Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SFE-YOLO: Spatial-Frequency Dual-Domain Enhanced YOLO for SAR Ship Detection

Huicong Ning
Anyang Institute of Technology
Ying Zhang
caroline_zy@nankai.edu.cn
Nankai University
Jiajia Chen
Anyang Embedded Intelligent Robot Co., Ltd., Anyang, Henan 455000, China
Peinan Hao
Anyang Institute of Technology
Xiaojing Zhao
Anyang Institute of Technology
Tianzi Yan
Beijing Xiao Mi Mobile Software Co., Ltd.

Research Article

Keywords: SAR ship detection, SPConv, EBlock, YOLOv8, Frequency-domain enhancement

Posted Date: August 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10573114/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SFE-YOLO: Spatial-Frequency Dual-Domain Enhanced YOLO for SAR Ship Detection

Huicong Ning[1], Ying Zhang[2,3*], Jiajia Chen[6], Peinan Hao[1], Xiaojing Zhao[4], Tianzi Yan[5]

[1]School of Computer Science and Information Engineering, Anyang Institute of Technology, Anyang, 455000, China .
[2]College of Electronic Information and Optical Engineering, Nankai University, Tianjin, 300350, China.
[3]Tianjin Key Laboratory of Optoelectronic Sensor and Sensing Network Technology, Tianjin, 300350, China.
[4]College of Mathematics and Information Science, Anyang Institute of Technology, Anyang, 455000, China.
[5]Beijing Xiao Mi Mobile Software Co., Ltd., Beijing, 100080, China.
[6]Anyang Embedded Intelligent Robot Co., Ltd., Anyang, 455000, China.

*Corresponding author(s). E-mail(s): caroline_zy@nankai.edu.cn;

Abstract

Although deep learning techniques have advanced SAR ship detection, existing detectors still suffer from two inherent limitations in complex marine environments: standard isotropic convolutions fail to capture multi-orientation slender hull features, and purely spatial modeling lacks global frequency perception to suppress heavy sea clutter, both leading to feature degradation and frequent small-vessel missed detection. To address these issues, we propose SFE-YOLO, a dual-domain enhanced YOLO framework tailored for SAR ship detection. Specifically, we integrate pinwheel convolution (SPConv) into the backbone to expand multi-directional receptive fields, which effectively extracts elongated ship contours of arbitrary orientation. Furthermore, we embed a frequency-domain enhancement block (EBlock) built upon FreUnitMLP into the feature fusion neck; it suppresses sea clutter and amplifies weak ship edges via spatial residual refinement and FFT-IFFT dual-domain calibration. Extensive evaluations on SSDD and HRSID datasets demonstrate that our method achieves mAP50 of 0.973 and 0.885, and mAP75 of 0.776 and 0.716, respectively, outperforming

ten mainstream detectors in clutter resistance, fine localization accuracy, and small-target generalization.

Keywords: SAR ship detection, SPConv, EBlock, YOLOv8, Frequency-domain enhancement

1 Introduction

The global maritime shipping and offshore port operations continue to expand, driving significantly increased demands for maritime safety management and resource/environmental regulation. Ship dynamic monitoring plays a vital role in supporting maritime protection, law enforcement, and emergency rescue operations. Traditional ship detection primarily relies on visible-light imaging and radar sensing: while visible images offer high resolution and detailed texture information, they are highly sensitive to lighting conditions and weather variations, leading to markedly reduced performance under nighttime or haze conditions; radar, although capable of all-weather detection, faces challenges due to significant differences between echo signatures and visual semantics, coupled with substantial noise interference in complex sea environments and maritime clutter [1–3]. Therefore, developing efficient and accurate ship target detection algorithms is crucial.

In recent years, with the rapid advancement of computer vision and artificial intelligence technologies, deep learning-based image processing methods have demonstrated significant potential in ship detection applications. Current deep learning-based object detection approaches primarily fall into two categories: the first comprises two-stage detection algorithms represented by the R-CNN series [4], which first generate candidate boxes containing potential targets through region-based candidate networks, followed by classification and position refinement of these boxes—these methods achieve high detection accuracy but exhibit relatively slow inference speeds. For instance, Subroto et al. [5] developed a coastal obstacle detection model using Faster R-CNN, achieving high accuracy across various ship recognition tasks; Fan et al. [6] addressed challenges posed by low SAR image resolution and strong background interference by proposing a hierarchical multi-scale Mask R-CNN that enhances multi-scale ship detection precision through cross-layer feature fusion. However, such architectures incur substantial computational costs, resulting in high model complexity and slow inference speeds, making them inadequate for the stringent real-time performance and edge deployment requirements of ship detection and tracking tasks. Another category comprises single-stage detection algorithms represented by YOLO and SSD [7], which eliminate the need for candidate box generation and enable direct end-to-end object classification and localization. These algorithms offer rapid detection speeds and are particularly well-suited for real-time detection requirements in maritime operations, making them a key focus of current research. Researchers have conducted extensive work on optimizing the lightweight nature and multi-scale adaptability of these single-stage algorithms, achieving significant breakthroughs: Zhang et al. [8] proposed the

lightweight ship detection architecture YOLO-Ships, which uses YOLOv5s as its baseline and incorporates modules such as Ghost, REGhost, and C3REGhost to enhance feature extraction. Combined with an improved feature pyramid and path aggregation network (FPN+PAN) along with enhanced Ghost coordinate attention (EGCA), this architecture effectively optimizes multi-scale feature fusion and improves detection performance for small-scale ships under complex sea conditions; Huang et al. [9] improved YOLOv8 by replacing the C2f module at the neck layer with an expansion residual module, enhancing the algorithm's ability to recognize multi-scale objects and represent diverse features; Muhammad et al. [10] replaced the network architecture of YOLOv8 with HGNetv2 and Slim-neck, significantly reducing computational costs while maintaining detection accuracy. Guan et al. [11] proposed a mixed-shuffling reparameterization module that combines reparameterized convolution with channel mixing operations, enhancing the model's ability to extract features of small maritime targets while reducing interference from ambient sea clutter and speckle noise. Zhu et al. [12] developed an incremental significance-guided approach utilizing position priors derived from signal saliency to progressively improve feature extraction of small-to-medium ships in large-scale SAR images, effectively addressing detection weaknesses such as low response rates and missed detections while improving feature representation. Xiao Zhenjiu et al. [13] designed a synthetic aperture radar ship detection framework integrating multiple mechanisms, incorporating a U-Net denoising module into YOLOv7 and refining the detection head to suppress image noise while enhancing multi-scale feature representation. Gao et al. [14] introduced an angle classification branch and a probability-distribution loss function into YOLOv5, addressing detection challenges caused by small and densely distributed targets in remote sensing images. Zhang et al. [15] implemented a frequency attention module based on YOLOv5 to adaptively exploit frequency-domain information, mitigating ocean wave interference and improving robustness and accuracy. Wang Yingjun et al. [16], optimizing the YOLO11 model, developed the AFEM module to enhance degradation feature extraction and added an MFP multi-scale perception module to reduce background interference and minimize detection errors.

In summary, while existing deep learning approaches have achieved competitive performance in SAR ship detection, but they still exhibit limited adaptability to complex marine scenarios. Standard isotropic convolutions struggle to capture multi-orientation slender hull features, easily causing feature degradation and missed detection of small ships. Moreover, original YOLO models focus merely on local spatial modeling and lack global contextual perception, making weak ship edge features susceptible to sea clutter noise. To better adapt the YOLO framework to practical SAR ship detection tasks, this paper optimizes the baseline network by embedding mature existing feature enhancement modules for scenario-specific performance improvement. The main contributions of this work are summarized as follows:

- We propose an orientation-adaptive feature extraction strategy by integrating SPConv pinwheel convolution into the YOLOv8 CSPLayer. This design addresses the fundamental mismatch between standard square convolution kernels and the slender, arbitrarily oriented shape of SAR ship targets, expanding multi-directional

receptive fields to capture fine-grained hull contours that are otherwise lost in conventional convolutions.

