Computer Science and Information Systems 23(2):729–756 https://doi.org/10.2298/CSIS251018019X

YOLO-BDM: An Improved Ship Detection Algorithm Based on YOLOv11n

Fangyuan Xiong[1], Dezhi Han[2], Xiang Shen[3], and Manlin Zhu[1]

[1] School Of Information Engineering, Shanghai Maritime University,1550 Haigang Avenue, Shanghai, 201306, Shanghai, China
xiongfangyuan@stu.shmtu.edu.cn (corresponding author)
[2] Shanghai Ship and Shipping Research Institute Co., Ltd., Shanghai, 200135, China
dzhan@shmtu.edu.cn
[3] School of Computer Science, The University of Sydney, Sydney, New South Wales, Australia
shenxiang1107@163.com

Abstract. Synthetic Aperture Radar (SAR) ship detection is crucial for maritime traffic management, search and rescue, and environmental monitoring but remains challenging due to small targets, blurred contours, and complex ocean backgrounds. To address these issues, this paper proposes YOLO-BDM, an improved detector based on YOLOv11. The Diverse Branch Block (DBB) is introduced into the backbone to enhance feature representation through multi-branch training and reparameterized inference. A Multi-scale Contextual Attention (MCA) mechanism is integrated into the backbone and neck to strengthen multi-scale semantic modeling and background discrimination. Additionally, a four-layer Bidirectional Feature Pyramid Network (BiFPN) is employed for efficient multi-scale feature fusion. Experiments on the SAR-Ship dataset show YOLO-BDM achieves 97.27% mAP, 94.11% Precision, and 93.07% Recall, surpassing the baseline and validating its effectiveness.

Keywords: Ship detection; BiFPN module; YOLOv11n; DBB module; MCA attention mechanism

1. Introduction

Object detection, as a core problem in computer vision, has garnered extensive attention in recent years across diverse application scenarios including intelligent transportation [38], medical imaging [17], industrial automation [10], and remote sensing image processing [35]. Synthetic Aperture Radar (SAR), with its all-weather, all-time imaging capabilities and robustness in complex environments, has emerged as a vital tool for ocean monitoring and vessel detection [37]. However, ship targets in SAR images often present challenges such as small size, blurred contours, complex backgrounds, and speckle noise interference [20], making high-precision detection difficult. Achieving robust and efficient SAR ship detection in complex marine environments has thus become a critical research topic in intelligent remote sensing perception, holding significant theoretical and practical value.

Early SAR ship detection methods primarily relied on manually designed features and shallow classifiers. However, constrained by noise interference and single-scale feature modeling, their robustness and generalization capabilities were limited [31]. To overcome this bottleneck, researchers gradually shifted toward detection frameworks driven

by convolutional neural networks (CNNs) [46]. Regarding two-stage detectors, the R-CNN series proposed by Girshick et al. laid the foundation for end-to-end detection. Ren et al. introduced the Region Proposal Network (RPN) through Faster R-CNN, significantly enhancing candidate box generation and detection efficiency. Subsequently, Lin et al. proposed the Feature Pyramid Network (FPN) [24], strengthening multi-scale feature representation; Cai and Vasconcelos proposed Cascade R-CNN [1], improving detection accuracy at high IoU thresholds through stepwise optimization; Pang et al. introduced Libra R-CNN [27], achieving improvements in sample allocation and feature utilization. While these methods excel in detection accuracy and semantic modeling, their slow inference speeds and high computational complexity hinder real-time detection requirements.

Against this backdrop, single-stage detectors have gradually emerged as a research hotspot. Frameworks such as YOLO [28] and SSD [25] have achieved significant improvements in inference speed while maintaining reasonable detection accuracy. However, early single-stage methods exhibited limitations in modeling complex textures and representing multi-scale features, particularly in low signal-to-noise ratio conditions where small objects were prone to detection failures. To address this, Dai et al. introduced Deformable Convolutions (DCN) [9], enhancing the adaptability of convolutions to geometric deformations and intricate textures. Hu et al. proposed Channel Attention Mechanisms [18], improving model discrimination in complex backgrounds through feature weighting. Driven by these advancements, SAR vessel detection has seen further development. For instance, Ma et al. proposed a free-bounding box detection method based on keypoint estimation and attention mechanisms, effectively suppressing false alarms in complex backgrounds [26]. However, it still suffers from insufficient discrimination between adjacent targets in high-density scenes due to ambiguous keypoint matching. Zhao et al. introduced the CRAS-YOLO model [43], achieving high-precision detection and classification of multi-category vessels, though its category coverage remains limited. Zhou et al. proposed the FGNet model [44], integrating a global context module and multi-scale feature enhancement module to improve target discrimination and multi-scale feature representation in complex scenes. Nevertheless, false negatives and false positives may still occur in strongly cluttered coastal environments. Overall, although notable progress has been achieved in SAR ship detection, existing methods still encounter intrinsic limitations when applied to complex marine environments [30], In particular, most YOLO-based detectors adopt conventional multi-scale feature fusion strategies that inadequately address semantic misalignment across feature levels, leading to suboptimal performance in small and densely distributed ship detection tasks. This issue is especially pronounced in SAR imagery, where speckle noise and weak target boundaries further degrade the reliability of shallow feature representations.

In summary, although existing methods have made some progress in SAR ship detection tasks, they still face numerous challenges in multi-scale modeling, feature focusing, and model deployment. This limitation is particularly pronounced in SAR ship detection scenarios involving small targets, where speckle noise and weak object boundaries further impair the reliability of shallow feature representations. Fundamentally, this problem stems from the insufficient correction of semantic bias during multi-scale feature fusion and the lack of effective contextual modeling to align features across different scales [5]. Furthermore, existing attention-enhanced YOLO-style architectures often focus on either channel-wise or spatial-wise feature modulation in isolation, lacking the capacity for joint

contextual modeling across multiple dimensions [4]. Finally, in pursuit of accuracy, certain detection frameworks incorporate extensive convolutional stacking and redundant modules, resulting in structurally complex architectures with substantial parameter scales. This hinders efficient deployment on edge computing or resource-constrained platforms. Consequently, achieving context-enhanced, lightweight modelling while maintaining detection precision has become an urgent research priority.More fundamentally, these limitations reflect a common design paradigm in existing YOLO-based SAR ship detectors, in which multi-scale feature fusion, attention-driven feature focusing, and structural efficiency are typically optimized independently rather than within a unified framework. Consequently, achieving a balanced integration of contextual representation enhancement, precise feature discrimination, and deployment-friendly efficiency remains an open challenge.

To address these challenges, this paper proposes the improved YOLO-BDM model based on the YOLOv11n framework. It aims to achieve precise feature extraction of small targets in complex SAR scenes, efficient fusion of multi-scale information, and suppression of background interference. Specific contributions include:

(1) The Diverse Branch Block (DBB) is introduced into the C3K2 architecture of the backbone network. During training, this multi-branch convolutional structure enriches feature representations. At inference time, it is equivalently transformed into a single convolutional layer through structural reparameterization. This approach balances model expressiveness with inference efficiency while mitigating the issue of edge weakening in small targets within SAR images.

(2) Embedding a Multi-scale Contextual Attention (MCA) mechanism at key nodes of the backbone and neck structures. This combines global average pooling with standard deviation pooling to extract multi-scale contextual information. Channel-wise dynamic weighting enhances responses in critical regions, effectively suppressing background noise interference and improving discrimination capabilities in complex environments;

(3) Introducing the BiFPN_Concat module into the neck network. It enhances interaction between shallow-layer details and deep-layer semantics through bidirectional feature propagation and learnable weight mechanisms. Simultaneously, feature concatenation preserves richer original information, improving flexibility in multi-scale modeling. This approach is particularly effective for detecting ship targets with significant scale variations.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed model and the techniques employed. Section 4 provides a detailed analysis of the experimental results. Section 5 concludes the paper and discusses future research directions.

2. Related Work

2.1. Traditional Ship Detection Algorithms

In SAR or optical imagery, traditional ship detection methods primarily rely on manually designed features and classical algorithms, typically including threshold segmentation, texture features, or keypoint descriptors. These methods are simple to implement,

computationally efficient, and can effectively detect targets in straightforward scenarios. However, their performance is limited in complex backgrounds, when dealing with small, multi-scale targets, or under low signal-to-noise ratio conditions [8]. To address these limitations, Constant False Alarm Rate (CFAR)-based detection methods have gained widespread application in complex maritime clutter environments. By adaptively estimating background noise, dynamically adjusting detection thresholds, suppressing false alarms, and maintaining high detection rates, CFAR-based approaches have become a key direction for improving traditional ship detection research [40].

