remote sensing

MDPI

Article

Ship Detection in SAR Images Based on Feature Enhancement Swin Transformer and Adjacent Feature Fusion

Kuoyang Li [1], Min Zhang [1,*], Maiping Xu [2], Rui Tang [2], Liang Wang [2] and Hai Wang [1]

[1] School of Aerospace Science and Technology, Xidian University, Xi'an 710126, China; kyli_1@stu.xidian.edu.cn (K.L.); wanghai@mail.xidian.edu.cn (H.W.)
[2] Shaanxi Academy of Aerospace Technology Application Co., Ltd., Xi'an 710199, China; linsheng@stu.xidian.edu.cn (M.X.); kxzhou@stu.xidian.edu.cn (R.T.); cheng_xi@stu.xidian.edu.cn (L.W.)
* Correspondence: minzhang@xidian.edu.cn

Abstract: Convolutional neural networks (CNNs) have achieved milestones in object detection of synthetic aperture radar (SAR) images. Recently, vision transformers and their variants have shown great promise in detection tasks. However, ship detection in SAR images remains a substantial challenge because of the characteristics of strong scattering, multi-scale, and complex backgrounds of ship objects in SAR images. This paper proposes an enhancement Swin transformer detection network, named ESTDNet, to complete the ship detection in SAR images to solve the above problems. We adopt the Swin transformer of Cascade-R-CNN (Cascade R-CNN Swin) as a benchmark model in ESTDNet. Based on this, we built two modules in ESTDNet: the feature enhancement Swin transformer (FESwin) module for improving feature extraction capability and the adjacent feature fusion (AFF) module for optimizing feature pyramids. Firstly, the FESwin module is employed as the backbone network, aggregating contextual information about perceptions before and after the Swin transformer model using CNN. It uses single-point channel information interaction as the primary and local spatial information interaction as the secondary for scale fusion based on capturing visual dependence through self-attention, which improves spatial-to-channel feature expression and increases the utilization of ship information from SAR images. Secondly, the AFF module is a weighted selection fusion of each high-level feature in the feature pyramid with its adjacent shallow-level features using learnable adaptive weights, allowing the ship information of SAR images to be focused on the feature maps at more scales and improving the recognition and localization capability for ships in SAR images. Finally, the ablation study conducted on the SSDD dataset validates the effectiveness of the two components proposed in the ESTDNet detector. Moreover, the experiments executed on two public datasets consisting of SSDD and SARShip demonstrate that the ESTDNet detector outperforms the state-of-the-art methods, which provides a new idea for ship detection in SAR images.

Keywords: synthetic aperture radar (SAR); ship detection; feature enhancement Swin transformer; adjacent feature fusion; Cascade R-CNN

check for updates

Citation: Li, K.; Zhang, M.; Xu, M.; Tang, R.; Wang, L.; Wang, H. Ship Detection in SAR Images Based on Feature Enhancement Swin Transformer and Adjacent Feature Fusion. *Remote Sens.* **2022**, *14*, 3186. https://doi.org/10.3390/rs14133186

Academic Editors: Ali Khenchaf and Jean-Christophe Cexus

Received: 24 May 2022
Accepted: 28 June 2022
Published: 2 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Synthetic aperture radar (SAR) has the advantages of all-weather, all-day, anti-jamming, far detection, and high concealment [1]. As a remote sensing image data source, SAR images are widely used in scientific research, military reconnaissance, disaster monitoring, resource planning, and natural environment protection [2]. SAR ship detection, as a fundamental marine mission, has an essential value in maritime resource management and maritime emergency rescue. Therefore, ship detection in SAR images is an attractive research topic.

In recent years, deep learning technology has presented a diversified development trend in the field of image processing [3–5]. Deep learning-based object detection techniques have been developing rapidly, and many common detection methods are classified

into one-stage and two-stage. The one-stage object detection algorithm is a direct regression to obtain the bounding box coordinates and class probabilities at the same time. The representative algorithms are YOLO series [6–9], SSD [10], RetinaNet [11], etc. The two-stage algorithm generates the region proposal bounding box and then classifies the bounding box. The representative algorithms are RCNN [12], Fast RCNN [13], Faster RCNN [14], Mask RCNN [15], Cascade R-CNN [16], etc. There are also some excellent functional modules for enhancing the above models, such as the feature pyramid network (FPN) [17] and the convolutional block attention module (CBAM) [18]. Researchers have applied transformer technology to computer vision in the last two years. The attention mechanism is the main module of the transformer, which is used to establish the global relationship between image pixels. Using the image serialization process, position embedding is introduced to describe the position relationship, thus saving the spatial information of the image. Due to the above advantages, transformer techniques are widely used in computer vision, consisting of image classification [19], object detection [20], and so on [21–23].

Relying on target detection technology, remote sensing image processing is flourishing [24–27], and ship detection of SAR images is an important research direction in remote sensing. The SAR ship detection data SSDD is provided by Li et al. [28] and is improved for Faster RCNN based on the dataset. Later, Wang et al. [29] proposed a publicly available dataset SARShip and proposed an optimized version based on SSD by reducing unnecessary convolutional layers, which enhances the detection of small ships while improving the detection speed. Zhang et al. [30] proposed a grid-based convolutional neural network (G-CNN) on the basis of YOLO, which uses a backbone convolutional neural network (B-CNN) and a detection convolutional neural network (D-CNN) for high-speed ship detection. Zhou et al. [31] designed the CSPMRes2 module and the feature pyramid network with fusion coefficients (FC-FPN) module based on YOLOv5 to improve the accuracy of multi-scale ship detection. Zhang et al. [32] proposed a quad feature pyramid network (Quad-FPN). The network is comprised of four FPNs: deformable convolution, content-aware feature reassembly, path aggregation spatial attention, and balanced scale global attention. It optimizes the complex background interference and multi-scale ship feature discrepancy problems. Cui et al. [33] densely connected the convolutional block attention module (CBAM) to a pyramidal network to form a dense attention pyramidal network (DAPN). It obtains richer semantic feature information in multi-scale ships and highlights the salient features on specific scales. After the appearance of a visual transformer, they soon applied it to the ship detection of SAR images. Xia et al. [34] proposed a visual transformer architecture, which is termed CRTransSar, in consideration of the contextual joint representation learning by combing the transformer technique and convolution neural network (CNN). Some researchers put forward the anchor-free frame method to detect ships in SAR images. Qu et al. [35] developed an anchor-free detection model by introducing a transformer encoder module, which not only enhances the dependency between ship objects but also focuses on the contextual relationship between objects and the global image. Feng et al. [36] designed a lightweight multiscale backbone based on YOLOX and proposed a new position-enhanced attention strategy to construct a lightweight anchor-free algorithm for SAR ship detection. All these methods improve the backbone and neck modules of the ship detection algorithm in SAR images to different degrees. In addition to the above methods, some scholars have studied few-shot ship detection in recent years and proposed very superior methods [37–39]. Zhang et al. [40] used a semantic embedding approach to align visual features and semantic features for zero-shot ship detection. Zhang et al. [41] proposed to detect ships of unknown classes with only a few labeled and designed a few-shot learning algorithm with an attention mechanism. However, there is still the problem of information loss because of the increased depth of feature extraction, as well as the problem of attention scattering due to the high similarity between background and object. Furthermore, there is the problem that the object area of the ship is not accurately localized during the detection process.

This paper describes a ship detection network combining transformer and CNN to achieve an excellent detection performance in SAR images. We first introduce a CNN based on Swin transformer [21] to build a feature enhancement Swin transformer (FESwin) module to better extract the features of ships. FESwin not only has the global context information perception and spatial information extraction capabilities of a transformer, but also the local information feature extractability and aggregate channel feature information of CNN. In addition, aiming to overcome the shortcoming of feature fusion, we construct a feature pyramid network architecture for ship detection, namely the adjacent feature fusion (AFF) module. A weighted selection fusion of high-level features and adjacent shallow features, with the fusion proportion using adaptive weights that can be learned during the training phase of the ship detection network, so that AFF has a powerful feature fusion capability. On the other hand, we introduce Cascade R-CNN [16] as the detection head to improve the detection accuracy of the ship detection network. Finally, we combine FESwin, AFF, and Cascade R-CNN methods to build an enhancement Swin transformer detection network (ESTDNet). We experimentally validate the design of ESTDNet on SSDD and SARShip datasets. The results show that ESTDNet has significantly improved multi-scale detection performance in complex backgrounds. This paper focuses on the optimal design of the backbone and neck parts of the object detection framework. Therefore, we use the Cascade R-CNN Swin framework as the baseline model of our method, and we can flexibly embed our methods in any other object detection framework as a functional module. We summarize the contributions of this work below.

1. A FESwin module is proposed as a backbone network to extract ship feature information. The module not only has the excellent spatial feature information processing capability of the Swin transformer but also uses CNN to enhance the association among feature map channels. It effectively suppresses the problem of insufficient feature extraction caused by strong scattering of SAR objects, obtaining more significant feature information at different scales, and enhances the transmission capability of feature information.
2. We construct an AFF module that allows shallow feature information in the feature pyramid to be selectively fused into adjacent higher-level feature information adaptively. The idea of learnable weights and proximity fusion reduces the huge information difference between the bottom and higher-level features and alleviate the problem of attentional dispersion in feature maps.
3. A ship detector with SAR images is constructed by combining the FESwin module with the AFF module. The effects of FESwin and AFF on ESTDNet were verified separately for both models to improve performance. Experiments on SSDD and SARShip datasets show that ESTDNet can detect ships better in SAR images with higher detection accuracy.

The rest of this paper is arranged as follows: Section 2 describes the proposed network. Section 3 analyses the experimental results of the network proposed in this paper and compares them with other algorithms. Section 4 discusses some phenomena according to the experimental results. Finally, Section 5 gives conclusions about this paper.

2. The Proposed Method

This paper proposes a SAR ship detection algorithm called ESTDNet based on feature enhancement Swin transformer and adjacent feature fusion, which uses Cascade R-CNN Swin as the benchmark model. Firstly, combining the advantages of CNN structure and Swin transformer, FESwin is innovatively proposed as a new backbone network. Second, the FPN is replaced with AFF to reconstruct the multi-scale feature pyramid. In the testing phase, we use the common objects in context (COCO) metrics as the evaluation standard. The overall framework is shown in Figure 1. The algorithm proposed in this paper will be explained in detail from three aspects.

