remote sensing

MDPI

Article

A Scale-Aware Pyramid Network for Multi-Scale Object Detection in SAR Images

Linbo Tang [1,2,†], Wei Tang [2,†], Xin Qu [3], Yuqi Han [4,*], Wenzheng Wang [2] and Baojun Zhao [2]

[1] Advanced Technology Research Institute, Beijing Institute of Technology, Jinan 250300, China; tanglinbo@bit.edu.cn
[2] Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; tgwi@bit.edu.cn (W.T.); wang_wenzheng@pku.edu.cn (W.W.); zbj@bit.edu.cn (B.Z.)
[3] Air and Space Defense System Lab, Beijing Institute of Electronic Engineering, Beijing 100074, China; qubuaa@gmail.com
[4] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
* Correspondence: yuqi_han@tsinghua.edu.cn
† These authors contributed equally to this work.

Abstract: Multi-scale object detection within Synthetic Aperture Radar (SAR) images has become a research hotspot in SAR image interpretation. Over the past few years, CNN-based detectors have advanced sharply in SAR object detection. However, the state-of-the-art detection methods are continuously limited in Feature Pyramid Network (FPN) designing and detection anchor setting aspects due to feature misalignment and targets' appearance variation (i.e., scale change, aspect ratio change). To address the mentioned limitations, a scale-aware feature pyramid network (SARFNet) is proposed in this study, which comprises a scale-adaptive feature extraction module and a learnable anchor assignment strategy. To be specific, an enhanced feature pyramid sub-network is developed by introducing a feature alignment module to estimate the pixel offset and contextually align the high-level features. Moreover, a scale-equalizing pyramid convolution is built through 3-D convolution within the feature pyramid to improve inter-scale correlation at different feature levels. Furthermore, a self-learning anchor assignment is set to update hand-crafted anchor assignments to learnable anchor/feature configuration. By using the dynamic anchors, the detector of this study is capable of flexibly matching the target with different appearance changes. According to extensive experiments on public SAR image data sets (SSDD and HRSID), our algorithm is demonstrated to outperform existing boat detectors.

Keywords: synthetic aperture radar; multi-scale object detection; feature pyramid network; convolutional neural network

Citation: 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. https://doi.org/10.3390/rs14040973

Academic Editors: Xichao Dong, Yuanhao Li, Cheng Hu, Andrea Monti Guarnieri, Dušan Gleich and Pedro Melo-Pinto

Received: 20 December 2021
Accepted: 11 February 2022
Published: 16 February 2022

1. Introduction

Synthetic Aperture Radar (SAR) has been found as a microwave remote sensing system capable of effectively acquiring SAR images at high resolutions under all-weather and complex environmental conditions. As spaceborne SAR advances sharply (e.g., distributed spaceborne SAR systems and spaceborne–airborne bistatic SAR), SAR turns out to be critical to the military and civilian fields. Object detection based on SAR images is critical to land reconnaissance, military intelligence acquisition and marine management. However, SAR images contain numerous targets at significantly different scales (e.g., ships, bridges and vehicles). Moreover, the shape of the target changes due to the cross-sidelobe blur and scattering points, especially the wide distribution of the object's aspect ratio. Accordingly, a robust object detector is required to exhibit the learning capabilities for scale-variant and shape-variant targets within SAR images.

According to the state-of-the-art algorithms for detection, the constant false-alarm rate (CFAR) approach [1] has been found as a highly common technique for SAR object detection.

The CFAR approach is capable of computing the adaptive threshold by complying with the set false-alarm rates and the background clutters' statistical distributions and subsequently distinguishing object pixels according to the background through the comparison of the pixel intensity based on the determined threshold. On the whole, the CFAR approaches can show high performance under simple scenes. Nevertheless, the CFAR algorithm has been significantly dependent of statistical areas of background. Overall, the description can be more accurate, as the statistical area of the background is broadened, whereas the mentioned approach is likely to cause more noticeable variations in background clutters and upregulate the rates of false alarms. For the mentioned reason, target detection for large-scale SAR images is unlikely to be achieved with the use of the conventional approaches.

Over the past few years, deep-learning-based detection algorithms have achieved satisfactory detection performance in SAR images and gained the attentions of many researchers. Most of these methods either employ Feature Pyramid Network (FPN) [2] structure for the extraction of features at multiple scales or exploit anchor-based/anchor-free schemes to align spatial features of SAR objects at various scales and with different shapes.

On one hand, FPN integrates the feature layers with various spatial resolutions and semantic features on the basis of top-to-bottom paths as well as horizontal connections. In this way, FPN could gather rich semantic information at all levels from a single-scale image. On the basis of the high multi-scale feature expression ability of FPN, some subsequent studies have adopted FPN in detection pyramid for detecting targets at various scales within SAR images. The following are some examples. Based on feature balance and a refined network, Fu et al. [3]. developed a multi-scale SAR ship anchorless frame detection method and improved the semantic feature information of small objects by setting a balanced pyramid structure under the attention mechanism. Zhao et al. [4] built an attention-receptive pyramid network to detect ships exhibiting different sizes under complex backgrounds. Zhang et al. [5] built a Balance Scale Global Attention FPN that could refine features at the respective feature level in the pyramid to solve the feature-level imbalance of different-scale ships. Given that the element-wise addition fusion method cannot fully exploit the guiding significance of the semantic features in the feature pyramid Network, Zhao et al. [6] developed the semantic attention module (SAM) to fuse features in various resolutions. Zhou et al. [7] proposed a cross-scale object detection method for SAR images based on a Scale Expansion Pyramid Network (SEPN) to address objects with large scale differences in SAR images.

Furthermore, anchor-based and anchor-free strategies are exploited to extract targets' spatial features in SAR detection applications. To be specific, anchor-based detectors generally employ various scales and aspect ratio anchor boxes to improve the ability of object detection algorithms to be generalized for a range of object shapes and scales. In addition, anchor-free methods follow the idea of image segmentation, through which objects are predicted based on the key points or the pixels of the object center point. For instance, Guo et al. [8] and Liu et al. [9] use CenterNet [10] as an object detection framework to detect multi-scale SAR targets, which is an anchor-free method based on key points. On the other hand, Cui et al. [11] and Zhao et al. [4] employed Faster RCNN [12] as an anchor-based detection framework to detect multi-scale SAR targets.

The CNN-based algorithms above have made significant breakthroughs in multi-scale SAR object detection, but some problems remain unsolved.

Two defects of the existing FPN structure should be noted. (1) Feature misalignment caused by inaccurate spatial sampling: For the existing FPN framework, the spatially coarser (higher-level) feature maps are generally upsampled prior to merging with the corresponding maps of features within the bottom-up path. However, impacted by common non-learnable properties of upsampling operations (e.g., the closed neighbors) and repeated applications of downsampling and upsampling, there is an inaccurate correspondence between bottom-up and upsampling features (i.e., feature misalignment). The wrong features will adversely affect the learning of subsequent layers, leading to the wrong classification of the final prediction. (2) Ignorance of inter-layer correlation between multi-

level scale features: In the feature pyramid, the semantic information covered in the features at different levels shows significant differences. A wide variety of feature fusion strategies are developed for maintaining the representation capabilities of the feature maps at different levels to be consistent. However, the mentioned methods tend to directly add features with identical resolutions, and they do not take the inherent properties of the feature pyramid into account. For instance, in the feature pyramid, feature maps with adjacent scales should have a strong correlation, whereas the correlation above is not considered in the state-of-the-art fusion method.

Moreover, the above CNN-based detection algorithms have defects in assigning anchors for the targets (label assignment strategy). For instance, the existing label assignment strategy performs inferiorly with many missed detections when handling densely arranged targets or the ones without central features (e.g., slender or crescent-shaped ship objects). The mentioned issues are mainly caused by the assignment by modern object detectors on the basis of CNN to anchors in terms of a ground-truth object according to object–anchor Intersection-over-Union (IoU) limitation [13]. Based on the premise that anchors spatially aligned with the objects are constantly acceptable for classification and localization, the respective anchor under the assignment can achieve independent monitoring of network learning to predict objects. However, the space alignment is unnecessarily the only standard for assigning anchors. First, the most typical characteristics exhibited by objects with acentric characteristics are far from their geometric cores. Thus, an anchor with spatial alignment is likely to be unable to comply with the most typical feature of the anchor, thereby leading to the reduction of the performance of localization and classification. Furthermore, when multiple objects are clustered together, it is not possible to use IoU standards for matching objects with appropriate anchors/features.

To address the limitations above, this study builds a scale-aware feature pyramid network (SARFNet) that largely covers a scale adaptive feature extraction sub-network and a self-learning anchor assigning scheme. To be specific, we exploit an enhanced feature pyramid and a scale-equalizing pyramid convolution to overcome the defects of feature misalignment and neglect of inter-layer correlation in FPN. Our proposed feature pyramid embeds a feature alignment module and feature selection module into the feature pyramid. The feature alignment module acquires knowledge for the alignment of the upsampled map of features toward several feature maps to be referenced through the adjustment of the respective sampling position within the convolution kernel using the learning offset. The feature selection module adaptively selects the underlying feature map with rich spatial details through channel-attention and spatial-attention mechanisms. Furthermore, a scale-balanced pyramid convolution (SEPC) is deployed into the detection framework to extract the correlation information between feature layers based on FPN. To be specific, it adopts 3-D convolution to correlate similar feature maps and explore the interaction between scales. On the other hand, we utilize a learning and matching (LTM) method to improve the existing IoU restriction, allowing the target to more flexibly match anchors. The LTM achieves an updating of "free" anchor matching from anchor assignments made by handicraft through the formulation of detector training on the basis of the structure of Maximum Likelihood Estimation (MLE).

The main novelties our proposed model can be summarized in threefold:

(1) This study develops a scale-aware pyramid object detection framework for SAR images, which can effectively detect objects with large-scale variation and appearance changes by considering the target's unique characteristics in SAR images.
(2) A novel feature pyramid structure is proposed in this study, which can address defects of feature matching schemes in state-of-the-art feature pyramids and enhance the feature modeling ability for targets in SAR images.
(3) Numerous experiments are carried out to verify the effectiveness of the proposed method as compared with the existing literature.

The remainder of this article is organized as follows. In Section 2, we briefly introduce the works related to SAR object detection and multi-scale object detection based on deep

learning. In Section 3, we introduce the model architecture of the SARFNet object detector. In Section 4, we discuss the performance of the proposed method and compare it with that of State-of-the-Art object detection methods on two SAR object detection data sets (SSDD and HRSID). Finally, we present a few concluding remarks in Section 5.