However, traditional CFAR methods still suffer from limitations such as high computational complexity and low efficiency when processing high-resolution SAR images or large-scale scenes. Wang et al. [36] proposed a fast CFAR algorithm based on density screening (DC). By rapidly eliminating high-density background clutter using superpixel density features, it performs local detection only on a small number of candidate targets. This approach reduces computational complexity by 75%–96% while maintaining or even improving detection accuracy, effectively suppressing false alarms. It represents a significant breakthrough in balancing CFAR detection efficiency and accuracy. Furthermore, for complex coastal environments, Li et al. [23] proposed a hierarchical detection scheme for airborne single-channel SAR. This approach employs a K-log-normal mixture distribution model with adaptive background windows for CFAR prescreening, enhancing the resolution between sea clutter and targets. Subsequently, it introduces fine-grained discrimination based on micro-Doppler motion characteristics, further suppressing false alarms through radial velocity and image entropy analysis. This work demonstrates that traditional CFAR methods can achieve high-precision, low-false-alarm real-time detection on lightweight platforms by integrating statistical modeling with motion features. Despite CFAR's excellence in adaptive background estimation and false alarm suppression, issues of fitting inaccuracy and efficiency limitations persist in heterogeneous backgrounds and complex nearshore scenarios. To further address these challenges, Chen et al. [7] proposed a multi-modal saliency-based (MMS) vessel detection method. By integrating enhanced CFAR, superpixel (MSER), local stability analysis, and sea-land segmentation, they constructed four complementary saliency maps, effectively resolving fitting inaccuracies and oversegmentation in heterogeneous backgrounds.

Overall, the limitations of traditional ship detection methods—such as high false alarm rates, inaccurate fitting in complex backgrounds, and oversegmentation—have been significantly mitigated through continuous improvements by previous researchers. However, even enhanced traditional ship detection algorithms still struggle with challenges like detecting extremely small targets, handling multi-scale ships, and processing large-scale high-resolution images [16]. This indicates that addressing the practical challenges of SAR vessel detection requires not only innovation at the algorithmic level, but also consideration of reliability, manageability, and overall system performance at the architectural level [15].Furthermore, the advancement of deep learning techniques, such as convolutional neural networks (CNNs), has provided effective means for automatic feature extraction, multi-scale modeling, and end-to-end training. These developments have collectively propelled the evolution of deep learning-based ship detection algorithms, gradually establishing them as the mainstream approach in this field.

2.2. Deep Learning-Based Ship Detection Algorithms

Deep learning-based ship detection algorithms leverage convolutional neural networks (CNNs) to automatically extract hierarchical features, enabling end-to-end target localization and classification. Compared with traditional methods, such approaches significantly improve detection performance in complex backgrounds, for small-scale and multi-scale targets, and in high-resolution SAR imagery.

Existing deep learning-based ship detection studies can be broadly analyzed from different technical perspectives, including multi-scale feature modeling and localization accuracy, contextual attention and semantic discrimination, as well as lightweight and efficient architectural design. From an implementation standpoint, these methods are commonly realized through two-stage or single-stage detection frameworks, each emphasizing different trade-offs between accuracy and efficiency. Together, these advances have laid a solid foundation for continuous performance improvements in SAR ship detection [11].

Multi-scale Modeling and Localization-Oriented Methods Accurate localization of vessels across varying scales is a fundamental challenge in SAR ship detection, particularly in dense maritime scenes and complex cluttered backgrounds. To address this issue, many studies emphasize multi-scale feature modeling and precise localization, with two-stage detection frameworks serving as a representative implementation due to their explicit region proposal and refinement mechanisms.

Zhou et al. proposed UltraHi-PrNet [45], which improves feature alignment across scales via scale transfer and expansion layers but suffers from the computational inefficiency and generalization limits of its Faster R-CNN backbone. Tang et al.'s PEGNet [32], enhances Faster R-CNN with modules for better multi-scale fusion and noise suppression, yet its horizontal anchor design limits effectiveness for rotated targets in dense scenes. To address rotation issues, Zhang et al. proposed ORPSD [41]. using an outer rectangular projection scheme, though it retains the high computational cost typical of two-stage detectors.

In summary, multi-scale modeling and localization-oriented methods achieve high detection accuracy by explicitly refining candidate regions and integrating scale-aware features. However, the computational inefficiency and limited real-time performance of two-stage frameworks remain unresolved challenges, particularly for large-scale or resource-constrained SAR applications [6].

Contextual Attention and Semantic Discrimination Methods To address the limitations of pure multi-scale modeling in capturing long-range dependencies and discriminative features, researchers have developed methods that explicitly incorporate contextual attention and semantic enhancement mechanisms. A fundamental challenge for single-stage detectors has been to balance fine-grained feature interaction with high speed, a trade-off inherently linked to model stability and flexibility [29]. The evolution of frameworks like the YOLO series reflects this ongoing pursuit.

Representative advances in this direction include YOLOv4 [42],which introduced a CSPDarknet backbone to reduce computational redundancy and enhance feature diversity

through a split-and-fusion strategy. However, its reliance on anchor boxes limits generalization in scenes with significant scale and aspect ratio variations. To overcome such limitations, Hu et al. proposed BANet [19], an anchor-free design that integrates Local and Non-Local Attention Modules. This architecture improves fine-grained modeling of multi-angle vessels and contextual reasoning in complex backgrounds, though preserving fine texture details for small targets remains challenging.

In summary, methods focusing on contextual attention and semantic discrimination effectively address the limitations of conventional multi-scale approaches by emphasizing spatial context and fine-grained feature interactions. These techniques enhance detection robustness in complex maritime scenes, particularly for vessels with diverse orientations and subtle features, laying the groundwork for subsequent improvements in lightweight and efficient network architectures.

Lightweight and Efficient Single-Stage Detection Methods To address the challenges of deploying high-performance SAR ship detection models under resource constraints, research has focused on developing lightweight and efficient single-stage network architectures. Representative methods include: SSD-YOLO [12], an anchor-free framework enhanced with a multidimensional feature module to sharpen small target boundaries while maintaining real-time performance, though it struggles with complex contexts and fine details for very small or clustered vessels. FD-Net [13], incorporates deformable convolutions across the network, combined with an Enhanced Feature Pyramid and adaptive fusion module, to better represent vessels of varying scales and shapes, but its semantic integration and real-time efficiency are limited. LKE-Det [11]employs a decomposable large kernel to capture long-range dependencies and embeds edge gradient features to improve contour delineation and suppress clutter, yet efficient multi-scale fusion for subtle targets remains a challenge. RepGFPN [2], introduces efficient cross-layer connections and bidirectional fusion to aggregate shallow and deep features directly, enhancing detection of small and coastal targets, though preventing feature degradation during deep propagation is still a core issue [14].

In summary, existing lightweight and efficient architecture methods have achieved notable improvements in balancing computational cost, model complexity, and detection accuracy. Nevertheless, current approaches still face challenges in fully integrating multi-scale features, preserving fine-grained vessel details, and capturing rich contextual information under complex maritime conditions [3]. To address these limitations, we propose the YOLO-BDM vessel detection model, which combines enhanced multi-scale feature fusion, semantic discrimination, and lightweight design to achieve robust, real-time performance for SAR ship detection.

3. Methods

3.1. Baseline Model YOLOv11n

YOLOv11n is a lightweight variant of the YOLO series proposed in recent years, designed to improve object detection accuracy and stability while reducing computational costs. Compared to its predecessors, YOLOv11n systematically optimizes both the backbone and neck networks, notably incorporating modules such as C3K2 and C2PSA to enhance

feature extraction and fusion capabilities. As shown in Figure 1 its overall architecture primarily consists of a CSP-based backbone network, the C3K2 module, and the C2PSA attention mechanism, achieving a superior balance between detection performance and inference efficiency.

