, achieves the best performance of almost all metrics on the two generic datasets. To enhance the visualization of the two tables, their corresponding histograms are shown in Figure 16.



(a) (b)

Figure 16. The histograms of ablation experimental results on the SSDD dataset (**a**) and the Instance-RSDD dataset (**b**).

From the two histograms in Figure 16, we can see that almost all the metrics of SRNet are higher than those metrics of DAM+ATSS, thus illustrating that by training the network with the AIoU calculation method, more rational positive samples are used to effectively train the network so that the performance can be improved. From the metrics of Conv+ATSS, DeformConv+ATSS, and DAM+ATSS, we can see that almost all the metrics of DAM+ATSS are higher than others. It indicates that, without training the network with the AIoU calculation method, the DAM+ATSS still achieves a higher performance than other networks with standard convolution and deformable convolution, respectively. However, the AP_L of DAM+ATSS on the SSDD dataset is lower than others, which means that the DAM cannot be trained effectively with the positive samples of large ships obtained by ATSS. In addition, the high metrics of DAM+ATSS illustrate that by considering the difference between the regression task and classification task, the DAM can improve the performance a lot.

4.5.2. Ablation for Different α in the AIoU

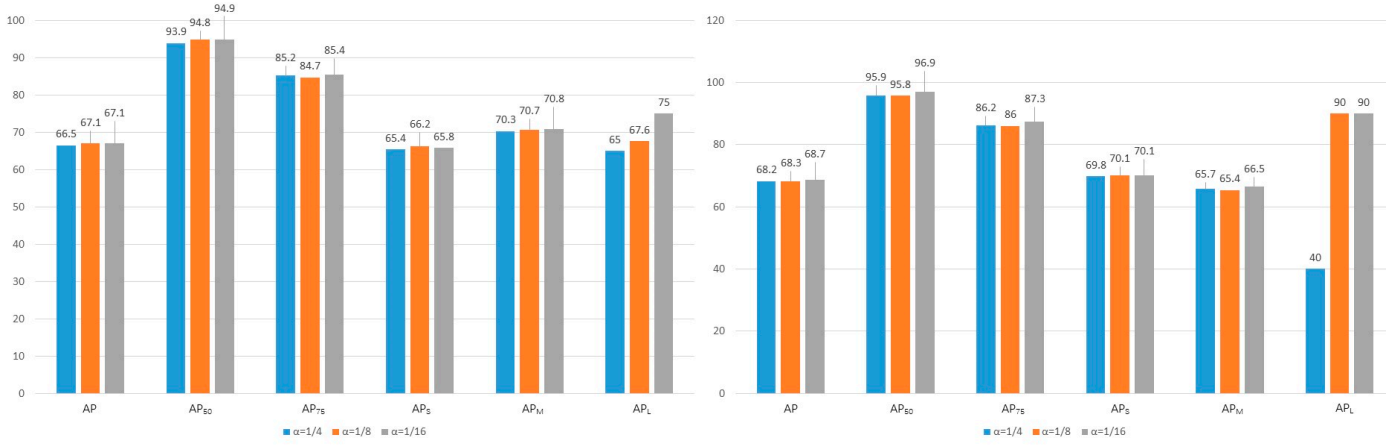
To verify the rationality of the value of factor α, the ablation experiments with different values of α were conducted on the SSDD dataset and Instance-RSDD dataset. As shown in Tables 7 and 8, the value of α is set to 1/4, 1/8, and 1/16. Take the network with α of 1/4 as the baseline, and with the α decreasing, the performance of the network raise. To enhance the visualization of the two tables, their corresponding histograms are shown in Figure 17. From the two histograms, we can see that almost all the metrics columns of $\alpha = 1/16$ are higher than those columns of other α settings. The metrics AP_L especially increase obviously. Thus, our parameter setting, $\alpha = 1/16$, is reasonable.

Table 7. Ablation experiment results of factor α on the SSDD dataset.

α	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1/4	66.5	93.9	85.2	65.4	70.3	65.0
1/8	67.1	94.8	84.7	**66.2**	70.7	67.6
1/16	**67.1**	**94.9**	**85.4**	65.8	**70.8**	**75.0**

Table 8. Ablation experiment results of factor α on the Instance-RSDD dataset.

α	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1/4	68.2	95.9	86.2	69.8	65.7	40.0
1/8	68.3	95.8	86.0	70.1	65.4	90.0
1/16	**68.7**	**96.9**	**87.3**	**70.1**	**66.5**	**90.0**



(a) (b)

Figure 17. Histograms of the ablation experimental results of factor α on the SSDD (**a**) and the Instance-RSDD dataset (**b**).

5. Conclusions

In this paper, we proposed a novel ship instance-segmentation network based on RB-Boxes called SRNet for ocean SAR images. To extract features aligned with the directions and shapes of ships, the DAM in the detection procedure was designed with two different alignment convolutions for the regression task and classification task. In the two alignment convolutions, the two sampling strategies of their kernels were proposed to enable the sampling locations of the kernels distributed on the area that needs to focus. By taking the regression feature maps and classification feature maps from the DAM as inputs, the DetHead achieves accurate rotated object detection. Then, with the RB-Boxes output from DetHead after the rotated NMS, the MaskHead in the segmentation procedure obtains

instance masks by two sampling processes and a convolution, and then it transfers them onto the original images. In the training process, the aspect ratio of each ship is taken into account to calculate the AIoU for the corresponding ground-truth box so that the training effect is improved. In experiments, the accuracies of RB-Boxes and instance masks are evaluated, respectively. The results of these comparison experiments demonstrate that not only does the RB-Boxes from the DetHead outperform the results of the eight rotated object detection networks, but also the instance masks from the MaskHead outperform the results of the five instance-segmentation networks. Moreover, ablation experiments were conducted to verify the effectiveness of the DAM and the AIoU and the rationality of the factor α setting, thus demonstrating that the DAM in feature extraction and the AIoU in network training can produce a positive effect on the network performance. Moreover, the factor α setting in AIoU is rational.

