remote sensing

MDPI

Article

An Adaptive Sample Assignment Strategy Based on Feature Enhancement for Ship Detection in SAR Images

Hao Shi [1,2], Zhonghao Fang [1], Yupei Wang [1,2,*] and Liang Chen [1,2]

[1] Radar Research Lab, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; shihao@bit.edu.cn (H.S.); 3120210732@bit.edu.cn (Z.F.); chenl@bit.edu.cn (L.C.)
[2] Chongqing Innovation Center, Beijing Institute of Techonology, Chongqing 401120, China
* Correspondence: 6120210201@bit.edu.cn

Abstract: Recently, ship detection in synthetic aperture radar (SAR) images has received extensive attention. Most of the current ship detectors preset dense anchor boxes to achieve spatial alignment with ground-truth (GT) objects. Then, the detector defines the positive and negative samples based on the intersection-over-unit (IoU) between the anchors and GT objects. However, this label assignment strategy confuses the learning process of the model to a certain extent and results in suboptimal classification and regression results. In this paper, an adaptive sample assignment (ASA) strategy is proposed to select high-quality positive samples according to the spatial alignment and the knowledge learned from the regression and classification branches. Using our model, the selection of positive and negative samples is more explicit, which achieves better detection performance. A regression guided loss is proposed to further lead the detector to select well-classified and well-regressed anchors as high-quality positive samples by introducing the regression performance as a soft label in the calculation of the classification loss. In order to alleviate false alarms, a feature aggregation enhancement pyramid network (FAEPN) is proposed to enhance multi-scale feature representations and suppress the interference of background noise. Extensive experiments using the SAR ship detection dataset (SSDD) and high-resolution SAR images dataset (HRSID) demonstrate the superiority of our proposed approach.

Keywords: synthetic aperture radar (SAR); ship detection; label assignment; convolutional neural network (CNN)

Citation: Shi, H.; Fang, Z.; Wang, Y.; Chen, L. An Adaptive Sample Assignment Strategy Based on Feature Enhancement for Ship Detection in SAR Images. *Remote Sens.* **2022**, *14*, 2238. https://doi.org/10.3390/rs14092238

Academic Editor: Alex Hay-Man Ng

Received: 6 April 2022
Accepted: 4 May 2022
Published: 7 May 2022

1. Introduction

Synthetic aperature radar (SAR) is an active microwave sensor. Its all-day and all-weather working characteristics mean that it has been used in various fields [1–7]. Among these applications, ship detection has attracted more and more scholars' attention because of its great value in the military and civilian fields [8–13]. However, it is difficult to achieve accurate ship detection due to the large variation in scales and the strong interference of the complex backgrounds.

In the early years, researchers mainly used traditional detection methods to extract the salient features of the ship targets manually according to the prior knowledge. Although traditional methods have made much progress, there are still many problems such as complex algorithm design, low detection efficiency and lack of generalization.

With the rapid development of deep learning, object detectors based on convolution neural networks (CNNs) have achieved excellent detection performance, which extently overcomes the shortcomings of traditional methods. Modern CNN-based object detectors can be divided into two categories: one-stage and two-stage detectors. The detection process of the two-stage detectors mainly consists of two steps: the model first generates the candidate regions that potentially contain targets, and then uses classification and regression sub-networks to classify and regress the targets in the candidate regions, such as

Faster R-CNN [14], R-CNN [15]. Since the two-stage detectors contain the step of candidate region generation, they usually have high detection accuracy but the detection speed is slow and cannot achieve real-time detection. The one-stage detectors perform classification and regression by one-step prediction, such as SSD [16], RetinaNet [17] and YOLO series [18–20]. Compared with the two-stage detectors, the detection accuracy of the one-stage detectors is lower but the detection speed is higher, and allows the real-time detection.

In recent years, a variety of CNN-based methods have been widely introduced into ship detection in SAR images and have achieved a level of high detection performance and efficiency. Ke et al. [21] replaced the ordinary convolution with a deformable convolution in the Faster R-CNN, which enables the model to better learn the geometric features of the ship targets. In order to detect ship targets in complex backgrounds, Du et al. [22] improved the feature extraction ability of the network by integrating salient features into the SSD. Yu et al. [23] designed a more reasonable feature extraction and fusion method to enhance the feature representations of ship targets. To improve the detection speed of detectors, Chang et al. [24] replaced the ordinary convolution with depthwise separable convolution to increase the detection speed of the model. To detect the multi-scale ship targets more accurately, Wei et al. [25] designed a high-resolution feature pyramid network, which makes full use of the semantic feature contained in the high-resolution feature map to improve detection performance. Yu et al. [26] designed a novel two-way structure to extract multi-scale context features efficiently.

However, for accurate ship detection, these methods first preset anchor boxes at various scales and aspect ratios to achieve spatial alignment with multi-scale ship targets. Anchor boxes whose intersection-over-union (IoU) with the ground-truth objects is greater than the specified threshold (generally 0.5) are then defined as positive samples in the training process. The rest of the anchor boxes are defined as negative samples. This process is also known as label assignment.

Methods that use dense anchor boxes have several shortcomings in the ship detection task of SAR images. Firstly, the distribution of ship targets in SAR images is relatively sparse. Most of the preset dense anchor boxes will be defined as negative samples under the IoU-based assignment guideline. As a result, the problem of an imbalance of positive and negative samples will be more serious, which degrades the detection performance. Secondly, the IoU values between the anchors and ground-truth boxes need to be calculated. Extremely dense anchor boxes not only increase the number of parameters of the model but also greatly increase the computational cost. Thirdly, a fixed set of scale and aspect ratio parameters of the anchor boxes are usually only suitable for a specific data set. If the current data set is replaced, the scale and aspect ratio parameters need to be recalculated, which makes the detector sensitive to the data sets. Our proposed method achieves more accurate classification and regression performance only using one anchor box.

We found that the original label assignment strategy based on the IoU between the anchor boxes and the ground-truth boxes cannot guarantee whether an anchor box is an accurate positive sample. As shown in Figure 1, the IoU value between the anchor box (water green bounding box) and the ground-truth box (red bounding box) is 0.34. According to the original label assignment strategy, the IoU is less than the specified threshold, thereby the anchor box was not be assigned as a positive sample. However, the detection result corresponding to the anchor box (yellow bounding box with score of 1.00) performs with excellent classification confidence and location accuracy. The anchor box has the potential to become a positive sample and should be defined as a positive sample participating in the training process. Therefore, we suggest that the original label assignment strategy that only considers the IoU values between anchors and ground-truth boxes has drawbacks for defining positive and negative samples.

In the SAR images, the ship targets usually have different scattering properties from their surrounding background areas. However, certain areas of the inshore scenes have similar scattering power distributions to the ship targets. This phenomenon makes the

detector misidentify these areas as ship targets, resulting in false alarms, as shown in Figure 2.

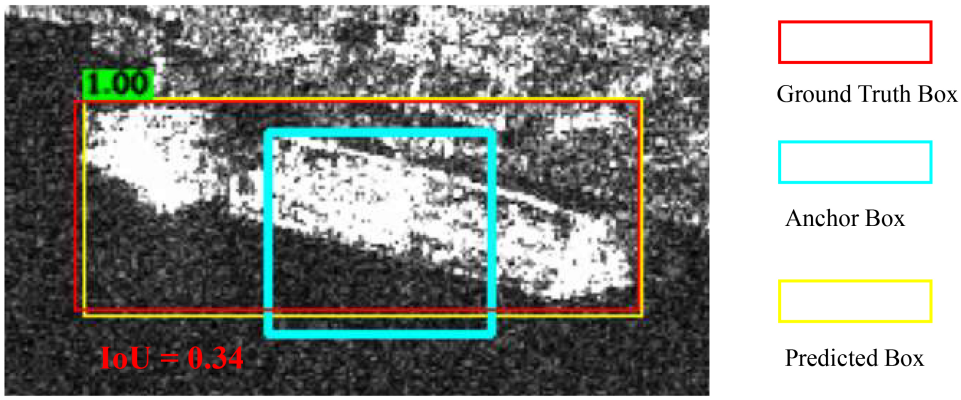

Figure 1. Visualization of the detection result of the original label assignment strategy. The red bounding box denotes the ground-truth object. The water green bounding box denotes the anchor box used to predict the ground-truth object. The yellow bounding box denotes the detection result corresponding to the anchor box.

According to the ground-truth objects (a) and the detection results (b) in Figure 2, the ground-truth objects (red bounding boxes) and the false alarm areas (orange ellipses) are found to have a similar scattering intensity. We believe that the reason for the false alarm phenomenon is that the ship targets and the false alarm areas have similar semantic representations at the feature scale, which makes it difficult to achieve accurate detection of ship targets.