2. Related Works

In this section, we briefly introduce SAR object detection and multi-scale object detection based on deep learning. These studies have greatly contributed to our method.

2.1. SAR Object Detection

Object detection in SAR images has aroused rising attention over the past decade. Conventional ship detection algorithms comply with CFAR. CFAR algorithms primarily carry out a pixel-by-pixel detection of SAR images under local sliding windows. The respective pixel in the SAR image participates in the sliding window operation multiple times, which makes the calculation speed of the algorithm generally low. However, as high-resolution SAR images emerge, complex ground interference and texture scenes are commonly accompanied by multi-scale targets, so the calculation of considerable background clutter pixels in large-scene SAR images turns out to be time-consuming. To address the problem of multi-scale target detection in SAR images, some scholars have proposed numerous improved CFAR target detection algorithms (e.g., the iterative CFAR object detection algorithm [14], the optimized super-pixel CFAR object detection algorithm [15] and the scale sliding window object detection algorithm [16]). To acquire more target information and eliminate the effect of coherent speckle noise, Li et al. [15] built a two-stage CFAR detection algorithm for object super-pixel detection, which has a better detection effect for ships in simple scenarios. The detection results of diverse targets at different scales are poor. To detect targets at large scales, Zhai et al. [17] proposed a ship detection algorithm for saliency and contextual information processing, which is capable of focusing on large ships and background targets with prominent features; however, this method ignores small ships. The mentioned methods are only for scenes with simple backgrounds in SAR images. Though the detection effect is high, the detection performance will be reduced in the case of complex large-scene SAR images.

Over the past few years, thanks to the autonomous feature learning ability of deep learning, researchers have applied object detection under deep learning to SAR images. In Du et al. [18], the saliency information is introduced to the SSD detection algorithm to more effectively guide the SSD deep learning network to independently learn the salient features of the SAR target and reduce false alarms. Cui et al. [9] substituted the spatial shuffle-group enhanced attention module into CenterNet as an SAR target feature enhancement module to reduce the false alarms attributed to the interference of offshore and inland. Fu et al. [3] yielded an anchor-free object detection algorithm in accordance with feature balance pyramid, which was found to be able to detect multi-scale SAR ships in complex scenes. Lin et al. [19] added the squeeze and excitation rank module to the backbone network in the Faster R-CNN target detection algorithm to improve the performance of Faster R-CNN in detecting SAR targets. Cui et al. [11] introduced the dense attention mechanism to the feature pyramid to improve FPN's multi-scale feature expression ability. Zhao et al. [4] built a two-stage attention receptive pyramid network to enhance the performance of recognizing multiscale ships in SAR pictures by improving links between nonlocal characteristics and refining information in distinct feature maps. Wei et al. [20] developed a detection method for the detection of ships in high-resolution SAR images by complying with a high-resolution ship detection network. In order to fuse SAR images with rich polarization information and optical images with spatial detail information to improve the performance of change detection, Li et al. [21] propose a deep-translation-based change detection network (DTCDN) for optical and SAR images

Although progress has been made, due to the inherent noise characteristics of SAR images, existing detection networks still have limitations in improving detection perfor-

mance. Therefore, before using SAR images for deep learning, a separate preprocessing step is required, such as Noise (despeckle). Therefore, SAR image denoising has become a current research hotspot. For example, Mukherjee Li et al. [22] propose interferogram denoising and coherence prediction using two separate CNN architectures to eliminate noise information in Interferometric Synthetic Aperture Radar (InSAR) images. To reduce the loss of target information caused by SAR image denoising, Shin et al. [23] propose a new target detection framework that combines an unsupervised denoising network into a traditional two-stage detection network and uses a strategy to fuse the region suggestions extracted from the original SAR image and the synthetic denoising SAR image.

Unlike the algorithms above, the feature alignment module and the SEPC module are substituted into the feature pyramid to address feature misalignment and the neglect of inter-layer correlation within the feature pyramid. Moreover, a learning-to-match anchors strategy is employed in anchor design to update hand-crafted anchor assignment to learnable anchor/feature configuration.

2.2. Multi-Scale Object Detection Based on Deep Learning

Scale variation across object instances has been treated as one of the most knotty problem in modern development of detection. To address this challenge, several approaches have been proposed. An image pyramid is an intuitive method, where SNIP [24] and SNIPER [25] select a specific scale for each resolution during multi-scale training. Nevertheless, the image pyramid approaches significantly increase the inference time, making them unsuitable for practical applications.

Another approach aims to employ in-network feature pyramids to approximate image pyramids with less computational costs. The idea was first demonstrated in the construction of a fast feature pyramid for object detection by interpolating some feature channels from nearby-scale levels. In the deep learning era, the approximation results have become even more accurate. For example, a single-shot detector (SSD) [26] utilizes multiscale feature maps from different layers and detects objects of various scales at each feature layer. DSSD [27] and MS-CNN [28] perform object detection at multiple layers for objects of different scales. However, these bottom-up sampling methods have low accuracy for small objects due to the insufficient semantic information in low-level features.

To compensate for the absence of semantics in low-level features, many feature pyramid structures [2,26,29–32] that make more effective use of multi-scale features have been proposed. The FPN [2] exploits a bottom-up pathway, a top-down pathway, and lateral connections to efficiently fuse features of various resolutions and scales. However, as reported by several recent studies, some problems remain in the FPN structure. For instance, the top-down pathway FPN only introduces high-level semantic information to low-level features, while ignoring the role of low-level features for localization. To address this issue, SA-FPN [29], combining Top-Down style FPN and Bottom-Up style FPN, absorbs the characteristics of both and becomes a more accurate module for scale variation perception. Researchers at FSAF [33] stressed a flaw in the FPN's heuristic-guided feature selection. Furthermore, in the training process, online feature selection should be exploited to dynamically determine the most appropriate number of features for the respective instance. Additionally, in FSAF, the feature representation of the respective region of interest (RoI) is extracted at a single feature level and any valid information existing on the pyramid's other feature levels is ignored. As opposed to the mentioned finding, Libra RCNN [34] and PANet [30] demonstrated that regardless of an object's size, all feature map layers present meaningful information for object detection. The methods above integrate valid information at all feature levels through a common operation (an elementwise max or sum). However, the fusion mechanism of simple addition does not consider semantic differences between various feature layers. To solve this problem, EfficientDet [35], a bidirectional-weighted FPN for simple and rapid feature fusion, has been presented. YOLO-ASFF [36], in a unique adaptive spatial feature fusion technique, learns an adaptive spatial fusion weight in the training process to filter out inconsistencies. M2Det [37] is used to discover that the

respective feature map in a pyramid comprises single-level features primarily or entirely, which influences detection performance. A multilevel FPN (MLFPN) is proposed. However, the MLFPN significantly complicates the model to design a more effective feature pyramid.

However, the following approaches still face certain difficulties. For M2Det, the MLFPN adds excessive parameters while neglecting the valid information in the low-level features. Though Libra-RCNN and PANet attempt to fuse valid information at all feature levels, their structure's feature fusion operation (e.g. BFP in Libra-RCNN or AFP in PANet) can merely take up a small portion of the feature fusion space and can be more effectively improved with a more flexible approach. Furthermore, EfficientDet and YOLO-ASFF ignore the multiscale information at the respective convolution layer. SEPC [38] proposes the PConv (Pyramid Convolution) theory, in which 3-D convolution is exploited to correlate similar feature maps and mine the interaction between scales. Moreover, since the features of the feature pyramid vary significantly between layers, SEPC performs deformable convolution on the high-level features of the feature pyramid, which can fit the scale changes in practice.

Inspired by SEPC and FaPN [39], this study develops a novel Scale-Aware Pyramid Network to build more effective feature representations for objects at different scales. First, by establishing a global context attention mechanism and a spatial context attention mechanism, contextual information is extracted from the channel and the space to improve the ability of multi-scale feature expression. Subsequently, this study employs a 3-D convolution to correlate similar feature maps and investigates the interactions between scales.

3. Proposed Framework

In this section, we introduce the model architecture of the SARFNet object detector and different designing choices in details.

3.1. Network Architecture

The designed scale-aware FPN (SARFNet) is mostly modified from RetinaNet [40] with the identical backbone (Figure 1). Compared with RetinaNet, the network structure of SARFNet involves three differences, as reflected in the feature pyramid structure, the head structure of the classification and regression branch, as well as the label assignment strategy. First, to address the feature mismatch problem, a feature alignment module and a feature selection module are introduced to the feature pyramid network structure to improve the multi-scale feature expression ability of the FPN. Second, the Scale-equalizing pyramid convolution is employed to modify the head design of RetinaNet as an attempt to enhance the information interaction between the classification branch and the regression branch and extract the Cross-scale correlation in the FPN pyramid. Lastly, to solve the problem that the label assignment by the hand-crafted IoU criterion in RetinaNet causes the low detection performance of SAR targets with various appearances and aspect ratios, this study introduces a learning-to-match strategy with the aim to update hand-crafted anchor assignments to learnable anchor/feature configurations.

3.2. Enhanced Feature Pyramid Network

As shown above, the design of the classic feature pyramid shows defects in the pixel-level upsampling and the feature-mapping level. Inspired by Feature Aligned Feature Pyramid Network [39], the Enhanced Feature Pyramid Network (E-FPN) is developed to extract robust multi-scale features. To be specific, the Enhanced FPN consists of a Feature Selection Module (FSM) and a Feature Alignment Module (FAM), as shown in Figure 2. This study defines the output of the i-th stage of the bottom-up network as C_i, which has a span of 2^i pixels relative to the input image, that is, $C_i \in R^{\frac{H}{2^i} \times \frac{W}{2^i}}$, where the size of the input image is $H \times W$. We use $\hat{C}_i$ to represent the output of the FSM layer given C_i input. The output feature after the fusion of the i-th feature from the top-down path is defined as P_i, and the upsampled and aligned features to C_{i-1} are P_i^{up} and $\hat{P}_i^{up}$, respectively.

