remote sensing

MDPI

Article

ATSD: Anchor-Free Two-Stage Ship Detection Based on Feature Enhancement in SAR Images

Canming Yao †, Pengfei Xie †, Lei Zhang * and Yuyuan Fang

School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China

* Correspondence: zhanglei57@mail.sysu.edu.cn

† These authors contributed equally to this work.

Abstract: Syntheticap erture radar (SAR) ship detection in harbors is challenging due to the similar backscattering of ship targets to surrounding background interference. Prevalent two-stage ship detectors usually use an anchor-based region proposal network (RPN) to search for the possible regions of interest on the whole image. However, most pre-defined anchor boxes are redundantly and randomly tiled on the image, manifested as low-quality object proposals. To address these issues, this paper proposes a novel detection method combined with two feature enhancement modules to improve ship detection capability. First, we propose a flexible anchor-free detector (AFD) to generate fewer but higher-quality proposals around the object centers in a keypoint prediction manner, which completely avoids the complicated computation in RPN, such as calculating overlapping related to anchor boxes. Second, we leverage the proposed spatial insertion attention (SIA) module to enhance the feature discrimination between ship targets and background interference. It accordingly encourages the detector to pay attention to the localization accuracy of ship targets. Third, a novel weighted cascade feature fusion (WCFF) module is proposed to adaptively aggregate multi-scale semantic features and thus help the detector boost the detection performance of multi-scale ships in complex scenes. Finally, combining the newly-designed AFD and SIA/WCFF modules, we present a new detector, named anchor-free two-stage ship detector (ATSD), for SAR ship detection under complex background interference. Extensive experiments on two public datasets, i.e., SSDD and HRSID, verify that our ATSD delivers state-of-the-art detection performance over conventional detectors.

Keywords: anchor-free two-stage detector; spatial insertion attention; weighted cascade feature fusion; ship detection; synthetic aperture radar (SAR)

Citation: Yao, C.; Xie, P.; Zhang, L.; Fang, Y. ATSD: Anchor-Free Two-Stage Ship Detection Based on Feature Enhancement in SAR Images. *Remote Sens.* **2022**, *14*, 6058. https://doi.org/10.3390/rs14236058

Academic Editor: Alin Achim

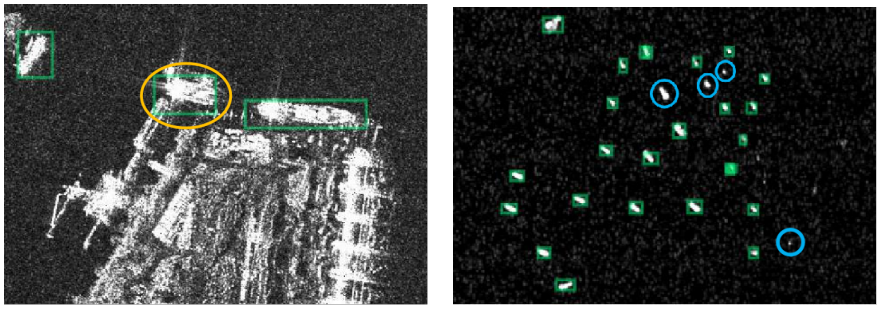
Received: 15 October 2022
Accepted: 24 November 2022
Published: 29 November 2022

1. Introduction

Synthetic aperture radar (SAR) ship detection plays a promising role in many fields such as port management [1–3], traffic monitoring [4–6], marine surveillance [7–9], etc. [10,11], and increasingly becomes an important means for safeguarding maritime rights and interests. However, considering practical situations for SAR ship detection, significant challenges often arise from ship targets surrounded by complex background interference. For example, non-target facilities in harbors often show backscattering characteristics close to that of ship targets, manifested as similar appearances (e.g., pixel intensity) in SAR images, making it challenging to accurately identify ships, even in high-resolution imagery. Many traditional methods have been proposed for SAR ship detection, such as the commonly used constant false alarm rate (CFAR) method and its variants [12–15]. Nevertheless, CFAR-based detectors often perform unsatisfactorily in the above-mentioned strong interference circumstances [13,14], because they only use low-order statistics to distinguish a target from its background interference, and therefore deem to fail if the target and the interference share similar backscattering characteristics.

In recent years, many anchor-based deep-learning methods have been proposed for SAR ship detection [16–18] to overcome the fragility of traditional methods under strong

interference. In general, these methods can be divided into (*i*) two-stage methods, like region-based CNN (R-CNN) algorithms with region proposal networks (RPNs) [19–22], and (*ii*) one-stage methods, like you only look once (YOLO) series [23–25]. Two-stage methods often have more accurate results thanks to their multiple anchor refinements, while one-stage methods usually have higher computational efficiency. It is worth noting that most existing anchor-based detectors usually suffer from the following two drawbacks in the scenario of SAR ship detection: (1) Most of the pre-defined anchor boxes are redundantly and randomly tiled on the image [19], manifested as low-quality generated proposals, likely leading to inaccurate recognition of ship targets as shown in Figure 1a. (2) A series of anchor-box hyper-parameters need to be carefully tuned [24,26,27]. Note that insufficiently tuned hyper-parameters usually affect the detection performance in the case of scale-rich ships [17], particularly for small ships, as the intersection over union (IoU) calculations are more sensitive to small bounding boxes.



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

Figure 1. Some undesired detection results under typical scenes from RetinaNet [26]. The spring-green rectangles represent the detection results. The orange and blue circles represent the false alarms and the missing ships, respectively. (**a**) False alarms in complex scenes. (**b**) The missing small ships.

Considering the above deficiencies of anchor-based methods, anchor-free deep-learning methods have become popular for SAR ship detection [28,29], as they eliminate the requirement for anchor boxes and avoid tuning the corresponding hyper-parameters. Most anchor-free methods follow the point-prediction fashion: First, locate a pair of keypoints [30,31] or center points [32,33] of objects, and then regress them to final bounding boxes. With the point-prediction fashion, there are several advantages for SAR ship detection: (1) It is friendly to detect small ships as they usually have little semantic information for localization in the high-level features. (2) Ships densely arranged near the shore can be well detected due to the simplification of non-maximum suppression (NMS) [28,34]. (3) Leveraging points around target centers can effectively suppress low-quality predictions that contain much background interference. More importantly, anchor-free methods obtain performance on par with anchor-based methods due to the applications of feature pyramid networks (FPNs) [35] and attention mechanisms [36]. FPNs are usually used to boost the detection performance of multi-scale targets, and attention mechanisms encourage paying increasing attention to salient features. However, despite their great success, some deficiencies need further investigation.

1. Most existing attention mechanisms squeeze spatial features into a single vector via global pooling [36,37], neglecting the positional information among spatial-wise levels [38]. Whereas in harbor scenarios with background interference appearing similar to target ships, the positional features are believed essential for locating ships.
2. Small ships are sensitive to background interference in low-level features due to their few pixels, while they are also semantically weak when mapped to high-level features. This imbalance across different levels makes the detector focus more on compelling larger ships than small ones [29]. However, conventional FPN-based methods [35,39] typically assume equal contributions between different feature levels

and ignore adaptive feature fusion, leading to small ships often missing detection, as shown in Figure 1b.

In this paper, we design a novel two-stage detection method to deal with the above-mentioned drawbacks of anchor-based and anchor-free methods under strong background interference in SAR images. Firstly, we present a new anchor-free detector (AFD) to generate reliable proposals in the first stage in a keypoint prediction manner. The AFD can generate fewer yet higher quality proposals around the target centers, significantly simplifying the fine regressions of the second stage. Secondly, to deal with the deficiency 1, a novel spatial insertion attention (SIA) module is proposed to enhance the feature discrimination between ship targets and background interference. It focuses on extracting positional information among two spatial directions to enhance the representation of the features while retaining the positional information. Thirdly, to address the deficiency 2, we leverage the proposed weighted cascade feature fusion (WCFF) module to adaptively fuses pyramidal features and utilizes the non-local network [40] to capture wide-range receptive fields, effectively boosting multi-scale ship detection performance in complex scenes. Finally, leveraging the newly developed AFD and SIA/WCFF modules, we propose a novel anchor-free two-stage ship detector (ATSD) for SAR ship detection under complex background interference. In summary, our ATSD is capable of inheriting both the simplicity promising of AFD and the effectiveness of two feature enhancement modules in the first stage, so as to facilitate the high-precision performance of further refinement in the second stage.

Our main contributions are listed below:

1. A flexible anchor-free detector (AFD) is proposed to generate fewer but higher quality proposals around the object centers than the RPN, which completely avoids the complex manual anchor-box settings as in RPN, outperforms the RPN and achieves a better speed-accuracy trade-off.
2. A novel spatial insertion attention (SIA) module is proposed to help the detector concentrate on the localization accuracy of ship targets by capturing valid position information in complex scenes.
3. An improved weighted cascade feature fusion (WCFF) module is proposed to ameliorate the imbalance across different feature levels on detecting ships by adaptively aggregating multi-scale semantic features.
4. The effectiveness of the proposed ATSD is empirically verified on two public datasets, i.e., SSDD [41] and HRSID [42], where our ATSD surpasses conventional CNN-based detectors by a large margin. A series of ablation experiments and qualitative analyses are conducted to demonstrate the effectiveness of each component in the ATSD.

The rest of this article is organized as follows. Section 2 introduces related work. Section 3 presents our ATSD in detail. The experimental results and analyses are provided in Section 4. Finally, conclusions are given in Section 5.

2. Related Work

In this section, we briefly review conventional CNN-based SAR ship detectors, attention mechanism algorithms, and multi-scale feature fusion methods.

2.1. CNN-Based SAR Ship Detectors

Current CNN-based SAR Ship detection consists of anchor-based and anchor-free detectors. The former can be divided into two-stage and one-stage methods, while the latter follows the point-prediction pipeline.

Anchor-based Detector. Faster R-CNN [19] establishes the predominant position of the two-stage framework in SAR ship detection. They usually use an anchor-based RPN to find numerous regions of interest (RoIs) in the first stage and then refine the filtered anchors in the second stage. Based on the Faster R-CNN, many modified ship detection techniques are proposed to boost its performance, including RoI feature normalization [41], multi-scale feature fusion [17,43], and feature enhancement and enrichment [16,29]. With the advent of single shot multi-box detector (SSD) [44], one-stage detectors have attracted wide

attention in the field of SAR ship detection because of their simple structures. They often preset the anchor boxes on the feature maps and directly infer the object category and box offsets. After that, many works are presented for SAR ship detection based on different architectures, including YOLO [24,45] and RetinaNet [46]. In general, one-stage methods can achieve performance on par with two-stage methods at a much higher efficiency.

Anchor-free Detector. Most anchor-free detectors follow the point-prediction pipeline that is different from anchor-based detectors. They regard objects as keypoints [30,47] or center-points [32,33] and then predict bounding boxes to detect objects. These methods are able to eliminate those hyper-parameters associated with anchor boxes and therefore offer higher efficiency and are becoming the mainstream of SAR ship detection. For instance, Cui et al. [28] and Guo et al. [34] add network modules based on CenterNet [31] to improve ship detection performance; Fu et al. [29] introduce an attention-guided balanced pyramid module to detect ships based on the FCOS [32] framework.

Some recent studies [32,48] show the limitations of the low-quality proposals generated by the RPN in the two-stage detectors, e.g., most proposals are redundant and will entail a heavy computational load. To address this issue, FCOS has proved that replacing the RPN can yield a higher recall rate; Zhou et al. [48] propose to use a variety of one-stage detectors [26,31,49,50] to build probabilistic detectors and demonstrate their feasibility. In this paper, we focus on detecting SAR ships under complex interference with higher efficiency. To this end, we explore building a flexible anchor-free detector (AFD), which generates reliable proposals that are tightly surrounded the object centers with reduced redundancy. Empirically, our AFD delivers fewer (128 vs. 1K) proposals with higher quality than the RPN, achieving higher accuracy and faster inference speed.