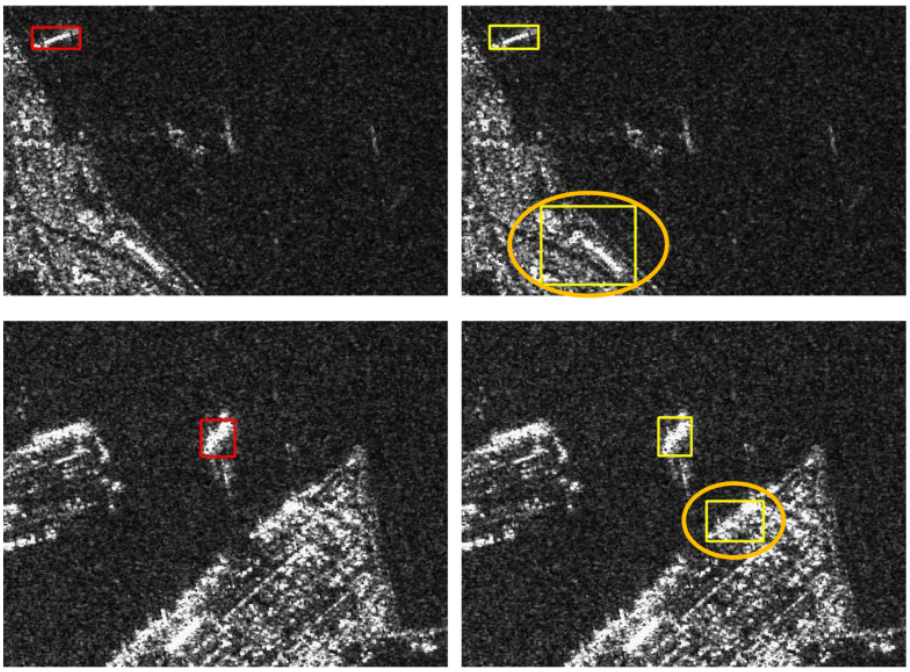

Figure 2. Visualization of the false alarm phenomena of the detector. (**a**) The ground-truth objects in SAR images. (**b**) Detection results of the model. The red bounding boxes denote the ground-truth objects. The yellow bounding boxes denote the detection results. The orange ellipses denote the false alarms.

Based on our analysis of the above problems, we propose a novel adaptive sample assignment (ASA) strategy to achieve high-quality training samples selection and improve the performance of the classification and regression. Firstly, we only generate one square anchor box on each cell of the feature maps, which greatly reduces the number of anchor boxes and computational overheads. Next, we preliminarily construct a candidate sample set for each ground-truth object to filter out anchor boxes which are at an inappropriate spatial scale. The candidate sample set could then adaptively filter the false candidates using knowledge learned from the regression and classification branches in the detector to ensure that only high-quality positive samples participate in the training process.

In addition, a regression guided loss function is proposed to further lead the detector to select well-classified and well-regressed samples as high-quality positive samples by introducing the regression performance of candidate samples as a soft label into the calculation of classification loss.

To alleviate the false alarm problems, we propose a feature aggregation enhancement pyramid network (FAEPN) to suppress the interference of noise and enhance the salient features of ship targets in complex backgrounds. FAEPN first uses the feature aggregation module to integrate multi-scale semantic information. Next, the integrated feature is fed into a fine-grained tuning module for further refinement. Considering that the feature pyramid structure is beneficial for the detector to detect multi-scale ships, we use the rebuild module to restore the hierarchical semantic features. Finally, the level fusion module is used to eliminate the feature aliasing problem caused by interpolation and pooling operations to achieve the enhancement of hierarchical features.

The experimental results for the SAR ship detection dataset (SSDD) and high-resolution SAR images dataset (HRSID) show that our proposed method achieves better detection performance compared to other detectors, proving the effectiveness of our method.

The contributions of this article can be summarized as follows:

1. A novel adaptive sample assignment (ASA) strategy is proposed to select high-quality positive samples based on the spatial alignment and the knowledge learned from the classification and regression branches of the network, which improves the classification and location performance.
2. A regression guided loss function is adopted to further guide the selection process of the model for high-quality positive samples by introducing a soft label in the classification loss function.
3. In order to alleviate the false alarm problems, a feature aggregation enhancement pyramid network (FAEPN) is proposed to enhance the salient features of the ship targets and suppress the noise interference from the complex backgrounds.

The rest of this paper is organized as follows. Section 2 presents the details of our proposed method. Section 3 shows the ablation experiments of our proposed method and the performance for different datasets. Section 4 discusses some problems arising from the detection results. Section 5 concludes the paper.

2. The Proposed Method

The overall structure of our proposed method is shown in Figure 3. The backbone of the network is ResNet-50 [27]. Firstly, we use FPN [28] to create hierarchical features. Then, we use a proposed feature aggregation enhancement pyramid network (FAEPN) to improve the salient features of ship targets and suppress the background noise by aggregating, refining, reconstructing and fusing multi-scale features. Next, the proposed adaptive sample assignment (ASA) strategy is used to select high-quality positive samples in the training process. Moreover, the regression guided loss is proposed to further lead the model to select well-classified and well-regressed samples as high-quality positive samples. The following sections introduce the components in detail.

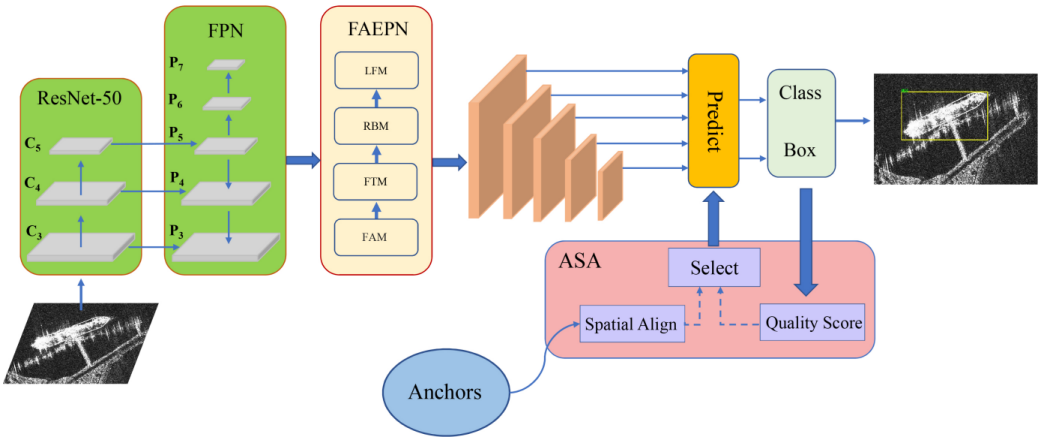

Figure 3. The overall structure of our model.

2.1. Adaptive Sample Assignment Strategy

Most of the current object detectors preset dense anchor boxes with various scales and aspect ratios in order to improve the detection performance. However, a large number of anchor boxes are redundant due to the sparse distribution of ships in the SAR images. As discussed in the introduction, redundant anchor boxes are defined as negative samples, which makes the imbalance of positive and negative samples more serious, and cause additional computational overheads. It has been proved in previous work that the label assignment strategy based on the IoU between the anchors and ground-truth boxes cannot provide the model with accurate positive samples, which constrains the model performance [29–33].

Inspired by the above works, we propose an adaptive sample assignment (ASA) strategy to assign the high-quality anchor boxes as positive samples. ASA strategy suggests that if an anchor box is defined as a positive sample, two conditions need to be met. The first is the spatial constraint condition: the larger the sample covering the ground-truth object area, the more likely it might be a positive sample; The second is consistent prediction condition: a sample with a higher classification score and better regression performance is more likely to be a positive sample. During the training process, the samples that satisfy these two conditions at the same time are defined as high-quality positive samples. The algorithm is shown in Algorithm 1.

Specifically, we first construct a candidate sample set A_{gt} for each ground-truth object according to the spatial constraint condition, which quickly filters out the samples that do not match the ground-truth object on the spatial scale. The spatial constraint condition we define is that the center point of the anchor box falls in the area of the ground-truth object and the IoU between the anchor box and the ground-truth box is greater than the specific threshold $\mathcal{T}$. We set $\mathcal{T} = 0.1$ in our experiments. Next, the candidate sample set filters the false candidate samples adaptively according to the knowledge learned from the classification and regression heads, which ensures only high-quality positive samples participate in the training process.

Formally, for a ground-truth object gt, given a sample j, which is in the candidate sample set of this gt. Its classification score is $score_j$ and the IoU between its prediction box and the gt box is IoU_j, we introduce the quality score (Q) to evaluate the quality of a candidate sample, that is:

$$Q_j = score_j^{\phi} \times IoU_j^{\lambda} \tag{1}$$

where ϕ and λ are weighting factors, which are used to control the proportion of classification and regression results in the quality score. We set $\phi = 2$ and $\lambda = 1$ in our experiments.