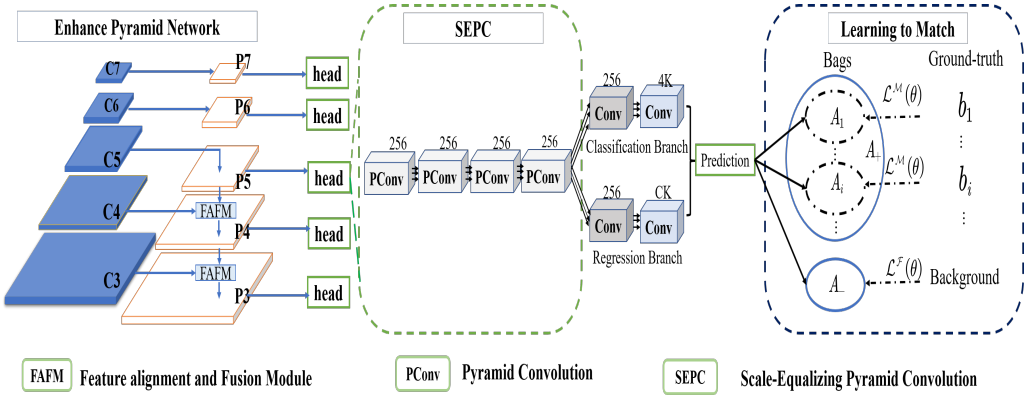


Figure 1. Workflow of the proposed scale-aware pyramid object detection network. There are three components: The Enhance FPN block, SEPC block, and learning-to-match strategy. According to the IoU criterion, anchors in A are categorized to multiple positive anchor bags $A_i \subseteq A_+$ and a negative anchor bag A_- , and $A = A_+ \bigcup A_-$. During detector training, minimizing the anchor matching loss $\mathcal{L}^{\mathcal{M}}(\theta)$ drives matching positive/negative anchors within the positive anchor bag in a "soft" manner.

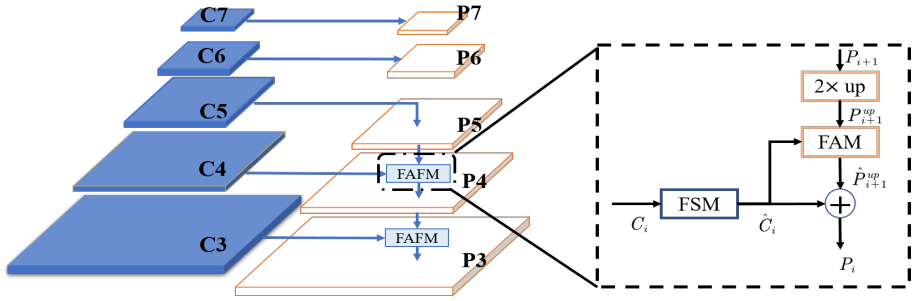


Figure 2. Feature-aligned Pyramid Network Structure.

Feature Selection Module. Classical FPN simply adopts a 1×1 convolution for channel compression before feature fusion to maintain the number of channels for high and low features to be unchanged. However, since the significance of the respective channel feature is not judged, the simple 1×1 convolution will eliminate vital features containing excessive spatial details when channel reduction is being carried out. To solve the problem above, a feature selection module (FSM) is proposed to explicitly model the importance of feature maps, suppress redundant feature maps and recalibrate them accordingly.

Figure 3 illustrates the structure of the proposed FSM module. First, the global information z_i of the respective input feature map c_i is extracted based on the global average pool operation. Subsequently, the global information z_i is sent to the feature importance modeling layer $f_m(\cdot)$ to learn the weight of the respective channel within the input feature map. The mentioned weights represent the significance of the respective feature map by the importance vector $\boldsymbol{u}$. Next, we use the important vector to scale the original input feature map, and then add the scaled feature map to the original feature map to form a re-scaled feature map. Lastly, the feature selection layer $f_s(\cdot)$ is introduced on the rescaled feature map to selectively maintain vital feature maps and delete useless feature maps to reduce channels. On the whole, the process of FSM is formulated as follows:

$$u = f_m(\boldsymbol{z}) \tag{1}$$

$$\hat{C}_i = f_s(C_i + u \times C_i) \tag{2}$$

where the global information $\boldsymbol{z} = [z_1, z_2, \cdots, z_D]$ can be calculated by $z_d = \frac{1}{H_i \times W_i} \sum_{h=1}^{H_i} \sum_{w=1}^{W_i} c_d(h, w)$. $\boldsymbol{u} = [u_1, u_2, \cdots, u_D]$ is the feature importance vector, and u_d is the importance of the d-th input feature map. $f_m(\cdot)$ is the feature importance modeling

layer, which can be modeled by a 1×1conv layer followed by a sigmoid activation function. $f_s(\cdot)$ is the feature selection layer, which is modeled by 1×1 convolution.

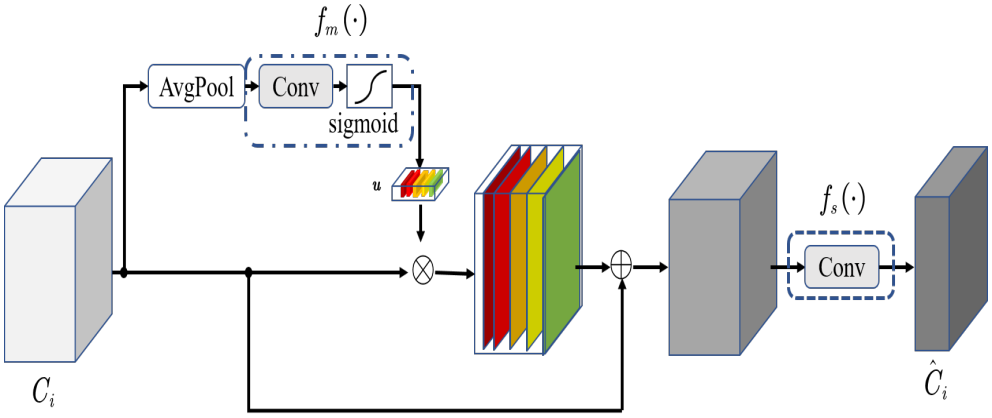


Figure 3. Feature Selection Module Structure.

Feature Alignment Module. A predictable contextual misalignment is caused for the upsampling high-level features P_i^{up} and the low-level features C_{i-1} since the classic FPN recursively draws upon downsampling operations. Accordingly, the feature fusion method by exploiting element superposition or channel splicing will impact the prediction of the target boundary and then cause misclassification during the prediction. Thus, the feature fusion method by applying element superposition or channel splicing will impact the prediction of the object boundary and subsequently cause misclassification during the prediction. To address the problem above, a feature alignment module [39] learning to align the upsampled feature map to a set of reference feature maps is proposed by regulating the respective sampling position in the convolution kernel under the learning offset. Figure 4 [39] illustrates the workflow of Feature Alignment Module. Before feature aggregation, we align the upsampled feature map P_i^{up} with its reference feature $\acute{C}_{i-1}$, i.e., the feature map P_i^{up} is regulated based on the spatial location information offered by $\acute{C}_{i-1}$.

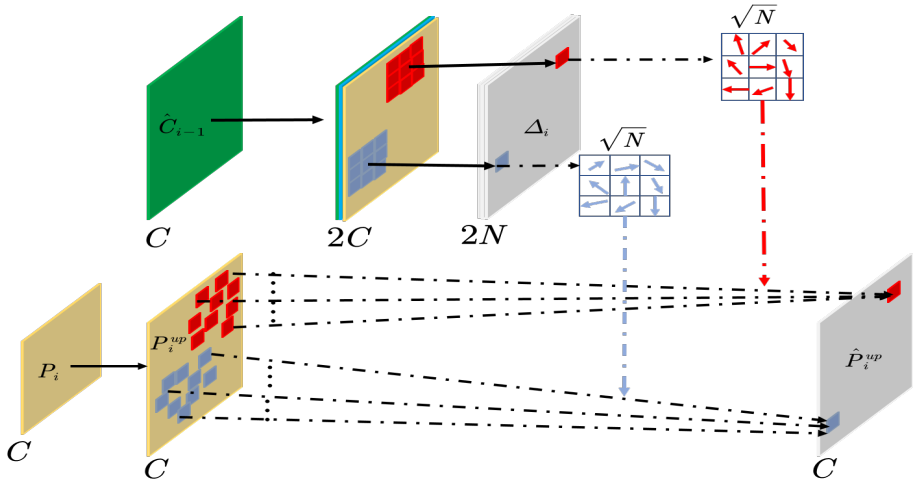


Figure 4. Schematic diagram of the structure of the feature alignment module. N denotes a convolutional kernel of N sample locations. C denotes the number of channels of the feature. Δ_i to represent the convolution kernel offset to be learned.

3.3. Scale-Equalizing Pyramid Network-SEPC

For the feature pyramid, feature maps of adjacent scales are required to be significantly correlated with each other, whereas this correlation is not considered in the state-of-the-art feature fusion method. Accordingly, Scale-Equalizing Pyramid Network-SEPC [38]

is employed for addressing the mentioned problems. SEPC can capture the interaction between scales through Pyramid Convolution (PConv).

The Pyramid Convolution (PConv) refers to a 3-D convolution that spans scales and spatial dimensions. If the feature at the respective level is expressed as a point in Figure 5, PConv can be denoted as N different 2-D convolution kernels. However, there is a problem of size mismatch between different-level feature maps within the feature pyramid. As the pyramid level rises, the size of the space will shrink. To adapt to the shape and size mismatch, a range of strides for K various kernels are set under convolution at diverse layers. For instance, for PConv with $N = 3$, the stride of the first kernel should be 2, and the stride of the last kernel should be 0.5. Then, the output of PConv can be defined as follows:

$$y^l = \omega_1 \cdot x^{l-1} + \omega_2 \cdot x^l + \omega_3 \cdot x^{l+1} \tag{3}$$

where l denotes pyramid level, ω_1,ω_2,ω_3 express 3 parameter-sharing convolution kernels containing 2, 1 and 0.5 strides, respectively. The kernel with a step size of 0.5 is further replaced by an ordinary convolution with a step size of 1 and a continuous bilinear upsampling layer.

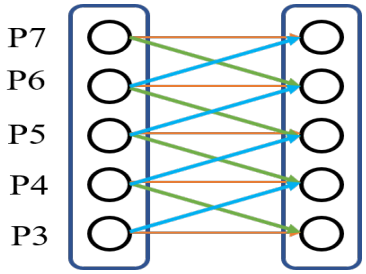


Figure 5. The Pyramid Convolution Structure.

The PConv, in addition to being able to extract scale-related features, also benefits from its compatibility with head design of RetinaNet. According to Figure 6a, the RetinaNet head is a PConv with a scale kernel of 1. Thus, the 4 convolution heads can be directly replaced with our PConv module with a scale kernel of 3. Nevertheless, the respective PConv still leads to additional calculations. To simplify the calculation and improve the connection between the regression branch and the classification branch, as an alternative, the classification and positioning branches are shared with 4 PConv modules to build a combined header structure (Figure 6b). An additional ordinary convolution is introduced after the shared 4 PConv modules to satisfy the differences in classification and positioning tasks. To solve the problem of feature mismatch during the upsampling process, a deformable convolution is added after the upsampling step in SEPC, and the problem of feature mismatch is alleviated to a certain extent through the learning method, which is the same as the principle of the Feature Alignment Module.