Author Contributions: Conceptualization, X.Y., Q.Z. and Q.D.; methodology, X.Y. and Q.Z.; software, X.Y.; validation, X.Y.; formal analysis, X.Y.; data curation, X.Y., Z.H., X.L. and D.W.; writing—original draft preparation, X.Y. and Q.Z.; writing—review and editing, X.Y. and Q.Z.; funding acquisition, Q.Z. and Q.D. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the National Natural Science Foundation of China (grant nos. 61403294 and U1805264).

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. 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]
2. 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. [CrossRef]
3. Su, H.; Wei, S.; Liu, S.; Liang, J.; Wang, C.; Shi, J.; Zhang, X. HQ-ISNet: High-Quality Instance Segmentation for Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 989. [CrossRef]
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
5. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
6. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *arXiv* **2019**, arXiv:1906.09756. [CrossRef] [PubMed]
7. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Loy, C.C. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.
8. Zhang, T.; Zhang, X. A Full-Level Context Squeeze-and-Excitation ROI Extractor for SAR Ship Instance Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4506705. [CrossRef]
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT PRESS: Cambridge, MA, USA, 2015; pp. 91–99.
10. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
11. Guo, Z.; Liu, C.; Zhang, X.; Jiao, J.; Ji, X.; Ye, Q. Beyond Bounding-Box: Convex-Hull Feature Adaptation for Oriented and Densely Packed Object Detection. In Proceedings of the CVPR 2021, Online, 19–25 June 2021; pp. 8792–8801.
12. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2849–2858.
13. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *AAAI Conf. Artif. Intell.* **2021**, *35*, 3163–3171. [CrossRef]
14. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2021; pp. 2785–2794.

15. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021.
16. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
17. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017.
18. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
19. Zhang, T.; Zhang, X. HTC+ for SAR Ship Instance Segmentation. *Remote Sens.* **2022**, *14*, 2395. [CrossRef]
20. Kang, M.; Ji, K.; Leng, X.; Lin, Z. Contextual Region-Based Convolutional Neural Network with Multilayer Fusion for SAR Ship Detection. *Remote Sens.* **2017**, *9*, 860. [CrossRef]
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation 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. 7132–7141.
22. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
23. 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]
24. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware Reassembly of Features. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2019; Volume 2019, pp. 3007–3016.
25. Marino, A.; Sanjuan-Ferrer, M.J.; Hajnsek, I.; Ouchi, K. Ship Detection with Spectral Analysis of Synthetic Aperture Radar: A Comparison of New and Well-Known Algorithms. *Remote Sens.* **2015**, *7*, 5416–5439. [CrossRef]
26. Zhang, T.; Quan, S.; Yang, Z.; Guo, W.; Zhang, Z.; Gan, H. A Two-Stage Method for Ship Detection Using PolSAR Image. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
27. Leng, X.; Ji, K.; Zhou, S.; Xing, X. Fast shape parameter estimation of the complex generalized Gaussian distribution in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1933–1937. [CrossRef]
28. Tang, G.; Zhao, H.; Claramunt, C.; Men, S. FLNet: A Near-shore Ship Detection Method Based on Image Enhancement Technology. *Remote Sens.* **2022**, *14*, 4857. [CrossRef]
29. Zhou, Q.; Yu, C. Point RCNN: An Angle-Free Framework for Rotated Object Detection. *Remote Sens.* **2022**, *14*, 2605. [CrossRef]
30. Maurice, W.; Gabriele, C. General E(2)-Equivariant Steerable CNNs. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, WA, USA, 8–14 December 2019; pp. 14334–14345.
31. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 9759–9768.
32. Yu, Y.; Yang, X.; Li, J.; Gao, X. A Cascade Rotated Anchor-Aided Detector for Ship Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
33. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 10183–10192.
34. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 2921–2929.
35. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
36. Kim, J.U.; Ro, Y.M. Attentive Layer Separation for Object Classification and Object Localization in Object Detection. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–29 September 2019; pp. 3995–3999.
37. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
38. 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.
39. Zhang, T.; Zhang, X.; Li, J.; Xu, 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]
40. Xu, C.; Su, H.; Li, J.; Liu, Y.; Yao, L.; Gao, L.; Yan, W.; Wang, T. RSDD-SAR: Rotated ship detection dataset in SARimages. *J. Radars* **2022**, *11*, 581–599.
41. Instance-RSDD. Available online: https://github.com/XIDIAN212Lab/SRNet (accessed on 23 December 2022).
42. Wang, P.; Niu, Y.; Wang, J.; Ma, F.; Zhang, C. Arbitrarily Oriented Dense Object Detection Based on Center Point Network in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1536. [CrossRef]

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. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
44. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. MMRotate: A Rotated Object Detection Benchmark using PyTorch. *arXiv* **2022**, arXiv:2204.13317.
45. 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.
46. 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.