2.2. Attention Mechanism Algorithms

In harbor situations, the backscattering characteristics of ships are often similar to those of nearby ports and reefs, making it challenging for a detector to detect ships accurately. Attention mechanisms can be good solutions to the issue, because (*i*) they enable extracting valid target features with strengthened feature representative [18,28], and (*ii*) they enable effective interaction of features between multiple dimensions [37,38,51,52]. One classic example is the SENet [36], which proposes a squeeze and excite (SE) module to recalibrate channel-wise features by explicitly modeling the inter-dependencies between feature channels. Inspired by this principle, the SAR ship detector proposed by [18] utilizes SENet to improve detection performance. Recently, the spatial shuffle-group enhance (SSE) module [28] is proposed to strengthen the relationship between channels for better feature extraction, and Bai et al. [53] design a discrete wavelet multi-scale attention method to help the detector focus on the object area. Moreover, The CBAM [37] leverages the channel and spatial dimensions for adaptive feature refinement. However, these approaches directly compress spatial features into a single vector via a global average pooling, lacking careful attention to spatial-wise features. To deal with it, the scSE [54] achieves the spatial-wise attention via a 1×1 convolution operation, but it is independent of the channel-wise attention, neglecting feature interaction between these two directions. CoordAtt [38] and CoAM [55] utilize two spatial dimensions for feature calibration while retaining location information. In contrast to directly separating features as they [38,55] do, the SIA module proposed in this paper inserts the positional information into the channel-wise attention through the calibration feature (see Section 3). The SIA module is readily applied to any backbone network in a plug-and-play manner.

2.3. Multi-Scale Feature Fusion Methods

One of the main challenges in SAR ship detection is to detect small ships surrounded by strong backscattering interference. Recently, multi-scale feature fusion has proved valuable in detecting small ships [29,34]. Conventional FPN [35] fuses multi-scale features by a top-down pathway, but it is inherently limited by the simple one-way information flow, making it difficult to fuse valid features. To deal with that, the path aggregation network (PANet) [39] adds an extra bottom-up pathway to enhance feature representation

on top of FPN. Nowadays, NAS-FPN [56] leverages neural architecture search (NAS) to design feature network topology automatically. Although NAS-FPN achieves improved performance, its network is irregular and hard to interpret or modify. EfficientDet [57] thus proposes a weighted bi-directional FPN to efficiently fuse multi-scale features in an intuitive and principled manner. Shamsolmoali et al. [58] propose a multiple patch FPN that leverages cross-scale connections for producing multi-scale representative features. Generally, most existing FPN-based SAR ship detection methods are simple connections between two features at the same level without considering the contributions of different features [17,34]. However, it is revealed in [57] that the semantic information from different input features contributes unequally to the output-fused features. Therefore, we propose a WCFF module to adaptively fuse features from each branch, aiming at an automatically multi-scale feature fusion optimization. More details are given in Section 3.

3. Methodology

The overall framework of our ATSD is shown in Figure 2. Two feature enhancement modules are involved in the first-stage AFD, that is, the SIA embedded in the backbone network and the WCFF embedded in the neck block. In the following, we first introduce the ATSD framework and its loss function. Next, we describe in detail the two feature enhancement modules, i.e., the SIA and WCFF.

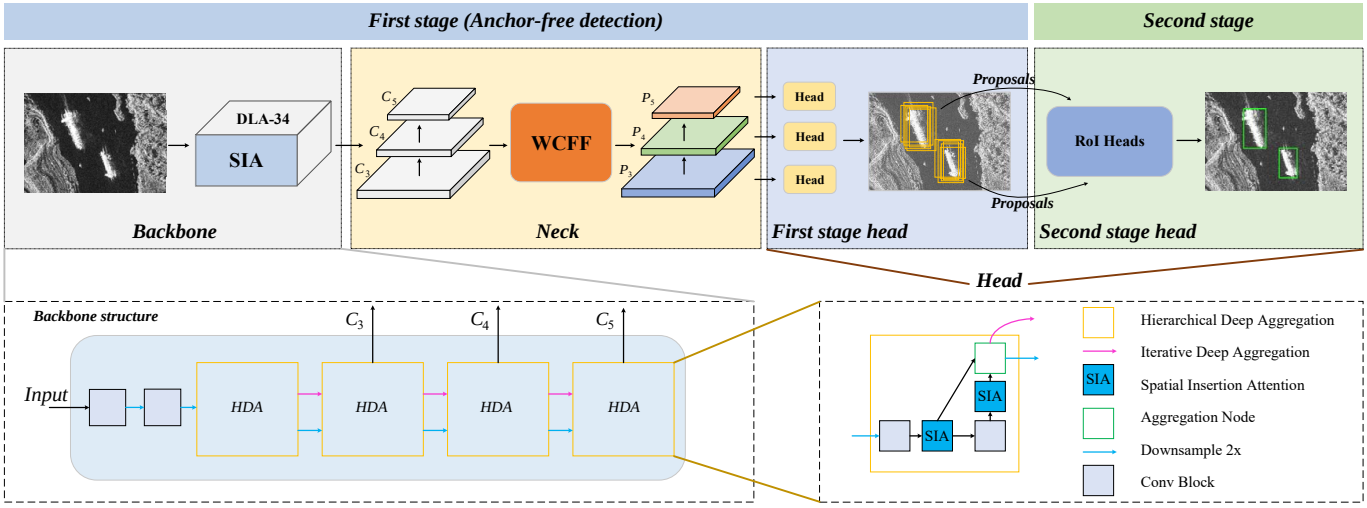


Figure 2. The overall framework of ATSD follows the usual object detector design manner: *"Backbone-Neck-Head"*. The SIA module embedded in the backbone network and the WCFF module embedded in the neck block are introduced in the first stage, and the head network spans the first and second stages.

3.1. Anchor-Free Two-Stage Ship Detector

We design an improved anchor-free detector (AFD) to generate proposals in the first stage, whose categories and bounding boxes are refined several times by cascaded RoI heads in the second stage. The integration of these two stages constitutes the proposed ATSD.

In the first stage, our AFD takes a SAR image $I \in \mathbb{R}^{H \times W}$ with height H and width W as input and then processes the image through the "Backbone-Neck-First stage head". In detail, as shown in Figure 2, we first extract features from an modified backbone DLA-34 [59] embedded by the SIA module, and then employ the last three outputs of hierarchical deep aggregation (HDA) from stride $\{8, 16, 32\}$ as bottom-up pathway input pyramids. We define them correspondingly as $\{C_3, C_4, C_5\}$, where C_l represents the feature with resolution $1/s_l = 1/2^l$ to the input image. The output pyramids are defined as $\{P_3, P_4, P_5\}$, where $P_l \in \mathbb{R}^{(H/s_l) \times (W/s_l) \times D}$ and D is the channel dimension of P_l. We empirically set D as 256 in our model. The output pyramids are then sent to the first stage head to predicts a set of rectangular object proposals, each with an objectness score (measuring the probability of an object being a target or background). We model this prediction process by attaching two branches in parallel (i.e., heads) to different pyramid feature layers, an example is

illustrated in Figure 3. Specifically, both branches stack five 3 × 3 convolutional layers to learn task-specific features for object classification and regression, respectively. Let $F_l \in [0,1]^{H/s_l \times W/s_l \times 1}$ with $l = 3,4,5$ be the classification map at layer l from the pyramid feature P_l. For each pixel point $p^l_{x,y}$ at location (x,y) on F_l, it can be mapped back onto the input image location (x',y'), where $x' = s_l(x+0.5)$ and $y' = s_l(y+0.5)$. A prediction $p^l_{x,y}$ = 1 corresponds to a detected keypoint, while $p^l_{x,y}$ = 0 represents background. Different from the CenterNet [31] directly regressing the object size (i.e., w,h), our AFD employs a 4D vector $\hat{\mathbf{t}} = (l^*, t^*, r^*, b^*)$ as the regression targets for the location, where $\hat{\mathbf{t}}$ depicts the relative distances from the location to the four sides of the bounding box, as shown in Figure 4. Formally, if location $p^l_{x,y}$ is associated to a ground-truth bounding box $\mathbf{g} = (x_0, y_0, x_1, y_1)$, its training regression targets can be formulated as:

$$\begin{aligned} l^* &= x' - x_0, t^* = y' - y_0, \\ r^* &= x_1 - x', b^* = y_1 - y'. \end{aligned} \tag{1}$$

Unlike the anchor-based detectors assigning anchors with different sizes in different pyramid levels, our AFD directly limits the range of bounding box regression for each level. Specifically, $[m_{l-1}, m_l]$ is the size range that level l is responsible for regression. If a regression target $\hat{\mathbf{t}}$ on any feature level is out of the range, then the corresponding location is regarded as a negative sample and is not required to regress a bounding box anymore. In this paper, m_2, m_3, m_4 and m_5 are set as 0, 64, 192 and ∞, respectively. Leveraging the keypoint prediction, most of the generated proposals are around the ship centers, which significantly reduces the redundant proposals when compared to the RPN; accordingly, we can feed much fewer proposals (128 vs. 1K) to the second stage with higher quality than the RPN.

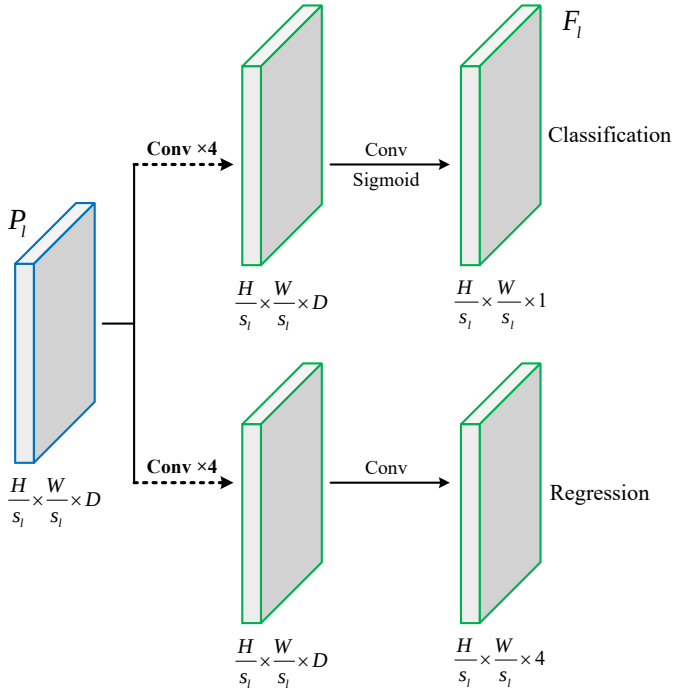


Figure 3. Structure of the AFD head. Both branches stack five 3 × 3 convolutional layers to learn task-specific features for object classification and regression, respectively.

In the second stage, we consider employing the single RoI head (SH) [19] and the cascaded RoI heads (CH) [21] to progressively refine the bounding boxes of ship proposals with different sizes. The main difference is that the former has only one head trained with a fixed IoU threshold (e.g., 0.5), while the latter has multiple sequentially connected heads trained with increasing IoU thresholds of $\{0.5, 0.6, 0.7\}$. The "refinement" idea of

CH is illustrated in Figure 5, representing that the output of a head trained with a certain IoU threshold is a good distribution to train the head of the next higher IoU threshold. Empirically, our ATSD with CH performs much better than SH, and therefore we employ CH in the second stage for our ATSD by default. In general, our ATSD inherits the advantage of the high accuracy of the two-stage detector while maintaining the efficiency of the anchor-free detector.

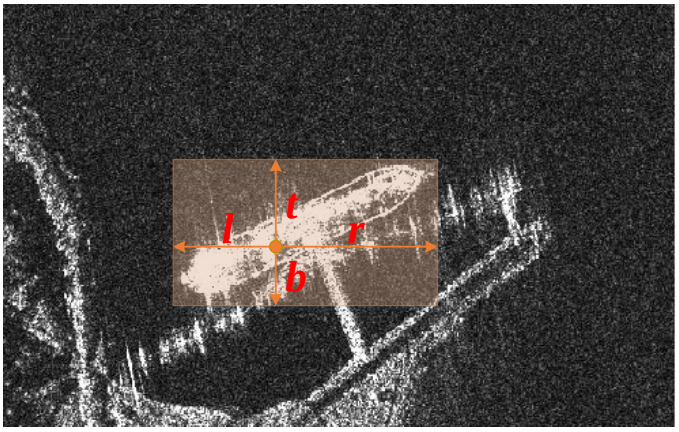


Figure 4. A 4D vector (l, t, r, b) encodes the bounding box. (l, t, r, b) depicts the relative distances from the object center to the four sides of the bounding box.