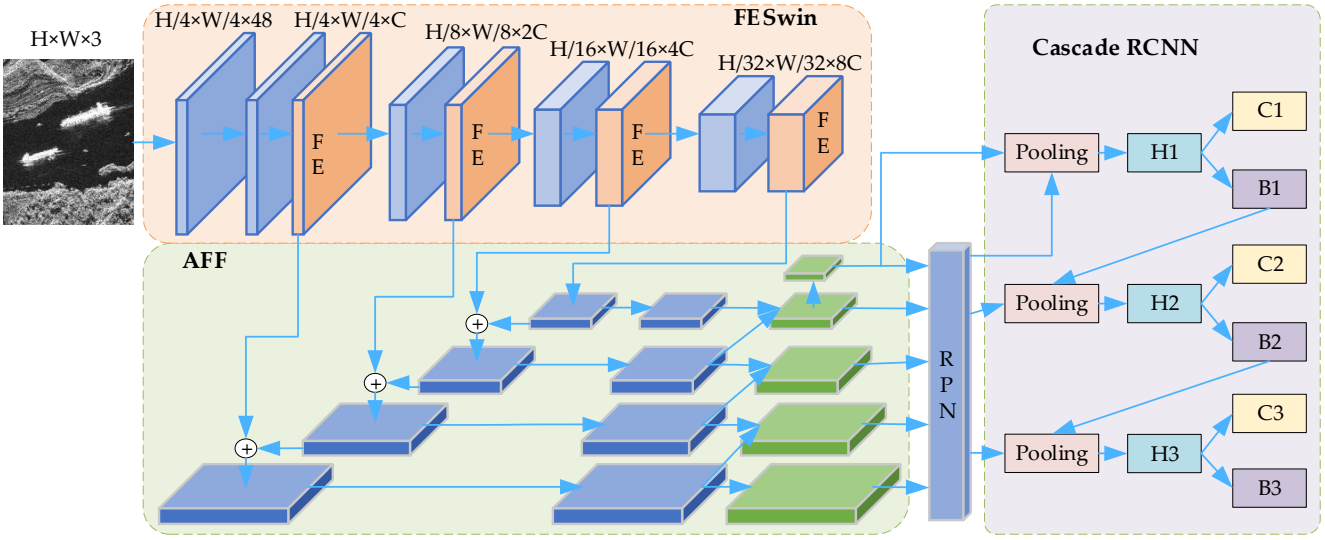

Figure 1. The overview of ESTDNet. Compared to Cascade R-CNN Swin, we mainly improve the feature extraction capability by using FESwin as the backbone network and employing AFF to replace the original FPN for feature fusion.

2.1. FESwin Backbone Network

In order to obtain a feature map with richer information about the ships, the feature information can be transferred to a deeper level of the model. In this paper, we propose the feature enhancement Swin transformer, called FESwin, as the backbone network for feature extraction, as shown in Figure 2. Our method is based on the use of the transformer idea, hierarchical structure design, and window attention mechanism to establish the association between image features. Firstly, the contextual information before and after the Swin transformer block of each stage is fused using a skip connection. Secondly, we enhance the inter-channel interaction of information after fusion. It optimizes the feature extraction capability and has better ship detection accuracy in SAR images.

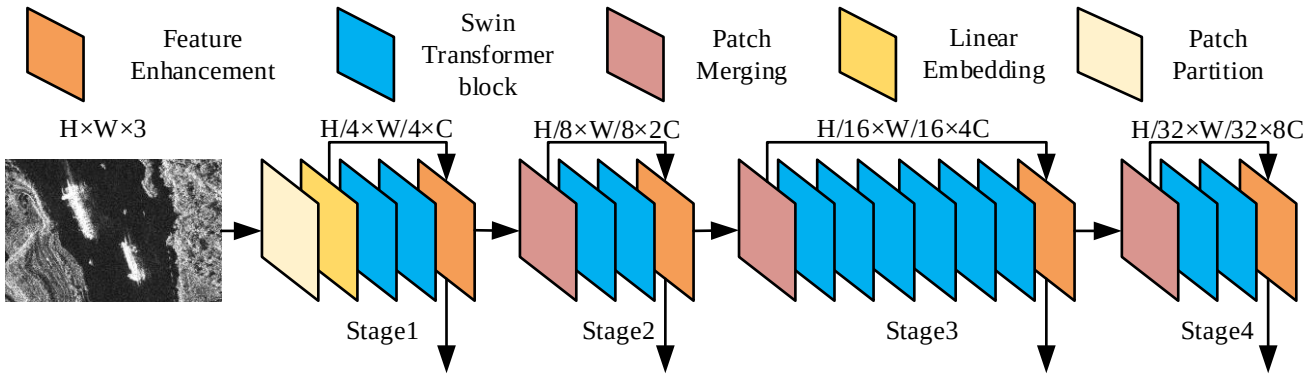

Figure 2. FESwin backbone network model. Swin transformer block [21] and feature enhancement for feature extraction, and the remaining three modules that are responsible for scaling the model.

As shown in Figure 2, the FESwin backbone network is composed of a Swin transformer and feature enhancement module. Relying on the hierarchical design of the Swin transformer, the feature enhancement module is introduced in each feature extraction stage. The feature maps of each stage are feature enhanced again, and the more expressive and informative feature maps are used as the output of the current stage. Meanwhile, the output of each stage of FESwin is used as the input of the next stage to obtain more advanced semantic information, enhancing the whole backbone network. In our study, we found that the Swin transformer performs attentional operations at each stage with relatively independent information between each dimension and weak interaction among

feature channels. Although it performs the integration of spatial and channel information, the correlation between feature channels of a single stage is weak. Therefore, the feature enhancement module is proposed in the backbone to aggregate the contextual information of different perceptions before and after the Swin transformer block and takes advantage of the CNN to enhance the channel information interaction. It enables further integration of channel and spatial information to enhance model representation.

We propose a feature enhancement module, as shown in Figure 3. The feature maps before and after feature extraction are fused by using a skip connection. After that, the fusion is carried out in equal proportion using single-point channel information interaction as the primary method and local spatial information interaction as the supplement. Local spatial information interaction using convolutional layers and activation functions to enhance local perception and obtain a larger perceptual field. Channel information interaction is performed by point-wise convolution at each spatial location, and cross-channel information aggregation is performed for each patch. Finally, the two are fused by weighting to enhance the feature extraction ability, which makes the model have a better expression ability.

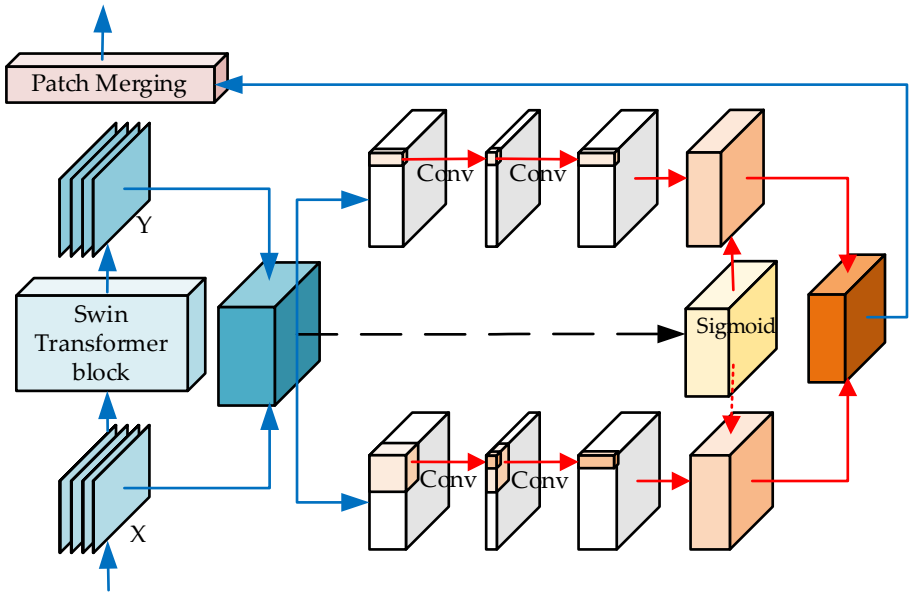

Figure 3. Feature enhancement module, which uses CNN to enhance the relationship between channels and improve the expression ability of the model.

We can represent the feature maps before and after the Swin transformer block as X and Y, and make channel information integration and spatial information integration after the fusion of the two feature maps. We use point-wise convolution (pwconv) in channel information integration, allowing point-by-point channel information at each spatial location to be used interactively. We denote the output by $C(X) \in R^{C\times H\times W}$ and the definition is shown in Equation (1). Besides the channel information, in terms of spatial information, we integrate each point in a single channel with the neighboring location points. We replace the output by $S(X) \in R^{C\times H\times W}$ and the definition is shown in Equation (2).

$$C(X) = LN(pwconv_2(\delta(LN(pwconv_1(X+Y))))) \quad (1)$$

$$S(X) = LN(conv_2(\delta(LN(conv_1(X+Y))))) \quad (2)$$

where $pwconv_1(X) \in R^{C/4\times H\times W}$ (4 is the channel compression rate) represents the dimensionality reduction and $pwconv_2(X) \in R^{C\times H\times W}$ for dimensionality increase, δ is ReLU and LN indicates layer normalization. The convolutional kernel sizes of conv1 and conv2 are $(C/4) \times C \times 3 \times 3$ and $C \times (C/4) \times 3 \times 3$, respectively. By calculating C(X) and S(X) after the same shape as the input features, both can retain the fine details in the original features to different degrees. We use Z to express the resultant feature map after the weighted

fusion of spatial and channel features, and the weights required for fusion are obtained by (X + Y) using the activation function sigmoid, denoted as $m(X+Y) \in R^{C\times H\times W}$. In the weighted fusion method, the sum of the weights of the feature mappings is restricted to 1 while using $\otimes$ to denote multiplication by elements, and the above calculation process is shown in Equation (3)

$$Z = m(X+Y) \otimes C(X) + (1 - m(X+Y)) \otimes S(X) \quad (3)$$

2.2. AFF Module

The architecture of the proposed AFF is shown in Figure 4, with two main optimizations based on the FPN. One part is the bottom-up augmentation of the adjacent layer features, and the bottom-up augmentation only associates the fusion of feature information between the current layer and its adjacent shallow layers. No association is made outside of adjacent feature layers, and the fused content is relatively independent. The other part is the weighted selection fusion between layers using learnable adaptive weights to obtain excellent fusion results. The AFF module combines the advantages of FPN and PANet [42], while the upward fusion of adjacent layers avoids the problem of excessive semantic information gap between multiple layers. The AFF module alleviates the loss of feature information and feature information attention dispersion. It enables the ship information to gain attention in feature maps with different scales.