- We design a frequency-domain clutter suppression module (EBlock) with FreUnitMLP at its core, embedded in the feature fusion neck. By combining spatial residual optimization with FFT-IFFT global calibration, the module adaptively suppresses sea speckle noise in the frequency domain while strengthening faint edge features of tiny ships — a capability inherently absent in purely spatial CNN architectures.
- We establish a dual-domain synergistic framework that unites orientation-adaptive spatial feature extraction and global frequency-domain clutter suppression, where the two modules complement each other: SPConv enhances local structural details of ships, and EBlock cleans global background noise. Comprehensive experiments on two public SAR benchmarks (SSDD and HRSID) validate the effectiveness of this synergy, with consistent improvements across mAP75, small-target detection, and clutter resistance compared to ten mainstream detection paradigms.

2 Related Work

2.1 Feature transformation between spatial and frequency domains

The feature transformation between the spatial domain and the frequency domain serves as the fundamental theoretical basis for multi-dimensional image representation and cross-domain optimization. Its essence lies in achieving a bidirectional, reversible mapping of image signals between the pixel space and the frequency space via the Fourier transform framework, thereby overcoming the limitations of single-dimensional feature representation and enabling complementary modeling of local spatial structures and global frequency distributions.

Images exist in spatial domain form both during raw acquisition and in the representation of shallow network features. The spatial domain uses pixel coordinates (x, y) as its fundamental representation unit, describing target locations, contours, and local structural relationships through neighborhood grayscale values, gradients, and texture variations, thereby providing precise spatial topological information. Pixel-level mathematical operations are directly applied to either the raw image or the network-generated feature map to extract features; typical operations include pixel grayscale mapping, gradient computation, and neighborhood convolution. The mathematical expression for convolutional feature extraction is as follows:

$$G(x,y) = \sum_{m} \sum_{n} K(m,n) \cdot f(x-m, y-n) \tag{1}$$

In this formula, f(x, y) represents the feature in the input spatial domain, and K(m, n) denotes the convolution kernel. This approach relies on correlations between neighboring pixels to capture local features such as contours and textures, enabling precise spatial representation. However, spatial operations are confined to a limited receptive field, making it unable to distinguish large-area backgrounds, essential target details,

and global noise from a global perspective. In scenarios with severe degradation—such as maritime fog or low-light blurring—the effectiveness of feature decoupling and interference suppression becomes significantly limited.

To address the limitations of global modeling in the spatial domain, the spatial features are transformed to the frequency domain using two-dimensional discrete Fourier transform (DFT), enabling global signal decomposition. The transformation formula is as follows:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-2\pi j\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{2}$$

After performing operations such as feature enhancement and filtering in the frequency domain, the results must be reconstructed into the spatial domain using the inverse discrete Fourier transform (IDFT), following a closed-loop transformation process. The reconstruction formula is as follows:

$$\hat{f}(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \hat{F}(u,v) e^{-2\pi j\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{3}$$

In this formula, $\hat{F}(u,v)$ represents the frequency-domain feature after adaptive adjustment, and $\hat{f}(x,y)$ denotes the reconstructed spatial feature map that accurately restores the pixel spatial structure, thereby overcoming the limitation of pure frequency-domain processing in losing target location information.

The frequency domain eliminates pixel-level spatial constraints and provides a global representation of images based on signal fluctuation characteristics. Using the two-dimensional discrete Fourier transform (DFT), it decomposes the spatial-domain signal into a superposition of multiple sets of sine and cosine basis functions with distinct frequencies, amplitudes, and phases, thereby achieving global feature extraction. This decomposition effectively decouples background information, target details, and degraded noise that were coupled in the spatial domain, enabling targeted feature enhancement and interference suppression.

2.2 Convolutions and their improvements

Regular convolution serves as the fundamental feature extraction operation in the spatial domain, directly modeling local features at image pixels or network feature maps, and constitutes a core component of mainstream CNNs. Let the input spatial feature map be denoted as $f(x,y)$ with dimensions $M \times N$; the convolution kernel of size $k \times k$ be represented by its weights $w(m,n)$; and the stride be 1 with no padding. The mathematical expression for the convolutional output feature $g(x,y)$ is as follows:

$$g(x,y) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} w(m,n) \cdot f(x+m, y+n) + b \tag{4}$$

Here, b denotes the bias term. The computational logic involves the convolution kernel sliding across the entire feature map to perform weighted summation with neighboring pixels in each window, thereby extracting local spatial features such as edges,

textures, and corner points. Conventional convolution relies solely on a limited number of neighboring pixels to establish dependencies, excelling at capturing target local contours and spatial position information; it features simple computation and easy deployment, making it widely used in backbone networks for object detection. However, it has inherent limitations: the receptive field is derived solely from stacking convolution kernel sizes, resulting in extremely low efficiency in modeling long-range global dependencies; all channels share the same set of convolution weights, preventing adaptive adjustment of feature extraction intensity for images with varying scales or degradation levels.

To address the limitations of conventional convolutional networks, Chen et al. proposed Frequency Domain Dynamic Convolution (FDConv)[17], which consists of three core components: Fourier Decoupling Weighting (FDW), Kernel Space Modulation (KSM), and Frequency Band Modulation (FBM). The FDW serves as the foundational element of FDConv, abandoning the traditional approach of learning convolution kernels in the spatial domain by directly training learnable spectral coefficients in the frequency domain. Fixed parameter sets are divided into non-overlapping frequency-domain index groups, each responsible for specific frequency bands; these spectral parameters are then mapped back to the spatial domain via inverse discrete Fourier transform, followed by cropping and recombination to generate independent convolution kernels. Each parameter group corresponds to distinct spectral regions, resulting in kernels with differentiated frequency-domain responses. This approach enables the generation of over ten complementary kernels without increasing parameter complexity, effectively addressing issues of spectral homogeneity and parameter redundancy in traditional dynamic convolutions. To overcome the coarse-grained nature of simple weight fusion, the KSM module introduces dual branches—local and global—which uses a lightweight one-dimensional convolution to extract local channel information while a fully connected layer captures global context, jointly predicting a dense modulation matrix to fine-tune each weight element. This achieves fine-grained adaptive adjustment of kernel frequency-domain responses, enhancing the model's adaptability to diverse input scenarios. The FBM further transcends the spatial invariance limitation inherent in conventional dynamic convolutions. This module decomposes the convolution kernel into independent low-frequency, intermediate-frequency, and high-frequency subbands via Fourier transform, performing feature convolution in the frequency domain based on the convolution theorem; it generates customized frequency-band modulation maps for different spatial regions of the image and applies spatially variable weighting to features across each frequency band. Although FDConv offers advantages in frequency diversity and parameter efficiency, it still suffers from drawbacks such as parameter redundancy, limited adaptability, and complex hyperparameter tuning.

Liu et al. proposed the Gated Bottleneck Convolution (GBC)[18], introducing a bottleneck convolution layer that employs a low-rank approximation method to map high-dimensional matrices into lower-dimensional spaces, thereby reducing computational complexity. This approach enables the model to efficiently extract features from crack images while minimizing resource consumption. Additionally, the gating mechanism assigns dynamic characteristics to each spatial position and channel, allowing the

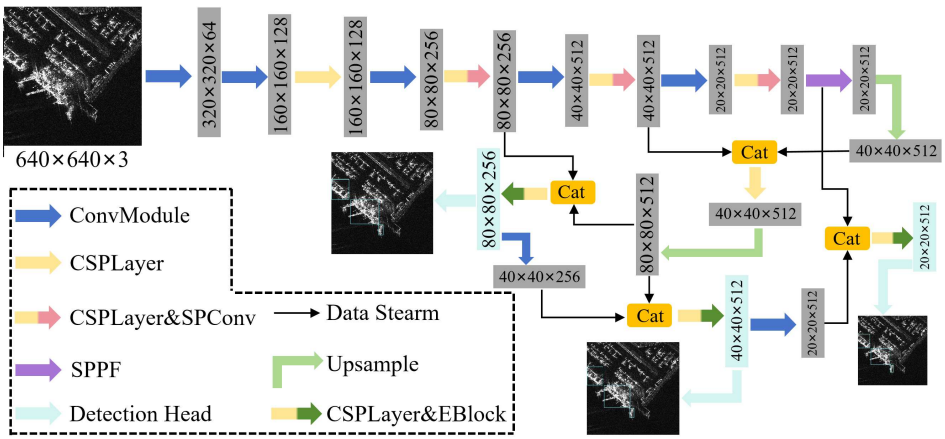


Fig. 1 The network architecture of SFE-YOLO.

model to adaptively adjust its behavior based on different features, enhancing detail capture and segmentation accuracy. GBC primarily consists of three key components: a bottleneck convolution layer, a gating mechanism, and residual connections. The input features first retain a portion for subsequent residual connections; then proceed through the bottleneck convolution layer, which reduces parameter count and computational load via specialized convolutional techniques. After processing, the features undergo normalization and activation functions to produce two sets of features. These features are combined via Hadamard product to form a gated feature map, which is further refined through additional bottleneck convolutions. Finally, the refined features are combined with the initially retained features via residual connections to obtain the final output. This architecture ensures effective feature extraction while preventing information loss, enhances model robustness, and achieves more precise segmentation results.