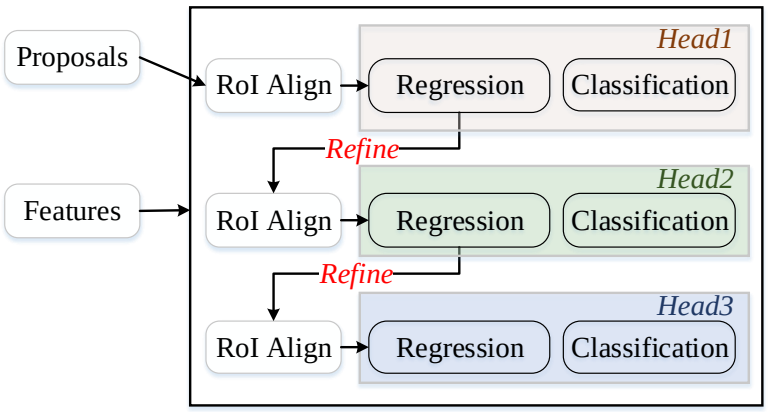


Figure 5. Cascaded RoI heads. Three heads (*Head 1, 2, 3*) are trained with IoU thresholds of {0.5, 0.6, 0.7}, respectively.

3.2. Loss Function

The total loss can be decomposed into the first stage loss and the second stage loss. As for the first stage, we perform binary classification in the form of keypoint prediction following [31]. The corresponding ground truth keypoint heatmap $G_l \in [0,1]^{H/s_l \times W/s_l \times 1}$ is established by a Gaussian kernel:

$$G_l = \exp\left(-\frac{(x-\tilde{z}_x)^2 + (y-\tilde{z}_y)^2}{2\sigma_z^2}\right) \tag{2}$$

where $\tilde{z} = \lfloor z/s_l \rfloor$ represents a low-resolution equivalent to the ground truth keypoint z and σ_z is an object size-adaptive standard deviation. We define the first stage loss function $\mathcal{L}_1$ as follows:

$$\mathcal{L}_1(p^l_{x,y}, \hat{\mathbf{t}}^l_{x,y}) = \frac{1}{N}\sum_l \sum_{x,y} \{\mathcal{L}_{obj}(p^l_{x,y}, g^l_{x,y}) + [g^l_{x,y} = 1]\mathcal{L}_{reg}(\mathbf{t}^l_{x,y}, \hat{\mathbf{t}}^l_{x,y})\} \tag{3}$$

where $g^l_{x,y}$ is the ground truth keypoint on G_l and $\mathbf{t}^l_{x,y}$ is the predicted vector from the regression branch of the head. N is the number of keypoints. $\mathcal{L}_{obj}$ is implemented as the focal loss [26] and $\mathcal{L}_{reg}$ is the GIoU loss [60]. $[g^l_{x,y} = 1]$ is the indicator function, being 1 if $g^l_{x,y} = 1$ and 0 otherwise.

As for the second stage, we minimize an objective function with cascade RoI heads following [21]. Each proposal x generated by the first stage contains a predicted bounding box $\mathbf{b}$ with its class label $y \in \mathcal{C}$, where $\mathcal{C} = \{0, 1\}$ is a predefined set of classes (corresponding to non-ship and ship, respectively), and the label y is obtained by calculating the IoU between $\mathbf{b}$ and the nearby ground truth box $\mathbf{g}$. As shown in Figure 5, each head t includes a classifier h_t and a regressor f_t for IoU threshold u_t, where $u_t > u_{t-1}$. $h_t(x)$ estimates the category to which proposal x belongs and $f_t(x, \mathbf{b})$ is used to regress a candidate bounding box $\mathbf{b}$ into a target bounding box $\mathbf{g}$. With these definitions, the second stage training loss $\mathcal{L}^t_2$ at each head t is defined as follows:

$$\mathcal{L}^t_2(x^t, \mathbf{b}^t) = \mathcal{L}^t_{cls}(h_t(x^t), y^t) + \gamma[y^t = 1]\mathcal{L}^t_{loc}(f_t(x^t, \mathbf{b}^t), \mathbf{g}) \tag{4}$$

where $\mathbf{b}^t = f_{t-1}(x^{t-1}, \mathbf{b}^{t-1})$ indicates the bounding box refined from the previous head. $\mathcal{L}_{cls}$ and $\mathcal{L}_{loc}$ are implemented as the focal loss [26] and smooth L_1 loss [61], respectively. $\gamma = 1$ is the trade-off coefficient. $[y^t = 1]$ is the indicator function, being 1 if $y^t = 1$ and 0 otherwise.

The total optimized loss function $\mathcal{L}$ is defined as,

$$\mathcal{L} = \mathcal{L}_1 + \sum_{t=1}^{T} \mathcal{L}^t_2 \tag{5}$$

where $T = 3$ is the number of cascaded RoI heads we employed in the experiments. In particular, $T = 1$ denotes employing a single RoI head.

3.3. Spatial Insertion Attention

When detecting ships under complex background interference in SAR images, it is likely that the scattering intensity of background interference is close to that of ships. Accordingly, the feature extractor may easily provide coarse or inaccurate features of ships for subsequent classification and regression, resulting in high false alarms. Many studies [36,62] have introduced attention mechanisms to enhance the channel relationships when extracting representative features for ship targets. Nevertheless, they only consider the relationship among feature channels but ignore the ship position information [38], which is crucial for extracting representative positional features under complex backgrounds.

Inspired by the idea of direction-aware attention [38] and the cross-channel attention [36], we develop a novel spatial insertion attention (SIA) module to improve localization accuracy, as shown in Figure 6. Formally, given an input feature $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_C] \in \mathbb{R}^{H \times W \times C}$, our goal is to find a transformation that can effectively augment the feature representations and output a new feature $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2 \dots \tilde{\mathbf{x}}_C]$ with the same size as $\mathbf{X}$. It is worth emphasizing that, the 2D global average pooling used in [36,37] squeezes global spatial information into a channel descriptor over the whole spatial dimension $H \times W$, leading to the missing of ship positional information that facilitates the detector in obtaining valid ship features.

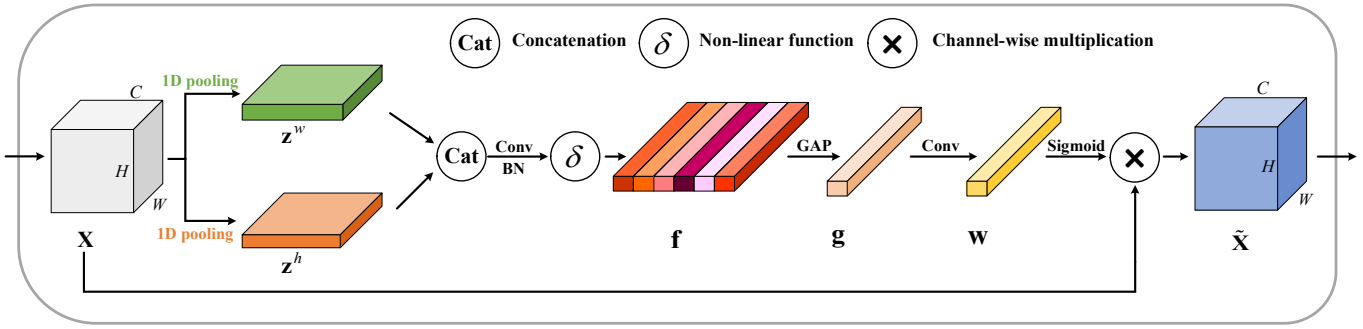


Figure 6. The design of the spatial insertion attention (SIA) module. "Conv" and "GAP" represent a convolutional layer and a global average pooling layer. "BN" denotes batch normalization."Sigmoid" is activation function.

To insert precise positional information into the channel descriptor and to enable larger target receptive fields, we factorize the 2D global pooling into two parallel 1D global pooling. Specifically, we utilize two 1D global pooling with the pooling kernels $(H, 1)$ or $(1, W)$ upon the input feature $\mathbf{X}$ to aggregate statistics along the vertical and horizontal directions, respectively. Formally, a statistic output $\mathbf{z}^h \in \mathbb{R}^{H\times C}$ is generated by shrinking $\mathbf{X}$ along the horizontal dimension, where the c-th element at height h can be calculated by:

$$z_c^h(h) = \frac{1}{W}\sum_{i=1}^{W} x_c(h, i) \tag{6}$$

Similarly, the statistic output $\mathbf{z}^w \in \mathbb{R}^{W\times C}$ is aggregated along the vertical dimension with its c-th element at width w formulated as:

$$z_c^w(w) = \frac{1}{H}\sum_{j=1}^{H} x_c(j, w) \tag{7}$$

The pair of features along two spatial directions obtained by the Equations (6) and (7), respectively, enable the network to capture long-range dependencies along either spatial direction while retaining the accurate positional information. Moreover, the captured positional information is crucial for representing the ship targets and boosting detection performance. We concatenate the two features and forward them to down-stream calculation, which can be calculated by:

$$\mathbf{f} = \delta(BN(Conv(Cat[\mathbf{z}^h, \mathbf{z}^w]))) \tag{8}$$

where $Cat[\cdot, \cdot]$ indicates a concatenation operation along the spatial dimension. $Conv(\cdot)$ represents a 1×1 convolutional layer. $BN(\cdot)$ is batch normalization and $\delta(\cdot)$ represents a non-linear activation function [63]. $\mathbf{f} \in \mathbb{R}^{(H+W)\times C/r}$ is the intermediate feature map that encodes positional information in both the horizontal direction and the vertical direction. $r = 16$ is a hyper-parameter that controls dimension reduction.

To effectively establish inter-channel relationships for re-weighting the importance of channels [36], we apply an average pooling to the concatenate feature $\mathbf{f}$ to produce a channel descriptor $\mathbf{g} \in \mathbb{R}^{C/r}$. In this way, the above captured positional information is naturelly inserted into the descriptor, enabling feature information from the spatial receptive field to be leveraged by its subsequent layers. After that, we design a simple self-gating mechanism to obtain the dependencies among channels:

$$\mathbf{w} = Conv(\mathrm{ReLU}(\mathbf{g})) \tag{9}$$

where ReLU and $Conv(\cdot)$ refer to an activation function and a convolutional layer, respectively. After applying a sigmoid activation function, the output $\mathbf{w} \in \mathbb{R}^{C\times1\times1}$ is then expanded as the same size of $\mathbf{X}$ and used as attention weights.

Finally, the output of the SIA module can be formulated as:

$$\tilde{\mathbf{x}}_c = \mathbf{x}_c \otimes w_c \tag{10}$$

where $\otimes$ represents the channel-wise multiplication between the feature map $\mathbf{x}_c \in \mathbb{R}^{H\times W}$ and the weight w_c from $\mathbf{w}$.

The proposed SIA module helps the backbone network enhance the feature representation of ships and advances the positioning ability under complex background interference and hence boosting the ship detection performance. The SIA module is placed after the end of each convolution block (conv block) in DLA-34 [59], as shown in Figure 2.

3.4. Weighted Cascade Feature Fusion

In complex SAR scenes, background interference often easily affects the network to extract distinguishable ship features [16,29]. Small ships only occupy a few pixels and are susceptible to background interference from low-level network features, which is detrimental to classification and localization. Moreover, small ships typically have little semantic information [24,34] when mapped to high-level features, making the model focus more on the compelling larger ships than small ones. Therefore, effectively integrating low-level and high-level features is crucial for multi-scale ship detection. We thus design a WCFF module to dynamically fuse multi-scale features, as shown in Figure 7. The WCFF contains two components: non-local network and optimized weighted connection. The non-local network works on the edge branch to capture wide-range receptive fields, and the optimized weighted connection is used to fuse each feature branch adaptively.