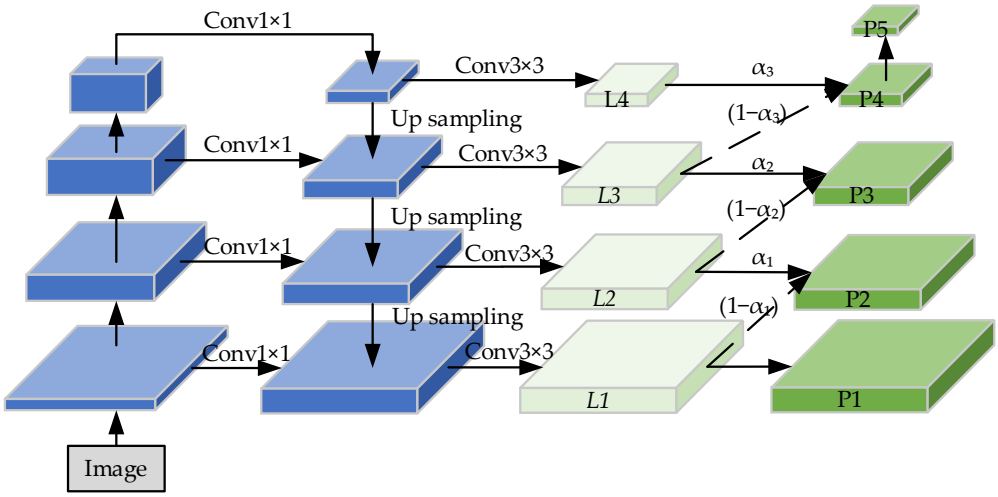

Figure 4. The AFF architecture. α1, α2, and α3 are the adaptive weights that can be learned for each layer.

Figure 4 shows that the result obtained from the initial feature pyramid can be expressed as {L1, L2, L3, L4}. For L2, L3, and L4 high-level feature mappings are fused with its proximity feature mappings, and the shallow-level features {L1, L2, L3} are expressed as {L1′, L2′, L3′} through the unified scale of downsampling. The L2, L3, and L4 layers are then used to generate the learnable weights α1, α2, and α3. Each weight is learned independently to form adaptive fusion parameters for each feature mapping. Finally, the two adjacent layers of feature mappings are multiplied by mutually opposing learnable weights respectively, and then the results are cumulated. The final feature mapping of the output results is represented by using {P1, P2, P3, P4}, and the computational structure can be written sequentially as:

$$P1 = L1 \quad (4)$$

$$P2 = \alpha 1 \times L2 + (1 - \alpha 1) \times L1' \quad (5)$$

$$P3 = \alpha 2 \times L3 + (1 - \alpha 2) \times L2' \quad (6)$$

$$P4 = \alpha 3 \times L4 + (1 - \alpha 3) \times L3' \quad (7)$$

In using the AFF module, it is important to note that the extension for adjacent fusion does not operate on L1. Second, the sum of the weights used for two adjacent feature mappings is controlled to be 1, to ensure the stability of the model training.

2.3. Architecture of ESTDNet

The detailed architecture of ESTDNet is shown in Figure 5, which is the application of FESwin and AFF modules in Cascade R-CNN Swin. Because FESwin is structurally complex, we show the network model characteristics in Table 1. FESwin is used as the backbone network for feature extraction, and the AFF module is used as the neck for feature fusion. The output of the FESwin module is the input of the AFF module, and the output of the AFF module is paired with Cascade R-CNN for category prediction and object position prediction to complete ship detection in SAR images.

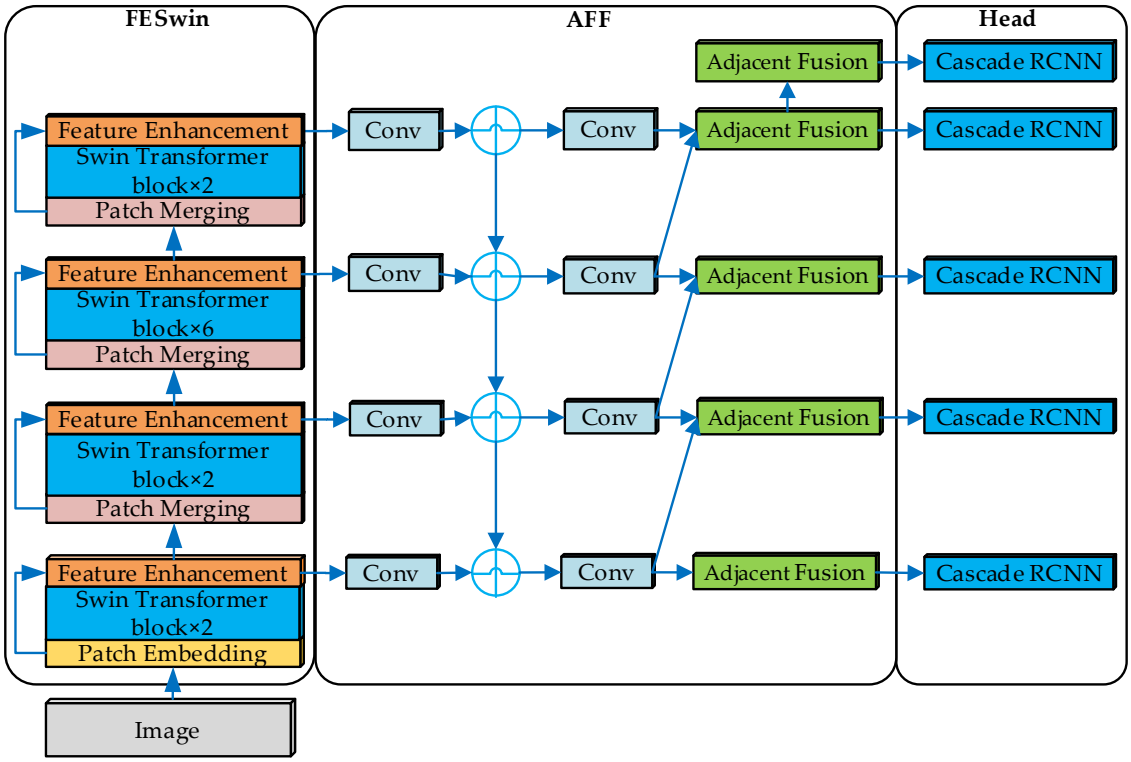

Figure 5. The architecture of ESTDNet. Using the FESwin and AFF modules is shown in detail.

Table 1. The FESwin model structure.

Layer_Name	Patch_Size	Layer		
Pretreatment	H/4 × W/4 × 48	Patch partition		
Stage1	H/4 × W/4 × C	Linear Embedding		
		Swin transformer block × 2	LayerNorm W-MSA/SW-MSA LayerNorm MLP	×2
		Feature enhancement	Conv 3 × 3 LayerNorm ReLU Conv 3 × 3 LayerNorm sigmoid	Conv 1 × 1 LayerNorm ReLU Conv 1 × 1 LayerNorm

Table 1. *Cont.*

Layer_Name	Patch_Size	Layer		
Stage2	H/8 × W/8 × 2C	PatchMerging		
		Swin transformer block × 2	LayerNorm W-MSA/SW-MSA LayerNorm MLP	×2
		Feature enhancement	Conv 3 × 3 LayerNorm ReLU Conv 3 × 3 LayerNorm sigmoid	Conv 1 × 1 LayerNorm ReLU Conv 1 × 1 LayerNorm
Stage3	H/16 × W/16 × 4C	PatchMerging		
		Swin transformer block × 6	LayerNorm W-MSA/SW-MSA LayerNorm MLP	×6
		Feature enhancement	Conv 3 × 3 LayerNorm ReLU Conv 3 × 3 LayerNorm sigmoid	Conv 1 × 1 LayerNorm ReLU Conv 1 × 1 LayerNorm
Stage4	H/32 × W/32 × 8C	PatchMerging		
		Swin transformer block × 2	LayerNorm W-MSA/SW-MSA LayerNorm MLP	×2
		Feature enhancement	Conv 3 × 3 LayerNorm ReLU Conv 3 × 3 LayerNorm sigmoid	Conv 1 × 1 LayerNorm ReLU Conv 1 × 1 LayerNorm

3. Results

In this paper, the method's effectiveness is verified by two datasets, SSDD and SAR-Ship. Our approach is also compared with other state-of-the-art object detection algorithms: Faster RCNN [14], YOLOv3 [8], Cascade R-CNN [16], PAA [43], ATSS [44], DETR [20], Deformable DETR [22], Tood [45], YOLOF [46]. All experiments were conducted using a PC with Intel(R) Xeon(R) Gold 5218 CPU@2.30 GHz × 16 and 64 GB of memory, and NVIDIA GeForce RTX 2080Ti GPU with 12 GB of memory. The operating system is a 64-bit Windows Server 2019 Standard.

3.1. Experiment Settings

Because ESTDNet is built on the Cascade R-CNN Swin, ESTDNet is an end-to-end networking model. We use the experimental result of Cascade R-CNN Swin as our baseline. The initial learning rate is set to 0.001, the optimizer is SGD, and the thresholds of NMS are set to 0.5. After statistical analysis of data sets, the input image size of the SSDD dataset was set as 672 × 672, and that of the SARShip dataset as 256 × 256. Although the image size setting of the two datasets is different, ESTDNet changes the size of the network layer as the image size grows. We use image flipping to enhance the number of samples during

the training process, which is used to improve the diversity of the training dataset. In addition, our model does not use a pre-trained model but is trained from scratch.

3.2. Experiment Datasets

We use two datasets, SSDD and SARShip, to verify the effectiveness of the proposed method in this paper. The SSDD dataset consists of 1160 images of a total of 2540 ships. The SARShip dataset has 39,729 images, consisting of a total of 50,885 ships, and is composed of 102 HSPA-3 images and 108 Sentinel 1 satellite images that have been processed and cropped to a size of 256 × 256 pixels. The resolution of the pictures is 3 m, 5 m, 8 m, and 10 m, respectively. Port terminals, offshore waters, and far seas are some of the scenes covered in the images. Ship types include tankers, cargo ships, large container ships, and small fishing boats. We randomly divide the two datasets into training and test sets in the ratio of 8:2. Next, we use the COCO dataset annotation format to process the bounding boxes and label annotations and convert the original dataset label storage file to JSON file format for storage.

In order to visualize the number of ships of different sizes in the dataset, this paper counts the ships of different sizes according to the definition of the COCO metric [47] and displays them in the form of histograms, as shown in Figure 6. Figure 6a,b shows the number of large, medium, and small ships in the SSDD and SARShip datasets.