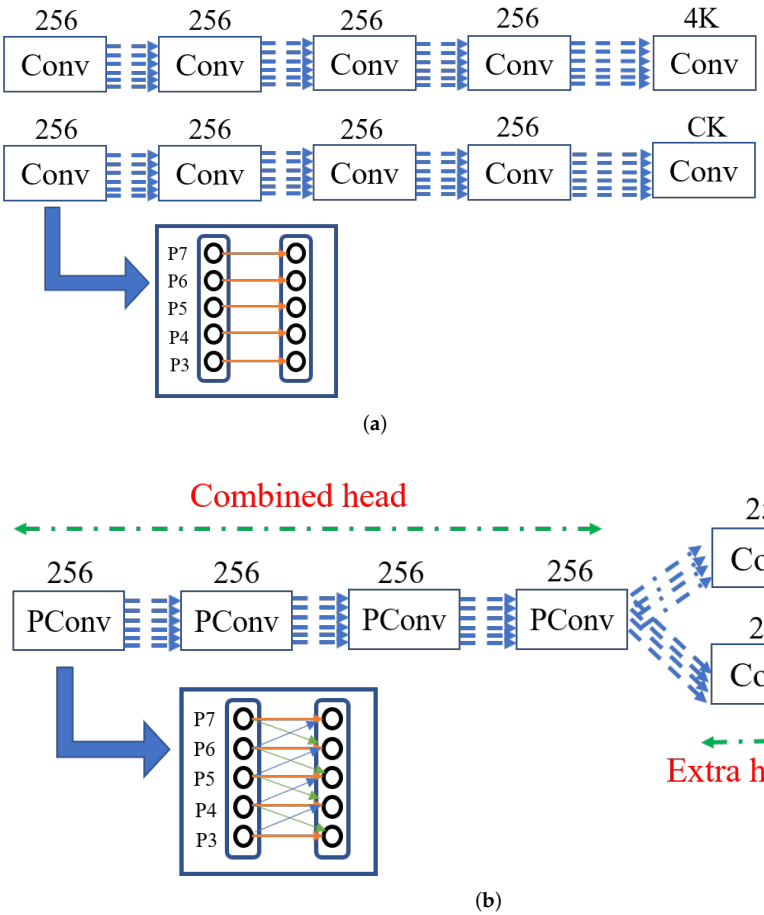


Figure 6. (**a**) Head design of the original RetinaNet; (**b**) Head design with PConv. In the final output convolution, 4 denotes the number of anchor-based methods, *K* is the number of anchor boxes, and *C* represents the class number in classification.

3.4. Label Assignment by Learning-to-Match Strategy

In addition to overcoming the large difference in SAR object scales, there is another problem to be solved in SAR object detection, which is the diverse appearance of SAR targets. SAR targets are objects that are generally slender or irregularly shaped due to interference (e.g., cross sidelobe blurring). In terms of anchor-based detectors, spatial alignment, i.e., the intersection (IoU) between the object and the anchor point, is exploited as the standard for allocating anchor points. In terms of anchor-based detectors, the respective assigned anchor independently supervises network learning to achieve object prediction, which follows the assumption that anchors aligned with the object space constantly apply to classification and positioning. For objects with non-central features (e.g., slender objects), however, the most typical feature is that they are far from their geometric centers. Spatially aligned anchor points are likely to correspond to poorly represented features, thereby reducing classification and localization performance.

To solve the problems above, the learning-to-match (LTM) approach [13] is introduced for label assignment in anchor-based detectors to detect SAR objects that exhibit different appearances. According to the principle of the LTM method, when the object is occluded or the feature is eccentric, the matching between features and objects becomes difficult to measure due to the intersection ratio (IoU) between the prediction box and the true value box. Accordingly, LTM transforms the matching problem between the target and the feature into the maximum likelihood estimate while converting the maximum likelihood probability into a loss function. The matching function is improved to maximize the

possibility of detection customized likelihood and select the optimal anchor point in a "soft" manner. The LTM strategy is elucidated below.

(1) Detector Training as Maximum Likelihood Estimation. For the original single-stage detector, the ground-truth annotations are denoted as B ,where the ground-truth box for the i-th object is denoted as $b_i \in B$. On the convolutional feature maps of X, a set of anchors A are defined as reference points at multiple scales and aspect ratios. After the forward propagation of the network, the respective anchor $a_j \in A$ will be predicted through classification and regression. By complying with the manual design rule of IoU indicator, the respective anchor frame will fall into an object or a background. The matching matrix $C_{ij} \in \{0,1\}$ indicates whether the object b_i is assigned to anchor box a_j. The set of positive anchor boxes $A_+ \subseteq A$ and the set of negative anchor boxes $A_- \subseteq A$ are written as $\{a_j | \sum_i C_{ij} = 1\}$and $\{a_j | \sum_i C_{ij} = 0\}$, respectively. The loss function $\mathcal{L}(\theta)$ of the detector is:

$$\mathcal{L}(\theta) = \sum_{a_j \in A_+} \sum_{b_i \in B} C_{ij} \mathcal{L}_{ij}^{cls}(\theta) + \beta \sum_{a_j \in A_+} \sum_{b_i \in B} C_{ij} \mathcal{L}_{ij}^{loc}(\theta) + \sum_{a_j \in A_-} \mathcal{L}_j^{bg}(\theta) \tag{4}$$

Among them, $L_{ij}^{cls}(\theta)$, $L_{ij}^{loc}(\theta)$, and $L_j^{bg}(\theta)$ represent the classification loss function of the target, the target regression positioning loss function, and the classification loss of the background class, respectively; β expresses a regularization factor; and θ expresses the parameter learned in the network.

According to the maximum likelihood estimation, the original objective loss function $L(\theta)$ can be transformed into likelihood probability:

$$\begin{aligned} \mathcal{P}(\theta) &= e^{-\mathcal{L}(\theta)} \\ &= \prod_{a_j \in A_+} \left(\sum_{b_i \in B} C_{ij} e^{-\mathcal{L}_{ij}^{cls}(\theta)} \right) \prod_{a_j \in A_+} \left(\sum_{b_i \in B} C_{ij} e^{-\beta \mathcal{L}_{ij}^{loc}(\theta)} \right) \prod_{a_j \in A_-} e^{-\mathcal{L}_j^{bg}(\theta)} \\ &= \prod_{a_j \in A_+} \left(\sum_{b_i \in B} C_{ij} \mathcal{P}_{ij}^{cls}(\theta) \right) \prod_{a_j \in A_+} \left(\sum_{b_i \in B} C_{ij} \mathcal{P}_{ij}^{loc}(\theta) \right) \prod_{a_j \in A_-} \mathcal{P}_j^{bg}(\theta) \end{aligned} \tag{5}$$

where $\mathcal{P}_{ij}^{cls}(\theta)$ and $\mathcal{P}_j^{bg}(\theta)$ represent the classification confidence, and $\mathcal{P}_{ij}^{loc}(\theta)$ is the positional confidence. In this way, the problem of minimizing the loss function Equation (4) in target detection can be transformed into the problem of maximizing the likelihood probability function Equation (5). Although the above process strictly takes into account the classification improvement as well as the anchor frame positioning, the anchor box-object matching improvement is not taken into account.

(2) Detection Customized Likelihood. FreeAnchor [13] takes into account the matching of bounding box and features under conventional object detection algorithms. First, select several bounding boxes with larger IoU for the respective object b_j in accordance with the spatial relationship between the anchor box and the object to form a set of bounding boxes $A_j \in A$. To improve the recall rate, for the respective target $b_i \in B$, the predicted value (a_j^{loc} and a_j^{cls}) of at least one candidate anchor $a_j \in A_j$ should be ensured to be close to the real label. On that basis, the likelihood function indicating the recall rate is:

$$\mathcal{P}_{recall}(\theta) = \prod_i \max_{a_j \in A_i} \left(\mathcal{P}_{ij}^{\text{cls}}(\theta) \mathcal{P}_{ij}^{\text{loc}}(\theta) \right) \tag{6}$$

Maximizing the value of $\mathcal{P}_{recall}(\theta)$ means that the respective target should be ensured to have its corresponding matching anchor frame and increase the recall rate of object detection.

To improve the detection accuracy, the detector should classify the poorly positioned anchor as the background class, and its likelihood probability is written below:

$$\mathcal{P}_{precision}(\theta) = \prod_j \left(1 - P\{a_j \in A_-\} \left(1 - \mathcal{P}_j^{bg}(\theta) \right) \right) \tag{7}$$

where $P\{a_j \in A_-\} = 1 - \max_i P\{a_i \to b_j\}$ is the probability that a_j misses all objects and $P\{a_j \to b_i\}$ denotes the probability that anchor a_j predicts object b_i correctly.