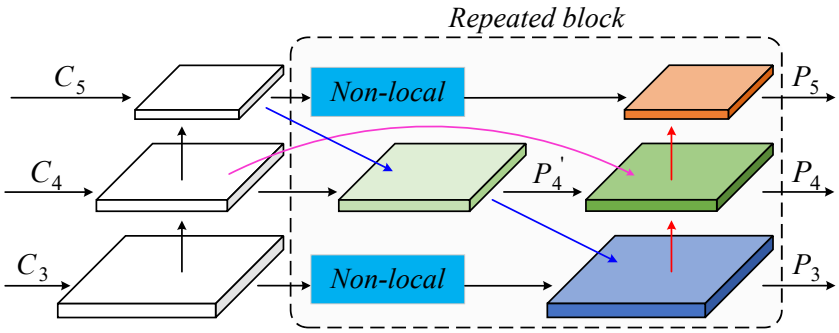


Figure 7. The structure of WCFF module. The non-local network is embedded in the edge branch.

3.4.1. Non-Local Network

A non-local operation [40] calculates the response at a position using a weighted sum of all positions ($\forall$) in the input features. From another perspective, non-local operations capture long-range dependencies directly by computing interactions between any two positions, regardless of their positional distance. Formally, it can be calculated by:

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})}\sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j) \tag{11}$$

where $\mathbf{x}_i$ is the input feature and $\mathbf{y}_i$ is the output response. The factor $\mathcal{C}(\mathbf{x})$ is used to normalize the response. The function $g(\cdot)$ computes a new representation of the input feature at the position j, which is defined as a linear embedding in this paper: $g(\mathbf{x}_j) = W_g\mathbf{x}_j$, where W_g is a weight matrix to be learned. The pairwise function $f(\cdot)$ computes

the similarity between the given features of $\mathbf{x}_i$ and $\mathbf{x}_j$ and we model this process by an embedded Gaussian function. The pairwise function $f(\cdot)$ is defined as:

$$f(\mathbf{x}_i, \mathbf{x}_j) = e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)} \tag{12}$$

where $\theta(\mathbf{x}_i) = W_\theta \mathbf{x}_i$ and $\phi(\mathbf{x}_j) = W_\phi \mathbf{x}_j$ are also two linear embeddings. We set the number of channels represented by W_g, W_θ and W_ϕ to be half of the number of channels in $\mathbf{x}$. In this paper, W_g, W_θ and W_ϕ are all easily implemented by a 1×1 convolutional layer. The normalization factor is set as $\mathcal{C}(\mathbf{x}) = \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j)$.

Next, we integrate the non-local operation in Equation (11) into a non-local network as:

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i \tag{13}$$

where $\mathbf{z}_i$ is the final output feature of the same size as $\mathbf{x}_i$ and $\mathbf{y}_i$ is calculated by Equation (11). The weight matrix W_z computes a position-wise embedding on $\mathbf{y}_i$. The structure of a non-local network is illustrated in Figure 8. The non-local network captures wide-range dependencies among each feature position to expand the target receptive fields. During this process, the global features of small ships can be well attended.

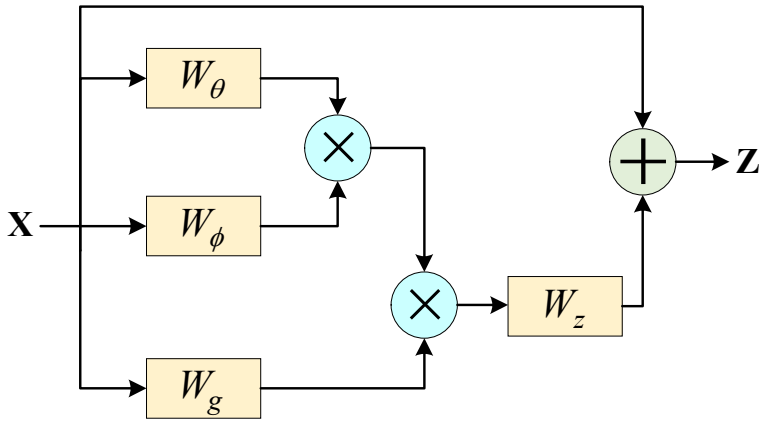


Figure 8. A Non-local network. W_θ, W_ϕ, W_g and W_z are weight matrices to be learned.

3.4.2. Optimized Weighted Connection

We first add an extra path from the original input feature to the output feature if they are at the same level. When merging pyramid features at different scales, a common way is to first resize them to a uniform scale (e.g., downsampling P_3 to P_4) and sum them up. In this work, different from the conventional FPN [35] directly employs a convolutional operation to sum pyramid features up, we further present a learnable weight to measure the importance of each input feature. The output-weighted feature P_i is calculated by:

$$P_i = \sum_j \lambda_{ij} \cdot F_j \tag{14}$$

where parameter λ_{ij} corresponds to the normalized scalar weight on each input feature F_j, which can be adaptively learned by the network. Specifically, we introduce a fast weighting mode to compute parameter λ_{ij}:

$$\lambda_{ij} = \frac{w_{ij}}{\varepsilon + \sum_j w_{ij}} \tag{15}$$

where w_{ij} is the learnable weight on the input feature and ε is a small value for numerical stability. Furthermore, we apply a ReLU function after w_{ij} to ensure $w_{ij} \geq 0$.

The output pyramid features by above several weighted optimizations can be summarized as:

$$\begin{aligned} P_5 &= Conv\left(\frac{w_{51}\cdot Nonlocal(C_5)+w_{52}\cdot Down(P_4)}{\varepsilon+w_{51}+w_{52}}\right)\\ P_4' &= Conv\left(\frac{w_{41}'\cdot C_4+w_{42}'\cdot Up(C_5)}{\varepsilon+w_{41}'+w_{42}'}\right)\\ P_4 &= Conv\left(\frac{w_{41}\cdot P_4'+w_{42}\cdot C_4+w_{43}\cdot Down(P_3)}{\varepsilon+w_{41}+w_{42}+w_{43}}\right)\\ P_3 &= Conv\left(\frac{w_{31}\cdot Nonlocal(C_3)+w_{32}\cdot Up(P_4')}{\varepsilon+w_{31}+w_{32}}\right) \end{aligned} \tag{16}$$

where $Up(\cdot)$ and $Down(\cdot)$ denote a up-sampling and down-sampling operation for resolution matching. $Nonlocal(\cdot)$ refers to a non-local network and $Conv(\cdot)$ represents a convolutional layer. The P_4' is the intermediate feature on the top-down pathway, ensuring that feature fusion has a top-down information flow.

We regard the improved fusion network after combining the above two components as an iterative block, as shown in the dashed part of Figure 7. We stack this block multiple times to obtain more valuable semantic features [56], where the number of iterations is set to 3 by balancing detection performance and computational complexity, as discussed in Section IV. Empirically, the proposed WCFF module is effective in extracting valid multi-scale ship features. The first stage heads are attached to the final output features $\{P_3, P_4, P_5\}$ for object-background classification and bounding box regression, as shown in Figure 2.

4. Results and Discussion

In this section, we first describe our experiment settings and evaluation metrics. Next, we verify the effectiveness of our improved anchor-free detector (AFD) and then conduct a series of ablation experiments to demonstrate the contribution of the SIA and WCFF modules. Third, we visualize the experimental results to further verify each proposed component in this paper. Finally, we empirically demonstrate that our proposed ATSD outperform other CNN-based detection methods.

4.1. Experiment Settings and Evaluation Metrics

4.1.1. Dataset and Settings

We conduct the experiments on two public datasets: SSDD [41] and HRSID [42]. The statistics of the SSDD and HRSID are summarized in Table 1 and the flow chart of preprocessing for SAR images is shown in Figure 9 The SSDD has 1160 SAR images containing 2358 ships lying in multiple complex scenes. Ships in SSDD have many different sizes, from the smallest size 7 × 7 to the biggest size 211 × 298, which can be utilized to evaluate the multi-scale detection performance. We randomly select 788 images and 140 images as the training and validation sets, and the remaining 232 images as the test set. To accommodate the diversity of image resolution, we use multi-scale training with the short edge in the range [256, 544] and the long edge up to 608. During testing, we use a fixed short edge at 512 and a long edge up to 608. The ablation studies are implemented on this dataset to demonstrate the effectiveness of each proposed improvement. We use the HRSID to validate the generalization of our method. It consists of 5604 SAR images with a fixed size of 800 × 800 and is divided into the training set with the amount of 65% images and the test set with 35% images. All the images are resized to a size of 1000 × 1000 for training and testing in this work.

We implement our method based on detectron2 [64]. The backbone network is initialized by the pretrained DLA-34 on the ImageNet. Specifically, our model is trained with the stochastic gradient descent (SGD) [65] optimizer with a batch size of 4 for 55 K iterations. The weight decay and momentum of the optimizer are set to 0.0001 and 0.9, respectively. The base learning rate is set as 0.02 and is dropped by 10× at iterations 30 K and 34 K. The

proposed method is performed [66,67] on an NVIDIA RTX 2080Ti GPU with PyTorch 1.6 and CUDA 10.0.

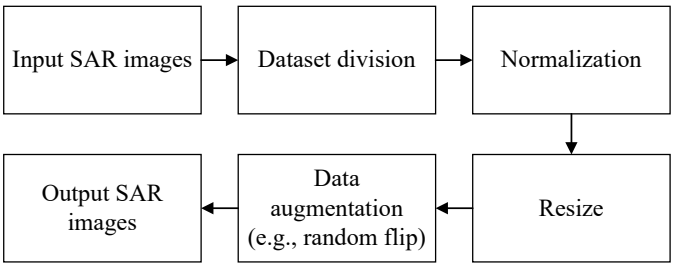


Figure 9. A flow chart of pre-processing for SAR images.

Table 1. Statistics of SSDD and HRSID datasets.

Dateset	Images	Satellite	Polarization	Resolution (m)
SSDD	1160	TerraSAR-X Sentinel-1 RadarSat-2	HH, VV, VH, HV	1–10
HRSID	5604	TerraSAR-X Sentinel-1B TanDEM	HH, VV, HV	0.5–3

4.1.2. Evaluation Metrics

We adopt precision, recall, F1 score, and the COCO evaluation protocol that includes AP (average precision), $AP_{0.5}$, $AP_{0.75}$, AP_s, AP_m, AP_l, FLOPs (G), runtime (ms) and parameters (M) as the metrics to evaluate the performance of our proposed method. Precision and recall are defined as,

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \end{aligned} \tag{17}$$

where TP (true positives) indicates the number of correctly detected ships, FP (false positives) is the number of false alarms, and FN (false negatives) denotes the number of missing ships. Generally, a detected bounding box is considered a true positive when its IoU with the ground truth is higher than 0.5. Otherwise, it is considered a false positive. Besides, we leverage the F1 score to reflect the comprehensive performance of the models, which is calculated as:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{18}$$

The precision-recall curve (PR curve) generally reveals the relation between precision and recall and can be drawn by calculating the precision-recall pair under different confidence thresholds. The AP metric is defined as the area under the PR curve and used to evaluate the overall quality of a detector, which is defined as:

$$\text{AP} = \int_0^1 P(R)\mathrm{d}R \tag{19}$$

where P represents precision, R denotes recall. $AP_{0.5}$ is AP calculated at IoU = 0.5. Similarly, the AP at the IoU threshold of 0.75 is denoted as $AP_{0.75}$, which can strictly reflect the localization accuracy. AP_s, AP_m and AP_l represent AP for small objects (areas $< 32^2$), medium objects ($32^2 <$ areas $< 96^2$), and large objects (areas $> 96^2$), respectively.

4.2. Performance of ATSD

4.2.1. Effectiveness of AFD

In order to explore the effectiveness of our novel anchor-free detector (AFD), a series of comparison experiments are designed in this part. Table 2 compares original RPN-based two-stage detectors with corresponding AFD-based two-stage detectors, where SH represents a single RoI head, and CH denotes cascaded RoI heads. The first block of the table shows the performance of two RPN-based two-stage models, defined as RPN-SH and

RPN-CH. The following block is the results of our AFD with its corresponding improved two-stage detectors obtained by combining SH or CH, termed AFD-SH and AFD-CH. As shown in Table 2, AFD-SH and AFD-CH both outperform their RPN-based two-stage detectors by up to 4.2%, 2.9% in $AP_{0.5}$ respectively, and also get a slight improvement on AP. More importantly, each improved two-stage detector with AFD has a faster runtime than the RPN-based counterpart, demonstrating the efficiency of our AFD. We argue that this is because the AFD avoids the complex IoU calculations associated with anchor boxes and makes predictions at only one location, rather than the anchor-based RPN [19] stacking multiple anchor boxes for each location. When the requirement for positioning accuracy is higher, the two-stage detector with a single RoI head gains an improvement of 3% and a significant increase of 4% with cascaded RoI heads in $AP_{0.75}$ compared to a pure AFD, revealing the necessity and effectiveness of utilizing RoI head(s) for further bounding-box refinement in the second stage to improve detection performance. Since cascade RoI heads perform better than a single RoI head, we thus employ it in the second stage to form our novel anchor-free two-stage detector (ATSD).

Table 2. Performance of anchor-based RPN and anchor-free detector (AFD) in two-stage detectors on the SSDD dataset.

Method	AP	$AP_{0.5}$	$AP_{0.75}$	Runtime (ms)
RPN-SH	0.5956	0.9137	0.7099	24.9
RPN-CH	0.6132	0.9301	0.7152	37.6
AFD	0.5941	0.9522	0.6755	**15.2**
AFD-SH	0.6089	0.9567	0.7049	16.5
AFD-CH	**0.6174**	**0.9597**	**0.7188**	23.1

In general, Table 2 demonstrates the validity of the proposed AFD employed in the first stage for boosting detection accuracy in SAR images. Based on the above experimental results, we set AFD-CH as the basic structure, followed by introducing two feature enhancement modules.

4.2.2. Ablation Study

In this part, we conduct a series of ablation experiments to analyze the influence of each feature enhancement module proposed in ATSD. For a fair comparison, all subsequent experiments are implemented with the same parameter settings. The overall results are reported in Table 3, and the corresponding PR curves under the IoU threshold of 0.75 are presented in Figure 10. It can be seen that both proposed modules have improved the detection performance to a certain extent with barely increased model parameters. Compared with the baseline, the final ATSD increases the AP by 2% and achieves a considerable improvement of 2.4% when considering $AP_{0.75}$, indicating that it can better boost the positioning accuracy. In the following parts, the effect of each module is analyzed in detail.

Table 3. Performance of each feature enhancement module in our ATSD on the SSDD dataset.

SIA	WCFF	Precision	Recall	F1	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l	Params (M)
×	×	0.9455	0.9305	0.9380	0.6174	0.9597	0.7188	0.5837	0.6656	0.5956	61.4
✓	×	0.9693	0.9404	0.9546	0.6225	0.9638	0.7366	0.5766	**0.6881**	0.6219	61.6
×	✓	0.9482	0.9444	0.9463	0.6255	0.9619	0.7202	0.5853	0.6793	0.6283	**61.3**
✓	✓	**0.9695**	**0.9484**	**0.9588**	**0.6373**	**0.9688**	**0.7435**	**0.5976**	0.6866	**0.6394**	61.5

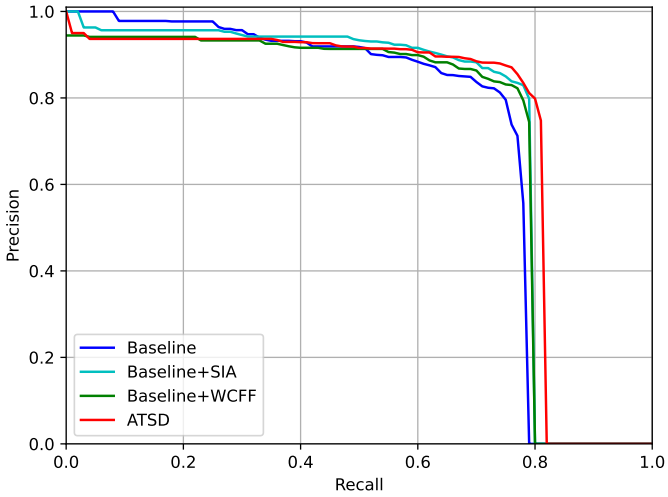


Figure 10. PR curves of different feature enhancement modules under the IoU threshold of 0.75 on the SSDD dataset.

Effectiveness of SIA. The experimental results in Table 3 justify the effectiveness of the SIA module. The $AP_{0.75}$ gains 2.2% higher than the baseline, manifesting the positive effect of SIA on achieving evident localization accuracy. Some ship detection results near the harbor are shown in Figure 11. It can be seen from Figure 11b,f that our SIA module pays attention to the ship centers and highlights the position of ships in the image while suppressing other interference areas in the heatmaps. More importantly, leveraging the positional information brought by the SIA module, the false alarms appearing in unexpected areas in strong interference environments are significantly reduced, as shown in Figure 11c,g. The red numbers in Figure 11 represent the calculated IoU between the predicted bounding box and the corresponding ground truth. The higher IoUs in Figure 11d,h indicate the SIA module can provide higher quality positioning ability than the baseline. Moreover, we empirically conduct comparison experiments on commonly-used non-linear activation functions (i.e., ReLU, swish [68], h-swish [63]) in the SIA module and show the results in Table 4. It can be seen that these activation functions have a minor influence on accuracies, indicating that all the tested activation functions can be utilized in the SIA module. However, considering the deployment advantages of the h-swish, e.g., it is faster to compute and more quantization-friendly [63], we employ it as the activation function in our SIA module. In addition, to verify whether our SIA module is more effective than other existing attention mechanisms, we compare it with five other methods, including SENet [36], scSE [54], CBAM [37], CoAM [55] and CoordAtt [38]. The comparison results are reported in Table 5. It can be observed that our method shows a better overall performance, particularly on the F1 score, $AP_{0.5}$ and $AP_{0.75}$. All these methods aim at making the model focus more on the ship targets, but our SIA has a more vital ability to extract positional information for ship location compared to the five reference methods, whose effectiveness is proven by the experimental results.

Table 4. Comparisons of various activation functions in SIA module.

Activation	AP	$AP_{0.5}$	$AP_{0.75}$
ReLU	0.6269	0.9614	0.7297
swish [68]	**0.6315**	0.9617	0.7317
h-swish [63]	0.6225	**0.9638**	**0.7366**

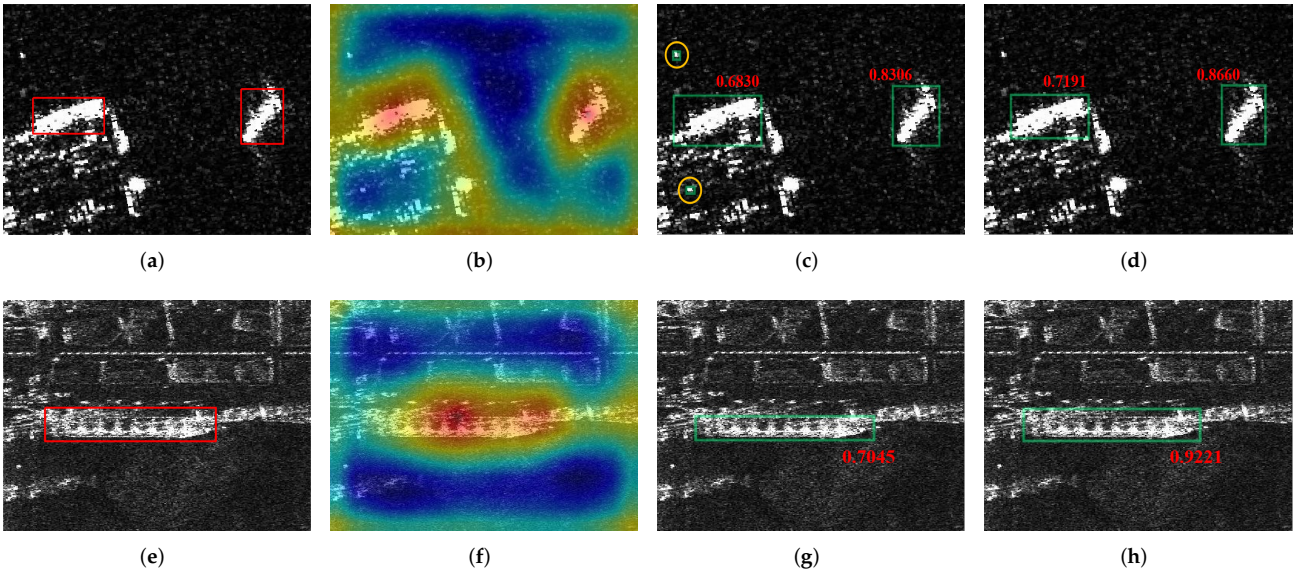


Figure 11. Detection results of the baseline with and without SIA. The spring-green rectangles represent the detection results, and the orange circles represent the false alarms. The red number reflects the IoU between the prediction result and the corresponding ground truth. (**a**,**e**) Ground truth. (**b**,**f**) Visualization of the confidence maps with SIA. (**c**,**g**) Detection results of the baseline. (**d**,**h**) Detection results of the baseline with SIA.

Table 5. The comparison results with SENet, scSE, CBAM, CoAM, CoordAtt and our SIA.

Method	Precision	Recall	F1	AP	$AP_{0.5}$	$AP_{0.75}$	Params (M)
Baseline	0.9455	0.9305	0.9380	0.6174	0.9597	0.7188	**61.4**
+SENet [36]	0.9520	0.9444	0.9482	0.6172	0.9616	0.7241	61.6
+scSE [54]	0.9653	0.9384	0.9517	0.6203	0.9507	0.7243	62.2
+CBAM [37]	0.9412	**0.9543**	0.9477	0.6196	0.9611	0.7283	61.5
+CoAM [55]	0.9613	0.9384	0.9497	0.6177	0.9609	0.7226	61.7
+CoordAtt [38]	0.9556	0.9384	0.9469	0.6189	0.9618	0.7329	61.7
+SIA	**0.9693**	0.9404	**0.9546**	**0.6225**	**0.9638**	**0.7366**	61.6

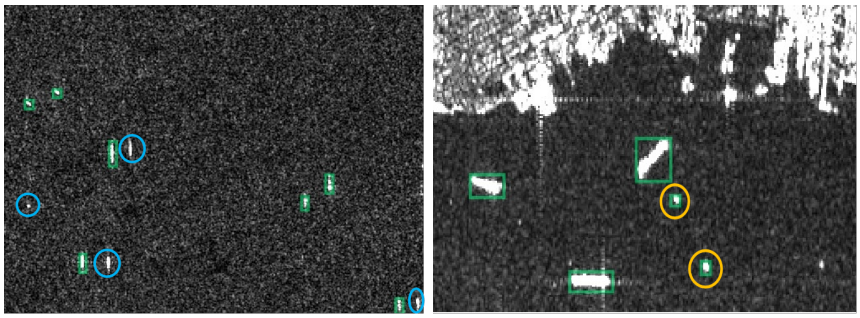
Effectiveness of WCFF. We first analyze the effectiveness of the non-local networks by successively adding them to the WCFF module. The comparison results are reported in Table 6. It can be observed that after adding the non-local network, the detector has an appreciable improvement in each AP metric with a slight increase in model parameters, indicating that the non-local network further improves the detection performance. In addition, we conduct several experiments with different settings of output channels (D) and iterations (N) in WCFF blocks to investigate the impact on the model performance while keeping other settings unchanged. The results are reported in Table 7. When only changing D from 160 to 256, the model gains a slight improvement of AP and $AP_{0.5}$, but yields 1% higher $AP_{0.75}$. When only changing N from 3 to 4, the performance remains consistent mainly. Consequently, it can be observed that WCFF is more sensitive to the number of output channels than the iterations of blocks. In order to balance the accuracy and computation cost, we finally chose D = 256 and N = 3 to build up our WCFF module. The effect of WCFF can also be intuitively verified in Figure 12. From the left column of Figure 12, the model with the WCFF module plays a positive role in detecting small ships. From the right column of Figure 12, the WCFF module is effective in identifying interference similar to the backscattering of small ships, reducing false alarms in strong background interference. In general, Table 3 indicates that the WCFF module achieves

considerable improvements over the baseline on the AP_s, AP_m and AP_l metrics, indicating the positive effectiveness of WCFF in feature fusion for multi-scale SAR ship detection.

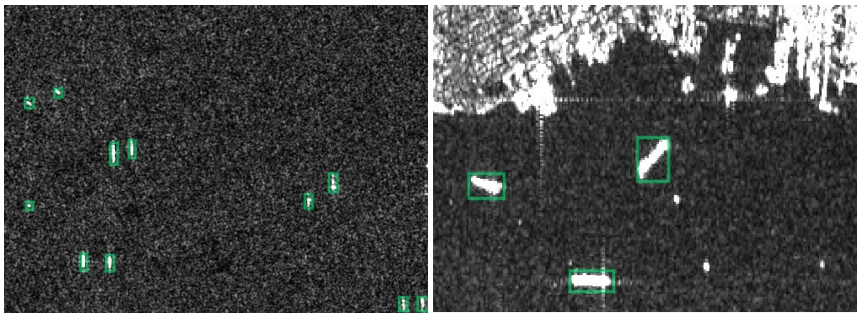
So far, we have demonstrated the validity of each feature enhancement module separately. The overall results in Table 3 show that combining these two modules can further boost the final detection performance. Next, we will conduct qualitative analyses of AFD and SIA/WCFF modules respectively.

Table 6. Effectiveness of the non-local network in WCFF module.

Non-Local	Precision	Recall	F1	AP	$AP_{0.5}$	$AP_{0.75}$	Params (M)
×	**0.9573**	0.9345	0.9457	0.6201	0.9609	0.7161	**60.5**
✓	0.9482	**0.9444**	**0.9463**	**0.6255**	**0.9619**	**0.7202**	61.3



(**a**)



(**b**)

Figure 12. Comparison results of the methods with and without WCFF. The spring-green rectangles represent the detection results. The orange and blue circles represent the false alarms and the missing ships. (**a**) Results of the baseline. (**b**) Results of the baseline with WCFF.

Table 7. Results of varying the output channels and iterations of WCFF blocks.

#Channels (D)	#Iterations (N)	AP	$AP_{0.5}$	$AP_{0.75}$	Params (M)
160	3	0.6194	0.9595	0.7114	**44.3**
160	4	0.6195	0.9555	0.7129	44.6
256	**3**	**0.6255**	**0.9619**	0.7202	61.3
256	4	0.6240	0.9597	**0.7245**	61.4

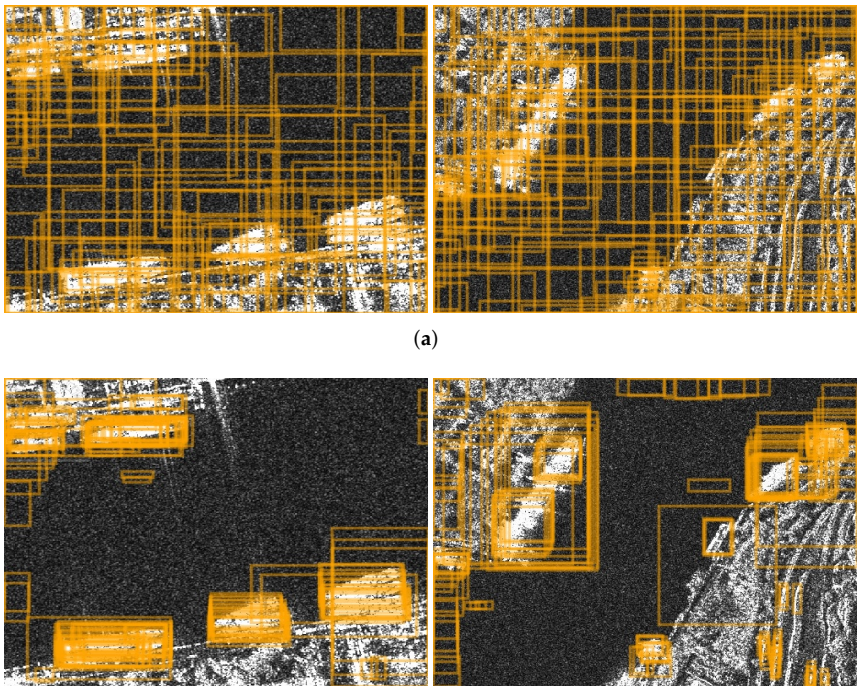
4.3. Visualization Analyses

In this section, we further explore and discuss the effectiveness of the various components in our ATSD. We also visually analyse each of the proposed components to evaluate their effectiveness.

4.3.1. AFD

The overall design idea of AFD is to reduce redundant proposals from the complex background of SAR images, thus producing higher quality proposals. With this in mind,

we design the AFD in terms of keypoint prediction fashion. To intuitively understand its effect, some visualizations of the proposals in coastal scenes are given in Figure 13. It can be seen that AFD can predict more reliable proposals around target centers, unlike the traditional RPN that searches for potential regions over the whole image. By leveraging the keypoint prediction of AFD, our detector can benefit significantly from the high-quality proposals and subsequentially feed much fewer proposals to the second stage than the RPN (128 vs. 1000). We conduct experiments to empirically study the influence of the top-k proposals for different k values in [16, 32, 64, 128, 256, 512, 1000], with the results shown in Table 8. It is clear that the accuracy increases with an increasing k value as expected. However, for $k \geq 128$, the performance gain is minimal. Therefore, we select $k = 128$ to balance the detection performance and computation cost.



(**a**)

(**b**)

Figure 13. Visualization of proposals generated by the conventional RPN and our AFD. (**a**) proposals from the RPN-based detector, for clarity, we only show proposals with its confidence score > 0.4. (**b**) proposals from the AFD-based detector.

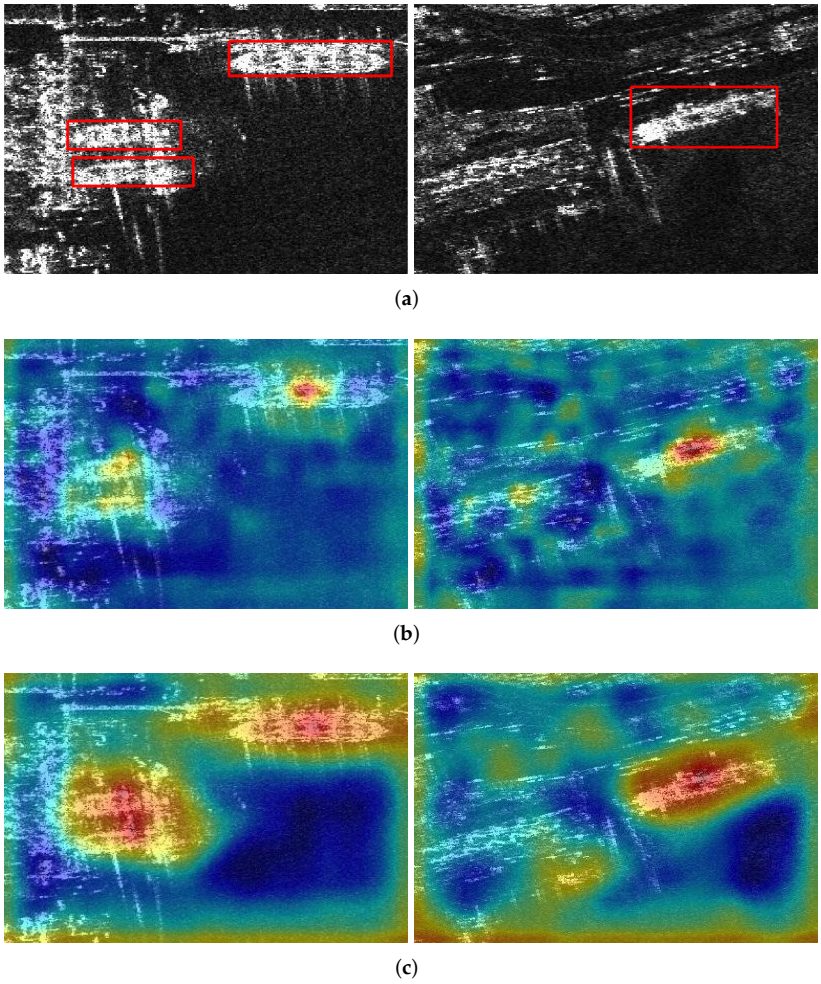
Table 8. Analysis of different values of top-k in the first stage.

Top-k	AP	$AP_{0.5}$	$AP_{0.75}$
16	0.6032	0.9316	0.7049
32	0.6112	0.9514	0.7121
64	0.6162	0.9597	0.7133
128	0.6174	0.9597	0.7188
256	0.6177	0.9597	0.7187
512	0.6178	0.9597	0.7190
1000	0.6184	0.9597	0.7190

4.3.2. SIA

The SIA module is designed to enhance ship positional information while suppressing interference in the surroundings with similar scattering characteristics to that of ships. Some visualization heatmaps are given in Figure 14 to verify the effect of the SIA module. We use Grad-CAM [69] as our visualization tool in this paper. The heatmaps are generated from the confidence scores of the last layer in the classification branch of the first stage

detection head. It can be seen from the Figure 14c that after the introduction of SIA, the ship centers have higher confidence scores, indicating that the model is noticeable to attend these positions. As shown in the Figure 14c, it can be observed that SIA can highlight the integrity of spatial structure of ships, as it extracts positional information from horizontal and vertical directions, enabling the detector to extract more distinguished features.



(**a**)

(**b**)

(**c**)

Figure 14. Visualization of the confidence maps. (**a**) Ground truth. (**b**) Visualization of the baseline without SIA. (**c**) Visualization of the baseline with SIA.

4.3.3. WCFF

The WCFF module aims to dynamically weight each pyramid feature and eliminate redundant information from background interference, so as to fuse more effective multi-scale semantic features. Table 3 and Table 6 have validated the effectiveness of WCFF. Furthermore, non-local networks are introduced on the edge branches to capture wide-range dependencies of targets and to improve the recognition accuracy between small ships and interference. To understand the effectiveness of non-local networks more intuitively, we visualize heatmaps with/without non-local networks in WCFF, as shown in Figure 15. Compared with the baseline, the WCFF without the non-local network can already globally notice small ships, as shown in Figure 15c. It can be concluded that the weighted cascade feature fusion effectively enhances the semantic representation of the small ships. Moreover, Figure 15d shows that the detector can obtain larger receptive fields after embedding the non-local network and focus more on the small ships overall.

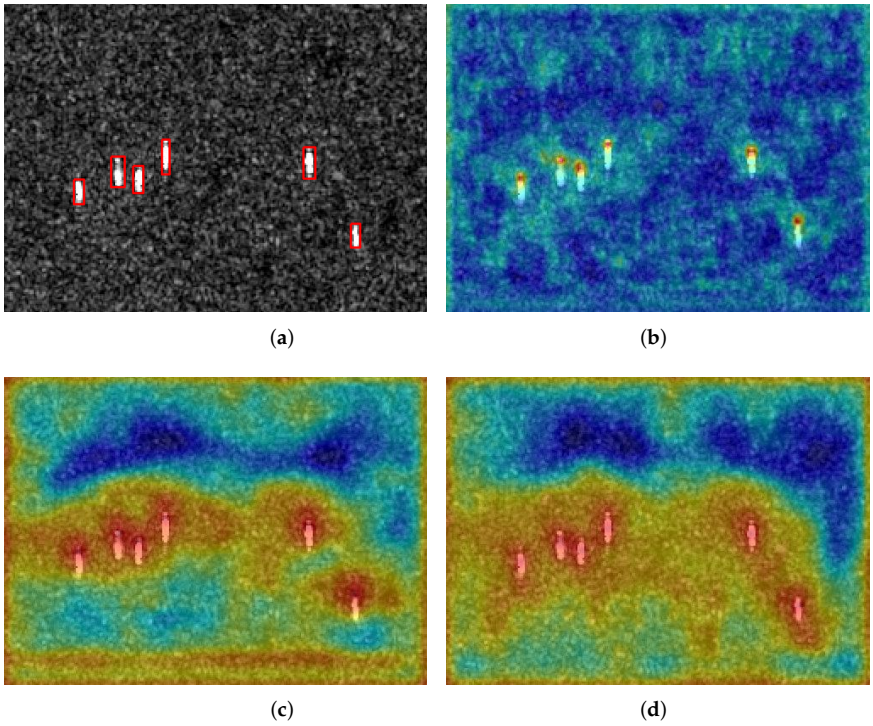


Figure 15. Visualization of the confidence maps. (**a**) Ground truth. (**b**) Visualization of the baseline. (**c**)Visualization without non-local in WCFF. (**d**) Visualization with non-local in WCFF.

Our ATSD has generally achieved satisfactory results under complex interference, as shown in Figure 16a. However, our improvement cannot effectively solve the detection problem under angle-transformed images to some extent, as shown in Figure 16b. We conjecture that a possible reason mainly causes the problem: the dataset is cropped from large-scale SAR images with a certain pixel overlap by sliding a fixed window, resulting in a sample of similar scenes. When randomly selecting training samples, the supervised learning will enable the model to learn specific memories of similar scenes and thus be sensitive to SAR images from different angles. In the future, we will consider leveraging semi-supervised or unsupervised learning to reduce the dependence on data and enhance the robustness of the model to image angle transformations.

4.4. Comparison with Other CNN-Based Detectors

To validate the effectiveness of the proposed ATSD, we compare it with six different methods on the SSDD dataset, including RetinaNet, Faster R-CNN, Cascade R-CNN, CenterNet-SSE, FCOS and YOLO v3. The CenterNet-SSE [28] combines a SSE module with CenterNet to focus on ships in complex SAR scenes. In particular, we use ResNet50 [70] in RetinaNet for easy implementation, and others use DLA-34. The comparison results are reported in Tables 9 and 10. The PR curves under the IoU threshold at 0.75 on the SSDD dataset are illustrated in Figure 17. It can be observed that our method obtains the best performance on all metrics. Compared with anchor-free detectors, the AP and the F1 score achieve more than 3% and 2% improvement over the CenterNet-SSE and outperform FCOS over 5% and 3%, respectively. Moreover, the significantly increased $AP_{0.75}$ reflects that the positioning ability in ATSD is more accurate than in the other six methods. Furthermore, the gains of AP_s, AP_m, and AP_l represent that our ATSD obtains a better performance on multi-scale ship detection. Some detection results are illustrated in Figure 18. Compared with other detection methods, our ATSD can effectively reduce false alarms under complex interference. Moreover, we calculate the IoUs of some predicted boxes with their corresponding ground-truths in some SAR images, as shown in Figure 18e,f. It can be seen that the bounding boxes predicted by our method have higher IoUs than

those predicted by CenterNet-SSE, indicating that the two feature enhancement modules, i.e., SIA/WCFF, play an important role in target localization. In addition, the PR curves in Figure 17 also confirm the distinguished improvement of our ATSD.

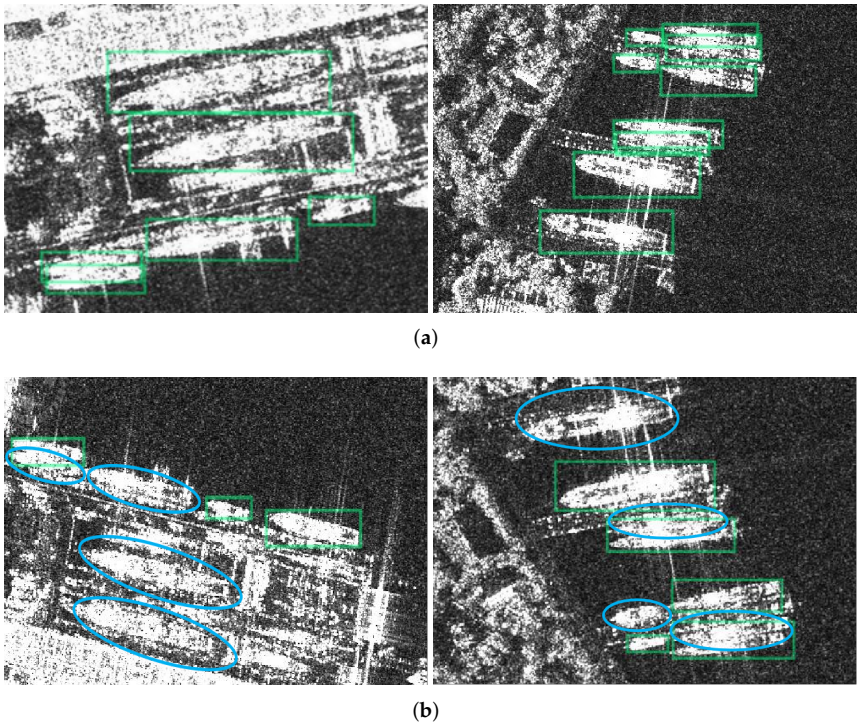


Figure 16. Comparison detection results of angle-transformed SAR images. The spring-green rectangles represent the detection results. The blue circles represent the missing ships. (**a**) Satisfactory results. (**b**) Undesirable results.

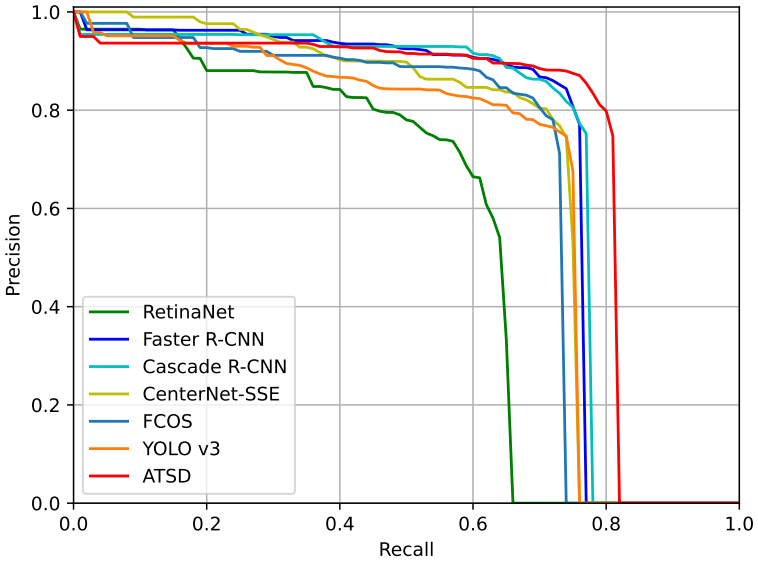


Figure 17. PR curves of different CNN-based methods under the IoU threshold of 0.75 on the SSDD dataset.

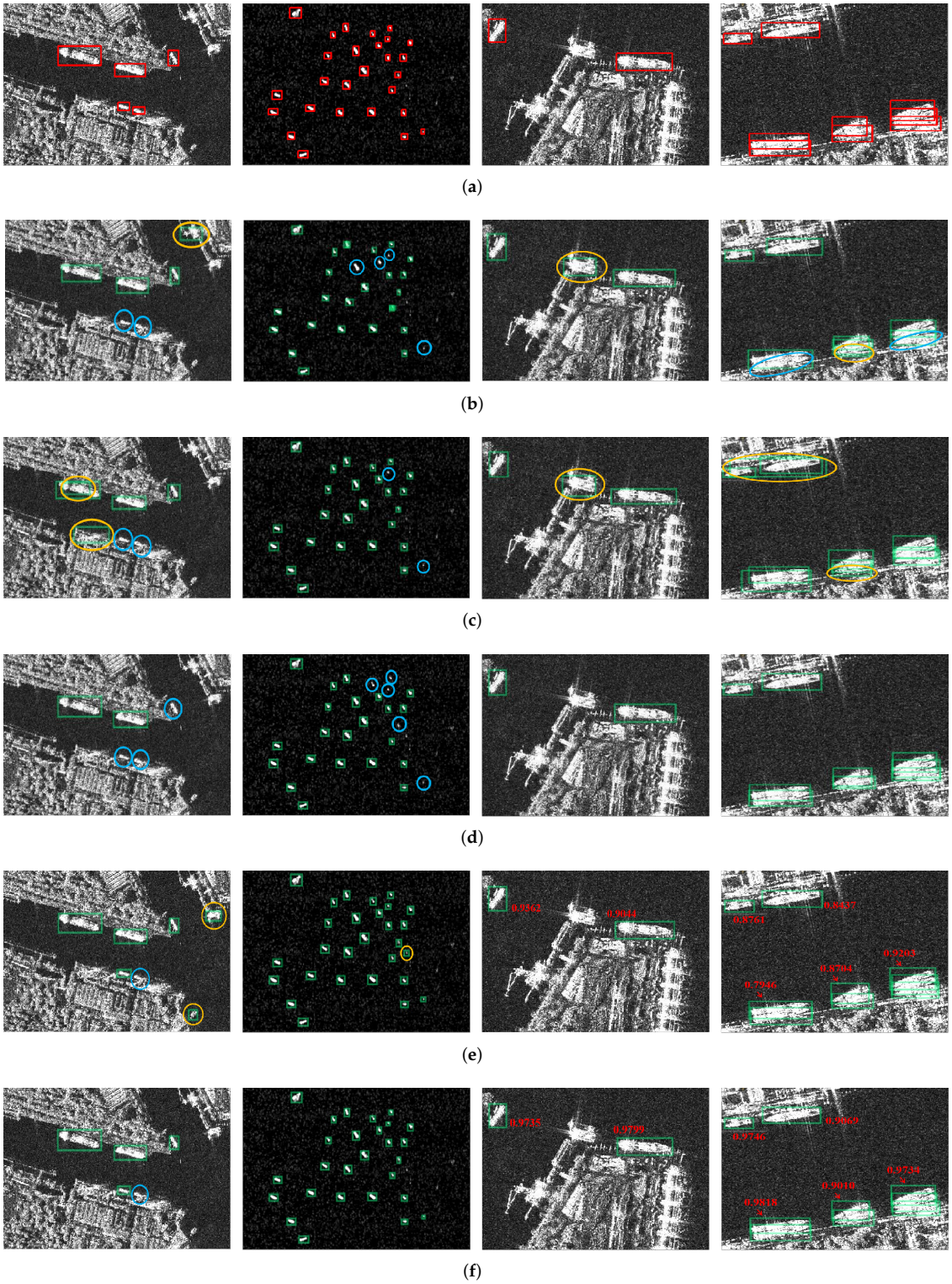


Figure 18. Comparison results of different methods on the SSDD dataset. The spring-green rectangles are the detection results. The blue and orange circles represent the missing ships and false alarms, respectively. The red value reflects the IoU between the predicted bounding box and the corresponding ground truth. (**a**) Ground truth. (**b**) Results of RetinaNet. (**c**) Results of Faster R-CNN. (**d**) Results of Cascade R-CNN. (**e**) Results of CenterNet-SSE. (**f**) Results of our ATSD.

Table 9. Comparison results of CNN-based detectors on the SSDD dataset on various AP metrics.

Method	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l
RetinaNet	0.5167	0.8934	0.5500	0.4533	0.6024	0.5014
Faster R-CNN	0.5956	0.9137	0.7099	0.5447	0.6634	0.6217
Cascade R-CNN	0.6132	0.9301	0.7152	0.5779	0.6708	0.6325
CenterNet-SSE	0.6043	0.9607	0.6859	0.5405	0.6818	0.6040
FCOS	0.5841	0.9433	0.6645	0.5480	0.6436	0.5034
YOLO v3	0.5790	0.9492	0.6619	0.5557	0.6210	0.5407
Ours	**0.6373**	**0.9688**	**0.7435**	**0.5976**	**0.6866**	**0.6394**

Table 10. Comparison results of CNN-based detectors on the SSDD dataset on other metrics.

Method	Precision	Recall	F1	Params (M)	FLOPs (G)	Runtime (ms)
RetinaNet	0.8523	0.8591	0.8557	36.9	1.59	23.4
Faster R-CNN	0.9410	0.9186	0.9297	31.8	10.51	24.9
Cascade R-CNN	0.9324	0.9305	0.9314	59.6	35.27	37.6
CenterNet-SSE	0.9584	0.9146	0.9360	**19.8**	**1.39**	18.2
FCOS	0.9445	0.9126	0.9283	31.8	4.17	19.1
YOLO v3	0.9282	0.9246	0.9264	61.5	2.96	**13.5**
Ours	**0.9695**	**0.9484**	**0.9588**	61.5	7.25	32.2

To validate the robustness of our proposed method, we also perform our ATSD and six other detectors on the HRSID dataset. Table 11 shows that our approach still maintains superior performance over the other CNN-based detectors, demonstrating the excellence of our ATSD.

Table 11. Comparison results of CNN-based detectors on the HRSID dataset

Method	Precision	Recall	F1	AP	$AP_{0.5}$
RetinaNet	0.8433	0.8034	0.8229	0.5850	0.8450
Faster R-CNN	0.8685	0.8571	0.8628	0.6301	0.8781
Cascade R-CNN	0.8906	0.8579	0.8740	0.6678	0.8740
CenterNet-SSE	0.8988	0.8360	0.8663	0.6178	0.8748
FCOS	0.8957	0.8328	0.8631	0.6027	0.8730
YOLO v3	0.8298	**0.8809**	0.8546	0.6055	0.8712
Ours	**0.9026**	0.8656	**0.8837**	**0.6726**	**0.8819**

5. Conclusions

This paper proposes an anchor-free two-stage detector named ATSD for SAR ship detection under complex background interference. A new AFD is present to generate fewer but higher quality proposals than the RPN around the target centers, achieving a better speed-accuracy trade-off. Moreover, two newly-designed feature enhancement modules, i.e., the SIA and WCFF, are proposed for effective feature enhancement. The SIA highlights the ship's positional information for extracting distinctive features under strong interference. The WCFF is employed to weight pyramid features adaptively for fusing more valid multi-scale semantic information. Extensive experiments on the SSDD and HRSID datasets demonstrate the effectiveness of our proposed method. Furthermore, the comparative results show that our ATSD achieves outstanding detection performance over the other mainstream CNN-based detectors.

Author Contributions: Conceptualization, C.Y. and L.Z.; methodology, C.Y.; software, C.Y.; validation, C.Y. and P.X.; formal analysis, C.Y. and P.X.; investigation, C.Y. and Y.F.; writing—original draft preparation, C.Y.; writing—review and editing, C.Y. and L.Z.; visualization, C.Y.; supervision, L.Z.; project administration, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China under Grant 62101603, in part by the Shenzhen Science and Technology Program under Grant KQTD20190929172704911, in part by the Aeronautical Science Foundation of China under Grant 2019200M1001, and in part by the Guangdong Key Laboratory of Advanced IntelliSense Technology under Grant 2019B121203006.

Data Availability Statement: Not applicable.

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

References

1. Liu, C.; Yang, J.; Zheng, J.; Nie, X. An Unsupervised Port Detection Method in Polarimetric SAR Images Based on Three-Component Decomposition and Multi-Scale Thresholding Segmentation. *Remote Sens.* **2022**, *14*, 205. [CrossRef]
2. Heiselberg, P.; Sørensen, K.A.; Heiselberg, H.; Andersen, O.B. SAR Ship–Iceberg Discrimination in Arctic Conditions Using Deep Learning. *Remote Sens.* **2022**, *14*, 2236. [CrossRef]
3. Hamze-Ziabari, S.M.; Foroughan, M.; Lemmin, U.; Barry, D.A. Monitoring Mesoscale to Submesoscale Processes in Large Lakes with Sentinel-1 SAR Imagery: The Case of Lake Geneva. *Remote Sens.* **2022**, *14*, 4967. [CrossRef]
4. Cerutti-Maori, D.; Klare, J.; Brenner, A.R.; Ender, J.H.G. Wide-Area Traffic Monitoring With the SAR/GMTI System PAMIR. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 3019–3030. [CrossRef]
5. Bi, H.; Zhu, D.; Bi, G.; Zhang, B.; Hong, W.; Wu, Y. FMCW SAR Sparse Imaging Based on Approximated Observation: An Overview on Current Technologies. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4825–4835. [CrossRef]
6. Ouchi, K. Recent Trend and Advance of Synthetic Aperture Radar with Selected Topics. *Remote Sens.* **2013**, *5*, 716–807. [CrossRef]
7. Xue, R.; Bai, X.; Zhou, F. Spatial–Temporal Ensemble Convolution for Sequence SAR Target Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1250–1262. [CrossRef]
8. Yeremy, M.; Campbell, J.; Mattar, K.; Potter, T. Ocean Surveillance with Polarimetric SAR. *Can. J. Remote Sens.* **2001**, *27*, 328–344. [CrossRef]
9. Shirvany, R.; Chabert, M.; Tourneret, J.Y. Ship and Oil-Spill Detection Using the Degree of Polarization in Linear and Hybrid/Compact Dual-Pol SAR. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 885–892. . 2012.2182760. [CrossRef]
10. Chen, J.; Zhang, J.; Wu, T.; Hao, J.; Wu, X.; Ma, X.; Zhu, X.; Lou, P.; Zhang, L. Activity and Kinematics of Two Adjacent Freeze–Thaw-Related Landslides Revealed by Multisource Remote Sensing of Qilian Mountain. *Remote Sens.* **2022**, *14*, 5059. [CrossRef]
11. Sahour, H.; Kemink, K.M.; O'Connell, J. Integrating SAR and Optical Remote Sensing for Conservation-Targeted Wetlands Mapping. *Remote Sens.* **2022**, *14*, 159. [CrossRef]
12. Gao, G.; Liu, L.; Zhao, L.; Shi, G.; Kuang, G. An Adaptive and Fast CFAR Algorithm Based on Automatic Censoring for Target Detection in High-Resolution SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1685–1697. [CrossRef]
13. Leng, X.; Ji, K.; Yang, K.; Zou, H. A Bilateral CFAR Algorithm for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1536–1540. [CrossRef]
14. Gao, G. A Parzen-Window-Kernel-Based CFAR Algorithm for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 557–561. [CrossRef]
15. Erfanian, S.; Tabataba Vakili, V. Introducing Excision Switching-CFAR in K Distributed Sea Clutter. *Signal Process.* **2009**, *89*, 1023–1031. [CrossRef]
16. 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]
17. Jiao, J.; Zhang, Y.; Sun, H.; Yang, X.; Gao, X.; Hong, W.; Fu, K.; Sun, X. A Densely Connected End-to-End Neural Network for Multiscale and Multiscene SAR Ship Detection. *IEEE Access* **2018**, *6*, 20881–20892. . [CrossRef]
18. 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]
19. 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.* **2017**, *39*, 1137–1149. [CrossRef]
20. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
21. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. [CrossRef]

22. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 821–830. [CrossRef]
23. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017; pp. 6517–6525. [CrossRef]
24. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-Yolov4: Scaling Cross Stage Partial Network. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Nashville, TN, USA, 20–25 June 2021; pp. 13024–13033. [CrossRef]
26. 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. [CrossRef]
27. Hong, Z.; Yang, T.; Tong, X.; Zhang, Y.; Jiang, S.; Zhou, R.; Han, Y.; Wang, J.; Yang, S.; Liu, S. Multi-Scale Ship Detection From SAR and Optical Imagery Via A More Accurate YOLOv3. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6083–6101. [CrossRef]
28. 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.* **2021**, *59*, 379–391. [CrossRef]
29. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1331–1344. [CrossRef]
30. Law, H.; Deng, J. Cornernet: Detecting Objects as Paired Keypoints. In Proceedings of the 15th European Conference on Computer Vision, ECCV, Munich, Germany, 8–14 September 2018; Volume 11218 LNCS, pp. 765–781. [CrossRef]
31. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
32. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9626–9635. [CrossRef]
33. 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]
34. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ Model for Ship Detection in SAR Images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]
35. Lin, T.Y.; Dollár, 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, 21–26 July 2017; pp. 936–944. [CrossRef]
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision, ECCV, Munich, Germany, 8–14 September 2018; Volume 11211 LNCS, pp. 3–19. [CrossRef]
38. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717. [CrossRef]
39. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [CrossRef]
40. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803. [CrossRef]
41. Li, J.; Qu, C.; Shao, J. Ship Detection in SAR Images Based on an Improved Faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017; pp. 1–6. [CrossRef]
42. 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]
43. 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]
44. 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 14th European Conference on Computer Vision, ECCV, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9905 LNCS, pp. 21–37. [CrossRef]
45. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. Depthwise Separable Convolution Neural Network for High-Speed SAR Ship Detection. *Remote Sens.* **2019**, *11*, 2483. [CrossRef]
46. Su, H.; Wei, S.; Wang, M.; Zhou, L.; Shi, J.; Zhang, X. Ship Detection Based on RetinaNet-Plus for High-Resolution SAR Imagery. In Proceedings of the 2019 6th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), Xiamen, China, 26–29 November 2019; pp. 1–5. [CrossRef]
47. Zhou, X.; Zhuo, J.; Krähenbühl, P. Bottom-Up Object Detection by Grouping Extreme and Center Points. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 850–859. [CrossRef]
48. Zhou, X.; Koltun, V.; Krähenbühl, P. Probabilistic Two-Stage Detection. *arXiv* **2021**, arXiv:2103.0746.

49. 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; pp. 9756–9765. [CrossRef]
50. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In Proceedings of the 34th Conference on Neural Information Processing Systems, NeurIPS, Online, 6–12 December 2020; Volume 2020.
51. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems, NeurIPS, Montreal, QC, Canada 2–8 December 2018; Volume 2018; pp. 9401–9411.
52. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3138–3147. [CrossRef]
53. Bai, J.; Ren, J.; Yang, Y.; Xiao, Z.; Yu, W.; Havyarimana, V.; Jiao, L. Object Detection in Large-Scale Remote-Sensing Images Based on Time-Frequency Analysis and Feature Optimization. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
54. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI, Granada, Spain, 16–20 September 2018; pp. 421–429. [CrossRef]
55. Yang, X.; Zhang, X.; Wang, N.; Gao, X. A Robust One-Stage Detector for Multiscale Ship Detection With Complex Background in Massive SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
56. Ghiasi, G.; Lin, T.Y.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7029–7038. [CrossRef]
57. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
58. Shamsolmoali, P.; Chanussot, J.; Zareapoor, M.; Zhou, H.; Yang, J. Multipatch Feature Pyramid Network for Weakly Supervised Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
59. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412. . 2018.00255. [CrossRef]
60. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. [CrossRef]
61. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
62. Li, X.; Hu, X.; Yang, J. Spatial Group-wise Enhance: Improving Semantic Feature Learning in Convolutional Networks. *arXiv* **2019**, arXiv:1905.09646.
63. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324. [CrossRef]
64. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 15 Octomber 2022).
65. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
66. Topcuoglu, H.; Hariri, S.; Wu, M.Y. Performance-Effective and Low-Complexity Task Scheduling for Heterogeneous Computing. *IEEE Trans. Parallel Distrib. Syst.* **2002**, *13*, 260–274. [CrossRef]
67. Zhang, Y.; Zhou, Y.; Lu, H.; Fujita, H. Spark Cloud-Based Parallel Computing for Traffic Network Flow Predictive Control Using Non-Analytical Predictive Model. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 7708–7720. [CrossRef]
68. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for Activation Functions. *arXiv* **2017**, arXiv:1710.05941.
69. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV, Venice, Italy, 22–29 October 2017; Volume 2017, pp. 618–626. [CrossRef]
70. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]