remote sensing

MDPI

Article

LPST-Det: Local-Perception-Enhanced Swin Transformer for SAR Ship Detection

Zhigang Yang [1], Xiangyu Xia [1], Yiming Liu [1], Guiwei Wen [1], Wei Emma Zhang [2] and Limin Guo [1,*]

[1] Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China; zgyang@hrbeu.edu.cn (Z.Y.); xiaxiangyu@hrbeu.edu.cn (X.X.); lym2018080415@hrbeu.edu.cn (Y.L.); 846699275@hrbeu.edu.cn (G.W.)

[2] School of Computer Science, University of Adelaide, Adelaide 5005, Australia

* Correspondence: guolimin@hrbeu.edu.cn

Abstract: Convolutional neural networks (CNNs) and transformers have boosted the rapid growth of object detection in synthetic aperture radar (SAR) images. However, it is still a challenging task because SAR images usually have the characteristics of unclear contour, sidelobe interference, speckle noise, multiple scales, complex inshore background, etc. More effective feature extraction by the backbone and augmentation in the neck will bring a promising performance increment. In response, we make full use of the advantage of CNNs in extracting local features and the advantage of transformers in capturing long-range dependencies to propose a Swin Transformer-based detector for arbitrary-oriented SAR ship detection. Firstly, we incorporate a convolution-based local perception unit (CLPU) into the transformer structure to establish a powerful backbone. The local-perception-enhanced Swin Transformer (LP-Swin) backbone combines the local information perception ability of CNNs and the global feature extraction ability of transformers to enhance representation learning, which can extract object features more effectively and boost the detection performance. Then, we devise a cross-scale bidirectional feature pyramid network (CS-BiFPN) by strengthening the propagation and integration of both location and semantic information. It allows for more effective utilization of the feature extracted by the backbone and mitigates the problem of multi-scale ships. Moreover, we design a one-stage framework integrated with LP-Swin, CS-BiFPN, and the detection head of R^3Det for arbitrary-oriented object detection, which can provide more precise locations for inclined objects and introduce less background information. On the SAR Ship Detection Dataset (SSDD), ablation studies are implemented to verify the effectiveness of each component, and competing experiments illustrate that our detector attains 93.31% in mean average precision (mAP), which is a comparable detection performance with other advanced detectors.

Keywords: local perception unit; feature pyramid network; swin transformer; ship detection; SAR object detection

Citation: Yang, Z.; Xia, X.; Liu, Y.; Wen, G.; Zhang, W.E.; Guo, L. LPST-Det: Local-Perception-Enhanced Swin Transformer for SAR Ship Detection. *Remote Sens.* **2024**, *16*, 483. https://doi.org/10.3390/rs16030483

Academic Editor: Domenico Velotto

Received: 24 December 2023
Revised: 19 January 2024
Accepted: 22 January 2024
Published: 26 January 2024

1. Introduction

Remote sensing observations provided by the Synthetic Aperture Radar (SAR) [1] are almost unaffected by light and weather conditions. In comparison to other types of satellite reconnaissance, SAR as an active microwave sensor is better suited for detecting marine ships in environments with rapidly changing conditions [2]. SAR ship detection, whose aim is to find specific ship objects and mark their locations from a satellite perspective, is of significant value in civilian and military fields, like maritime rescuing and marine traffic management [3]. SAR ship detection has obtained considerable concerns among scholars and has emerged as a significant research emphasis in recent years [4].

Extensive works have been investigated with regard to SAR object detection. Generally, deep-learning-based methods [5] get better detection performance than traditional

manual-feature-based methods [6,7] owing to the significant growth of convolutional neural networks (CNNs) [8]. Most CNN-based methods are divided into two-stage and one-stage methods. In the two-stage framework, region proposals are first generated and then a head network performs regression and classification, such as Faster R-CNN [9], and Mask R-CNN [10]. While a one-stage detector directly predicts the category and regresses the location coordinates without generating proposal regions, such as SSD [11], RetinaNet [12], and YOLO [13–16] series. Moreover, the attention-based transformer has recently been introduced into computer vision, demonstrating significant potential in fundamental tasks like object detection and image classification [17].

Although significant efforts have been made, SAR object detection still faces many challenges. Due to the distinct imaging mechanism, SAR images exhibit some unique characteristics such as unclear contour, discontinuous texture, sidelobe interference, speckle noise, and complex background (especially in the case of inshore) [18], which pose significant obstacles to SAR object detection. Since SAR targets are more indistinctive than objects in optical remote sensing images, we are aiming to enhance the SAR ship detection performance in three key aspects, as listed below:

1. A strong backbone is necessary to extract object features more effectively.
2. After feature extraction by the backbone, the integration and propagation of both semantic and location information in the feature pyramid network (FPN) need to be strengthened for more effective feature representation.
3. Oriented bounding boxes are preferred to horizontal ones in order to provide more precise locations for arbitrary-oriented objects and reduce the introduction of background information.

As for the first aspect, we design a transformer-based backbone integrated with CNN modules, which makes full use of the CNN's ability to extract local features and the Swin Transformer's ability in capturing long-range dependencies [19]. Although the Swin Transformer acquires excellent performance in optical object detection, challenges still exist when applying it in SAR ship detection. Swin Transformer does well in capturing long-range dependencies but exhibits a limited capability in local feature extraction, which is crucial for detecting small SAR ships. Considering CNN's ability in local feature extraction, we design a convolution-based local perception unit (CLPU) comprising dilated convolution [20] and depth-wise convolution [21]. Attributed to these techniques, the convolution-based local perception unit reinforces local feature extraction and enhances the representation learning of local relations. We insert the convolution-based local perception unit before the Swin Transformer block, forming a hybrid network named the local-perception-enhanced Swin Transformer (LP-Swin).

As for the second aspect, enhancing feature representation in the feature pyramid improves the utilization of feature extracted from the backbone, so we introduce a cross-scale bidirectional feature pyramid network (CS-BiFPN) based on BiFPN [22]. In contrast to the standard FPN [23] structure, BiFPN integrates top-down and bottom-up pathways for the fusion of multi-scale features. Additionally, BiFPN introduces learnable weights, enabling the network to learn the significance of different input features and focus on those that have a significant contribution to the output features. By incorporating additional longitudinal cross-scale connections to the existing BiFPN network, the integration and propagation of both semantic and location information can be strengthened, and CS-BiFPN achieves a more comprehensive representation of multi-scale features.

As for the third aspect, in order to implement detection with oriented bounding boxes and reduce the introduction of background information, we utilize R^3Det [24] as the baseline, which is an effective one-stage detector for arbitrary-oriented objects. R^3Det incorporates a novel module in order to address feature misalignment issues that existed in refined one-stage detectors. R^3Det achieves superior extraction speed and accuracy.

Based on the above three improvements, we propose a transformer-based detector called LPST-Det, which is a one-stage approach built on the architecture of R^3Det. We first design a convolution-based local perception unit to introduce CNN's ability of local

feature extraction into the Swin Transformer, composing a hybrid backbone named local-perception-enhanced Swin Transformer (LP-Swin). Then, we devise a CS-BiFPN to obtain more powerful feature representations. Finally, we apply our LPST-Det to a public SAR dataset and further analyze the accuracy gain of LP-Swin and CS-BiFPN in the overall performance. The main contributions of this paper are as follows:

1. We propose a local-perception-enhanced Swin Transformer backbone called LP-Swin, which combines the advantage of CNN in collecting local information and the advantage of Swin Transformer to extract long-distance dependencies, so that more powerful object features can be extracted from SAR images. Specifically, we introduce a convolution-based local perception unit termed CLPU consisting of dilated convolution and depth-wise convolution to facilitate the extraction of local information in the vision transformer structure and improve the overall detection performance.
2. We introduce a cross-scale bidirectional feature pyramid network termed CS-BiFPN with the aim of enhancing the utilization of features extracted by the backbone and mitigating the challenges posed by multi-scale ships. The incorporation of longitudinal cross-scale connections strengthens the integration and propagation of both semantic and location information, which benefits both classification and location tasks.
3. We construct a one-stage framework integrated with LP-Swin and CS-BiFPN for arbitrary-oriented SAR ship detection. Ablation tests are conducted to assess the effect of our proposed components. Experiments on the public SSDD dataset illustrate that our approach attains comparable detection results with other advanced strategies.

