Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042311

ARTICLE

Swin-PAFF: A SAR Ship Detection Network with Contextual Cross-Information Fusion

Yujun Zhang*, Dezhi Han and Peng Chen

School of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China

*Corresponding Author: Yujun Zhang. Email: zhangyujun000329@163.com

Received: 25 May 2023 Accepted: 15 August 2023 Published: 29 November 2023

ABSTRACT

Synthetic Aperture Radar (SAR) image target detection has widespread applications in both military and civil domains. However, SAR images pose challenges due to strong scattering, indistinct edge contours, multi-scale representation, sparsity, and severe background interference, which make the existing target detection methods in low accuracy. To address this issue, this paper proposes a multi-scale fusion framework (Swin-PAFF) for SAR target detection that utilizes the global context perception capability of the Transformer and the multi-layer feature fusion learning ability of the feature pyramid structure (FPN). Firstly, to tackle the issue of inadequate perceptual image context information in SAR target detection, we propose an end-to-end SAR target detection network with the Transformer structure as the backbone. Furthermore, we enhance the ability of the Swin Transformer to acquire contextual features and cross-information by incorporating a Swin-CC backbone network model that combines the Spatial Depthwise Pooling (SDP) module and the self-attentive mechanism. Finally, we design a cross-layer fusion neck module (PAFF) that better handles multi-scale variations and complex situations (such as sparsity, background interference, etc.). Our devised approach yields a noteworthy AP@0.5:0.95 performance of 91.3% when assessed on the HRSID dataset. The application of our proposed technique has resulted in a noteworthy advancement of 8% in the AP@0.5:0.95 scores on the HRSID dataset.

KEYWORDS

Transformer; deep learning; SAR object detection; ship detection

1 Introduction

Synthetic Aperture Radar (SAR) is an advanced active microwave sensor used for high-resolution remote sensing observations of the Earth [1]. It has an all-weather, all-day operation capability and performs an essential role in ocean monitoring [2]. SAR ship detection is a fundamental maritime task with important applications in offshore traffic management, fisheries management, and offshore emergency rescue [3,4]. Therefore, more and more scholars have devoted their attention to this topic [5–16].

Previous standard SAR ship detection methods require manual design of ship features, such as constant false alarm rate (CFAR) [1], saliency [2], superpixel [3], and transformation [4]. However, these traditional methods usually have complex algorithms, limited migration capability, and require tedious manual design, which limit their application in practical scenarios. In addition, these methods generally use a limited number of ship images for theoretical analysis to define ship features. Moreover, they cannot capture the ship features of diverse sizes in various backgrounds [5]. Therefore, their multi-scale and multi-scene detection performance is usually poor.

In recent years, with the development of deep learning and the rise of convolutional neural networks (CNNs), the current state-of-the-art CNN-based SAR ship detectors have addressed some of these problems to some extent [17]. However, recent research has shown that the size of the actual receptive field in convolutional neural networks is much smaller than the theoretical receptive field, which means that CNNs may not be able to make full use of the contextual information of the input data. In addition, the feature capture capability of CNNs is also limited, and they cannot extract global representation features well. To enhance the global capture capability of CNNs, deeper convolutional layers can be stacked, but this results in models that are too large, with too many learning parameters to train and converge effectively, which causes a dramatic increase in computation and reducing the timeliness of the models. Therefore, there is a requirement to explore alternative model structures and techniques to address these issues.

The Transformer structure [18] proposed by Google has been gradually applied to computer vision tasks after achieving good results in natural language processing. In 2020, Vision Transformer (ViT) [19] became the first Transformer structure adopted in computer vision, and it achieved state-of-the-art results in optical natural scene recognition, demonstrating the feasibility of Transformers in this field. Subsequently, other Transformer-based network models have emerged, such as Detection Transformer (DETR) [20] and Swin Transformer [21].

Swin Transformer is a novel Transformer architecture used for computer vision tasks such as image classification, object detection, and segmentation [22]. It employs a hierarchical attention mechanism by dividing the input image into fixed-sized patches, allowing each patch to capture global information effectively. This approach helps extract more contextual and global features. Compared to traditional Transformer networks, Swin Transformer introduces a local window mechanism, transforming global attention computation into local window-based computation, significantly reducing computational complexity. This enables Swin Transformer to operate more efficiently when dealing with large-scale images while reducing storage requirements. Therefore, the main reason for choosing Swin Transformer is its ability to extract rich contextual features while reducing computational overhead compared to other Transformer networks.

To better handle the extracted features from Swin Transformer, this paper proposes an innovative cross-level fusion neck module called PAFF, which offers several advantages including multi-scale receptive fields, feature fusion, enhanced contextual information, and end-to-end training. PAFF constructs a feature pyramid to obtain rich multi-scale features, facilitating the detection of objects at different scales. Additionally, PAFF can fuse features from different levels, integrating information from both low-level and high-level features to improve object detection performance. It also propagates contextual information through an upsampling path, enhancing the semantic representation of features. Importantly, PAFF supports end-to-end training, enabling the entire network to learn collaboratively and enhance model performance.

This paper focuses on optimizing the design of the backbone and neck parts of the target detection framework, using YOLOX [23] as the baseline. The integration of the Transformer model and the single-stage model in our research presents a novel approach that enhances the existing methodology for ship detection in Synthetic Aperture Radar (SAR) imagery. This innovative combination offers significant contributions to the field, bringing about advancements in the current SAR ship detection framework, and its main contributions include the following:

1. In order to solve the existing problem of SAR target detection which lacks weak contextual information of the sensed images, we propose an end-to-end SAR target detection network with the Transformer structure as the backbone.

2. Combining strategies such as the SDP module and the self-attentive mechanism, a Swin-CC backbone model is proposed to improve the ability of the Swin Transformer to acquire contextual features and cross information.

3. A cross-layer fusion neck module PAFF is designed, which can better cope with multi-scale changes and complex practical situations. The module can effectively fuse features from different layers to improve the performance and capability of the model, and it also improves the model's ability to detect and recognise objects at different scales.

The remainder of the paper is organised as follows: Section 2 of this paper presents an expanded elucidation of the primary framework employed herein. The subsequent Section 3 delineates the core methodology employed in this study, as well as the evaluation metrics utilized to gauge the experimental outcomes. Section 4 provides comprehensive details regarding the experimental configurations, presenting the results of the conducted ablation experiments and a substantial number of comparative experiments. Finally, Section 5 serves as the conclusive section of this paper, summarizing the overall findings and presenting prospects for advancing the research.

2 Related Work

2.1 YOLOX

The SAR ship detection problem poses challenges due to the unclear edge information of targets and the limitations of traditional anchor-based methods. In order to overcome these shortcomings and leverage the benefits of anchor-free detection frameworks, we draw inspiration from the latest anchor-free model called Exceeding YOLO Series in 2021 (YOLOX) [23], which represents a significant improvement over the YOLO series. YOLOX introduces an explicit definition of positive sample regions by projecting the 3×3 region of the ground truth bounding box onto the center of the feature graph. It predicts four values for each target, namely the offset distance of the upper left corner and the height and width of the bounding box. To handle fuzzy samples more effectively, YOLOX adopts the simOTA algorithm for positive and negative sample matching [23]. The simOTA algorithm involves calculating the matching degree for each pair, selecting the top k prediction boxes with the smallest cost in a fixed central area, and marking the grids associated with these positive samples as positive. Given the sparsity, small sample characteristics, and target scattering in SAR images, YOLOX's characteristics of a decoupling head, a new tag allocation strategy, and an anchor-free mechanism make it an ideal baseline detector with a trade-off between accuracy and speed. Therefore, we have chosen YOLOX as the benchmark network for our research.

2.2 Swin Transformer

In SAR images, small target ships may lose information during downsampling. The Swin Transformer infrastructure can address this issue with its large throughput and massively parallel processing capability. To extract features from an image, firstly, input image is first divided into tokens with high-resolution properties of the image. The Swin Block employs the Shift Window approach to limit the self-focused computation to non-overlapping partial windows, enabling cross-window connections and improving efficiency [22].

The Swin Transformer Block is an elementary module within the Swin Transformer that extracts features and transfers information in the input image. It consists of two primary components: the Swin Self-Attention (MSA) and the Swin Transformer Feed Forward (FFN). The MSA captures the relative position encoding between elements within each window by computing global contextual information. It then performs a multi-headed self-attention calculation to enable each element to aggregate information from others. This attention calculation considers the relative position relationships between elements and uses scaling parameters to balance the importance between different parts of the attention calculation. The FFN comprises two linear layers and an activation function that perform a non-linear transformation on the elements within each window. The output of the FFN calculation is added to the output of the MSA and normalized by Layer Normalization to produce the final output of the Swin Transformer Block. In the Swin Transformer, the input and output of each Swin Transformer Block consist of a set of windows, rather than individual pixels or feature vectors. This approach enables the Swin Transformer to start with locally aggregated information and gradually expand to globally aggregated information, thus capturing the structural and contextual information of the input image better. Additionally, the Swin Transformer uses a window grouping strategy that allows multiple Swin blocks to share windows, enhancing the coherence of information flow while maintaining efficiency.

The Swin Transformer adopts the concept of Windows Multi-Head Self-Attention (W-MSA) to partition the feature map into multiple non-overlapping regions, or windows, and performs Multi-Head Self-Attention within each window. This approach reduces the computational cost compared to applying Multi-Head Self-Attention (MSA) directly to the entire global feature map as in the Vision Transformer, particularly when the shallow feature map is large.

3 Main Methods

In this section, Swin Transformer is selected as the basic architecture of the backbone network, based on which the Contextual Intersection Information Extractor (CSC) module and the backbone Swin-CC network based on contextual cross-information fusion are designed. In addition, to better extract multi-scale SAR target features, an adaptive spatial feature fusion feature pyramid structure with an enhancement neck (PAFF) is designed. The general framework diagram is shown in Fig. 1.

3.1 General Architecture of Swin-PAFF

This paper uses YOLOX as the basic architecture and introduces Swin Transformer as the backbone model of the network, and designs the feature extraction module CSC to fully capture contextual cross-information. In particular, the CSC module achieves more comprehensive and finer information interaction and better contextual information capture by cross-linking different blocks in the target feature map and can help the network to better capture the relationship between the target and its surroundings, thus improving the accuracy of target detection. The module has higher computational efficiency and better interpretability, can optimizes the utilization of contextual and cross-position information, can extract richer feature information, and improves the multi-scale SAR target multi-characterization and description capability.

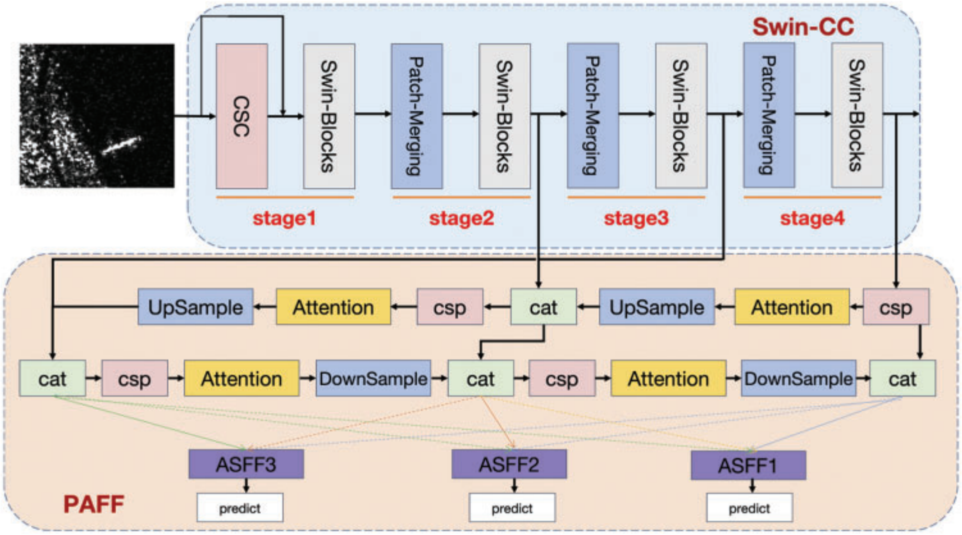

Figure 1: General framework of Swin-PAFF

In this paper, a novel enhanced neck module called PAFF (Cross-Layer Augmented Features and Adaptive Spatial Fusion) is proposed. The module incorporates a feature pyramid network [24] to facilitate the transfer of robust semantic features from the top to the bottom layers. Employing top-down and bottom-up feature pyramid networks, it enhances multi-scale connectivity operations. Furthermore, the proposed module leverages an attention mechanism to generate a weight map of equal dimensions from various layer parameters. Each pixel in the weight map corresponds to a specific coordinate, which is then multiplied element-wise with the corresponding element in the feature map to generate a new feature map [25]. This approach effectively captures spatial relationships within the feature map, leading to improved model accuracy. Moreover, unlike traditional attention mechanisms, explicit calculation of attention weights is unnecessary, thereby reducing computational costs. Finally, the spatial adaptive fusion module facilitates better adaptation of the model to target scale variations and positional changes, resulting in enhanced feature fusion.

Swin-PAFF consists of three main components: Swin-CC, PAFF, and YOLOX-Head. First, features are extracted from the input image using the designed Swin-CC backbone network, and the obtained feature maps are fused at multiple scales using the PAFF module. The YOLOX-Head module receives the multi-scale feature maps and performs multi-scale feature fusion.

3.2 Swin-CC Backbone Network Based on Contextual and Cross-Information Fusion

For the characteristics of SAR targets such as strong scattering, sparse and multi-scale, a target detection backbone network based on context and cross information fusion, Swin-CC, is designed by combining the advantages of Transformer, which can extract richer contextual features and cross information and improve multi-scale SAR target characterization. First, this paper uses the Swin Transformer, a current state-of-the-art model in the field of target detection and instance segmentation, as the base backbone network. Next, inspired by CCNet [26] and CSP_Darknet [27], this paper extends the field of perception, depth, and resolution by introducing the CSC attention mechanism module in the Patchembed section, which enhances the feature-aware domain, strengthens the comparison between different windows on the feature map, and performs the ability to fuse contextual and cross-information.

3.2.1 Context-Based and Cross-Information Fusion CSC Module

In this study, we integrated the CSC attention mechanism module into the Patchembed module and restructured it. The structure of the network for extracting the CSC attention mechanism is depicted in Fig. 2. By incorporating the CSC module into the Patchembed module, we expanded the field of perception and the depth of the network. As a result, we improved the resolution of the network and its overall performance.

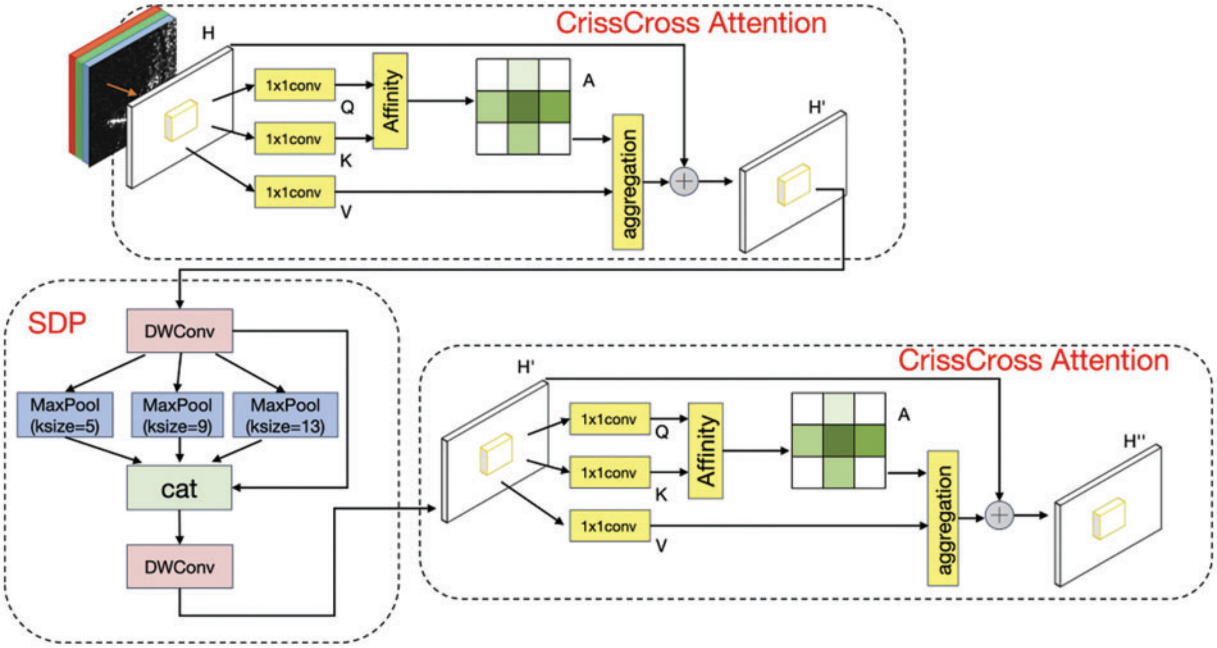

Figure 2: CSC module network structure

The figure presented in this paper, Fig. 3 illustrates that the CSC feature extraction module comprises two components: the CrissCross Attention (CCA) [26] mechanism module and the SDP module. The CrissCross Attention is a multi-headed attention mechanism that is designed to improve a model's ability to model spatial and channel relationships between features. It operates by mapping input features into multiple query, key, and value vectors and then computing an attention score by taking the inner product of queries and keys to assign a weight to each location and channel. Then, the weighted value vectors are summed to produce a weighted feature representation. The advantage of CrissCross Attention is that it can effectively capture spatial relationships and cross-channel dependencies between features, thus improving the model performance while consuming less GPU memory.

We have innovatively introduced the CSC operation to address the limitation of the CCA module, which can only capture contextual information both horizontally and vertically. The CSC operation first fuses the feature information extracted from the previous CCA module with the SDP operation information features, reducing the number of computed parameters, before adding another layer of CCA modules. In the first CCA attention module, the feature map H extracted from each patch's feature picture serves as the input, producing an output feature map $H^{'}$ of the same shape. In the second CCA attention module, the input is $H^{'}$ that has undergone feature fusion in the SDP module, producing an output feature map $H^{''}$. Fig. 3 illustrates this process. By fusing features from two CCA modules and one SDP module, we obtain full image contextual information from all pixel points and generate new features that are contextually dense and rich in information.

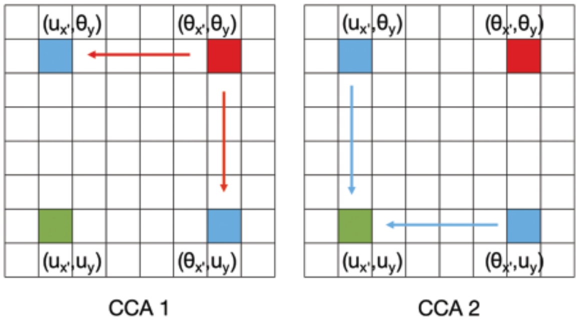

Figure 3: Example diagram after two CCA propagations

Overall, our CSC module compensates for the fact that cross-information cannot be obtained from all pixels with dense contextual information. Compared to the CCA module, the CSC module does not bring in additional parameters and can obtain improved performance with lower computational effort. So we introduce the CSC module to enhance the network's ability to extract contextual features and fuse cross-information while reducing the computational effort.

3.2.2 Multi-Scale Feature Extraction

The SDP method is a novel approach to enhancing the receptive field of neurons and achieving multi-scale feature extraction by utilizing maximum pooling and depth separable convolution (DWconv) techniques. Through the integration of convolution kernels of varying sizes across different layers, the neural network is capable of capturing a broader range of features and improving its receptive field, resulting in improved target detection. Furthermore, the employment of multi-layer features helps to mitigate the impact of environmental factors and enhance the background data at varying scales, thus facilitating the detection of small targets.

3.3 Feature Pyramid and Spatial Adaptive Fusion Module PAFF

The architecture of the feature pyramid and spatial adaptive fusion module (PAFF) proposed in this paper consists of two main parts, PAFPN and ASFF [28], and its structural diagram is shown in Fig. 4.

PAFPN is a feature extraction and integration network that can detect targets of various sizes at different scales. It improves on the Feature Pyramid Network (FPN) by not only upsampling and fusing features between each level but also combining features between levels to enhance sampling and fusion. The path aggregation module in PAFPN captures the importance between different feature layers and fuses their features with the corresponding up/down-sampled features into a unified feature pyramid. This approach allows PAFPN to better handle targets of different sizes and improve detection performance for both large and small targets.

Research on attention mechanisms has revealed that inter-channel attention significantly improves model performance. However, it often neglects inter-pixel location information. Therefore, this paper incorporates Coordinate Attention (CA) [29] operations into the feature map before up/downsampling by PAFPN. As shown in Fig. 5, CA encodes the relationship between channel and remote location information. The overall structure consists of two steps: Coordinate Information Embedding and Coordinate Attention Generation.

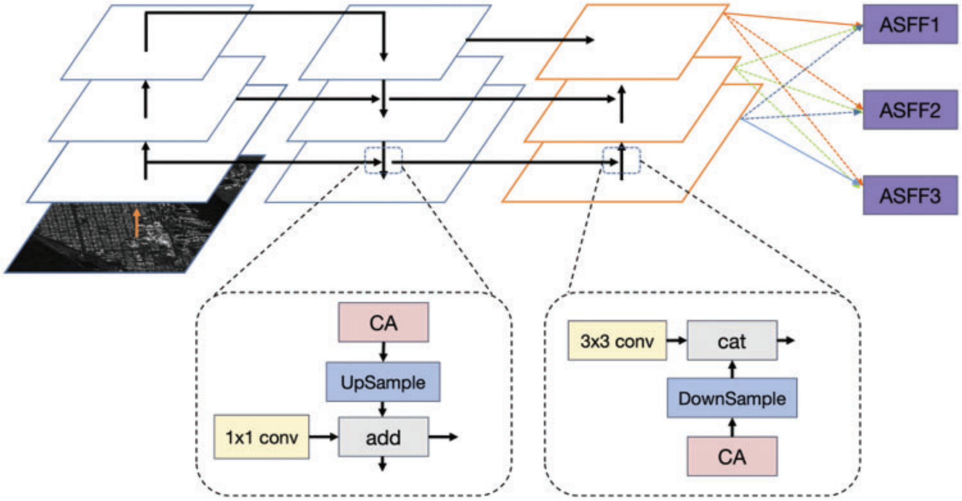

Figure 4: Structure of the PAFF model

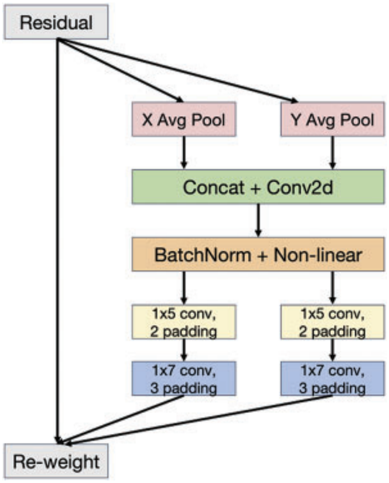

Figure 5: CA model structure

Modeling the relationships between channels with ordinary convolution is difficult. Explicitly constructing interdependencies between channels can enhance the model sensitivity to informative channels, thereby facilitating the final classification decision. Furthermore, global average pooling can help the model capture global information, which is not available through convolution. The compression step for the cth channel given an input X can be expressed as shown in Eq. (1).

$$z^c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{1}$$

Feature pyramids are a commonly used technique in object detection but suffer from inconsistency between features at different scales, particularly for first-level detectors. To address this issue, we propose the Attentional Spatial Fusion Network (ASFF) feature fusion strategy to improve target detection performance within the feature pyramid in single-level detectors. The ASFF network module filters and fuses feature maps from different levels to retain valuable information, adaptively blending features from different levels at each spatial location. From Fig. 4, we show that fusion vectors are a weighted combination of vectors from the first three feature maps, with network-learned coefficients shared across all channels. We set the feature maps to different levels (l ∈ 0, 1, 2) based on the input mapping's dimensionality, with the corresponding feature maps referred to as Xl. In this paper, we set level l (l ∈ 0, 1, 2) to 192, 384, and 768, respectively, depending on the input size and the characteristics of the SAR ship target. The output of the ASFF module is defined as follows:

$$y^l = \alpha^l \times x^{1\to l} + \beta^l \times x^{2\to l} + \gamma^l \times x^{3\to l} \tag{2}$$

where $x^{n\to l}$ denotes the feature vector after adjusting the features at level n to level l. α, β and γ are the learning rates at level n, level n+1 and level n+2, respectively, defined as follows:

$$\alpha^l + \beta^l + \gamma^l = 1 \tag{3}$$

$$\alpha^l = \frac{e^l_\alpha}{e^l_\alpha + e^l_\beta + e^l_\gamma} \tag{4}$$

3.4 Loss Function

Swin-PAFF is derived from the YOLOX model, with improvements and optimizations made to the original model. As a result, the loss function used in Swin-PAFF is similar to that of the YOLOX model. Below are the formulas for the operations of L_{cls}, L_{reg}, L_{obj}, and Loss, respectively:

$$Loss = \frac{L_{cls} + L_{reg} + L_{obj}}{N_{pos}} \tag{5}$$

To calculate the L_{cls} classification loss, we utilize the Varifocal loss. This loss function allows for more flexible control of the model's attention on samples of different difficulty levels by adjusting the α and β parameters accordingly. The formula for the classification loss is as follows:

$$VFL(p,q) = \begin{cases} -q(qlog(p) + (1-q)log(1-p) & q > 0 \\ -\alpha p^\gamma log(1-p) & q = 0 \end{cases} \tag{6}$$

L_{reg} localization loss function, we use GIoU Loss. GIoU Loss is a useful metric for target detection as it makes the predicted bounding box as close as possible to the true bounding box in terms of position and size. This helps to improve the accuracy of the target detection model, especially when dealing with objects with complex shapes or high occlusions. Its formula is as follows:

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{\mathrm{A^c} - \mathrm{u}}{\mathrm{A^c}} \tag{7}$$

For the L_{obj} loss function, we use the binary cross-entropy loss function (BCELoss). It is calculated as follows:

$$Loss_{obj} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p(y_i)) + (1-y_i)log(1-p(y_i)) \tag{8}$$

For the final 15 rounds, we disabled the data augmentation mode and employed the SmoothL1Loss function on the unencoded predictions of the positive sample bounding boxes and their corresponding ground truth bounding boxes. The calculation of the SmoothL1Loss is presented in Eq. (12).

$$smoothL1(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{9}$$

4 Experimental Results

The present section presents the experimental evaluation of our proposed detection method. The HRSID dataset is utilized as the experimental data, and the model's effectiveness is validated using data from the SSDD [29] and a subset of the HRSID [30] dataset. Subsequently, we assess the effectiveness of the Swin-CC attention mechanism module and the PAFF module on the model by performing ablation experiments, as described in the following section. Finally, we compare our approach with other existing methods to validate its effectiveness. The experimental environment includes the MMdetection [31] based framework, NVIDIA RTX3090 GPU with 24 GB video memory, and Ubuntu 18.04 operating system. The AdamW optimizer is used with a batch size of 8 and the model epoch set to 100.

4.1 Calculation Methods of Evaluation Metrics

This paper utilizes the COCO evaluation metric, which employs average precision (AP) as the primary metric. AP is calculated by determining the mean precision at ten intersections with an IoU threshold that ranges from 0.50 to 0.95 in intervals of 0.05. AP_{50} refers to the AP with an IoU threshold of 0.50, while AP_{75} represents the AP with an IoU threshold of 0.75. In addition, AP_S and AP_M denote the AP for small and medium ships, respectively. The evaluation metric uses true positives (TP), false positives (FP), and false negatives (FN) to determine the number of samples in each category. Recall (R), precision (P), and mean average precision (MAP) are defined as follows:

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{11}$$

$$\mathrm{mAP} = \int_0^1 \mathrm{P(R)dR} \tag{12}$$

4.2 Ablation Experiments

This paper presents a series of ablation experiments to verify the efficacy of the Swin-CC and PAFF modules proposed in this study. The experiments use the HRSID dataset published by Wei et al. [30] in 2020, which comprises 5604 images of SAR ships with resolutions ranging from 0.1 to 3 m and featuring HH, HV, and VV polarisations.

Our selection of YOLOX as the current baseline is supported by the data presented in Table 1. We conducted training on the HRSID dataset using several widely used target detection models, including YOLOF [32], YOLOv3 [33], YOLOv7 [34], YOLOv8 [35], and YOLOX [23]. Upon comparing their performance, we observed that YOLOX achieved a slightly lower AP_{50} value than YOLOv8, but with a smaller number of parameters. This finding indicates that YOLOX can deliver satisfactory results while requiring relatively fewer computational resources. As a result, we have chosen YOLOX as our benchmark model for further analysis.

Table 1: Comparison with the state-of-the-art detectors on HRSID dataset

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	Params
YOLOF	0.459	0.682	0.529	0.45	0.544	42.06
YOLOv7	0.489	0.756	0.559	0.519	0.325	37.62
YOLOv3	0.476	0.823	0.499	0.519	0.38	61.52
YOLOv8	0.614	0.853	0.691	0.627	0.607	43.7
YOLOX	0.567	0.835	0.639	0.591	0.534	27.53

The benchmark model is augmented with the Swin-CC and PAFF modules, and their impact on the model's performance is evaluated. The experimental results presented in Table 2 indicate that the Swin-CC module significantly improves detection accuracy. That is because the Swin Transformer's remote modeling capability and the CSC module's powerful feature extraction capability enable the model to capture target object features more effectively. Furthermore, the PAFF module is used as a feature enhancement extraction tool, and the results demonstrate that incorporating the PAFF module improves the model's detection accuracy by 1%. Most notably, the combined use of the Swin-CC and PAFF modules leads to an exponential increase of approximately 8% in feature extraction performance compared to the baseline model. These experimental results establish the proposed method's scalability, applicability, and efficacy in enhancing detection performance.

Table 2: Ablation experiments

Method	Swin-CC	PAFF	AP	AP_{50}	AP_{75}	AP_S	AP_M
YOLOX (baseline)	–	–	0.567	0.835	0.639	0.591	0.534
YOLOX+Swin-CC	√	–	0.583	0.842	0.665	0.602	0.547
YOLOX+PAFF	–	√	0.576	0.848	0.651	0.598	0.545
Swin-PAFF	√	√	0.646	0.913	0.732	0.657	0.678

Table 2 illustrates that the Swin-CC module, serving as the backbone network module, has superior feature extraction capabilities due to its combination of global-local and spatial-location characteristics. Furthermore, the use of the Swin-CC module improves the detection performance of small, medium, and large ships compared to the baseline YOLOX. Specifically, the AP_S and AP_M for small and medium-sized vessels increased by 1.1% and 1.3%, respectively, while AP, AP_{50}, and AP_{75} increased by 1.6%, 0.7%, and 2.6%, respectively. In addition, the PAFF module improves inter-feature fusion and adaptive spatial fusion compared to the baseline YOLOX, resulting in improvements of 0.9%, 1.3%, and 1.2% for AP, AP_{50}, and AP_{75}, and 0.7% and 1.1% for AP_S and AP_M. When combined, the complete Swin-PAFF yields significant improvements of 7.9%, 7.8%, 9.3%, 6.6%, and 14.4% in each metric compared to the baseline. These findings suggest that both the Swin-CC module and the PAFF module of Swin-PAFF improve model detection performance and their combination results in a more robust model.

4.3 Comparative Experiments

To substantiate the effectiveness of the proposed approach, this paper conducts a comparative analysis with state-of-the-art SAR ship detection methods, utilizing HRSID dataset as presented in Table 3. The Swin-PAFF model is a hybrid model combining a single-stage model and a transformer model. Hence, we selected approaches primarily based on single-stage and transformer models for comparison, including YOLOF [32], YOLOv3 [33], YOLOv7 [34], YOLOv8 [35], RetinaNet [36], SSD [37], FCOS [38], TOOD [39], VFNet [40], PVT [41], and deformable DETR [42]. Furthermore, we evaluated the performance of our model against classical one-stage models using the LS-SSDD-v1.0 dataset, with the results reported in Table 4. Additionally, to assess the model's generalization capability, we directly applied the weights of the aforementioned models to predict the accuracy of the SSDD dataset. We conducted predictions for the test set, training set, nearshore hull, and offshore hull, and the outcomes are presented in Tables 5–8. The Swin-PAFF model exhibited promising results across all scenarios, highlighting its robustness and superior performance.

Table 3: Results of ten network models with the HRSID dataset

Method	AP	AP$_{50}$	AP$_{75}$	AP$_{S}$	AP$_{M}$	Params
YOLOF	0.459	0.682	0.529	0.45	0.544	42.06
SSD	0.448	0.735	0.49	0.44	0.594	36.04
YOLOv7	0.489	0.756	0.559	0.519	0.325	37.62
DeformableDETR	0.532	0.798	0.599	0.537	0.586	39.84
PVT	0.538	0.815	0.587	0.551	0.571	71.1
YOLOv3	0.476	0.823	0.499	0.519	0.38	61.52
VFNet	0.593	0.83	0.663	0.6	0.657	32.67
RetinaNet	0.586	0.851	0.646	0.592	0.659	36.33
YOLOv8	0.614	0.853	0.691	0.627	0.607	43.7
FCOS	0.577	0.861	0.649	0.591	0.622	50.96
TOOD	0.62	0.872	0.702	0.63	0.642	50.97
Swin-PAFF (ours)	0.646	0.913	0.733	0.657	0.679	45.7

Table 4: Results of ten network models with the LS-SSDD-v1.0 dataset

Method	AP	AP$_{50}$	AP$_{75}$	AP$_{S}$	AP$_{M}$	Params
YOLOF	0.166	0.507	0.039	0.150	0.318	42.06
YOLOv3	0.208	0.580	0.078	0.202	0.315	37.62
RetinaNet	0.219	0.607	0.072	0.202	0.367	36.33
YOLOv7	0.281	0.713	0.116	0.266	0.384	37.62
YOLOv8	0.297	0.74.4	0.150	0.281	0.432	43.7
Swin-PAFF (ours)	0.301	0.752	0.161	0.293	0.447	45.7

Table 5: Results of ten network models under the train dataset in SSDD

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M
FCOS	0.304	0.608	0.238	0.338	0.173
YOLOv3	0.275	0.658	0.162	0.337	0.234
DeformableDETR	0.302	0.692	0.217	0.32	0.211
TOOD	0.339	0.72	0.264	0.362	0.243
YOLOF	0.299	0.722	0.177	0.329	0.147
VFNet	0.35	0.726	0.274	0.384	0.235
RetinaNet	0.362	0.741	0.305	0.391	0.24
PVT	0.331	0.748	0.237	0.36	0.226
SSD	0.287	0.77	0.349	0.397	0.391
YOLOv7	0.343	0.753	0.304	0.374	0.375
YOLOv8	0.447	0.784	0.455	0.478	0.398
Swin-PAFF (ours)	0.464	0.803	0.476	0.492	0.41

Table 6: Results of ten network models under val dataset in SSDD

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M
FCOS	0.294	0.633	0.235	0.33	0.176
DeformableDETR	0.284	0.643	0.211	0.304	0.2
YOLOv3	0.272	0.645	0.171	0.328	0.227
YOLOF	0.283	0.682	0.167	0.314	0.152
TOOD	0.33	0.687	0.252	0.354	0.251
VFNet	0.33	0.688	0.245	0.362	0.23
PVT	0.334	0.716	0.254	0.366	0.227
RetinaNet	0.351	0.721	0.302	0.385	0.241
SSD	0.377	0.753	0.33	0.384	0.388
YOLOv7	0.384	0.767	0.345	0.396	0.393
YOLOv8	0.449	0.773	0.465	0.468	0.408
Swin-PAFF (ours)	0.462	0.78	0.479	0.489	0.41

Tables 4 and 5 depict that the proposed Swin-PAFF approach outperforms other models in detecting ships on the HRSID dataset and the LS-SSDD-v1.0 dataset. Additionally, Tables 3–6 exhibit that the Swin-PAFF framework not only achieves outstanding results on the HRSID dataset but also showcases strong generalization capabilities on the SSDD dataset, especially in complex nearshore scenes. The Swin-CC's ability to extract contextual cross-information and the PAFF enhancement neck's spatially adaptive fusion capability are the two main factors behind this remarkable performance, which improves the feature extraction performance of the Swin-PAFF framework and enhances the network's robustness.

Table 7: Results of the ten network models under the inshore dataset in SSDD

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M
DeformableDETR	0.195	0.417	0.165	0.211	0.164
YOLOv3	0.191	0.426	0.158	0.323	0.138
PVT	0.234	0.493	0.199	0.266	0.187
VFNet	0.288	0.548	0.296	0.336	0.201
TOOD	0.306	0.555	0.332	0.364	0.183
YOLOF	0.311	0.575	0.289	0.366	0.167
FCOS	0.29	0.577	0.255	0.325	0.221
SSD	0.308	0.595	0.286	0.345	0.276
RetinaNet	0.311	0.598	0.286	0.325	0.299
YOLOv7	0.282	0.584	0.268	0.329	0.289
YOLOv8	0.368	0.596	0.418	0.409	0.306
Swin-PAFF (ours)	0.37	0.603	0.426	0.409	0.311

Table 8: Results of the ten network models with the offshore dataset in SSDD

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M
FCOS	0.295	0.655	0.227	0.334	0.156
YOLOF	0.286	0.738	0.142	0.311	0.161
DeformableDETR	0.325	0.745	0.231	0.344	0.246
TOOD	0.352	0.758	0.234	0.36	0.328
VFNet	0.358	0.769	0.242	0.377	0.288
YOLOv3	0.326	0.779	0.188	0.334	0.317
RetinaNet	0.374	0.789	0.312	0.416	0.214
PVT	0.38	0.808	0.294	0.406	0.273
SSD	0.423	0.853	0.364	0.412	0.487
YOLOv7	0.453	0.812	0.348	0.454	0.466
YOLOv8	0.487	0.864	0.456	0.492	0.458
Swin-PAFF (ours)	0.507	0.877	0.508	0.523	0.476

The analysis presented in Table 4 reveals that our model exhibits a higher number of parameters in comparison to other YOLO models. While the fusion of YOLO and Swin Transformer for static target detection is a novel approach, the significant increase in the number of parameters after incorporating the Swin Transformer raises concerns regarding the practicality of real-time detection and tracking tasks. It is imperative to acknowledge this issue as a crucial area that requires further development in the future.

Moreover, Fig. 6 presents that the Swin-PAFF model possesses powerful generalization capabilities and surpasses other models in detecting near-shore ships without relying on the SSDD dataset for training. Furthermore, Fig. 7 demonstrates that the Swin-PAFF model exhibits exceptional performance in detecting small nearshore objects on the HRSID dataset, further highlighting its exceptional generalization capabilities.

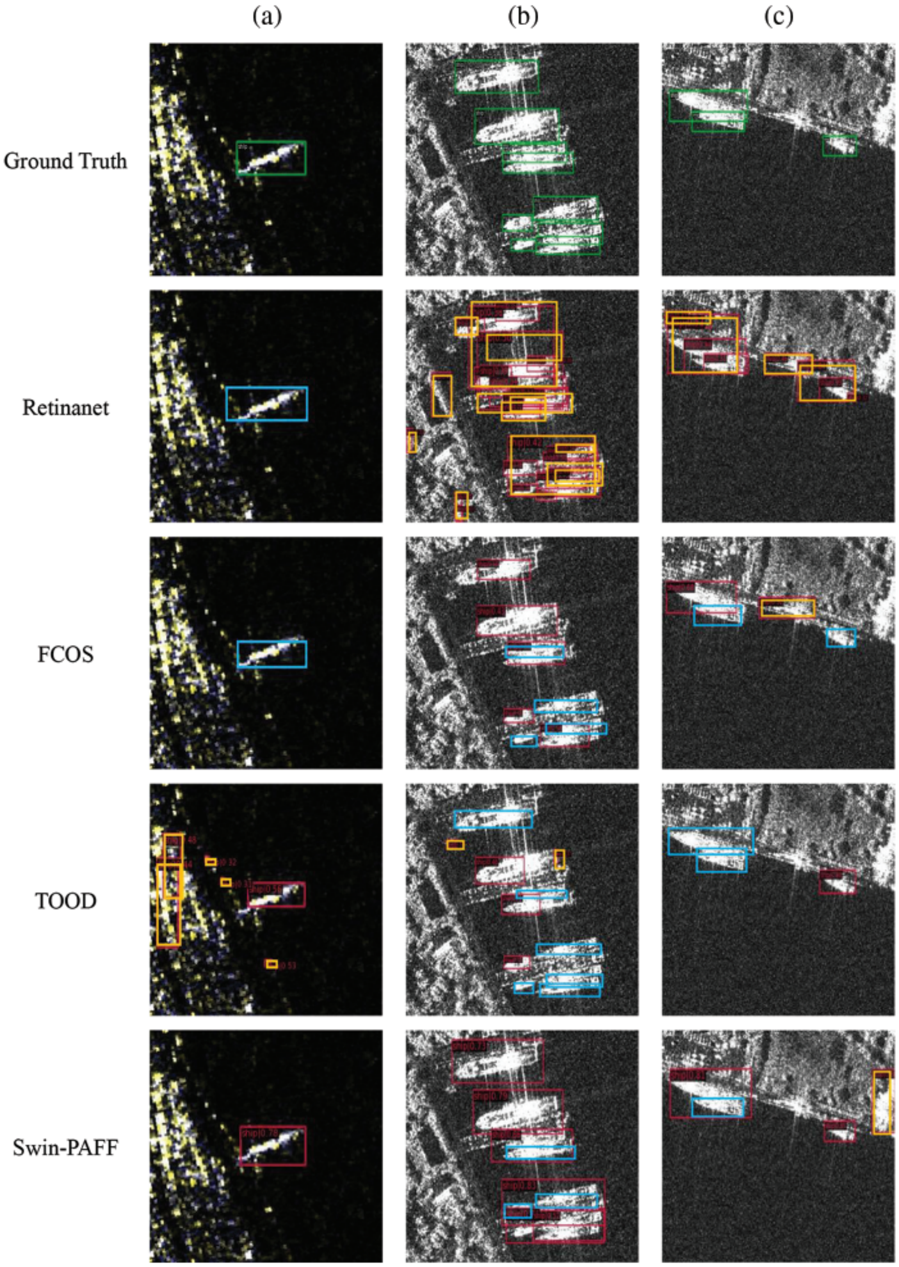

Figure 6: Detection nearshore in the SSDD dataset. (a) Green markers indicate true boxes. (b) Orange markers indicate targets with false detections (i.e., false positives, FP). (c) Blue markers indicate missed detections (i.e., false negatives, FN)

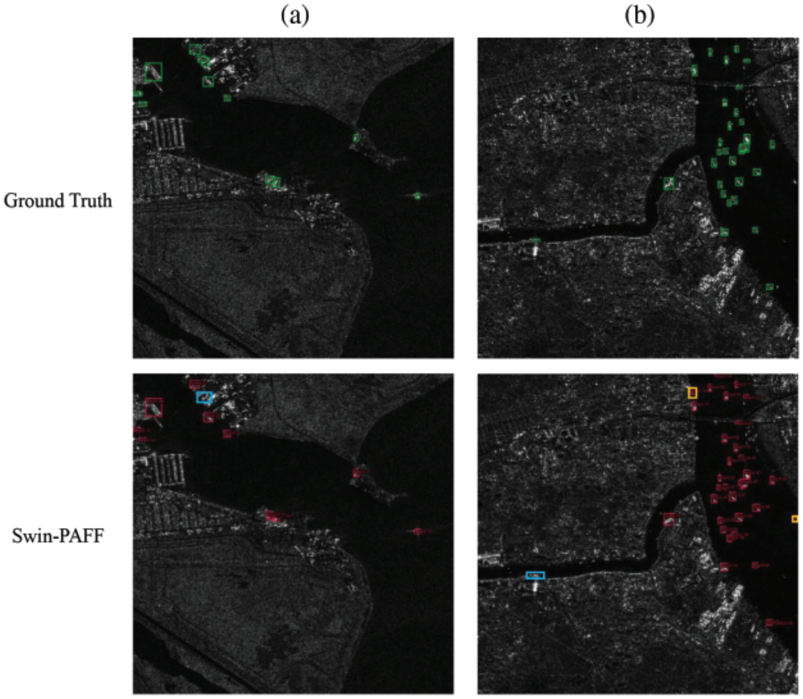

Figure 7: HRSID dataset detection Swin-PAFF module (a) Green markers indicate true boxes. (b) Orange markers indicate targets with false detections (i.e., false positives, FP). (c) Blue markers indicate missed detections (i.e., false negatives, FN)

5 Conclusion

In this paper, we present Swin-PAFF, a Transformer network designed for SAR ship detection, which incorporates contextual cross-information fusion and adaptive spatial feature fusion. Swin-PAFF builds on the YOLOX, an advanced single-stage target detection algorithm, and addresses challenges related to complex backgrounds, large scale differences, and dense distribution of small targets in SAR ship detection.

In order to enrich the extraction of comprehensive contextual intersection information, we have developed a dedicated feature extraction module, referred to as the Contextual Intersection Information Extractor (CSC), and seamlessly integrated it with the Swin Transformer architecture. Furthermore, we propose a novel technique, named Adaptive Spatial Feature Fusion (PAFF), for enhancing the feature pyramid structure, thereby bolstering the efficacy of feature extraction. Our methodology was rigorously evaluated on the HRSID dataset, yielding an accuracy improvement of approximately 8%. Additionally, Swin-PAFF exhibits exceptional performance as a single-stage network, thereby underscoring its robustness and practical utility. Nonetheless, the incorporation of the Transformer architecture significantly escalates the computational requirements of our model, necessitating the exploration of strategies to alleviate this computational burden.

Although Swin-PAFF combines Transformer and CNN detection methods, we intend to explore ways to reduce the computational complexity associated with the Transformer model by integrating it with blockchain in future research [43,44]. In addition, we plan to explore ways to make the Transformer model lighter.

Acknowledgement: None.

Funding Statement: This research is supported by the National Key Research and Development Program of China under Grant 2021YFC2801001, the Natural Science Foundation of Shanghai under Grant 21ZR1426500, and the 2022 Graduate Top Innovative Talents Training Program at Shanghai Maritime University under Grant 2022YBR004.

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

Availability of Data and Materials: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. Liu, J. Zhang, G. Gao, J. Yang and A. Marino, "CFAR ship detection in polarimetric synthetic aperture radar images based on whitening filter," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 58–81, 2019.

[2] M. Yang, C. Guo, H. Zhong and H. Yin, "A curvature-based saliency method for ship detection in SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 9, pp. 1590–1594, 2021.

[3] H. Lin, H. Chen, K. Jin, L. Zeng and J. Yang, "Ship detection with superpixel-level fisher vector in high-resolution SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 247–251, 2020.

[4] C. P. Schwegmann, W. Kleynhans and B. P. Salmon, "Synthetic aperture radar ship detection using haar-like features," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 2, pp. 154–158, 2017.

[5] S. D. Khan, L. Alarabi and S. Basalamah, "A unified deep learning framework of multi-scale detectors for geo-spatial object detection in high-resolution satellite images," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 9489–9504, 2022.

[6] X. Qian, X. Cheng, G. Cheng, X. Yao and L. Jiang, "Two-stream encoder GAN with progressive training for co-saliency detection," *IEEE Signal Processing Letters*, vol. 28, pp. 180–184, 2021.

[7] T. Zhang and X. Zhang, "A mask attention interaction and scale enhancement network for SAR ship instance segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[8] Y. Zhou, K. Fu, B. Han, J. Yang, Z. Pan *et al.,* "D-MFPN: A doppler feature matrix fused with a multilayer feature pyramid network for sar ship detection," *Remote Sensing*, vol. 15, no. 3, pp. 626, 2023.

[9] G. Tang, H. Zhao, C. Claramunt, W. Zhu, S. Wang *et al.,* "PPA-Net: Pyramid pooling attention network for multi-scale ship detection in SAR images," *Remote Sensing*, vol. 15, no. 11, pp. 2855, 2023.

[10] Y. Du, L. Du, Y. Guo and Y. Shi, "Semisupervised SAR ship detection network via scene characteristic learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[11] S. Li, Z. Pan and Y. Hu, "Multi-aspect convolutional-transformer network for SAR automatic target recognition," *Remote Sensing*, vol. 14, no. 16, pp. 3924, 2022.

[12] M. Yasir, S. Liu, M. Xu, H. Sheng, S. H. Md *et al.,* "Multi-scale ship target detection using SAR images based on improved YOLOv5," *Frontiers*, vol. 9, pp. 1086140, 2022.

[13] Y. Yang, Y. Ju and Z. Zhou, "A super lightweight and efficient sar image ship detector," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[14] Z. Wang, R. Wang, J. Ai, H. Zou and J. Li, "Global and local context-aware ship detector for high-resolution SAR images," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 4, pp. 1–9, 2023.

[15] Y. Gong, Z. Zhang, J. Wen, G. Lan and S. Xiao, "Small ship detection of SAR images based on optimized feature pyramid and sample augmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 7385–7392, 2023.

[16] Z. Sun, C. Meng, T. Huang, Z. Zhang and S. Chang, "Marine ship instance segmentation by deep neural networks using a global and local attention (GALA) mechanism," *PLoS One*, vol. 18, no. 2, pp. e0279248, 2023.

[17] H. Yan, "Detection with fast feature pyramids and lightweight convolutional neural network: A practical aircraft detector for optical remote images," *Journal of Applied Remote Sensing*, vol. 16, no. 2, pp. 024506, 2022.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998– 6008, 2017.

[19] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo *et al.,* "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov *et al.,* "End-to-end object detection with Transformers," In: A. Vedaldi, H. Bischof, T. Brox and J. M. Frahm (Eds.), *Computer Vision—ECCV 2020, Lecture Notes in Computer Science*, pp. 213–229, Cham: Springer International Publishing, 2020.

[21] X. Ke, X. Zhang, T. Zhang, J. Shi and S. Wei, "SAR ship detection based on swin transformer and feature enhancement feature pyramid network," in *IEEE Int. Geoscience and Remote Sensing Symp.*, Kuala Lumpur, Malaysia, pp. 2163–2166, 2022.

[22] S. Paul and P. Y. Chen, "Vision Transformers are robust learners," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 36, no. 2, pp. 2071–2081, 2022.

[23] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei *et al.,* "CCNet: Criss-cross attention for semantic segmentation," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 603–612, 2019.

[24] Z. Ge, S. Liu, F. Wang, Z. Li and J. Sun, "YOLOX: Exceeding YOLO series in 2021," arXiv:2107.08430, 2021.

[25] Z. Li, E. Li, T. Xu, A. Samat and W. Liu, "Feature alignment FPN for oriented object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[26] H. Zhu, S. Guo, W. Sheng and L. Xiao, "SCM: A searched convolutional metaformer for SAR ship classification," *Remote Sensing*, vol. 15, no. 11, pp. 2904, 2023.

[27] Y. Qiao, Y. Guo and D. He, "Cattle body detection based on YOLOv5-ASFF for precision livestock farming," *Computers and Electronics in Agriculture*, vol. 204, pp. 107579, 2023.

[28] Q. Hou, D. Zhou and J. Feng, "Coordinate attention for efficient mobile network design," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Online, pp. 13708–13717, 2021.

[29] T. Zhang, X. Zhang, J. Li, X. Xu, B. Wang *et al.,* "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sensing*, vol. 13, no. 18, pp. 3690, 2021.

[30] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su *et al.,* "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[31] C. Kai, W. Jia, P. Jiang, C. Yu, X. Yu *et al.,* "MMDetection: Open mmlab detection toolbox and benchmark," arXiv:1906.07155, 2019.

[32] X. Zhang, X. Yang, D. Yang, F. Wang and X. Gao, "A universal ship detection method with domain-invariant representations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[33] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng *et al.,* "You only look one-level feature," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Online, pp. 13034–13043, 2021.

[34] C. Y. Wang, A. Bochkovskiy and H. Y. Mark, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv:2207.02696, 2022.

[35] H. Lou, X. Duan, J. Guo, H. Liu, J. Gu *et al.,* "DC-YOLOv8: Small-size object detection algorithm based on camera sensor," *Electronics*, vol. 12, no. 10, pp. 2323, 2023.

[36] L. Ale, N. Zhang and L. Li, "Road damage detection using RetinaNet," in *2018 IEEE Int. Conf. on Big Data (Big Data)*, Seattle, WA, USA, pp. 5197–5200, 2018.

[37] A. Nambiar, A. Vaigandla and S. Rajendran, "Efficient ship detection in synthetic aperture radar images and lateral images using deep learning techniques," in *OCEANS 2022*, pp. 1–10, 2022.

[38] S. Yang, W. An, S. Li, G. Wei and B. Zou, "An improved FCOS method for ship detection in SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8910–8927, 2022.

[39] C. Feng, Y. Zhong, Y. Gao, M. R. Scott and W. Huang, "TOOD: Task-aligned one-stage object detection," in *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 3490–3499, 2021.

[40] H. Guo, H. Bai, Y. Yuan and W. Qin, "Fully deformable convolutional network for ship detection in remote sensing imagery," *Remote Sensing*, vol. 14, no. 8, pp. 1850, 2022. https://doi.org/10.3390/rs14081850

[41] H. Shi, B. Chai, Y. Wang and L. Chen, "A local-sparse-information-aggregation Transformer with explicit contour guidance for SAR ship detection," *Remote Sensing*, vol. 14, no. 20, pp. 5247, 2022.

[42] Z. Xing, J. Ren, X. Fan and Y. Zhang, "S-DETR: A Transformer model for real-time detection of marine ships," *Journal of Marine Science and Engineering*, vol. 11, no. 4, pp. 696, 2023.

[43] X. Ren, Y. Bai, G. Liu and P. Zhang, "YOLO-Lite: An efficient lightweight network for sar ship detection," *Remote Sensing*, vol. 15, no. 15, pp. 3771, 2023.

[44] Z. Xu, J. Zhai, K. Huang and K. Liu, "DSF-Net: A dual feature shuffle guided multi-field fusion network for SAR small ship target detection," *Remote Sensing*, vol. 15, no. 18, pp. 4546, 2023.