Algorithm 1: Adaptive Sample Assignment algorithm

Input: $\mathcal{GT}, \mathcal{A}, \mathcal{T}, \mathcal{K}$
- $\mathcal{GT}$ is a set of ground-truth objects
- $\mathcal{A}$ is a set of all anchors across all pyramid levels
- $\mathcal{T}$ is the IoU threshold
- $\mathcal{K}$ controls the number of the positive samples for each ground-truth object

Output: $\mathcal{P}, \mathcal{N}$
- $\mathcal{P}$ is a set of positive samples
- $\mathcal{N}$ is a set of negative samples

$\mathcal{P}, \mathcal{N} \leftarrow \varnothing$

for every ground-truth $gt \in \mathcal{GT}$ **do**
- $\mathcal{A}_{gt} \leftarrow \{i \in \mathcal{A}$: the center of anchor box i is in gt box and IoU$(i, gt) >= \mathcal{T}\}$
- $\mathcal{Q}_{gt} \leftarrow \{ j \in \mathcal{A}_{gt}$: compute $\mathbf{Q}_j$ by Equation (1) $\}$
- $indices = argsort(\mathcal{Q}_{gt})$ // Sort in descending order
- $\mathcal{P} \leftarrow \mathcal{P} \cup indices[0 : \mathcal{K}]$

end for

$\mathcal{N} \leftarrow (\mathcal{A} - \mathcal{P})$

return $\mathcal{P}, \mathcal{N}$

After obtaining the quality scores of all candidate samples of the gt, we sorted them in descending order and take the first $\mathcal{K}$ candidate samples as high-quality positive samples of the gt in the training process.

2.2. Regression Guided Loss

In order to further improve the model's ability to select high-quality positive samples, we designed a regression guided loss to guide the model's selection process. We first introduce a metric of the regression quality, called regression precision (r_p). Considering that the IoU values between the candidate samples and the ground-truth boxes and the IoU values between the ground-truth boxes and the predicted boxes of the candidate samples, we measure the regression precision using an effective combination of two types of IoU values. Specifically, considering the different epochs of the model in the training process, we designed the following metric to compute the regression precision for each instance:

$$r_p = \delta \times exp(t-1) + \eta \times exp(t) + (1-\delta) \tag{2}$$

where η denotes the mean IoU value between the candidate samples and the corresponding ground-truth. Using mean value can better reflect the degree of spatial matching between the candidate samples and the ground-truth object. δ denotes the max IoU value between the predicted boxes of the candidate samples and the corresponding ground-truth. Using max value can better describe the optimal regression effect of the prediction boxes. $t = epoch/total_{epoch}$ indicates the different stages of the training process.

To guide the model to select high-quality positive samples with excellent classification and regression performance, we use r_p as a soft label to replace the binary label in the original Focal Loss [17]. We only assign r_p to a set of positive samples, with others labeled as 0. We explicitly increased classification scores for the well-regressed samples and reduced classification scores for the poorly-regressed samples. The classifier will output more accurate classification results through introducing the regression precision of candidate samples, which further guides the selection process of the high-quality positive samples. The regression guided loss can be written as:

$$L_{cls} = -\frac{1}{N_{pos}}\left(\sum_{i=1}^{N_{pos}} \alpha(1-s_i)^{\gamma} BCE(s_i, r_p) + \sum_{j=1}^{N_{neg}} (1-\alpha) s_j^{\gamma} BCE(s_j, 0)\right) \tag{3}$$

where N_{pos} and N_{neg} indicate the number of all positive and negative samples, respectively. BCE denotes the binary cross entropy loss, α and γ are adjustment factors. We set $\alpha = 0.25$, $\gamma = 2$ in our experiments.

2.3. Feature Aggregation Enhancement Pyramid Network

There are many areas with very complex backgrounds in SAR images. Therefore, enhancing the features of ship targets and suppressing background noise is particularly important. In order to make low-level and high-level features complement each other, FPN [28] constructed a top-down information propagation path to improve the detection performance. However, Refs. [34–36] mentioned that the semantic gaps exist between features at different scales, meanwhile the high-level semantic information would be gradually diluted as the propagation progress. To achieve more effective fusion of different scale feature information, we propose a feature aggregation module (FAM) that aggregates different semantic features. Motivated by Cao et al. [37], who integrates global feature representations via attention mechanism, we adopted global context block as the fine-grained tuning module (FTM) to realize the refinement of the aggregated feature. For the better detection performance of multi-scale ship targets, we used the rebuild module (RBM) to reconstruct the hierarchical feature structure. We note that the interpolation operations are used to construct the top-down information propagation path in the FPN and generate the aggregated feature, meanwhile the pooling operations are used to reconstruct the hierarchical features. Lin et al. [28] proposed that the multi-scale feature fusion using interpolation may cause feature aliasing, which interferes with the classification and regression process. Correspondingly, we consider that pooling operations may also introduce a similar problem. Therefore, we propose a level fusion module (LFM) consisting of the feature aggregation channel attention module (FACAM) and the feature refinement channel attention module (FRCAM) to eliminate the effect of feature aliasing before multi-scale feature fusion by the attention mechanism. The overall structure of the FAEPN is shown in Figure 4.

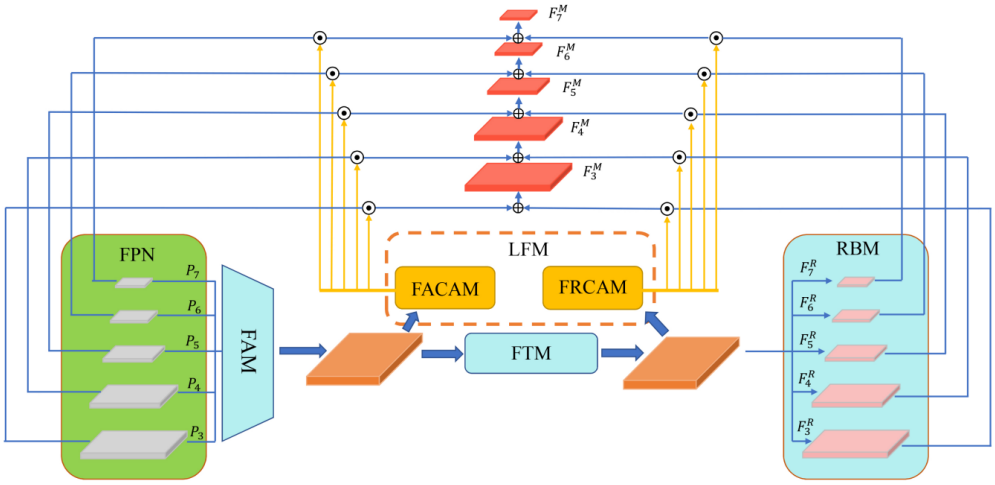

Figure 4. The overall structure of the FAEPN.

(1) Feature Aggregation Module: We feed the last three stages features extracted by the backbone network into the FPN to generate $\{\mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5\}$ and denote the input features as $\{\mathbf{C}_3, \mathbf{C}_4, \mathbf{C}_5\}$ according to the size of the feature maps. The entire process is the same as in [28]:

$$\mathbf{M}_i = \begin{cases} \mathbf{Conv_{1\times1}}(\mathbf{C}_i) + \mathbf{Up}(\mathbf{M}_{i+1}) & i = 3,4 \\ \mathbf{Conv_{1\times1}}(\mathbf{C}_5) & i = 5 \end{cases} \tag{4}$$

$$\mathbf{P}_i = \mathbf{Conv_{3\times3}}(\mathbf{M}_i) \qquad i = 3,4,5 \tag{5}$$

where $\mathbf{Conv_{1\times1}}$ and $\mathbf{Conv_{3\times3}}$ denote the 1×1 and 3×3 convolutional layer, respectively. **Up** denotes the bilinear interpolation with the upsampling factor of 2. $\mathbf{M}_i$ denotes the integrated feature map produced by 1×1 convolution and interpolation. In order to obtain richer semantic feature information, two independent 3×3 convolution layer are appended on $\mathbf{P}_5$ and $\mathbf{P}_6$ to generate $\mathbf{P}_6$ and $\mathbf{P}_7$. This process can be formulated as follows:

$$\begin{aligned} \mathbf{P}_6 &= \mathbf{Conv_{3\times3\_S2}}(\mathbf{P}_5) \\ \mathbf{P}_7 &= \mathbf{Conv_{3\times3\_S2}}(\mathbf{P}_6) \end{aligned} \tag{6}$$

where $\mathbf{Conv_{3\times3\_S2}}$ represents the 3×3 convolution layer with the stride of 2.

In deep convolution neural networks, high-level feature maps contain richer contextual information and more abstract feature representations, but lose many details for localization. Compared with high-level feature maps, shallow feature maps contain relatively weak semantic information, but retain more spatial location information. For the optimal balance between semantic and spatial location information, we chose $\mathbf{P}_3$ as the intermediate stage and resize the feature map of other levels to the same size as $\mathbf{P}_3$ with the interpolation operation. The resized features were then fed together into the FAM to generate the aggregated feature $\mathbf{F_A}$. The whole process can be described as:

$$\mathbf{F_A} = (\sum_{i=4}^{7} \mathbf{Upsample}(\mathbf{P}_i) + \mathbf{P}_3)/5 \tag{7}$$

(2) Fine-grained Tuning Module: In order to enhance the salient features of ship targets and suppress noise interference, we adopt the global context block [37] as the fine-grained tuning module to achieve refinement of the aggregated feature $\mathbf{F_A}$. The structure of the fine-grained tuning module is shown in Figure 5.

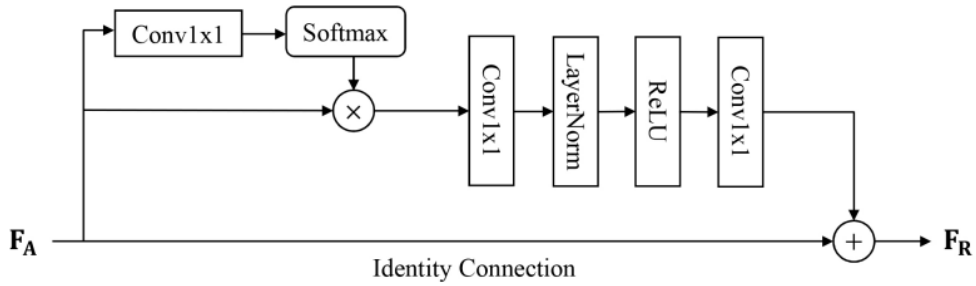

Figure 5. Structure of the fine-grained tuning module.

Given aggregated feature $\mathbf{F_A}$, we constructed the refined feature $\mathbf{F_R}$ as follows:

$$\begin{aligned} \mathbf{F_C} &= \mathbf{SPR}(\mathbf{F_A}) \otimes \mathbf{F_A} \\ \mathbf{F_R} &= \mathbf{T}(\mathbf{F_C}) + \mathbf{F_A} \end{aligned} \tag{8}$$

where $\otimes$ denotes matrix multiplication. Firstly, the spatial position relationship map **SPR** of the $\mathbf{F_A}$ is obtained as follows:

$$\mathbf{SPR}(\mathbf{F_A}) = \psi(\mathbf{Conv_{1\times1}}(\mathbf{F_A})) \tag{9}$$

where $\mathbf{Conv_{1\times1}}$ denotes the 1×1 convolution layer and ψ denotes the softmax activation function. Next, the feature map $\mathbf{F_C}$ containing larger receptive field and global context feature is obtained by multiplying the aggregated feature $\mathbf{F_A}$ with the spatial position relationship map **SPR**. Then, the feature $\mathbf{F_C}$ is fed into the bottleneck transform network to capture the channel-wise dependencies. The transformed feature is denoted as $\mathbf{F_T}$ and the whole pipeline can be formulated as:

$$\mathbf{F_T} = \mathbf{W_1}(\beta(\sigma(\mathbf{W_0}(\mathbf{F_C})))) \tag{10}$$

where $\mathbf{W}_0 \in \mathbb{R}^{C/r \times C}$ and $\mathbf{W}_1 \in \mathbb{R}^{C \times C/r}$ are the weights of 1×1 convolution layer, which are used to, respectively, reduce and restore the number of the feature channels and r denotes the channel reduction ratio. σ represents the layer normalization and β represents the ReLU activation function. Finally, the refined feature $\mathbf{F_R}$ is obtained by element-wise addition of the transformed feature $\mathbf{F_T}$ and the aggregated feature $\mathbf{F_A}$.

(3) Rebuild Module: To achieve better detection performance for multi-scale ships, we reconstruct the feature hierarchy through the refined feature using a max pooling operation. The reconstructed hierarchical features have the same size as the hierarchical features of the FPN. Hierarchical features are constructed as follows:

$$\mathbf{F}_i^{\mathbf{R}} = \mathbf{Pool}(\mathbf{F_R}) \qquad i = 3, 4, 5, 6, 7 \tag{11}$$

(4) Level Fusion Module: We fused the reconstructed hierarchical features with the hierarchical features of the FPN to further enhance the feature of each level. To alleviate the feature aliasing problem caused by interpolation and pooling operations, we introduce the feature aggregation channel attention module (FACAM) and feature refinement channel attention module (FRCAM) to, respectively, get the channel-wise weight $\mathbf{W_A}$ of the $\mathbf{F_A}$ and the channel-wise weight $\mathbf{W_R}$ of the $\mathbf{F_R}$. The FACAM and the FRCAM have the same structure, as shown in Figure 6.

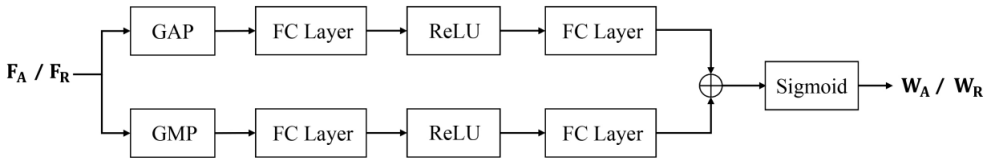

Figure 6. Structure of the FACAM and the FRCAM.

We first extracted two different spatial semantic information of input feature by global average pooling (GAP) and global max pooling (GMP). Next, the two spatial semantic vectors are, respectively, fed into fully connected layers. Finally, the output feature results are added by element-wise and sigmoid activation function is used to obtain the channel-wise weight.

After getting the channel-wise weight results, we performed the channel weighting to complete the final hierarchical feature fusion, the process can be described as follows:

$$\mathbf{F}_i^{\mathbf{M}} = \mathbf{F}_i^{\mathbf{R}} \odot \mathbf{W_R} + \mathbf{P}_i \odot \mathbf{W_A} \qquad i = 3, 4, 5, 6, 7 \tag{12}$$

where $\odot$ denotes the element-wise multiplication.

Through our proposed feature aggregation enhancement pyramid network, the original FPN hierarchical features were aggregated, refined, reconstructed and fused to further enhance semantic features at different scales and effectively suppress the occurrence of false alarms.

3. Experiments

3.1. Datasets

3.1.1. SSDD

SSDD [38] is an open SAR ship detection dataset, which contains multi-scale SAR ships in different sensors, polarization modes, image resolutions and scenes. In the SSDD dataset, there are 1160 images and 2540 ships in total and the average image size is 500×500. The total dataset is divided into training set and test set, including 928 and 232 images, respectively. The images are resized to 512×512 in our experiments.

3.1.2. HRSID

HRSID [39] is a high-resolution SAR images dataset, which contains 5604 images collected from Sentinel-1 and TerraSAR-X. The SAR ships in HRSID dataset are provided

with HH, HV and VV polarization modes and various resolutions from 0.1 m to 3 m. We follow the original reports in HRSID dataset to divide the entire dataset into training set and test set as 13:7. The image sizes are 800×800 in our experiments.

3.2. Experimental Details

We use the pretrained ResNet-50 on ImageNet to initialize the backbone network. We use the feature pyramid of $\{P_3^M, P_4^M, P_5^M, P_6^M, P_7^M\}$ to detect multi-scale ships. For each cell of the feature maps, we only set one anchor to classify and regress targets. We conduct ablation studies on the SSDD dataset.

Our model is trained with Adam optimizer and the batch size is set to 32 on Titan RTX GPU. The initial learning rate is set to 10^{-4} and is divided by 10 at every decay step. The total iterations of SSDD and HRSID are 2 k, 17 k, respectively.

3.3. Evaluation Metrics

We use PASCAL VOC object detection challenge [40] evaluation metrics including precision (p), recall (r) and mean average precision (mAP) to evaluate the model detection performance. Precision (p) and recall (r) can be defined as:

$$p = \frac{TP}{TP + FP} \tag{13}$$

$$r = \frac{TP}{TP + FN} \tag{14}$$

where TP, FP and FN denote the number of true positives, false positives and false negatives, respectively.

The mean average precision (mAP) measures the comprehensive performance of a detector by considering both precision and recall. The $P(r)$ denotes the precision-recall curve and mAP can be calculated as:

$$mAP = \int_0^1 P(r) \times dr \tag{15}$$

3.4. Ablation Study

3.4.1. Evaluation of Different Components

We conducted a component-wise experiment on the SSDD dataset to verify the contribution of the proposed components. The experimental results are shown in Table 1. Since we only set one anchor box on each cell of the feature maps, which makes the detection of ground-truth objects restricted, our baseline model only achieves a precision of 75.5% and mAP of 86.88%. Using the FAEPN module, the precision increased by 11.6% and the mAP increased by 6.43%, indicating that the FAEPN module significantly improves the feature representations of the ship targets by aggregating, refining, rebuilding and fusing multi-scale semantic features and suppresses the background noise. The mAP increased by 6.23% and the precision increased by 10.54% using the ASA strategy, which indicates that our label assignment strategy accurately provides the model with high-quality positive samples, so that even one anchor box can achieve excellent performance. In addition, using the ASA strategy and regression guided loss at the same time, the mAP and precision are further improved, indicating that the regression guided loss we designed can further guide the model to select high-quality positive samples. Finally, using all the proposed components simultaneously, our final model reaches a precision of 88.54% with an increase of 13.04% and a mAP of 95.19% with an increase of 8.31%, demonstrating the effectiveness of our proposed method. Subsequent sections will describe each of the proposed components in detail.

Table 1. Effects of the proposed components on the SSDD dataset.

FAEPN	ASA	Regression Guided Loss	mAP	Precision	Recall
×	×	×	86.88	75.5	89.74
✓	×	×	93.31	87.1	94.3
×	✓	×	93.11	86.04	94.29
×	✓	✓	93.39	86.73	94.31
✓	✓	✓	**95.19**	**88.54**	**95.94**

The best results in the table are shown in bold.