To comply with the NMS process, $P\{a_j \to b_i\}$ is implemented by a saturated linear function, which is expressed as:

$$P\{a_j \to b_i\} = Saturated\ linear\left(IoU_{ij}^{loc}, t, max_j IoU_{ij}^{loc}\right) \tag{8}$$

$$Saturated\ linear(x, t_1, t_2) = \begin{cases} 0, x \leqslant t_1 \\ \frac{x-t_1}{t_2-t_1}, t_1 < x < t_2 \\ 1, x \geqslant t_2 \end{cases} \tag{9}$$

In summary, the likelihood probability function of the detector can be redefined as $\acute{\mathcal{P}}(\theta)$, which is shown in Equation (10). In the detector training process, the free matching of the detection anchor frame and the target is realized by maximizing $\mathcal{P}_{recall}(\theta)$ and $\mathcal{P}_{precision}(\theta)$:

$$\begin{aligned} \mathcal{P}'(\theta) &= \mathcal{P}_{recall}(\theta) \times \mathcal{P}_{precision}(\theta) \\ &= \prod_i \max_{a_j \in A_i} \left(\mathcal{P}_{ij}^{cls}(\theta) \mathcal{P}_{ij}^{loc}(\theta)\right) \times \prod_j \left(1 - P\{a_j \in A_-\}\left(1 - \mathcal{P}_j^{bg}(\theta)\right)\right) \end{aligned} \tag{10}$$

(3) Anchor Matching Mechanism. To achieve the fusion of the self-defined likelihood function and the target detection method based on CNN, the likelihood function $\acute{\mathcal{P}}(\theta)$ defined by Equation (10) should be converted back to the required loss function $\acute{\mathcal{L}}(\theta)$:

$$\begin{aligned} \mathcal{L}'(\theta) &= -\log \mathcal{P}'(\theta) \\ &= -\sum_i \log\left(\max_{a_j \in A_i}\left(\mathcal{P}_{ij}^{cls}(\theta) \mathcal{P}_{ij}^{loc}(\theta)\right)\right) - \sum_i \log\left(1 - P\{a_j \in A_-\}\left(1 - \mathcal{P}_j^{bg}(\theta)\right)\right) \end{aligned} \tag{11}$$

The *max* function in Equation (11) is employed to select the most matching anchor for the respective target, whereas the confidence of all anchor is relatively low at the initial training stage. Moreover, since the initialization method of the network parameters is random initialization, the anchor frame with the highest confidence may not be the most matching anchor at this time. Thus, FreeAnchor exploits the $Mean - max$ function to select the anchor box, as defined below:

$$Mean - max(X) = \frac{\sum_{x_j \in X} \frac{x_j}{1-x_j}}{\sum_{x_j \in X} \frac{1}{1-x_j}} \tag{12}$$

We substitute the max function of Equation (11) into $Mean - max$, add balance factor w_1 *and* w_2, and apply focal loss [40] to the second term in Equation (11). On that basis, a FreeAnchor detector's customized loss function is summarized:

$$\mathcal{L}''(\theta) = -w_1 \sum_i \log(Mean - max(X_i)) + w_2 \sum_j FL\left(P\{a_j \in A_-\}\left(1 - \mathcal{P}_j^{bg}(\theta)\right)\right) \tag{13}$$

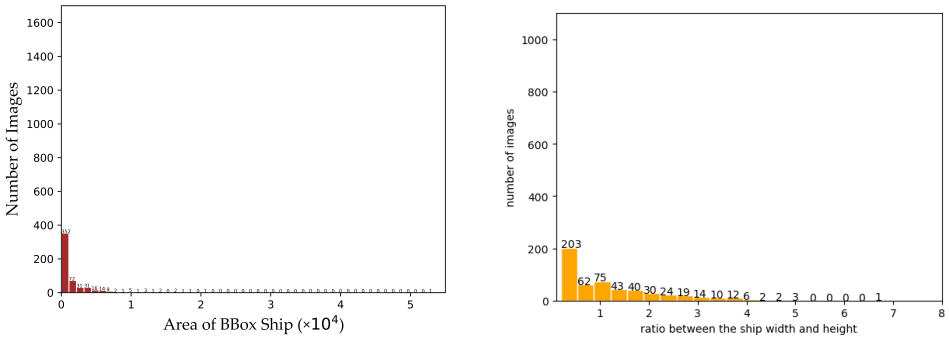
where the weights w_1 and w_2 are expressed as $w_1 = \frac{\alpha_f}{\|B\|}$ and $w_2 = \frac{1-\alpha_f}{\eta_a \|B\|}$, respectively. η_a is the number of anchor boxes in the candidate anchor box set, and $\|B\|$ is the number of objects. The parameters α and τ are from the focal loss, and $FL(x) = -x^\gamma \log(1-x)$. $X_i = \left\{\mathcal{P}_{ij}^{cls}(\theta) \mathcal{P}_{ij}^{loc}(\theta) | \alpha_j \in A_i\right\}$ is a likelihood set conforming to the anchor bag A_i.

4. Experiments

In the present section, the effectiveness of the proposed method is verified, extensive experiments are performed on two SAR object detection data sets (SSDD and HRSID) to compare the proposed method with other cutting-edge detectors (one-stage and two-stage).

4.1. Datasets and Evaluation Metrics

SSDD dataset [41]: SSDD is the first publicly available data set for SAR image ship target detection at home and abroad, which consists of 1160 SAR images with 500 × 500 pixels containing 2358 ships under wide resolutions (1–15 m). The SSDD data set target detection in SAR images poses a great challenge because the SSDD data set not only contains multi-scale ship targets but also contains two complex scenes of the ocean and inshore, as well as a large number of densely distributed small boats. To train the proposed method with the use of SSDD, the regulations of the SSDD data set are followed: The images with the last numbers one and nine of the file number are determined as the test set, and the rest are considered the training set. Figure 7 [41] illustrates the target distribution of the targets in SSDD dataset. Specifically, Figure 7a shows the aspect ratio distribution of the target in the SSDD dataset, which intuitively reflects the large difference in the aspect ratio of the target in the SSDD dataset. Figure 7b reflects the area distribution of the target frame, and the area is an important indicator reflecting the scale change. It can be seen from Figure 7b that the scale of the targets in the SSDD dataset is also very different.

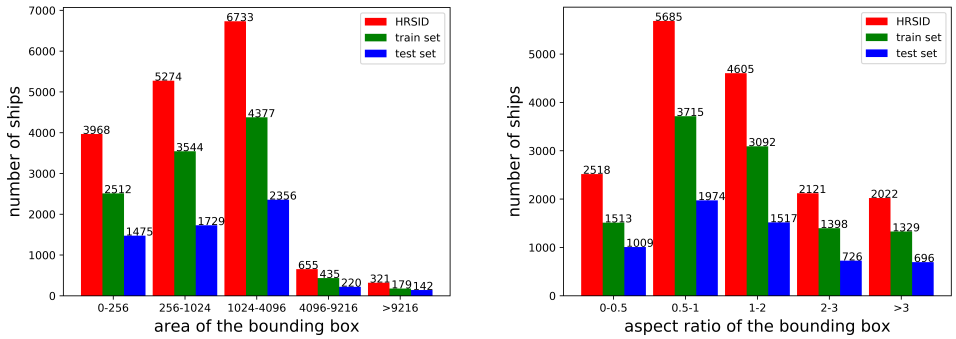


(**a**) (**b**)

Figure 7. Geometry characteristics of targets in SSDD. (**a**) Histogram of the area distribution of ship targets; (**b**) Histogram of the aspect ratio of ship targets.

HRSID dataset [42]: HRSID refers to a large-scale SAR ship detection data set, which covers SAR images under diverse properties (i.e., sea area, polarization, sea condition, resolution and coastal port). Furthermore, it covers ships at multiple scales that are marked with bounding boxes within diverse environments (e.g., mode of polarization, sensor type, as well as scene). According to statistics, there are 5604 cropped SAR images and 16,951 annotated ships in HRSID. Figure 8 [42] visually shows the distribution characteristics of the target scale and aspect ratio in the HRSID data set according to their original study.

Evaluation Metrics: We have taken AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L to characterize the performance of the detectors on the test set of this study. The definition of this evaluation index is identical to the metric standard of MS COCO object detection challenge [43], and it has been extensively employed to assess various object detection tasks.



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

Figure 8. The aspect ratio and area distribution map of the object in HRSID: (**a**) bar area of the bounding box; (**b**) bar aspect ratio of the bounding box.

4.2. Implementation Details

SARFNet is implemented on the single-stage detector RetinaNet [40], with the use of ResNet [44] to be the network of backbone. By replacing the FPN, head design and loss function in RetinaNet with the modules introduced above, we update the RetinaNet detector to the SARFNet detector. In terms of the classification subnet's last convolutional layer, the bias initialization is set to $b = -\log\left(\frac{(1-\rho)}{\rho}\right)$ with $\rho = 0.02$. To maintain the fairness of the experiment, we chose the mmdetection [45] tool for code development. Mmdetection is a flexible toolkit for reimplementing existing methods. It is convenient for us to use State-of-the-Art Methods such as Faster RCNN to complete comparative experiments. After that, to maintain the consistency between the hyperparameter settings and HRSID [42], the SAR image was scaled to 1000×1000 pixels during the training and testing process. All detectors were trained using 1 GPU, and in the 12th Completed in epoch; the momentum and weight decay are set to 0.9 and 0.0001, respectively. When training and testing strictly filter the low-precision bounding box, the IoU threshold is set to 0.7. We choose SGD with an initial learning rate of 0.0025 as the optimizer and set other hyper-parameters to the default values in mmdetection. Experiments are performed under the Pytorch framework on a server with NVIDIA Titan XP.

4.3. Ablation Studies

For verifying the contribution of the proposed components, this study carried out Ablation tests according to the HRSID data set. Table 1 lists the experimentally achieved results. To evaluate the contribution of the respective module, we present several comparisons in Table 1, where the Enhanced Feature Pyramid Network, Scale-Equalizing Pyramid Convolution and learning-to-match strategy correspond to SARFNet. First, this study assesses the contribution of several elements to our baseline recognizer as a reference. According to Table 1, all the techniques contribute to an accuracy gain, and the final baseline acquires an AP score of 64.1%.

(1) Learning-to-Match Strategy. By formulating the detector training as a maximum likelihood estimation (MLE) framework, LTM updates the hand-made anchor point assignments to "free" object-anchor point correspondences. According to the experiment, after adding the LTM strategy, the AP value increases by 1.6 % as compared with the baseline.

(2) Enhanced Feature Pyramid Network. The E-FPN Network a feature alignment module and a feature selection module to the feature pyramid network, which addresses feature misalignment and improves the expression ability of multi-scale features. According to the experiment, replacing FPN with E-FPN will significantly upregulate the AP value by 1.7%.

(3) Scale-Equalizing Pyramid Convlution. SEPC significantly improves the box AP from 63.3% to 64.1%. This validates that the representation of high-resolution features is largely improved based on the proposed adaptive fusion strategy.

Table 1. Ablation study on the components of the scale-aware pyramid network.(LTM: Learning-to-Match; SEPC: Scale-Equalizing Pyramid Convolution; E-FPN: Enhanced Feature Pyramid Network; Y: Yes; N: No).

Baseline	LTM	E-FPN	SEPC	AP
Y	N	N	N	60.0
Y	Y	N	N	61.6
Y	Y	Y	N	63.3
Y	Y	Y	Y	64.1

4.4. Comparison with State-of-the-Art Methods

Results on SSDD. To verify the effectiveness of the proposed algorithm, our proposed method is compared with 16 state-of-the-art methods on the SSDD data set. The backbone applied by these comparison methods and the test results is listed in Table 2. To be specific, HR-SDNet, ISASDNet, FBR-Net and CenterNet++ are methods from well-known journals in wide remote-sensing fields. On the whole, the rest methods are state-of-the-art methods in natural scene object detection. Notably, all the above methods use multi-scale feature information to solve the problem of large-scale differences. FPN-Faster RCNN, ISASDNet, FCOS, FSAF, Free-anchor, FoveaBox, ATSS, AutoAssign and RetainNet pertain to classical feature pyramid network (FPN) methods. The methods of HR-SDNet, FBR-Net, CenterNet++ and Lira RCNN have reduced the defects of the classic feature pyramid network. For instance, FBR-Net and CenterNet++ add a full-scale feature fusion module under FPN to address the problem of imbalanced scale features. AP indicators are capable of reflecting the overall performance of target detection. The AP_{50} and AP_{75} indicators indicate the detection rate of target detection at IoU = 0.5 and IoU = 0.75, respectively. According to Table 2, the proposed method is also superior to the above method in terms of AP_{50} and AP_{75} indicators. Indices AP_S, AP_M and AP_L represent the detection performance of target detection on small, medium and large targets, respectively, and reflect the multi-scale object detection performance of object detection. According to the results in Table 3, whether it is the target detection method by employing the classic FPN structure or the target detection method by exploiting the improved FPN structure, significant differences exist in the performance of these three indicators, which reveals that the method above in scale adaptive ability can be further improved. The proposed method is better than the method above in the performance of these three indicators, and the values of these three indicators are insignificantly different, which reveals that the proposed method exhibits a strong scale adaptive ability.

Results on HRSID. To further verify the effectiveness of the proposed method, we compare the performance of the proposed model with the classic detection model on the HRSID data set. The results of the performance comparison are shown in Table 3. According to Table 3, the proposed method achieves 64.1% AP on the HRSID data set. The results are not only better than anchor-free detectors (e.g., 6.2%, 1.5%, 3.6%, 3.1% and 5.8% higher than FCOS, FASF, FoveaBox, ATSS and AutoAssign, respectively) but are also better than anchor-based detectors (e.g., 0.6%, 0.4%, 4.1% and 2.5% higher than faster than FPN-Faster R-CNN, Lira RCNN, RetinaNet and Free-anchor, respectively). As indicated by the quantitative results, the proposed method exhibits superior performance, in particular for objects at significantly different scales. Accordingly, the proposed scale-aware feature golden character detection network is verified to have effectiveness. Furthermore, from the three indicators (i.e., AP_S, AP_M as well as AP_L), the proposed method exhibits superior

performance, which outperforms the methods above. Thus, the proposed scale-aware feature golden character detection network is verified with effectiveness.

Table 2. Comparison of the performance of different detection models on SSDD.

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Anchor-free methods							
FBR-Net [3]	ResNet50	–	94.1	59.1	–	–	–
CenterNet++[8]	DAL-34	–	92.7	–	–	–	–
SEPN [7]	ResNet101	–	97.2	–	–	–	–
FCOS [46]	ResNet50	64.4	93.6	75.2	65.8	62.0	17.4
	ResNet101	63.6	94.3	73.7	65.0	60.1	44.4
FASF [33]	ResNet50	63.1	94.0	74.3	64.4	59.5	55.1
	ResNet101	63.9	93.5	76.2	65.4	60.3	40.1
FoveaBox [47]	ResNet50	64.6	92.9	76.9	66.4	60.5	33.8
	ResNet101	65.5	94.8	79.3	66.8	64.0	34.5
ATSS [48]	ResNet50	65.0	94.0	76.8	65.8	63.8	39.9
	ResNet101	66.1	94.5	78.8	66.5	66.5	52.7
AutoAssign [49]	ResNet50	64.9	94.9	78.7	66.5	61.3	45.2
	ResNet101	54.3	89.0	62.4	55.9	52.1	48.4
Anchor-based methods							
HRSDNet [42]	HRFPN-W32	61.1	93.9	70.1	56.6	67.7	58.8
	HRFPN-W40	60.9	94.4	69.7	56.2	67.8	58.9
ISASDNet [50]	ResNet50	61.0	95.4	67.7	62.4	60.5	55.2
	ResNet101	62.7	96.8	68.5	63.6	60.3	52.5
FPN Faster RCNN [2]	ResNet50	59.1	93.8	68.6	55.2	66.0	47.3
	ResNet101	58.4	94.0	65.9	54.5	64.7	51.9
Cascade R-CNN [51]	ResNet50	59.7	93.1	67.6	54.8	67.1	57.8
	ResNet101	60.3	94.0	69.6	56.0	66.6	59.3
Mask R_CNN [52]	ResNet50	58.9	93.4	66.6	55.3	64.9	49.7
	ResNet101	59.4	93.9	67.7	54.9	66.2	53.9
Libra RCNN [34]	ResNet50	66.5	95.2	81.2	68.8	63.9	22.6
	ResNet101	66.7	95.9	83.7	69.5	66.6	20.3
RetinaNet [40]	ResNet50	55.5	90.2	62.3	51.2	62.6	45.4
	ResNet101	55.2	90.8	60.2	50.9	62.2	49.7
Free_anchor [13]	ResNet50	64.0	94.1	76.8	65.1	61.2	63.8
	ResNet101	64.6	95.0	77.1	65.7	61.8	49.6
Our	ResNet50	67.3	96.3	80.5	68.4	66.7	48.3
	ResNet101	67.6	96.8	81.2	68.6	66.9	49.5

Table 3. Comparison of the performance of different detection models on HRISD.

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Anchor-free methods							
FCOS [46]	ResNet50	57.9	84.4	64.9	60.2	55.5	16.1
	ResNet101	61.1	86.5	68.5	62.3	61.7	14.8
FASF [33]	ResNet50	62.6	88.3	71.6	64.1	59.4	16.4
	ResNet101	63.0	88.7	72.2	64.3	61.9	12.7
FoveaBox [47]	ResNet50	60.5	83.9	68.8	62.0	59.4	24.5
	ResNet101	58.0	84.8	64.0	58.9	61.1	23.6
ATSS [48]	ResNet50	61.0	84.6	69.0	62.3	63.2	13.6
	ResNet101	58.9	82.5	65.9	59.9	63.6	8.80
AutoAssign [49]	ResNet50	58.3	85.4	64.9	59.9	61.4	22.6
	ResNet101	54.3	89.0	62.4	55.9	52.1	48.4
Anchor-based methods							
HRSDNet [42]	HRFPN-W32	68.6	88.4	79.0	69.6	70.0	25.2
	HRFPN-W40	69.4	89.3	79.8	70.3	71.1	28.9
FPN Faster RCNN [2]	ResNet50	63.5	86.7	73.3	64.4	65.1	16.4
	ResNet101	63.9	86.7	73.6	64.8	66.2	24.2
Cascade R-CNN [51]	ResNet50	66.6	87.7	76.4	67.5	67.7	28.8
	ResNet101	66.8	87.9	76.6	67.5	68.8	27.7
Mask R_CNN [52]	ResNet50	65.0	88.0	75.2	66.1	66.1	17.3
	ResNet101	65.4	88.1	75.7	66.3	68.0	23.2
Libra RCNN [34]	ResNet50	63.8	86.2	73.6	65.0	65.4	17.5
	ResNet101	64.2	86.6	73.3	65.0	67.2	23.8
RetinaNet [40]	ResNet50	60.0	84.7	67.2	60.9	60.9	26.8
	ResNet101	59.8	84.8	67.2	60.4	62.7	26.5
Free_anchor [13]	ResNet50	61.6	86.4	72.6	61.9	62.6	25.4
	ResNet101	62.7	87.3	73.7	62.6	63.1	27.6
Our	ResNet50	64.1	88.2	73.8	62.3	632	27.5
	ResNet101	64.2	88.3	74.1	62.5	63.5	30.4

Ship detection in offshore and inshore scenarios on HRSID dataset. In order to test the performance of the detection algorithm in complex scenarios, we conduct experiments according to the ocean and nearshore scenarios provided by HRSID. The detection results of the two scenarios are shown in Table 4. It can be clearly seen that whether it is state-of-the-art detectors or the method in this paper, the detection index of the ocean scene is higher than that of the nearshore scene.As can be seen from Table 4, the method in this paper is better than one-stage start-art-of methods such as RetinaNet, Free_anchor and ATSS and is close to the performance of two-stage methods such as HRSDNet and Mask R_CNN.

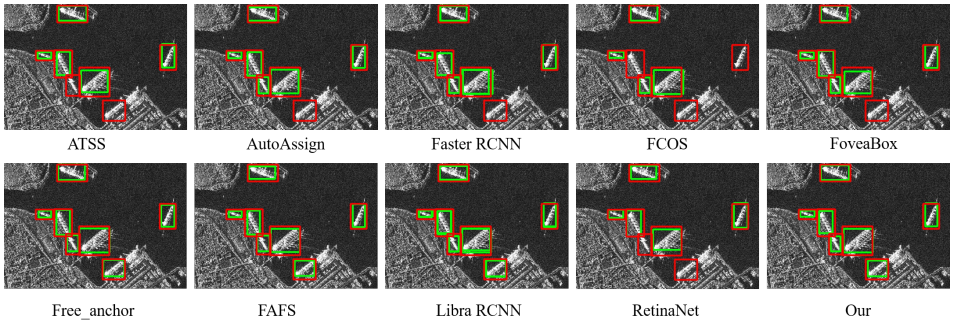
Table 4. Ship detection in the inshore and offshore scenes of HRSID.

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Anchor-free methods							
FCOS [46]	Inshore	38.5	67.2	38.9	39.3	47.1	16.6
	Offshore	75.3	97.7	89.3	77.8	67.9	25.6
FASF [33]	Inshore	43.5	69.2	46.8	43.1	55.2	24.7
	Offshore	79.2	98.6	93.3	80.8	72.9	45.9
FoveaBox [47]	Inshore	48.6	78.0	53.0	48.4	55.4	19.9
	Offshore	79.6	97.7	92.7	81.0	74.0	65.6
ATSS [48]	Inshore	43.6	70.0	46.0	42.4	59.8	18.0
	Offshore	81.2	97.9	93.7	82.8	78.4	42.3
AutoAssign [49]	Inshore	41.4	72.7	41.0	40.0	57.6	28.2
	Offshore	77.9	97.9	91.6	80.1	77.3	55.9
Anchor-based methods							
HRSDNet [42]	Inshore	58.9	81.3	68.3	57.7	72.3	30.1
	Offshore	85.7	98.6	96.0	86.1	82.3	68.2
FPN Faster RCNN [2]	Inshore	51.4	78.3	58.1	50.4	64.0	24.1
	Offshore	80.7	98.0	94.5	82.0	78.2	31.3
Cascade R-CNN [51]	Inshore	55.9	79.6	63.6	54.5	69.6	32.7
	Offshore	83.6	98.0	95.5	84.9	81.1	65.4
Mask R_CNN [52]	Inshore	53.1	79.0	60.7	52.5	63.6	20.0
	Offshore	81.0	98.8	94.6	82.3	79.0	44.9
Libra RCNN [34]	Inshore	50.2	73.8	57.0	49.3	60.9	27.6
	Offshore	81.5	97.9	94.6	82.8	79.3	57.3
RetinaNet [40]	Inshore	41.3	69.0	42.5	39.4	57.9	28.4
	Offshore	79.6	98.6	93.2	81.2	75.1	57.4
Free_anchor [13]	Inshore	45.1	69.9	50.2	43.5	58.8	28.6
	Offshore	79.3	98.6	94.4	80.8	73.5	58.5
Our	Inshore	51.2	76.3	54.4	45.6	60.6	24.8
	Offshore	81.6	98.6	94.7	82.5	76.3	58.8

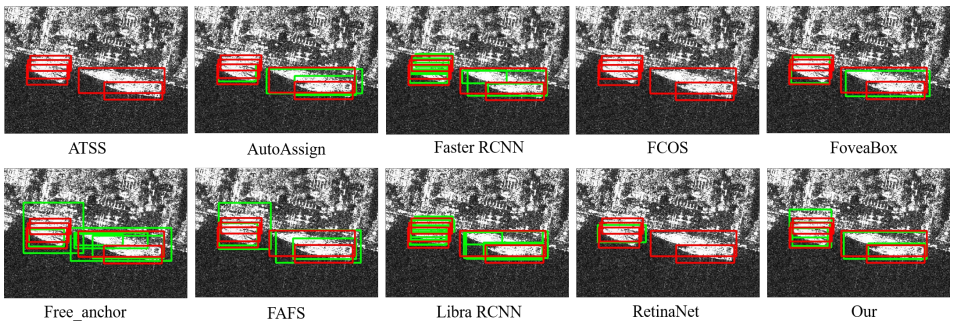
4.5. Discussion

To visually demonstrate the performance of the proposed method, we analyze it from the qualitative perspective. In the qualitative analysis, we visualized the detection results in the SSDD and HRSID data set, and the results are shown in Figures 9–14. Figures 9 and 10 show the visualization results of ours and other state-of-the-art methods. It can be seen intuitively that our method is superior to other methods. The overall visual results of our method are illustrated in Figures 11–14, where Figures 11–14 are the ship detection results on SSDD and HRSID respectively. According to Figures 12 and 14, the proposed method shows better performance under different scale targets and backgrounds. Notably, SARFNet performs well in detecting multi-scale targets and further targets with various appearance changes (e.g., interference from cross side lobes). In addition, SARFNet exhibits better detection performance for densely arranged ship targets. As revealed by the above phenomenon, the proposed method has the ability of scale adaptation and shape adaptation, which benefit from the use of scale adaptive features here and the strategy of learning to match anchor.

However, the proposed method exploits a detection method based on a horizontal rectangular frame. When facing a scene where ships and other targets are rotating and densely packed, the detection method of the horizontal rectangular frame cannot eliminate the interference between the background and neighboring targets, thereby causing missed inspections for the ship and other objects. Accordingly, the subsequent research direction is how to extract rotation invariance information and improve the ability of target detection.

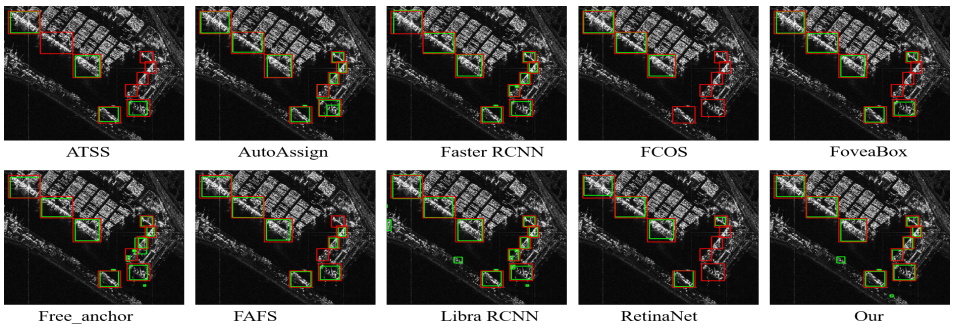


(**a**)



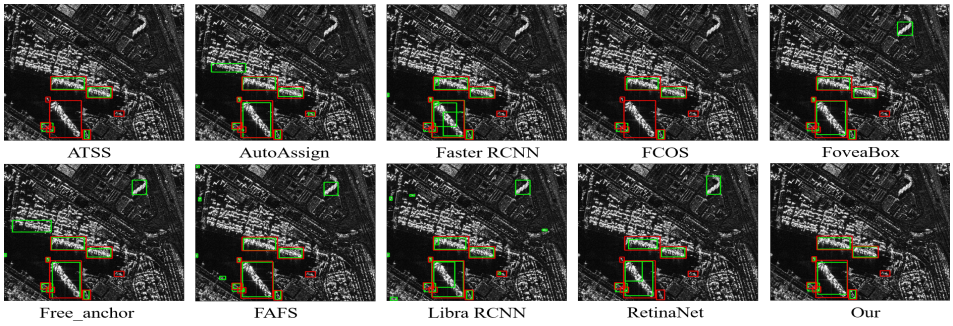
(**b**)

Figure 9. Comparison of the detection results by different methods on SSDD. Red bounding box denotes ground truth and Green bounding box denotes predicted results. (**a**) and (**b**) represent the detection results of multi-scale objects and dense objects, respectively.



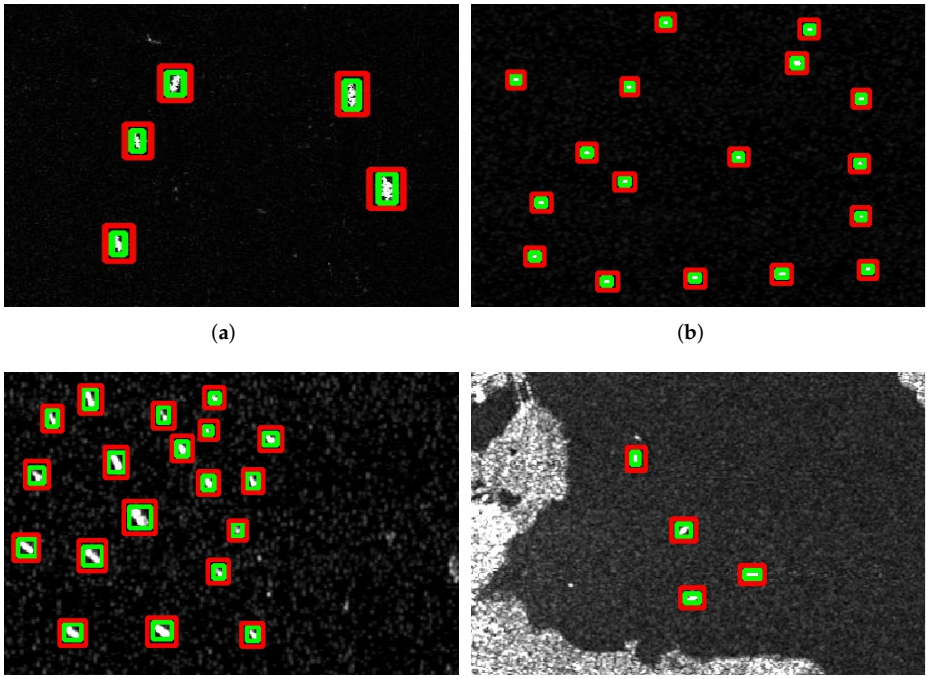
(**a**)

Figure 10. *Cont.*



(**b**)

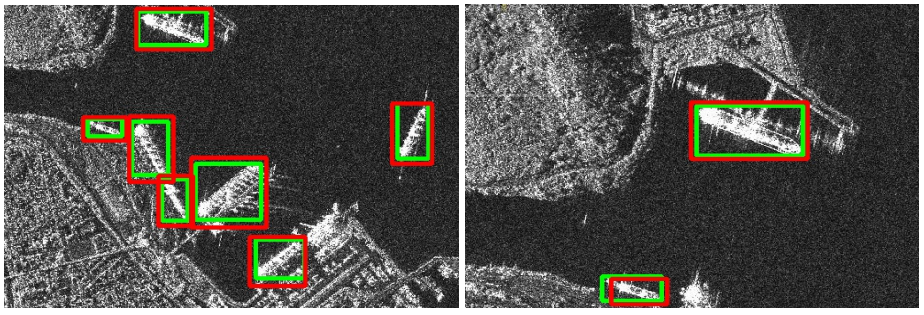
Figure 10. Comparison of the detection results by different methods on HRSID. Red bounding box denotes ground truth and Green bounding box denotes predicted results. (**a**,**b**) represent the detection results of multi-scale ships in complex inshore scenes.



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

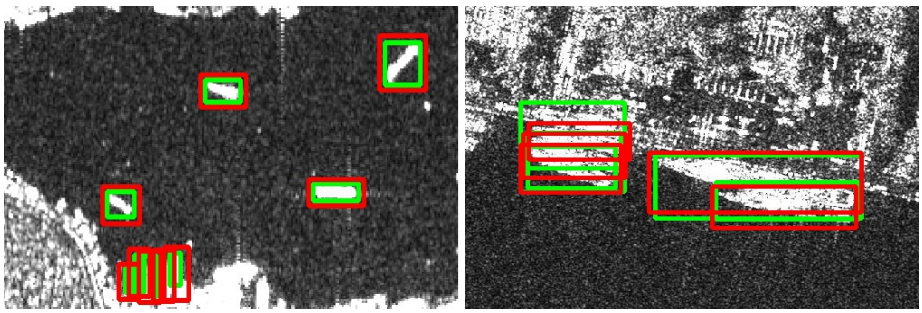
(**c**) (**d**)

Figure 11. The detection resultsof offshore scene on SSDD. The red box represents the ground-truth, and the green box represents the detection result. (**a**,**b**) represent the detection results of ships in clean ocean scenes. (**c**,**d**) represent the detection results of ships in a scenario with sea clutter interference.



