remote sensing

MDPI

Article

Stepwise Attention-Guided Multiscale Fusion Network for Lightweight and High-Accurate SAR Ship Detection

Chunyuan Wang [1], Xianjun Cai [2], Fei Wu [1,*], Peng Cui [3], Yang Wu [2] and Ye Zhang [4]

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201602, China; wangcy@sues.edu.cn
[2] Shanghai Institute of Satellite Engineering, Shanghai 200240, China; caixianjun888@163.com (X.C.); cyuanw@163.com (Y.W.)
[3] Department of Computer Engineering, Dongshin University, Naju-si 58245, Republic of Korea; cuipengsql@sohu.com
[4] School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China; zhye@hit.edu.cn
* Correspondence: wufei@sues.edu.cn

Abstract: Many exceptional deep learning networks have demonstrated remarkable proficiency in general object detection tasks. However, the challenge of detecting ships in synthetic aperture radar (SAR) imagery increases due to the complex and various nature of these scenes. Moreover, sophisticated large-scale models necessitate substantial computational resources and hardware expenses. To address these issues, a new framework is proposed called a stepwise attention-guided multiscale feature fusion network (SAFN). Specifically, we introduce a stepwise attention mechanism designed to selectively emphasize relevant information and filter out irrelevant details of objects in a step-by-step manner. Firstly, a novel LGA-FasterNet is proposed, which incorporates a lightweight backbone FasterNet with lightweight global attention (LGA) to realize expressive feature extraction while reducing the model's parameters. To effectively mitigate the impact of scale and complex background variations, a deformable attention bidirectional fusion network (DA-BFNet) is proposed, which introduces a novel deformable location attention (DLA) block and a novel deformable recognition attention (DRA) block, strategically integrating through bidirectional connections to achieve enhanced features fusion. Finally, we have substantiated the robustness of the new framework through extensive testing on the publicly accessible SAR datasets, HRSID and SSDD. The experimental outcomes demonstrate the competitive performance of our approach, showing a significant enhancement in ship detection accuracy compared to some state-of-the-art methods.

Keywords: ship object detection; multiscale feature fusion; synthetic aperture radar (SAR); attention mechanism; lightweight model; FasterNet

Citation: Wang, C.; Cai, X.; Wu, F.; Cui, P.; Wu, Y.; Zhang, Y. Stepwise Attention-Guided Multiscale Fusion Network for Lightweight and High-Accurate SAR Ship Detection. *Remote Sens.* **2024**, *16*, 3137. https://doi.org/10.3390/rs16173137

Academic Editor: Dusan Gleich

Received: 1 July 2024
Revised: 21 August 2024
Accepted: 23 August 2024
Published: 25 August 2024

1. Introduction

Synthetic aperture radar (SAR) offers a distinct advantage over traditional optical or infrared imaging sensors, benefiting from the ability to operate under all weather and all time conditions. SAR is not hindered by weather variations, enabling consistent and dependable ship monitoring. This capability is pivotal for maritime traffic management, ensuring safety and facilitating prompt emergency response. Therefore, the development of accurate and robust SAR ship detection algorithms is meaningful research [1–4].

Traditional SAR object detection algorithms mainly rely on manually designed features and limited shallow learning representations, where the constant false alarm rate (CFAR) [5] stands as the most prevalent detection technique. Essentially, it is based on segmentation to categorize pixels into ship or non-ship classes and amalgamate ship pixel clusters into coherent ship regions. Conventional ship detection approaches mainly include the following classes: statistical features derived from gray-scale information, visual saliency,

template matching, and classification learning [6]. However, these algorithms are overly dependent on manually designed features, which restricts their applicability and makes them highly susceptible to the influence of background statistical variations.

In recent years, deep learning has achieved substantial advancements in domains such as object detection and image recognition [7,8]. Deep convolutional neural networks (CNNs) excel at extracting image features through training and learning a large amount of sample data to fulfill the requirements of object detection tasks across diverse scenarios. Two-stage detectors, such as RCNN [9], Fast RCNN [10], Faster RCNN [11], and Mask RCNN [12], operate by generating candidate object regions through a region proposal network (RPN). Then, these regions undergo classification and bounding box regression to yield the final outcomes. On the other hand, one-stage detectors like RetinaNet [13], SSD [14], and the YOLO series [15–18] employ a unified neural network to concurrently determine the positional coordinates and category probabilities of objects, thus having faster detection speed.

Deep CNNs, while achieving positive results in processing natural scene images, encounter challenges when applied directly to ship detection in SAR imagery. The primary obstacles include the following: (1) Most ship detection methods enhance accuracy by increasing model complexity, which often leads to a proliferation of model parameters, thereby diminishing the practical efficiency. (2) SAR images may contain complex background interferences, such as coastlines, ports, and buoys, which can be mistaken for ship objects. This confusion can degrade the accuracy of detection algorithms. (3) The size and orientation of a ship can fluctuate with variations in distance, attitude, and motion status. Consequently, the SAR reflection characteristics of ships may alter under different conditions, potentially causing an increase in false positive or false negative detections.

To overcome the above challenges, we propose a lightweight and robust detector with superior performance for multiscale ship detection in complex SAR image backgrounds, designated as SAFN. Firstly, the feature extraction module LGA-FasterNet is introduced, which embeds a lightweight global attention (LGA) mechanism into the FasterNet backbone, effectively capturing long-distance contextual relationships and enriching the model's feature representation while simultaneously reducing the model's parameters. Secondly, to achieve enhanced feature fusion, a deformable attention bidirectional fusion network (DA-BFNet) is proposed based on a deformable location attention (DLA) block and a deformable recognition attention (DRA) block. It is designed to capture multiscale contextual information, mitigate the impact of complex backgrounds in SAR images, and, thereby, further enhance the detection performance of multiscale ships.

The primary contributions of this paper are as follows:

- According to the characters of different feature levels, this study presents a stepwise attention mechanism, which robustly detects multiscale objects against complex backgrounds by extracting their discriminative features and then fusing them step by step.
- The incorporation of LGA-FasterNet facilitates both expressive feature extraction and a reduction in model parameters. Moreover, DA-BFNet innovatively integrates the DLA block with the DRA block through bidirectional connections, achieving an enhanced feature fusion.
- Extensive experiments were conducted on the SAR image datasets HRSID and SSDD. Our proposed SAFN exhibits state-of-the-art performance, with detection accuracies reaching 91.33% and 98.18%, respectively.

The rest of this paper is organized as follows. The second section reviews the relevant literature. The third section introduces the specific methodology of SAFN. The fourth section presents the experimental outcomes and analysis. The paper concludes with a summary in the fifth section.

2. Related Work

Complex models require more computing resources and hardware costs. Reducing model complexity enables the model to be deployed in environments with limited resources. Zhang et al. [19] proposed an efficient depthwise separable convolutional neural network (DS-CNN), leveraging an anchor box mechanism, concatenation strategy, and multi-scale detection approach to create a lightweight model tailored for object detection tasks. Pang et al. [20] designed a lightweight module MNEBlock based on MobileNetV3 to reduce the parameter quantity of the model and introduce a lightweight attention mechanism to improve detection accuracy while ensuring detection speed. Zhang et al. [21] proposed a lightweight detection network called ShipDeNet-20, which reduces model complexity by reducing the number of convolution layers and kernel size and using depthwise separable convolutions. Yang et al. [22] proposed a boundary-aware context module and a frequency self-attention refinement module to compensate for low precision caused by lightweight neural networks. Yang et al. [23] advanced an efficient and lightweight target detection network incorporating soft quantization, a split bidirectional feature pyramid network, and a linear transformation module. The robustness of the model is validated in several publicly available datasets.

SAR images often contain complex backgrounds, such as coastlines, ports, and buoys, which can be mistaken for ships. Additionally, natural phenomena like ocean surface waves, wind waves, and swells can interfere with SAR imagery, complicating the task of detecting ships. Xiao et al. [24] used power-based convolutional blocks to suppress speckle noise and coastal interference and designed feature alignment guidance blocks to address ship misalignment issues to enhance ship contour features and reduce speckle noise. Yao et al. [25] utilized the proposed spatial insertion attention module to enhance the feature discrimination between ships and background, encouraging the detector to focus on the localization accuracy of ship objects. Moreover, a new weighted cascaded feature fusion module was proposed to adaptively aggregate multiscale semantic features to improve the detection performance of multiscale ships in complex scenes. Lin et al. [26] improved the performance and generalization ability of traditional anchor box settings by using a keypoint-based strategy to predict bounding boxes. Meanwhile, a global context-guided feature balancing pyramid was introduced to balance semantic information at different levels to reduce the interference of speckle noise. Feng et al. [27] designed a new position-enhanced attention strategy that suppresses background interference by adding position information to the channel attention that highlights object features.

Different types of ships come in various sizes and shapes, which adds to the difficulty of detecting and identifying them in SAR images. Yu et al. [28] proposed an improved scheme based on YOLOv5, which combines coordinate attention blocks and uses a bidirectional feature pyramid network to better fuse feature information, thereby improving the detection performance. Wang et al. [29] proposed a novel NAS-YOLOX, which utilizes neural architecture search feature pyramid network and multiscale attention mechanism to solve the problems of object scattering, multiscale, and background interference. Tang et al. [30] further enhanced the ability of multiscale ship detection by introducing the ASPP module, which extracts features from multiple scales using hole convolution and spatial pyramid pooling, enabling the ship detection of different scales. Lin et al. [31] designed a Feature Enhancement Pyramid, which includes a Spatial Enhancement Module and a Feature Alignment Module to reduce the impact of scattered noise on feature extraction. The Shallow Feature Reconstruction module was introduced to enhance semantic information extraction and position description, significantly improving the accuracy of ship detection.

3. Methodology

For the purpose of achieving effective ship detection in SAR images, a lightweight detector termed SAFN is constructed, taking into account the unique characteristics of SAR imagery. The comprehensive architecture of SAFN is shown in Figure 1. Specifically, the LGA mechanism is integrated into the FasterNet backbone to more effectively capture

all relevant information about ships in complex backgrounds. Then, the proposed DLA and DRA blocks are embedded into the path fusion process to obtain an enhanced feature fusion. Further details of this architecture and its components will be displayed in the following sections.

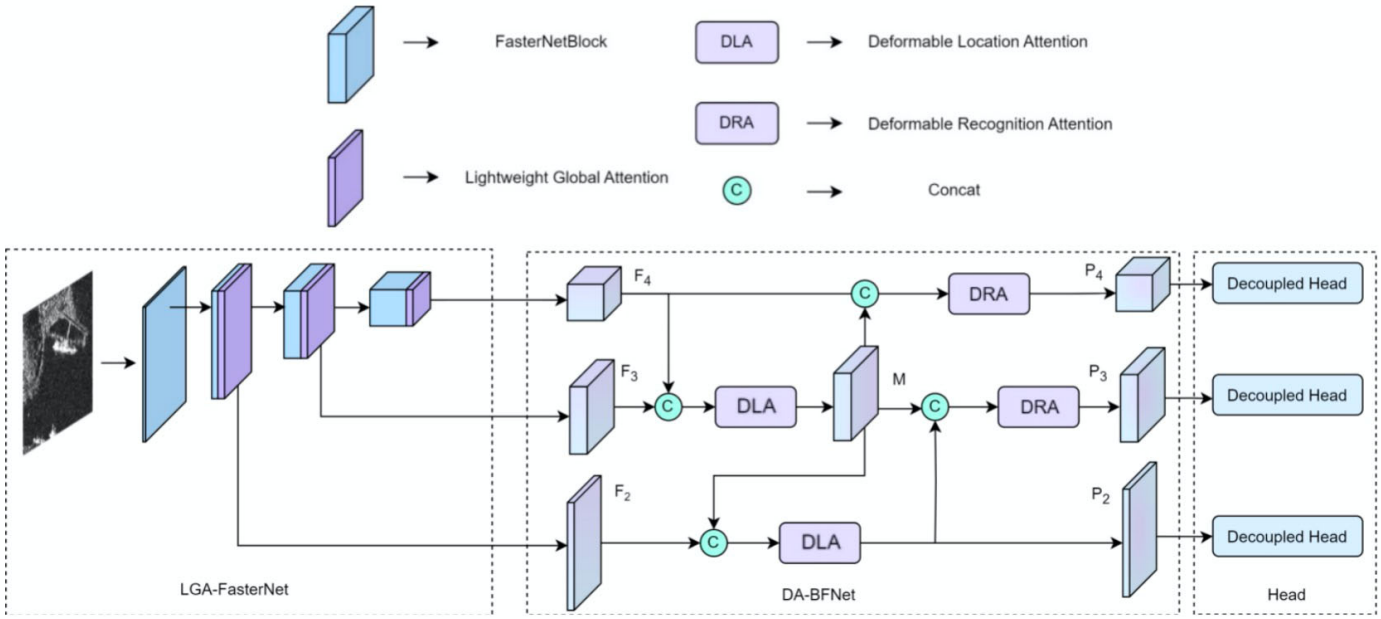


Figure 1. Overall architecture of SAFN.

3.1. Feature Extraction Module

3.1.1. Lightweight Global Attention

The scarcity of ships in SAR images renders them susceptible to interference from ground and sea clutter, especially in large scenes. Deep CNNs conventionally learn features based on local receptive fields, often ignoring the correlations that exist between non-local regions within an image. Traditional methods for modeling local context are insufficient for capturing the long-distance dependencies between an object and pixels that are distant from it. In contrast, non-local attention mechanisms excel at capturing these long-distance dependencies, providing a more comprehensive representation of the correlations between objects and their backgrounds by using global contextual information. However, conventional non-local attention mechanisms necessitate the computation of correlation weights across all positions, which can significantly amplify computational complexity. Therefore, inspired by global context attention [32], LGA is introduced by combining a non-local network [33] with the Squeeze-and-Excitation (SE) network [34]. Its structure is shown in Figure 2.

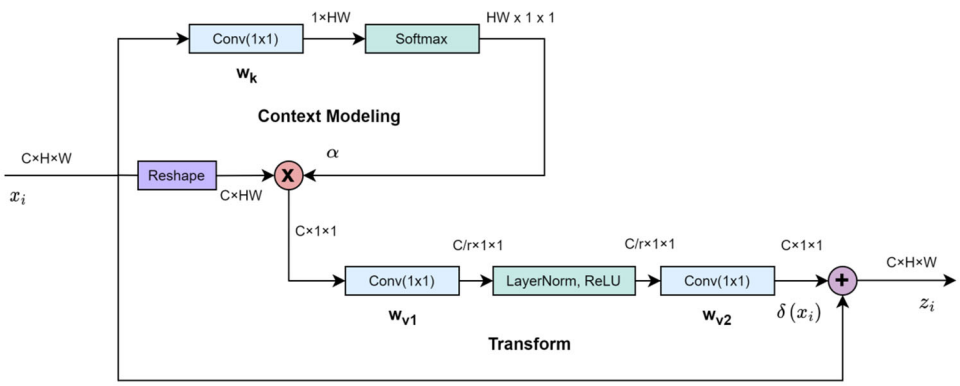


Figure 2. Structure of the LGA.

The LGA block receives the feature maps generated by the backbone network for each layer. While these feature maps are rich in local information, they often lack global contextual information. Therefore, LGA aims to introduce global contextual information to better understand the input data.

Given an input feature set $I = [x_1, x_2, \ldots, x_c] \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels, H is the height, and W is the width of the feature maps, the process begins by reshaping the feature set to $C \times HW$. A 1×1 convolution operation, denoted as W_k, is then applied, followed by a softmax function to derive the attention weights. The formulation for this process is as follows:

$$\alpha_j = \frac{e^{W_k x_j}}{\sum_{m=1}^{N_P} e^{W_k\ x_m}} \tag{1}$$

where α_j represents the jth weight of the global attention pooling, and N_p represents the total number of positions within the feature map. The global contextual features are obtained through attention pooling, which involves calculating the correlation weights between each position and all other positions across the feature map. Then, inspired by the SE mechanism, the global feature descriptor is transformed into a weight vector with the same number of channels through a 1×1 convolution, denoted as W_v. The feature transformation δ can be defined by

$$\delta(x_i) = W_{v2}\ \mathrm{RELU}\ (\mathrm{LN}(W_{v1}(\textstyle\sum_{j=1}^{N_P} \alpha_j))) \tag{2}$$

where RELU denotes the Rectified Linear Unit, and LN is LayerNorm.

Finally, the obtained global context weights are added to the original feature map to aggregate the global context features across every position within the map. The expression for LGA is as follows:

$$z_i = x_i + \delta\ (x_i) \tag{3}$$

where z_i represents the output feature layer.

3.1.2. FasterNet with Lightweight Global Attention

This work introduces FasterNet-T0, the minimal configuration of the FasterNet series [35], designed to create a lightweight and swift network specifically for SAR ship detection. The architecture of FasterNet-T0 comprises four hierarchical stages, and each stage is preceded by an embedding or merging layer that facilitates spatial down-sampling and channel expansion. The stages consist of a series of FasterNet blocks, with a higher concentration of blocks and computational resources allocated to the latter two stages. Each FasterNet block is composed of a PConv layer followed by two 1×1 Conv layers, forming an inverted residual block.

The PConv layer operates on the principle of applying filters to a subset of input channels to extract spatial features, leaving the remaining channels unaltered. For the sake of continuous or regular memory access, either the first or last set of continuous channels is used in calculations, representing the entire feature map. Assuming that the input and output feature maps maintain the same number of channels, the computational complexity of PConv, measured in terms of floating-point operations (FLOPs), is given by $h \times w \times k^2 \times {c_p}^2$, where h and w represent the height and width of the feature map, k is the kernel size, and c_p is the number of partial convolution channels used. If only $1/4$ of the channels are utilized, the FLOPs are reduced to $1/16$ of those required by traditional convolutions.

FasterNet effectively reduces the number of network parameters by incorporating PConv but essentially still uses stacked convolution operations for feature extraction, which has limitations when addressing long-distance dependencies. Therefore, we embed the LGA mechanism into FasterNet to obtain LGA-FasterNet as the backbone feature extraction network, as illustrated in Figure 3. In LGA-FasterNet, local information is captured through the stacking of FasterNet blocks, while the integrated LGA layer is tasked with acquiring global information for each feature layer. This integrated design empowers the object

detector to more effectively capture global information and long-distance dependencies of ships, compensating for the limitations of convolution operations and enhancing the ship recognition capability while suppressing background interference.

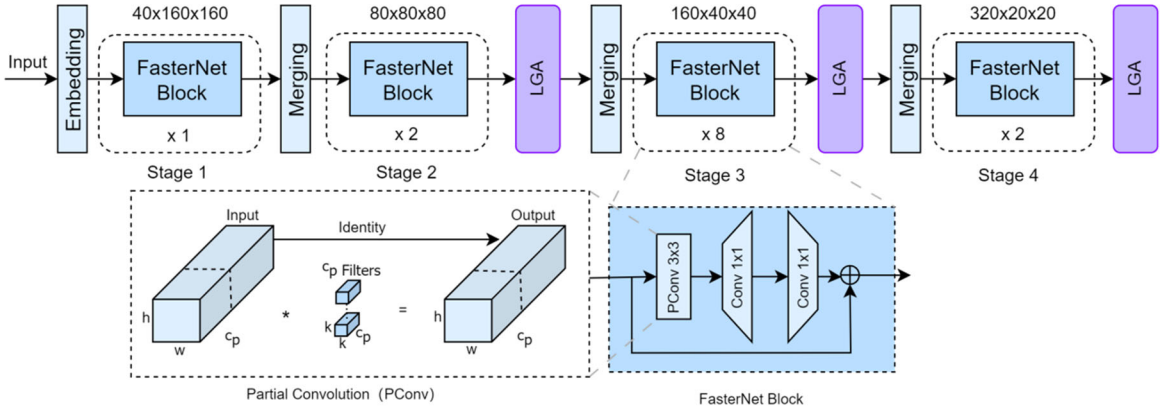


Figure 3. Structure of the LGA-FasterNet.

3.2. Feature Fusion Module

Cross-scale feature fusion is recognized as an effective strategy for achieving multiscale receptive fields. However, features at different scales exhibit significant semantic disparities, and a direct fusion of these features can ignore the contextual information inherent in the image. Moreover, it is a very challenging task to extract fine-grained features and positional details from small-scale objects within large-scale remote sensing imagery since convolution, sampling, and aggregation operations may result in the loss of critical information. It is essential to enhance the network's feature extraction capabilities while suppressing irrelevant background details. Drawing inspiration from the Path Aggregation Network (PAN) [36], we introduce our proposed DA-BFNet, as depicted in the middle part of Figure 1. By designing DLA and DRA blocks and bidirectionally integrating these attention mechanisms into both the shallow and deep layers of the network, the interaction between spatial and channel information is improved.

3.2.1. Deformable Convolution Network

Ships in SAR images are affected by a multitude of factors, including variations in perspective, attitude, and deformation. A deformable convolution network (DCN) [37] incorporates learnable offset parameters that enable the convolution kernel to adjust its sampling positions dynamically on the input feature map, thus accommodating irregular shapes and object deformations. The architecture of the DCN is shown in Figure 4. For each position p_0 in the output feature map y, the eigenvalue $y(p_0)$ is calculated using the following equation:

$$y(p_0) = \sum_{Pn \in R} w(p_n)(p_0 + p_n + \Delta p_n) \quad (4)$$

where $w(p_n)$ is the convolutional kernel weights, p_n represents the set of all sampling positions within the receptive field, and Δp_n signifies the offsets for each pixel, which are generated by an additional convolution operation on the input feature map and typically range from -1 to 1. The addition of offsets allows the convolutional kernel to calculate weights not only at fixed positions but also dynamically with other pixels, enhancing the flexibility of the convolution operation. As illustrated in Figure 5, the front image displays regular convolutional sampling points, and the latter one displays deformable convolutional sampling points. This mechanism demonstrates a robust modeling capability, particularly with objects undergoing shape changes.

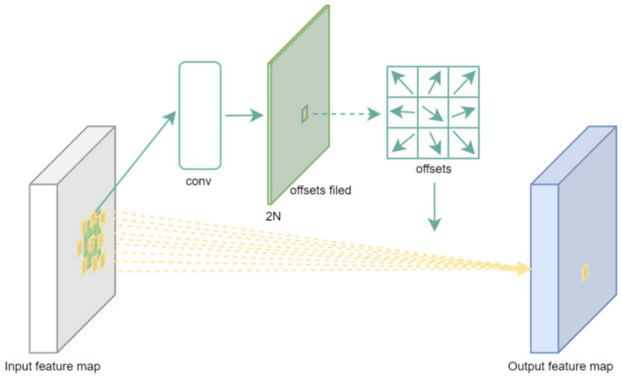


Figure 4. Deformable convolution Network.

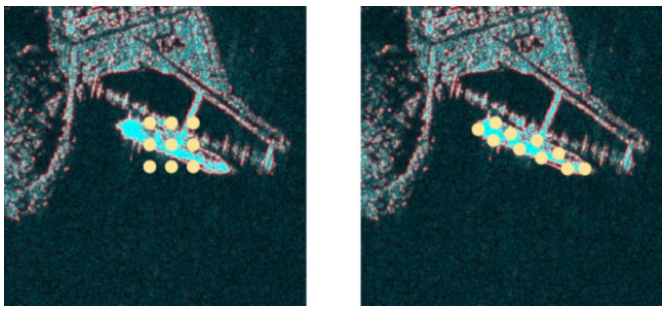


Figure 5. Comparison of sampling points.

3.2.2. Deformable Location Attention Block

Shallow feature maps excel in capturing local details and localization accuracy, making them suitable for object localization and bounding box regression tasks. To enhance the network's ability to detect and locate ships, a DLA block is designed and integrated into the shallow feature fusion stage. The structure of the DLA block is illustrated in Figure 6.

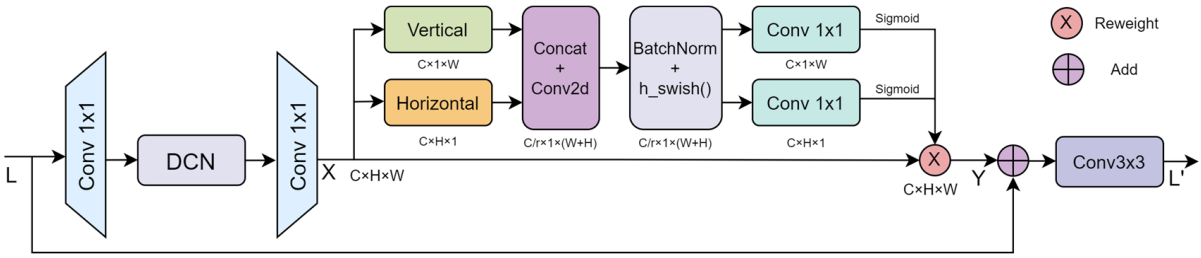


Figure 6. Structure of the DLA.

For the input feature layers L, they sequentially pass through a 1×1 convolution, a DCN, and another 1×1 convolution, where the 1×1 convolution is used to adjust the number of the input feature layers to reduce computational complexity, and the DCN is used to improve the model's accuracy and robustness in representing ships with diverse shapes and within complex scenes. The resulting feature map X undergoes spatial attention decoupling [38], initiated by two separate one-dimensional pooling operations along the horizontal and vertical directions. The outcome Y of the spatial attention is then concatenated in parallel with the input feature L through a residual structure and connected in series with a 3×3 convolution to produce the final feature map L'. The entire process is summarized as follows:

$$\begin{aligned} X &= \mathrm{Conv}_{1\times1}(\mathrm{DCN}(\mathrm{Conv}_{1\times1}(L))) \\ Y &= \mathrm{SA}(X) \\ L' &= \mathrm{Conv}_{3\times3}(L + Y) \end{aligned} \tag{5}$$

where $\text{Conv}_{1\times1}$ and $\text{Conv}_{3\times3}$ represent convolutions with convolution kernels of 1 and 3, respectively. DCN(·) denotes the DCN, and SA(·) represents the spacial attention operation. X is the input feature map subjected to pooling operations using kernels of size (H, 1) and (1, W) along the horizontal and vertical directions, respectively. This results in two feature maps with directional perception, as

$$\begin{aligned} z_c^h(h) &= \frac{1}{W}\sum 0 \le i < Wxc(h,j) \\ z_c^w(h) &= \frac{1}{H}\sum 0 \le j < Hxc(j,w) \end{aligned} \tag{6}$$

where x_c is the cth channel at vertical position h or horizontal position w. These transformations aggregate features along two spatial directions, capturing long-distance dependencies in one direction while retaining precise positional information in the other, aiding in the accurate localization of objects of interest. The two transformations are concatenated along the spatial dimension, and a 1 × 1 convolutional function is used to compress the channels, resulting in

$$f = \delta\left(\text{Conv}_{1\times1}([z^h, z^w])\right) \tag{7}$$

where $[\cdot,\cdot]$ represents the concatenation operation along the spatial dimension, δ is a nonlinear activation function, and f represents the intermediate feature map encoding spatial information in both horizontal and vertical directions. r is the reduction ratio used to control the number of output channels. The feature map f is then split into two independent tensors, f^h and f^w, along the spatial dimension. These tensors are processed through separate 1 × 1 convolutional layers to restore the original number of channels and are normalized and weighted fused using a sigmoid function σ, yielding

$$\begin{aligned} g^h &= \sigma\left(\text{Conv}_{1\times1}\left(f^h\right)\right) \\ g^w &= \sigma\left(\text{Conv}_{1\times1}\left(f^w\right)\right) \\ y_c(i,j) &= x_c(i,j) \times {g_c}^h(i) \times {g_c}^w(j) \end{aligned} \tag{8}$$

where (i,j) represent pixel position coordinates. g^h and g^w represent the attention weights for the two different spatial directions. Finally, the DLA feature is output through a residual connection and a 3 × 3 convolution, depicted as the Formula (5).

3.2.3. Deformable Channel Attention Block

Deep feature layers excel in abstracting semantic information and capturing global context, making them well-suited for object recognition and classification tasks. To further differentiate between object and interference information, a DRA block is designed and integrated into the deep feature fusion stage. The structure of the DRA block is illustrated in Figure 7.

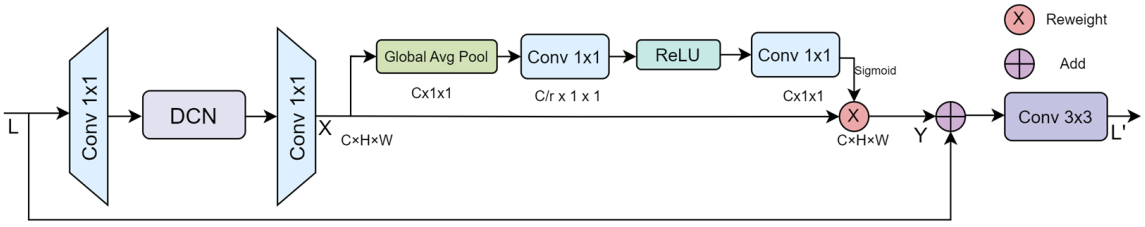


Figure 7. Structure of the DRA.

For the input features L, they undergo a series of sequential processing steps, including a 1 × 1 convolution, a DCN block, another 1 × 1 convolution, and a lightweight channel

attention module [39]. Then, a parallel residual structure is constructed, followed by a 3×3 convolution. Similarly, the entire process can be summarized as

$$\begin{aligned} X &= \mathrm{Conv}_{1\times1}(\mathrm{DCN}(\mathrm{Conv}_{1\times1}(L))) \\ Y &= \mathrm{CA}(X) \\ L' &= \mathrm{Conv}_{3\times3}(L + Y) \end{aligned} \tag{9}$$

where $\mathrm{Conv}_{1\times1}$ and $\mathrm{Conv}_{3\times3}$ represent convolutions with convolution kernels of 1 and 3, respectively. $\mathrm{CA}(\cdot)$ represents the channel attention operation. A global average feature vector is derived from global average pooling across the feature map, calculated as

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{10}$$

Then, the resulting vector z is processed using two 1×1 convolutions to generate channel weight values, given as

$$\boldsymbol{s} = \sigma\left(\mathrm{Conv}(\delta\left(\mathrm{Conv}(z)\right))\right) \tag{11}$$

where δ represents the ReLU activation function and σ represents the sigmoid activation function. The weights, represented by $\boldsymbol{s}$, are used to modulate the importance of different channels. Each channel's feature map is then multiplied by its corresponding weight to either enhance or suppress the channel's information. The output feature, which reflects the weighted contribution of each channel, can be expressed as

$$Y = \boldsymbol{s} \cdot X \tag{12}$$

Finally, the DRA feature is output through a residual connection followed by a 3×3 convolution, and the calculation formula is as shown in Equation (9).

3.2.4. Deformable Attention Bidirectional Fusion Network

Multiscale object detection methods usually use the hierarchical structure of deep CNNs to extract features, which are then sequentially passed to the detector for predicting object bounding boxes. The foundation is that deep feature layers have larger receptive fields and rich semantic information and global context awareness, making them more suitable for object recognition and classification. Meanwhile, shallow feature layers have smaller receptive fields and finer details and precise localization, making them more effective for object localization and bounding box regression. The fact is that there are substantial semantic differences across feature layers at various scales, and direct fusion of them could ignore the contextual information within the image.

Therefore, in order to enhance the information expression of feature maps, DLA and DRA blocks are embedded into the shallow and deep layers of the bidirectional fusion network, respectively. This integration allows for multilevel information to interact and enhance each other, culminating in a comprehensive feature map and adapting the model to complex scenes and multiscale object detection. The DA-BFNet, shown in the central section of Figure 1, excludes resize blocks used for image concatenation to simplify the illustration. The flow can be summarized as

$$\begin{aligned} M &= \mathrm{DLA}(\mathrm{Concat}(F_3, \mathrm{Upsampling}(F_4))) \\ P_2 &= \mathrm{DLA}(\mathrm{Concat}(F_2, \mathrm{Upsampling}(M))) \\ P_3 &= \mathrm{DRA}(\mathrm{Concat}(\mathrm{Downsampling}(P_2), M)) \\ P_4 &= \mathrm{DRA}(\mathrm{Concat}(P_4, \mathrm{Downsampling}(M))) \end{aligned} \tag{13}$$

where M denotes the middle feature maps, F is the feature maps derived from the LGA-FasterNet, P expresses the aggregated feature maps.

The architecture of this network is designed to maintain high-resolution feature extraction capabilities while effectively capturing semantic information. It places a strong emphasis on the interrelationships among different positions within feature maps. By constructing bidirectional connections that leverage diverse receptive fields, the network is capable of extracting multiscale spatial contextual information. This capability is particularly advantageous for multiscale object detection tasks, offering significant benefits for small ship detection.

4. Experimental Results and Analysis

4.1. Datasets and Training Settings

A thorough evaluation is conducted using the SSDD and HRSID datasets to ascertain SAFN reliability. These datasets encompass a variety of scenarios, including not only typical views of ships on distant seas but also in nearshore, port, and island environments. They comprise images from different sensors, resolutions, and polarizations, featuring a diverse set of scenes. The SSDD dataset includes 1160 images from three distinct types of satellites, totaling 2456 ship objects. The images vary in size, ranging from 256 $\times$ 256 pixels to 608 $\times$ 608 pixels, with resolutions that span from 1 to 15 m. We adhere strictly to the partitioning guidelines stipulated by the SSDD dataset publisher to establish the training and testing datasets for SSDD. That is, images with file suffixes 1 and 9 are used as test sets. We employ the 7:1:2 ratio commonly utilized in existing research. This allocation reserves 70% of the dataset for training, 10% for validating, and the final 20% for testing. The input image size is uniformly resized to 640 $\times$ 640. The HRSID dataset, known for its high-resolution images, contains 5604 images and 16,965 ship objects, with resolutions of 0.5 m, 1 m, and 3 m. These images originate from TanDEM-X, TerraSAR-X, and Sentinel-1B sensors. For this dataset, a 6.5:3.5 ratio for data division is adopted, dedicating 65% of the data for training and 35% for testing.

All experiments are conducted using PyTorch 1.9, implemented on an AMD Ryzen 7 4800 HS processor and a NVIDIA GeForce RTX 2060 6 GB. For data augmentation, we employ a combination of techniques, including random horizontal flipping, color jitter, and multiscale cropping, along with advanced Mosaic and MixUp enhancement strategies. The batch size is configured at 8, with the training conducted over a total of 300 epochs. The learning rate uses $lr \times \text{BatchSize}/64$, with an initial value lr of 0.01 and the cosine lr schedule. The weight decay coefficient is set to 0.0005, and the momentum for stochastic gradient descent is set to 0.9. In order to ensure the rigor and fairness of our experiments, the training details remain consistent throughout the experimental process. Figure 8 shows the loss curve with the training epochs.

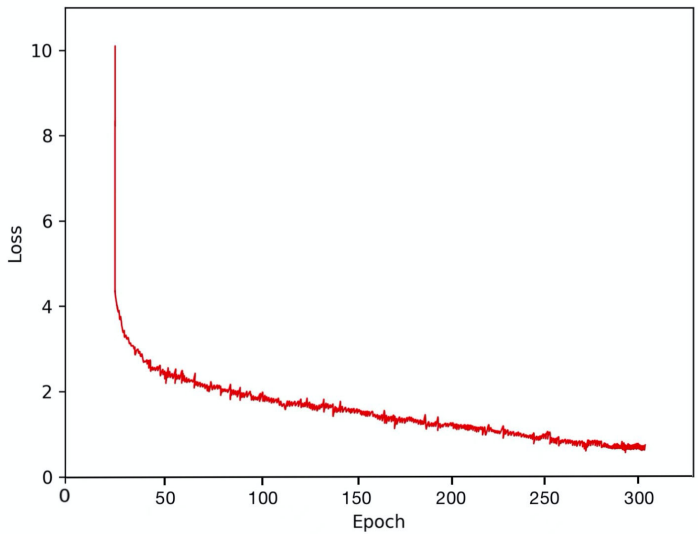


Figure 8. Training loss curve.

4.2. Evaluation Criteria

The following evaluation metrics are employed to validate the detection performance of our proposed method, including precision, recall, average precision (*AP*), model size, and the number of parameters. The metrics are calculated based on the following definitions:

$$\begin{aligned} P &= TP/(TP+FP) \\ R &= TP/(TP+FN) \\ AP &= \int_0^1 P(R)dR \end{aligned} \tag{14}$$

where *TP* represents the correctly detected ships, *FP* represents the incorrectly detected ships, and *FN* represents the missed ships. Precision is defined as the ratio of accurately predicted ships to the total number of predictions. Recall is the ratio of accurately predicted ships to the actual number of ships present. *AP* is quantified as the area under the precision–recall curve. AP_{50} is used as our evaluation metric, which is the *AP* measured at an IoU threshold of 0.5.

In addition, there are some metrics used to assess the model's complexity. Parameters represent the number of learnable parameters within the model. FLOPs represent the number of floating-point operations performed by the model during inference, with a higher FLOPs count typically indicating a more complex model

4.3. Comparison Experiment

To demonstrate the effectiveness of the proposed SAFN, we conducted comparative evaluations against several classic deep networks on the SSDD and HRSID datasets. The networks include Faster R-CNN, YOLOv3, CenterNet, YOLOv5-s, YOLOX-s, and YOLOv7-s. The experimental results, as shown in Table 1, illustrate the performance using AP^S, which denotes the AP_{50} on the SSDD dataset, and AP^H, which denotes the AP_{50} on the HRSID dataset. It can be concluded that the SAFN performs the best in terms of detection performance, computational complexity, and model parameter quantity. The SAFN's AP^S and AP^H exceed those of YOLOv7-s by 0.78% and 1.63%, respectively, indicating that our proposed approach can obtain higher-quality bounding boxes and more precise object localization.

Table 1. Comparisons with classical object detectors.

Methods	AP^S (%)	AP^H (%)	FLOPs (G)	Params (M)
Faster R-CNN	95.40	82.12	91.41	41.12
YOLOv3	96.20	85.20	61.62	77.54
CenterNet	95.10	86.59	20.40	14.21
YOLOv5-s	96.28	87.34	16.54	7.23
YOLOX-s	96.97	87.50	26.92	8.96
YOLOv7-s	97.40	89.70	18.38	5.68
Proposed method	98.18	91.33	10.98	3.70

Furthermore, to substantiate the performance of our proposed SAFN, comparisons were made with existing SAR ship detection methods, including the balance learning for ship detection method (BLNet) proposed in [40], the improved YOLOv5 and BiFPN proposed in [28], the anchor-free two-stage ship detection algorithm (ATSD) proposed in [25], the global context guide feature balance pyramid and unified attention detection algorithm (FBUANet) proposed in [26], the lightweight position-enhanced anchor-free algorithm (LPEDet) proposed in [27], the neural architecture search and multiscale attention detection algorithm (NAS-YOLOX) proposed in [29], the pyramid pooling attention network (PPANet) proposed in [30] and the feature enhancement pyramid and shallow feature reconstruction network (FEPSNet) proposed in [31]. The comparative results are detailed in Table 2.

Table 2. Comparisons with other SAR ship detectors.

Methods	AP^S (%)	AP^H (%)	FLOPs (G)	Params (M)
BL-Net [40]	95.25	88.67	41.17	47.81
YOLOv5 + BiFPN [28]	95.02	85.11	-	-
ATSD [25]	96.88	88.19	7.25	61.5
FBUA-Net [26]	96.20	90.30	71.11	36.54
LPEDet [27]	97.40	89.70	18.38	5.68
NAS-YOLOX [29]	97.2	91.10	-	44.4
PPA-Net [30]	95.19	89.27	-	-
FEPS-Net [31]	96.00	90.70	-	37.31
Proposed method	98.18	91.33	10.98	3.70

The experimental results on the SSDD dataset demonstrate the competitiveness of our method. Compared to other state-of-the-art methods, including BLNet, YOLOv5-BiFPN, ATSD, FBUA Net, LPEDet, NAS-YOLOX, FEPSNet, and PPANet, our method achieves an increase in AP^S of 2.93%, 3.16%, 1.3%, 1.98%, 0.78%, 0.98%, 2.99%, and 2.18%, respectively.

The HRSID dataset, with its more complex image backgrounds and a higher number of small ship objects, better showcases the effectiveness of our proposed SAFN. Specifically, the SAFN improves upon these state-of-the-art methods by approximately 0.23% to 6.22%. This enhancement is attributed to the designed stepwise attention mechanism, which strengthens the interaction of multilevel information and captures multiscale contextual features. By addressing the impact of multiscale objects and complex backgrounds, our method achieves optimal detection results. Compared to BLNet, YOLOX-BiFPN, ATSD, FBUANet, LPEDet, NAS-YOLOX, FEPSNet, and PPANet, improvements in AP^H are 2.66%, 6.22%, 3.14%, 1.03%, 1.63%, 0.23%, 2.06%, and 0.63%, respectively.

In addition, our SAFN has the smallest number of parameters among all compared algorithms. Although it has slightly higher computational complexity than ATSD, the AP^S on both datasets are 1.3% and 3.14% higher than ATSD, with particularly notable improvements on the HRSID dataset. These results indicate that our proposed method not only attains high detection accuracy but also maintains a relatively low parameter count and model complexity, thereby validating the effectiveness of the SAFN presented in this paper.

4.4. Ablation Experiment

To substantiate the efficacy of the blocks introduced in this paper, a comprehensive suite of ablation experiments is conducted, as detailed in Table 3. In order to illustrate the lightweight and effectiveness of the LGA block, comparative analyses are made with SE and CBAM modules, as depicted in the second set of experiments in Table 3. It is observed that LGA and SE possess a similar parameter count, yet LGA demonstrates superior detection accuracy. Although CBAM exhibits a marginal gain in accuracy, ranging from 0.04% to 0.06%, it comes at the cost of a higher parameter volume. Consequently, LGA achieves an optimal equilibrium between model conciseness and precision. Furthermore, to showcase the efficiency and potency of the FasterNet backbone, it is compared with MobileNetV3 and GhostNet. The third set of experiments tabulated in Table 3 reveals that FasterNet sustains the highest detection accuracy while preserving the most streamlined parameter count. Additionally, the fourth set of experiments in Table 3 isolates the impact of incorporating the DCN block, our proposed DLA block, and the DRA block into the network individually. Each block contributes positively to the network's detection accuracy.

Table 3. Ablation experiments of some blocks used in this paper.

Experiment	AP^S (%)	AP^H (%)	FLOPs (GMac)	Params (M)
YOLOX(CSPDarknet + FPN)	96.97	87.50	26.92	8.96
LGA-FasterNet + FPN	97.64	89.06	10.70	4.50
SE-FasterNet + FPN	97.54	88.94	10.60	4.50
CBMA-FasterNet + FPN	97.68	89.12	11.00	4.70
FasterNet + FPN	97.20	88.14	10.40	4.30
MobilenetV3 + FPN	95.79	86.36	15.60	6.25
GhostNet+FPN	96.12	86.94	10.40	4.25
LGA-FasterNet + DCN-BFNet	97.82	89.93	10.80	3.65
LGA-FasterNet + DLA-BFNet	98.01	90.57	10.80	3.65
LGA-FasterNet + DRA-BFNet	97.98	90.41	10.80	3.65
SAFN(LGA-FasterNet + DA-BFNet)	98.18	91.33	10.98	3.70

In order to distinctly compare the advantages of the proposed modules, a series of ablation experiments were designed. The first experiment establishes YOLOX as the baseline. The second experiment replaces the YOLOX backbone with FasterNet, compressing the convolutional channels in the neck and detection head. The third experiment built upon the second by integrating the LGA module to assess the effectiveness of the network with embedded attention mechanisms. The fourth experiment introduced our DA-BFNet, replacing FPN in the first experiment to verify the effectiveness of multilevel information interaction and enhancement strategy. The fifth experiment represented our proposed SAFN. These sequential experiments stepwise verified the effectiveness and superiority of our method, with all experiments maintaining consistent training, validation, and test sets, as well as hyperparameters. The comparative results are presented in Table 4.

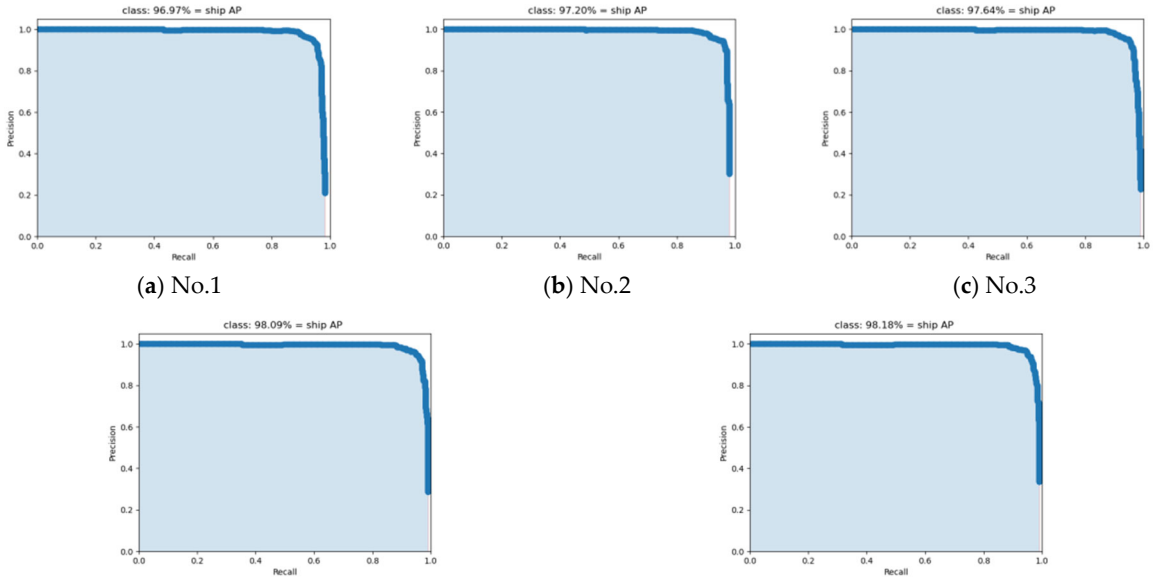
Table 4. Ablation experiment of two network modules proposed in this paper.

Experiment	AP^S (%)	AP^H (%)	FLOPs (GMac)	Params (M)
YOLOX(CSPDarknet + FPN)	96.97	87.50	26.92	8.96
FasterNet + FPN	97.20	88.14	10.40	4.30
LGA-FasterNet + FPN	97.64	89.06	10.70	4.50
CSPDarknet + DA-BFNet	98.09	89.36	26.98	8.70
SAFN(LGA-FasterNet + DA-BFNet)	98.18	91.33	10.98	3.70

In the second experiment, the backbone was switched from Darknet-53 to FasterNet-T0. This change led to a modest improvement in AP^S from 96.97% to 97.20%, with an increase of 0.23%, and in AP^H from 87.50% to 88.14%, with an increase of 0.64%. Additionally, the FLOPs are significantly reduced from 26.92 to 10.40, with a decrease of 16.52. These results indicate that the introduced backbone not only improves detection accuracy in SAR ship detection tasks but also reduces the number of parameters, resulting in a lighter model suitable for deployment on resource-limited embedded devices. In the third experiment, the addition of the LGA module to the backbone, despite a minimal increase in parameters, improves AP^S by 0.44% and AP^H by 0.92%, enhancing overall detection performance. In the fourth experiment, on the more complex HRSID dataset, the AP_{50} reaches 89.36%, which is 1.86% higher than the original model. This improvement is attributed to DA-BFNet's ability to effectively enhance information exchange across different feature layers, thereby strengthening the network's capability to model ships at multiple scales. The fifth experiment demonstrates that our proposed SAFN significantly outperforms the third experiment in terms of detection accuracy, with only a slight increase in model complexity. Notably, on the complex HRSID dataset, the AP^H improves by 2.27%. Compared to the

baseline, the total parameter count decreases by 64%, with the AP^S increasing by 1.21% and AP^H by 3.83%.

The precision–recall curves for these experiments on the SSDD dataset are depicted in Figure 9, with recall on the horizontal axis, precision on the vertical axis, and the shaded area representing the AP value. The area under the curve for our proposed SAFN is the largest, indicating the optimal performance of the novel method.



(**a**) No.1 (**b**) No.2 (**c**) No.3

(**d**) No.4 (**e**) No.5

Figure 9. Precision–recall curves of five ablation experiments.

4.5. Visualization

To further demonstrate the superiority of our SAFN, a series of visualization results are presented in Figure 10. In these visualizations, GT stands for the ground truth, and experiments No.1 to No.5 correspond to those detailed in Table 4. The red circle represents objects that are not detected, while the yellow circle signifies objects with detection inaccuracies.

Examples 1, 2, and 3 illustrate scenarios with complex backgrounds, where ships in a nearshore environment are significantly obscured due to land scattering effects.

Examples 1 and 3 depict densely packed ships, while Examples 3 and 4 showcase a multitude of small-scale ships. In Example 2, there is a considerable variation in the scale of the ships. It is evident that our method outperforms others in all instances. Specifically, for the detection of multiscale and small ships, the incorporation of LGA into the lightweight backbone markedly enhances the detection outcomes. Faced with the challenge of complex backgrounds, it is clear that our proposed DA-BFNet effectively mitigates this issue, enabling more accurate detection of nearshore ships and those in close proximity, thereby delivering superior performance. Concurrently, the DA-BFNet module has made a significant contribution to addressing the challenges of multiscale object detection. Specifically, the issues of missed and false detections, which were prevalent in objects of varying scales, have been notably mitigated. This enhancement is evident when comparing the visualization results of index No.4 with those of No.1, and similarly, when contrasting the visualization results of No.5 with those of No.3.

The visualization results provide a clear visual testament to our method's capability to effectively address key challenges inherent in SAR ship detection, exhibiting commendable performance across multiple datasets. In essence, the integration of the LGA-FasterNet and the DA-BFNet allows the model to harness multiscale contextual information and facilitate multilevel feature interaction. This novel network culminates in the precise localization and recognition of ships of varying sizes, even when set against complex backgrounds.

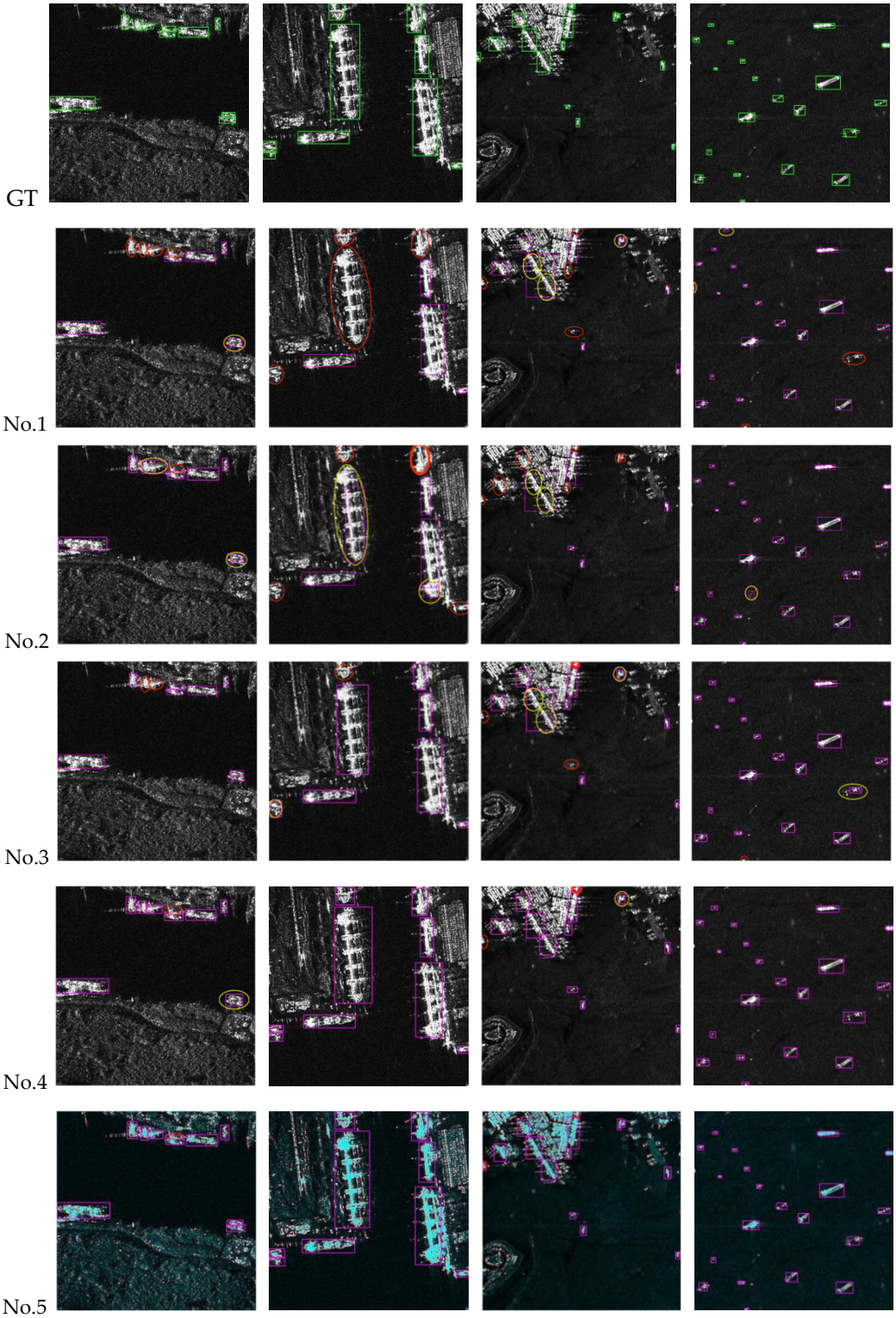


Figure 10. Visual ablation results. GT represents ground truth; labels 1–5 correspond to the row IDs 1–5 in Table 4.

5. Conclusions

Considering the distinctive attributes of SAR imagery and ship objects, and the demand for lightweight models, a novel framework SAFN is proposed based on a stepwise attention strategy. Firstly, the LGA-FasterNet is constructed by integrating global attention into FasterNet to achieve robust feature extraction while reducing the model's parameters.

Furthermore, the proposed DA-BFNet strategically integrates a DLA block with a DRA block by using bidirectional fusion, balancing the position and channel features across the network to more effectively detect multiscale ship objects within complex scenes. The experimental outcomes demonstrate that SAFN offers relatively fewer parameters and lower model complexity, concurrently enhancing detection accuracy and effectiveness.

Author Contributions: Conceptualization, C.W. and X.C.; methodology, C.W. and F.W.; software and validation, C.W., Y.W. and P.C.; writing—original draft preparation, C.W. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China under Grants 61801286 and 42375140, and Key Science and Technology Innovation Projects in Shanghai "Research on Supplementary imaging Path Planning and Analysis for 3D Realistic Construction" under Grant 22DZ1100803.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments on the paper and the builders of the datasets.

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

References

1. Yasir, M.; Wan, J.; Xu, M.; Sheng, H.; Zeng, Z.; Liu, S.; Colak, A.T.I.; Hossain, M.S. Ship detection based on deep learning using SAR imagery: A systematic literature review. *Soft Comput.* **2023**, *27*, 63–84. [CrossRef]
2. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual attention inception network for remote sensing visual question answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
3. Chen, Y.; Zhou, L.; Pei, S.; Yu, Z.; Chen, Y.; Liu, X.; Du, J.; Xiong, N. KNN-BLOCK DBSCAN: Fast clustering for large scale data. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *51*, 3939–3953. [CrossRef]
4. Jin, K.; Chen, Y.; Xu, B.; Yin, J.; Wang, X.; Yang, J. A patch to-pixel convolutional neural network for small ship detection with PolSAR images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6623–6638. [CrossRef]
5. Robey, F.C.; Fuhrmann, D.R.; Kelly, E.J. A CFAR adaptive matched filter detector. *IEEE Trans. Aerosp. Electron. Syst.* **1992**, *28*, 208–216. [CrossRef]
6. Li, J.; Xu, C.; Su, H.; Gao, L.; Wang, T. Deep learning for SAR ship detection: Past, present and future. *Remote Sens.* **2022**, *14*, 2712. [CrossRef]
7. Wang, J.; Wang, Y.; Wu, Y.; Zhang, K.; Wang, Q. FRPNet: A feature reflflowing pyramid network for object detection of remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1–5. [CrossRef]
8. Guo, H.; Gu, D. Closely arranged inshore ship detection using a bi-directional attention feature pyramid network. *Int. J. Remote Sens.* **2023**, *44*, 7106–7125. [CrossRef]
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
10. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
11. 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]
12. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
13. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
14. 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 (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, realtime object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
16. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
17. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4:Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
18. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.

19. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. Depthwise separable convolution neural network for high-speed SAR ship detection. *Remote Sens.* **2019**, *11*, 2483. [CrossRef]
20. Pang, L.; Li, B.; Zhang, F.; Meng, X.; Zhang, L. A lightweight YOLOv5-MNE algorithm for SAR ship detection. *Sensors* **2022**, *22*, 7088. [CrossRef]
21. Zhang, T.; Zhang, X. ShipDeNet-20: An only 20 convolution layers and <1-MB lightweight SAR ship detector. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1234–1238. [CrossRef]
22. Yang, X.; Zhang, S.; Duan, S.; Yang, W. An effective and lightweight hybrid network for object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 1–11. [CrossRef]
23. Yang, X.; Zhang, J.; Chen, C.; Yang, D. An efficient and lightweight CNN model with soft quantification for ship detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
24. Xiao, M.; He, Z.; Li, X.; Lou, A. Power transformations and feature alignment guided network for SAR ship detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
25. Yao, C.; Xie, P.; Zhang, L.; Fang, Y. ATSD: Anchor-Free Two-Stage Ship Detection Based on Feature Enhancement in SAR Images. *Remote Sens.* **2022**, *14*, 6058. [CrossRef]
26. Bai, L.; Yao, C.; Ye, Z.; Xue, D.; Lin, X.; Hui, M. A novel anchor-free detector using global context-guide feature balance pyramid and united attention for SAR ship detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
27. Feng, Y.; Chen, J.; Huang, Z.; Wan, H.; Xia, R.; Wu, B.; Sun, L.; Xing, M. A lightweight position-enhanced anchor-free algorithm for SAR ship detection. *Remote Sens.* **2022**, *14*, 1908. [CrossRef]
28. Yu, C.; Shin, Y. SAR ship detection based on improved YOLOv5 and BiFPN. *ICT Express* **2023**, *10*, 28–33. [CrossRef]
29. Wang, H.; Han, D.; Cui, M.; Chen, C. NAS-YOLOX: A SAR ship detection using neural architecture search and multi-scale attention. *Connect. Sci.* **2023**, *35*, 1–32. [CrossRef]
30. Tang, G.; Zhao, H.; Claramunt, C.; Zhu, W.; Wang, S.; Wang, Y.; Ding, Y. PPA-Net: Pyramid Pooling Attention Network for Multi-Scale Ship Detection in SAR Images. *Remote Sens.* **2023**, *15*, 2855. [CrossRef]
31. Bai, L.; Yao, C.; Ye, Z.; Xue, D.; Lin, X.; Hui, M. Feature enhancement pyramid and shallow feature reconstruction network for SAR ship detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1042–1056. [CrossRef]
32. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Global context networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *45*, 6881–6895. [CrossRef]
33. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. Chen, J.; Kao, S.H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
36. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
37. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308.
38. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
39. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Effificient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
40. Zhang, T.; Zhang, X.; Liu, C.; Shi, J.; Wei, S.; Ahmad, I.; Zhan, X.; Zhou, Y.; Pan, D.; Li, J.; et al. Balance learning for ship detection from synthetic aperture radar remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *182*, 190–207. [CrossRef]