Journal of
Marine Science and Engineering

MDPI

Article

Enhanced Detection Method for Small and Occluded Targets in Large-Scene Synthetic Aperture Radar Images

Hui Zhou [1], Peng Chen [2,*], Yingqiu Li [1] and Bo Wang [1]

[1] School of Computer and Software, Dalian Neusoft Information University, Dalian 116023, China; zhouhui@neusoft.edu.cn (H.Z.); liyingqiu@neusoft.edu.cn (Y.L.); wangbo@neusoft.edu.cn (B.W.)
[2] Navigation College, Dalian Maritime University, Dalian 116026, China
* Correspondence: chenpeng@dlmu.edu.cn

Abstract: Ship detection in large-scene offshore synthetic aperture radar (SAR) images is crucial in civil and military fields, such as maritime management and wartime reconnaissance. However, the problems of low detection rates, high false alarm rates, and high missed detection rates of offshore ship targets in large-scene SAR images are due to the occlusion of objects or mutual occlusion among targets, especially for small ship targets. To solve this problem, this study proposes a target detection model (TAC_CSAC_Net) that incorporates a multi-attention mechanism for detecting marine vessels in large-scene SAR images. Experiments were conducted on two public datasets, the SAR-Ship-Dataset and high-resolution SAR image dataset (HRSID), with multiple scenes and multiple sizes, and the results showed that the proposed TAC_CSAC_Net model achieves good performance for both small and occluded target detection. Experiments were conducted on a real large-scene dataset, LS-SSDD, to obtain the detection results of subgraphs of the same scene. Quantitative comparisons were made with classical and recently developed deep learning models, and the experiments demonstrated that the proposed model outperformed other models for large-scene SAR image target detection.

Keywords: large-scene SAR image; occlusion targets detection; small target detection; multi-attention mechanism

check for updates

Citation: Zhou, H.; Chen, P.; Li, Y.; Wang, B. Enhanced Detection Method for Small and Occluded Targets in Large-Scene Synthetic Aperture Radar Images. *J. Mar. Sci. Eng.* **2023**, *11*, 2081. https://doi.org/10.3390/jmse11112081

Academic Editors: Fausto Pedro García Márquez and Coro Gianpaolo

Received: 23 September 2023
Revised: 27 October 2023
Accepted: 27 October 2023
Published: 30 October 2023

1. Introduction

As an important target for maritime monitoring, maritime management, and wartime tracking, the accuracy requirements for ship detection at sea are increasing [1]. Synthetic aperture radar (SAR), which is not affected by weather, has a large imaging area and a constant resolution when it is far away from the observed target, and has become an important means of detecting ship targets at sea [2].

Most traditional ship detection methods are based on manually extracted features [3]. With the popularization and development of deep learning theory, deep learning models have been widely used in ship detection based on SAR images [4]. Compared with the general optical image target detection task, SAR images are usually acquired in bad weather and complex marine environments, and there is a large amount of background interference in the images. Additionally, in the application of SAR image target detection, the ship target will exist near islands, offshore ports, and buildings. Owing to the complex environment, there are varying degrees of occlusion, the occluded ship image shows irregular shapes, and the detection accuracy is significantly reduced. To address the practical problems in the application of SAR image ship detection, numerous scholars have conducted relevant research. Wenxu et al. (2020) proposed a multi-scale feature fusion single-shot ship target detection model that used deconvolution and pooling layers to enhance the accuracy of feature extraction [5]. Using YOLOv5, Li et al. (2021) added feature refinements and multi-feature fusion to reduce false-alarm rates [6]. Wenping et al. (2022) applied pixel-level denoising and semantic enhancement to reduce missed and false detections [7], and Liu et al. (2021) added a coordinate attention mechanism (AM) to YOLOv5 to handle high

aspect ratios and dense arrangements and optimize its loss function [8]. Occlusions at sea are problematic for deep learning object detection for these two reasons: multiple targets may occlude one another, or they may be occluded by geographical features and other interfering signals. Researchers have proposed several solutions. Tian et al. provided a pool of convolutional neural network (CNN) components that act as subdetection networks. The final integrated results were then characterized [9]. Ouyang et al. used pattern mining to extract the local features of a target to further train the local feature detector. These iteratively trained detectors can be embedded in a CNN to overcome occlusions [10].

Despite the progress in target detection research for the occlusion problem, a series of problems, such as the unsatisfactory optimization effect and high time complexity of the algorithm, remain. Ship detection in SAR images with a large number of small occluded targets has a high false detection rate when the target is occluded by a building onshore or port, or when there is mutual overlap between ships. If the scale of the occluded ship is small (less than 100×100 pixels), it will likely not be detected accurately. To solve the problem of small-scale ship detection in multi-scene SAR images, Jiao et al. introduced a densely connected network for multi-scale feature fusion and reduced the weight of non-small target samples in the loss function using focus loss [11]. Sun et al. added the atrous convolutional pyramid module and the multi-scale attention mechanism module for multi-scale marine ships' description and segmentation, and the proposed category-position module optimized position regression [12]. Yang et al. enhanced the RetinaNet architecture for forecasting rotatable bounding boxes. They employed diverse techniques to tackle challenges such as feature scale inconsistency, incongruity among distinct learning tasks, and an imbalanced distribution of positive samples within SAR ship detection [13].

However, in practical engineering applications, ship detection in large-scene SAR images is closer to the actual application of global ship surveillance, and the fast detection of multi-scale, occluded ship targets derived from large-scene SAR images remains a challenge. The aforementioned studies were only performed on small-scene datasets with small SAR image slices, such as BBox-SSDD, SSDD, and the high-resolution SAR image dataset (HRSID), which means that the detection models trained on these small-scene datasets are difficult to directly apply to large-scene marine surveillance images with wide mapping areas in real engineering applications, which affects model practicability. Additionally, in large-scene marine surveillance SAR images with wide mapping areas, ship sizes tend to be smaller; however, with the variety of ship sizes in other existing datasets, they do not correspond to small-size scenarios in real-world scenarios, which can lead to accuracy degradation of the detection model when migrating to generalization in large scenes. Therefore, this study primarily focuses on the relatively difficult-to-detect, occluded, and small-sized targets in large scenes and proposes a target detection model that fuses multiple attention mechanisms. First, we establish the backbone network of multi-feature fusion and the self-attention mechanism module, the transform attention component (TAC), in the backbone network to deal with global information to obtain better perceptual ability and target object feature abstraction ability. For the feature mapping subgraph, a multi-scale feature complex fusion structure is used to integrate shallow localization features with deeper semantic features, and the channel and spatial attention component (CSAC) is added to integrate the feature space and channel information in two dimensions. A GIoU-based loss function is also used. Finally, the model is tested on the large-scale SAR image dataset LS-SSDD, high-resolution SAR image dataset (HRSID), and multi-scale SAR-Ship-Dataset. The experimental results demonstrate that the improved model can automatically recognize and detect small targets in SAR images under various scenarios with high accuracy.

2. Related Work

With the increasing amount of available data and the rapid development of computing power, deep learning is playing an increasingly important role in SAR image target detection. Scholars have continuously improved and optimized the algorithm and model

structure based on CNNs to enhance the detection effect. In order to solve the problems of multiple scales, small targets, occluded targets, and complex scenes in images, different attention mechanisms and methods are brought into the model. For illustration, Dense Attention Pyramid Networks (DAPN) [14], the Attention Receptive Pyramid Network (ARPN) [15], the Convolutional Block Attention Module R-CNN (CBAM Faster R-CNN) [16], and the Quad Feature Pyramid Network (Quad-FPN) [17] adopt the attention mechanism to enhance the local features, and Double-Head R-CNN [18] is used to focus on the classification and localization tasks by utilizing the fully connected head and convolutional head, respectively. Compare these to our model, which introduces both self-attention and global attention mechanisms. Table 1 shows the descriptions of the state of the art.

Table 1. Descriptions of existing models and our method.

Models	Characteristics
DAPN	DAPN utilizes a pyramid structure in which the Convolutional Block Attention Module (CBAM) is densely connected to each concatenated feature map, creating a network that extends from top to bottom. This design aids in the filtration of negative objects and the suppression of interference from the surrounding environment in the top-down pathway of lateral connections.
ARPN	ARPN is a two-stage detector designed to improve the performance of detecting multi-scale ships in SAR images. It represents the Receptive Fields Block (RFB) and utilizes it to capture characteristics of multi-scale ships with different directions. RFB enhances local features with their global dependences.
Double-Head R-CNN	R-CNN based detectors often use Double-Head R-CNN (fully connected head and convolutional head) for classification and localization tasks. A Double-Head method is proposed where one fully connected head is responsible for classification, while one convolutional head is used for bounding box regression.
CBAM Faster R-CNN	CBAM Faster R-CNN utilizes channel and spatial attention mechanisms to enhance the significant features of ships and suppress interference from surroundings.
Quad-FPN	Quad-FPN is a two-stage detector designed to improve the performance of detecting ships in SAR images. It consists of four unique Feature Pyramid Networks (FPNs). These FPNs are well-designed improvements that guarantee Quad-FPN's excellent detection performance without any unnecessary features. They enable Quad-FPN's excellent ship scale adaptability and detection scene adaptability.
Our model	Unlike the previous model, we establish a backbone network of multi-feature fusion and a self-attention mechanism module. We also introduce the transform attention component and the channel and spatial attention component. Additionally, we use a GIoU-based loss function.