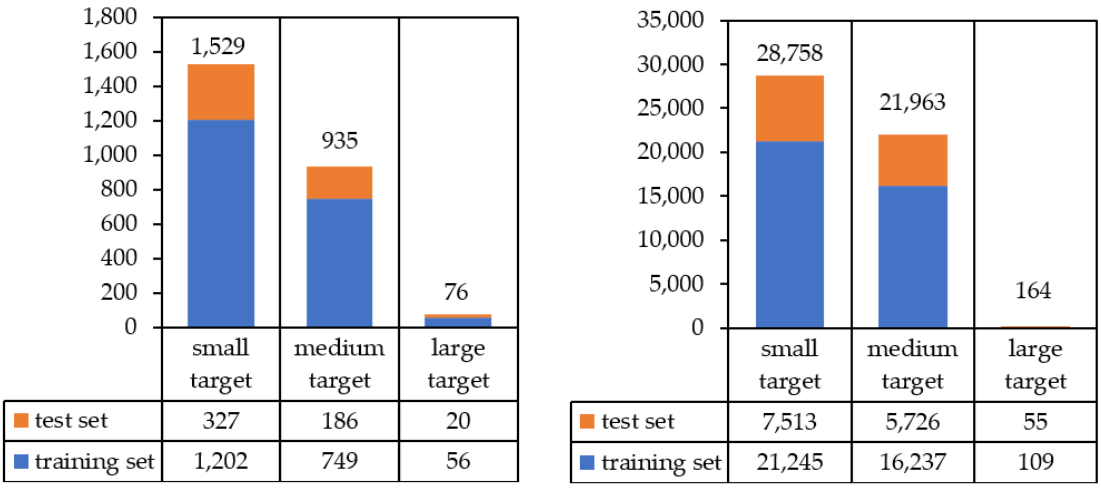

Figure 6. Statistics of different size ships in the datasets. (**a**) Statistics of the SSDD ships; (**b**) statistics of the SARShip ships.

3.3. Experiments on the SSDD Dataset

We conducted a large number of experiments on the SSDD dataset with the aforementioned configured parameter settings, and the corresponding experimental results are displayed in Table 2. The results of ESTDNet were all better than the baseline results. Among them, AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L increased by 2.8%, 2.3%, 4.4%, 1.9%, 3.2%, and 9%, respectively. Compared to Cascade R-CNN Swin, ESTDNet can effectively improve the detection performance of ships in SAR images. According to the 2.8% and 4.4% improvements in AP and AP_{75}, ESTDNet improves positioning accuracy and makes positioning more accurate in ship detection.

Table 2. The experimental result of ESTDNet on the SSDD dataset.

Methods	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)
Cascade R-CNN Swin	56.6	91.5	64.7	53.5	63.5	51.3
ESTDNet	59.4	93.8	69.1	55.4	66.7	60.3

As shown in Table 3, we performed the experiments on the SSDD dataset for some state-of-the-art object detection methods. The experimental results showed that ESTDNet obtained the best AP results of 59.4% compared to other methods. Compared to other methods, YOLOF achieved better performance on medium ships, with AP_M reaching 67.4%, 0.6% higher than ESTDNet. However, the AP_S and AP_L of small and large ships are lower than ESTDNet by 3.6% and 3.9%. In addition, ESTDNet is 6.4% higher than YOLOF on AP_{75}, indicating that ESTDNet can obtain more accurate ship position information than YOLOF. Therefore, according to the experimental results of multiple detection methods on the SSDD dataset, ESTDNet can effectively detect ships in SAR images, and can obtain more accurate ship position information, and its comprehensive detection performance surpasses other excellent methods.

Table 3. The results of different methods on the SSDD dataset.

Methods	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)
Faster RCNN	53.5	90.7	56.8	52.2	56.8	53.6
YOLOv3	57.7	93.8	64.8	54.5	63.6	60.2
Cascade R-CNN	56.9	91.8	63.3	53.4	64.8	53.1
PAA	56.0	91.6	64.0	51.1	65.7	53.1
ATSS	55.2	92.4	59.9	51.9	60.9	52.2
DETR	50.2	91.1	52.7	41.7	64.3	59.3
Deformable DETR	52.6	93.3	55.0	46.9	61.8	58.2
Tood	56.4	91.1	66.0	52.0	65.2	41.0
YOLOF	57.2	93.2	62.7	51.8	**67.4**	56.9
Cascade R-CNN Swin	56.6	91.5	64.7	53.5	63.5	51.3
ESTDNet	**59.4**	**93.8**	**69.1**	**55.4**	66.7	**60.3**

Figure 7 presents the experimental results of both the proposed and compared methods. The ground truths, detection results, missed detection results and the false detection results are indicated with green, red, yellow, and blue boxes, respectively. In order to show the detection effect of various methods more intuitively, we selected five images with complex backgrounds in the near-shore region to demonstrate the detection effect. Since the distribution of ships on the near shore is sparse compared to the distribution of ships in far-sea areas, the ships in SAR images are incredibly similar to the coastal background. Therefore, ship detection is more complicated and better reflects the performance of the detection method. As shown in Figure 7, all methods suffer from some degree of missed detection. Tood and ATSS have the highest number of missed detections and the worst performance in the near-shore region. YOLOv3, YOLOF, and DETR missed detection significantly when small and medium ships were dense. Secondly, PAA and Faster RCNN have more false detections. Because of the high similarity between the ship and the coastal background, Faster RCNN detected the background as the ship object in some images. The remaining methods all have a small number of false detections caused by overlapping detection at dense ship locations. The experimental visualization shows that the effectiveness of the detection method is somewhat compromised in near-shore areas, especially when ships are densely arranged. Especially the small and medium-sized ships near the coast are more difficult to accurately identify. Among all the algorithms, ESTDNet has more detection coverage and has the best performance when only individual images are missed and no background is falsely detected as a ship, which indicates that the overall detection performance of ESTDNet is better than other methods.

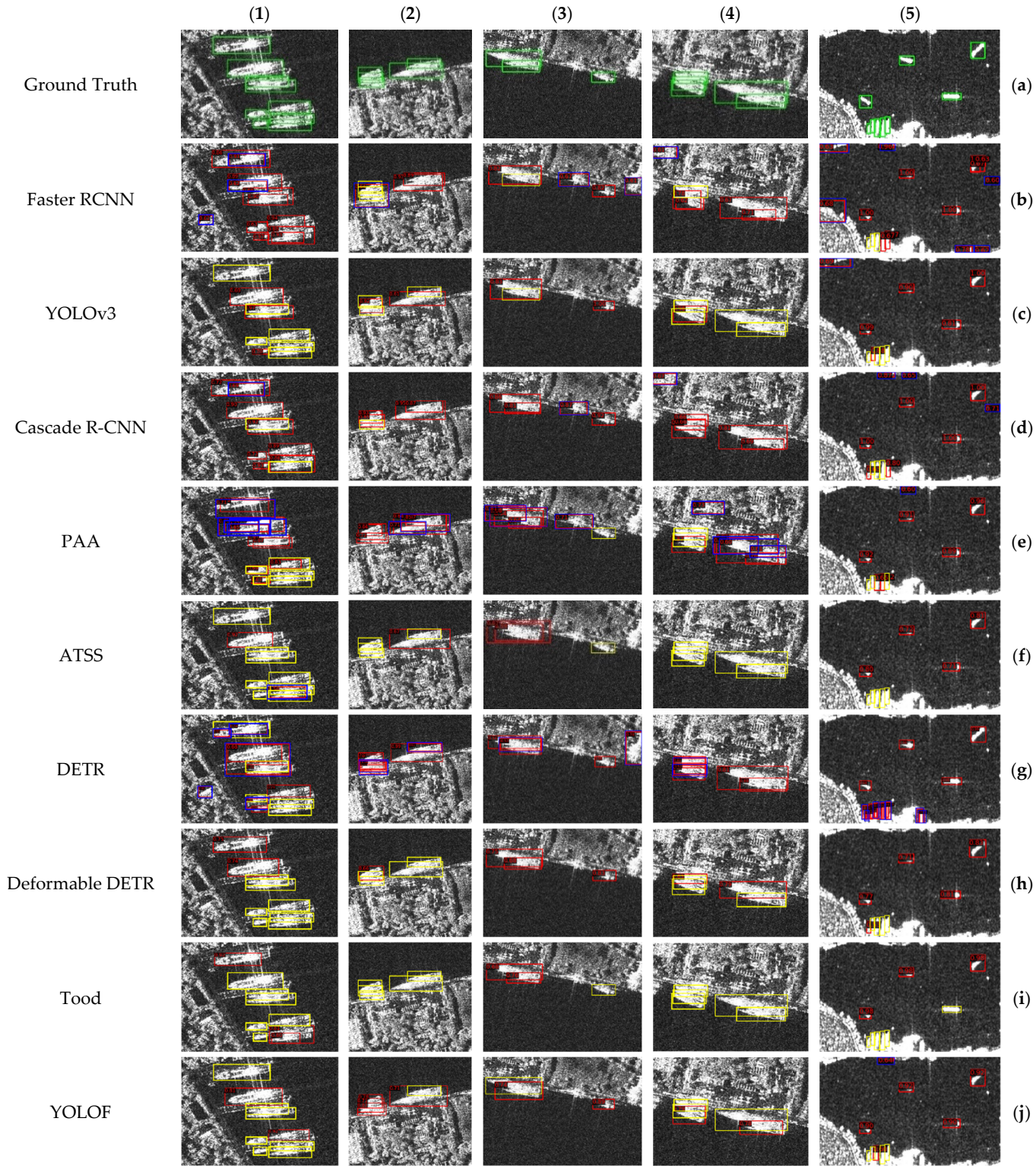

Figure 7. *Cont.*

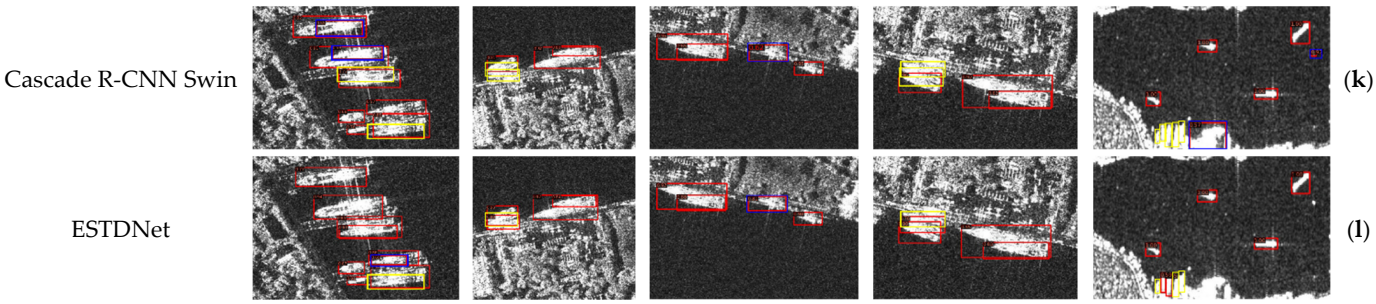

Figure 7. Experimental results for the SSDD dataset. (**a**): ground truth images; from (**b**–**l**): predicted results of Faster RCNN, YOLOv3, Cascade R-CNN, PAA, ATSS, DETR, Deformable DETR, Tood, YOLOF, Cascade R-CNN Swin, and ESTDNet, respectively. The ground truths, detection results, missed detection results and the false detection results are indicated with green, red, yellow, and blue boxes, respectively.