Yang proposed the Pinwheel-shaped Convolution (PConv)[19], which significantly expands the receptive field through asymmetric multi-directional parallel convolutions while requiring only a minimal increase in parameter size. This module features a two-stage convolutional architecture: The first stage comprises four parallel branches of unidirectional convolutions, each employing 1×3 or 3×1strip kernels with asymmetric four-edge padding, creating four distinct sampling paths—left-right widening padding for horizontal features and top-bottom widening padding for vertical features—that extend outward in both horizontal and vertical directions to simulate the Gaussian pixel distribution pattern with attenuation from the center. The outputs from each branch are normalized by BN and activated by SiLU before being concatenated to integrate multi-directional local details; the second stage employs 2×2 convolutional integration to merge features, unify channels and dimensions, enable multi-directional feature interaction, inherently incorporate channel attention mechanisms, and adaptively balance contributions from different feature directions.

3 Method

3.1 Overall architecture of the SFE-YOLO

The architecture of the proposed detection network SE-YOLO is illustrated in Figure. 1. The model adheres to the standard three-stage pipeline of YOLOv8, with the utilisation of normalized SAR images resized to 640×640 as inputs to the network. In the context of feature extraction, the architecture incorporates ConvModules and a modified CSPLayer integrated with SPConv to facilitate progressive downsampling and hierarchical feature aggregation. This integration serves to expand the effective receptive field, thereby enhancing the model's sensitivity to fine-grained spatial details, including slender ship contours and differently oriented small targets. At the final stage of the process, the SPPF module is attached to the backbone terminus. The purpose of this attachment is to aggregate long-range global contextual information.

For the neck network, we follow the YOLO bidirectional multi-scale fusion paradigm, using bottom-up upsampling to convey deep semantic information and top-down downsampling to convey shallow-level details, while leveraging channel concatenation to achieve complementary fusion of these two types of features. In addition, our method incorporates improved units, including CSPLayer and EBlock, after each feature concatenation, with EBlock featuring a built-in FreUnitMLP frequency-domain enhancement sub-module. This sub-module uses a two-dimensional fast Fourier transform to transform features into the frequency domain and performs global adaptive weighting, which specifically enhances the high-frequency textures of small ships while suppressing low-frequency noise from the sea surface.

During the prediction, YOLO outputs three fused feature maps with resolutions of 80×80, 40×40, and 20×20, corresponding to small, medium, and large ships, respectively. Each feature map is fed into a separate detection head to perform object classification and bounding box regression tasks.

Overall, our method relies on the collaborative optimization of the lightweight spatial module SPConv and the frequency-domain module EBlock to achieve a bidirectional complementarity between the extraction of local spatial details and the enhancement of global frequency-domain features, thereby improving the detection accuracy of small ships on complex SAR sea surfaces without compromising the model's real-time inference performance.

3.2 Components

3.2.1 YOLOv8

YOLOv8 is a representative single-stage object detection model released by Ultralytics in 2023, inheriting the efficient inference characteristics of previous YOLO algorithms. Its backbone network, neck feature fusion module, and decoupled detection heads have been specifically optimized to reduce computational costs and better balance detection accuracy and speed. Due to its reliable performance, YOLOv8 has been widely adopted in various real-time visual detection applications and has achieved strong results on mainstream public datasets. Although many newly proposed state-of-the-art detection

models have emerged since YOLOv8, extensive comparative experiments have demonstrated that, when specifically applied to SAR ship detection tasks, YOLOv8 still delivers better detection performance than its subsequent advanced counterparts. To address detection requirements under varying hardware conditions, YOLOv8 offers five models with different parameter scales: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Among all versions, YOLOv8n features the lightest network architecture and the fewest parameters. It requires fewer computational resources and less storage space, making it more suitable for devices with limited computational performance. Furthermore, YOLOv8n maintains high inference speed without significant accuracy loss, effectively meeting the requirements for real-time detection. Therefore, we selected YOLOv8n as the baseline model and developed SE-YOLO for SAR ship detection.

3.2.2 SPConv

Traditional symmetric convolution kernels are limited by their isotropic receptive fields and have inherent shortcomings in capturing elongated textures oriented in multiple directions; they struggle to generate balanced feature responses to horizontal, vertical, and oblique edges. However, naval targets typically exhibit a long, narrow, weakly scattering shape. Conventional square convolution cannot fully extract fine-grained contour information from the hull in any orientation, which can easily lead to the weakening of target features and, consequently, the failure to detect small ships. To address this problem, we introduced the SPConv [20] into the feature extraction module. This convolution decomposes standard 2D convolution into parallel horizontal $1 \times k$ and vertical $k \times 1$ strip convolution operations, which extract texture features from different directions, respectively. At the same time, we introduced a feature cyclic shift operation to correct spatial deviations in the receptive field, thereby achieving multi-directional feature aggregation. This approach expands the network's effective receptive field while maintaining a lightweight architecture, thereby enhancing the model's ability to recognize features of slender ship targets oriented in different directions.

Given an input feature map $X \in R^{B \times C \times H \times W}$, where B denotes the batch size, C refers to the number of feature channels, H and W represent the height and width of feature maps respectively, the complete computation pipeline of SPConv can be mathematically formulated as follows. First, ZeroPad2d is applied to the input feature to preserve consistent spatial resolution after strip convolution. The padded feature is then split into four parallel branches for 1D convolution operations: the first two branches implement horizontal $Conv_{1\times k}$, where parameter k denotes the extended length of 1D convolution kernels, and two shift configurations (k,0,1,0) and (0,k,0,1) are set to capture horizontally oriented textures with different inclinations. The remaining two branches adopt vertical $Conv_{k\times 1}$ with shift configurations (0,1,k,0) and (1,0,0,k) to extract multi-angle vertical contours. The convolution process of four branches is defined as:

$$\begin{cases} X_1, X_2 = Conv_{1\times k}(ZeroPad(X)) \\ X_3, X_4 = Conv_{k\times 1}(ZeroPad(X)) \end{cases} \tag{5}$$

where X_1, X_2 stand for feature outputs of two horizontal 1D convolution branches, andX_3, X_4denote feature outputs of two vertical 1D convolution branches. Subsequently, the outputs of horizontal and vertical convolution branches are concatenated along the channel dimension and processed by cyclic shifting operation $roll(\bullet)$ to align the center of receptive fields. Specifically, fused horizontal features are shifted along the height dimension via $roll(\bullet, dims = 2)$, where dims=2 corresponds to the height axis of feature maps; fused vertical features are shifted along the width dimension via $roll(\bullet, dims = 3)$, where dims=3 represents the width axis of feature maps, which is formulated as:

$$\begin{cases} F_{hori} = roll(Cat(X_1, X_2), dim = 2) \\ F_{vert} = roll(Cat(X_3, X_4), dim = 3) \end{cases} \tag{6}$$

where F_{hori} and F_{vert} are aligned fused features for horizontal and vertical directions after cyclic shifting. Finally, the calibrated directional feature maps are concatenated along the channel axis to generate the module output $F_{out} = Cat(F_{hori}, F_{vert})$. Such a decomposition strategy replaces integral $k \times k$ 2D convolution with separated 1D strip convolutions, which drastically reduces floating-point computational cost with negligible parameter growth. When embedded into the backbone network, SPConv strengthens the extraction of multi-scale fine-grained features without sacrificing real-time inference performance.