3. A Target Detection Model Incorporating Multiple Attention Mechanisms

3.1. Multi-Feature Fusion-Based Backbone Network

The fundamental idea of a multi-feature fusion backbone network is to fuse the features extracted in the deep network at different scales to form a feature pyramid that makes the models more receptive to multimodal target attributes. First, a backbone ResNet network is used to obtain a feature map via bottom-to-top convolutions (C1, C2, C3, C4, C5). A feature pyramid layer then upsamples the results in a top-down manner and laterally connects them using a 1×1 convolution kernel (256 channels) to form a new feature map (M2, M3, M4, M5), where

M5 = C5.Conv (256, (1, 1))
M4 = Upsampling (M5) + C4.Conv (256, (1, 1))
M3 = Upsampling (M4) + C3.Conv (256, (1, 1))
M2 = Upsampling (M3) + C2.Conv (256, (1, 1))

Finally, to eliminate several confounding effects, a 3×3 convolution is used to obtain the feature maps from M2–M5, and a new feature map (P2, P3, P4, P5); the structure of the multi-feature fusion backbone network is illustrated in Figure 1.

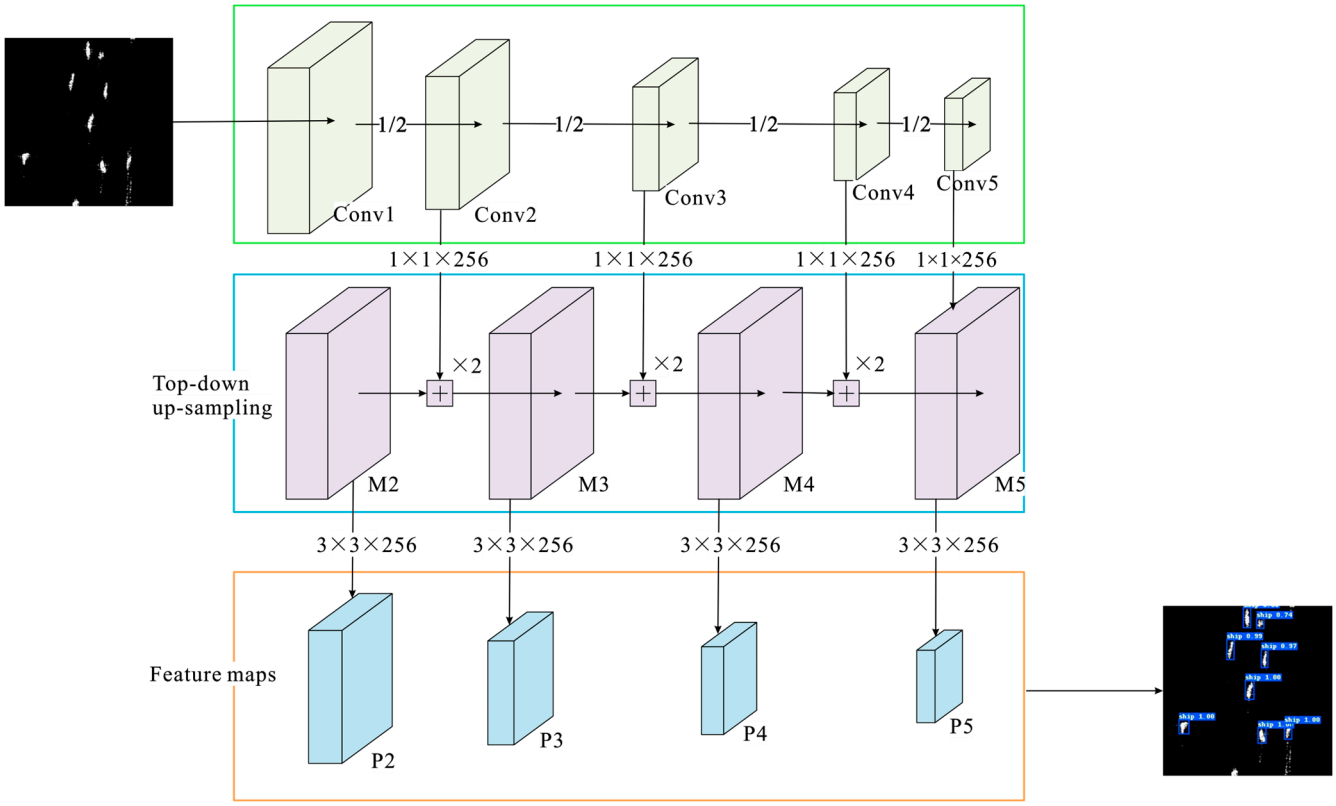

Figure 1. Multi-feature fusion-based backbone network.

3.2. Transform Attention Component (TAC)

The self-attentive component transformer is a stacked model architecture of multiple encoders and decoders that computes global correspondences between outputs and inputs using a unique multi-head attention mechanism. The extracted feature map and position encoding are used as inputs. The feature map is expanded into a one-dimensional sequence of $H \times W$ (height $\times$ width) features to be passed to the encoder, and the target-level information is extracted by the mechanism of mutual attention. This not only improves attention to the target region but also reduces background interference by focusing on the overall input information [19].

The working principle of TAC is to reconstruct image features and key matrices by multiplying the input feature sequence with different weighting matrices. $\boldsymbol{Q}$ is the query matrix, that is, the feature matrix of the image; $\boldsymbol{K}$ is the key matrix; and $\boldsymbol{V}$ is the value matrix.

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{Feature} \cdot \boldsymbol{W}_q, \\ \boldsymbol{K} &= \boldsymbol{Feature} \cdot \boldsymbol{W}_k, \\ \boldsymbol{V} &= \boldsymbol{Feature} \cdot \boldsymbol{W}_v, \end{aligned} \tag{1}$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ are the learnable weights of different matrices. The feature matrix $\boldsymbol{Q}$ and key matrix $\boldsymbol{K}$ are multiplied using the softmax method to obtain the attention matrix. To further realize numerical aggregation weighted by the attention weights, the attention matrix was multiplied by $\boldsymbol{V}$ to obtain the correlation between the targets in the SAR image. Finally, it can be represented in the TAC by multiple heads of attention.

$$\mathrm{TAC}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sum_{m=1}^{M} \left[\left(S_{soft\ \max} \left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^T}{\sqrt{d_k}} \right) \right) \cdot \boldsymbol{V} \right]. \tag{2}$$

where M represents the number of attention heads and d_k is the dimension of $\boldsymbol{K}$. In the feature extraction phase of the backbone network, the input image is segmented into multiple subfeature maps. Some subfeature maps are selected for processing by the TAC

module, which utilizes a self-attention mechanism to model the associations and extract features from different regions in the feature maps. The TAC module can capture global contextual information and learn the dependencies between different locations in the feature maps, which can convert the input feature maps into more expressive subfeature maps. As shown in Figure 2.

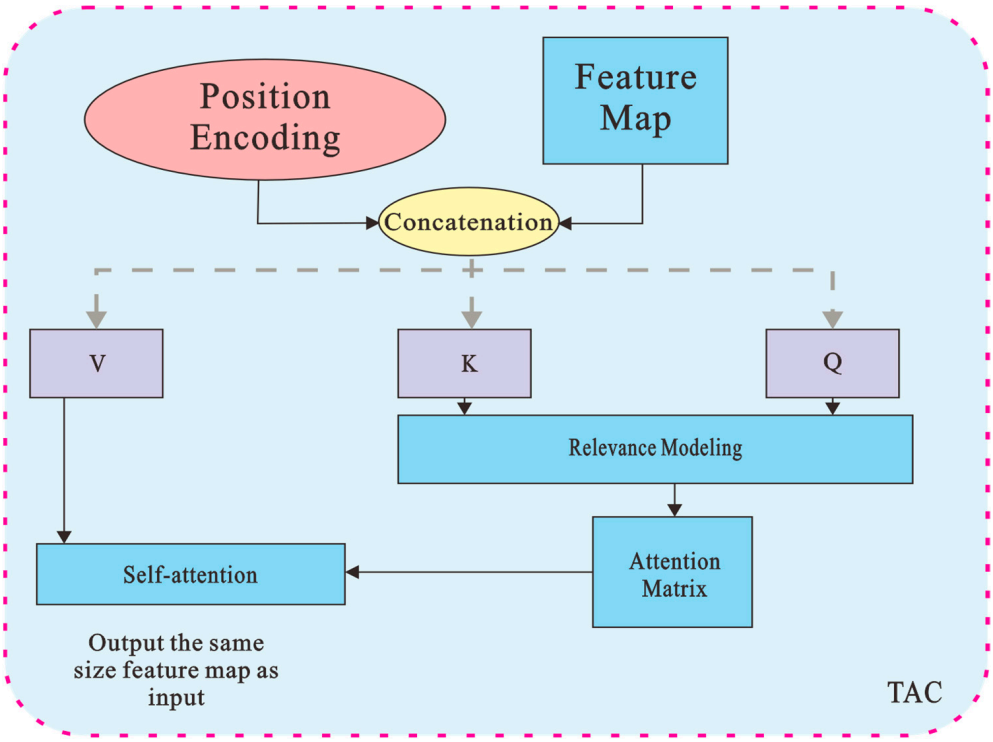

Figure 2. Transform attention component.