2. Literature Review

2.1. CNN-Based SAR Object Detection

CNN has demonstrated significant success in SAR object detection. Li et al. [25] enhanced the Faster R-CNN by hard negative mining and transfer learning for SAR ship detection. The authors of [26] introduced a G-CNN to enhance the efficiency of the detector by combining the depth-wise separable convolution. Jiao et al. designed a DCMSNN [27] to solve multi-scale and multi-scene problems. Xu et al. [28] employed network pruning to introduce a lightweight architecture termed Lite-YOLOv5, which excels in on-board ship detection. Xu et al. [29] designed the GWFEF-Net, utilizing dual-polarization features for ship detection, and achieved superior detection performance. The authors of [30] presented a unique approach based on YOLOv7 for SAR instance segmentation. Zheng et al. [31] incorporated the MobileNetV1 backbone into the YOLOv4 algorithm to devise a lightweight network for ship recognition.

However, the above methods all adopt horizontal bounding boxes to regress the location of objects, which cannot accurately annotate the objects. Therefore, oriented bounding boxes are more suitable for SAR object detection. MSR2N, proposed by [32], employs oriented bounding boxes to reduce background information interference in ship detection. Wang et al. [33] replaced the horizontal bounding box with the rotating bounding boxes on the SSD framework for the localization and detection of ships in a single stage. An et al. [34] proposed DRBox-v2 for rotating SAR target detection, which combines a multi-layer rotational bounding box generation mechanism, a modified encoding scheme, focal loss, and hard negative mining. The authors of [35] constructed a rotation detector, which combines both horizontal and rotating anchors to address scenes with dense distribution. In our study, the R^3Det framework is adopted to detect ships with oriented bounding boxes.

The above methods all adopt CNN as the backbone to capture features, but it is hard for the CNN to learn the long-distance dependencies owing to the constrained receptive field of an individual convolution. Although the CNN's global feature extraction ability can be enhanced by multiple stacking convolutional layers, it will lead to a much deeper network with numerous parameters, experiencing difficulty in converging [36]. In our study, the transformer network is employed to augment the global feature extraction.

2.2. Transformer in Object Detection

Transformer architecture has become prevalent in recent years. The transformer is good at collecting global features and establishing long-range dependencies because of its self-attention mechanism. Due to its significant success in the NLP field, scholars have attempted to incorporate a transformer network into computer vision, yielding several notable works. DETR [37] is the first transformer architecture applied to object detection. DETR combined CNN and transformer encoder-decoder structure to build an end-to-end detector, avoiding many hand-designed components. Deformable DETR [38] presented a deformable attention mechanism to address issues related to restricted spatial resolution of features and slow convergence.

Applying transformers to computer vision has two main research lines. One is the hybrid structure that combines CNNs and transformers, the other is the pure transformer architecture. The conformer [39] designed a concurrent structure to combine the CNN with visual transformer, which achieves enhanced representation learning. CMT [40] took advantage of both transformer and the CNN to promote its feature extraction ability. As a pioneering research, the Vision Transformer (ViT) [17] directly employed the standard transformer for the image recognition task. ViT simply treats an image as non-overlapping image patches, achieving encouraging results in image classification. However, because of the quadratic complexity to image size, it is difficult to directly apply ViT to downstream tasks like semantic segmentation and object detection, whose inputs are high-resolution images. To address the above issues, Swin Transformer was proposed as a general-purpose network. Notably, its linear complexity enhances the adaptation to downstream vision tasks.

Researchers are actively exploring the Swin Transformer's utilization in the remote sensing field. CRTransSar [41] innovatively proposed a vision transformer combined with CNN architecture on the basis of contextual joint-representation learning. Shi et al. [42] presented a novel transformer, which mainly replaces the original attention mechanism with a deformable attention mechanism. ESTDNet [43] proposed a FESwin to strengthen the capture of SAR ship features. Ke et al. [44] have implemented a SAR ship detection algorithm on the basis of the Swin Transformer, which leverages the capabilities of the transformer to model long-range dependencies. Xu et al. [45] combined the Swin Transformer with convolutional layers that comprise dilated convolution to take advantage of the CNN and the Swin Transformer. Inspired by CMT and Xu's work, we incorporate dilated convolution and depth-wise convolution to strengthen the extraction of local information. Moreover, we implement detailed experiments on the dilation rate, analyzing its influence on the transformer backbone.

2.3. Multi-Scale Feature Representations

Since FPN can effectively augment the features extracted by the backbone, it holds great significance in object detection. As a fashionable work, FPN [23] introduced a top-down connection path to utilize multi-scale features, enabling more accurate object detection. The original FPN has performance bottlenecks due to the single top-down fusion pathway, so PANet [46] added an additional bottom-up path aggregation network to solve this issue. However, there are only adjacent-scale feature fusion connections in the above two structures. NAS-FPN [47] designed a better cross-scale feature pyramid structure by employing neural architecture search technology, but it is difficult to explain the irregular network. Moreover, BiFPN [22] repeatedly applied top-down and bottom-up architecture and introduced learnable weights to help the model control the contribution of different input features.

In SAR object detection, some studies are focusing on the improvement of feature pyramid network. As an illustration, Liu et al. [48] introduced a scale-transfer pyramid network by incorporating lateral connections and establishing connections between multi-scale features. Hu et al. [49] introduced a DenseFPN that employs distinct methods for processing shallow and deep features. Quad-FPN [50], which consists of four novel FPNs, was proposed to address the small ship detection and multi-scale feature problems. Unlike

conventional approaches, its innovative structural design not only achieves excellent detection performance but also provides novel ideas for researchers. Chen et al. [51] designed a multi-scale fusion detector to detect small targets in a road scene. The authors of [52] devised a DFANet to extract and fuse dual-modality features. In this study, to overcome the limitation that there are only transverse cross-scale connections in the BiFPN, we introduce additional longitudinal cross-scale connections to improve the extraction of location information.

3. Materials and Methods

This section introduces LPST-Det, a SAR ship detector built on the basis of vision transformer. Initially, we present the overall framework of the LPST-Det. Subsequently, we provide a detailed elaboration of the two essential components of LPST-Det, including the local-perception-enhanced Swin Transformer backbone and the cross-scale bidirectional feature pyramid network.

3.1. Overview of the Proposed Method

The workflow of our LPST-Det is illustrated in Figure 1. LPST-Det is built on the R^{3}Det framework, an advanced one-stage model for inclined object detection. The proposed detector comprises three components, including LP-Swin, CS-BiFPN, and R^{3}Det-Head. Initial feature maps are extracted by the LP-Swin backbone, and then CS-BiFPN augments them into powerful multi-scale feature representations. Subsequently, the R^{3}Det-Head performs regression and classification, producing the final results.

To enhance the extraction of both local perception and global representation, we devise the LP-Swin backbone, which combines the global feature capture capability of the Swin Transformer and the local feature extraction ability of the CNN. Furthermore, to obtain a more effective feature representation, we design the CS-BiFPN, which enhances the original BiFPN by introducing longitudinal cross-scale connections. These longitudinal cross-scale connections can facilitate the transmission and integration of both semantic information and location information.

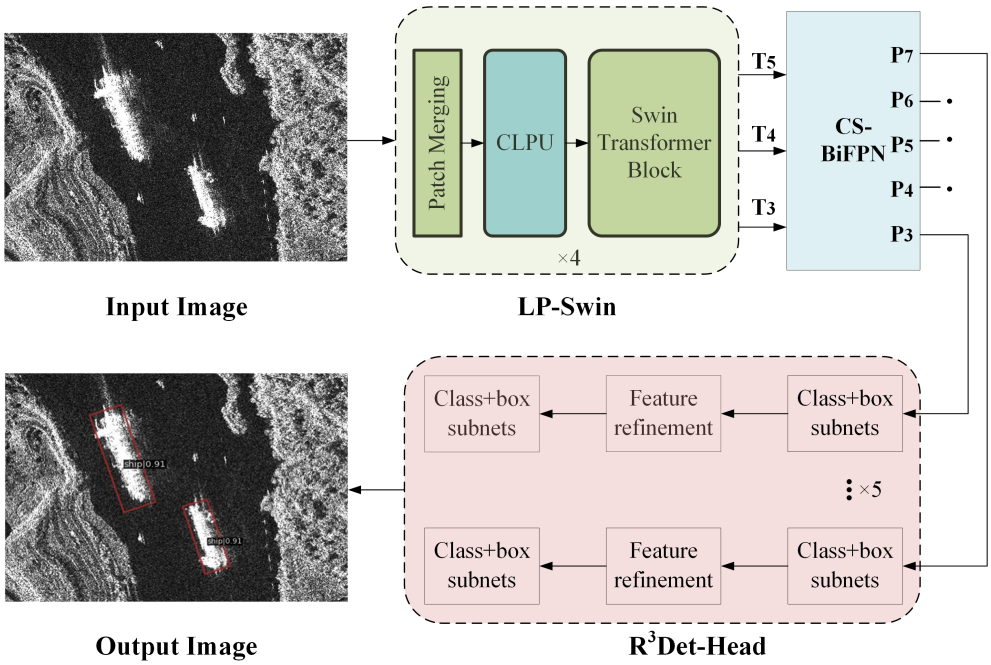


Figure 1. Workflow of the proposed LPST-Det.