3.4. Experiments on the SARShip Dataset

The results of the proposed method on SARSship are shown in Table 4. The table shows that ESTDNet exceeds the Cascade R-CNN Swin baseline in all COCO metrics. Among them, ESTDNet has improved the accuracy of AP_M and AP_S by 3.7% and 2.9%, and APL possesses a considerable improvement of 11.1%, proving that ESTDNet can improve the inspection performance of different scales of ships at the same time. Next, the AP_{75} as well as AP_{50} metrics improve by 6.7% and 1.6%, indicating that the ESTDNet method is able to obtain more accurate information about the ship's position. The AP accuracy is higher than the Cascade R-CNN Swin benchmark model by 3.5%, indicating the excellent overall performance of ESTDNet. The above results validate the FESwin and AFF proposed in this paper, which make an important contribution to the extraction, fusion, and transmission of feature information in SAR images.

Table 4. The experimental result of ESTDNet on the SARShip dataset.

Methods	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)
Cascade R-CNN Swin	57.3	93.4	63.1	53.0	62.8	56.4
ESTDNet	60.8	95.0	69.8	55.9	66.5	67.5

We compare the detection performance of ESTDNet with other methods for the SAR-Ship dataset in Table 5. Among all methods, ESTDNet has the best AP metric result of 60.1%, indicating that ESTDNet has a better AP value. The index accuracy rates of AP, AP_{50}, and AP_{75} show that ESTDNet can obtain more accurate ship position information and detect more ship objects than other methods. Secondly, the numerical displays of the indicators of AP_S, AP_M, and AP_L also show that ESTDNet has a more robust detection performance for large, medium, and small ships. In conclusion, compared with other methods, ESTDNet's evaluation of the COCO performance index is relatively balanced, effectively detected most ships, and obtained more accurate detection position information.

The experimental results in terms of compared methods, which were conducted on the SARShip dataset, are illustrated in Figure 8. The ground truths, detection results, missed detection results and the false detection results are indicated with green, red, yellow, and blue boxes, respectively. To make the detection results more representative, we selected five detection images of objects at different scales with different backgrounds. From the results, in the near-shore large and medium ship detection, most methods can detect ships, but there are many false detections. The Faster RCNN, PAA, and ATSS detection methods have a large number of false detections. The Tood, YOLOv3, Deformable DETR, and Cascade R-CNN Swin methods have missed detections. DETR, YOLOF, and ESTDNet have excellent detection results. Because of the relatively dense distribution of small ships, it is difficult to detect all the ships. The compared detection methods all have a certain degree

of missed detections, with Tood, ATSS, and Deformable DETR detection methods having the most serious missed detections. Compared with other detection methods, ESTDNet has a relatively low number of false and missed detections, and the prediction accuracy per ship is higher than most detection methods, indicating the superiority of ESTDNet.

Table 5. The results of different methods on the SARShip dataset.

Methods	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)
Faster RCNN	50.8	92.7	50.5	47.2	55.4	46.9
YOLOv3	46.6	90.9	42.6	42.8	52.1	43.1
Cascade R-CNN	58.1	93.4	65.1	53.7	63.8	57.7
PAA	53.6	93.2	56.1	49.5	58.7	50.5
ATSS	53.7	93.5	56.2	49.4	59.2	52.7
DETR	56.5	94.5	62.7	49.1	65.2	64.2
Deformable DETR	56.8	94.2	63.3	50.2	64.1	52.2
Tood	57.7	94.4	64.1	53.2	63.4	66.1
YOLOF	54.4	94.7	56.3	48.2	62.2	54.0
Cascade R-CNN Swin	57.3	93.4	63.1	53.0	62.8	56.4
ESTDNet	**60.8**	**95.0**	**69.8**	**55.9**	**66.5**	**67.5**

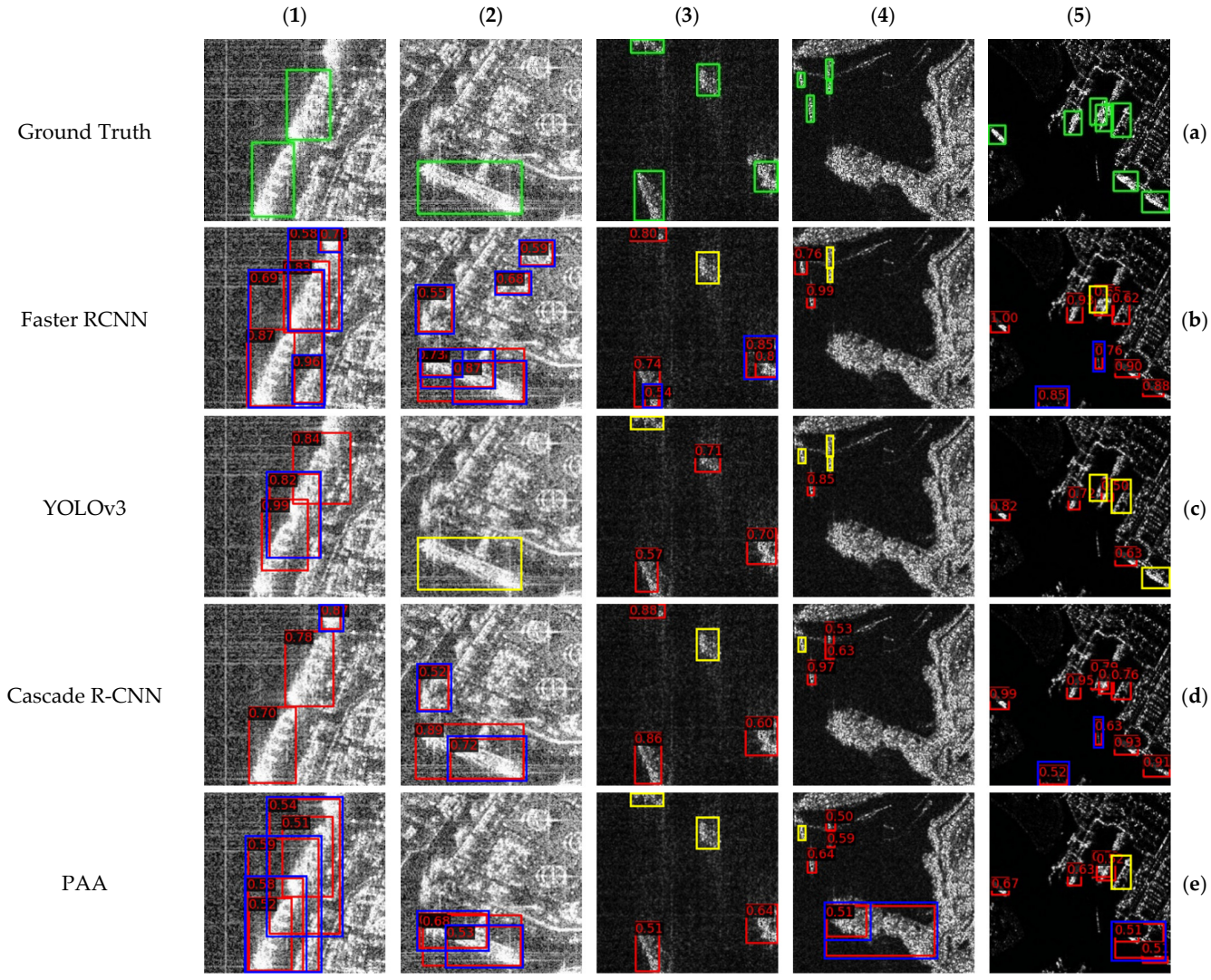

Figure 8. *Cont.*

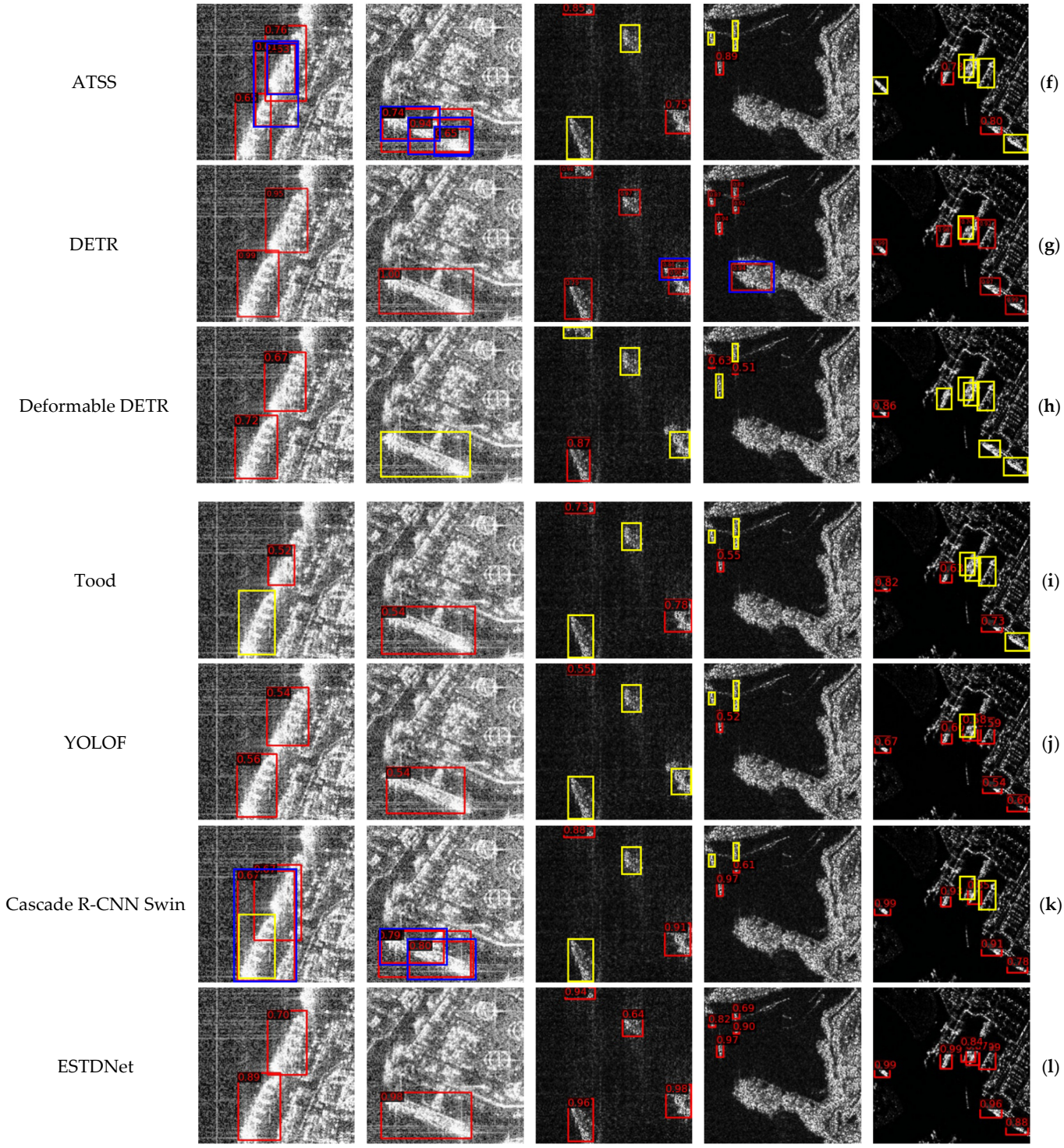

Figure 8. Experimental results in the SARShip dataset. (**a**): ground truth images; from (**b**–**l**): predicted results of Faster RCNN, YOLOv3, Cascade R-CNN, PAA, ATSS, DETR, Deformable DETR, Tood, YOLOF, Cascade R-CNN Swin, and ESTDNet, respectively. The ground truths, detection results, missed detection results and the false detection results are indicated with green, red, yellow, and blue boxes, respectively.