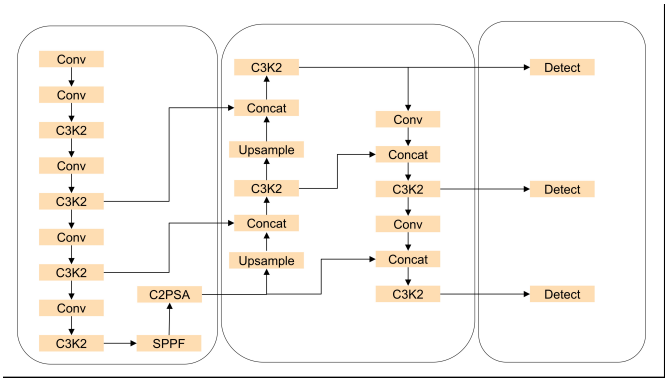


Fig. 1. YOLOv11n network architecture

In the backbone network, the C3K2 module proposed in YOLOv11n enhances the feature extraction architecture through two key innovations: dual-path residual connections and a dynamic receptive field mechanism. The dual-path design preserves lightweight characteristics while introducing a deformable convolution branch. The primary path retains standard convolutions to ensure computational efficiency, whereas the novel secondary path equips the model with stronger adaptability to multi-scale objects. The dynamic receptive field mechanism further enhances feature representation by leveraging multi-scale deformable convolutions. Moreover, unlike the channel compression strategies commonly employed in mainstream lightweight solutions, C3K2 adopts feature reorganization techniques to better preserve critical spatial information. The C3K2 module is illustrated in Figure 2.

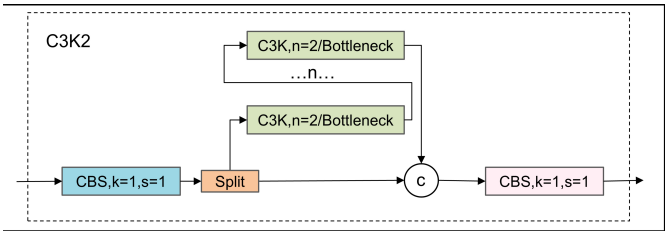


Fig. 2. C3K2 Module Architecture

In the Neck section, YOLOv11n employs the C2PSA module to enhance spatial modeling capabilities during multi-scale feature fusion. This module combines the CSP ar-

chitecture with the PSA mechanism, preserving efficient gradient flow pathways while strengthening the model's perception of target spatial locations. C2PSA first splits the input features into two branches: one follows the conventional convolutional path to preserve the original spatial structure, while the other introduces a parallel attention mechanism to model contextual regions at different scales. PSA captures semantic information within distinct receptive fields by constructing multiple sub-branches in parallel. These are then compressed and aggregated to guide the model's focus toward key areas, thereby enhancing robustness for objects with varying scales and complex backgrounds. The fused features from both branches ultimately generate representations with higher discriminative power. This module enhances spatial dependency modeling while mitigating the lack of global perception in shallow features. The C2PSA architecture is illustrated in Figure 3.

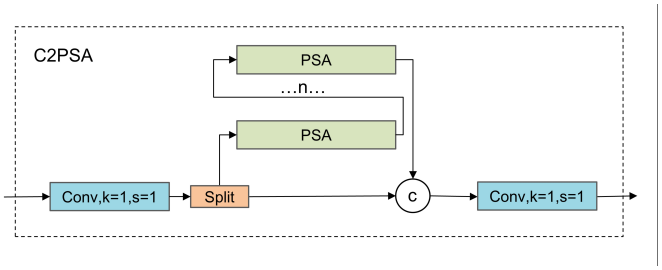


Fig. 3. C2PSA Module Architecture

During the feature fusion stage, YOLOv11n retains the dual-path architecture of FPN and PAN while adjusting network width and depth to meet lightweight deployment requirements. By introducing optimized activation functions (such as SiLU) and normalization operations in certain connection pathways, unnecessary computational redundancy is reduced. This strategy not only accelerates inference speed but also maintains a favorable balance between feature representation capability and model convergence performance, providing more stable high-level semantic support for subsequent modules.

With its compact structure, fast inference speed, and high detection accuracy, YOLOv11n demonstrates excellent adaptability in practical applications. Its optimizations in feature extraction capabilities, multi-scale modeling effects, and edge deployment efficiency make it particularly suitable for tasks sensitive to real-time performance and computational resources.

3.2. Improved Model YOLO-BDM

The overall architecture of the YOLO-BDM model is shown in Figure 4. Building upon YOLOv11n, this model introduces three key modules-DBB, MCA, and BiFPN-aimed at enhancing feature extraction and fusion capabilities to improve the accuracy and adaptability of vessel detection.

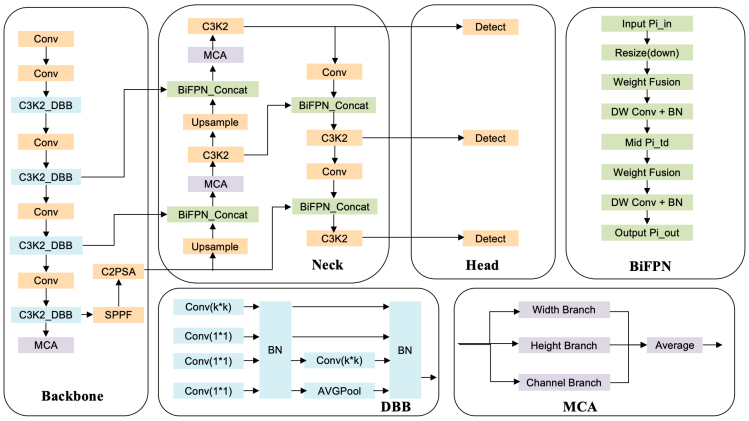


Fig. 4. YOLO-BDM network architecture

Within the YOLO-BDM model, systematic improvements to the backbone, neck, and feature fusion structures achieve unified optimization of multi-scale feature modeling and context-guided processing. In the backbone network, the Diversified Branch Block (DBB) replaces the original Bottleneck convolution, enabling multi-branch convolution fusion. This enhances the representation of objects across different scales and shapes, particularly boosting detection performance for small targets. Multiscale Contextual Attention Mechanism (MCA) is embedded at critical nodes in the backbone and neck layers. By dynamically adjusting attention distribution across channels, MCA effectively amplifies responses in target regions while suppressing background interference. The neck layer employs the BiFPN_Concat feature fusion module, which preserves multiscale information through concatenation operations, improving detection capabilities for objects with significant scale variations. Through the synergistic integration of DBB, MCA, and BiFPN, YOLO-BDM achieves significant improvements in ship detection accuracy and robustness while maintaining lightweight architecture. It is particularly well-suited for remote sensing scenarios involving small targets, ambiguous contours, and complex backgrounds.

3.3. Bidirectional Feature Pyramid Network (BiFPN)

Small vessels in remote sensing imagery typically exhibit characteristics such as compact dimensions, blurred edges, and indistinct textural features. These traits make it challenging for traditional object detection models to accurately locate and identify such targets within complex backgrounds, significantly compromising overall detection accuracy. To enhance the model's adaptability across multi-scale scenarios, this paper introduces the Bidirectional Feature Pyramid Network (BiFPN). This approach strengthens effective interactions between features at different levels and further optimizes the fusion strategy for multi-scale features.

Compared to conventional feature fusion methods like Feature Pyramid Network (FPN) and Path Aggregation Network (PANet), BiFPN introduces both top-down and bottom-up

information propagation pathways in its structural design, enabling efficient coupling of multi-level features. Furthermore, this architecture enhances cross-scale information aggregation through repeated stacking and incorporates a learnable weighted fusion mechanism. This allows the model to dynamically adjust the contribution ratios of different input features during the fusion process, thereby improving the selectivity and robustness of overall feature representation.

Given these advantages, this paper introduces the BiFPN module into the Neck section of YOLOv11n, replacing the original PANet structure to achieve more precise and stable multi-scale feature representation. The overall structure of this module is shown in Figure 5.