To accurately annotate the object, we utilize the detection head of the R^{3}Det algorithm to implement detection with oriented bounding boxes. R^{3}Det employs horizontal anchors to improve speed and the recall rate in the first step, while refined rotation anchors are applied

in dense scenes during the refinement step. Moreover, to weaken feature misalignment caused by bounding box location changes, the feature refinement module (FRM) performs position information re-encoding corresponding to the refined bounding box. This process achieves feature map reconstruction and enables precise target detection.

LP-Swin and CS-BiFPN serve as two main components of our method, and a comprehensive description of the backbone and neck will follow in the subsequent section. One can refer to [24] for details of other parts in this detection framework.

3.2. Local-Perception-Enhanced Swin Transformer Backbone

3.2.1. Structure of Local-Perception-Enhanced Swin Transformer

We introduce a local-perception-enhanced Swin Transformer backbone called LP-Swin, incorporating the strengths of both vision transformer and CNN architecture. While the CNN architecture excels at extracting local features, it exhibits disadvantages in capturing long-distance feature dependencies due to convolution operations. The vision transformer is expert in capturing long-distance feature dependencies but misses local features due to the self-attention mechanisms. The proposed LP-Swin backbone can combine the advantages of the self-attention mechanism and convolution operation to enhance representation learning.

Figure 2 presents the LP-Swin's overall structure. Swin Transformer, specifically the Swin-T version, is chosen as the basic four-stage architecture because of its similar hierarchy and similar complexity to ResNet-50 [53] for fair comparison. Then, we redesign the architecture of the Swin Transformer by adding CNN-based modules to strengthen the ability of extracting local feature. Inspired by CMT [40] and Xu's work [45], we adopt dilated convolution and depth-wise convolution to design a convolution-based local perception unit called CLPU and insert it before each Swin Transformer block to enrich local information. For clarity, T2, T3, T4, and T5 are defined as the output features of four stages.

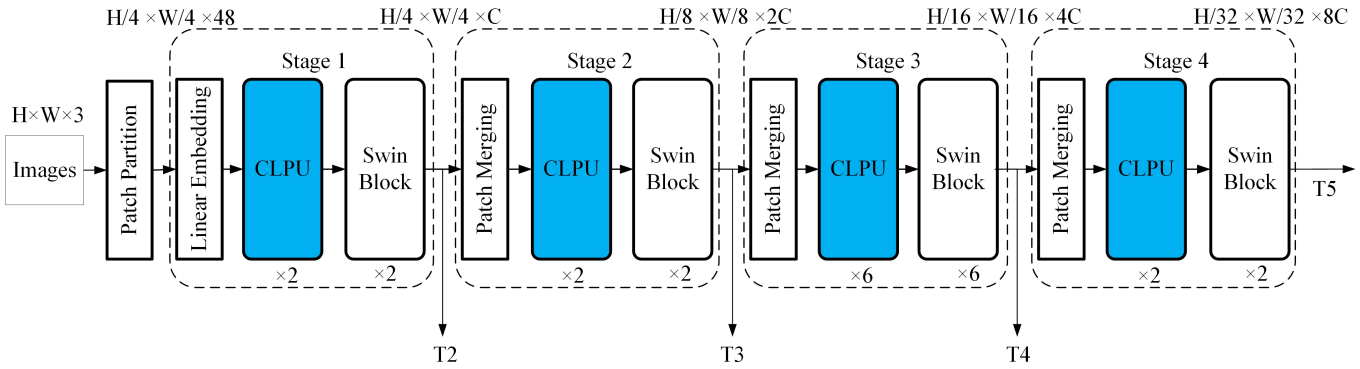


Figure 2. Overall structure of the LP-Swin backbone. (Swin block is short for Swin Transformer block).

Next, we will describe the principle and workflow of the Swin Transformer block in Section 3.2.2, while more details about CLPU will be left to Section 3.2.3.

3.2.2. Swin Transformer Block

Figure 3 illustrates two Swin Transformer blocks, consisting of a window MSA (W-MSA) module, a shifted window MSA (SW-MSA) module, MLP modules with a GELU activation function, LayerNorm (LN) functions, and residual connections. MSA calculation in the standard Transformer leads to quadratic complexity, while the W-MSA in Swin Transformer performs self-attention within a local window rather than the entire area, achieving linear complexity, as follows:

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C \quad (1)$$

$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC \quad (2)$$

Although calculation within local windows can reduce W-MSA's complexity, these windows cannot connect with each other resulting in model performance degradation. Cross-window connections can be achieved by SW-MSA. Therefore, W-MSA alternates with SW-MSA in two successive blocks.

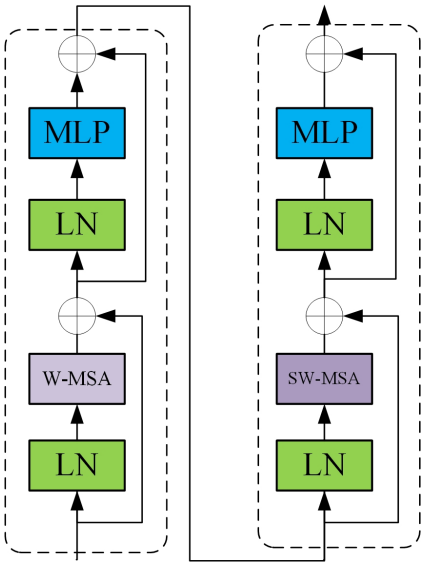


Figure 3. Swin Transformer blocks [19].

Figure 4 illustrates the shifted window approach. W-MSA partitions the 8 × 8 feature map into 2 × 2 windows of size 4 × 4 (M = 4). Subsequently, SW-MSA generates 3 × 3 non-overlapping windows from the 2 × 2 windows by displaying them with (M/2, M/2) pixels. The shifted window method enhances the cross-window exchange of information and also increases the receptive field.

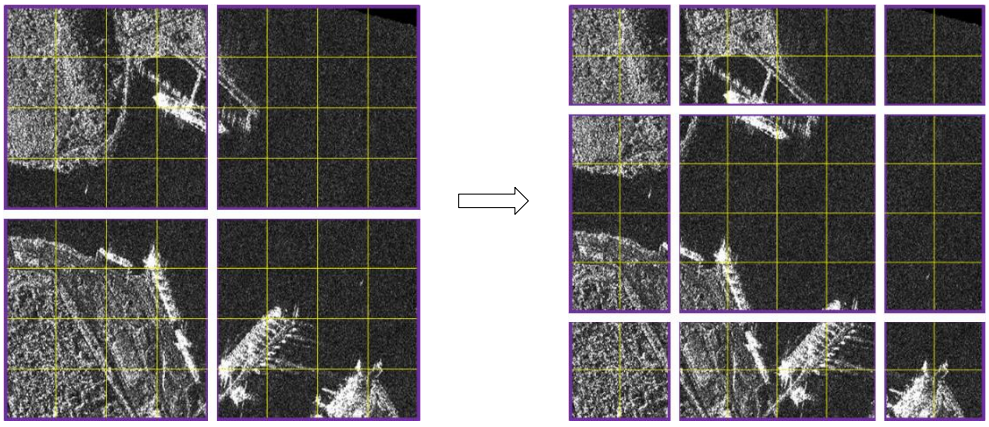


Figure 4. Shifted window in Swin Transformer.