3.3. Channel and Spatial Attention Component (CSAC)

Transformer-based attentional mechanisms with deep semantic features have a larger sensory field; however, a larger downsampling factor results in a loss of positional information. In addition to the transformer-based self-attention mechanism used to form a feature map that focuses on interrelationships, attentional mechanisms include channel attention, pixel attention, multilevel attention, and other methods of focusing on key features [8,20].

Channel attention is used to bring the attention of the CNN to the channel dimensions. Hence, the input feature layer $\mathbf{F}$ ($H \times W \times C$) provides the average pooling (Avg-Pool) and maximum pooling (Max-Pool) operations, after which it is compressed into a vector ($H \times W \times 1$). Subsequently, using two fully connected (TFC) layers, the vector is mapped to the weight and bias vectors. Finally, the corresponding weight of each channel is calculated using the activation function, and a new feature map, $\mathbf{F_c}$, is generated, which accounts for the importance of different feature channels.

$$\begin{aligned} \mathbf{F_{avg}} &= \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Avg - Pool(\mathbf{F}), \\ \mathbf{F_{max}} &= \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Max - Pool(\mathbf{F}), \\ \mathbf{F_c} &= sigmoid(TFC(\mathbf{F_{avg}}) + TFC(\mathbf{F_{max}})) \end{aligned} \tag{3}$$

Based on the calculation results from the channel attention, the spatial attention performs average and maximum pooling operations on $\mathbf{F_c}$. The compressed feature layer is then focused on the most useful data of the spatial region, and this vector is converted into a weight matrix using TFC. A new feature map, $\mathbf{F_s}$, indicates the importance of different spatial positions. Figure 3 depicts the computation process of each attention map.

$$\begin{aligned} \mathbf{F_{c\_avg}} &= \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Avg-Pool(\mathbf{F_c}), \\ \mathbf{F_{c\_max}} &= \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Max-Pool(\mathbf{F_c}), \\ \mathbf{F_s} &= sigmoid(TFC(\mathbf{F}_{c\_avg})+TFC(\mathbf{F_{c\_max}})) \end{aligned} \tag{4}$$

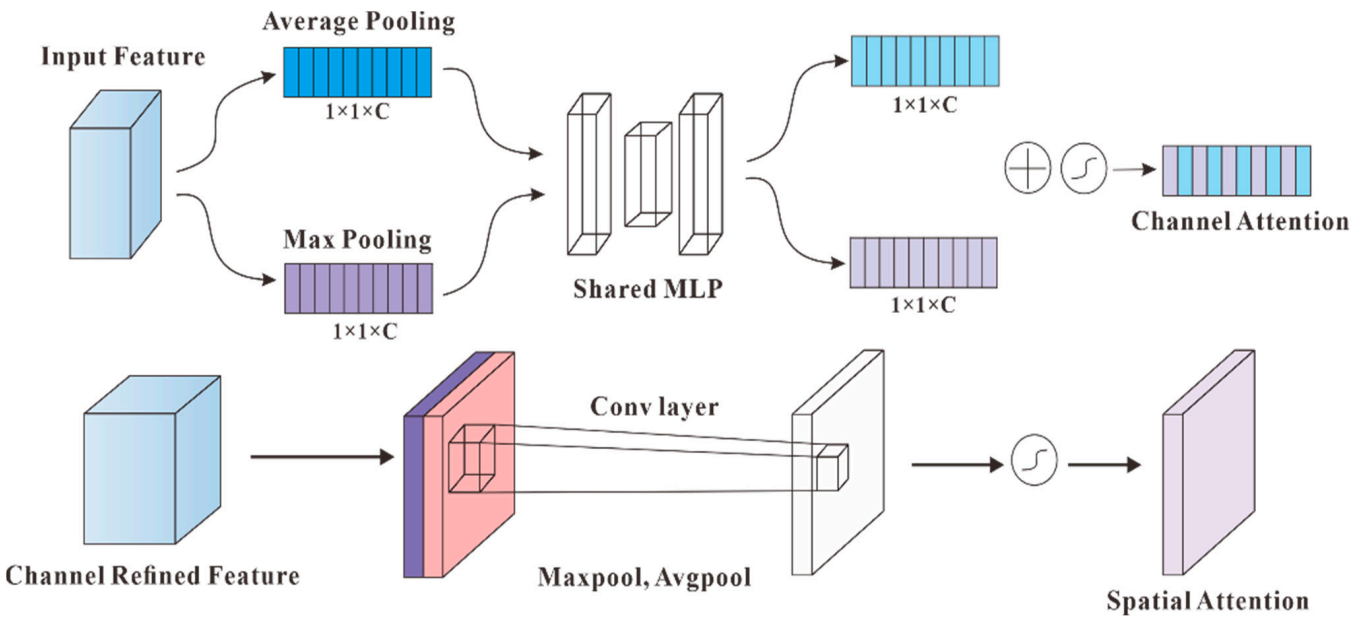

Figure 3. Channel attention and spatial attention.

The CSAC module integrates feature space and feature channel information in two dimensions by introducing spatial and channel attention, as shown in Figure 4. In the target detection model, the feature mapping layers at various scales employ the channel attention mechanism to determine the feature dependencies between different channel maps and calculate the weighted values for all channel maps. The spatial attention mechanism was used to weigh each spatial location in the feature mapping layer to strengthen the model's ability to perceive and utilize spatial location features.

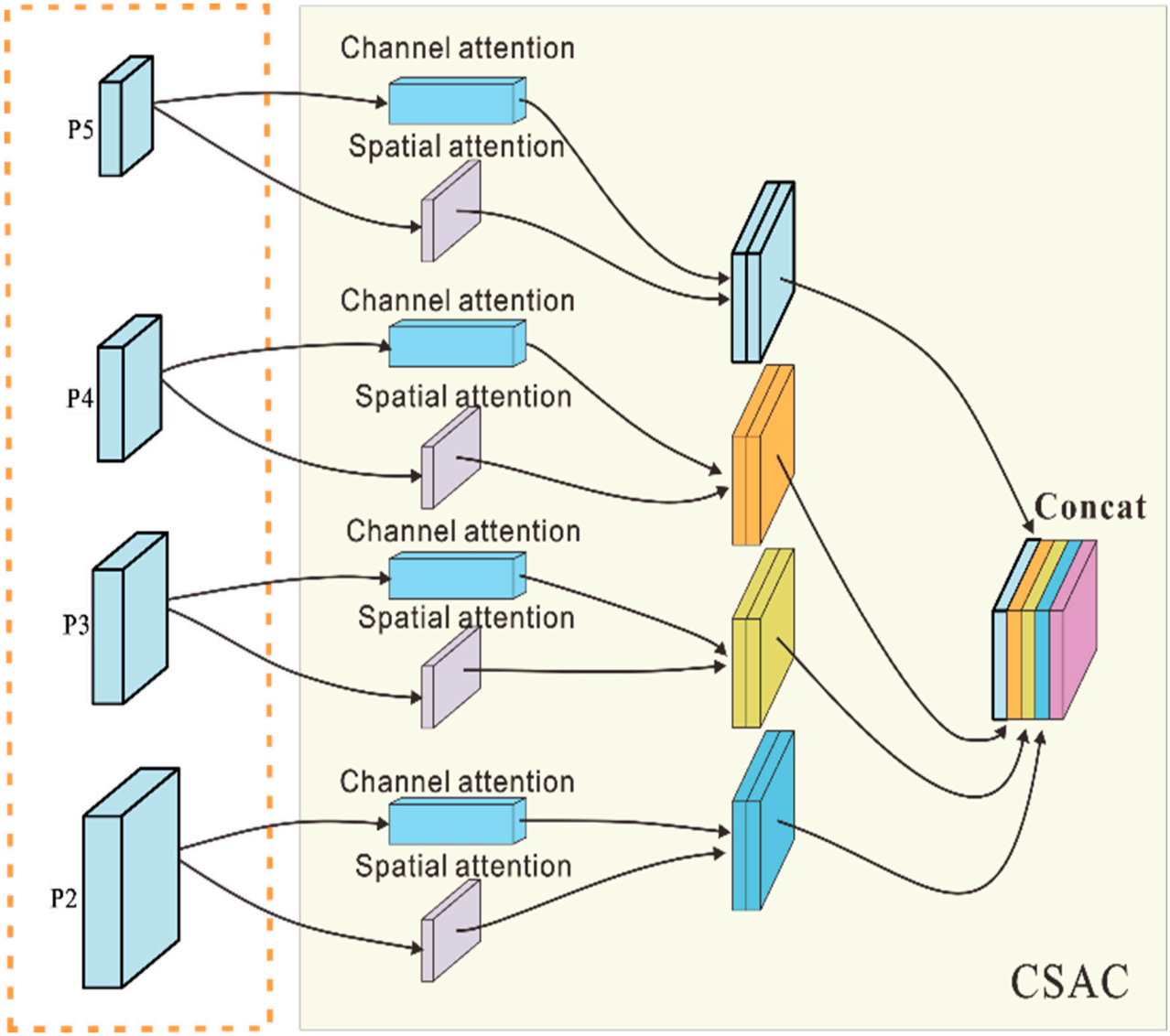

Figure 4. Channel and spatial attention component.