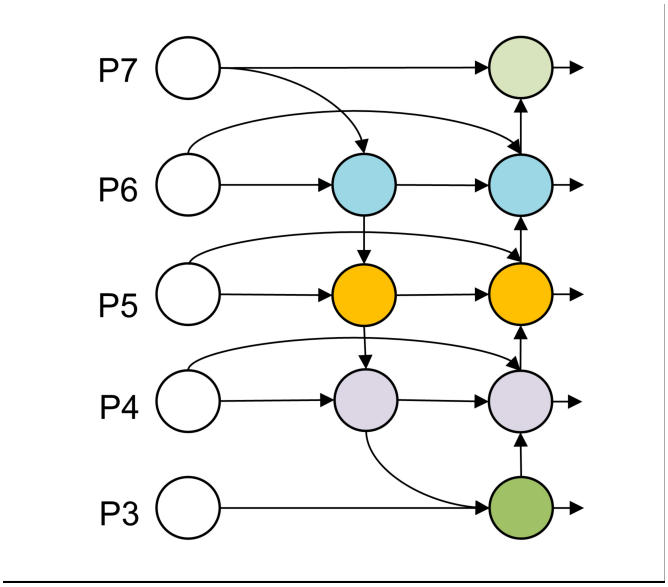


Fig. 5. BiFPN Architecture

Traditional feature pyramid structures typically employ uniform or static weighting when fusing multi-scale features, failing to dynamically adjust based on the varying contributions of different features in object detection tasks. BiFPN addresses this issue by introducing a learnable weighting fusion mechanism. This enables the network to adaptively allocate fusion weights for feature maps at different scales during training, thereby more effectively integrating shallow-layer detail information with deep-layer semantic representations. This enhances the network's modeling capability for objects at varying scales. During feature fusion, BiFPN assigns distinct trainable weights to multiple input feature maps. The fusion process can be represented as:

$$\boldsymbol{y} = \frac{w_1 \boldsymbol{x}_1 + w_2 \boldsymbol{x}_2}{w_1 + w_2 + \epsilon}. \tag{1}$$

In the formula, x_1 and x_2 represent input feature maps from different scales, w_1 and w_2 are their corresponding non-negative learnable weight parameters, and ϵ is a minimal constant introduced to prevent the denominator from becoming zero. This normalization mechanism not only ensures numerical stability but also enables the model to automatically learn the relative importance of different features during the fusion process. For scenarios with multiple input features, this formula can be naturally extended to:

$$\mathbf{y} = \frac{\sum_i w_i \cdot \mathbf{x}_i}{\sum_i w_i + \epsilon}. \tag{2}$$

This method avoids the shortcoming of simple weighted averaging, which treats all features equally, thereby endowing the network with stronger scale adaptability.

Structurally, FPN employs a top-down, unidirectional feature propagation approach, while PANet introduces bottom-up pathways to enhance the supplementation of deep-level semantics by shallow features. BiFPN further optimizes this framework by designing bidirectional pathways as independent fusion modules and processing feature maps through repeated multi-level stacking, thereby achieving efficient cross-level information exchange. Additionally, this architecture incorporates lateral cross-scale connections, enabling direct information exchange between feature maps at the same layer during fusion to enhance semantic consistency. Typically, BiFPN fuses feature maps from layers P3 to P7 to cover full-scale feature information spanning low-level details to high-level semantics.

To further enhance fusion efficiency and reduce redundant computations, BiFPN introduces a structural pruning strategy during the network construction phase. Specifically, for fusion nodes receiving input from only a single path (i.e., connected by only one upstream edge), the system determines that they do not constitute effective information exchange. Consequently, these nodes are directly skipped during fusion layer construction, thereby avoiding the introduction of ineffective fusion operations. This strategy can be formally described as follows:

$$\text{if } \text{fan}(n) = 1 \Rightarrow \text{prune}(n). \tag{3}$$

In the formula, fan(n) denotes the number of n inputs to node . When only a single input path exists, the system skips constructing this fusion node. This mechanism effectively simplifies the network architecture, reducing computational complexity and parameter size. It is particularly suitable for embedding BiFPN structures into lightweight object detection models, such as the YOLOv11n framework adopted in this paper.

Furthermore, the pruning operation does not affect feature flow along critical paths, as it only removes redundant "pseudo-fusion" nodes. Consequently, it significantly enhances computational efficiency while preserving detection performance.

As an enhanced feature fusion network, BiFPN plays a pivotal role in the YOLO-BDM model. Through strategies including weighted feature fusion, top-down and bottom-up bidirectional feature propagation, cross-scale connections, and pruning optimization, it enhances the model's detection capabilities for multi-scale objects and strengthens feature representation. This results in significantly improved detection accuracy in complex scenes.

3.4. Diverse Branch Block (DBB)

In object detection tasks, models need to simultaneously capture fine details of small objects and global semantics of large objects. Traditional convolutional layers, however, have limited feature representation due to their monolithic structure, making it difficult to effectively model multi-scale and diverse features in complex scenes. Multi-branch architectures, such as the Inception series, can enrich the feature space through branches of varying scales and complexities, but they incur significant inference overhead, limiting practical deployment. The Diverse Branch Block (DBB) addresses this issue by introducing diverse branches during training to enhance representation, which are equivalently merged into a single convolution at inference via structural re-parameterization, achieving "enhanced training representation with zero additional inference cost."

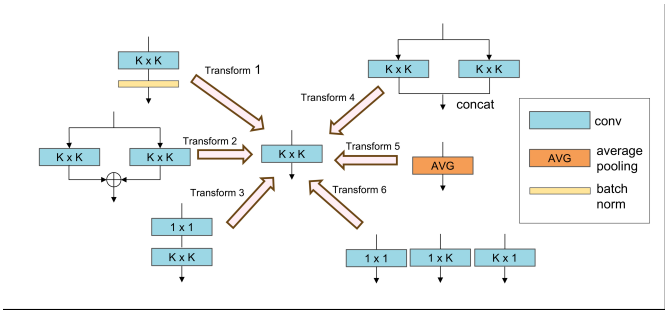


Fig. 6. DBB Architecture

The structure of the DBB module is illustrated in Figure 6. This module consists of multiple complementary branches, including a standard $k \times k$ convolution branch, a sequentially stacked 1×1–$k \times k$ convolution branch, a standalone convolution branch, and an average pooling (AVG Pooling) branch. Additionally, the module allows the incorporation of non-square convolutions (e.g.$1 \times k$, $k \times 1$,) to further expand the receptive field. Each branch is equipped with a batch normalization (BN) layer, introducing nonlinearity during training and significantly enhancing representational capability. Unlike Inception, DBB can fold all branches into a single convolution layer at inference through strict numerical equivalence transformations, avoiding the extra computational and memory overhead of multi-branch structures. This characteristic enables DBB to improve feature representation while maintaining efficient inference speed.

In mathematical modeling, the core of DBB lies in leveraging the linearity of convolutions and structural re-parameterization to equivalently transform the multi-branch structure during training into a single convolution at inference. Let the input feature be $I \in \mathbb{R}^{C\times H\times W}$, the convolution kernel $F \in \mathbb{R}^{D\times C\times K\times K}$, and the bias $b \in \mathbb{R}^{D}$. The basic convolution operation can be expressed as:

$$O = I * F + \mathrm{REP}(b) \tag{4}$$

where $\mathrm{REP}(b)$ denotes bias expansion. The key principle of DBB is based on the homogeneity and additivity of convolutions:

$$I * (pF) = p(I * F), \quad \forall p \in \mathbb{R}, \tag{5}$$

$$I * F^{(1)} + I * F^{(2)} = I * \left(F^{(1)} + F^{(2)}\right). \tag{6}$$

Leveraging this property, DBB designs six types of equivalent transformations (Transform I–VI) to progressively fold multi-branch operations into a single convolution. First, Transform I illustrates the fusion of convolution and batch normalization (BN). By absorbing the scaling and shifting parameters of BN, the convolution kernel and bias can be redefined as:

$$F'_j = \frac{\gamma_j}{\sigma_j} F_j, \quad b'_j = -\frac{\mu_j \gamma_j}{\sigma_j} + \beta_j, \tag{7}$$

thus eliminating the need for an additional BN layer during inference. Building on this, Transform II utilizes the additivity of convolutions: if multiple convolution branches share the same configuration, their weights and biases can be directly summed:

$$F' = F^{(1)} + F^{(2)}, \quad b' = b^{(1)} + b^{(2)}, \tag{8}$$

For more complex cases, Transform III addresses sequentially stacked 1×1 and $k \times k$ convolutions. Since the former only performs channel mixing, it can be merged with the latter into a single equivalent convolution:

$$F' = F^{(2)} * TRANS\left(F^{(1)}\right), \quad b' = \widehat{b} + b^{(2)}, \tag{9}$$

Transform IV corresponds to the channel concatenation commonly seen in Inception structures. Essentially, it concatenates the convolution kernels and biases of multiple branches along the output channel dimension, equivalent to a wider convolutional layer:

$$F' = \mathrm{CONCAT}\left(F^{(1)}, F^{(2)}\right), \quad b' = \mathrm{CONCAT}\left(b^{(1)}, b^{(2)}\right), \tag{10}$$

Additionally, DBB supports mapping non-convolution operations into convolutions. Transform V shows that average pooling can be regarded as a convolution with fixed weights:

$$F'_{d,c,:,:} = \begin{cases} \dfrac{1}{K^2}, & d = c, \\ 0, & d \neq c. \end{cases} \tag{11}$$

Finally, Transform VI demonstrates how non-square convolutions (e.g.$1 \times k$, $k \times 1$,) can be expanded via zero-padding into standard $k \times k$ convolutions, ensuring that all branches can be merged into a uniform form.