3.2.3. Convolution-based Local Perception Unit

Vision transformer is good at capturing long-range feature dependencies but easily misses the local relation and the structure information inside the image patch due to the self-attention mechanisms. We introduce a convolution-based local perception unit termed the CLPU to facilitate the extraction of local information in the vision transformer structure. The equation below clarifies the CLPU:

$$CLPU(X) = DConv(DWConv(X)) + X \quad (3)$$

where X is the input, DWConv denotes depth-wise convolution, and DConv represents dilated convolution. To reinforce the model's representation learning capability, a CLPU is integrated before each Swin Transformer block.

Figure 5 illustrates our CLPU's composition, comprising a 3 × 3 depth-wise convolution, a 3 × 3 dilated convolution, a GELU activation function, and a residual connection. In the CLPU module, the depth-wise convolution is employed to extract local features. Subsequently, the dilated convolution increases the receptive field, facilitating the extraction of context information surrounding the ship target. Finally, the GELU activation function and the residual connection improve the learning ability of this convolutional module.

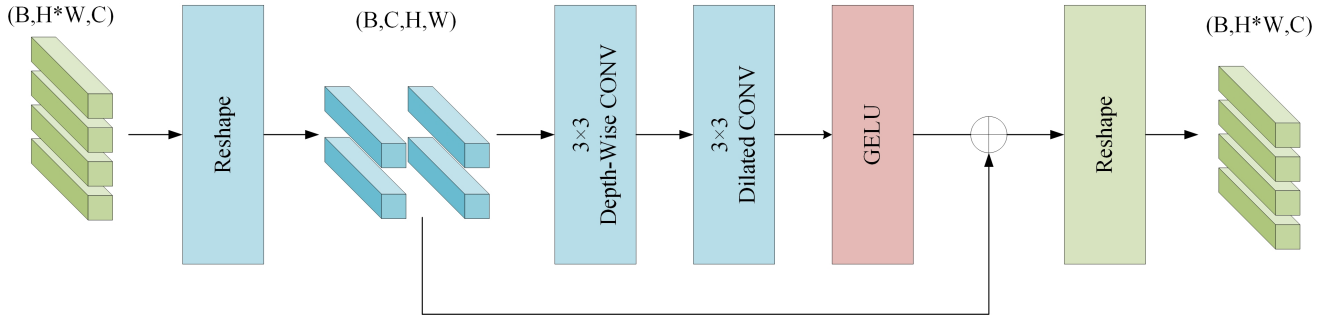


Figure 5. Structure of CLPU.

Each input channel undergoes a single filter operation in depth-wise convolution. Every single convolution kernel conducts convolution operations on an input channel to obtain the corresponding output channel. This approach makes the depth-wise convolution highly efficient, as it significantly lowers the computational cost and reduces the model size in comparison to standard convolution.

By adjusting the dilation rate (defined as d), dilated convolution enlarges the receptive field. When d = 1, the dilated convolution degenerates into standard convolution. In SAR ship detection, the utilization of dilated convolution increases the receptive field, thereby effectively capturing the contextual information around the ship target. Figure 6 gives intuitive illustrations of 3 × 3 dilated convolution with d = 1, 2, and 3, respectively. In Section 4.3.1, we determine that d = 3 is the optimum parameter in our model based on the ablation experiments.

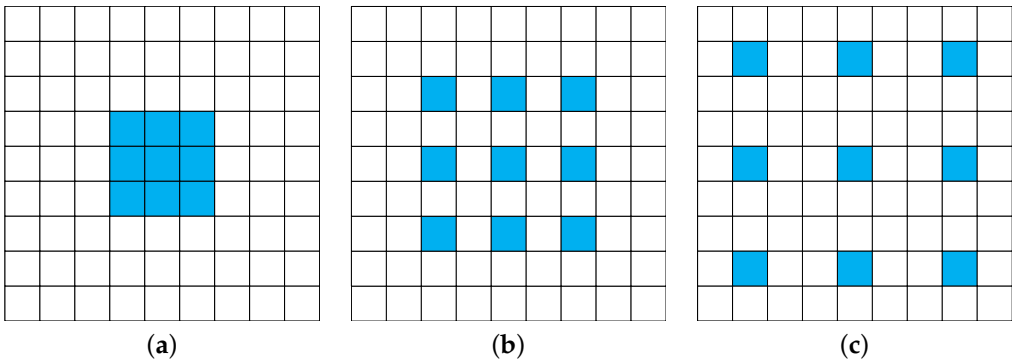


Figure 6. Illustrations of dilated convolution under different dilation rates: (**a**) d = 1; (**b**) d = 2; (**c**) d = 3.

3.3. Cross-Scale Bi-Directional Feature Pyramid Network

We devise CS-BiFPN, a cross-scale bi-directional feature pyramid network. This design is motivated by the structure design of NAS-FPN [47] and BiFPN [22]. Our innovative network aims to augment the extraction of both multi-scale semantic information and position information. In the original BiFPN, there are only transverse cross-scale connections

between the input and the output, illustrated in Figure 7a. We introduce additional longitudinal cross-scale connections to enhance the transmission and integration of both semantic information and location information, benefiting both classification and location tasks.

We add two kinds of longitudinal cross-scale connections, including two top-down and three bottom-up cross-scale connections, as shown in Figure 7b. T_3, T_4, T_5, T_6, and T_7 are input feature maps, where T_3, T_4, and T_5 are extracted by the LP-Swin backbone (as illustrated in Figure 2), T_6 is derived from a 3×3 stride-2 convolution on T_5, and T_7 is obtained in a similar way on T_6. Then, through multiple transverse and longitudinal connections, they are combined in various resolutions to construct a feature pyramid P_3, P_4, P_5, P_6, P_7. To enhance clarity, we illustrate the implementation process using P_6 as an example.

$$P_6' = Conv(T_6, Resize(T_7)) \tag{4}$$

$$P_6 = Conv(T_6, P_6', Resize(P_4), Resize(P_5)) \tag{5}$$

where P_6' is the intermediate feature on the top-down pathway, and P_6 is the output feature on the bottom-up pathway. In the first step (i.e., top-down), T_7 is upsampled by a factor of 2 and then fused with T_6. Subsequently, the fused feature goes through a convolution process, generating the intermediate feature P_6'. In the second step (i.e., bottom-up), P_4 and P_5 are downsampled by a factor of 4 and 2, respectively. Then P_6 is obtained by fusing T_6, P_6' and the two downsampled features. Subsequent feature maps are constructed following a similar approach.

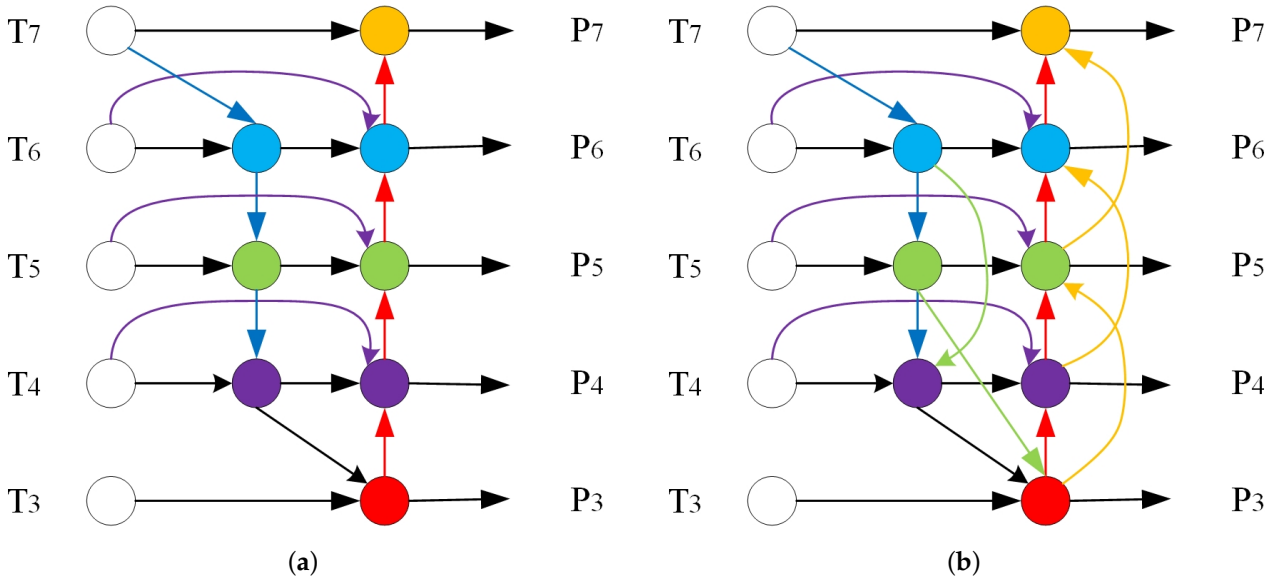


Figure 7. Comparison of different feature pyramid networks. (**a**) The original BiFPN; (**b**) the proposed CS-BiFPN.