3.5. Ablation Experiments

The ablation experiments of ESTDNet, as exhibited in Table 6, are performed on the SSDD dataset, with the Cascade R-CNN acting as the baseline. We use the FESwin module and AFF module for the ablation study. The detection performance metrics of ESTDNet

were all effectively improved using the FESwin module, and the improvement was even more pronounced for large ships compared to the Cascade R-CNN Swin. Compared with Cascade R-CNN Swin, the detection performance of the ESTDNet with only the AFF module can acquire significant improvement with regard to the large ships, while there is a slight enhancement in detection performance for the other. In addition, in order to ensure the stability of the experimental results. We conducted 20 experiments on ESTDNet and recorded the average detection accuracy of multiple experiments in Table 6. The values in parentheses indicate the standard deviation of the results of multiple experiments.

Table 6. The experimental result of ESTDNet on the SSDD dataset.

Methods	FESwin	AFF	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)
Cascade R-CNN Swin			56.6	91.5	64.7	53.5	63.5	51.3
ESTDNet	√		58.8	93.8	68.3	54.6	66.6	58.9
		√	57.5	92.3	66.2	54.3	64.3	54.2
	√	√	59.4 (0.1%)	93.8 (0.3%)	69.1 (0.2%)	55.4 (0.3%)	66.7 (0.2%)	60.3 (0.4%)

In this paper, FESwin is the ESTDNet backbone network module with superior feature extraction capability, due to the fact that both global-local and spatial-channel characteristics are comprehensively considered. As shown in Table 6, the detection performance of large, medium, and small ships is improved using the FESwin module compared to the Cascade R-CNN Swin. Among them, the most significant improvement in large ship inspection AP_L increased by 7.6%, and the AP_S and AP_M of small and medium-sized ship inspection also increased by 1.1% and 3.1%. In addition, the targeting is more precise, with AP, AP_{50}, and AP_{75} all boasting 2.2%, 2.3%, and 3.6% improvements. Using the AFF module compared with Cascade R-CNN Swin increases the information mobility between the underlying features and the higher-level features, with 0.9%, 0.8%, 1.5%, 0.8%, and 0.8% improvement in AP, AP_{50}, AP_{75}, AP_S, and AP_M metrics, and 2.9% improvement in AP_L for large ship objects. The full ESTDNet brings a 2.8%, 2.3%, 4.4%, 1.9%, 3.2%, and 9% improvement in each metric compared to Cascade R-CNN Swin. The above phenomenon shows that the use of the FESwin module has an important contribution to the detection accuracy improvement of ESTDNet, and the combination with the AFF module has improved the detection accuracy to different degrees without depleting the improvement effect of the FESwin module. Therefore, both the FESwin module and the AFF module of ESTDNet can improve the model detection performance, and the combination of the two can yield a more robust ship detection model for SAR images.

3.6. Comparison of Inference Time

We compare the inference time of ESTDNet with other methods. Table 7 shows that the inference time of ESTDNet is higher than the baseline Cascade R-CNN Swin, with about nine milliseconds more inference per image and 1.5 images per second less processing. This is because ESTDNet increases some computation modules on the Cascade R-CNN Swin baseline for better accuracy. In addition, ESTDNet is 13.5 milliseconds faster than the inference time of Deformable DETR on the SSDD dataset and 11 milliseconds slower than the inference time of Deformable DETR on the SARShip dataset. The inference speed of CNN's methods is slightly higher than transformer's methods, and ESTDNet is comparable to other the inference times of transformer methods are similar.

Table 7. Comparison of methods inference time.

Methods	FPS (Image/Seconds)		Inference Time (Milliseconds/Image)	
	SSDD	**SARShip**	**SSDD**	**SARShip**
Faster RCNN	16.9	21	59.1	47.6
YOLOv3	47.4	41	21.1	24.4
Cascade R-CNN	13.7	15.7	73.1	63.7
PAA	13.2	12.9	75.6	77.8
ATSS	18.9	23.2	53	43.1
DETR	15.7	15.8	63.7	63.4
Deformable DETR	10.6	13.1	94.8	76.2
Tood	19.1	16.3	52.4	61.4
YOLOF	29	37.2	34.5	26.9
Cascade R-CNN Swin	13.7	12.8	73	78.1
ESTDNet	12.3	11.5	81.3	86.9

4. Discussion

We verify the superiority of ESTDNet by conducting several experiments using the SSDD and SARship datasets. The ablation experiments of FESwin and AFF modules on the SSDD dataset have proved that each of them can improve ship detection performance, and the combination of both can achieve better detection results. In order to observe more intuitively the enhancement effect of the two modules, we show the output of the two modules separately with heat maps.

4.1. FESwin Module Effect Validation

In this paper, we compare the effect of feature extraction between FESwin and Swin transformers, and use a heat map for verification. Figure 9 visualizes the results of feature extraction at four different scales for Swin transformer and FESwin, respectively. The more highlighted color in the heat map indicates the more feature attention received in the feature map. In order to illustrate the effect more cleanly, we selected six images of large, medium, and small ship objects including far sea and near shore. As shown in Figure 9, FESwin enhances the feature information that originally existed only in the first stage for medium and small ships, solving the problem of losing feature information as the model deepens. FESwin makes it possible for medium and small ship objects to have ship feature information of interest in at least stages one, two, and three, providing multiple scales of information support for detection. For large ship objects, FESwin expands the feature information focus that initially existed only in phases 1 and 4 to feature information focus in all four phases. Second, near-shore images are more difficult to detect than far-sea images. When the Swin transformer performs feature extraction, there are many false concerns as the depth of the model increases because of the extensive similarity between the object and the background. FESwin dramatically eases these problems and reduces the need for erroneous focus in the coastal context. In summary, the heat map shows that FESwin optimizes the performance of the feature extraction model, brings the ship into focus in more multi-scale feature maps, enhances the utilization of ship information in SAR images, and improves the feature representation capability of the feature extraction model.