In summary, DBB achieves a strict mapping from "multi-branch during training" to "single convolution during inference" through these six transformations. This allows the model to exploit diverse paths for enhanced representation during training while maintaining the same computational cost as standard convolutions during inference, balancing performance and efficiency.

3.5. Multi-Scale Contextual Attention (MCA)

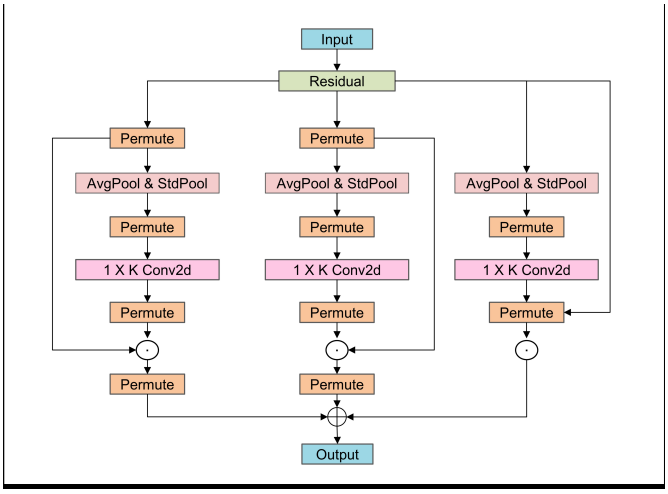


Fig. 7. MCA Architecture

The Multi-scale Contextual Attention (MCA) module is a lightweight architecture designed for visual perception tasks, widely applied in image segmentation, object detection, and similar applications. Its core objective is to enhance the model's ability to understand multi-scale semantic information and local details within images. By jointly modeling multi-scale contextual information and incorporating an attention mechanism for adaptive feature enhancement, the MCA module enables the network to focus more precisely on critical regions within images. This significantly improves the robustness and accuracy of object detection.

Traditional multi-scale feature fusion methods (such as FPN and PANet) integrate feature information across different levels to some extent. However, they often overlook deep semantic correlations between contextual information during feature fusion, leading to potential information redundancy or loss. Furthermore, these methods typically employ static or equal-weight fusion strategies, lacking the ability to dynamically model the contribution of feature maps at different scales. This makes it difficult to flexibly adjust based on target size or image complexity, thereby impacting detection performance.

To overcome the aforementioned issues, the MCA module introduces a multi-scale contextual modeling mechanism during the feature modeling stage, employing two statistical pooling operations: Global Average Pooling (AvgPool) and Global Standard Deviation Pooling (StdPool). The former models the overall background semantic information of the image, while the latter enhances sensitivity to edge textures and local structures.Given an input feature map $F \in \mathbb{R}^{C\times H\times W}$, the computation of the two contextual descriptor vectors is expressed as:

$$F_{avg} = \mathrm{AvgPool}(F), \quad F_{std} = \mathrm{StdPool}(F), \tag{12}$$

where $F_{avg}, F_{std} \in \mathbb{R}^C$ denote the channel-wise average and standard deviation features, respectively. This dual-branch structure provides complementary information at both global and local scales, establishing a context-aware basis for subsequent feature weighting.

After obtaining the contextual features, the MCA module introduces a channel attention mechanism to dynamically adjust the response strength of different features. Specifically, the two contextual vectors are passed through a shared-parameter fully connected subnetwork to generate channel weights, which are then normalized using the Sigmoid activation function:

$$\alpha_{avg} = \sigma(\mathrm{FC}(F_{avg})), \quad \alpha_{std} = \sigma(\mathrm{FC}(F_{std})), \tag{13}$$

Here, $\alpha_{avg}, \alpha_{std} \in \mathbb{R}^C$ are the attention weight vectors derived from the average and standard deviation branches, respectively, and $\sigma(\cdot)$ denotes the Sigmoid function, constraining the attention weights within the range [0, 1]. In this way, MCA effectively emphasizes feature channels that contribute to target regions while suppressing redundant information.

Finally, MCA fuses the attention-weighted signals from the two contextual branches with the original feature map to produce the context-enhanced output feature map:

$$F_{MCA} = (\alpha_{avg} + \alpha_{std}) \otimes F \tag{14}$$

where $\otimes$ denotes element-wise multiplication along the channel dimension. The fused feature map not only preserves the original semantic representation but also introduces prominent responses to critical regions along the channel dimension, enabling the network to focus more effectively on target areas.

In terms of fusion strategy, MCA does not treat all scale features equally but instead adapts feature map importance modeling based on contextual information: for small object detection tasks, the module prioritizes enhancing information from high-resolution feature layers to preserve more detailed textures; whereas in large object recognition scenarios, MCA emphasizes deeper, lower-resolution features with stronger semantic expressiveness. This mechanism significantly enhances the model's flexibility in feature modeling across different-scale objects, contributing to improved overall detection performance.

In summary, by introducing multi-scale context awareness and attention regulation mechanisms, the MCA module enhances the network's ability to synergistically understand both the spatial structure and semantic content of images. This compact and versatile module is particularly well-suited for challenging scenarios in remote sensing imagery, such as dense small objects and complex backgrounds. Integrating MCA into the high-level semantic layer of the backbone network effectively improves the model's object discrimination capability and localization accuracy in complex backgrounds.

4. Experimental Results and Analysis

4.1. Dataset

This experiment utilizes the SAR image vessel detection dataset constructed by the Chinese Academy of Sciences' Institute of Remote Sensing and Digital Earth, along with the

SeaShip dataset, as experimental data sources. SAR-ShipData primarily employs domestically produced GF-3 SAR data and Sentinel-1 SAR data as its main sources. It features diverse vessel slice types and varied backgrounds, making it suitable for multiple SAR image application scenarios. The SeaShip dataset, constructed from multi-source optical remote sensing imagery, encompasses diverse vessel types including container ships, oil tankers, passenger ferries, and fishing vessels. It features rich scale variations and complex background interferences, enabling research on model transferability and generalization performance between optical and SAR scenarios. Regarding data partitioning strategies, both datasets were randomly divided. The SAR-Ship dataset was split into training, validation, and test sets at an 8:1:1 ratio, while the SeaShip dataset was partitioned at a 7:2:1 ratio for model training, parameter tuning, and performance evaluation, as shown in Table 1.

Table 1. Distribution of dataset quantities

	SAR-Ship	SeaShip
Number of train set images	31783	4900
Number of test set images	3974	1400
Number of val set images	3972	700

4.2. Experimental Setup and Parameter Configuration

We conducted experiments on a system running Ubuntu 18.04.5, utilizing PyCharm as the software environment. Table 2 details the configuration information of the experimental platform used for training.

Table 2. Experimental platform configuration information

Item	Value
CPU	Intel(R) Xeon(R) Platinum 8362 @ 2.80GHz
GPU	NVIDIA GeForce RTX 3090
CUDA Version	11.1
Data processing	Python 3.8.10
Deep learning framework	PyTorch 1.8.1

The hyperparameters used during training are as follows: The input image size for all experiments was 640×640 pixels. All other parameters were set to the default values of the YOLOv11n model. To accelerate model convergence, mosaic data augmentation was disabled during the final 10 epochs of training. Detailed parameters of the trained model are shown in Table 3.