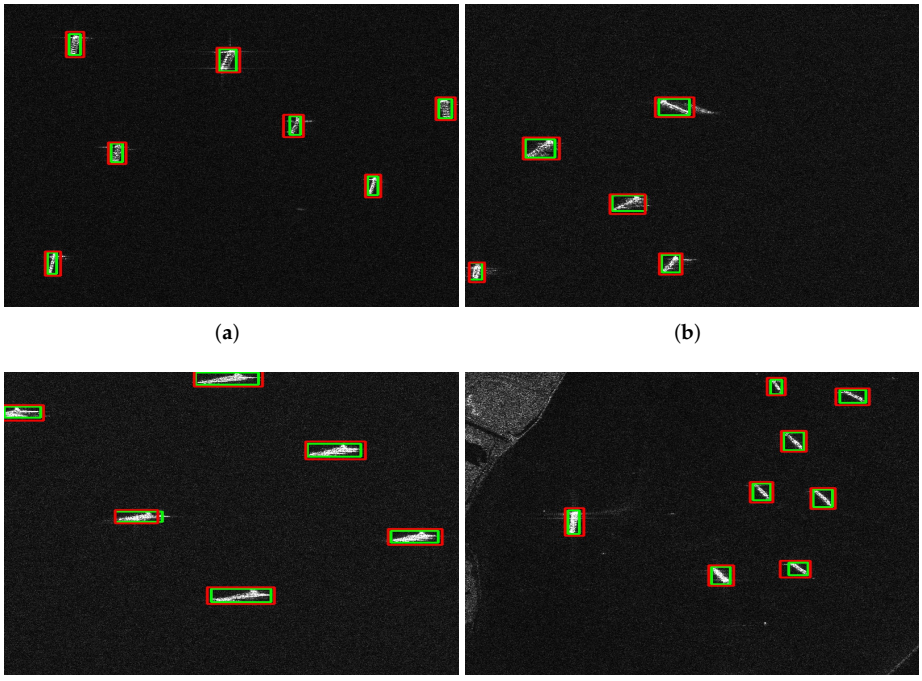
(**a**) (**b**)

Figure 12. *Cont.*



(**c**) (**d**)

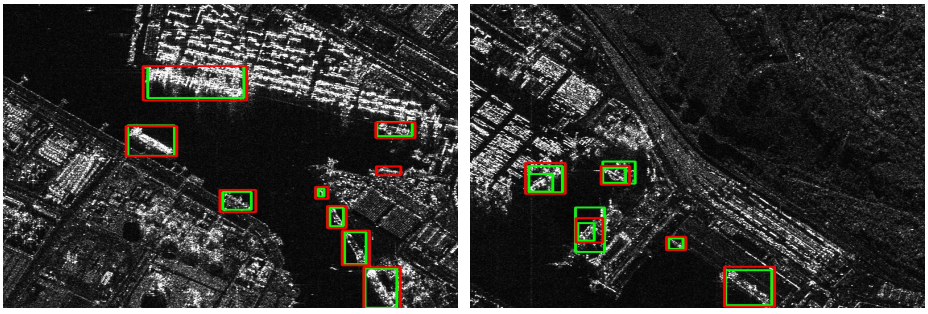
Figure 12. The detection results of inshore scene on SSDD. The red box represents the ground-truth, and the green box represents the detection result. (**a**,**b**) represent the detection results of multi-scale ships in complex inshore scene. (**c**,**d**) represent the detection results of dense ships in complex inshore scene.



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

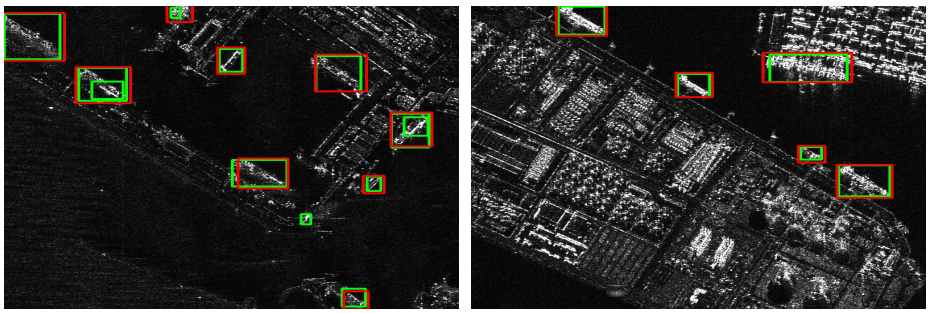
(**c**) (**d**)

Figure 13. The detection results of offshore scene on HRSID. The red box represents the ground-truth, and the green box represents the detection result. (**a**–**d**) represent the ship detection results of our method in typical offshore scenes.



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

Figure 14. *Cont.*



(**c**) (**d**)

Figure 14. The detection results of inshore scene on HRSID. The red box represents the ground-truth, and the green box represents the detection result. (**a**–**d**) represent the ship detection results of our method in typical inshore scenes.

5. Conclusions

This study proposes a novel and effective learning approach for detecting objects in SAR images, called the scale-aware pyramid network (SARFNet), which adaptively selects useful and discriminative features for objects of various scales. Compared with other state-of-the-art methods, the quantitative comparison results on two public data sets for SAR object detection show that the proposed SARFNet approach achieves the highest detection accuracy. In subsequent studies, we hope to introduce a rotation-invariant feature extraction module to the network to adaptively mine rotation-direction features to locate targets more accurately with multiple directions in SAR images.

Author Contributions: Funding acquisition, Y.H. and B.Z.; methodology, L.T. and W.T.; supervision, B.Z. and L.T.; validation, X.Q. and W.W.; writing—original draft, W.T. and Y.H. All authors have read and agreed to the published version of this manuscript.

Funding: This study receives support from China Postdoctoral Science Foundation under Grant 2021TQ0177 and the Shandong Natural Science Foundation under Grant ZR2021MF021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were used in this study. SSDD datasets can be found here: https://github.com/TianwenZhang0825/Official-SSDD. HRSID datasets can be found here: https://github.com/chaozhong2010/HRSID.

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

References

1. Gao, G.; Kuang, G.; Zhang, Q.; Li, D. Fast detecting and locating groups of targets in high-resolution SAR images. *Pattern Recognit.* **2007**, *40*, 1378–1384. [CrossRef]
2. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017.
3. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 1331–1344. [CrossRef]
4. 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]
5. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A Novel Quad Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
6. Zhao, D.; Zhu, C.; Qi, J.; Qi, X.; Su, Z.; Shi, Z. Synergistic Attention for Ship Instance Segmentation in SAR Images. *Remote Sens.* **2021**, *13*, 4384. [CrossRef]
7. Zhou, Z.; Guan, R.; Cui, Z.; Cao, Z.; Pi, Y.; Yang, J. Scale Expansion Pyramid Network for Cross-Scale Object Detection in Sar Images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 5291–5294. [CrossRef]

8. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]
9. Cui, Z.; Wang, X.; Liu, N.; Cao, Z.; Yang, J. Ship Detection in Large-Scale SAR Images Via Spatial Shuffle-Group Enhance Attention. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 379–391. [CrossRef]
10. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
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. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
13. Zhang, X.; Wan, F.; Liu, C.; Ji, X.; Ye, Q. Learning to Match Anchors for Visual Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
14. An, W.; Xie, C.; Yuan, X. An Improved Iterative Censoring Scheme for CFAR Ship Detection With SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4585–4595.
15. Li, T.; Liu, Z.; Xie, R.; Ran, L. An Improved Superpixel-Level CFAR Detection Method for Ship Targets in High-Resolution SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 184–194. [CrossRef]
16. Hui, D.; Lan, D.; Yan, W.; Wang, Z. A Modified CFAR Algorithm Based on Object Proposals for Ship Target Detection in SAR Images. *IEEE Geoence Remote Sens. Lett.* **2016**, *13*, 1925–1929.
17. Zhai, L.; Li, Y.; Su, Y. Inshore Ship Detection via Saliency and Context Information in High-Resolution SAR Images. *IEEE Geoence Remote Sens. Lett.* **2016**, *13*, 1870–1874. [CrossRef]
18. 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.* **2020**, *58*, 3366–3376. [CrossRef]
19. 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]
20. 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]
21. Li, X.; Du, Z.; Huang, Y.; Tan, Z. A deep translation (GAN) based change detection network for optical and SAR remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *179*, 14–34. [CrossRef]
22. Mukherjee, S.; Zimmer, A.; Kottayil, N.K.; Sun, X.; Ghuman, P.; Cheng, I. CNN-Based InSAR Denoising and Coherence Metric. In Proceedings of the 2018 IEEE SENSORS, New Delhi, India, 28–31 October 2018; pp. 1–4. [CrossRef]
23. Shin, S.; Kim, Y.; Hwang, I.; Kim, J.; Kim, S. Coupling Denoising to Detection for SAR Imagery. *Appl. Sci.* **2021**, *11*, 5569. [CrossRef]
24. Singh, B.; Davis, L.S. An Analysis of Scale Invariance in Object Detection-SNIP. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3578–3587. [CrossRef]
25. Singh, B.; Najibi, M.; Davis, L.S. SNIPER: Efficient Multi-Scale Training. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Montreal, QC, Canada, 3–8 December 2018; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 9333–9343.
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 21–37.
27. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
28. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 354–370.
29. Ni, F.; Yao, Y. Multi-Task Learning via Scale Aware Feature Pyramid Networks and Effective Joint Head. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 4265–4272. [CrossRef]
30. 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, 18–23 June 2018; pp. 8759–8768.
31. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
32. Zhao, B.; Zhao, B.; Tang, L.; Han, Y.; Wang, W. Deep Spatial-Temporal Joint Feature Representation for Video Object Detection. *Sensors* **2018**, *18*, 774. [CrossRef]
33. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.
34. 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; pp. 821–830.
35. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *Int. Conf. Mach. Learn.* **2019**, *97*, 6105–6114.
36. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
37. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6569–6578.

38. Wang, X.; Zhang, S.; Yu, Z.; Feng, L.; Zhang, W. Scale-Equalizing Pyramid Convolution for Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
39. Huang, S.; Lu, Z.; Cheng, R.; He, C. FaPN: Feature-aligned Pyramid Network for Dense Image Prediction. In Proceedings of the International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
40. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
41. 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]
42. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
43. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*; Springer International Publishing: Berlin, Germnay, 2014.
44. 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.
45. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
46. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9627–9636.
47. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **2020**, *29*, 7389–7398. [CrossRef]
48. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
49. Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; Sun, J. AutoAssign: Differentiable Label Assignment for Dense Object Detection. *arXiv* **2020**, arXiv:2007.03496.
50. Wu, Z.; Hou, B.; Ren, B.; Ren, Z.; Jiao, L. A Deep Detection Network Based on Interaction of Instance Segmentation and Object Detection for SAR Images. *Remote Sens.* **2021**, *13*, 2582. [CrossRef]
51. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498. [CrossRef] [PubMed]
52. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.