3.4.2. Evaluation of Feature Aggregation Enhancement Pyramid Network

To verify the effectiveness of the proposed feature aggregation enhancement pyramid network, we implemented some ablation experiments on the SSDD dataset. The experimental results are shown in Table 2. The optimal performance is in the case of using a level fusion module (LFM) and the channel reduction ratio of 1:8, achieving a mAP of 93.31%, precision of 87.1% and recall of 94.3%. According to the experimental results, irrespective of which channel reduction ratio is used, the mAP, precision and recall are steadily improved by using the level fusion module. Especially for the channel reduction ratio of 1:16, using the level fusion module achieves a 3.53% mAP and 3.48% recall improvement compared to not using the level fusion module. The above results demonstrate that the proposed level fusion module can effectively eliminate the feature aliasing problem caused by interpolation and pooling operations.

Table 2. Analysis of the feature aggregation enhancement pyramid network on the SSDD dataset.

Ratio	mAP			Precision			Recall		
	w/LFM	w/o LFM	Δ	w/LFM	w/o LFM	Δ	w/LFM	w/o LFM	Δ
1:4	92.73	92.19	0.54	84.56	84.07	0.49	94.29	93.92	0.37
1:8	**93.31**	92.46	0.85	**87.1**	85.54	1.56	**94.3**	93.91	0.39
1:16	90.31	86.78	3.53	82.53	81.09	1.44	92.28	88.8	3.48

The best results in the table are shown in bold.

Some visualization detection results are shown in Figure 7. The red bounding boxes denote the ground-truth objects and the yellow bounding boxes denote the detection results. It shows that the baseline model will identify areas with strong scattering power in the land background as the ship targets (see the second column in Figure 7). In contrast, the baseline model with FAEPN module can alleviate the interference of noise during the detection process and achieve more accurate detection (see the third column in Figure 7). Notably, even if the region with similar geometric shapes and strong scattering power in the offshore scenes, our proposed FAEPN module also can accurately distinguish the background noise from the ships, which further proves the effectiveness of the FAEPN. (see the second and third columns of the third row in Figure 7).

3.4.3. Evaluation of Adaptive Sample Assignment

To explore the contribution of the ASA strategy, we implemented the element-based ablation experiments, with the experimental results shown in Table 3. In order to ensure that each ground-truth object has a certain number of high-quality positive samples during the training process, we roughly set three sets of $\mathcal{K}$ values: 25, 30 and 35 to verify the robustness of the model to the parameter $\mathcal{K}$. Next, we investigate the impact of different values of ϕ and λ on model detection performance, where ϕ and λ are used to control the proportion of classification score and regression accuracy in the quality score.

The experimental results show that the mAP varies from 92.2% to 93.2% when different $\mathcal{K}$ values are set, indicating that the performance of the model is not sensitive to the parameter $\mathcal{K}$. To achieve an optimal balance between the precision and recall, we conduct

a rough search for the parameters ϕ and λ for every different $\mathcal{K}$ to explore the relationships among $\mathcal{K}$, ϕ and λ. Finally we set $\mathcal{K} = 30$, $\phi = 2$ and $\lambda = 1$ in our experiments.

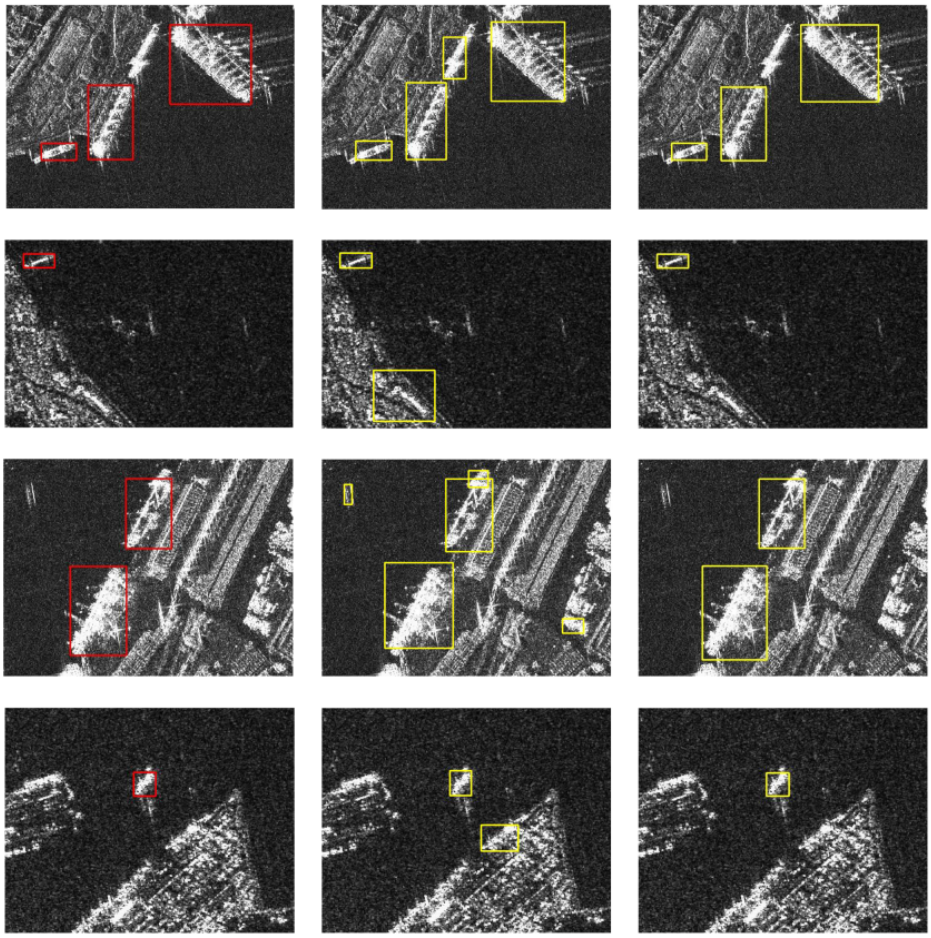

Figure 7. Visualization detection results of the model with and without FAEPN. (**a**) Ground-truth objects. (**b**) Detection results of the baseline. (**c**) Detection results of the baseline with FAEPN. The red bounding boxes denote the ground-truth objects. The yellow bounding boxes denote the detection results.

Table 3. Analysis of adaptive sample assignment on the SSDD dataset.

ID	$\mathcal{K}$	ϕ	λ	mAP	Precision	Recall
1	25	1	1	93.04	82.64	94.66
2		2	1	92.81	83.7	94.27
3		2	2	92.99	82.66	94.84
4		4	2	92.84	83.84	94.29
5	30	1	1	92.69	84.83	94.11
6		2	1	**93.11**	**86.04**	**94.29**
7		2	2	92.67	84.08	93.93
8		4	2	92.97	85.89	94.29
9	35	1	1	92.42	81.83	93.9
10		2	1	92.21	82.52	93.92
11		2	2	92.47	82.1	93.94
12		4	2	92.88	86.74	93.91

The best results in the table are shown in bold.

3.4.4. Evaluation of Regression Guided Loss

To verify the effectiveness of our proposed regression guided loss, we carried out the comparative experiments based on the parameter settings of $\mathcal{K} = 30$, $\phi = 2$ and $\lambda = 1$. The experimental results are shown in Table 4. The experimental results show that the model has further improved the mAP, precision and recall by replacing the focal loss function with the proposed regression guided loss function. This suggests that it is critical to use the regression performance as the soft label to train the classification branch, which makes the classifier pay more attention to learning the samples with excellent classification and regression performance.

Table 4. Analysis of regression guided loss on the SSDD dataset.

Regression Guided Loss	mAP	Precision	Recall
×	93.11	86.04	94.29
✓	**93.39**	**86.73**	**94.31**

The best results in the table are shown in bold.

3.5. Main Results and Analysis

3.5.1. Results for the SSDD Dataset

The confusion matrix of our proposed method in the entire, inshore and offshore scenes of the SSDD dataset is shown in Table 5, where TP, FN, FP and TN means the true positives, false negatives, false positives and true negatives, respectively. Since the true negatives are not used for the evaluation of detection metrics, we set the TN term as * in the confusion matrix. From the results of the confusion matrix, we can find that our proposed method has the excellent performance in both inshore and offshore scenes. Additionally, our proposed method can detect the ship targets accurately and has less false alarm problems. The detection results of different methods on the SSDD dataset are shown in Table 6. Anchor Number denotes the number of anchors at each cell of the feature maps. Our ASAFE, Faster R-CNN, Double-Head R-CNN, PANet and RetinaNet reaches a mAP of 95.19%, 90.01%, 91.17%, 91.73% and 86.37% in the entire scenes, respectively. The detection results of the entire scenes show that our proposed method achieves the optimal detection performance compared to other algorithms. It can be found from the detection results of the inshore scenes that our ASAFE surpasses the second-best Faster R-CNN by 7.01% on mAP. It indicates that benefiting from the proposed FAEPN module, our model effectively suppresses the strong interference of background noise in the inshore scenes and achieves the best detection performance. Additionally, it is worth noting that our method only sets one square anchor box at each cell of the feature maps but achieves the excellent detection results compared with the Faster R-CNN, Double-Head R-CNN, PANet and RetinaNet with nine anchor boxes. Comparison results show that the number of the anchor boxes does not improve the detection performance of the model. Instead, selecting the high-quality positive samples in the training process is more important and also proves the effectiveness of the proposed ASA strategy and the regression guided loss function.

Table 5. The confusion matrix of the SSDD dataset in the entire, inshore and offshore scenes.

Entire Scenes		Inshore Scenes		Offshore Scenes	
TP = 521	FN = 22	TP = 139	FN = 33	TP = 367	FN = 4
FP = 67	TN = *	FP = 45	TN = *	FP = 8	TN = *

Since the true negatives are not used for the evaluation of detection metrics, we set the TN term as *.

Table 6. Comparisons with different methods on the SSDD dataset.

Methods	Anchor Number	Entire Scenes			Inshore Scenes			Offshore Scenes		
		mAP	P	R	mAP	P	R	mAP	P	R
Faster R-CNN [14]	9	90.01	88.02	90.71	73.81	71.28	74.91	97.04	96.49	97.72
Double-Head R-CNN [41]	9	91.17	86.82	92.01	73.68	69.68	76.74	97.75	96.75	98.26
PANet [42]	9	91.73	87.43	92.19	73.04	70.37	77.32	97.64	96.19	97.98
RetinaNet [17]	9	86.37	85.12	88.09	70.35	69.04	72.09	96.08	95.07	96.95
ASAFE (ours)	1	**95.19**	88.54	95.94	**80.82**	75.66	81.01	**98.78**	97.87	98.89

The best results in the table are shown in bold.

Some visualized detection results are shown in Figure 8, where the first three columns from left to right are inshore scenes and the forth column is an offshore scene. The red bounding boxes denote the ground-truth objects and the green bounding boxes indicate the detection results. Benefiting from the proposed ASA strategy, our method has more excellent classification and regression performance compared to other methods (see the first and forth columns in Figure 8). In the inshore scenes, Double-Head R-CNN, PANet and RetinaNet all show interference by strong scattered areas on the land, resulting in false detections (see the second, third and forth rows of the third column in Figure 8). Faster R-CNN, Double-Head R-CNN, PANet and RetinaNet also have a certain degree of missed detections in the inshore scenes (see the second column in Figure 8). Visualized detection results show that all the methods get poor results in the inshore scenes, particularly for densely arranged ship targets. Our method effectively suppresses the interference of background noise and achieves the best performance with no false detections and less missed detections than other methods.

The learning curve of our proposed method on the SSDD dataset is shown in Figure 9. Sub-figure (a) indicates the relationship between classification loss, regression loss, total loss and epoch. Sub-figure (b) indicates the relationship between the precision, recall, mAP and epoch. According to the learning curve, it can be found that our proposed method has a fast convergence speed, and achieves a precision and mAP of more than 0.8 within 30 epochs of iteration, which proves that our model has a good learning ability.

3.5.2. Results for the HRSID Dataset

The confusion matrix of our proposed method in the entire, inshore and offshore scenes of the HRSID dataset is shown in Table 7. Based on the entire, inshore and offshore results in the confusion matrix, it can be found that our algorithm has excellent precision and recall on the HRSID dataset. Especially for the offshore scenes, our algorithm only generates 40 false positives, which shows the effectiveness of our proposed method. The detection results of different methods on the HRSID dataset are shown in Table 8. Our ASAFE reaches the best detection performance among all methods with only one anchor box and achieves a mAP of 85.18% in the entire scenes, a mAP of 68.91% in the inshore scenes and a mAP of 96.92% in the offshore scenes, outperforming the second-best Double-Head R-CNN algorithm with nine anchor boxes by 4.77% mAP in the entire scenes, 8.73% mAP in the inshore scenes, and 0.94% mAP in the offshore scenes. The comparison results show the superiority of our proposed method, and also verify once again that the selection of high-quality samples can determine the detection performance of a detector more than the number of anchor boxes. Additionally, we observe that Faster R-CNN, Double-Head R-CNN and PANet have a similar performance for precision, recall and the mAP in the entire, inshore and offshore scenes. In addition, there is a big gap between RetinaNet and other algorithms in the detection performance of inshore scenes, which indicates that RetinaNet is greatly affected by the complex backgrounds in inshore scenes. It is worth noting that in the inshore scenes, our method outperforms other algorithms by a large margin on the mAP and recall metrics, indicating that our proposed FAEPN module effectively suppresses the

complex background noise in SAR images and enhances the salient features of ship targets to improve the detection performance of the model.

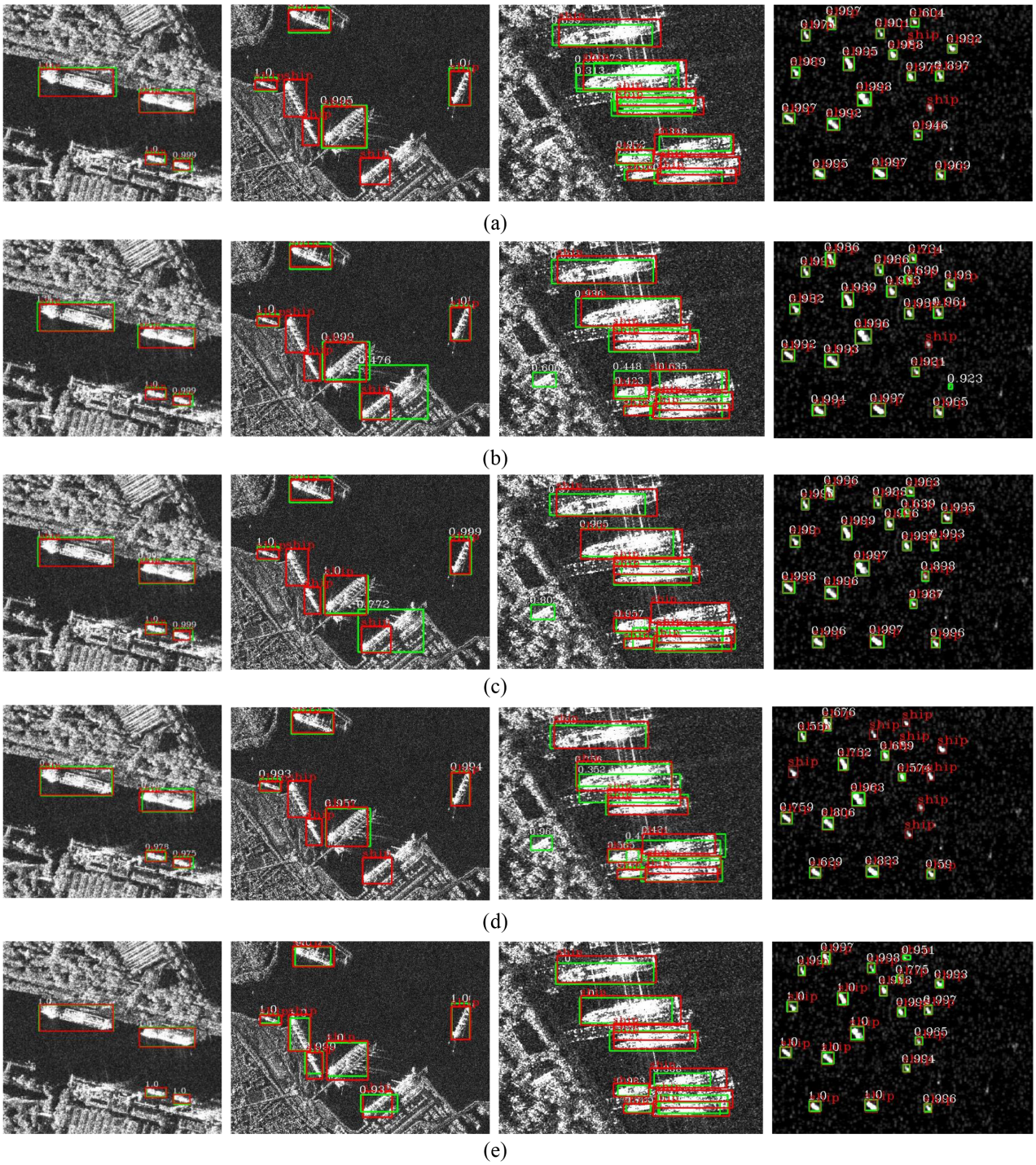

Figure 8. Comparison detection results of different methods in the inshore and offshore scenes of the SSDD dataset. (**a**) Detection results of the Faster R-CNN. (**b**) Detection results of the Double-Head R-CNN. (**c**) Detection results of the PANet. (**d**) Detection results of the RetinaNet. (**e**) Detection results of the ASAFE. Green bounding boxes denote the detection results. Red bounding boxes denote the ground-truth objects.

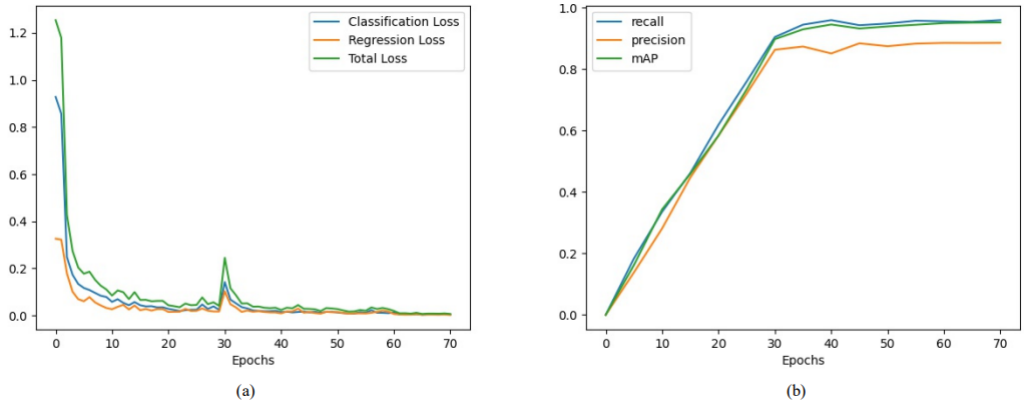

Figure 9. The learning curve on the SSDD dataset. (**a**) The relationship between classification loss, regression loss, total loss and epoch. (**b**) The relationship between the precision, recall, mAP and epoch.

Table 7. The confusion matrix of the HRSID dataset in the entire, inshore and offshore scenes.

Entire Scenes		Inshore Scenes		Offshore Scenes	
TP = 5134	FN = 784	TP = 2044	FN = 797	TP = 2972	FN = 105
FP = 1017	TN = *	FP = 808	TN = *	FP = 40	TN = *

Since the true negatives are not used for the evaluation of detection metrics, we set the TN as *.

Table 8. Comparisons of different methods on the HRSID dataset.

Methods	Anchor Number	Entire Scenes			Inshore Scenes			Offshore Scenes		
		mAP	P	R	mAP	P	R	mAP	P	R
Faster R-CNN [14]	9	79.35	80.03	80.43	59.07	71.04	62.27	96.12	97.58	96.29
Double-Head R-CNN [41]	9	80.41	81.55	81.45	60.18	70.13	64.62	95.98	97.46	96.04
PANet [42]	9	80.11	80.9	81.05	59.91	63.83	64.59	96.26	96.93	96.33
RetinaNet [17]	9	77.32	77.94	79.96	53.46	59.11	61.81	95.76	96.17	95.81
ASAFE (ours)	1	**85.18**	83.46	86.75	**68.91**	71.67	71.95	**96.92**	98.67	96.59

The best results in the table are shown in bold.

Figures 10 and 11 present the detection results of our method and other methods in the inshore and offshore scenes of the HRSID dataset. The red bounding boxes denote the ground-truth objects and the green bounding boxes denote the detection results. According to the detection results in Figure 10, all methods have the problem of missed detections, but Faster R-CNN, Double-Head R-CNN, PANet and RetinaNet are more serious. Some detection results of Double-Head R-CNN and PANet have a poor regression performance. Our method has better detection performance for extremely small ship targets. As shown in Figure 11, RetinaNet has the worst performance of all methods. Although other methods detect all ship targets, our method outperforms other methods in classification score and regression accuracy, which illustrates that the proposed ASA strategy and regression guided loss function improve the detection performance through providing the model with well-classified and well-regressed samples during the training process.

The learning curve of our proposed method on the HRSID dataset is shown in Figure 12. Sub-figure (a) indicates the relationship between classification loss, regression loss, total loss and epoch. Sub-figure (b) indicates the relationship between the precision, recall, mAP and epoch. We can find that although the loss curve has a large fluctuation, which affects the learning effect of the model, the precision and mAP of about 0.8 are still achieved at about 30 epochs. It shows that our model still has good learning ability on the HRSID dataset and good generalizability.

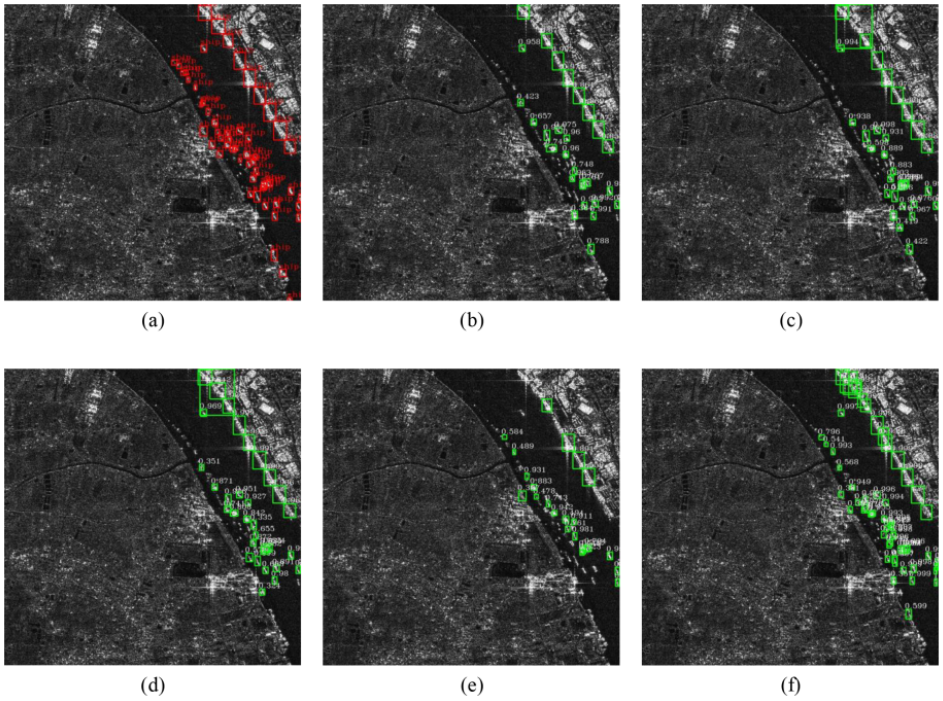

Figure 10. Comparison of detection results for different methods in the inshore scenes of the HRSID. (**a**) Ground-truth objects. (**b**) Detection results of the Faster R-CNN. (**c**) Detection results of the Double-Head R-CNN. (**d**) Detection results of the PANet. (**e**) Detection results of the RetinaNet. (**f**) Detection results of the ASAFE. Green bounding boxes denote the detection results. Red bounding boxes denote the ground-truth objects.

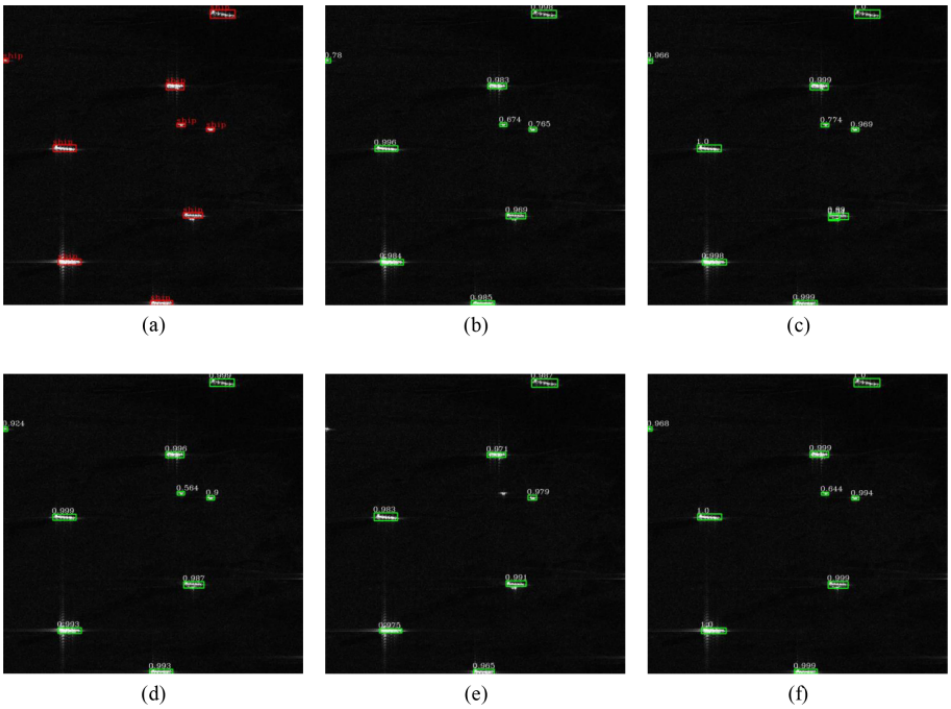

Figure 11. Comparison of detection results for different methods in the offshore scenes of the HRSID. (**a**) Ground-truth objects. (**b**) Detection results of the Faster R-CNN. (**c**) Detection results of the Double-Head R-CNN. (**d**) Detection results of the PANet. (**e**) Detection results of the RetinaNet. (**f**) Detection results of the ASAFE. Green bounding boxes denote the detection results. Red bounding boxes denote the ground-truth objects.