3.4. Evaluation Metrics

We adopt the precision rate (P), recall rate (R), and mean average precision (mAP) to quantitatively assess the performances of SAR ship detectors. The precision and the recall are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

where TP denotes the amount of accurately recognized ship targets, FP indicates the amount of false alarms and FN represents the amount of missed ship targets. Average precision (AP) is calculated by the area under the Precision-Recall (PR) curve, as follows:

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

We employ the AP to comprehensively assess the models' performances under different Intersection over Union thresholds, with 0.5 being the default. In our detection tasks, the mAP remains equivalent to the AP as the task involves a single object category. Furthermore, we calculate the mAP for both inshore and offshore test sets to assess our method's performance in different SAR scenarios, referred to as Inshore and Offshore, respectively.

4. Results and Discussion

We evaluate our model through ablation studies and comparative experiments on a public SAR ship dataset. First, we give a brief introduction to the benchmark dataset and elaborate on the details of the implementation. Subsequently, we validate the competitiveness of our LPST-Det by comparing its performance with several advanced approaches. Next, ablation tests are performed to assess the performance of two essential parts: the LP-Swin backbone and the CS-BiFPN neck. Finally, we describe the limitations of our LPST-Det in the discussion section, analyzing reasons and providing potential solutions.

4.1. Dataset and Implementation Details

Within the SAR image community, the SAR Ship Detection Dataset (SSDD) [54] serves as the first publicly available dataset. It comprises diverse images with resolutions varying from 1 to 15 m, captured by different sensors such as RadarSat-2, TerraSAR-X, and Sentinel-1, derived from various complex scenes including both inshore and offshore scenarios. There are 2456 ships in total among 1160 images in SSDD. The sizes of the images vary from 214 $\times$ 160 to 668 $\times$ 526. Moreover, it comprises several typical hard-detected samples, such as ships under severe speckle noise, densely distributed small ships, and densely parallel ships at ports, which pose challenges in the detection task. The SSDD dataset provides three types of annotations, including a horizontal bounding box, rotatable bounding box, and polygon segmentation. To achieve a more precise localization of ship targets, we utilize the rotatable bounding box annotations and employ the R^3Det detector for ship detection. The training dataset versus the test dataset ratio is established as 8:2 in accordance with the protocol specified in [54].

We adopt the AdamW optimizer to train our model for 72 epochs, employing a batch size of 8. A weight decay of 0.05 is adopted and the initial learning rate is set as 0.0001. We implement the warm-up strategy for 500 iterations, and the learning rate undergoes a 10-fold reduction at each decay step. Moreover, we apply LoadImageFromFile, LoadAnnotations, and RRandomFlip strategies to enhance the data. In our experiments, the input images are resized into 672 $\times$ 672. In addition, we use the Swin-T pretraining model in our study.

4.2. Experimental Results on the SSDD

We integrate the proposed LP-Swin backbone and the CS-BiFPN neck with the R^3Det head to design our LPST-Det. Table 1 reports the comparison results between LPST-Det and other advanced methods.

Table 1. Performance comparison with different advanced methods. Items marked with * mean that the data come from the original paper.

Method	Stage	mAP_{50}	Inshore AP	Offshore AP	FPS
KeyShip * [55]	-	90.72	-	-	13.3
Cascade RCNN * [56]	Multiple	88.45	-	-	2.8
ROI Transformer [57]	Multiple	90.51	75.11	96.40	-
MSR2N * [32]	Two	90.11	-	-	9.7
Gliding Vertex [58]	Two	91.88	75.23	98.35	-
CSL [59]	Two	92.16	76.15	98.87	7.0
SCRDet++ [60]	Two	92.56	77.17	99.16	8.8
RetinaNet-R [12]	One	86.14	59.35	97.10	30.6
R^3Det [24]	One	90.45	76.36	97.21	24.5
DRBox-v2 * [34]	One	92.81	-	-	18.1
S^2A-Net [61]	One	92.08	75.79	98.79	29.0
Jiang's * [62]	One	92.50	-	-	-
LPST-Det	One	**93.85**	**81.28**	**99.20**	16.4

We carry out experiments on both inshore and offshore test datasets to analyze the performance of LPST-Det in various scenarios. Our LPST-Det attains Inshore AP, Offshore AP, and mAP of 81.28%, 99.20%, and 93.85%, respectively. We conduct a comparative analysis between LPST-Det and existing two-stage, one-stage, and SAR object detection methods. In comparison to the advanced two-stage method SCRDet++ and one-stage method S2A-Net [61], our approach achieves a 1.29% and 1.77% improvement in mAP, respectively. And our LPST-Det outperforms other SAR target detectors such as DRBox-v2, KeyShip [55] and Jiang's method [62]. Additionally, compared with the original R^3Det method, LPST-Det improves the performance of inshore and offshore scenario by 4.92% and 1.99%, which proves the effectiveness of our proposed LP-Swin and CS-BiFPN in diverse complex scenes.

Figure 8 presents some visualized results of ships in the SSDD offshore dataset. For SAR ships in simple offshore scene, due to the relatively minor background information interference, LPST-Det achieves both accurate and stable detection, resulting in a quite high accuracy in the SSDD offshore dataset.

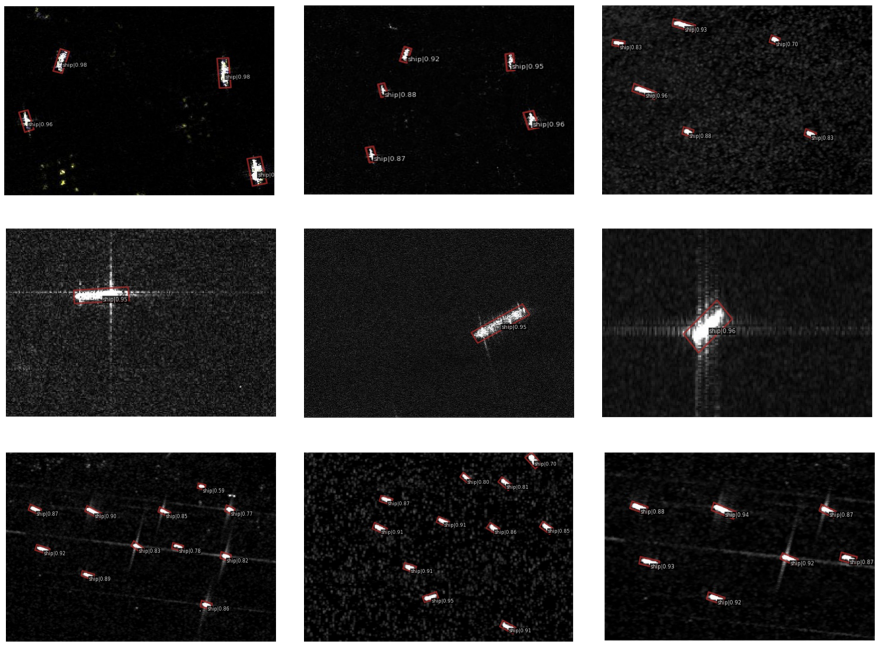


Figure 8. Detection results of SSDD offshore ship. The red oriented boxes indicate correct detection outcomes.

Figure 9 exhibits several visualized detection results of inshore ships. It is evident that LPST-Det mitigates background interference arising from the port and accurately locates ships near the port. Both the detection precision and the visualized results verify the effectiveness of LPST-Det in offshore and inshore scenes.

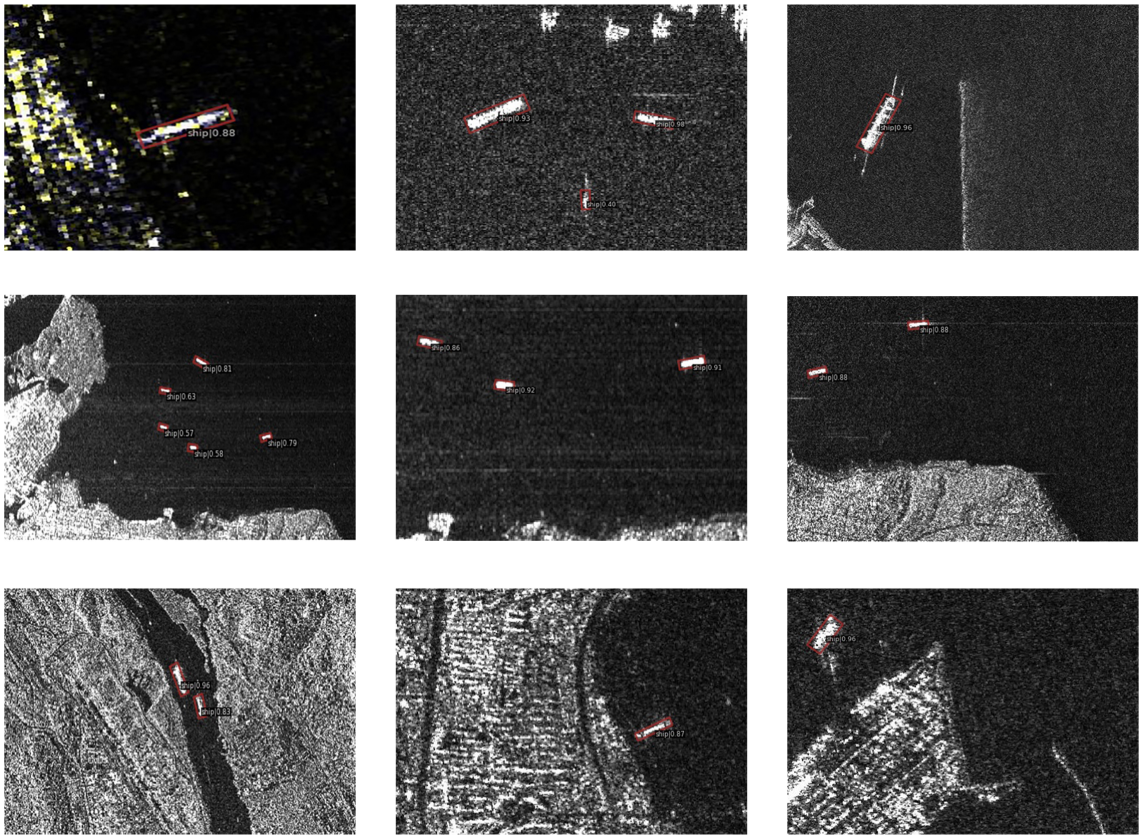


Figure 9. Detection results of SSDD inshore ship. The red oriented boxes indicate correct detection outcomes.

4.3. Ablation Experiments

4.3.1. Experiments for the Dilation Rate in CLPU

As mentioned in Section 3.2.3, we have introduced the CLPU module to augment the vision transformer's local feature extraction. The dilation rate (d) serves as a key hyper-parameter in the CLPU because it controls the receptive field. In other words, it determines how much contextual information can be captured by the dilated convolution. To determine the ideal value of d, experiments are conducted by varying d from 1 to 4. Table 2 provides the comparative results.

Compared to the standard 3 × 3 convolution (i.e., d = 1), dilated convolutions with d = 2, 3, and 4 yield gains of 0.2%, 0.39%, and 0.08% in mAP, respectively, while the recall rate is almost the same. It proves that the technique of expanding the receptive field is beneficial for local feature extraction. The highest recall rate of 0.944 and the best mAP of 93.34% indicate that d = 3 is the optimal value, so we use this criterion in the CLPU of the LP-Swin backbone for the subsequent experiments.

Table 2. Experiments for the dilated convolution in CLPU.

dilation Rate	mAP_{50}	Inshore AP	Offshore AP	Precision	Recall
d = 1	92.95	79.96	97.98	95.5	94.2
d = 2	93.15	80.90	98.05	96.1	94.2
d = 3	**93.34**	**80.95**	**98.26**	**96.3**	**94.4**
d = 4	93.03	80.56	98.00	95.7	94.2

4.3.2. Experiments for LP-Swin Backbone

As mentioned in Section 3.2.1, we have proposed an LP-Swin backbone that combines the advantages of both vision transformer and CNN architecture. We select the R^3Det detection algorithm as the baseline and conduct three sets of ablation experiments. Our baseline comprises ResNet-50 and the standard FPN structure. Initially, ResNet-50 is replaced by Swin Transformer while maintaining consistency in the neck part. Subsequently, we integrate the LP-Swin into the network. Table 3 shows the comparative results.

In comparison to the ResNet-50 backbone, the Swin Transformer backbone can bring a significant improvement of 2.38% in mAP, 3.21% in Inshore AP, and 0.61% in Offshore AP, which proves the superiority of the Swin transformer over the CNN. Furthermore, after introducing our CLPU, LP-Swin can increase mAP by 0.51% and further improve Inshore AP and Offshore AP by 1.38% and 0.46%. The highest results indicate the superiority of the LP-Swin and its detection ability in complex scenes.

Table 3. Comparison of the LP-Swin with standard backbones.

Method	mAP_{50}	Inshore AP	Offshore AP	Precision	Recall
ResNet-50	90.45	76.36	97.21	93.5	91.7
Swin-T	92.83	79.57	97.80	95.2	93.8
LP-Swin	**93.34**	**80.95**	**98.26**	**96.3**	**94.4**

Figure 10 depicts the visualization of detection results using three backbones: ResNet-50, Swin-T, and LP-Swin. In the initial scenario of Figure 10, the ResNet-50 network suffers from one missed detection, while the Swin Transformer network experiences one false alarm. In contrast, the LP-Swin network accurately identifies and localizes the ship object without any false alarms or missed detections. In the second scenario, the ResNet-50 network encounters one false alarm, while both the Swin Transformer network and the LP-Swin network achieve correct detection and precise localization. In the final scenario, all three networks achieve the correct classification. Notably, in scenes where all classification results are correct, LP-Swin attains the highest confidence score among the networks under comparison.

4.3.3. Experiments for CS-BiFPN Network

As mentioned in Section 3.3, we have proposed a CS-BiFPN neck, which facilitates the integration and propagation of both semantic and location information. Similarly, we carry out three sets of experiments to assess our CS-BiFPN. Firstly, we substitute the standard FPN with the standard BiFPN while employing the LP-Swin as the backbone. Subsequently, we improve the neck part by using the CS-BiFPN. Table 4 exhibits the comparative outcomes.

In contrast to the traditional FPN structure, BiFPN yields enhancements of 0.3% in mAP, 0.22% in Inshore AP, and 0.34% in Offshore AP. Furthermore, the CS-BiFPN contains both transverse and longitudinal cross-scale connections, leading to an additional 0.21% improvement in mAP, 0.11% improvement in Inshore AP, and 0.6% improvement in Offshore AP. It indicates that incorporating longitudinal cross-scale connections is essential for the FPN structure.

Table 4. Performance comparison of different feature pyramid networks.