3.4. TAC_CSAC_Net

The backbone network of feature fusion is established in the TAC_CSAC_Net model, which adopts a multi-attention mechanism to maximize the mining of target features and their associations with each other in a large-scene image, whereas the GIoU loss function is used to optimize the traditional target detection loss function. Its multi-attention mechanism is primarily reflected in the use of TAC to process the global information and obtain the correlation between different pixel points and in the use of CSAC to integrate the feature space and feature channel information in two dimensions. The structure of TAC_CSAC_Net is shown in Figure 5.

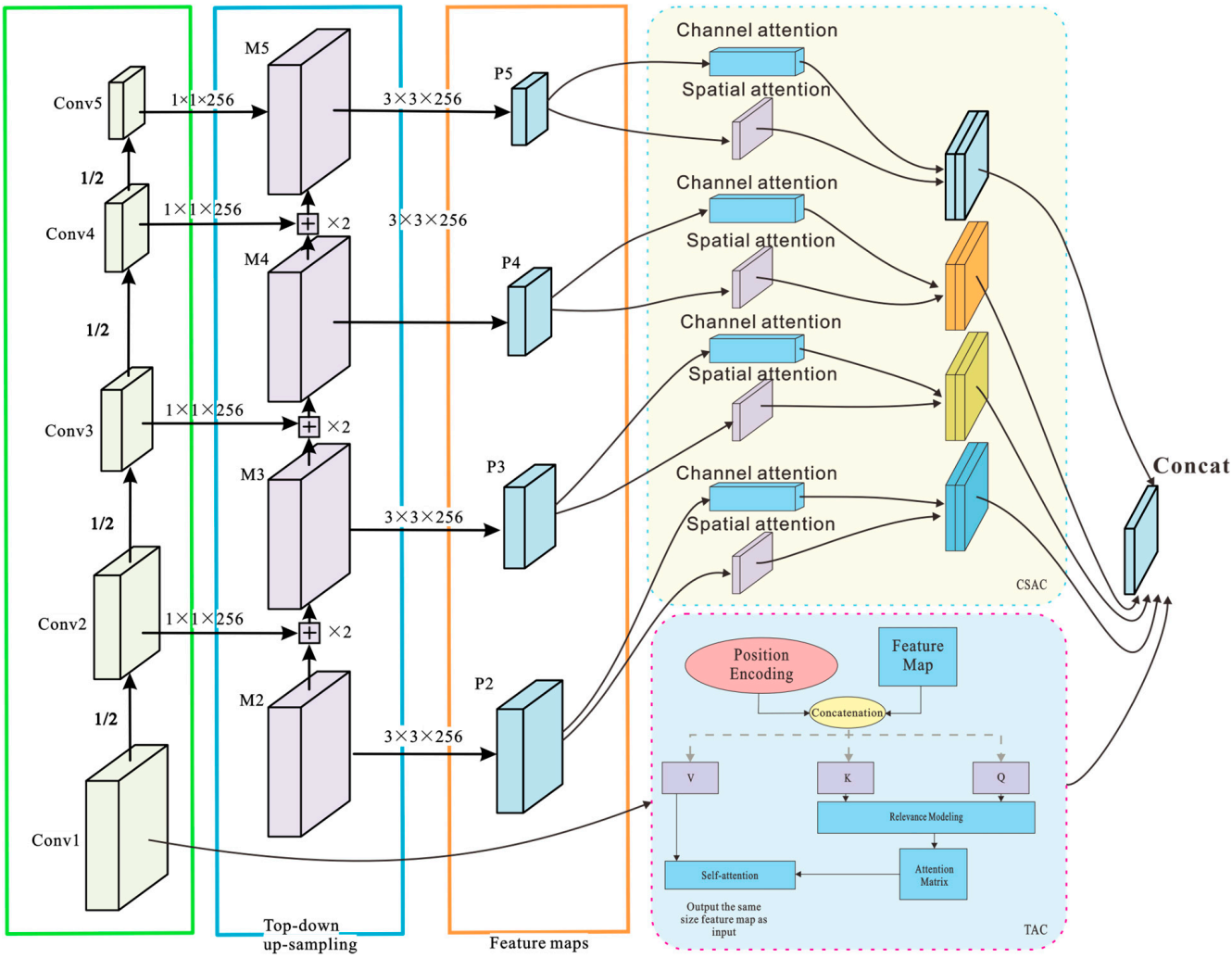

Figure 5. Illustration of TAC_CSAC_Net ship detection in large-scene framework for SAR image. TAC_CSAC_Net is designed with transformer self-attention component (TAC) and channel and spatial attention component (CSAC) at the bottom of the encoder–decoder architecture to perform integration of local features.

The overall loss of the model included L_{class} and L_{box} bounding-box losses. The cross-entropy loss feeds the classification loss, and the bounding box loss includes the L_1 loss between the true value, b_i, predicted $\hat{b}_{\sigma(i)}$, and L_{GIoU} losses.

$$L_{class} = -\frac{1}{N}\sum_{i=1}^{N}\left(p_i \log \hat{p}_{\hat{\sigma}_i} + (1-p_i)\log\left(1-\hat{p}_{\hat{\sigma}_i}\right)\right) \tag{5}$$

$$L_{box}\left(y_i, \hat{y}_{\sigma(i)}\right) = \sum_{1}^{N}\left[\lambda_1\left\|b_i - \hat{b}_{\sigma(i)}\right\| + \lambda_{GIoU} L_{GIoU}\left(b_i, \hat{b}_{\sigma(i)}\right)\right] \tag{6}$$

The IoU reflects the intersection and concatenation ratio between the prediction frame and the real frame; the larger the IoU, the greater the coincidence of the predicted and real boxes. Hence, IoU can be used as an optimization function [21]. The GIoU describes the minimum bounding box required for optimization when the gradient is zero (i.e., the predicted and real boxes do not overlap). Assuming that the coordinates of the true box are **gt** and the coordinates of the predicted box are **pb**, the IoU is obtained by Algorithm 1.

Algorithm 1 Calculating the IoU and GIoU loss functions

Input: Coordinates of the prediction frame **pb**, and the real frame coordinates **gt**:

$$\begin{aligned}\mathbf{pb} &= (x^p_{min}, x^p_{max}, y^p_{min}, y^p_{max}),\\ \mathbf{gt} &= (x^g_{min}, x^g_{max}, y^g_{min}, y^g_{max})\end{aligned}$$

Output: IoU, L_{GIoU}

1:

$$\begin{aligned}A_p &= (x^p_{max} - x^p_{min}) \times (y^p_{max} - y^p_{min}),\\ A_g &= (x^g_{max} - x^g_{min}) \times (y^g_{max} - y^g_{min})\end{aligned}$$

2: I_{pg} is the intersection of the prediction frame and the true frame, U_{pg} is a union:

$$\begin{aligned}I_{pg} &= \begin{cases} (x^I_2 - x^I_1) \times (y^I_2 - y^I_1) & \text{if } x^I_2 > x^I_1, y^I_2 > y^I_1 \\ 0 \quad \text{otherwise} \end{cases},\\ U_{pg} &= A_p + A_g - I_{pg}\end{aligned}$$

Where,

$$\begin{aligned}x^I_1 &= max(x^p_{min}, x^g_{min}), x^I_2 = min(x^p_{min}, x^g_{min}),\\ y^I_1 &= max(y^p_{min}, y^g_{min}), y^I_2 = min(y^p_{min}, y^g_{min})\end{aligned}$$

3:

$$\begin{aligned}x^c_{min} &= min\left(x^p_{min}, x^g_{min}\right), x^c_{max} = max\left(x^p_{max}, x^g_{max}\right),\\ y^c_{min} &= min\left(y^p_{min}, y^g_{min}\right), y^c_{max} = max\left(y^p_{max}, y^g_{max}\right)\end{aligned}$$

4:

$$A_c = (x_{\max^c} - x_{\min^c}) \times (y_{\max^c} - y_{\min^c})$$

5:

$$\begin{aligned}IoU &= \tfrac{I_{pg}}{U_{pg}},\\ GIoU &= IoU - \tfrac{|A_c - U_{pg}|}{|A_c|}\end{aligned}$$

6:

$$L_{GIoU} = 1 - GIoU$$

4. Results and Discussion

4.1. Experimental Procedure

4.1.1. Datasets

In this study, three public datasets—LS-SSDD [22], SAR-Ship-Dataset [23], and HRSID [24]—were used as experimental test datasets.

(1) LS-SSDD adopts Sentinel-1 satellite data and contains a total of 30 large-scene SAR images. The large-scene images were taken from 30 original large-scene satellite-based SAR images. The polarization modes included two modes—VV and VH, and the IW—which has the distinctive features of large-scene ocean observation, small-scale ship detection, a variety of pure backgrounds, a fully automated detection process, and a variety of standardized benchmarks. Figure 6 shows a sample large-scene image from the LS-SSDD dataset. The large-scene image from the LS-SSDD dataset has a size of 24,000 × 16,000 pixels with ~250 km cover width, three-channel grayscale image format, 24-bit depth JPG, and XML annotation format, which record the target position information. During the experiment, the first 20 original large-scene SAR images were selected as the training set, and the remaining images were selected as the test set. Each 24,000 × 16,000 pixel SAR image was directly cropped into 800 × 800 pixels sub-images without processing.

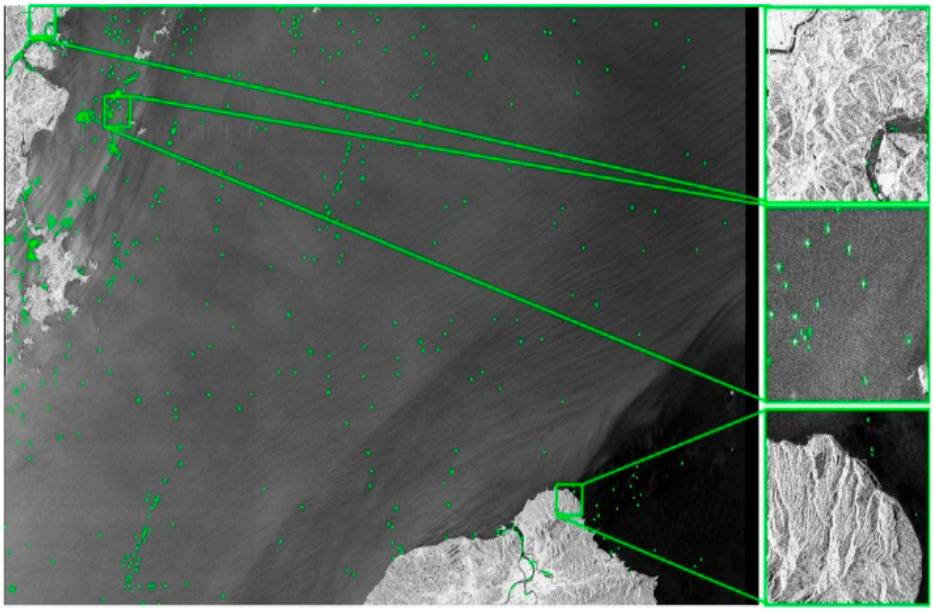

Figure 6. Sample large scene images in LS-SSDD dataset. Real ships are marked in green boxes.

(2) The SAR-Ship-Dataset, created by Wang et al. and labeled by SAR experts, is the most extensive publicly available dataset for multi-scale ship detection. It comprises 102 Chinese Gaofen-3 images and 108 Sentinel-1 images, totaling 43,819 ship chips. The chips have a resolution of 256 pixels and contain ships of various scales and backgrounds. The Gaofen-3 images were captured using Ultrafine Strip Chart (UFS), Fine Strip Chart 1 (FSI), Fully Polarized 1 (QPSI), Fully Polarized 2 (QPSII), and Fine Strip Chart 2 (FSII) imaging modes, with resolutions ranging from 3 to 10 m. Sentinel-1 images were acquired in S3 strip map (SM), S6 SM, and IW modes. The dataset also includes ships in complex scenes such as offshore, island, and harbor environments. Furthermore, the dataset covers scenarios with high ship densities and small target sizes (less than 15 $\times$ 15 pixels). Table 2 provides an overview of the dataset, with the training, validation, and test sets accounting for 70%, 20%, and 10% of the dataset, respectively.

Table 2. SAR-Ship-Dataset division.

Datasets	Total Number of Images	Occlusion	Small Target
Training set	21,420	12,840	8580
Verification set	6120	3660	2460
Testing set	3060	1836	1224

(3) The HRSID dataset is designed specifically for ship detection, semantic segmentation, and instance segmentation tasks in high-resolution SAR images. It consists of a total of 5604 images, including 99 Sentinel-1B, 36 TerraSAR-X, and 1 TanDEM-X images. Within these images, there are 16,951 ship instances, with small target scenes accounting for approximately 54.8% of all ships present. Similar to the construction process of the Microsoft COCO (Common Objects in Context) dataset, the HRSID dataset incorporates SAR images with varying resolutions, polarizations, sea states, sea areas, and coastal ports. This diversity enables researchers to benchmark and evaluate their methods effectively. The SAR images in the HRSID dataset have resolutions of 0.5 m, 1 m, and 3 m. To facilitate the development of algorithms, the dataset is split into three subsets: a 70% training set, a 20% validation set, and a 10% test set. This partitioning allows researchers to train their models, tune hyperparameters, and evaluate performance in a controlled manner. Figure 7 shows selected image samples from both the SAR-Ship-Dataset and HRSID datasets.

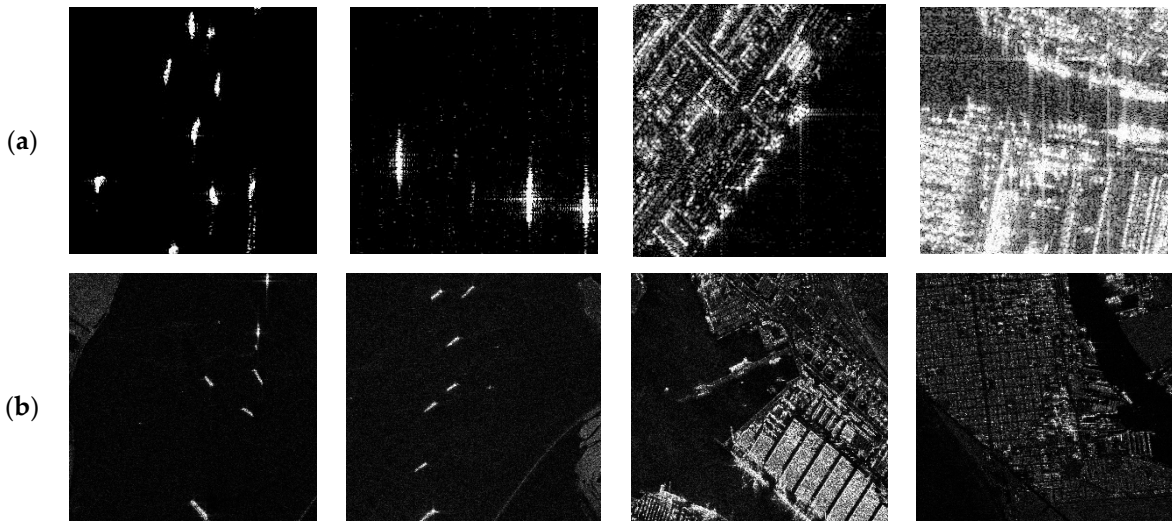

Figure 7. Image samples. (**a**) SAR-Ship-Dataset (**b**) HRSID.

4.1.2. Evaluation Metrics

The main evaluation metrics used in the detection model were precision, recall, and F1-score, which are defined as follows:

$$\begin{aligned} P &= N_{TD}/N, \\ R &= N_{TD}/N_{GT}, \\ F1-score &= \tfrac{2\times Precision\times Recall}{Precision+Recall}. \end{aligned} \tag{7}$$

where N_{TD} denotes the number of correctly detected ship targets, N_{GT} is the actual number of ship targets, and N is the total number of detected ship targets. Different IoU thresholds were used to calculate different numbers of ship targets $P(R)$. This value represents the precision–recall curve, and the purpose of the AP is to find the area under the precision–recall curve because it is the core index used to measure detection accuracy. The mean average precision (mAP) is the average value of all detection types. Because only one type of ship target is detected, mAP and AP have the same value.

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{9}$$

4.2. Experimental Analysis

All the experiments in this study were performed on an NVIDIA Tesla V100 graphics card. The number of training epochs was set to 200. Stochastic gradient descent (SGD) served as the optimizer with a 0.1 learning rate, a 0.9 momentum, and a 0.0001 weight decay. A soft non-maximum suppression (Soft-NMS) algorithm was used to suppress duplicate detections with an intersection over union (IoU) threshold of 0.5. The experiments focused on whether occluded targets as well as small targets in large-scene SAR images could be detected correctly. Experiments were first conducted on SAR-Ship and HRSID with numerous occluded scenes and small target objects to verify the effectiveness of the model. The model was then tested on the real large-scene dataset LS-SSDD, and the detection results of sub-images from the same scene during the detection process were directly spliced into a large-scene image without any other human involvement.

First, the experiments verified the effects of different attention mechanisms on the detection results. The TAC_CSAC_Net backbone network utilized Resnet50 and Resnet101

for the ablation experiments. The results using the dataset SAR-Ship-Dataset are shown in Tables 3 and 4. Taking the Resnet101 backbone network with better identification results, the performance metrics F1-score and mAP were improved by 0.002 and 1.1% in small target detection and 0.004 and 0.4% in occluded target detection, respectively, after the introduction of the TAC module compared to the original model. The introduction of the CSAC module improved the performance metrics F1-score and mAP by 0.005 and 2.1%, respectively, for small target detection and 0.005 and 1.6%, respectively, for occluded target detection, compared to the original model. After fusing the TAC_CSAC multi-attention mechanism, the F1-score of small target detection was improved by 0.012, mAP was improved by 3.5% compared with the original model, the F1-score of occluded target detection was improved by 0.026, and mAP was improved by 4.3%, which indicates that the TAC mechanism can capture the correlation between the features efficiently in the small target and occluded scenarios. Meanwhile, the complex background information blurs the position information of the ship target. The localization information of the target is not obvious after multilayer convolution, and it is very important to use CSAC to enhance the position and feature information. With respect to the Resnet101 backbone network with better recognition results, using the final improved model TAC_CSAC_Net versus the original model, the small target evaluation metrics precision, recall, F1-score, and mAP were improved by 2.6%, 0.7%, 0.017, and 3.7%, respectively, and the occlusion target evaluation metrics precision, recall, F1-score, and mAP were improved by 7.0%, 0.3%, 0.037, and 6.6%, respectively. Experimental results demonstrate that the proposed method is effective in detecting both small and occluded targets.

Table 3. Detection results of SAR-Ship-Dataset in small targets scenes.

Backbone Network (+Multi-Feature Fusion)	Attention Mechanism	P (%)	R (%)	F1-Score	mAP (%)
Resnet50		92.7	95.4	0.940	92.3
Resnet50	TAC	92.9	96.2	0.945	92.9
Resnet50	CSAC	92.9	96.4	0.946	92.8
Resnet50	TAC + CSAC	93.9	96.5	0.952	93.5
Resnet50	TAC_CSAC_Net	94.5	96.9	0.957	94.2
Resnet101		93.0	96.3	0.946	91.6
Resnet101	TAC	93.3	96.4	0.948	92.7
Resnet101	CSAC	93.7	96.5	0.951	93.7
Resnet101	TAC + CSAC	95.1	96.5	0.958	95.1
Resnet101	TAC_CSAC_Net	95.6	97.0	0.963	95.3

Table 4. Detection results of SAR-Ship-Dataset in occluded targets scenes.

Backbone Network (+Multi-Feature Fusion)	Attention Mechanism	P (%)	R (%)	F1-Score	mAP (%)
Resnet50		90.5	97.0	0.936	90.5
Resnet50	TAC	90.9	97.1	0.939	90.9
Resnet50	CSAC	92.4	97.4	0.948	91.4
Resnet50	TAC + CSAC	94.2	97.4	0.958	94.2
Resnet50	TAC_CSAC_Net	97.5	98.0	0.977	97.3
Resnet101		90.7	98.0	0.942	91.1
Resnet101	TAC	91.2	98.3	0.946	92.5
Resnet101	CSAC	91.5	98.1	0.947	92.7
Resnet101	TAC + CSAC	95.4	98.3	0.968	95.4
Resnet101	TAC_CSAC_Net	97.7	98.3	0.979	97.7

The results obtained using the HRSID are presented in Tables 5 and 6. Compared with the original initial model, the F1-score and mAP of TAC_CSAC_Net increased by 0.004 and 0.7% in the small target scenario and 0.066 and 0.5% in the occluded target scenario, respectively, indicating that both attentional mechanisms work accordingly and achieve close detection performance in both the occluded target and small target scenarios.

Table 5. Detection results of HRSID in small targets scenes.

Backbone Network (+Multi-Feature Fusion)	Attention Mechanism	P (%)	R (%)	F1-Score	mAP (%)
Resnet50		88.2	92.1	0.901	88.2
Resnet50	TAC	88.7	93.0	0.907	88.7
Resnet50	CSAC	89.3	92.6	0.909	89.3
Resnet50	TAC + CSAC	89.2	93.2	0.911	89.2
Resnet50	TAC_CSAC_Net	89.2	93.4	0.912	89.2
Resnet101		89.1	93.0	0.910	89.1
Resnet101	TAC	89.3	93.0	0.911	89.3
Resnet101	CSAC	89.7	93.0	0.913	89.7
Resnet101	TAC + CSAC	89.6	93.1	0.913	89.6
Resnet101	TAC_CSAC_Net	89.6	93.3	0.914	89.8

Table 6. Detection results of HRSID in occluded targets scenes.

Backbone Network (+Multi-Feature Fusion)	Attention Mechanism	P (%)	R (%)	F1-Score	mAP (%)
Resnet50		81.2	88.7	0.851	81.7
Resnet50	TAC	82.6	87.9	0.852	79.6
Resnet50	CSAC	84.2	89.4	0.867	84.2
Resnet50	TAC + CSAC	84.5	89.8	0.871	84.5
Resnet50	TAC_CSAC_Net	84.5	89.9	0.871	84.5
Resnet101		81.2	88.8	0.848	81.2
Resnet101	TAC	81.7	88.0	0.847	80.1
Resnet101	CSAC	84.8	89.4	0.870	84.8
Resnet101	TAC + CSAC	89.5	93.1	0.913	89.5
Resnet101	TAC_CSAC_Net	89.6	93.3	0.914	89.6

The model was used on the real large-scene dataset LS-SSDD, and the best model parameters from the training were used as the initial parameters to start the training by migration learning. In the LS-SSDD, without any additional embellishments, the large-scale images were fragmented into 9000 sub-images; that is, a large number of pure background sub-images were simultaneously involved in the training at the same time. From the final results, the direct-cut image was very close to the actual application. Table 7 presents the evaluation metrics of the detection results of the TAC_CSAC_Net model for a large-scene dataset. Compared with the SSDD and HRSID, the targets to be recognized in large scenes are smaller and more difficult to detect when they are occluded. Referring to the detection results graph in Figure 8, in the SAR image of the large scene, even at sea level where the target is small, it can have a high detection accuracy. However, in the occluded scene, owing to the double influence of the interference background and the small target in the large scene, although the accuracy is obviously improved, missed targets increased, which leads to the decline of the recall, as shown in Figure 9. The model evaluation metrics, F1 and mAP, both gradually improved, indicating that the multi-attention mechanism played an important role in feature capture. Taking the more effective Resnet101 backbone network as an example, F1 with the introduction of TAC increased from 0.724 to 0.747, an increase of 3.17%, and mAP increased from 70.9% to 72.2%, an increase of 1.83%; F1 with the introduction of CSAC increased from 0.724 to 0.766, an increase of 5.8%; mAP

increased from 70.9% to 74.8%, with an increase of 5.8%; with the final TAC_CSAC_Net model compared to the initial model, F1 increased from 0.724 to 0.822, with an increase of 13.5%; mAP increased from 70.9% to 78.6%, with an increase of 10.8%. Figure 10 shows the partial recognition results for a large scene. A test was performed to determine whether the target was included, and the final result was directly stitched as a large-scene SAR image.

Table 7. Detection results of LS-SSDD.

Backbone Network (+Multi-Feature Fusion)	Attention Mechanism	P (%)	R (%)	F1-Score	mAP (%)
Resnet50		73.1	65.8	0.693	63.0%
Resnet50	TAC	73.7	71.4	0.725	69.2%
Resnet50	CSAC	78.3	72.3	0.752	72.6%
Resnet50	TAC + CSAC	82.5	71.3	0.765	75.3%
Resnet50	TAC_CSAC_Net	83.8	73.6	0.784	76.4%
Resnet101		73.5	71.3	0.724	70.9%
Resnet101	TAC	75.1	74.4	0.747	72.2%
Resnet101	CSAC	78.4	74.9	0.766	74.8%
Resnet101	TAC + CSAC	83.5	72.7	0.777	75.3%
Resnet101	TAC_CSAC_Net	87.7	77.3	0.822	78.6%

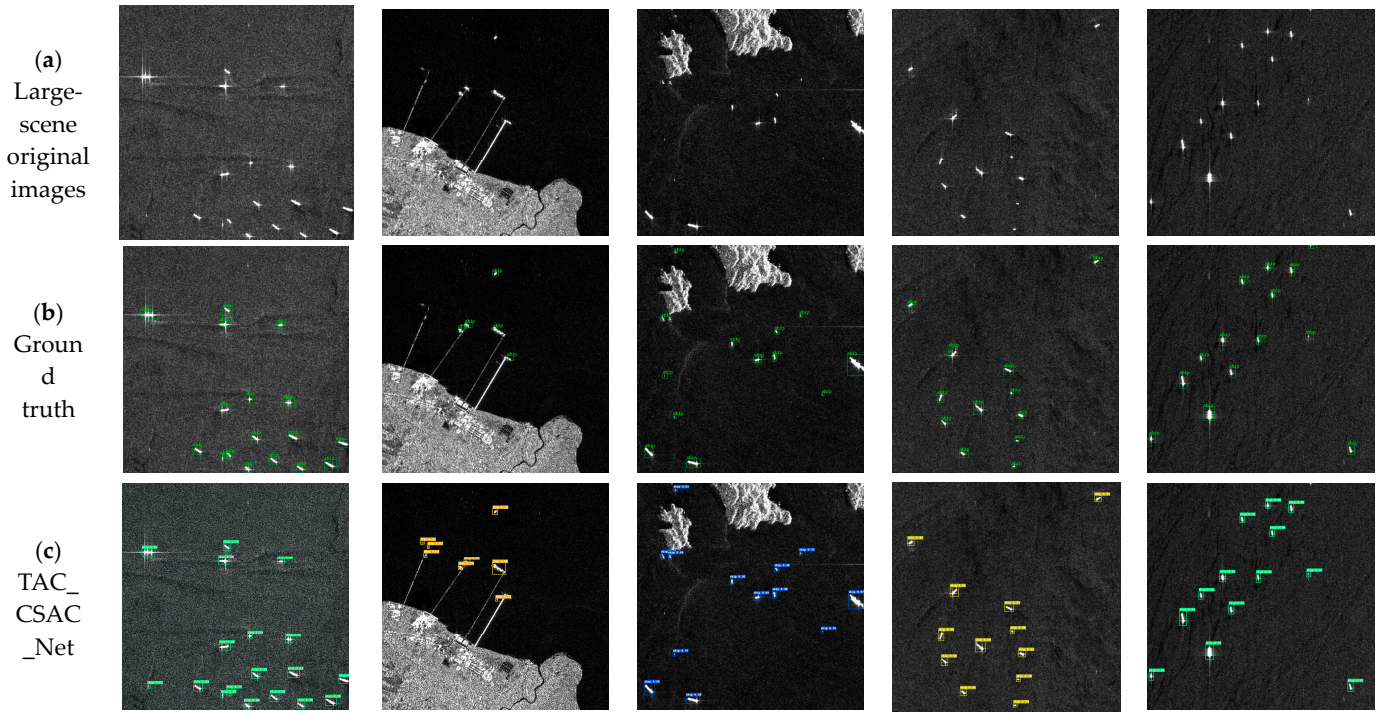

Figure 8. TAC_CSAC_Net's ship target detection results in a large scene; (**a**) shows large-scene original images, (**b**) shows the ground truth, (**c**) shows the outcomes of predictions of TAC_CSAC_Net model.

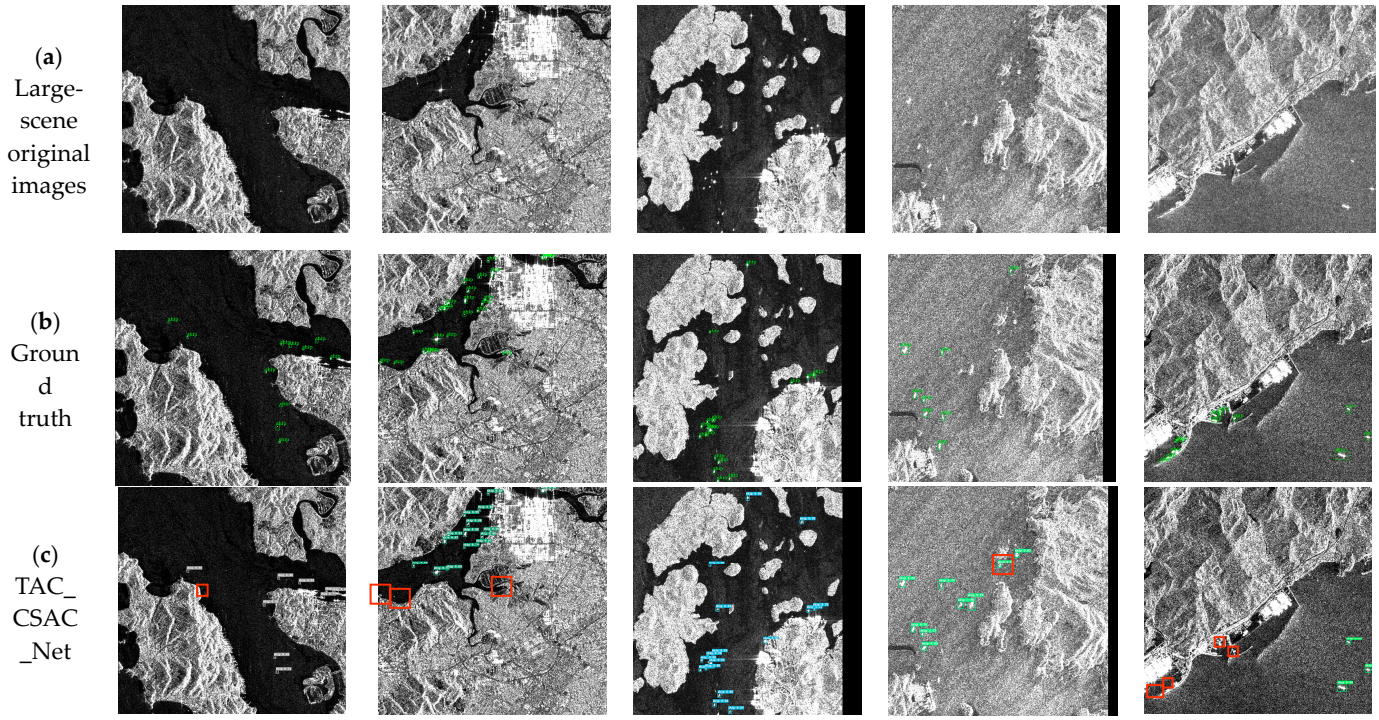

Figure 9. TAC_CSAC_Net ship target detection results in the occlusion scene of a large scene; (**a**) shows original images, (**b**) shows the ground truth, (**c**) shows the outcomes of prediction of TAC_CSAC_Net model. Missing detection is highlighted in red.

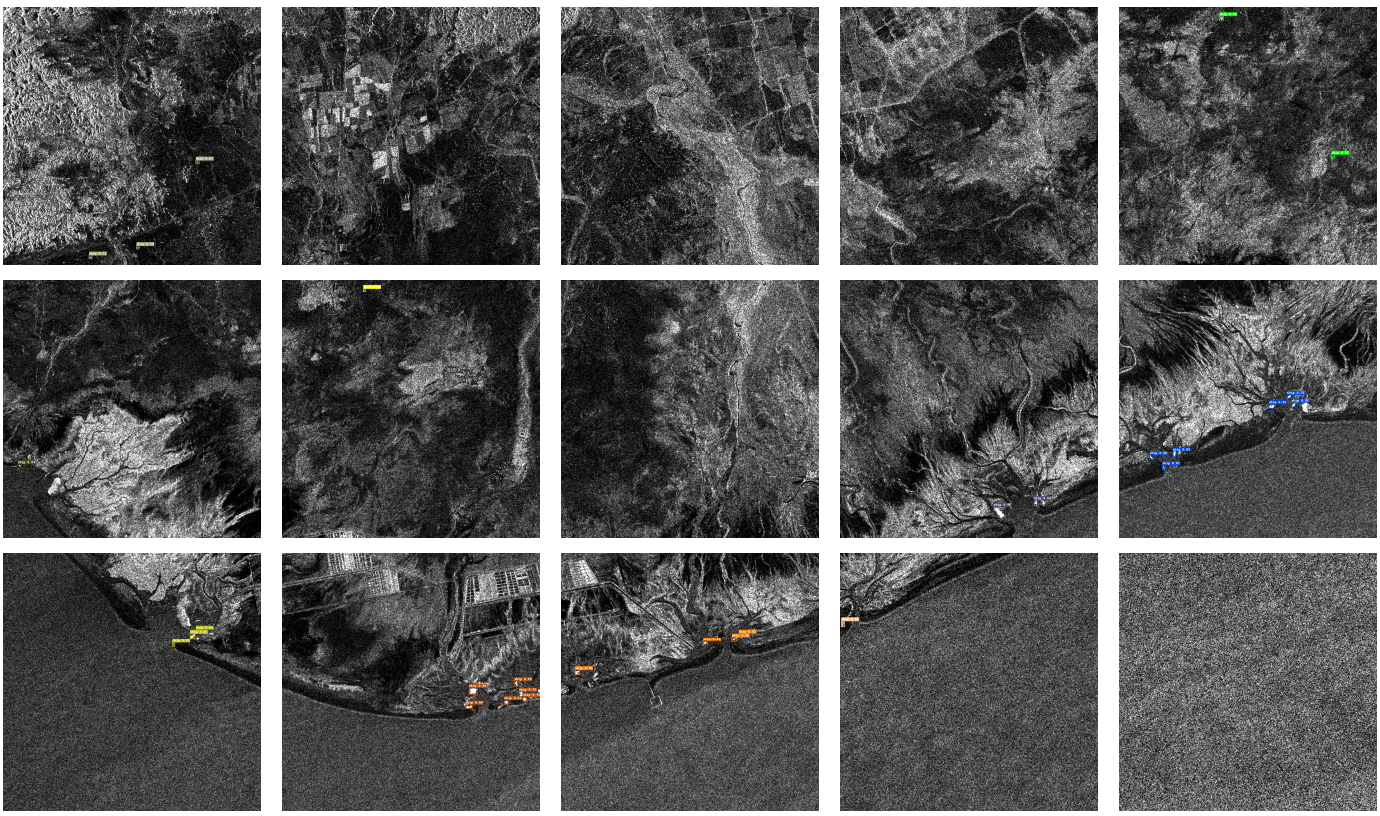

Figure 10. Large-scene partial recognition results. The final result can be stitched as a large-scene SAR image.