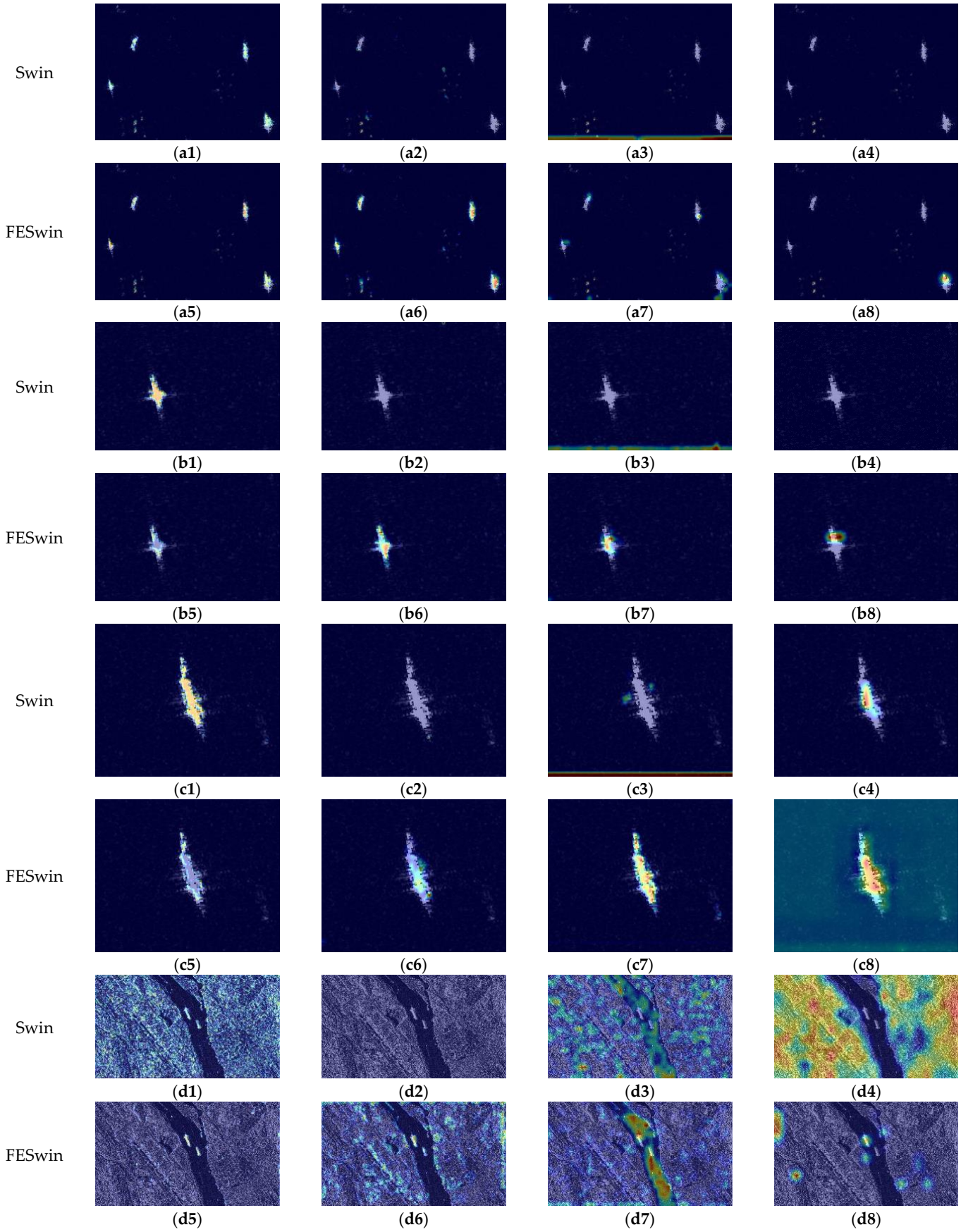

Figure 9. *Cont.*

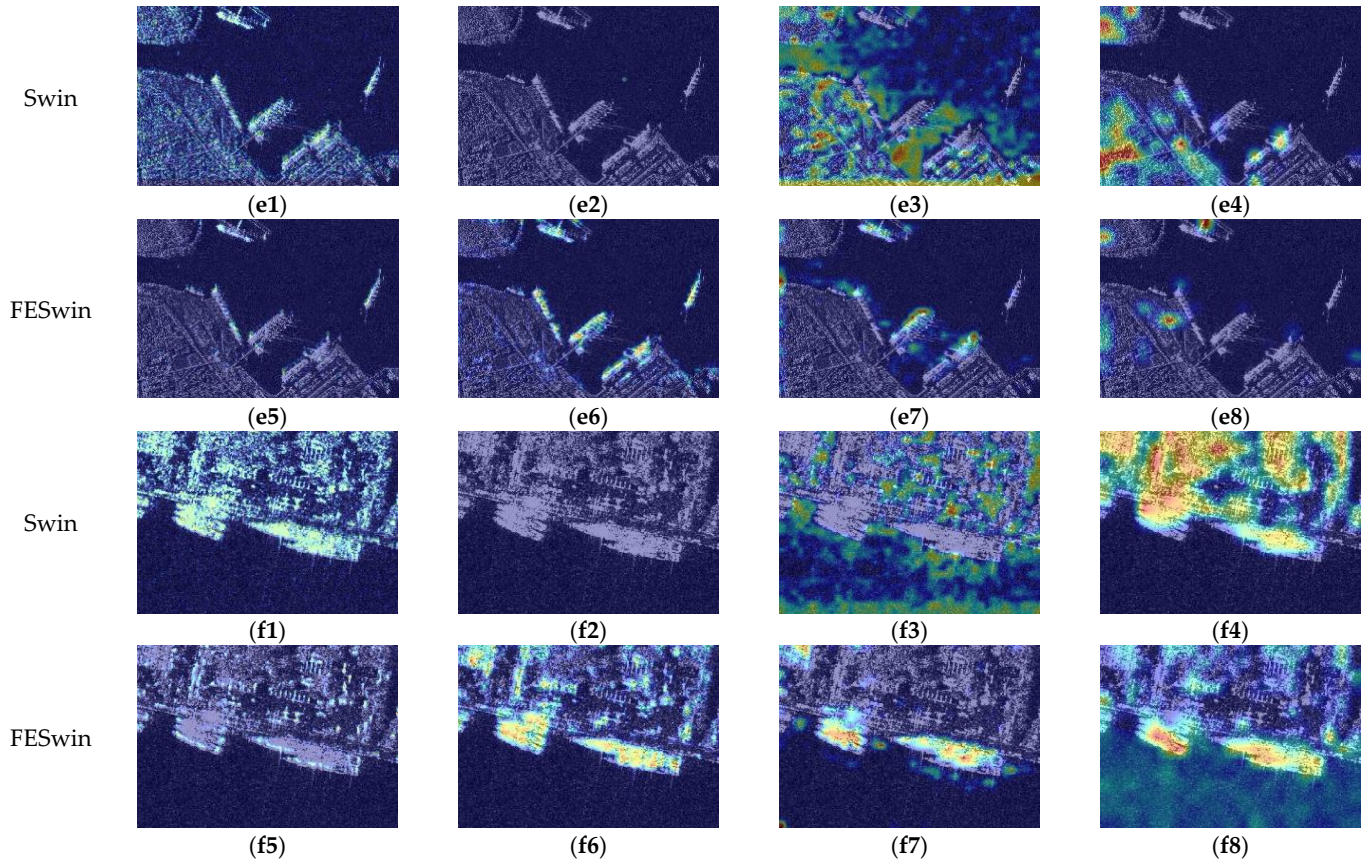

Figure 9. Results of FESwin effect comparison. (**a**–**c**): images of small, medium, and large ships in the far-sea region; (**d**–**f**): images of small, medium, and large ships in the near-shore region. (**1**–**4**): feature extraction results of four stages of Swin transformer; (**5**–**8**): feature extraction results of four stages of FESwin.

4.2. AFF Module Effect Validation

For the AFF effect, this paper uses the heat map for verification. Figure 10 visualizes the results of the five-scale feature fusion for FPN and AFF, respectively. The more highlighted color in the heat map indicates the more feature attention received in the feature map. In order to illustrate the effect more cleanly, we selected six images of large, medium, and small ship objects including far-sea and near-shore areas. The Swin transformer is used as the backbone network to ensure the fairness of the experiment. From the results in Figure 10, the useful feature information concerns in the FPN when detecting far-sea images are only present in the fourth and fifth layers, and the feature information mainly comes from the feature extraction in the fourth stage. FPN suffers from long feature information transmission paths, excessive information gaps between high and low layers, and attention scattering in the first, second, and third layers after feature fusion, resulting in the inability to provide effective ship feature information. AFF uses adjacent lower-level features to complement the higher-level features, so that attention to ship feature information appears in all the second, third, fourth, and fifth-level feature maps, effectively alleviating the problem of attention dispersion. Secondly, the degree of attention received by the ship is consistent with the scale distribution of the ship, which is beneficial to ship detection. Because the near-shore background is complex, ship detection in near-shore is not friendly to the feature extraction model. When using FPN for multi-scale feature fusion, there are many incorrect feature focus sites, and there is also the problem that the practical feature focus information is concentrated in four or five layers. AFF performs adjacency fusion of feature maps from bottom to top, reducing the focus on relative errors in each layer, avoiding the focus on coastal background feature information, and focusing attention

effectively. Taken together, AFF enhances the information flow between feature maps at each scale, reduces the information difference between the bottom features and the top features, and alleviates the problem of attention scattering in multiple scales. A multi-scale feature map with richer feature information is obtained using AFF, allowing ship information in SAR images to be used effectively.

We can see from the results in Figures 9 and 10 that the increase in background complexity and ship density impacts the detection results. In the future, we will consider combining geometric features to construct a ship detection network with a stronger feature representation to solve the background complexity problem. On the other hand, ship objects in SAR images show the multi-angle distribution and no overlap between objects, and we will try to use a rotate anchor detection method to solve the problem better.