3.2.3 EBlock

Despite the widespread application of convolutional neural networks in SAR detection, standard convolutions suffer from inherent drawbacks. Limited local receptive fields disable standard convolutions to capture long-range global contextual information. In practical marine SAR scenes, small ships present weak high-frequency edge features, which are easily submerged by homogeneous low-frequency sea clutter. Accordingly, conventional convolution methods fail to distinguish small ship targets from complex ocean background interference effectively. To solve the above problems and improve the detection performance of small SAR ships, we embeds a frequency-domain enhancement module (EBlock) into the network backbone, with FreUnitMLP as its core unit for feature enhancement. Specifically, spatial features are first optimized via a three-branch residual structure, spatial global-aware (SG) module and spatial-channel attention (SCA) mechanism. The refined spatial features are then transformed into frequency domain via fast Fourier transform (FFT), where lightweight 1×1 convolutions adaptively re-weight different frequency components. After frequency modulation, inverse fast Fourier transform (IFFT) converts features back to spatial domain. Learnable scaling parameters are further adopted to calibrate spatial and frequency features, realizing efficient dual-domain feature fusion.

Given the input feature $X \in R^{B \times C \times H \times W}$,where B denotes the batch size, C represents the total number of feature channels, and $H$and W refer to the height and width of input feature maps, respectively. The computation pipeline of EBlock is divided into spatial preprocessing and frequency-domain adjustment within FreUnitMLP, which is fully described by mathematical formulas. The input $X$is first calibrated by BatchNorm, then processed by stacked 3×3 convolution and 1×1 convolution to generate base feature X_{base}. The base feature is split into three residual branches (Branch1,

Branch2, Branch3) and fused via element-wise summation:

$$X_{base} = Conv_{1\times1}(Conv_{3\times3}(BatchNorm(X))) \tag{7}$$

$$X_{res} = Branch_1(X_{base}) + Branch_2(X_{base}) + Branch_3(X_{base}) \tag{8}$$

The fused residual feature X_{res} passes through SG module and SCA attention layer to eliminate redundant channels, generating preprocessed spatial featureX_{sp}. A learnable parameter Param is applied to fine-tune feature amplitude, producing the input feature X_{in} for frequency submodule:

$$X_{sp} = Conv_{1\times1}(SCA(SG(X_{res}))) \tag{9}$$

$$X_{in} = X_{sp} \oplus Param \tag{10}$$

Subsequently, the feature X_{in} is fed into FreUnitMLP for adaptive frequency optimization. The real-valued two-dimensional fast Fourier transform(rfft2) decomposes the input data into the real part$Real(\bullet)$ and the imaginary part $Imag(\bullet)$. After cascading the two components along the channel dimension, the frequency weights are adjusted using 1×1 convolution, BatchNorm, and the ReLU activation function. The frequency-domain data is then converted back to the spatial domain via the real-valued two-dimensional inverse Fourier transform (irfft2). The complete frequency transformation process is shown below:

$$X_{freq} = rfft2(X_{in}) \tag{11}$$

$$X_{cat} = Cat(Real(X_{freq}), Imag(X_{freq})) \tag{12}$$

$$X_{ftune} = ReLU(BatchNorm(Conv_{1\times1}(X_{cat}))) \tag{13}$$

$$F_{out} = irfft2(X_{ftune}) \tag{14}$$

where $rfft2$ and $irfft2$ are forward and inverse real 2D Fourier transform operators, Cat denotes channel-wise concatenation, and F_{out} is the final output feature of EBlock. Generally, embedded in the multi-scale feature fusion, the designed module mitigates the critical local receptive field limitation of standard convolution operations, which is a long-standing bottleneck in SAR object detection tasks. Different from heavy feature enhancement structures, the frequency branch is constructed based on compact convolutional components to cut down redundant computational costs. With this pragmatic structural optimization, the network can effectively excavate global contextual information from SAR images to enrich feature representations, and maintain high-efficiency inference throughout detection without obvious latency increase.

4 Experiment

4.1 Experimental Environment

In this work, we conduct all experiments on an Ubuntu 22.04 platform. We implement the model based on PyTorch 2.1.0 and Python 3.10, and adopt CUDA 12.1 to accelerate computational operations. The hardware configuration consists of a 14-vCPU Intel(R) Xeon(R) Gold 6348 CPU running at 2.60 GHz, 100 GB RAM, and an NVIDIA A800 GPU.

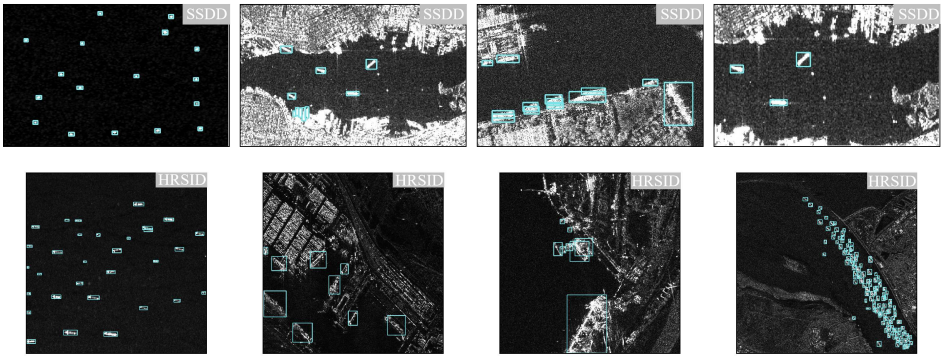


Fig. 2 (a) offshore multi-target (b) nearshore multi-target (c) nearshore multi-scale (d) nearshore small target

4.2 Datasets

SSDD: The SSDD dataset is a prevalent benchmark for SAR ship detection, which comprises 1160 images and 2456 annotated ship targets. It covers varied imaging resolutions (1–15 m), multiple polarization patterns, and typical marine scenarios ranging from open sea to complex inshore harbors. Several inherent difficulties hinder accurate detection on this dataset. First, ships exhibit drastic intra-class scale differences, posing challenges for multi-scale feature modeling. Second, coastal infrastructures and reefs produce high backscattering clutter similar to ship signatures, which easily induces false positives in near-shore areas. Third, the relatively limited sample size restricts the generalization capability of detection models.

HRSID: The HRSID dataset is a high-resolution and challenging benchmark for SAR ship detection and instance segmentation, which comprises 5604 images and 16951 annotated ships. It covers fine imaging resolutions (0.5–3 m), diverse polarization modes, and complex marine scenarios including open seas and busy inshore ports across global shipping lanes. Several inherent difficulties hinder accurate detection on this dataset. First, ships present extreme scale imbalance with dominant tiny targets, which complicates robust multi-scale feature modeling. Second, high-resolution imaging amplifies intricate sea clutter and background noise, interfering with target feature extraction. Third, densely clustered ship targets in harbor areas suffer from severe overlap and occlusion, greatly increasing the risk of missed detection and false alarms.

4.3 Performance Metrics

To evaluate the effectiveness of the improved SE-YOLO method for SAR ship detection, we employ four key performance evaluation metrics commonly adopted in object detection: precision (P), recall (R), average precision (AP), and mean average precision at the 0.5 IoU threshold (mAP50). These metrics are calculated based on three basic evaluation definitions, namely true positive (TP), false positive (FP), and false negative (FN). TP refers to the correctly localized ships that satisfy the preset IoU threshold, FP represents background mistakenly identified as ships, while

Table 1 Ablation results for each components, where the optimal value is indicated in **bold**.

NO.	Model	mAP50	mAP75	mAP50:95	mAPS	mAPM	mAPL
1	YOLOv8	0.959	0.725	0.632	0.588	0.722	0.584
2	SPConv	0.962	0.726	0.640	0.590	0.719	0.685
3	EBlock	0.970	**0.778**	0.656	0.601	0.742	0.692
4	**Ours**	**0.973**	0.776	**0.659**	**0.603**	**0.747**	**0.753**

FN corresponds to real ships that are not successfully detected by the network. The mathematical expressions of P and R are presented as follows:

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

$$AP = \int_0^1 P(R)dR \tag{17}$$

AP is defined as the integral area enclosed by the precision-recall curve, which reflects the comprehensive detection performance of the model under different confidence thresholds. In addition, mAP50 is introduced as the overall evaluation index, which averages the AP values of all target categories at a fixed IoU threshold of 0.5. Because we focus on single-class SAR ship detection tasks, mAP50 can effectively evaluate the overall detection accuracy and generalization performance of the proposed method. The mathematical formulas for AP and mAP50 are given as follows:

$$mAP50 = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{18}$$

where N indicates the total number of target categories, and AP_i represents the average precision of the i-th detection category.

4.4 Experiment Comparison and Analysis

4.4.1 The impact of each module

Table.1 summarizes the ablation results of different modules adopted in this study. Compared with the vanilla YOLOv8 model, both SPConv and EBlock yield consistent performance improvements, demonstrating the effectiveness of individual optimization for SAR ship detection. SPConv produces only minor accuracy gains, indicating that orientation-adaptive convolution contributes limited performance improvement when used alone, as it merely enhances feature representation for arbitrarily oriented ship targets. In contrast, EBlock achieves a more substantial performance promotion across mainstream metrics. The apparent improvements in mAP50, mAP75 and mAP50:95 indicate that frequency-domain enhancement is more capable of suppressing complex

Table 2 Ablation results for SPConv, where the optimal value is indicated in **bold**.

NO.	Model	mAP50	mAP75	mAP50:95	mAPs	mAPm	mAPl
1	FDConv	0.968	0.733	0.650	0.593	0.739	0.734
2	GBConv	0.960	0.735	0.649	0.601	0.725	0.716
3	PSConv	0.968	0.767	0.657	0.593	**0.748**	**0.797**
4	GSConv	0.971	0.755	0.656	0.599	0.740	0.746
5	Ours	**0.973**	**0.776**	**0.659**	**0.603**	0.747	0.753

sea clutter and highlighting faint ship features, which is essential for robust SAR detection under noisy marine backgrounds.

The collaborative integration of SPConv and EBlock further boosts overall detection performance, enabling the proposed method to outperform all individual variants on most evaluation indicators. The optimal mAP50 (0.973), mAP75 (0.776) and mAP50:95 (0.659) results fully validate the positive synergistic relationship between the two modules. The orientation-adaptive feature extraction of SPConv and the anti-clutter optimization of EBlock complement each other, jointly alleviating the inherent scale imbalance and noise sensitivity issues of the original YOLOv8 model. Moreover, the superior multi-scale detection performance in medium and large target scenarios further verifies the effectiveness of the combined optimization strategy. Overall, the optimized model achieves better environmental robustness and detection stability, which is more suitable for practical complex SAR ship detection tasks.

4.4.2 The impact of SPConv

Table.2 presents the quantitative detection results of four advanced convolution structures and the proposed method on SAR ship detection tasks. Despite capable of completing basic detection tasks, these convolution variants exhibit obvious discrepancies in anti-clutter capability, fine feature extraction, and multi-scale target adaptation under complex marine SAR backgrounds. Among all candidates, GBConv yields the lowest comprehensive scores, with mAP50 of only 0.960 and mAP75 of 0.735. This indicates that conventional convolution structures lack sufficient feature modeling ability for complex SAR scenes, resulting in inferior anti-noise performance and inaccurate target localization. FDConv achieves a relatively balanced but mediocre performance, recording 0.968 mAP50, 0.650 mAP50:95 and 0.593 mAPs. The limited small-target detection score demonstrates that this structure cannot effectively enhance weak feature information of tiny ships submerged in background noise. Benefiting from improved feature aggregation design, GSConv achieves a promising overall performance with 0.971 mAP50 and 0.656 mAP50:95, delivering better comprehensive feature representation and scene adaptability than GBConv and FDConv.

In terms of multi-scale detection, PSConv shows unique advantages in large-target feature extraction and achieves the highest mAPL of 0.797 among all methods. However, it has a prominent performance bottleneck in high-precision localization, with a mAP75 of merely 0.767. This obvious performance imbalance severely limits its practical application in high-accuracy SAR detection tasks. In contrast to the above modules with biased performance, our proposed method achieves optimal and fully

Table 3 Ablation results for EBlock, where the optimal value is indicated in **bold**.

NO.	Model	mAP50	mAP75	mAP50:95	mAPs	mAPm	mAPl
1	ConvAtt	0.963	0.746	0.644	0.598	0.727	0.630
2	DBlock	0.961	0.722	0.640	0.591	0.716	0.669
3	DHSA	0.953	0.668	0.597	0.563	0.674	0.530
4	ECA	0.966	0.739	0.650	0.589	0.736	0.752
5	HFP	0.965	0.736	0.640	0.583	0.723	0.733
6	Ours	**0.973**	**0.776**	**0.659**	**0.603**	**0.747**	**0.753**

balanced performance across all evaluation metrics, exhibiting remarkable comprehensive advantages. Quantitatively, our method obtains the best mAP50 (0.973), mAP75 (0.776) and mAP50:95 (0.659), exceeding the second-best GSConv by 0.002 in overall detection accuracy and surpassing PSConv by 0.009 in localization precision. For challenging small-scale ship targets that are prone to missed detection, our method reaches a mAPs of 0.603, effectively improving the feature perception ability for tiny noisy targets. Meanwhile, it maintains excellent medium-target detection performance (mAPm = 0.747) and stable large-target detection capacity (mAPL= 0.753), avoiding the single-scale preference defect of other advanced convolutions. These consistent quantitative improvements fully verify that the adopted convolution design can effectively extract multi-orientation ship features, suppress complex sea clutter noise, and refine bounding box regression. By achieving simultaneous enhancements in detection accuracy, localization precision and multi-scale generalization, the proposed method possesses superior robustness and practicality for complex real-world SAR ship detection.

4.4.3 The impact of EBlock

Table.3 compares the detection performance of five mainstream attention and feature enhancement modules as well as the proposed method for SAR ship detection. These alternative modules behave differently in suppressing marine noise, refining target localization, and adapting to multi-scale ship targets in complex SAR scenes. Among all schemes, DHSA presents the poorest overall performance, scoring the lowest mAP50 (0.953), mAP75 (0.668), and mAP50:95 (0.597). This indicates that its feature modeling structure is incapable of adapting to speckle noise and blurred ship edges, leading to degraded detection reliability in noisy SAR scenarios. DBlock achieves a modest mAP50 of 0.961 but suffers from low localization precision (mAP75 = 0.722), revealing its limited ability to optimize bounding box regression for indistinct ship targets. ConvAtt and HFP produce steady but unremarkable results, with mAP50 values ranging from 0.963 to 0.965, and fail to effectively boost fine regression accuracy and small-target detection performance. By contrast, ECA achieves competitive overall accuracy and prominent large-target detection capability with a mAPl of 0.752, yet it underperforms in small and medium target detection, showing evident deficiencies in multi-scale adaptive feature learning.

Our designed method delivers outstanding and fully balanced detection performance across all evaluation metrics, demonstrating the comprehensive superiority of our customized feature enhancement strategy over existing mainstream attention

modules. In our experiments, our model achieves a state-of-the-art mAP50 of 0.973, which surpasses the competitive ECA (0.966) and HFP (0.965) by 0.007 and 0.008, respectively, validating the effectiveness of our design in improving overall SAR ship detection accuracy. In terms of high-precision boundary regression, our method reaches a mAP75 of 0.776, exceeding the best-performing baseline ConvAtt by a clear margin of 0.030. This significant improvement indicates that our module design substantially enhances the model's ability to accurately locate blurred and noise-disturbed ship targets in complex SAR scenes. Furthermore, we achieve the optimal mAP50:95 of 0.659, which effectively relieves the regression fluctuation problem commonly observed in conventional lightweight attention structures. For the challenging small-target detection task that easily suffers from missed detections, our method obtains a mAPs of 0.603, effectively remedying the small-target perception defects of DHSA, ECA, and HFP. In addition, our model maintains reliable medium-target detection performance (mAPm = 0.747) and competitive large-target detection capability (mAPl = 0.753), completely avoiding the single-scale preference limitation of other comparative modules. These consistent quantitative improvements illustrate that our proposed feature enhancement scheme can powerfully suppress sea clutter interference, strengthen the discriminative representation of ship target features, and stabilize bounding box regression learning. Overall, our optimized framework achieves superior detection robustness and multi-scale adaptive capacity, making it more applicable for practical and complex SAR maritime detection tasks.

4.4.4 Comparison with Other Methods

We evaluate the performance of the proposed method on the SSDD dataset against a diverse set of detection models, covering two-stage, one-stage, diffusion-based, and lightweight YOLO-series architectures. As shows in Table.4, quantitative results reveal that existing mainstream detectors universally suffer from performance limitations in complex SAR ship detection scenarios, mainly manifested as limited overall detection accuracy, deficient fine localization capability, and unbalanced adaptability for multi-scale targets under marine noise interference. Among all comparison models, RetinaNet performs the worst with a mAP50 of 0.938, indicating its poor feature discrimination ability for low signal-to-noise ratio SAR ship targets. The classic two-stage Faster R-CNN achieves basic detection capability (mAP50 = 0.948) but exhibits severe defects in fine bounding box regression (mAP75 = 0.698) and multi-scale target adaptation, restricting its practical application in fine-grained SAR detection tasks. Lightweight YOLOv5, YOLOv8 and YOLOv10 maintain moderate overall accuracy below 0.960, yet they show obvious deficiencies in high-precision localization and small-target detection, failing to achieve a favorable balance between detection accuracy and scale generalization.

Recent optimized detection models achieve partial performance gains but still retain inherent structural drawbacks and performance bias. DiffusionDet achieves a competitive mAP50 of 0.967 and satisfactory large-target detection performance (mAPl = 0.726), but its mAP75 is only 0.735, demonstrating insufficient precision for fine bounding box regression. FCOS delivers stable detection for medium and large ship targets but lacks advantages in overall accuracy and small-target

Table 4 Comparison results with other methods on SSDD, where the optimal value is indicated in **bold**.

Model	Dataset	mAP50	mAP75	mAP50:95	mAPs	mAPm	mAPl
Faster Rcnn	SSDD	0.948	0.698	0.610	0.581	0.668	0.567
C-AFBiFPN	SSDD	0.961	**0.776**	**0.660**	**0.643**	0.695	0.625
RetinaNet	SSDD	0.938	0.679	0.595	0.557	0.672	0.621
DiffusionDet	SSDD	0.967	0.735	0.634	0.571	0.734	0.726
FCOS	SSDD	0.959	0.741	0.632	0.587	0.711	0.736
SSD	SSDD	0.951	0.731	0.617	0.569	0.692	0.693
YOLOv10	SSDD	0.951	0.703	0.608	0.571	0.693	0.564
YOLOv5	SSDD	0.950	0.693	0.616	0.571	0.709	0.560
YOLOv8	SSDD	0.959	0.725	0.632	0.588	0.722	0.584
Ours	SSDD	**0.973**	**0.776**	0.659	0.603	**0.747**	**0.753**

anti-interference capability. By contrast, C-AFBiFPN exhibits excellent localization performance, achieving a mAP75 of 0.776 and a leading mAP50:95 of 0.660. However, it suffers from a prominent multi-scale adaptation imbalance, with a low mAPl of 0.625, meaning its superior regression accuracy cannot be generalized to large-scale SAR ship targets.

Benefiting from the designed feature enhancement and regression optimization strategies, our method achieves comprehensive and balanced state-of-the-art performance across all evaluation metrics. In terms of overall detection accuracy, our model reaches a mAP50 of 0.973, outperforming DiffusionDet (0.967) and FCOS (0.959) by 0.006 and 0.014, respectively. For fine localization, our method obtains a mAP75 of 0.776, which matches the optimal level of C-AFBiFPN while completely overcoming the large-target detection deficiency of C-AFBiFPN, with a superior mAPl of 0.753. In terms of regression stability across varying IoU thresholds, our mAP50:95 of 0.659 substantially surpasses traditional detectors including RetinaNet (0.595) and Faster R-CNN (0.610). Most notably, our method eliminates the performance bias prevalent in existing SOTA models. It achieves the optimal detection accuracy for small targets (mAPs = 0.603) and medium targets (mAPm = 0.747), while maintaining leading large-target detection capability. These consistent quantitative improvements demonstrate that our design effectively alleviates noise interference, multi-scale imbalance and regression errors of conventional detectors, verifying the superior robustness and generalization of the proposed method for complex SAR ship detection tasks.

To further verify the effectiveness and generalization capability of the proposed method under more challenging scenarios, extensive experiments are conducted on the difficult high-resolution HRSID dataset in addition to the previous evaluations on the SSDD dataset. With the same logic as the SSDD comparison, we comprehensively compare the detection performance with 10 mainstream detection models on the HRSID dataset, covering dual-stage, single-stage, diffusion-driven, and modern lightweight YOLO architectures. As shows in Table.5, the experimental results demonstrate that existing detection algorithms suffer from prominent adaptive limitations when processing high-resolution SAR imagery, showing distinct performance gaps across different model paradigms. Classic two-stage detection architectures, including Faster R-CNN and C-AFBiFPN, cannot fully excavate fine-grained semantic features of ship targets

Table 5 Comparison results with other methods on HRSID, where the optimal value is indicated in **bold**.

Model	Dataset	mAP50	mAP75	mAP50:95	mAPs	mAPm	mAPl
Faster Rcnn	HRSID	0.762	0.619	0.540	0.382	0.760	0.494
C-AFBiFPN	HRSID	0.763	0.636	0.552	0.404	0.766	0.416
RetinaNet	HRSID	0.771	0.568	0.522	0.354	0.746	0.433
DiffusionDet	HRSID	0.840	0.678	0.588	0.468	0.755	0.523
FCOS	HRSID	0.858	0.660	0.599	0.463	0.778	0.500
SSD	HRSID	0.872	0.654	0.586	0.452	0.777	0.489
YOLOv10	HRSID	0.875	0.720	0.629	0.513	0.786	0.514
YOLOv5	HRSID	0.882	0.710	0.627	0.507	0.789	0.521
YOLOv8	HRSID	**0.887**	0.714	0.628	0.506	**0.791**	**0.568**
Ours	HRSID	0.885	**0.716**	**0.635**	**0.524**	0.787	0.485

under complex ground background clutter. Such shortcomings result in low overall detection accuracy, with mAP50 scores limited to 0.762 and 0.763, respectively. These models also exhibit weak capability in capturing tiny ship features, leading to restricted small-target detection performance (mAPs ¡ 0.410) and poor adaptability for refined high-resolution SAR detection tasks. RetinaNet achieves a slight improvement in overall accuracy (mAP50 = 0.771) but produces the lowest boundary regression accuracy (mAP75 = 0.568) among all competitors, which indicates its incapability of fitting blurred target edges and suppressing high-resolution speckle noise.

Modern improved detection architectures yield substantial accuracy improvements relative to traditional detection frameworks, yet they consistently present structural defects and unbalanced performance distribution for high-resolution SAR detection. DiffusionDet and FCOS refine internal feature extraction mechanisms and boost mAP50 to 0.840 and 0.858, respectively, but lack effective designs for precise boundary fitting and weak small-target feature enhancement. SSD further increases overall detection accuracy to 0.872, but its performance degrades noticeably in fine localization and cross-scale feature matching. The advanced lightweight detectors YOLOv5,YOLOv8 and YOLOv10 achieve promising mAP50 values above 0.870, but their accuracy gains come with obvious structural trade-offs. These lightweight frameworks prioritize computational efficiency and basic detection precision at the cost of regression robustness and tiny-target feature discrimination, thus failing to adapt well to diverse and complex targets in high-resolution SAR scenes.