Method	mAP_{50}	Inshore AP	Offshore AP	Precision	Recall
FPN	93.34	80.95	98.26	96.3	94.4
BiFPN	93.64	81.17	98.60	96.5	94.5
CS-BiFPN	**93.85**	**81.28**	**99.20**	**97.1**	**94.7**

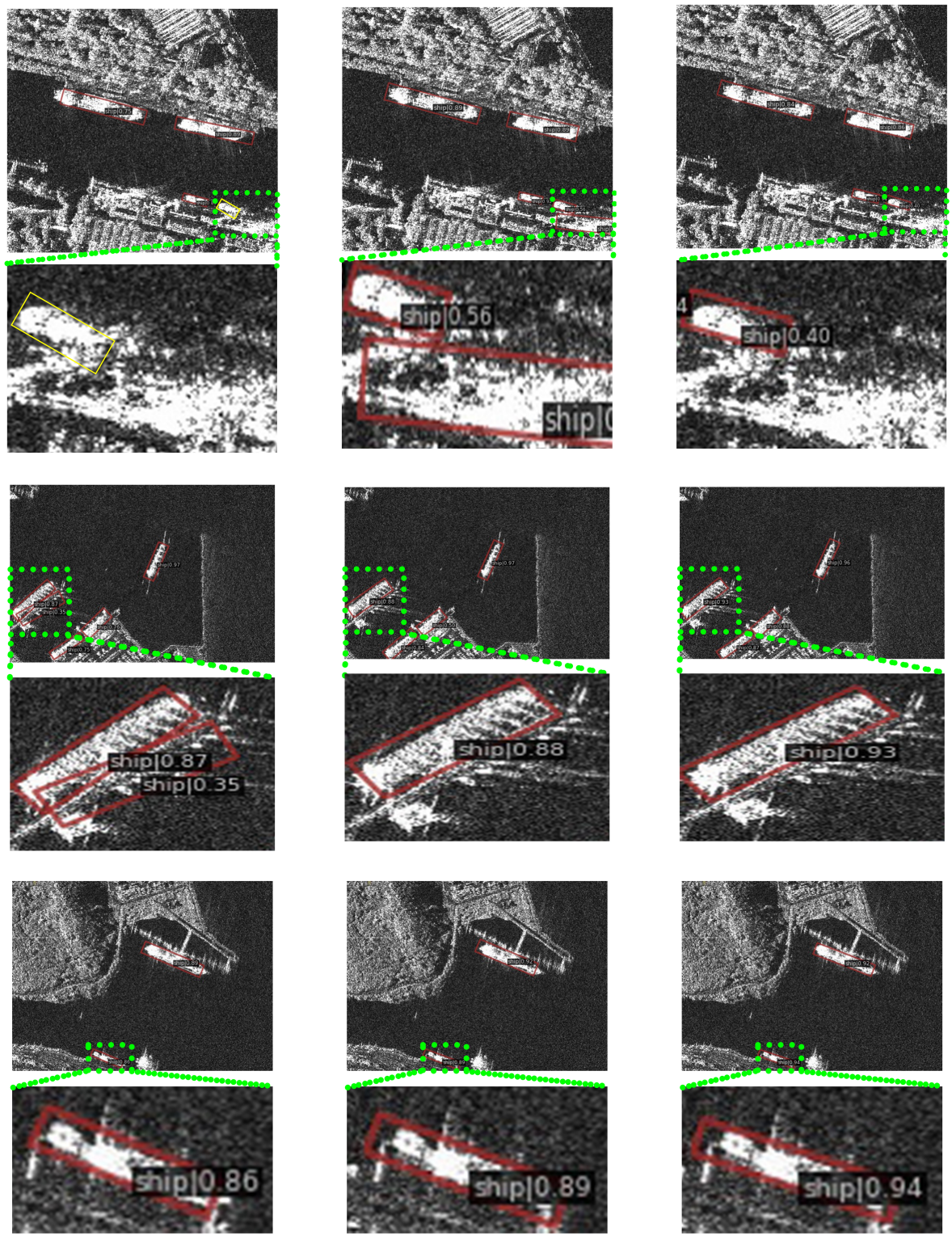


Figure 10. Visualization results of different backbones. (**a**) ResNet-50; (**b**) Swin Transformer; (**c**) LP-Swin. The yellow oriented boxes represent missed detection.

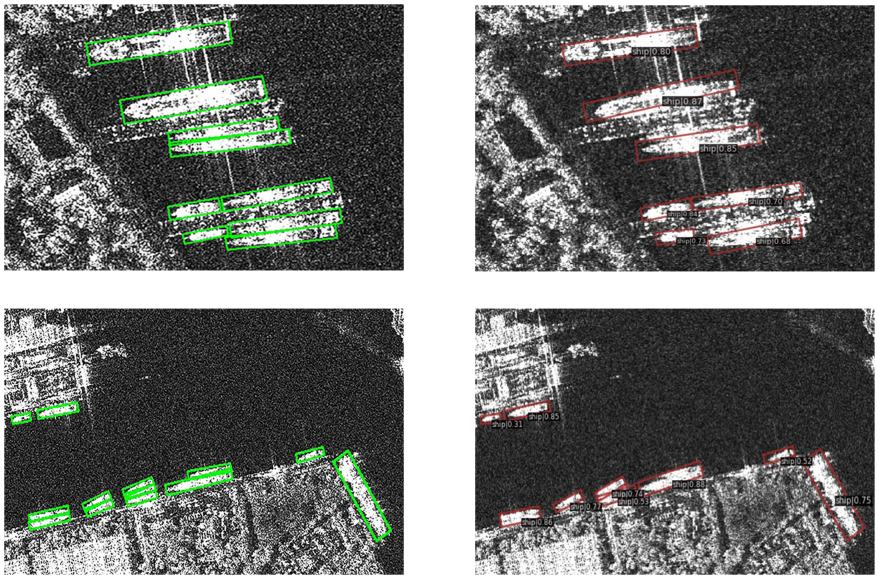
4.4. Limitation and Discussion

Although we have designed LP-Swin and CS-BiFPN to enhance feature extraction and fusion, and improved the detection performance in some inshore scenes, there are still limitations in some hard-detected samples, as shown in Figure 11. When ships are densely

distributed in parallel at inshore ports, LPST-Det may encounter missed detections, thereby degrading the detector's performance. The reasons for this phenomenon are as follows. (1) Out of a total of 928 training images in the SSDD dataset, only about 13 of them are related to the scene of ships densely parallel berthing at ports, which is not conducive to the model learning in the training stage. We consider this to be the root cause of the missed detection. (2) The distinctive SAR imaging mechanism results in unclear image contour, making it challenging to distinguish the boundaries of two tightly packed ships or ships very close to a port. (3) The interaction of cruciform sidelobes worsens the overlapping effect between targets.

In response to the above issues, further research can be conducted from the following aspects.

Firstly, for the problem of insufficient training samples, the generative adversarial network (GAN) is an effective strategy for data augmentation and can be used to generate scenes of multiple parallel ships at a port or an individual ship at different ports. Secondly, to address the unclear contour issue, it would be helpful to introduce a segmentation task that emphasizes edge information to enhance ship detection. In addition, most existing SAR detectors are following the framework of optical image object detection, ignoring some unique characteristics in SAR images, like cruciform sidelobes and sea clutter. An additional denoise subnetwork may be an attempt to alleviate these interferences. Lastly, ships berthed side by side will cause serious hull overlap effects, while the prow and stern are relatively easy to recognize. Therefore, imbalanced semantic feature extraction strategies may be beneficial for the detection head to predict the target location in these densely distributed scenarios.



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

Figure 11. Detection results of densely parallel ships. (**a**) Ground truth; (**b**) LPST-Det.

5. Conclusions and Future Work

This paper introduces LPST-Det, an effective one-stage detector designed for arbitrary-oriented SAR object detection. To enhance the extraction of object features in SAR images, we present a local-perception-enhanced Swin Transformer backbone termed LP-Swin, which combines the global feature extraction capability of Swin Transformer and the local feature capture ability of the CNN to enhance the feature extraction ability. We augment the transformer structure by incorporating a convolution-based local perception unit, referred to as the CLPU, to improve the capture of local features and boost the overall detection performance. Furthermore, we construct CS-BiFPN, a cross-scale feature fusion structure that enhances the extraction of multi-scale semantic information and position information

through the introduction of additional longitudinal cross-scale connections, which can enhance the fusion of extracted multi-scale features. In addition. We integrate the proposed LP-Swin and CS-BiFPN into the R^3Det detector, ensuring high accuracy while maintaining rapid detection speed. We perform comparison experiments and ablation tests on the SSDD dataset. In comparison to existing object detection methods, our method exhibits improvements over both general detectors and SAR ship detectors. LPST-Det attains 93.31% in mAP, with 81.28% in Inshore AP, and 99.20% in Offshore AP, illustrating that it achieves accurate classification and localization of ships in complex detection scenarios. Both visualized results and detection accuracy verify the detection performance of LPST-Det in complex scenes.

Although our detector succeeds in most detection scenes, it exhibits limitations in some hard-detected samples. When ships are densely distributed in parallel at inshore ports, LPST-Det may encounter missed detections, thereby degrading the detector's performance. Our future work will explore more effective solutions for inshore scenes and the design of a lightweight transformer-based backbone in SAR object detection.

Author Contributions: Conceptualization, X.X. and Z.Y.; methodology, X.X.; software, X.X.; validation, X.X.; formal analysis, X.X. and Z.Y.; investigation, X.X. and Z.Y.; resources, X.X. and Z.Y.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X., Z.Y., Y.L., G.W., and W.E.Z.; visualization, X.X.; supervision, Z.Y. and L.G.; project administration, X.X.; funding acquisition, Z.Y. and L.G. 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, grant number 61201238 and the Aeronautical Science Foundation of China, grant number 201801P6002.

Data Availability Statement: Data are contained within the article.

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

References

1. Liu, T.; Zhang, J.; Gao, G.; Yang, J.; Marino, A. CFAR ship detection in polarimetric synthetic aperture radar images based on whitening filter. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 58–81. [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. Schwegmann, C.P.; Kleynhans, W.; Salmon, B.P.; Mdakane, L.W.; Meyer, R.G.V. Very deep learning for ship discrimination in synthetic aperture radar imagery. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 104–107.
4. Shao, Z.; Zhang, X.; Zhang, T.; Xu, X.; Zeng, T. RBFA-Net: A Rotated Balanced Feature-Aligned Network for Rotated SAR Ship Detection and Classification. *Remote Sens.* **2022**, *14*, 3345. [CrossRef]
5. 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.* **2008**, *47*, 1685–1697. [CrossRef]
6. Cao, X.; Wu, C.; Yan, P.; Li, X. Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos. In Proceedings of the 2011 IEEE International Conference Image Processing (ICIP), Brussels, Belgium, 11–14 September 2011; pp. 2421–2424.
7. Zhou, G.; Tang, Y.; Zhang, W.; Liu, W.; Jiang, Y.; Gao, E. Shadow Detection on High-Resolution Digital Orthophoto Map (DOM) using Semantic Matching. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20.
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
12. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.

13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
14. 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, 21–27 July 2017; pp. 6517–6525.
15. Joseph, R.; Ali, F. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
16. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
18. Chen, S.-Q.; Zhan, R.-H.; Zhang, J. Robust single stage detector based on two-stage regression for SAR ship detection. In Proceedings of the International Conference on Innovation in Artificial Intelligence (ICIAI), Shanghai, China, 9–12 March 2018; pp. 169–174.
19. 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.
20. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
21. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
22. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 3163–3171.
25. 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: Models, Methods and Applications, Beijing, China, 13–14 November 2017; pp. 1–6.
26. Zhang, T.; Zhang, X. High-speed ship detection in SAR images based on a grid convolutional neural network. *Remote Sens.* **2019**, *11*, 1206. [CrossRef]
27. 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 multi-scale and multiscene SAR ship detection. *IEEE Access* **2018**, *6*, 20881–20892. [CrossRef]
28. Xu, X.; Zhang, X.; Zhang, T. Lite-YOLOv5: A Lightweight Deep Learning Detector for On-Board Ship Detection in Large-Scene Sentinel-1 SAR Images. *Remote Sens.* **2022**, *14*, 1018. [CrossRef]
29. Xu, X.; Zhang, X.; Shao, Z.; Shi, J.; Wei, S.; Zhang, T.; Zeng, T. A Group-Wise Feature Enhancement-and-Fusion Network with Dual-Polarization Feature Enrichment for SAR Ship Detection. *Remote Sens.* **2**022, *14*, 5276. [CrossRef]
30. Yasir, M.; Zhan, L.; Liu, S.; Wan, J.; Hossain, M.S.; Isiacik Colak, A.T.; Liu, M.; Islam, Q.U.; Raza Mehdi, S.; Yang, Q. Instance segmentation ship detection based on improved Yolov7 using complex background SAR images. *Front. Mar. Sci.* **2023**, *10*, 1113669. [CrossRef]
31. Zheng, Y.; Liu, P.; Qian, L.; Qin, S.; Liu, X.; Ma, Y.; Cheng, G. Recognition and Depth Estimation of Ships Based on Binocular Stereo Vision. *J. Mar. Sci. Eng.* **2022**, *10*, 1153. [CrossRef]
32. Pan, Z.; Yang, R.; Zhang, Z. MSR2N: Multi-stage rotational region based network for arbitrary-oriented ship detection in SAR images. *Sensors* **2020**, *20*, 2340. [CrossRef]
33. 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]
34. An, Q.; Pan, Z.; Liu, L.; You, H. DRBox-v2: An Improved Detector with Rotatable Boxes for Target Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8333–8349. [CrossRef]
35. Chen, S.; Zhang, J.; Zhan, R. R2FA-Det: Delving into High-Quality Rotatable Boxes for Ship Detection in SAR Images. *Remote Sens.* **2020**, *12*, 2031. [CrossRef]
36. Yang, M.; Wang, H.; Hu, K.; Yin, G.; Wei, Z. IA-Net: An Inception–Attention-Module-Based Network for Classifying Underwater Images From Others. *IEEE J. Ocean. Eng.* **2022**, *47*, 704–717. [CrossRef]
37. 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: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
38. Zhou, X.Z.; Su, W.J.; Lu, L.W.; Li, B.; Wang, X.G.; Dai, J.F. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2020.
39. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 367–376.
40. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.

41. 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]
42. Shi, H.; Chai, B.; Wang, Y.; Chen, L. A Local-Sparse-Information-Aggregation Transformer with Explicit Contour Guidance for SAR Ship Detection. *Remote Sens.* **2022**, *14*, 5247. [CrossRef]
43. 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. [CrossRef]
44. Ke, X.; Zhang, X.; Zhang, T.; Shi, J.; Wei, S. SAR ship detection based on an improved Faster R-CNN using deformable convolution. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 11–16 July 2021; pp. 3565–3568.
45. Xu, X.; Feng, Z.; Cao, C.; Li, M.; Wu, J.; Wu, Z.; Shang, Y.; Ye, S. An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation. *Remote Sens.* **2021**, *13*, 4779. [CrossRef]
46. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
47. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7029–7038.
48. Liu, N.; Cui, Z.; Cao, Z.; Pi, Y.; Lan, H. Scale-Transferrable Pyramid Network for Multi-Scale Ship Detection in SAR Images. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1–4.
49. Hu, W.; Tian, Z.; Chen, S.; Zhan, R.; Zhang, J. Dense feature pyramid network for ship detection in SAR images. In Proceedings of the Third International Conference on Image, Video Processing and Artificial Intelligence, Shanghai, China, 23–24 October 2020.
50. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A novel quad feature pyramid network for SAR ship detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
51. Chen, J.; Wang, Q.; Peng, W.; Xu, H.; Li, X.; Xu, W. Disparity-Based Multiscale Fusion Network for Transportation Detection *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 18855–18863. [CrossRef]
52. Zhang, R.; Li, L.; Zhang, Q.; Zhang, J.; Xu, L.; Zhang, B.; Wang, B. Differential Feature Awareness Network within Antagonistic Learning for Infrared-Visible Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**. [CrossRef]
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
54. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
55. Ge, J.; Tang, Y.; Guo, K.; Zheng, Y.; Hu, H.; Liang, J. KeyShip: Towards High-Precision Oriented SAR Ship Detection Using Key Points. *Remote Sens.* **2023**, *15*, 2035. [CrossRef]
56. 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, Seattle, WA, USA, 18–23 June 2018; pp. 6154–6162.
57. Ding, J.; Xue, N.; Long, Y.; Xia, G.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2849–2858.
58. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef] [PubMed]
59. Yang, X.; Yan, J. Arbitrary-Oriented Object Detection with Circular Smooth Label. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 677–694.
60. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2384–2399. [CrossRef]
61. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [CrossRef]
62. Jiang, X.; Xie, H.; Chen, J.; Zhang, J.; Wang, G.; Xie, K. Arbitrary-Oriented Ship Detection Method Based on Long-Edge Decomposition Rotated Bounding Box Encoding in SAR Images. *Remote Sens.* **2023**, *14*, 3599. [CrossRef]