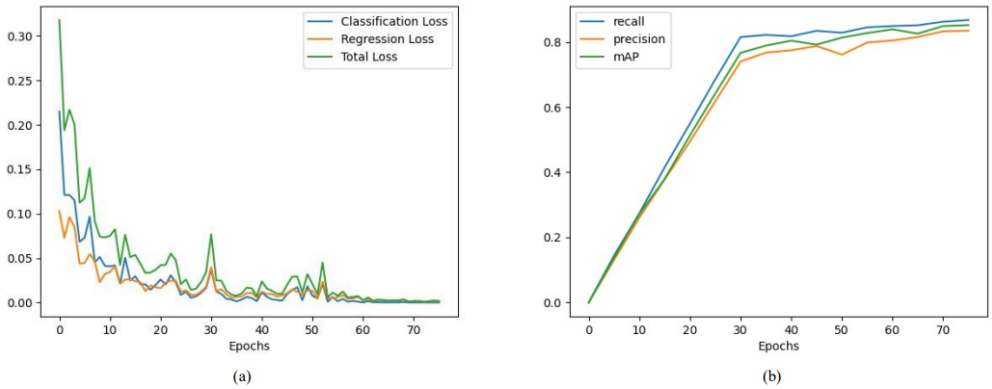

Figure 12. The learning curve on the HRSID dataset. (**a**) The relationship between classification loss, regression loss, total loss and epoch. (**b**) The relationship between the precision, recall, mAP and epoch.

4. Discussion

The experimental results for the SSDD and HRSID datasets illustrate the excellent performance of our proposed method. The extensive ablation experiments of the FAEPN, ASA and Regression Guided Loss show that high-quality sample selection can improve the detection performance and the aggregation, refinement, reconstruction and fusion of multi-scale features can enhance the feature representations and suppress the interference of background noise. As shown in Figures 8 and 10, our method does not perform well when detecting densely arranged and extremely small scale ship targets. In the future, we will consider how to design a more effective method for defining positive and negative samples to improve the detection performance of the model in the case of small scale and densely arranged ship targets.

5. Conclusions

In this article, we first analyze the shortcomings of the original label assignment strategy based on the IoU between anchor boxes and ground-truth objects in SAR images and propose the adaptive sample assignment (ASA) strategy. The ASA strategy generates only one square anchor box on each cell of the feature maps, which reduces the parameters and computational overheads of the model. Additionally, the detection performance of the model is improved by selecting the high-quality positive samples during the training process by the proposed spatial constraint condition and consistent prediction condition. Specifically, spatial constraint condition is used to quickly filter out anchor boxes that do not match the ground-truth objects at the spatial scale to construct the candidate sample set. Consistent prediction condition adaptively filters the false candidate sample based on the knowledge learned from the classification and regression branches. Moreover, the regression guided loss function is proposed to further guide the selection process of the high-quality positive samples by introducing regression performance of the samples as a soft label into the classification loss function. We also propose the feature aggregation enhancement pyramid network (FAEPN) to alleviate the false alarms by enhancing the salient features of the ship targets and suppressing the interference of background noise. FAEPN consists of Feature Aggregation Module (FAM), Fine-grained Tuning Module (FTM), Rebuild Module (RBM) and Level Fusion Module (LFM). FAM is used to achieve the aggregation of multi-scale feature information. FTM is proposed to achieve refinement of feature map. RBM is used to reconstruct the feature hierarchy to achieve better detection performance for multi-scale ship targets. Finally, LFM achieves multi-scale feature enhancement by alleviating the feature aliasing problem caused by interpolation and pooling operations. Extensive experiments on the SSDD dataset and the HRSID dataset verify the effectiveness of our proposed method. We reached a mAP of 95.19%, precision of 88.54% and the recall of 95.94 on the SSDD dataset. On the HRSID dataset, we reached a mAP of

85.18%, precision of 83.46% and the recall of 86.75%. The detection results on two datasets outperform many other methods. In the future, we will design the more effective and accurate label assignment strategy to achieve better performance in a case of small scale and densely arranged ship targets.

Author Contributions: Conceptualization, Z.F. and H.S.; methodology, Z.F.; software, Z.F.; validation, Z.F., H.S., L.C. and Y.W.; formal analysis Z.F. and H.S.; investigation, L.C. and Y.W.; writing—original draft preparation, Z.F., H.S. and L.C.; writing-review and editing, Z.F., H.S., Y.W. and L.C.; visualization, Z.F. and L.C.; funding acquisition, H.S. and L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in National Natural Science Foundation of China No. 6210010891, the Chang Jiang Scholars Program under Grant T2012122 and the Hundred Leading Talent Project of Beijing Science and Technology under Grant Z141101001514005.

Data Availability Statement: Publicly available datasets were used in this study. SSDD data can be found here: https://github.com/TianwenZhang0825/Official-SSDD, accessed on 5 April 2022 HRSID data can be found here: https://github.com/chaozhong2010/HRSID, accessed on 5 April 2022.

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

References

1. Habibollahi, R.; Seydi, S.T.; Hasanlou, M.; Mahdianpari, M. TCD-Net: A novel deep learning framework for fully polarimetric change detection using transfer learning. *Remote Sens.* **2022**, *14*, 438. [CrossRef]
2. Wang, J.; Wang, Y.; Liu, H. Hybrid Variability Aware Network (HVANet): A self-supervised deep framework for label-free SAR image change detection. *Remote Sens.* **2022**, *14*, 734. [CrossRef]
3. Liu, S.; Kong, W.; Chen, X.; Xu, M.; Yasir, M.; Zhao, L.; Li, J. Multi-scale ship detection algorithm based on a lightweight neural network for spaceborne SAR images. *Remote Sens.* **2022**, *14*, 1149. [CrossRef]
4. Krek, E.V.; Krek, A.V.; Kostianoy, A.G. Chronic oil pollution from vessels and its role in background pollution in the Southeastern Baltic Sea. *Remote Sens.* **2021**, *13*, 4307. [CrossRef]
5. Tang, L.; Tang, W.; Qu, X.; Han, Y.; Wang, W.; Zhao, B. A scale-aware pyramid network for multi-scale object detection in SAR images. *Remote Sens.* **2022**, *14*, 973. [CrossRef]
6. Zhang, J.; Zhang, W.; Hu, Y.; Chu, Q.; Liu, L. An improved sea ice classification algorithm with Gaofen-3 dual-polarization SAR data based on deep convolutional neural networks. *Remote Sens.* **2022**, *14*, 906. [CrossRef]
7. Chen, F.; Zhang, Y.; Zhang, J.; Liu, L.; Wu, K. Rice false smut detection and prescription map generation in a complex planting environment, with mixed methods, based on near earth remote sensing. *Remote Sens.* **2022**, *14*, 945. [CrossRef]
8. Gierull, C.H. Demystifying the capability of sublook correlation techniques for vessel detection in SAR imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2031–2042. [CrossRef]
9. Pappas, O.; Achim, A.; Bull, D. Superpixel-level CFAR detectors for ship detection in SAR imagery. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1397–1401. [CrossRef]
10. 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]
11. 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]
12. Yang, R.; Wang, G.; Pan, Z.; Lu, H.; Zhang, H.; Jia, X. A novel false alarm suppression method for CNN-based SAR ship detector. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1401–1405. [CrossRef]
13. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A novel quad feature pyramid network for SAR ship detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2016**, *39*, 1137–1149. [CrossRef]
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
17. 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.
18. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
19. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.

20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Ke, X.; Zhang, X.; Zhang, T.; Shi, J.; Wei, S. SAR ship detection based on an improved faster r-cnn using deformable convolution. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 3565–3568.
22. Du, L.; Li, L.; Wei, D.; Mao, J. Saliency-guided single shot multibox detector for target detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3366–3376. [CrossRef]
23. Yu, J.; Zhou, G.; Zhou, S.; Qin, M. A fast and lightweight detection network for multi-scale SAR ship detection under complex backgrounds. *Remote Sens.* **2022**, *14*, 31. [CrossRef]
24. Chang, Y.L.; Anagaw, A.; Chang, L.; Wang, Y.C.; Hsiao, C.Y.; Lee, W.H. Ship detection based on YOLOv2 for SAR imagery. *Remote Sens.* **2019**, *11*, 786. [CrossRef]
25. Wei, S.; Su, H.; Ming, J.; Wang, C.; Yan, M.; Kumar, D.; Shi, J.; Zhang, X. Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet. *Remote Sens.* **2020**, *12*, 167. [CrossRef]
26. Yu, L.; Wu, H.; Zhong, Z.; Zheng, L.; Deng, Q.; Hu, H. TWC-Net: A SAR ship detection using two-way convolution and multiscale feature mapping. *Remote Sens.* **2021**, *13*, 2558. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
29. Ming, Q.; Miao, L.; Zhou, Z.; Dong, Y. CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5605814. [CrossRef]
30. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic anchor learning for arbitrary-oriented object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Palo Alto, CA, USA, 2–9 February 2021; Volume 35, pp. 2355–2363.
31. Kim, K.; Lee, H.S. Probabilistic anchor assignment with iou prediction for object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 355–371.
32. Li, H.; Wu, Z.; Zhu, C.; Xiong, C.; Socher, R.; Davis, L.S. Learning from noisy anchors for one-stage object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10588–10597.
33. Ke, W.; Zhang, T.; Huang, Z.; Ye, Q.; Liu, J.; Huang, D. Multiple anchor learning for visual object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10206–10215.
34. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12595–12604.
35. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
36. Wang, X.; Zhang, S.; Yu, Z.; Feng, L.; Zhang, W. Scale-equalizing pyramid convolution for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13359–13368.
37. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea, 27–28 October 2019; pp. 1971–1980.
38. 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]
39. 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]
40. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
41. 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, 13–19 June 2020; pp. 10186–10195.
42. 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.