4.3. Comparative Experiments with Different Models

The models presented in this paper are compared with several classical and recently developed deep learning models used on the real large-scene dataset LS-SSDD, all executed on a Tesla V100. Iterations numbering 10^5 were performed on the SAR ocean dataset using the same training strategy. The batch size was set to 32, and the initial learning rate was set to 0.0001. The detection comparison results are listed in Table 8, which indicate that the proposed model has more powerful feature extraction and better target detection in large scenes. By evaluating the *p*-value, R-value, F1-score, and mAP-value of the respective algorithms on the LS-SSDD dataset, the models proposed in this study exhibited the highest detection accuracy. In the large scenario, the F1-score is higher than that of the suboptimal model, CBAM Faster R-CNN, which also contains an attention mechanism, by 0.019, which indirectly reflects the effectiveness of the proposed model with its global attention mechanism. Meanwhile, the detection precision was 4.7% higher than that of the suboptimal model, which indicates an improvement of the proposed model in terms of small target detection performance. Moreover, the frames per second (FPS) rate is used to measure the detection speed. Compared to other popular target detection algorithms for real-life large-scene SAR images, it is more capable of recognizing small and edge-featured fuzzy targets, which is a valuable contribution to practical applications in this field.

Table 8. Comparison of detection results from multiple models with the LS-SSDD dataset.

Models	P (%)	R (%)	F1-Score	mAP (%)	FPS
Faster-RCNN (Ren et al., 2015) [25]	72.8	72.1	0.724	74.4	4.82
SER Faster R-CNN (Lin et al., 2018) [26]	73.5	71.6	0.725	75.2	7.15
PANET (Liu et al., 2018) [27]	72.9	73.2	0.730	72.9	9.45
Cascade R-CNN (Cai and Vasconcelos, 2018) [28]	74.0	72.8	0.733	74.1	8.83
DAPN (Cui et al., 2019) [14]	73.8	75.1	0.744	74.1	12.22
ARPN (Zhao et al., 2020) [15]	73.5	71.6	0.725	75.2	12.15
Double-Head R-CNN (Wu et al., 2020) [16]	81.4	77.7	0.795	79.9	6.25
CBAM Faster R-CNN [17]	83.0	77.9	0.803	75.2	7.39
Quad-FPN (Zhang et al., 2021a) [18]	80.1	78.9	0.794	77.1	11.37
YOLOv5 (Jocher et al., 2021) [29]	72.8	77.1	0.748	74.4	21.76
YOLOv7 (Wang et al., 2022) [30]	78.2	76.1	0.771	76.3	22.43
Our model	87.7	77.3	0.822	78.6	8.06

5. Conclusions

The proposed method for detecting occluded targets and small target ships in large-scene SAR images focuses on the use of a multi-attention mechanism. By incorporating the transformer self-attention mechanism into the backbone network, a better target feature abstraction capability was obtained. Using channel attention and spatial attention to integrate the feature space and feature channel information in two dimensions can enhance the attention of the CNN in the channel dimension and strengthen the model's ability to perceive and utilize spatial location features. Experiments on publicly available multi-scale and multi-scene ship detection datasets, the SAR-Ship-Dataset and high-resolution SAR images, show that the improved model can significantly improve the detection performance of SAR images in different complex scenes and at different scales. Different attentional mechanisms can improve detection performance, and a model incorporating multiple attentional mechanisms has better detectability. The experimental results on the large-scene SAR images dataset show that the model can effectively improve ship detection accuracy in large-scene SAR images with a strong large-scene migration generalization capability. The experimental results also show that the proposed method has better detection performance and can reduce false alarms. However, it cannot completely eliminate missed detections in large-scene images. Further analysis and research on this topic are required, such as incorporating speckle noise removal methods before applying the model.

Author Contributions: H.Z. conceived and designed the algorithm and contributed to the manuscript and experiments; P.C. was responsible for the construction of the ship detection dataset, constructed the outline of the manuscript, and made the first draft of the manuscript; Y.L. supervised the experiments and was also responsible for the dataset; B.W. performed ship detection using deep learning methods. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Dalian Neusoft Institute of Information Joint Fund Project LH-JSRZ-202203, the Fundamental Scientific Research Project for Liaoning Education Department LJKMZ20222006, and the National Natural Science Foundation of CHINA 52271359.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Owing to the nature of this research, the participants in this study did not agree that their data can be publicly shared; therefore, supporting data are not available.

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

References

1. Xiao, X.; Zhou, Z.; Wang, B.; Li, L.; Miao, L. Ship Detection under Complex Backgrounds Based on Accurate Rotated Anchor Boxes from Paired Semantic Segmentation. *Remote Sens.* **2019**, *11*, 2506. [CrossRef]
2. 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.
3. An, Q.; Pan, Z.; You, H. Ship Detection in Gaofen-3 SAR Images Based on Sea Clutter Distribution Analysis and Deep Convolutional Neural Network. *Sensors* **2018**, *18*, 334. [CrossRef] [PubMed]
4. Yue, B.; Zhao, W.; Han, S. SAR Ship Detection Method Based on Convolutional Neural Network and Multi-Layer Feature Fusion. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*; Spinger: Berlin, Germany, 2020; Volume 1, pp. 41–53.
5. Shi, W.; Jiang, J.; Bao, S. Ship Detection Method in Remote Sensing Image Based on Feature Fusion. *Acta Photonica Sin.* **2020**, *49*, 57–67.
6. Y Li, Y.; Zhu, W.; Li, C.; Zeng, C. SAR Image Near-Shore Ship Target Detection Method in Complex Background. *Int. J. Remote Sens.* **2023**, *44*, 924–952. [CrossRef]
7. Ma, W.; Li, N.; Zhu, H.; Jiao, L.; Tang, X.; Guo, Y.; Hou, B. Feature Split–Merge–Enhancement Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
8. Ting, L.; Baijun, Z.; Yongsheng, Z.; Shun, Y. Ship Detection Algorithm Based on Improved YOLO V5. In Proceedings of the 2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE), IEEE, Dalian, China, 15–17 July 2021; pp. 483–487.
9. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Deep Learning Strong Parts for Pedestrian Detection. In Proceedings of the IEEE International Conference on Computer Vision, Araucano Park, Chile, 11–18 December 2015; pp. 1904–1912.
10. Ouyang, W.; Zhou, H.; Li, H.; Li, Q.; Yan, J.; Wang, X. Jointly Learning Deep Features, Deformable Parts, Occlusion and Classification for Pedestrian Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1874–1887. [CrossRef]
11. Jiao, J.; Zhang, Y.; Sun, H.; Yang, X.; Gao, X.; Hong, W.; Fu, K.; Sun, X. A Densely Connected End-to-End Neural Network for Multiscale and Multiscene SAR Ship Detection. *IEEE Access* **2018**, *6*, 20881–20892. [CrossRef]
12. Sun, Z.; Meng, C.; Cheng, J.; Zhang, Z.; Chang, S. A Multi-Scale Feature Pyramid Network for Detection and Instance Segmentation of Marine Ships in SAR Images. *Remote Sens.* **2022**, *14*, 6312. [CrossRef]
13. Yang, R.; Pan, Z.; Jia, X.; Zhang, L.; Deng, Y. A Novel CNN-Based Detector for Ship Detection Based on Rotatable Bounding Box in SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1938–1958. [CrossRef]
14. Cui, Z.; Li, Q.; Cao, Z.; Liu, N. Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8983–8997. [CrossRef]
15. 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]
16. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
17. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A Novel Quad Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
18. 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/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2020; pp. 10186–10195.
19. Amjoud, A.B.; Amrouch, M. Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review. *IEEE Access* **2023**, *11*, 35479–35516. [CrossRef]

20. Khan, A.; Rauf, Z.; Sohail, A.; Rehman, A.; Asif, H.; Asif, A.; Farooq, U. A Survey of the Vision Transformers and Its CNN-Transformer Based Variants. *arXiv* **2023**, arXiv:2305.09880.
21. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
22. Zhang, T.; Zhang, X.; Ke, X.; Zhan, X.; Shi, J.; Wei, S.; Pan, D.; Li, J.; Su, H.; Zhou, Y.; et al. LS-SSDD-v1.0: A Deep Learning Dataset Dedicated to Small Ship Detection from Large-Scale Sentinel-1 SAR Images. *Remote Sens.* **2020**, *12*, 2997. [CrossRef]
23. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]
24. 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]
25. Wan, S.; Goudos, S. Faster R-CNN for Multi-Class Fruit Detection Using a Robotic Vision System. *Comput. Netw.* **2020**, *168*, 107036. [CrossRef]
26. Lin, Z.; Ji, K.; Leng, X.; Kuang, G. Squeeze and Excitation Rank Faster R-CNN for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 751–755. [CrossRef]
27. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
28. Cai, Z.; Vasconcelos, N. Cascade R-Cnn: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
29. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Changyu, L.; Hogan, A.; Hajek, J.; Diaconu, L.; Kwon, Y.; Defretin, Y.; et al. Ultralytics/Yolov5: V5. 0-YOLOv5-P6 1280 Models, AWS, Supervise. Ly and YouTube Integrations. *Zenodo* **2021**. [CrossRef]
30. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, USA, 18–22 June 2023; pp. 7464–7475.