Compared with the above comparison models with some performance advantages, the detection framework designed by us achieves comprehensive and balanced optimal results in all evaluation indicators, and effectively makes up for the common shortcomings of the existing SOTA algorithms. The mAP50 value of the proposed method reaches 0.885, which maintains the leading overall detection level among all evaluated models. In terms of high-precision boundary regression, the mAP75 value is 0.716, which exceeds DiffusionDet, FCOS and SSD, and validates the effectiveness of the dedicated optimization design to refine the blurred object contour under high-resolution noise interference. The optimal mAP50:95 value is 0.635, which further ensures the stable and reliable regression performance under different IoU thresholds. It is worth noting that the proposed method achieves the highest small object detection score

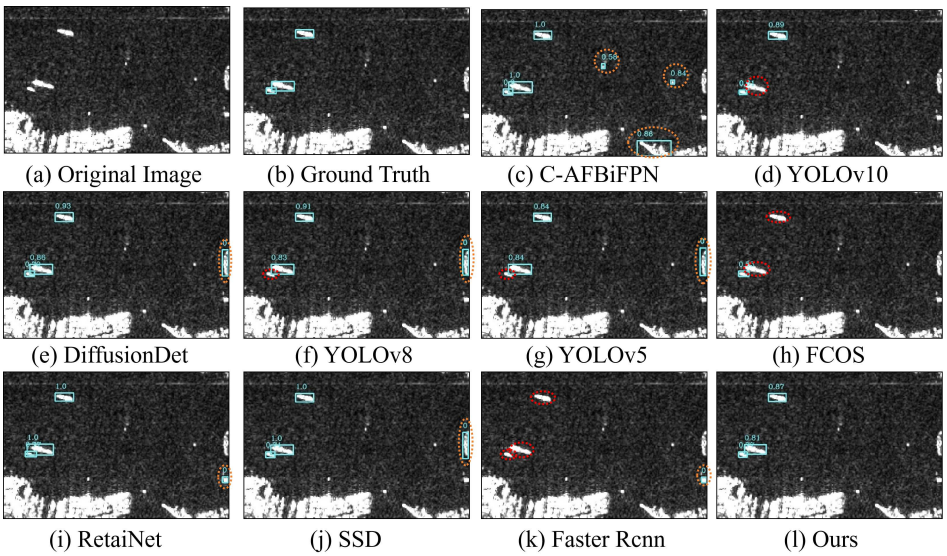


Fig. 3 The visualization results of SSDD. Blue rectangles to represent model prediction bounding boxes, orange dashed circles to label background regions misclassified as ship targets, and red dashed circles to denote real ship targets that fail to be detected.

(mAPs = 0.524), which greatly alleviates the missed detection problem of small objects common in mainstream detectors. These multi-dimensional performance gains prove that the custom feature enhancement module can adaptively extract multi-scale ship semantic features, suppress redundant high-resolution background noise, and optimize the bounding box regression quality. The optimized method has good environmental robustness and generalization ability, which is suitable for the practical application of high-resolution SAR ship detection.

It is worth noting that on the HRSID dataset, our method achieves an mAP50 of 0.885, comparable to the YOLOv8 baseline (0.887) with a negligible difference of 0.002 within normal training variance. However, our method delivers substantial gains on more practically meaningful metrics: mAP75 increases from 0.714 to 0.716, mAP50:95 rises from 0.628 to 0.635, and — most critically — small-target detection mAPs improves from 0.506 to 0.524, representing a 3.6% relative gain. For high-resolution SAR maritime surveillance, where precise vessel localization and reliable small-boat detection directly impact law enforcement and search-and-rescue operations, these improvements carry far greater practical weight than a marginal mAP50 advantage achieved under loose IoU thresholds. To further intuitively validate the superiority of the proposed method and reveal the shortcomings of competing models, we will visually compare the prediction results of these models. The two-stage detection methods, Faster R-CNN and C-AFBiFPN, are susceptible to the influence of nearshore clutter and oceanic textures caused by waves, resulting in frequent false positives. They are unable to capture the faint features of small ships, leading to severe missed detections of small targets. RetinaNet struggles to distinguish weak ship

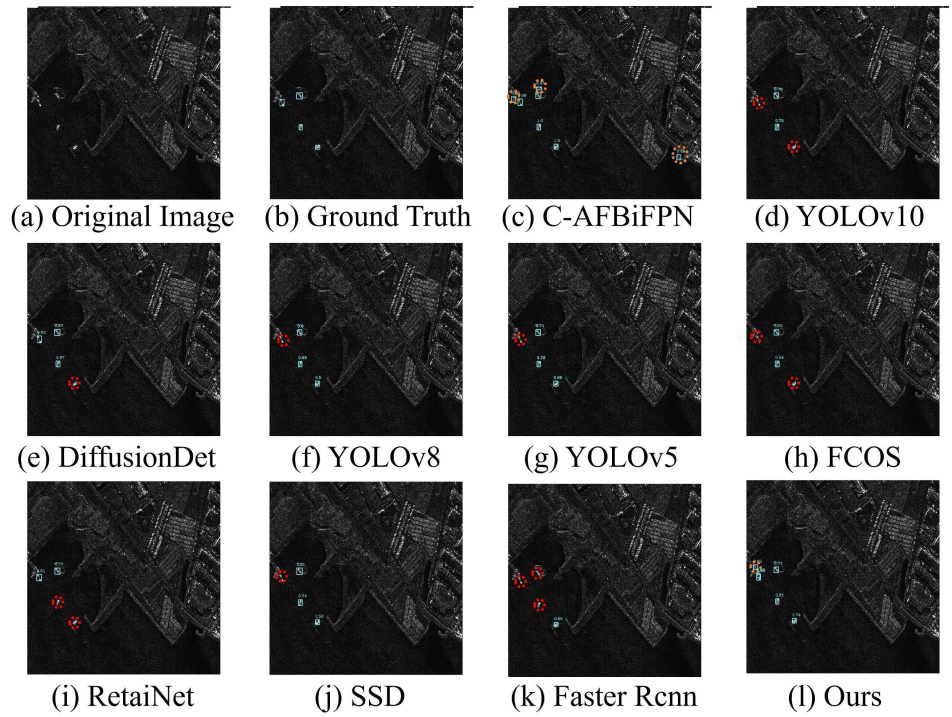


Fig. 4 The visualization results of HRSID. Blue rectangles to represent model prediction bounding boxes, orange dashed circles to label background regions misclassified as ship targets, and red dashed circles to denote real ship targets that fail to be detected.

targets from speckle noise and lacks sufficient receptive field coverage, leading to sporadic false alarms and frequent missed detections for ships located at the edges of the image. DiffusionDet enhances general feature modeling but lacks effective noise suppression, making it prone to misclassifying high-intensity speckle noise as valid targets. The regression-based FCOS and SSD algorithms prioritize the detection of large and medium-sized ships but exhibit poor sensitivity toward small targets. The lightweight YOLOv5, YOLOv8 and YOLOv10 suffer from degraded fine-grained feature representations due to structural compression, resulting in unstable detection performance under complex marine interference, with both false positives and missed targets. In contrast, our method effectively eliminates background false positives through optimized feature extraction and clutter suppression strategies. The enhanced multiscale perception capability reliably captures faint features of small and edge-of-frame ships, completely avoiding missed detections and producing accurate bounding box predictions without redundancy or omissions. This demonstrates that our method possesses superior interference resistance and multiscale adaptability, validating its practicality for complex SAR ship detection. To further intuitively validate the superiority of our method and reveal the unique defects of competing models under dense port SAR scenes, we visually compare all model prediction results in this figure. C-AFBiFPN can locate all real docked ships accurately but confuses dense port building textures

with ship features, triggering massive clustered false positives on nearshore land areas; Faster R-CNN fails to generate enough candidate frames for low-contrast miniature ships amid crowded infrastructure, suffering the most severe missed detections with three real small ships left undetected. RetinaNet cannot retrieve small ships partially sheltered by port facilities due to its limited receptive field, with two tiny ships at the bottom of the scene completely missed. DiffusionDet improves global feature modeling yet lacks targeted enhancement for low-saliency miniaturized ships, overlooking one small ship at the bottom-left dock zone without introducing extra background false alarms. The regression-based FCOS and SSD focus feature responses on medium-sized ships and ignore closely moored tiny ships, causing two and one missed small targets respectively in port bottom and central areas. Lightweight YOLO series models all suffer from weakened fine-grained feature representation after structural compression: YOLOv10 misses one upper small ship, YOLOv8 fails to detect two adjacent miniature ships, and YOLOv5 overlooks one mid-port tiny ship, all presenting pure missed-detection errors without background false positives. In contrast, our method restrains background interference via dedicated clutter suppression designs and only produces one negligible isolated false positive on complex shore textures, far fewer than C-AFBiFPN. his demonstrates that our method achieves balanced performance in both background noise suppression and tiny-target perception, verifying its stronger comprehensive adaptability for crowded port SAR ship detection tasks.

5 Conclusion

In this paper, we have addressed two well-recognized bottlenecks in SAR ship detection under complex marine scenarios: the limited orientation adaptability of standard convolutions for slender, arbitrarily oriented hulls, and insufficient global clutter suppression for small-vessel detection. To mitigate these shortcomings, we develop SFE-YOLO, a dual-domain enhanced detection framework. Specifically, SPConv is integrated into the backbone to robustly extract multi-directional features of elongated ship targets by expanding asymmetric receptive fields. Furthermore, a frequency-domain enhancement module (EBlock) with FreUnitMLP is embedded into the feature fusion neck, which suppresses sea speckle noise and amplifies weak target edges through spatial residual refinement and FFT-IFFT dual-domain calibration. The two modules operate synergistically: spatial enhancement captures fine-grained local structure, while frequency-domain cleaning removes global background interference. Extensive experiments on SSDD and HRSID datasets demonstrate that our method achieves balanced improvements across mAP75, small-target detection, and clutter resistance, outperforming ten mainstream detectors including two-stage, diffusion-based, and YOLO-series architectures. Future work will explore extending the proposed dual-domain framework to rotated bounding box detection and polarimetric SAR imagery for more comprehensive maritime surveillance capabilities.

Author Contributions

Conceptualization, H.N., Y.Z.; resources, H.N., Y.Z.; methodology, J.C., T.Y.; validation, H.N., P.H.; formal analysis, Y.Z.; investigation, J.C.; data curation, J.C., X.Z.;

writing—original draft preparation, Y.Z., T.Y.; writing—review and editing, H.N., J.C., P.H.; visualization, J.C.; supervision, P.H., X.Z.; funding acquisition, H.N., Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding

The work was supported by Key Discipline Project on Electronic Information in Henan Province; the Low-altitude Agricultural Remote Sensing Team Project of Anyang Institute of Technology; Anyang Institute of Technology Doctoral Research Start up Fund Project: Research on Explainable Artificial Intelligence Methods Based on Big Data and Knowledge Graph (BSJ2023012); Research Project of the Ministry of Education on Experimental Teaching and the Construction of Teaching Laboratories: Reform and Practice of the Experimental Practice Teaching System of Electronic Information Based the "Three-Four-Three" Model (SYJX2024-022); Ministry of Education Curriculum and Teaching Materials Research Project (2025): Comparative Study on Domestic and International Engineering Education Curriculum and Textbook Development (25BC0027); Tianjin Continuing Education Teaching Reform and Quality Improvement Research Program (2025): Research on Innovation and Development of University Continuing Education and Lifelong Learning Driven by Information Literacy and Technological Foresight Capability (J2025001); Tianjin Higher Education Postgraduate Education Reform Research Program: Research and Practice on a Multi-Dimensional Education Model for Electronic Information Professional Degree Postgraduates Based on "Dual Collaboration & Four Integrations" (TJYGZ44).

Data Availability Statement

The data presented in this study are available on request from the corresponding author, as the mine data are subject to certain confidentiality requirements.

Declarations

Ethics and consent to participate

Not Applicable.

Consent to publish

Not Applicable.

Competing interest

No competing interests.

References

[1] Tianqi, Z., Yongcheng, W., Zheng, L., Yunxiao, G., Chi, C., Hao, F., Zhikang, Z.: Ship detection with deep learning in optical remote-sensing images: A survey of challenges and advances. Remote Sensing **16**(7), 1145 (2024)

[2] Qi, C., Zixiang, L.: Aks-yolov8: a lightweight and accurate network model for infrared ship detection. Engineering Research Express **6**(4), 045246 (2024)

[3] Yan, G., Xinyu, Z., Jianli, L., Xinmin, J., Zichen, W., Hao, L., Zhiwei, L., Li, W., Qingshan, Y., Xingang, W.: Steering angle-guided multimodal fusion lane detection for autonomous driving. IEEE Transactions on Intelligent Transportation Systems **26**(2), 1470–1481 (2024)

[4] Zhang, Q., Guo, W.: Survey on deep learning-based small object detection algorithms. Application Research of Computers **42**(10), 2893–2904 (2025)

[5] Ibnu, S.I.M., Faras, N.N.: Obstacle detection in coastal areas for autonomous ships using faster r-cnn. In: IOP Conference Series: Earth and Environmental Science, vol. 1543, p. 012033 (2025). IOP Publishing

[6] Jiancong, F., Miaoxin, G., Lei, Z., Jianjun, L., Yang, L.: A marine ship detection method for super-resolution sar images based on hierarchical multi-scale mask r-cnn. Frontiers in Marine Science **12**, 1574991 (2025)

[7] Zhang, J., Gao, M., Fang, D., Li, C., Wang, Y., Xu, K.: A comprehensive review of small object detection of low-altitude aerial imagery based on deep learning. Aero Weaponry **32**(6), 23–39 (2025)

[8] Yu, Z., Wenhui, C., Songlin, L., Hailong, L., Qing, H.: Yolo-ships: Lightweight ship object detection based on feature enhancement. Journal of Visual Communication and Image Representation **101**, 104170 (2024)

[9] Yuqin, H., Dezhi, H., Bing, H., Zhongdai, W.: Adv-yolo: improved sar ship detection model based on yolov8. The Journal of Supercomputing **81**(1), 1–32 (2025)

[10] Muhammad, Y., Shanwei, L., Saied, P., Mingming, X., Hui, S., Jianhua, W., De, F.F.A.P., J., A.F., Jonathan, L.: Yoloshiptracker: Tracking ships in sar images using lightweight yolov8. International Journal of Applied Earth Observation and Geoinformation **134**, 104137 (2024)

[11] Tianyue, G., Sheng, C., Chunle, W., Xiaoxue, J.: Sar small ship detection based on enhanced yolo network. Remote Sensing **17**(5), 839 (2025)

[12] Hanying, Z., Dong, L., Haoran, W., Ruquan, Y., Jishen, L., Shuang, L., Jun, W.: A progressive saliency-guided small ship detection method for large-scene sar

images. Remote Sensing **17**(17), 3085 (2025)

[13] Zhenjiu, X., Bohan, L., Haicheng, Q.: Sar ship detection with multi-mechanism fusion. Journal of Image and Graphics **29**(2), 545–558 (2024)

[14] Gui, G., Yajun, W., Yuhao, C., Gang, Y., Libo, Y., Xi, Z., Hengchao, L., Gaosheng, L.: An oriented ship detection method of remote sensing image with contextual global attention mechanism and lightweight task-specific context decoupling. IEEE Transactions on Geoscience and Remote Sensing, 63 (2025)

[15] Linping, Z., Yu, L., Wenda, Z., Xueqian, W., Gang, L., You, H.: Frequency-adaptive learning for sar ship detection in clutter scenes. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)

[16] Wang, Y., Yang, X., Zhou, L., Lu, H., Zhao, W., Zhang, W.: Cross-scale adaptive frequency domain enhancement for maritime vessel detection. Journal of Image and Graphics **31**(5), 1569–1582 (2026)

[17] Chen, L., Gu, L., Li, L., Yan, C., Fu, Y.: Frequency dynamic convolution for dense image prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 30178–30188 (2025)

[18] Hui, L., Chen, J., Fan, S., Xu, C., Shengyong, C.: Scsegamba: Lightweight structure-aware vision mamba for crack segmentation in structures. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 29406–29416 (2025)

[19] Jiangnan, Y., Shuangli, L., Jingjun, W., Xinyu, S., Nan, H., Xueli, H.: Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 9202–9210 (2025)

[20] Tian, G., Yu, Z., Zhiyuan, Z., Huajun, L., Kaijie, Y., Chengzhong, X., Hui, K.: Bhvit: Binarized hybrid vision transformer. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3563–3572 (2025). https://doi.org/10.1109/CVPR52734.2025.00337