Table 3. Detailed parameters of the trained model

Parameters	Value
Epochs	200
Batch size	16
Input image size	640×640
Learning rate	0.01
Momentum	0.937
Weight decay	0.0005

4.3. Evaluation Metrics

To evaluate model performance, the following metrics are used: Mean Average Precision (mAP), target recognition accuracy per category, number of model parameters (Parameters), model floating-point operations (GFLOP), and model size. These metrics assess both model accuracy and efficiency. Here, TP denotes correctly predicted positive samples (correctly classified vessels), while FP denotes correctly predicted negative samples (incorrectly classified vessels). FN denotes correctly predicted false ship samples, while TN denotes incorrectly predicted false ship samples. AP represents the accuracy of target detection. Additionally, the average precision across all classes is denoted as mAP. mAP50 indicates the average of AP50 at an IoU threshold of 50%, where N represents the number of classes. The mAP@50-95 metric calculates the average AP values computed across IoU thresholds from 0.50 to 0.95 (in 0.05 increments). The number of model parameters, the model's floating-point operations, and its size reflect the computational complexity and resource requirements of the model. The formula is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$mAP = \frac{1}{k} \sum_{i=1}^{k} AP_i \tag{18}$$

$$mAP@0.5 = \frac{1}{N} \sum_{t=1}^{N} AP_t, \quad IoU = 0.5 \tag{19}$$

4.4. Training Results and Analysis

To validate the advantages of YOLO-BDM in object detection, we compared the proposed model with YOLOv11n under identical experimental conditions on the SAR-Ship dataset. The results are shown in Table 4.

Table 4. Performance Comparison of YOLO-BDM and YOLOv11n on SAR-Ship

	YOLOv11n	YOLO-BDM
Params	2.60MB	2.90MB
FLOPs	6.4G	7.2G
Precision	93.47%	94.11%
Recall	91.81%	93.07%
F1	92%	94%
mAP@0.5	96.53%	97.27%

From Table 4, it can be seen that YOLO-BDM has a parameter size of 2.90 MB, slightly higher than YOLOv11n's 2.60 MB, indicating that our model incorporates additional parameters to enhance feature extraction capabilities. The computational cost of YOLO-BDM is slightly greater than that of YOLOv11n, but still remains within an acceptable range. In addition, Precision, Recall, and F1 increased by 0.64%, 1.26%, and 2%, respectively, demonstrating that the improved model can reduce false positives and false negatives, achieving a better balance between precision and recall and improving detection accuracy. The mAP@0.5 increased by 0.74%, indicating an overall enhancement in YOLO-BDM's detection performance. To systematically evaluate the vessel recognition capability of YOLO-BDM, we compared the two models using Precision, Recall, and mAP@0.5 metrics on the SAR-Ship dataset through bar charts, and plotted their P-R curves, as shown in Figures 8 and 9.

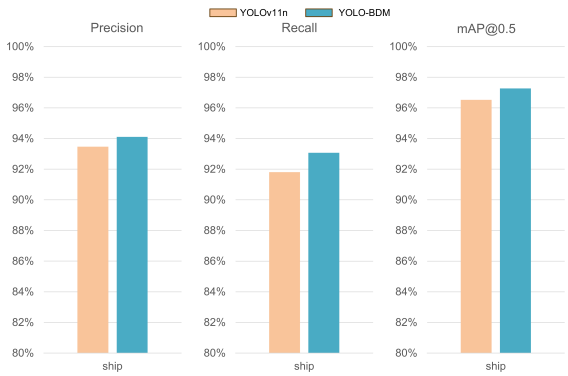
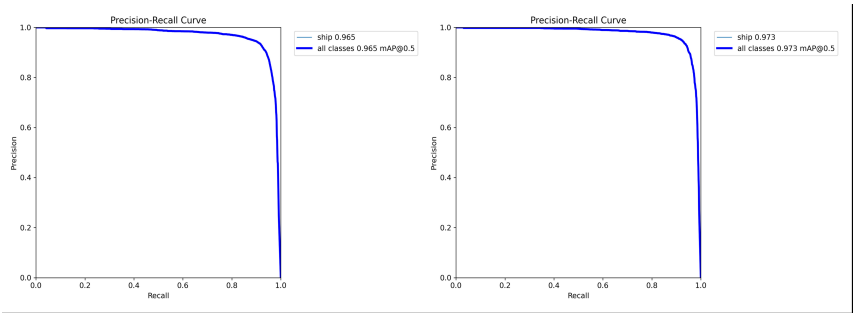


Fig. 8. Test Results for Different Types of Vessels

As shown in Figures 8, YOLO-BDM outperforms the baseline model YOLOv11n across all three metrics, indicating that our model effectively enhances vessel detection performance. The improvement in Precision demonstrates that YOLO-BDM excels at reducing false positives. The increase in Recall indicates that YOLO-BDM can detect

more targets and reduce missed detections, particularly under complex backgrounds or partial occlusions, reflecting stronger model robustness. The improvement in mAP@0.5 signifies that YOLO-BDM achieves superior overall detection performance, indicating a better balance across multiple evaluation metrics.

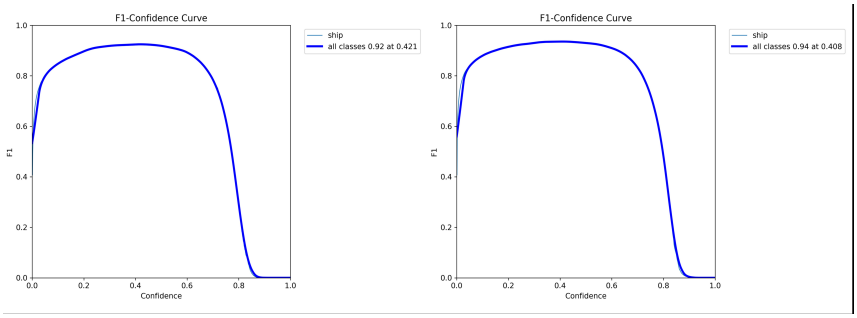


(a) P-R Curve of YOLOv11n Model (b) P-R Curve of YOLO-BDM Model

Fig. 9. Comparison of P-R Curves of YOLOv11n and YOLO-BDM on the SAR-Ship Dataset

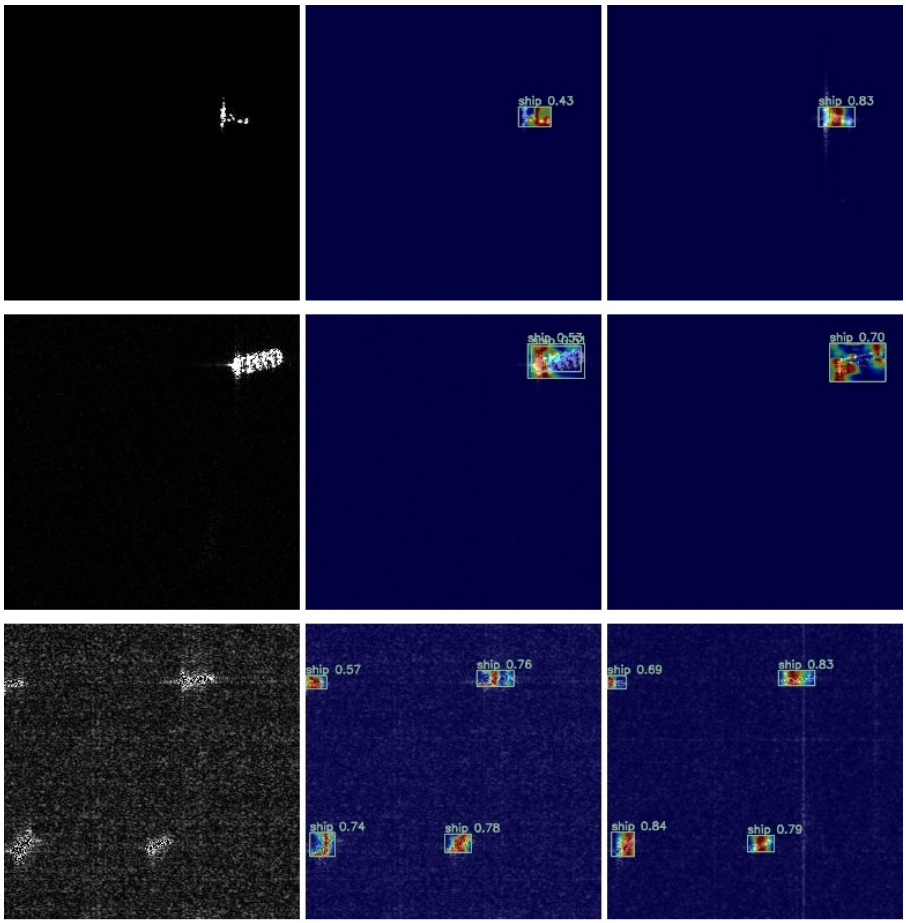
Figures 9 shows the precision-recall curves for YOLOv11n and the improved model YOLO-BDM, used to evaluate the performance of both models in the ship detection task. The closer the precision-recall curve is to the (1,1) corner, the stronger the model's detection capability. YOLOv11n's P-R curve exhibits a noticeable decline in the high Recall region, indicating that the baseline model introduces a significant number of false detections when identifying targets. In contrast, YOLO-BDM's P-R curve approaches the upper-right corner more closely, demonstrating that the YOLO-BDM model maintains high Recall while preserving high Precision, resulting in more stable performance. To compare the comprehensive performance metrics of the models, we plotted the corresponding F1 curves, as shown in Figures 10.

Analysis of the figure above reveals that YOLO-BDM achieves a 2% improvement in F1 score ($0.92 \rightarrow 0.94$) compared to YOLOv11n. This indicates that the enhanced model strikes a better balance between Precision and Recall, resulting in superior overall detection performance. The YOLO-BDM model achieves its highest F1 score at a lower confidence threshold of 0.408, meaning it maintains good detection performance at lower confidence levels and is more sensitive to target vessels. Furthermore, to visually compare the difference in feature attention regions between the two models, this paper presents heatmap visualizations on test samples from the SAR-Ship dataset, as shown in Figures 11. The visualization reveals that YOLOv11n exhibits scattered or misfocused responses in certain complex backgrounds, whereas YOLO-BDM demonstrates more concentrated high-response zones within the ship target areas. This indicates that YOLO-BDM more effectively captures ship feature regions under complex background conditions, reducing interference from irrelevant backgrounds and thereby enhancing detection accuracy and robustness.



(a) F1 Curve of YOLOv11n Model (b) F1 Curve of YOLO-BDM Model

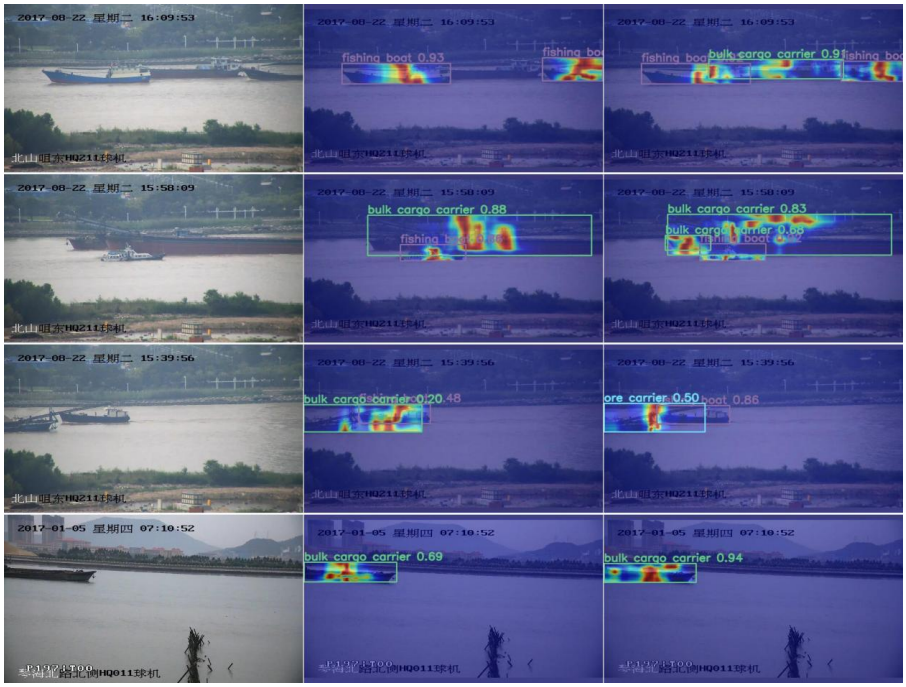
Fig. 10. Comparison of F1 Curves of YOLOv11n and YOLO-BDM on the SAR-Ship Dataset



(a) original image (b) YOLOv11n (c) YOLO-BDM

Fig. 11. Heatmaps of YOLOv11n and YOLO-BDM on the SAR-Ship Dataset

After completing training and performance analysis based on the SAR-Ship dataset, supplementary experiments were conducted using the optical remote sensing ship dataset SeaShip to further validate the generalization capability and robustness of the YOLO-BDM model in cross-modal ship detection tasks. The SeaShip dataset exhibits significant differences from SAR-Ship in imaging mechanisms, spectral characteristics, and background textures, while also encompassing a broader range of vessel types and more complex background interferences.



(a) original image (b) YOLOv11n (c) YOLO-BDM

Fig. 12. Heatmaps of YOLOv11n and YOLO-BDM on the SeaShip Dataset

Figures 12 presents a heatmap comparison between YOLOv11n and YOLO-BDM on the SeaShip dataset. Visual analysis reveals that compared to the baseline model YOLOv11n, the proposed YOLO-BDM model demonstrates improvements in several aspects. First, it exhibits enhanced localization accuracy for small object detection, capturing feature responses of minute targets with greater precision. Second, it demonstrates stronger interference resistance in complex backgrounds, effectively suppressing false detections. Third, it exhibits superior discrimination capabilities for detecting densely arranged targets.

4.5. Ablation Studies

To validate the effectiveness of each module within YOLO-BDM, we conducted a series of ablation experiments. The specific experimental results are shown in Table 5.

Table 5. Comparison of the effectiveness of different modules for vessel detection

Number	BiFPN	DBB	MCA	Precision (%)	Recall (%)	mAP@0.5 (%)	Params (M)	FLOPs (G)
1	×	×	×	93.47	91.81	96.53	2.59	6.4
2	✓	×	×	93.03	92.97	96.70	2.60	6.5
3	×	✓	×	93.74	92.95	97.08	2.62	6.6
4	×	×	✓	93.61	92.74	97.09	2.85	7.1
5	✓	✓	×	94.02	92.83	97.12	2.63	6.6
6	✓	×	✓	93.80	93.00	97.07	2.87	7.1
7	×	✓	✓	93.99	93.01	97.22	2.88	7.2
8	✓	✓	✓	94.11	93.07	97.27	2.90	7.2

In Table 5, the table presents the impact of different improvement modules (BiFPN, DBB, MCA) on vessel detection performance. Model 1 is the baseline model YOLOv11n, where "✓" indicates the module being incorporated.

Model 2 only introduces the DBB module; the Recall increases by 1.16%, but Precision slightly decreases, indicating that while more vessels are detected, the number of false positives also rises. This shows that DBB mainly improves feature extraction capability, helping detect more vessels.Model 3 only incorporates the MCA module, resulting in Precision rising to 93.74%, Recall increasing by 1.14%, and mAP@0.5 improving by 0.55%. These data demonstrate that MCA effectively enhances fine-grained feature representation and contextual modeling, improving vessel detection under complex backgrounds.Model 4 only introduces BiFPN, primarily strengthening multi-scale feature fusion and enhancing target discrimination, but it may cause some difficult-to-detect targets to be missed, so the Recall shows little change. This indicates that BiFPN mainly contributes to multi-scale feature aggregation, improving accuracy for detectable targets.

Models 5, 6, and 7 correspond to the pairwise combinations of the three modules. The combination of DBB and BiFPN increases both parameter size and computational cost, but the performance improvement is less than expected, indicating potential redundancy in the fused features. The combination of MCA and BiFPN not only enhances multi-scale feature fusion but also improves detection accuracy. The combination of DBB and MCA results in a significant increase in Precision, but Recall decreases, possibly due to stricter target selection caused by feature enhancement.

Model 8, which integrates all three modules, achieves the best results across all metrics, demonstrating that their combination can effectively enhance the accuracy and robustness of vessel detection. Among them, the MCA mechanism is the key factor in improving detection performance, particularly evident in the increases in Recall and overall mAP. The BiFPN module primarily strengthens multi-scale feature fusion, while the DBB module mainly enhances the model's feature extraction capability for vessels.

4.6. Comparative Experiments

To comprehensively evaluate the performance of YOLO-BDM, we conducted comparative experiments with several lightweight YOLO models (YOLOv4-tiny [21], YOLOv5s, YOLOv7-tiny [34], YOLOv8n [39], YOLOv10n [33], and YOLOv11n [22]). Analysis was conducted across four metrics: Precision, Recall, mAP@0.5, and Parameters. The results are as follows:

Table 6. Performance and Parameters of Different Models on SAR-SHIP Dataset

Model	**Precision (%)**	**Recall (%)**	**mAP@0.5 (%)**	**Params** ($\times 10^6$)
YOLOv4-tiny	87.41	90.28	86.01	5.92
YOLOv5s	91.12	88.15	88.77	7.19
YOLOv7-tiny	91.61	91.97	92.44	6.30
YOLOv8n	93.74	93.75	96.23	3.01
YOLOv10n	93.12	93.42	96.46	2.71
YOLOv11n	93.47	91.81	96.53	2.59
YOLO-BDM	94.11	93.07	97.27	2.90

As shown in Table 6 on the SAR-SHIP dataset, YOLO-BDM demonstrates superior performance across all metrics, particularly achieving the highest mAP@0.5 among all models. Compared with YOLOv4-tiny, YOLO-BDM improves precision by 6.7%, recall by 2.79%, and mAP@0.5 by 11.26%, indicating significant advantages in detection accuracy and stability. Compared with YOLOv8n and YOLOv10n, YOLO-BDM achieves improvements of 1.04% and 0.81% in mAP@0.5, respectively, demonstrating stronger overall detection capability.

YOLO-BDM has a parameter size of 2.90M, substantially smaller than YOLOv4-tiny, YOLOv5s, and YOLOv7-tiny, indicating a more lightweight model suitable for resource-constrained scenarios. Although its parameter count is slightly higher than YOLOv8n, YOLOv10n, and YOLOv11n, the gains in mAP@0.5 and recall justify the additional computational cost.

Table 7. Performance and Parameters of Different Models on SAR-SHIP Dataset

Model	**Precision (%)**	**Recall (%)**	**mAP@0.5 (%)**	**Params** ($\times 10^6$)
YOLOv4-tiny	94.36	91.61	95.87	5.92
YOLOv5s	97.69	97.73	97.53	7.19
YOLOv7-tiny	98.12	97.46	98.23	6.30
YOLOv8n	98.21	97.29	98.76	3.01
YOLOv10n	98.37	95.59	98.59	2.71
YOLOv11n	98.70	97.94	98.81	2.59
YOLO-BDM	99.39	98.06	98.97	2.90

To further validate the generalization capability of the model across different scenarios, this study compared the detection performance of several mainstream models on the SeaShip dataset (see Table 7). The results indicate that YOLO-BDM demonstrates excellent performance across all metrics. Its Precision reaches 99.39%, representing an increase of 5.03% over YOLOv4-tiny and improvements of 1.18% and 0.69% compared to YOLOv8n and YOLOv11n, respectively. In terms of mAP@0.5, YOLO-BDM achieves 98.97%, which is 3.10% higher than YOLOv4-tiny and shows gains of 0.21% and 0.38% over YOLOv8n and YOLOv10n, respectively. Furthermore, YOLO-BDM attains a Recall of 98.06%, exceeding YOLOv10n by 2.47%, reflecting stronger object detection capability. These results indicate that YOLO-BDM not only maintains leading overall accuracy but also exhibits clear advantages in detection stability and recall.

Overall, compared to YOLOv11n, YOLO-BDM achieves superior detection accuracy with only a slight increase in parameter count, demonstrating an effective balance between lightweight design and high-performance detection. In SAR ship detection tasks, it offers higher Precision, Recall, and mAP@0.5 while maintaining relatively low computational overhead.

5. Conclusion

This paper proposes YOLO-BDM, an enhanced SAR vessel detection model based on the YOLOv11 framework. To address typical challenges such as small target detection, blurred target contours, and complex ocean background interference, three key improvement modules are designed: DBB (Diversified Branch Block) enriches feature extraction and multi-scale modeling, MCA (Multi-scale Contextual Attention) enhances context-guided target attention capabilities, and BiFPN_Concat optimizes multi-scale feature fusion. Experimental results demonstrate that YOLO-BDM significantly outperforms existing lightweight YOLO models in core metrics including Precision, Recall, and mAP@0.5, while maintaining a low parameter count. This achieves a favorable balance between detection accuracy and computational efficiency, exhibiting strong robustness and adaptability particularly in detecting small, blurry-contoured vessels within complex backgrounds.

Although YOLO-BDM demonstrates strong performance in SAR vessel detection tasks, several areas warrant further investigation. Future work can be pursued in the fol-

lowing areas: Further optimizing the model architecture to reduce computational complexity, thereby adapting to resource-constrained devices such as drones and satellites for real-time detection requirements. Expanding the dataset scale by incorporating samples under more complex sea conditions (e.g., wave and wind effects, occlusions, camouflaged targets) to enhance the model's generalization capabilities.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
2. Cao, R., Sui, J.: A dynamic multi-scale feature fusion network for enhanced sar ship detection. Sensors 25(16), 5194 (2025)
3. Chen, C., Han, D., Chang, C.C.: Caan: Context-aware attention network for visual question answering. Pattern Recognition 132, 108980 (2022)
4. Chen, C., Han, D., Chang, C.C.: Mpcct: Multimodal vision-language learning paradigm with context-based compact transformer. Pattern recognition 147, 110084 (2024)
5. Chen, C., Han, D., Guo, Z., Chang, C.C.: Towards bias-aware visual question answering: Rectifying and mitigating comprehension biases. Expert Systems with Applications 264, 125817 (2025)
6. Chen, C., Han, D., Shen, X.: Clvin: Complete language-vision interaction network for visual question answering. Knowledge-Based Systems 275, 110706 (2023)
7. Chen, Z., Ding, Z., Zhang, X., Wang, X., Zhou, Y.: Inshore ship detection based on multimodality saliency for synthetic aperture radar images. Remote Sensing 15(15), 3868 (2023)
8. Cui, J., Jia, H., Wang, H., Xu, F.: A fast threshold neural network for ship detection in large-scene sar images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 6016–6032 (2022)
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
10. Duan, R., Deng, H., Tian, M., Deng, Y., Lin, J.: Soda: A large-scale open site object detection dataset for deep learning in construction. Automation in Construction 142, 104499 (2022)
11. Feng, Y., Zhang, Y., Zhang, X., Wang, Y., Mei, S.: Large convolution kernel network with edge self-attention for oriented sar ship deteciton. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024)
12. Fu, X., Zhou, Z., Meng, H., Li, S.: A synthetic aperture radar small ship detector based on transformers and multi-dimensional parallel feature extraction. Engineering Applications of Artificial Intelligence 137, 109049 (2024)
13. Guo, H., Bai, H., Yuan, Y., Qin, W.: Fully deformable convolutional network for ship detection in remote sensing imagery. Remote Sensing 14(8), 1850 (2022)
14. Han, D., Shi, J., Zhao, J., Wu, H., Zhou, Y., Li, L.H., Khan, M.K., Li, K.C.: Lrcn: Layer-residual co-attention networks for visual question answering. Expert Systems with Applications 263, 125658 (2025)
15. Han, D., Zhu, Y., Li, D., Liang, W., Souri, A., Li, K.C.: A blockchain-based auditable access control system for private data in service-centric iot environments. IEEE Transactions on Industrial Informatics 18(5), 3530–3540 (2021)
16. He, J., Su, N., Xu, C., Liao, Y., Yan, Y., Zhao, C., Hou, W., Feng, S.: A cross-modality feature transfer method for target detection in sar images. IEEE Transactions on Geoscience and Remote Sensing 61, 1–15 (2023)

17. Hu, B., Liu, