remote sensing

MDPI

Article

A Feature-Enhanced Network-Based Target Detection Method for SAR Images of Ships in Complex Scenes

Yunsheng Ba, Nan Xia *, Weijia Lu and Junqiao Liu

School of Information Science and Engineering, Dalian Polytechnic University, Dalian 116034, China
* Correspondence: xianan@dlpu.edu.cn

Highlights

What are the main findings?

- The proposed feature-aware enhancement network improves edge detection accuracy, and the dynamic hierarchical extraction module enhances CNN feature capture. The use of attention gating enables effective fusion of original and enhanced features, and the entire framework is integrated with a Multigranularity Feature Weaving Network.

What is the implication of the main finding?

- This work enhances detection accuracy by augmenting edge detail information within input images, while preserving the original detection framework. It supports integration into various YOLO-based detection frameworks, thereby enhancing its practicality.

Abstract

In the context of ship target detection with Synthetic Aperture Radar (SAR) images, misdetection and missed detection are often caused by complex background interference and the variability in target size. To address these challenges, this paper proposes an innovative method based on image enhancement and feature fusion to reduce background noise and effectively handle the detection confusion caused by differences in ship sizes. Firstly, a feature-aware enhancement network is introduced, which preserves and strengthens the edge information of the target objects. Secondly, during the feature extraction phase, a dynamic hierarchical extraction module is proposed, significantly improving the feature capture ability of convolutional neural networks and overcoming the limitations of traditional fixed kernel receptive fields. Finally, a feature fusion module based on attention gating is employed to fully leverage the complementary information between the original and enhanced images, achieving precise modeling and efficient fusion of inter-feature correlations. The proposed method is integrated with the YOLOv8 detection framework for target detection. Experimental results in the publicly available SSDD and HRSID datasets demonstrate detection accuracies of 97.9% and 93.2%, respectively, thus validating the superiority and robustness of the proposed method.

Keywords: synthetic aperture radar (SAR); ship detection; image enhancement; dynamic hierarchical extraction; attention gated fusion

Check for updates

Academic Editor: Haipeng Wang

Received: 15 November 2025
Revised: 21 December 2025
Accepted: 30 December 2025
Published: 5 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Synthetic Aperture Radar (SAR), as an advanced active microwave remote sensing technology, has been extensively employed across multiple domains, including environmental monitoring, maritime surveillance, and military intelligence reconnaissance, due to

its all-weather, day-and-night operational capabilities and high-resolution imaging characteristics [1,2]. SAR can penetrate clouds, smoke, and partial vegetation cover while providing high-resolution ground imaging capabilities, making it particularly effective for target detection and recognition in complex environments. However, despite significant technological advancements in SAR over the past few decades, its application in ship detection still faces numerous challenges. Issues such as complex background interference, varying ship sizes, and low contrast between targets and backgrounds significantly impact the accuracy and robustness of detection algorithms.

Traditional SAR ship detection methods typically employ a multi-step cascade architecture: first performing preprocessing operations to improve image quality, followed by land–sea segmentation, and then using Constant False Alarm Rate [3–5], detectors or texture analysis algorithms to extract potential target regions. These methods heavily rely on manually designed features and require extensive parameter tuning to achieve optimal detection performance, resulting in limited adaptability to different scenarios and generally poor detection accuracy.

In recent years, deep learning-based detection frameworks have achieved substantial progress in SAR image ship detection through the hierarchical feature extraction capability of convolutional neural networks (CNNs) [6,7], and have been widely applied in object detection [8–10], target classification [11], and other fields. Compared with traditional methods, deep learning-based object detection algorithms demonstrate superior performance, particularly in handling factors such as variations in ship size, shape, and orientation. Their end-to-end detection approach eliminates the need for manual intervention and assumptions throughout the process.

Currently, deep learning-based methods have become mainstream for object detection, as they leverage multi-layer network architectures to learn abstract features from vast datasets, thereby improving target capture and detection accuracy. CNN-based object detection models can be categorized into single-stage and two-stage detection models, each with its own advantages. Two-stage models, represented by algorithms such as Faster R-CNN [12], Mask R-CNN [13], and Cascade R-CNN [14], are widely used due to their refined feature decoupling and cascaded optimization strategies. These models employ region proposal networks or selective search algorithms to probabilistically model the spatial distribution of targets, generating high-recall candidate regions from input images before performing target category classification and bounding box regression. Two-stage models excel in detection accuracy and robustness against interference. Additionally, their clear gradient propagation paths facilitate the integration of extended modules such as attention mechanisms and multi-task supervision. However, they still suffer from limitations such as potential quantization errors and high computational complexity, resulting in slower detection speeds that struggle to meet real-time requirements.

On the other hand, single-stage models prioritize end-to-end efficiency and structural simplicity by eliminating the region proposal step, improving the capability of real-time object detection. Representative algorithms include RetinaNet [15], SSD [16], and the YOLO series [17–19]. These models offer fast inference speeds, strong adaptability, and streamlined architectures suitable for deployment on edge devices. However, their pursuit of real-time performance often comes at the cost of reduced accuracy in detecting small targets or complex scenes. Due to the special characteristics of SAR images, the influence of wave clutter and coherent spot noise included in them brings great inconvenience to the target detection work. Moreover, in near-shore scenarios, densely arranged ships, irregular shapes and sizes, and varying orientations further impact detection performance. To address these challenges, researchers have proposed various solutions, such as SAR image

denoising and despeckling networks [20,21], context information fusion, and anchor-free detection methods [22,23], aiming to enhance detection accuracy.

Several enhanced SAR ship detection methods have also played a key role in this research area. For example, incorporating attention mechanisms or designing loss functions enables models to focus more intently on regions of interest during detection, thereby reducing false negatives [24,25]. Other methods enhance the representation capabilities of target objects, optimising feature characterisation across different levels [26,27]. Alternatively, they combine complementary information types to achieve a more comprehensive representation of features [28,29]. Furthermore, deeper integration of statistical learning with deep learning has significantly advanced developments in object detection [30].

These methods strive to capture better and utilize useful feature information for prediction, but it cannot be ignored that they may lead to the omission of useful information or the amplification of irrelevant information. To address the aforementioned issues, this paper proposes a SAR image ship target detection method based on a feature enhancement network for complex backgrounds. The enhanced fusion method proposed in this paper effectively preserves the original feature information while minimizing the impact of background noise. This approach is crucial for improving the accuracy of ship detection and localization. By connecting a preprocessing network for image enhancement and feature fusion at the front end of the detector, we enhance edge detail information of target objects and perform importance-weighted fusion of features. This paper proposes two networks: a Feature-Aware Enhancement Network (FAEN), which enhances object edges against background noise; and a Multigranularity Feature Weaving Network (MFWN), which effectively extracts and fuses features via a Dynamic Hierarchical Extraction Module (DHEM) and a Gate-control Interactive Fusion Module (GIFM). The main contributions are as follows:

1. This paper proposes an FAEN to address detection interference caused by densely arranged ships in nearshore environments. By incorporating detail enhancement, feature-guided refinement, and edge-capturing operations into the input image processing, the FAEN effectively amplifies discriminative features of target objects while suppressing background noise, thereby improving detection robustness in cluttered maritime scenes.
2. In order to solve the problem of the diversity of the SAR ship scale changes and the large number of small-sized targets, we propose an MFWN consisting of a DHEM and a GIFM. By conducting multiscale deep information mining of images and performing weighted fusion using different granularity information, this network effectively preserves complete original information while better integrating high-order features such as edge details, thereby enhancing the feature representation capability of targets.

2. Related Work

This section provides an overview of image processing and feature extraction methods for SAR images. Most target detection algorithms are improved through the use of advanced convolution techniques and appropriate attention mechanisms, which endeavor to refine the extraction and utilization of feature information. The method proposed in this paper enhances the detailed feature information and fuses it using a more efficient fusion approach for data integration between features. This method maximally exploits useful information while simultaneously minimizing interference from noise.

2.1. SAR Image Denoising and Enhancement Methods

Denoising and image enhancement of SAR images have been a hot research topic in this field. To solve the speckle noise introduced by the coherent imaging mechanism of SAR

and to improve the effectiveness and usability of the information, so as to provide a reliable basis for detection and classification tasks. Liu et al. [31] proposed a multiscale residual dense dual-attention network to better retain the texture information while denoising and to generate the denoised image by extracting different levels of features to carry out residual learning. Thakur et al. [32] designed an attention and gradient-based denoising network for SAR images in order to retain more and finer detailed features while removing the scattering noise, by connecting the features with the gradient information of the noisy image to generate feature-preserved denoised images. Tan et al. [33] proposed to recover spatial detail information by a hybrid loss function, avoiding the need for noiseless images using the CNN methods that require noiseless images for training. By proposing an end-to-end self-supervised denoising model, the quality of the image is improved while the noise reduction and feature retention outperform the traditional methods. Liu et al. [34] proposed a local and global-based two-branch SAR image denoising network, which suppresses the scattering noise while completely retaining the information of the original image, and better preserves the texture of the image while denoising. Pan et al. [35] integrated the Swim transformer block into the denoising diffusion probabilistic model to construct a noise prediction network, which improves the realism of the processed image as well as the ability to preserve the image details.

The aforementioned methods employ self-supervised models or attention mechanisms to enhance the representation of target features; however, excessive focus on a single feature may result in the loss or omission of information from other features. In this paper, we do not use complex denoising transformations when processing SAR images. Instead, we enhance the edge details of the target object to improve the perception of the information, thereby suppressing noise. We use the FAEN to enhance detailed information about the edges of the target object and attenuate interference caused by noise, while fully preserving the original feature information. This minimizes the loss of features caused by noise removal, achieving a more refined feature representation.

2.2. Multiscale Feature Extraction Fusion Method

In response to the requirement of feature representation optimization in complex scenes, multiscale feature extraction and fusion mechanisms have become a key technology to enhance the perceptual capability of the model. Through the iterative innovation of the adaptive fusion strategy, the complementary and discriminative properties of the cross-scale features have been significantly enhanced. Yu et al. [36] enhanced the standard feature pyramid network by incorporating a bottom-up augmentation path, which facilitates more effective integration of precise localization information. This design was complemented by a fully convolutional channel attention mechanism that amplifies object contours and suppresses background interference. It significantly improves detection performance while having minimal impact on computational speed. Liu et al. [37] proposed a detection network based on a multiscale feature pyramid, which introduces a spatial information focusing module to enhance the target features in the feature fusion process in order to inhibit the background interference and reduce the misjudgment. Zhou et al. [38] designed a network with fewer parameters that achieves higher detection accuracy while maintaining low memory and computational cost. The network incorporates a spatial pyramid pooling module and a weighted pyramid network known as the bidirectional feature pyramid network. This design enables better adaptation to the ship's shape characteristics and improves the accuracy of target area localization. Luo et al. [39] introduced an attention module to the YOLOv8 architecture, and proposed a reparameterized neck structure and a new loss function to improve the localization ability of the detection frame. Feng et al. [40] proposed the adoption of a backbone network based on large convolutional

kernels to enhance the remote dependencies of ships by utilizing a broader receptive field. This approach integrates a multiscale parallel convolutional kernel module with a large convolutional kernel attention module, enabling the capture of feature information at different scales while ensuring the acquisition of sufficient contextual information.

The above methods aim to better combine target information at different scale levels and avoid missing information as well as target omission, but the enhancement and fusion of key information about the target still need to be improved. In this paper, we propose the MFWN to ensure better extraction and fusion of feature information through dynamic hierarchical extraction and gated interactive fusion. The original feature information can be retained for fusion after the feature information has been fully mined, so that the feature characterization used for detection is more accurate.

3. Proposed Model

In this section, we introduce the proposed framework in detail. This paper improves upon SAR images from various sources, imaging modes and incident angle ranges in different scenarios. The proposed method achieves improved detection performance for densely clustered vessels in coastal harbours and small targets in open ocean environments. In order to reduce the influence of background interference on the detection accuracy of ship targets in SAR images, this paper proposes an SAR ship target detection method based on a feature enhancement network in a complex background, and the core improvement includes the proposed FAEN and MFWN. In this approach, the FAEN module serves as a feature enhancement network to augment the input image. Subsequently, after multiscale feature extraction by the DHEM module, the GIFM module is used to perform feature correlation fusion. The network structure is shown schematically in Figure 1. For the SAR image that will be subjected to target detection, it first passes through the image enhancement network to weaken the influence of background noise while strengthening the edge and detail information of the target, and it outputs the enhanced feature map. Secondly, in the feature extraction stage, the MFEN designed in this paper carries out deeper target feature extraction on the enhanced feature map and fully explores and extracts the potential detailed information. The extracted enhanced features are fused with the features of the original input image with importance weighting, which fully retains the original information while highlighting the key information more. Finally, the obtained enhanced images are input into the YOLOv8 model for target detection.

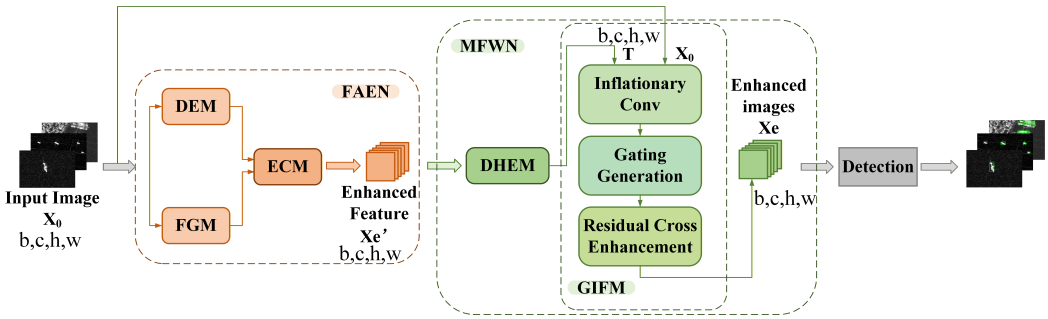


Figure 1. Flowchart of our proposed method.

3.1. Feature-Aware Enhancement Network

The design of the proposed FAEN primarily consists of three key components: the Detail Enhancement Module (DEM), the Feature Guidance Module (FGM), and the Edge Capture Module (ECM). Due to the complex scattering mechanism and the presence of speckle noise in SAR images, the signal-to-noise ratio at fine structural details is often reduced, making high-frequency detail information susceptible to being lost. To address

this issue, this paper employs a combination of these three modules to enhance the features of SAR images, as illustrated in Figure 2. The input tensors are the feature maps denoted as feature dimensions, where b represents batches, c represents channels, h and w are for height and width, respectively.

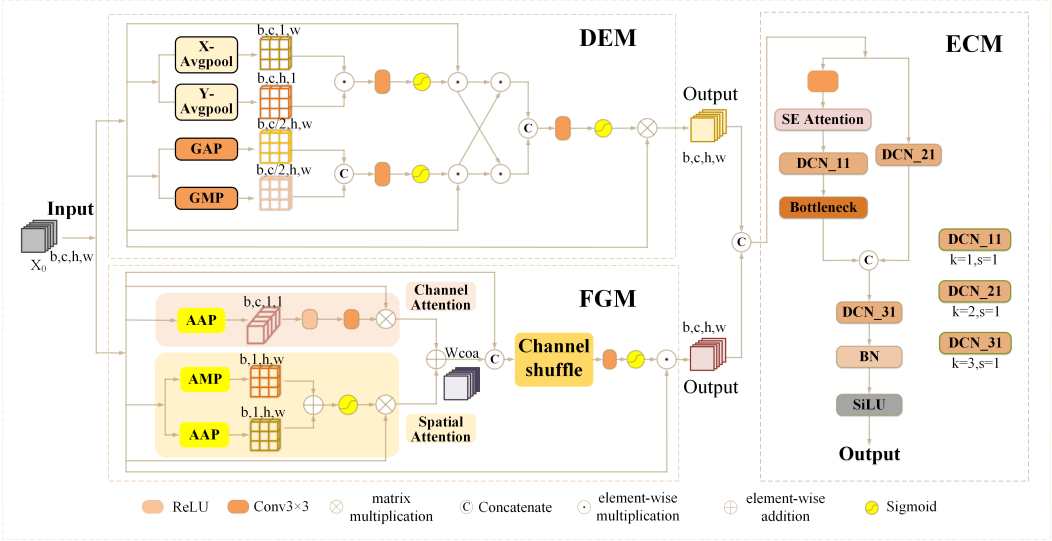


Figure 2. Diagram of FAEN.

3.1.1. Detail Enhancement Module

The DEM is designed to enhance the detailed representation of the input images. Specifically, multi-directional pooling captures edge and texture information along different orientations, while max pooling and average pooling provide global contextual information, thereby improving the overall feature representation. Through a combination of various pooling and convolutional operations, the DEM enhances image details at multiple scales and orientations, resulting in a more comprehensive feature representation. The feature process is shown in the following:

$$F_1 = \sigma(Conv_{1\times1}(X_AvgPool(X_o) \odot Y_AvgPool(X_o))) \quad (1)$$

$$F_2 = \sigma(Conv_{1\times1}(Concat[GAP(X_o), GMP(X_o)])) \quad (2)$$

where F_1 represents the features obtained after multi-directional pooling, F_2 represents characteristics obtained by different pooling methods, $Conv_{i\times i}(\cdot)$ represents a convolution with kernel i, X_o represents input image features, σ is the activation function Sigmoid, $X_AvgPool(\cdot)$ represents average pooling operation along the horizontal direction, $Y_AvgPool(\cdot)$ represents average pooling operation along the vertical direction, $\odot$ represented element-wise multiplication operation, $GAP(\cdot)$ is the global average pooling operation, $GMP(\cdot)$ is the global maximum pooling operation, respectively.

Additionally, the use of activation functions enables the compression and selection of relevant features, explicitly amplifying important characteristics. Element-wise multiplication and matrix multiplication facilitate the fusion of diverse features, preserving original information while significantly enhancing detail representation. The process of information interaction is shown in the following formulas:

$$F_1' = (F_1 \odot X_o) \odot F_2 \quad (3)$$

$$F_2' = (F_2 \odot X_o) \odot F_1 \quad (4)$$

where F_1' represents the original weighted crossover features of F_1, and F_2' represents the original weighted crossover features of F_2.

The results derived from various feature processing methods are element-wise multiplied with the input features, ensuring that the obtained outcomes align more closely with the original features while simultaneously enhancing detailed feature information. Subsequently, cross-multiplication is applied to selectively retain and choose the features, thereby strengthening the interaction between features processed through different methods.

The final output of the DEM is shown in Equation (5):

$$F = \sigma(Conv_{1\times1}(Concat[F_1^{'}, F_2^{'}])) \times X_o \tag{5}$$

where F represents output of DEM, and $Concat[\cdot]$ is Concatenate operation, $\times$ represents matrix multiplication.

Through multiscale pooling, the network learns to recognise features of the target vessel that exhibit spatial continuity and multiscale similarity. These characteristics are absent in false alarm sources. Combining multiple pooling methods simultaneously smooths local noise and distinguishes between persistent and transient target features. This preserves the most prominent target characteristics.

The results of the two processes are concatenated to more effectively capture the information in the image, and are then multiplied once again with the original features to ensure maximal preservation of the features.

3.1.2. Feature Guidance Module

The significance of the FGM is to realize the retention and mining of different levels of information in the image by generating attentional features and fully integrating them with the original features. The module is composed of a channel attention branch and a spatial attention branch, which adaptively select the more concerned feature information from the channel and spatial perspectives to process the feature information in their respective domains, so as to emphasize the selected features more. By jointly employing these techniques, we achieve a more complete and discriminative feature representation in the captured images.

$$G_1 = ReLU(Conv_{1\times1}(AP(X_o))) \odot X_o \tag{6}$$

$$G_2 = \sigma(Concat[MP(X_o), AP(X_o)]) \odot X_o \tag{7}$$

where G_1 represents the output of channel attention, G_2 represents the output of spatial attention, $ReLU$ represents the activation function ReLU, $AP(\cdot)$ is the average pooling operation, and $MP(\cdot)$ is the maximum pooling operation.

Coarse-to-Fine operation [41] is an incremental information processing approach, whose strategy is to optimize feature representations step by step from global semantics to local details by using multi-stage gradual refinement, denoted as $W_{coa}(\cdot)$ in this paper. The final output of the FGM performs the following fusion operations:

$$G = \sigma(Conv_{1\times1}(Ch_s(W_{coa}(Concat[G_1, G_2], X_o)))) \odot X_o \tag{8}$$

where G represents output of FGM, and $Ch_s(\cdot)$ represents Channel shuffle.

The proposed method first conducts a global analysis to estimate the target's coarse location and semantic category, followed by integrating local contextual information to optimize boundary localization and fine-grained feature representation. Channel attention suppresses background-related channels, while spatial attention highlights areas that may contain vessels and suppresses regions that are prone to false alarms. When combined with the channel shuffle, only locations that are significant in both the spatial and channel dimensions receive enhancement, thereby preventing misjudgements.

This hierarchical approach reduces computational overhead associated with high-resolution processing, minimizes redundant operations, enhances robustness in complex environments, and addresses key challenges in SAR-based ship detection, including multi-scale variability, high noise levels, and real-time processing requirements.

3.1.3. Edge Capture Module

The ECM integrates processing results from DEM and FGM while further enhancing feature representation. Through synergistic design of multiscale dynamic convolution and attention mechanisms, ECM constructs a hierarchical edge feature extraction and enhancement network. It operates through the Squeeze-and-Excitation (SE) [42] process, effectively modeling the interdependencies between channels. This enables the SE block to dynamically highlight informative features while diminishing the influence of less relevant ones. The schematic diagram of the SE module is shown in Figure 3. Joint attention guidance amplifies high-response edge channels, improving edge continuity for multiscale targets.

The deformable convolution's kernel offset mechanism efficiently models ship target rotation scaling and deformation while preserving edge details through local refinement. Three-level deformable convolution enables dynamic adaptation of receptive fields to target geometry, reducing convolutional edge localization errors. The ECM employs two asymmetric branches for feature extraction, effectively leveraging their complementary characteristics to achieve more comprehensive feature representation. The complex branch is capable of performing intricate transformations, while the simple branch helps preserve richer original information, thereby enhancing feature diversity and representation robustness.

The ECM module begins with a two-branch processing path. It employs Deformable ConvNet (DCN) at different levels, fused with the SE attention mechanism, to process one part of the input, resulting in M_1.

$$M_1 = Concat[Bott(DCN_{(1,1)}(SE(Conv_{1\times1}(G))), DCN_{(2,1)}] \tag{9}$$

where $Bott(\cdot)$ represents bottleneck, $DCN_{(i,j)}$ represents Deformable ConvNets with convolution kernel of size i and stride of j, and $SE(\cdot)$ represents SE attention mechanism.

Subsequently, another $k = 3$ deformable convolution, followed by Batch Normalization (BN) and a Sigmoid-Weighted Linear Unit (SiLU) activation, is applied to obtain the final output of the ECM, denoted as X'_e. The specific process is described in the following equations:

$$X'_e = SiLU(BN(DCN_{(3,1)}(M_1))) \tag{10}$$

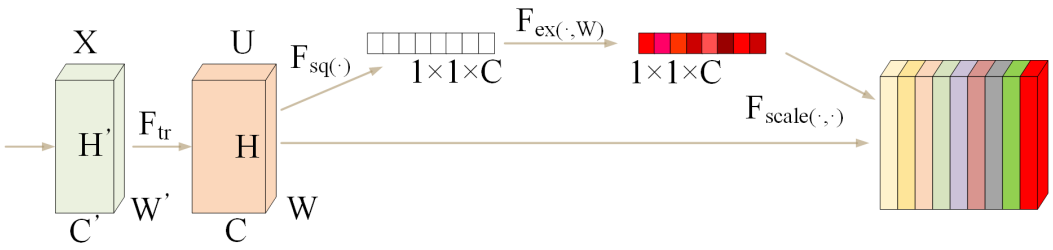


Figure 3. Schematic of the SE module.

3.2. Multigranularity Feature Weaving Network

MFWN, composed of a DHEM and a GIFM, facilitates comprehensive information interaction and fusion between the input image and its enhanced counterpart. The resulting weighted feature maps are subsequently fed into the detector for target identification. The MFWN module creates a structured, multi-path gradient enhancement system. Its core

mechanism uses parallel heterogeneous convolutional layers (7×7, 5×5 and 3×3) to extract features at multiple scales, dynamically fusing them through a learnable gated weighting network. The gating is modulated by attention weights generated by the sigmoid function. The module also incorporates deformable convolutions for adaptive spatial sampling. Combined with batch normalisation, this stabilises activation distributions and ensures gradient pass-through.

At the gradient propagation level, the design establishes a gradient multi-path distribution and convergence mechanism: gated weights enable differential adjustment of contributions across branches, residual connections guarantee identity mapping for low-frequency gradients, and offset learning in deformable convolutions adapts gradients further to geometric deformations. Overall, the MFWN integrates structured gradient path diversification with adaptive gating to enhance feature representation capabilities, ensuring the richness, flexibility and stability of gradient flow during training.

The architecture of MFWN is illustrated in Figure 4, with the following technical details:

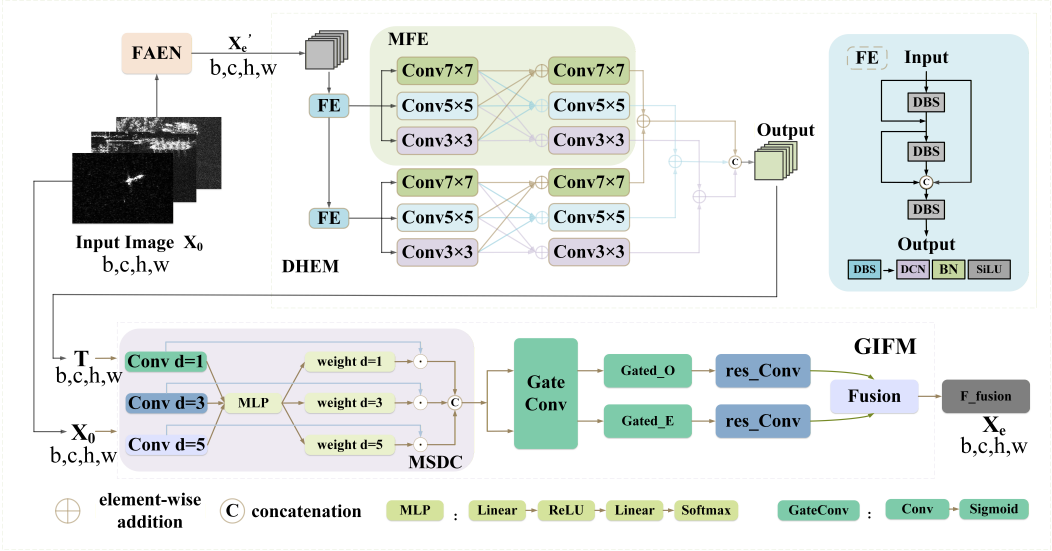


Figure 4. Detailed diagram of MFWN.

3.2.1. Dynamic Hierarchical Extraction Module

DHEM comprises four key components: the Feature Extraction (FE) module, the Multiscale Feature Extraction (MFE) module, matrix addition, and channel concatenation.

The FE module integrates three parallel Deformable Block Structures (DBSs) at distinct hierarchical levels, each composed of BN and a SiLU activation. By employing deformable convolutions with varying kernel sizes, the FE module adaptively captures multiscale edge structures, significantly improving the network's robustness to edges of varying thickness. The mathematical formulation of the DBS module is provided in the following equation, and x represents the input information processed by the model.

$$DBS(x) = SiLU(BN(DCN_{(3,1)}(x))) \tag{11}$$

The proposed MFEN employs a two-stage FE module architecture for hierarchical feature extraction. The outputs from both stages are subsequently processed by the MFE module to capture discriminative features at varying scales. To address challenges of edge blurring and low image quality in conventional edge detection methods, the MFE module is specifically designed to enhance feature robustness while suppressing noise interference. Compared to ResNet's residual connections [43], our architecture demonstrates superior performance in reducing network depth while maintaining multiscale coverage.

$$H_1 = Concat[DBS(X'_e), X'_e] \tag{12}$$

$$H_2 = Concat[DBS(H_1), H_1, X'_e] \tag{13}$$

where X'_e represents the features of the enhanced image.

The results obtained from the processing of the above equation are further utilized for more refined feature extraction through DBS, enhancing the richness of the feature representation extracted by the FE module. The FE module is described in the following equation:

$$H = DBS(H_2) \tag{14}$$

where H represents the output of FE.

The MFE module utilizes parallel convolutional branches with kernel sizes of (3, 5, 7) to extract multiscale features, where the hybrid receptive fields significantly improve contextual modeling and feature discriminability. Through progressive refinement of these multiscale convolutional layers, the module achieves hierarchical feature optimization, complementary multi-path feature preservation and adaptive scale fusion through learned weighting. The first stage of the MFE process is given in the following equation:

$$\begin{aligned} T_1 &= \mathrm{Conv}_{7\times7}(\mathrm{Conv}_{7\times7}(H') + \mathrm{Conv}_{5\times5}(H') + \mathrm{Conv}_{3\times3}(H')) \\ T_2 &= \mathrm{Conv}_{5\times5}(\mathrm{Conv}_{7\times7}(H') + \mathrm{Conv}_{5\times5}(H') + \mathrm{Conv}_{3\times3}(H')) \\ T_3 &= \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{7\times7}(H') + \mathrm{Conv}_{5\times5}(H') + \mathrm{Conv}_{3\times3}(H')) \end{aligned} \tag{15}$$

Similarly, the MFE process in the second stage is given by the following equation:

$$\begin{aligned} T'_1 &= \mathrm{Conv}_{7\times7}(\mathrm{Conv}_{7\times7}(H'') + \mathrm{Conv}_{5\times5}(H'') + \mathrm{Conv}_{3\times3}(H'')) \\ T'_2 &= \mathrm{Conv}_{5\times5}(\mathrm{Conv}_{7\times7}(H'') + \mathrm{Conv}_{5\times5}(H'') + \mathrm{Conv}_{3\times3}(H'')) \\ T'_3 &= \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{7\times7}(H'') + \mathrm{Conv}_{5\times5}(H'') + \mathrm{Conv}_{3\times3}(H'')) \end{aligned} \tag{16}$$

where H' denotes the output after passing through one FE module, while H'' represents the output processed by two consecutive FE modules.

The features obtained after multiscale extraction are subjected to different levels of summing and splicing operations to realize the full exploitation of the semantics contained in the enhanced image. The specific process is shown in the following equation:

$$T = Concat[(T_1 + T'_1), (T_2 + T'_2), (T_3 + T'_3)] \tag{17}$$

where T represents the fused features after DHEM.

The resulting multiscale feature maps from DHEM, enriched with comprehensive semantic information, are then fed into the subsequent feature fusion stage.

3.2.2. Gate-Control Interactive Fusion Module

The processed feature maps obtained through FAEN and DHEM contain rich global high-level and low-level information, where both prominent object details and typically overlooked critical features are significantly enhanced. To effectively integrate these features, we propose a GIFM for advanced feature fusion. The fusion process operates on the semantically enriched features T and the original input image X_0 with preserved raw characteristics. These two distinct features are processed independently through GIFM. This further optimises their respective feature representations to prepare them for subsequent, more effective fusion.

GIFM employs multiscale dynamic convolution (MSDC) with parallel dilated convolutions (d = 1, 3, 5) to capture local details, regional context, and global structures, respectively. A Multi-Layer Perceptron (MLP)-based weighting mechanism (comprising dimension reduction, SiLU activation, restoration, and softmax normalization) dynamically adjusts branch contributions, enabling comprehensive multi-resolution fusion while maintaining original feature integrity and enhancing detail representation capabilities. Specifically, the process is shown in Formula (18):

$$P = Softmax(Lin_{up}(ReLU(Lin_{down}(F)))) \tag{18}$$

where F denotes the input features for this processing stage, Lin_{down} performs dimensionality reduction, and Lin_{up} executes dimensionality restoration.

The multiscale features are then adaptively fused through dynamic weighting, with the complete fusion process mathematically formulated in Equation (19):

$$F_{ms} = \sum_{r} P^r \cdot Conv^r_{1\times1}(F) \qquad (r = 1, 3, 5) \tag{19}$$

where $d = r$ represents the dilation rate of convolution r, and P^r represents the weights associated with the convolution at a dilation rate of r.

The proposed architecture employs expanded receptive fields during feature fusion to prevent small-target information loss typically caused by downsampling operations. By implementing multiscale dynamic convolution, our method overcomes the limitations of conventional fixed-weight fusion approaches through scene-adaptive feature scale selection. The enhanced multiscale outputs from both original and augmented features are channel-wise concatenated and processed through a novel cross-feature guided gating mechanism, which facilitates effective information interaction while preserving spatial weighting capabilities and eliminating the local information blurring associated with pooling operations. This gating system dynamically generates spatial attention weights through a combination of 1×1 convolutions and Sigmoid activation, with the complete gating signal formulation provided in the following equations:

$$Gated_{(o)} = \sigma(Conv_{1\times1}(Concat[x_{ms_o}, T])) \tag{20}$$

$$Gated_{(e)} = \sigma(Conv_{1\times1}(Concat[x_{ms_e}, X_0])) \tag{21}$$

where $Gated_{(\cdot)}$ represents features after gating fusion, x_{ms_o} represents the result obtained by MSDC after processing the original feature X_0, and x_{ms_e} represents the result obtained by MSDC processed after processing the enhanced features T.

The proposed module demonstrates superior performance in simultaneously preserving fine-grained details of small targets, enabling content-adaptive feature selection, and maintaining spatial fidelity during cross-feature information flow. The selected features through gating are further processed via residual cross-enhancement, a mechanism that simultaneously maintains the integrity of original semantic representations and promotes cross-feature interaction. The process is described in the following equations:

$$F_{res(o)} = Gated_{(o)} + X_0 \tag{22}$$

$$F_{res(e)} = Gated_{(e)} + T \tag{23}$$

$$X_e = Concat[F_{res(o)}, F_{res(e)}] \tag{24}$$

where $F_{res(\cdot)}$ represents the result obtained by residual concatenation of the remaining original features after gating fusion, and X_e represents the final output features.

Final feature fusion is realized by GIFM processing, which allows better attention to target detail information in subsequent target detection.

The MFWN performs full information interactive fusion between the input image and the enhanced image, which consists of the DHEM and the GIFM, then inputs the feature maps resulting from the weighted fusion into the detector for target detection.

To summarize, the proposed framework enables efficient processing via the synergistic collaboration of the aforementioned modules. For a comprehensive description of its execution flow, the full pseudocode is outlined in Algorithm 1.

Algorithm 1 Algorithm of the proposed method.

Input image X_o
Obtain enhanced image X'_e with FAEN
$F \leftarrow \mathrm{DEM}(X_o)$, Eqn. (1) to (5).
$G \leftarrow \mathrm{FGM}(X_o)$, Eqn. (6) to (8).
$X'_e \leftarrow \mathrm{ECM}(F + G)$, Eqn. (9) to (10).
Perform MFWN that consists of DHEM and GIFM
$T \leftarrow \mathrm{DHEM}(X'_e)$, Eqn. (11) to (17).
The operation of GIFM includes the following steps:
$F_{ms} \leftarrow$ Inflationary Convolution(T), Eqn. (18) to (19).
$Gated_{(\cdot)} \leftarrow$ Gating Generation(F_{ms}), Eqn. (20) to (21).
$F_{res} \leftarrow$ Residual Cross Enhancement$(Gated_{(\cdot)})$, Eqn. (22) to (23).
$X_e \leftarrow \mathrm{Fuse}(F_{res})$, Eqn. (24).
Detection
Output **Detection Result** $\leftarrow \mathrm{Detection}(X_e)$

4. Experiments

In this section, we present the experimental results in detail. We will provide comprehensive information regarding the dataset, the parameters of the equipment used in the experiments, the evaluation metrics, and a comparative analysis of the results obtained from different algorithms. Additionally, ablation studies will be conducted to further evaluate the performance.

4.1. Datasets

We conducted training and testing on two image datasets: SSDD [44] and HRSID [45]. The SSDD dataset is the first publicly available benchmark for detecting vessel targets using synthetic aperture radar (SAR) imagery. It comprises a substantial collection of images of vessel targets captured in various locations, including the East and South China Seas, Bohai Bay, European ports and the Singapore Strait. These images were acquired using multiple satellites, including Sentinel-1 and TerraSAR-X, and feature a variety of non-uniform resolutions. The dataset contains 1160 images and 2456 instances of vessels, covering various environments such as coastal, offshore, and port areas.

In contrast, the HRSID dataset comprises 5604 high-resolution SAR images collected from various sensor sources, primarily at resolutions of 0.5 m, 1 m, and 3 m. It contains 16,951 vessels, with smaller targets rendered more distinctly. This dataset presents a diverse collection of vessel target imagery predominantly sourced from critical ports and waterways in China, Europe and the Americas. The images were generated by various radar systems and employ multiple polarisation methods within complex scenes. On average, each image in the HRSID dataset contains three vessel targets, distributed as follows: 2.0% large, 43.5% medium, and 54.5% small.

Details of the number of images and ship instances in each dataset are provided in Table 1. The experimental dataset is divided into training, validation and test sets according to the ratio of 7:2:1, and the size of the input images is 646 × 640.

Table 1. The number of images in the dataset and the number of target objects.

Datasets	Number of Images	Number of Ships
SSDD	1160	2358
HRSID	5604	16,951

4.2. Evaluation Criteria

Numerous target detection algorithms based on neural networks have been developed to meet the requirements of various application scenarios. These algorithms are often evaluated based on different criteria, primarily focusing on accuracy. In this study, five evaluation metrics are employed to assess the network's performance. These metrics are as follows:

1. Recall and Precision: P represents precision, and R represents recall. TP refers to the instances where both the prediction and the label are identified as ships, while FP denotes cases where the prediction is a ship but the label corresponds to the background. Conversely, FN represents instances where the prediction is background, but the label is a ship. The specific formulas for calculating these metrics are provided below:

$$P = \frac{TP}{TP + FP} \tag{25}$$

$$R = \frac{TP}{TP + FN} \tag{26}$$

2. F_1 is the harmonic average of precision and recall. The larger the F_1 value, the better the performance of the model. The F_1 score can be expressed as follows:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{27}$$

3. Mean Average Precision (mAP): mAP represents the area under the recall–precision curve. It serves as a comprehensive performance metric for the network, considering both precision (P) and recall (R). As such, mAP is regarded as a more authoritative evaluation criterion in target detection, with higher values indicating better detection accuracy. mAP provides a holistic measure of the network's overall performance improvement. The calculation formula for mAP is as follows:

$$AP = \int P(R) \mathrm{d}R \tag{28}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{29}$$

4. Parameters: Parameters describe the complexity of the overall model. In general, the smaller the number of network parameters, the easier it is for them to embed in the device.
5. Floating Point of Operations: FLOPs quantify the computational complexity of a model, they represent the total number of floating-point operations required to execute a model's forward pass. A lower FLOP count generally indicates a more computationally efficient model, requiring less processing power and memory, making it more suitable for deployment on resource-constrained devices.

4.3. Implementation Details

All experiments were performed on a hardware platform featuring an NVIDIA GeForce RTX 3090 GPU (24GB memory), utilizing a software environment comprising PyTorch 1.12.1, CUDA 11.3, and Ubuntu 18.04.6 LTS. The model was optimized using

Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, a momentum of 0.9, a batch size of 4, and a weight decay coefficient of 0.0005. The training protocol involved 100 epochs for each dataset, ensuring convergence while preventing overfitting through the application of early stopping monitoring.

4.3.1. Comparison with Other Methods

We compared the proposed framework with several state-of-the-art detection methods for SSDD, including Faster R-CNN [12], SSD [16], YOLOv7-tiny [17], YOLOv8n [18], YOLOv12s [19], FASC-NET [36], MFPNet [37], CRAS-YOLO [38], SHIP-YOLO [39] and LKE-Det [40]. To ensure consistency and reliability in our comparison, all the aforementioned methods were trained using the same datasets, with identical hardware and software conditions. The comparative results are presented in Table 2.

Table 2. Comparison results of existing methods on the SSDD dataset.

Method	P (%)	R (%)	mAP (%)	F_1 (%)	Params (M)	FLOPs (G)
Faster R-CNN [12]	86.3	88.6	87.5	87.4	32.8	792.8
SSD [16]	86.2	87.8	87.5	87.0	23.8	34.3
YOLOv7-tiny [17]	95.7	91.1	96.1	93.3	6.0	13.0
YOLOv8n [18]	95.4	94.4	96.9	94.9	3.0	8.1
YOLOv12s [19]	97.8	96.9	97.9	97.4	9.3	21.4
FASC-NET [36]	95.6	90.0	96.6	92.7	**0.64**	–
MFPNet [37]	97.0	91.0	97.1	93.9	1.0	–
CRAS-YOLO [38]	97.3	95.5	98.6	96.4	10.3	19.7
SHIP-YOLO [39]	97.1	94.5	97.1	95.8	2.5	**7.2**
LKE-Det [40]	96.0	**97.6**	97.4	96.8	33.53	42.7
Ours	**97.9**	96.1	**98.8**	**96.9**	3.4	11.6

The optimal results of the corresponding indicators are expressed by thickening.

Based on the quantitative analysis of the SSDD dataset presented in Table 2, CRAS-YOLO [38] exhibits strong detection performance; however, it has a higher number of parameters and FLOPs compared to the proposed network. SHIP-YOLO [39], with the lowest number of FLOPs, and FASC-NET [36], with the lowest number of parameters, demonstrate lower detection accuracy than the proposed method. Thus, the proposed method outperforms theirs in terms of detection performance. Although YOLOv12 [19] achieves comparable detection performance to the proposed method, its substantial computational demands compromise real-time capabilities.

Furthermore, to validate the effectiveness of the proposed method, we conducted additional experiments on the HRSID dataset using the same strategy. This enables a more thorough evaluation of the algorithm's performance by examining its behaviour on an alternative dataset. We employed precision, recall, mean average precision (mAP) and the F1 score as evaluation metrics. In this section, we compared the proposed method against superior representative algorithms. We evaluated the performance of the algorithms by comparing the corresponding metrics. Detailed metrics are presented in Table 3.

As previously noted, our algorithm outperformed the comparison algorithms in nearly all metrics, while only slightly increasing the number of parameters and computational complexity. We also demonstrate superior performance on different datasets, thereby validating the effectiveness and feasibility of our proposed approach.

Table 3. Comparison results of existing methods on the HRSID dataset.

Method	*P* (%)	*R* (%)	*mAP* (%)	F_1 (%)
Faster R-CNN [12]	81.5	73.1	77.6	77.1
SSD [16]	80.1	73.2	81.5	76.5
YOLOv7-tiny [17]	87.5	78.4	88.1	82.7
YOLOv8n [18]	93.1	83.6	92.6	88.1
YOLOv12s [19]	93.1	82.9	**92.8**	87.7
FASC-NET [36]	88.1	74.9	84.7	81.0
MFPNet [37]	92.8	79.5	89.8	85.7
CRAS-YOLO [38]	93.0	83.5	88.5	88.1
SHIP-YOLO [39]	92.9	83.1	88.0	87.7
LKE-Det [40]	92.1	82.9	91.8	87.2
Ours	**93.2**	**83.7**	**92.8**	**88.2**

The optimal results of the corresponding indicators are expressed by thickening.

4.3.2. Analysis and Discussions

This study focuses on image enhancement and processing techniques to improve detection performance by processing input images. In contrast to the aforementioned contrast-based algorithms, the proposed method maintains detection accuracy while keeping the number of parameters to a minimum. Crucially, it preserves the underlying principles of the detection mechanism, thereby enhancing the algorithm's applicability. Consequently, this study uses YOLOv8n as the baseline model for optimisation design. Furthermore, Section 4.3.3 presents experiments employing different types of YOLO models as detection heads. In comparison with other algorithms, our approach delivers a significant improvement in detection performance while keeping the increase in parameters and FLOPs to a minimum, thus showing clear advantages.

As can be seen from the results in Table 2, for the SSDD dataset, the CRAS-YOLO, SHIP-YOLO and YOLOv12s algorithms achieved detection accuracies of 97.3%, 97.1% and 97.8%, respectively. These results are relatively close to those of the method proposed in this paper. Furthermore, the proposed algorithm achieved an mAP value of 98.8%, a metric more focused on network performance in the object detection domain. Similar results were achieved by CRAS-YOLO and YOLOv12s with values of 98.6% and 97.9%, respectively. These results demonstrate the robust performance of the proposed method on the SSDD dataset.

Compared to the SSSD dataset, the HRSID dataset exhibits generally lower detection accuracy than the SSDD dataset due to its significantly greater number of images and instances. Furthermore, the HRSID dataset's higher scene complexity also contributes to its comparatively lower detection accuracy.

Even under these circumstances, the proposed method still achieves superior results compared to other benchmark algorithms. With an accuracy rate of 93.2% and a MAP value of 92.8%, it demonstrates outstanding performance. Our results outperform those of the YOLOv12s method, which also achieves a MAP value of 92.8%, across all other metrics. Similarly, both the YOLOv12s and CRAS-YOLO algorithms achieve excellent accuracy results, with values reaching 93.1% and 93.0%, respectively.

The specific visualization results are presented in Figure 5. We conducted a qualitative comparison of five detectors across several representative scenarios. The first row presents visualizations of detection results in densely arranged backgrounds. The second row shows visualizations of detection results for targets of varying sizes. The third row displays visualizations of detection results under significant background interference. The results demonstrate that our detector outperforms others in these complex scenarios, achieving detection results that are closer to the ground truth. In contrast, the other detectors exhibit

certain issues. For instance, Faster-RCNN [12] and YOLOv8n [18] are prone to false alarms in near-shore scenarios, while RetinaNet [15] experiences more significant missed detections when identifying densely arranged ships. Additionally, for targets that other methods fail to detect accurately, our network effectively eliminates the influence of similar objects, ensuring correct detection.

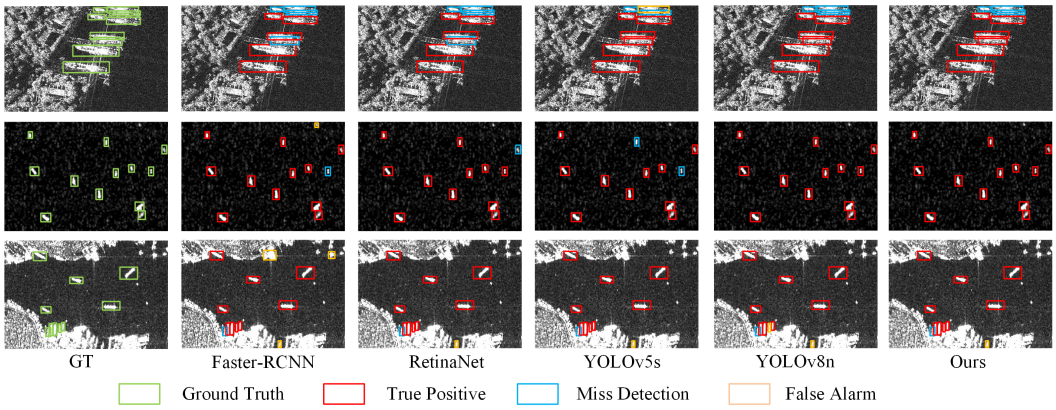


Figure 5. Comparison of visual detection results.

4.3.3. Ablation Analysis

As previously discussed, the incorporation of the FAEN, DHEM, and GIFM modules into the network led to varying degrees of improvement in the experimental results. As illustrated in Table 4, it is evident that, with the addition of these modules, the network parameters increased only slightly, while the average accuracy showed a modest improvement.

For the first set of experiments, we selected YOLOv8n as the baseline detection model. The input images were first processed using FAEN, then adjusted and fed into the detection model for analysis, yielding the first set of comparative results. This achieved a 0.2% improvement, which validates the effectiveness of FAEN. Next, the input images underwent DHEM feature extraction. After reshaping these features, they were fed into the detection network for the second set of control experiments. A noticeable decline in accuracy was observed, which was caused by erroneous detections resulting from excessive feature extraction. After the input images were subjected to FAEN enhancement, the original and enhanced images were fused using GIFM before being input into the detection model. This yielded a 0.5% improvement in accuracy, thus validating the effectiveness of GIFM. Next, the images underwent simultaneous FAEN and DHEM processing, followed by simple stitching and adjustment with the original images, before being input into the detection model. The results showed a further improvement on the original detection model, demonstrating the efficacy of combining both methods. Applying DHEM to the original image and then fusing the resulting features with those of the original image via GIFM before inputting the combined result into the detection model also yielded a certain degree of improvement. Finally, incorporating all modules achieved the highest detection accuracy to date, demonstrating the effectiveness of the proposed approach.

To ensure the statistical reliability of the evaluation results, five independent replicate experiments were conducted under strictly controlled identical conditions, including identical dataset partitions, learning rates and training iterations. In the results presentation, the mAP and corresponding Sample Variance (denoted by S^2) were employed as evaluation metrics. The variance was calculated from the mean average precision (mAP) values obtained across multiple runs. This serves as an indicator of the degree of fluctuation in the results, thereby demonstrating the model's stability. The results are presented in Table 4.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{30}$$

where n denotes the number of repeated experiments, $\bar{x}$ represents the average mAP across the experiments, and x_i represents the mAP results obtained from each experiment individually.

Table 4. The results of the ablation study of the proposed method are presented. We chose the yolov8n detection framework as the baseline model, in which the proposed module is gradually added and experimentally analyzed on the SSDD dataset.

FAEN	DHEM	GIFM	Params (M)	FLOPs (G)	*mAP* (%)	S^2
×	×	×	**3.0**	**8.1**	96.9	0.2237
✓	×	×	3.1	8.5	97.1	0.2335
×	✓	×	3.2	10.7	96.3	0.2613
✓	×	✓	3.2	9.0	97.4	0.2317
✓	✓	×	3.3	11.1	97.5	0.2449
×	✓	✓	3.3	11.2	97.9	0.2301
✓	✓	✓	3.4	11.6	**98.8**	0.2261

The optimal results of the corresponding indicators are expressed by thickening.

Although the number of parameters in the enhanced model increases from 3.0 M in the baseline model to 3.4 M, the corresponding mAP shows a 1.9% improvement. mAP serves as a more comprehensive and representative metric of network performance compared to precision (P) and recall (R). As we incorporate innovative modules into the network incrementally, this disturbs the overall convergence stability to some extent. However, comparative experiments reveal that, following the gradual addition of modules, the variance in experimental results shifts from increasing to decreasing, eventually converging with those of the original model. This shows that the proposed network model significantly improves detection accuracy without compromising the network's convergence properties. Experiments demonstrate that all three of the proposed modules have a positive impact on model performance. However, while they enhance accuracy, they also increase model size and complexity.

Based on the results of these ablation experiments, it is evident that our proposed method enhances mAP without significantly increasing the number of parameters. This suggests that the introduced modules contribute meaningfully to the overall network improvement.

Building upon the aforementioned comparisons, we employ a method of progressively incorporating the proposed modules in order to visualise the results, thereby enabling a more intuitive demonstration of detection performance. Successively integrating the FAEN, DHEM and GIFM modules into YOLOv8n for separate detection tasks illustrates the efficacy of the proposed approach. Figure 6 presents a comparison of the different networks in the ablation experiment. It is evident that incorporating the network processing module proposed in this paper enhances the performance of the network, resulting in improved predictions and higher scores. The method presented in this paper, when all modules are included, demonstrates superior performance in both prediction accuracy and overall score.

From a visual perspective, the validity of the proposed modules can be verified by generating corresponding heatmaps from the weights of different network models, as shown in Figure 7. Red areas in the heatmaps indicate greater focus on the target region, while blue areas indicate lesser focus. The first group shows heatmaps generated by the network after adding only the FAEN module. The second group shows heatmaps generated by the network after adding only the DHEM module. The third group displays heatmaps

generated by the complete model incorporating all three modules: FAEN, DHEM and GIFM. The FAEN module enhances detailed information to distinguish more clearly between targets and backgrounds, while the DHEM module thoroughly extracts information about the target object, thereby enriching representational capacity. Finally, the fusion of these two modules via GIFM achieves more precise localisation of information. Clearly, the collaborative integration of the proposed modules enhances the model's ability to capture discriminative information, thereby improving the neural network's representational capability for feature information.

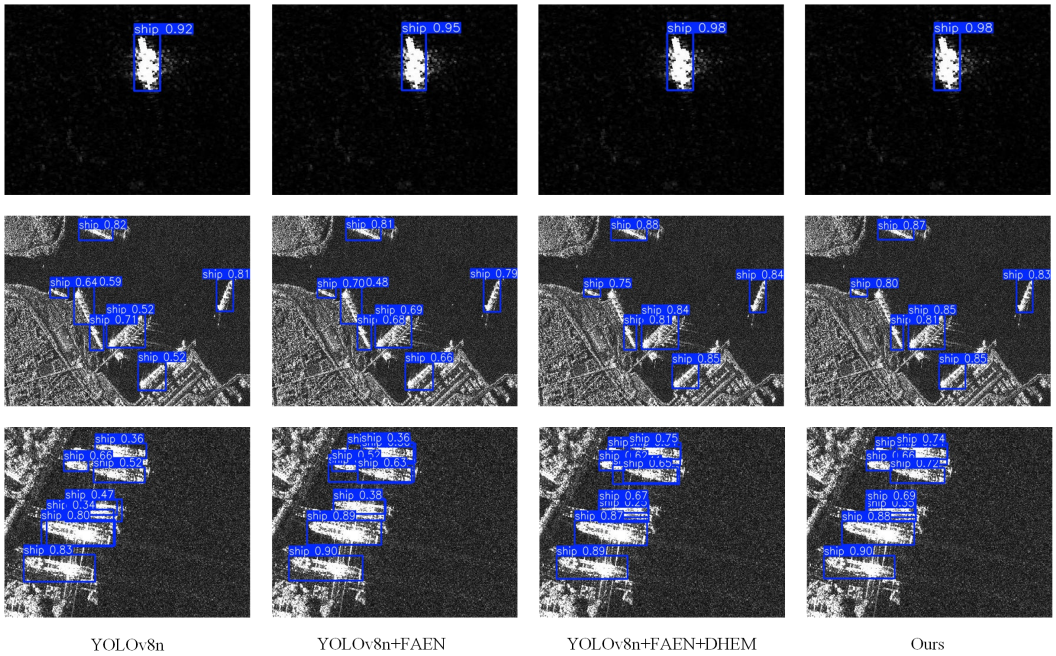


Figure 6. Comparison diagram of the ablation experiment.

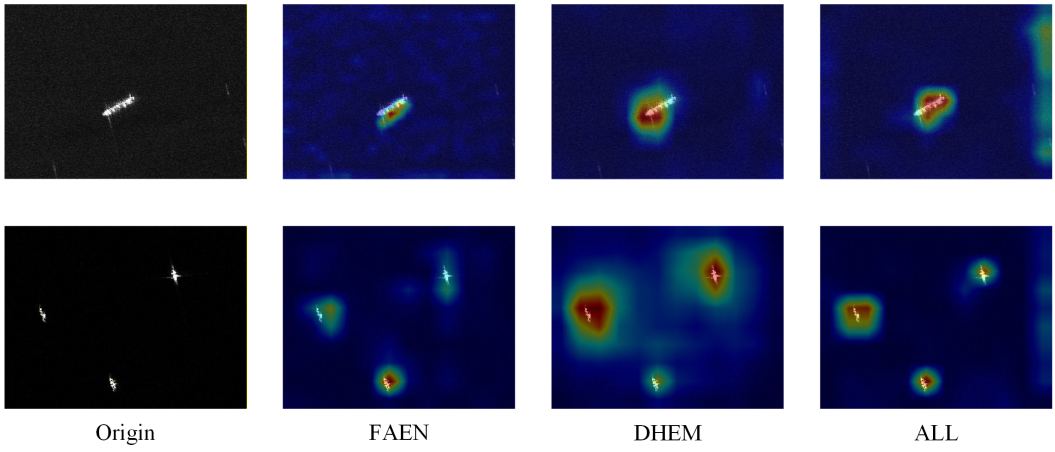


Figure 7. Visualization comparison of the ablation experiment.

Furthermore, to validate the effectiveness and generalisability of the proposed feedforward network, we designed a method to assess its efficacy while keeping the network itself unchanged and integrating it with various detection networks. This method was integrated with YOLOv5s, YOLOv7-tiny, YOLOv8n and YOLOv12s to perform detection tasks. The results were consistently positive, confirming the method's superior performance and generalisability. These findings are presented in Table 5.

As can be seen from the results in Table 5, combining different detection networks with the method proposed in this paper improves detection performance across all networks.

Furthermore, compared to the latest YOLOv12 version, our selected YOLOv8n demonstrates a significant computational advantage while maintaining comparable detection accuracy; the mAP value is merely 0.1 lower.

Table 5. The impact of identical networks on different detectors. We conducted an experimental analysis of the SSDD dataset using the same network with different detectors.

	mAP	Params (M)	FLOPs (G)
Ours + YOLOv5s	97.6	7.4	19.4
Ours + YOLOv7-tiny	97.3	6.4	**11.5**
Ours + YOLOv8n	98.8	**3.4**	11.6
Ours + YOLOv12s	**98.9**	9.7	24.9

The optimal results of the corresponding indicators are expressed by thickening.

5. Conclusions

In this paper, we propose a method that combines image enhancement and multiscale feature extraction for ship target detection in SAR images. The FAEN introduced in this work effectively preserves the original semantic information of the image while suppressing noise, thereby enhancing edge and detail representation. The DHEM proposed here maximizes the utilization of feature information within the image, overcoming issues related to information loss and insufficient feature extraction during the process. Additionally, the Gated Interaction Fusion Module enables a more refined fusion of multi-level information, ensuring comprehensive information integration and allowing the detection model to focus more precisely on the target of interest. Comparative experiments with mainstream methods and classical optimization algorithms, conducted on publicly available datasets such as SSDD and HRSID, demonstrate the superior performance of the proposed method. However, while the method proposed in this paper demonstrates strong performance in ship detection, its generalization capability requires further enhancement, as it has only been validated on two datasets. Additionally, the evaluation of its performance on ship features not present in the datasets, as well as in varied contextual environments, has yet to be addressed.

Future work will focus on further optimizing the model to achieve real-time performance and ensuring the detection efficiency of ships in large-scale SAR images. Although this method has demonstrated outstanding performance on two benchmark datasets, subsequent research will focus on three areas to further enhance its practical value: Firstly, to improve the model's generalization capability, plans include introducing SAR datasets from more diverse scenarios to explore its adaptability to unknown vessel characteristics and complex sea conditions, such as high seas and strong interference. At the same time, the research will focus on integrating different sources of information and using methods such as incremental learning to help models adapt to increasingly complex and dynamic real-world environments. Secondly, to improve computational efficiency, efforts will focus on model light weighting and hardware deployment optimization to enable real-time or near-real-time processing of wide-swath SAR imagery. Thirdly, with regard to system reliability, uncertainty quantification analysis will be integrated to improve the credibility of decisions. Furthermore, we will explore the development of an operational application platform that integrates detection, tracking and early warning capabilities, thereby advancing the technology from laboratory validation to practical engineering implementation.

Author Contributions: Conceptualization, N.X. and Y.B.; Data curation, W.L.; Formal analysis, Y.B.; Funding acquisition, N.X.; Investigation, Y.B.; Methodology, Y.B.; Project administration, Y.B.; Resources, W.L. and Y.B.; Supervision, N.X.; Validation, J.L.; Visualization, Y.B.; Writing—original draft, Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Liaoning Provincial Science and Technology Plan Joint Program (No. 2025-MSLH-049).

Data Availability Statement: The data supporting the results of this study can be obtained by contacting the corresponding author.

Conflicts of Interest: The authors declare that they have no known conflicts of interest.

References

1. Li, H.; Wang, D.; Hu, J.; Zhi, X.; Yang, D. FANT-Det: Flow-Aligned Nested Transformer for SAR Small Ship Detection. *Remote Sens.* **2025**, *17*, 3416. [CrossRef]
2. Xue, Y.; Zhan, L.; Liu, Z.; Bing, X. SAR Ship Target Instance Segmentation Based on SISS-YOLO. *Remote Sens.* **2025**, *17*, 3118. [CrossRef]
3. Wang, C.; Guo, B.; Song, J.; He, F.; Li, C. A Novel CFAR-Based Ship Detection Method Using Range-Compressed Data for Spaceborne SAR System. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5215515. [CrossRef]
4. Zeng, T.; Zhang, T.; Shao, Z.; Xu, X.; Zhang, W.; Shi, J.; Wei, S.; Zhang, X. CFAR-DP-FW: A CFAR-Guided Dual-Polarization Fusion Framework for Large-Scene SAR Ship Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 7242–7259. [CrossRef]
5. Li, M.D.; Cui, X.C.; Chen, S.W. Adaptive Superpixel-Level CFAR Detector for SAR Inshore Dense Ship Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4010405. [CrossRef]
6. Hao, Y.; Wu, J.; Yao, Y.; Guo, Y. A Robust Anchor-Free Detection Method for SAR Ship Targets with Lightweight CNN. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5028319. [CrossRef]
7. Lu, W.; Lan, C.; Niu, C.; Liu, W.; Lyu, L.; Shi, Q.; Wang, S. A CNN-Transformer Hybrid Model Based on CSWin Transformer for UAV Image Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1211–1231. [CrossRef]
8. Du, X.; Wu, X. Small Object Detection in Synthetic Aperture Radar with Modular Feature Encoding and Vectorized Box Regression. *Remote Sens.* **2025**, *17*, 3094. [CrossRef]
9. Wen, Z.; Zhang, W.; Qian, M. A comprehensive review of deep reinforcement learning for object detection. In Proceedings of the 2021 International Symposium on Artificial Intelligence and Its Application on Media (ISAIAM), Xi'an, China, 21–23 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 146–150.
10. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
11. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2094–2107. [CrossRef]
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
13. Li, M.; Chen, D.; Liu, S.; Liu, F. Prior Mask R-CNN Based on Graph Cuts Loss and Size Input for Precipitation Measurement. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5010015. [CrossRef]
14. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef]
15. Devi, S.K.; Baranidharan; Saranya, P. RetinaNet with Harmony Search Algorithm for Dynamic and Static Object Detection and Classification Model. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 16–18 March 2022; pp. 1030–1036.
16. 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 Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
17. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
18. Deng, H.; Zhang, Y. FMR-YOLO: Infrared Ship Rotating Target Detection Based on Synthetic Fog and Multiscale Weighted Feature Fusion. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5001717. [CrossRef]
19. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
20. Wang, X.; Wu, Y.; Shi, C.; Yuan, Y.; Zhang, X. ANED-Net: Adaptive noise estimation and despeckling network for SAR image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 4036–4051. [CrossRef]
21. He, K.; Li, P.; Wang, G.; Song, Z.; Li, L.; Wang, J. A novel SAR image denoising method based on self-supervised network. In Proceedings of the 2024 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Zhuhai, China, 22–24 November 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5.

22. Yan, S.; Chen, R.; Sang, H.; Zhou, Y.; Long, J.; Cai, N.; Xu, S.; Chen, J. Multihop Anchor-Free Network with Tolerance-Adjustable Measure for Infrared Tiny Target Detection. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5502011. [CrossRef]
23. Gao, F.; Cai, C.; Tang, W.; He, Y. A Compact and High-Efficiency Anchor-Free Network Based on Contour Key Points for SAR Ship Detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 4002705. [CrossRef]
24. Zhou, L.; Wan, Z.; Zhao, S.; Han, H.; Liu, Y. BFEA: A SAR Ship Detection Model Based on Attention Mechanism and Multiscale Feature Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11163–11177. [CrossRef]
25. Huang, P.; Tian, S.; Su, Y.; Tan, W.; Dong, Y.; Xu, W. IA-CIOU: An Improved IOU Bounding Box Loss Function for SAR Ship Target Detection Methods. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 10569–10582. [CrossRef]
26. Ai, J.; Tian, R.; Luo, Q.; Jin, J.; Tang, B. Multi-Scale Rotation-Invariant Haar-Like Feature Integrated CNN-Based Ship Detection Algorithm of Multiple-Target Environment in SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10070–10087. [CrossRef]
27. Mao, Q.; Li, Y.; Zhu, Y. A Hierarchical Feature Fusion and Attention Network for Automatic Ship Detection From SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 13981–13994. [CrossRef]
28. Xue, W.; Ai, J.; Zhu, Y.; Chen, J.; Zhuang, S. AIS-FCANet: Long-Term AIS Data Assisted Frequency-Spatial Contextual Awareness Network for Salient Ship Detection in SAR Imagery. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 15166–15171. [CrossRef]
29. Zhou, R.; Gu, M.; Hong, Z.; Pan, H.; Zhang, Y.; Han, Y.; Wang, J.; Yang, S. SIDE-YOLO: A Highly Adaptable Deep Learning Model for Ship Detection and Recognition in Multisource Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1501405. [CrossRef]
30. Liu, W.; Cao, H.; Xiong, G.; Liu, J.; Qi, C. GLRT-based detectors with enhanced selectivity for mismatched signals through a random-signal approach. *Signal Process.* **2026**, *240*, 110375. [CrossRef]
31. Liu, S.; Lei, Y.; Zhang, L.; Li, B.; Hu, W.; Zhang, Y.D. MRDDANet: A multiscale residual dense dual attention network for SAR image denoising. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5214213. [CrossRef]
32. Thakur, R.K.; Maji, S.K. Agsdnet: Attention and gradient-based sar denoising network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4506805. [CrossRef]
33. Tan, S.; Zhang, X.; Wang, H.; Yu, L.; Du, Y.; Yin, J.; Wu, B. A CNN-based self-supervised synthetic aperture radar image denoising approach. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5213615. [CrossRef]
34. Liu, S.; Tian, S.; Zhao, Y.; Hu, Q.; Li, B.; Zhang, Y.D. LG-DBNet: Local and global dual-branch network for SAR image denoising. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5205515. [CrossRef]
35. Pan, Y.; Zhong, L.; Chen, J.; Li, H.; Zhang, X.; Pan, B. SAR Image Despeckling Based on Denoising Diffusion Probabilistic Model and Swin Transformer. *Remote Sens.* **2024**, *16*, 3222. [CrossRef]
36. Yu, J.; Zhou, G.; Zhou, S.; Qin, M. A fast and lightweight detection network for multi-scale SAR ship detection under complex backgrounds. *Remote Sens.* **2021**, *14*, 31. [CrossRef]
37. Liu, S.; Chen, P.; Zhang, Y. A multiscale feature pyramid SAR ship detection network with robust background interference. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 9904–9915. [CrossRef]
38. Zhao, W.; Syafrudin, M.; Fitriyani, N.L. CRAS-YOLO: A novel multi-category vessel detection and classification model based on YOLOv5s algorithm. *IEEE Access* **2023**, *11*, 11463–11478. [CrossRef]
39. Luo, Y.; Li, M.; Wen, G.; Tan, Y.; Shi, C. SHIP-YOLO: A Lightweight Synthetic Aperture Radar Ship Detection Model Based on YOLOv8n Algorithm. *IEEE Access* **2024**, *12*, 37030–37041. [CrossRef]
40. Feng, Y.; Zhang, Y.; Zhang, X.; Wang, Y.; Mei, S. Large Convolution Kernel Network with Edge Self-Attention for Oriented SAR Ship Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 2867–2879. [CrossRef]
41. James, S.; Wada, K.; Laidlow, T.; Davison, A.J. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13739–13748.
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
45. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.