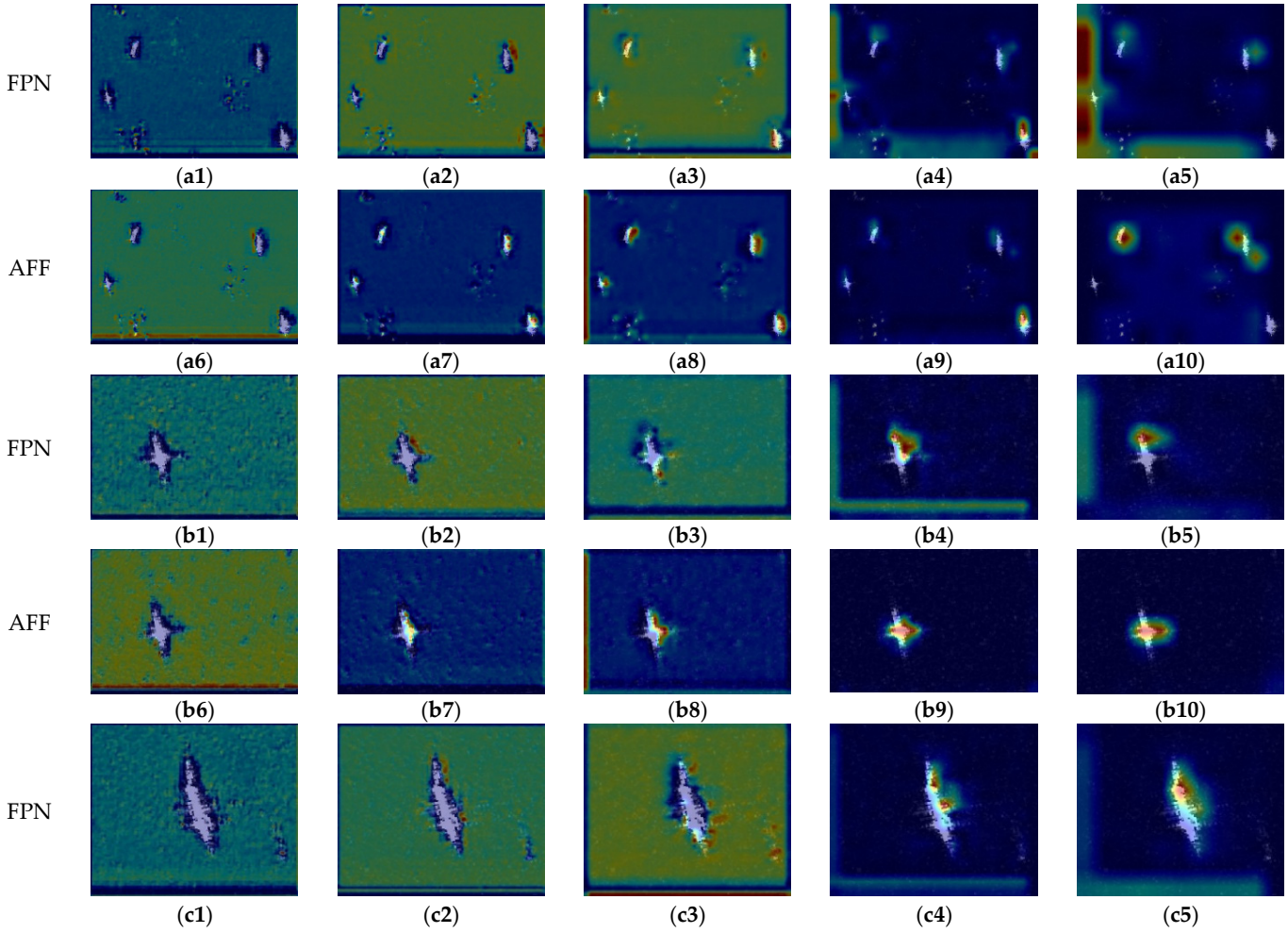

Figure 10. *Cont.*

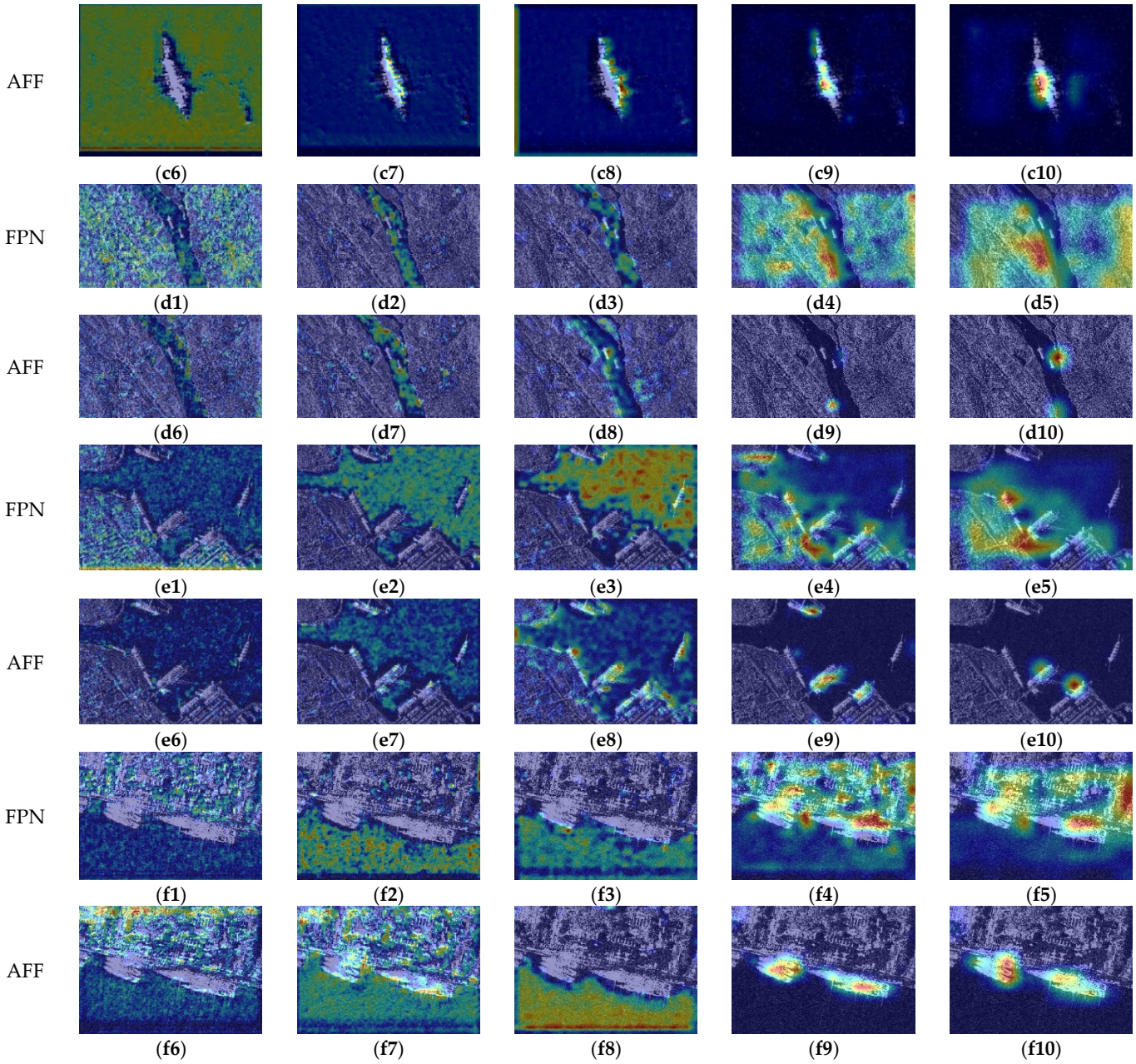

Figure 10. Results of AFF effect comparison. (**a**–**c**): images of small, medium, and large ships in the far-sea region; (**d**–**f**): images of small, medium, and large ships in the near-shore region. (**1**–**5**): feature fusion results of five scales of FPN; (**6**–**10**): feature fusion results of five scales of AFF.

5. Conclusions

In this paper, we propose ESTDNet for ship detection of SAR images. FESwin and AFF are essential components of ESTDNet, where the FESwin module is responsible for the feature extraction work of the images to obtain more feature information. The AFF module is more beneficial for fusing the extracted ship feature information. We use ablation experiments to confirm the effectiveness of these two modules. The ESTDNet based on FESwin and AFF can improve the accuracy of ship detection in SAR images. Moreover, we conduct experiments on the SSDD and SARShip datasets. The results reveal that ESTDNet achieves higher other detection performance than other methods and is a superior ship detection method in SAR images. This is of great importance in aviation, aerospace, military, and civil fields.

ESTDNet is a combination of transformer and CNN detection methods. In the future, our research will consider reducing the computational complexity caused by the transformer model. In addition, we will investigate the light-weighting of the transformer model.

Author Contributions: K.L., M.Z. and H.W. provided the ideas; K.L., M.X., R.T. and L.W. implemented this algorithm; K.L. wrote this paper; M.Z. and H.W. revised this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 12003018), Fundamental Research Funds for the Central Universities (No. XJS191305) and China Postdoctoral Science Foundation (No. 2018M633471).

Data Availability Statement: The public datasets are used in this study, no new data are created or analyzed. Data sharing is not applicable to this article.

Acknowledgments: We thank the authors of SARShip and SSDD for providing the experimental datasets, and the authors of Swin Transformer and the comparison method. Our experiments were supported by thr High-performance Computing Platform of XiDian University.

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

References

1. Fan, Y.; Wang, F.; Wang, H. A Transformer-Based Coarse-to-Fine Wide-Swath SAR Image Registration Method under Weak Texture Conditions. *Remote Sens.* **2022**, *14*, 1175. [CrossRef]
2. Zhang, X.; Wang, H.; Xu, C.; Lv, Y.; Fu, C.; Xiao, H.; He, Y. A Lightweight Feature Optimizing Network for Ship Detection in SAR Image. *IEEE Access* **2019**, *7*, 141662–141678. [CrossRef]
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
4. Qian, X.; Cheng, X.; Cheng, G.; Yao, X.; Jiang, L. Two-Stream Encoder GAN With Progressive Training for Co-Saliency Detection. *IEEE Signal Process. Lett.* **2021**, *28*, 180–184. [CrossRef]
5. Lin, S.; Zhang, M.; Cheng, X.; Wang, L.; Xu, M.; Wang, H. Hyperspectral Anomaly Detection via Dual Dictionaries Construction Guided by Two-Stage Complementary Decision. *Remote Sens.* **2022**, *14*, 1784. [CrossRef]
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
7. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
8. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
10. 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; pp. 21–37. [CrossRef]
11. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
13. 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.
14. 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]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
16. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
17. 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, 21–26 July 2017; pp. 936–944.
18. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.

20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: New York, NY, USA, 2020; pp. 213–229.
21. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
22. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
23. Dai, Z.; Cai, B.; Lin, Y.; Chen, J. UP-DETR: Unsupervised Pre-Training for Object Detection with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1601–1610.
24. Wang, J.; Lu, C.; Jiang, W. Simultaneous Ship Detection and Orientation Estimation in SAR Images Based on Attention Module and Angle Regression. *Sensors* **2018**, *18*, 2851. [CrossRef] [PubMed]
25. Chang, Y.-L.; Anagaw, A.; Chang, L.; Wang, Y.; Hsiao, C.-Y.; Lee, W.-H. Ship Detection Based on YOLOv2 for SAR Imagery. *Remote Sens.* **2019**, *11*, 786. [CrossRef]
26. Qian, X.; Lin, S.; Cheng, G.; Yao, X.; Ren, H.; Wang, W. Object Detection in Remote Sensing Images Based on Improved Bounding Box Regression and Multi-Level Features Fusion. *Remote Sens.* **2020**, *12*, 143. [CrossRef]
27. Su, N.; He, J.; Yan, Y.; Zhao, C.; Xing, X. SII-Net: Spatial Information Integration Network for Small Target Detection in SAR Images. *Remote Sens.* **2022**, *14*, 442. [CrossRef]
28. Li, J.; Qu, C.; Shao, J. Ship Detection in SAR Images Based on an Improved Faster R-CNN. In Proceedings of the SAR in Big Data Era (BIGSARDATA), Beijing, China, 13–14 November 2017.
29. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]
30. Zhang, T.; Zhang, X. High-Speed Ship Detection in SAR Images Based on a Grid Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 1206. [CrossRef]
31. Zhou, K.; Zhang, M.; Wang, H.; Tan, J. Ship Detection in SAR Images Based on Multi-Scale Feature Extraction and Adaptive Feature Fusion. *Remote Sens.* **2022**, *14*, 755. [CrossRef]
32. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A Novel Quad Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
33. 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]
34. Xia, R.; Chen, J.; Huang, Z.; Wan, H.; Wu, B.; Sun, L.; Yao, B.; Xiang, H.; Xing, M. CRTransSar: A Visual Transformer Based on Contextual Joint Representation Learning for SAR Ship Detection. *Remote Sens.* **2022**, *14*, 1488. [CrossRef]
35. Qu, H.; Shen, L.; Guo, W.; Wang, J. Ships Detection in SAR Images Based on Anchor-Free Model With Mask Guidance Features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 666–675. [CrossRef]
36. Feng, Y.; Chen, J.; Huang, Z.; Wan, H.; Xia, R.; Wu, B.; Sun, L.; Xing, M. A Lightweight Position-Enhanced Anchor-Free Algorithm for SAR Ship Detection. *Remote Sens.* **2022**, *14*, 1908. [CrossRef]
37. Rostami, M.; Kolouri, S.; Eaton, E.; Kim, K. Deep Transfer Learning for Few-Shot SAR Image Classification. *Remote Sens.* **2019**, *11*, 1374. [CrossRef]
38. Hao, P.; He, M. Ship Detection Based on Small Sample Learning. *J. Coast. Res.* **2020**, *108*, 135–139. [CrossRef]
39. Zhang, H.; Zhang, X.; Meng, G.; Guo, C.; Jiang, Z. Few-Shot Multi-Class Ship Detection in Remote Sensing Images Using Attention Feature Map and Multi-Relation Detector. *Remote Sens.* **2022**, *14*, 2790. [CrossRef]
40. Zhang, Z.; Zhoa, J.; Liang, X. Zero-shot Learning Based on Semantic Embedding for Ship Detection. In Proceedings of the 2020 3rd International Conference on Unmanned Systems (ICUS), Harbin, China, 27–28 November 2020; pp. 1152–1156.
41. Zhang, X.; Zhang, H.; Jiang, Z. Few shot object detection in remote sensing images. In *Image and Signal Processing for Remote Sensing XXVII*; Bruzzone, L., Bovolo, F., Eds.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2021; Volume 11862, pp. 76–81. [CrossRef]
42. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
43. Kim, K.; Lee, H.S. Probabilistic Anchor Assignment with oU Prediction for Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 355–371.
44. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9759–9768.
45. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-Aligned One-Stage Object Detection. In Proceedings of the 2021 IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada; 2021; pp. 3490–3499.

46. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-Level Feature. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13039–13048.
47. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.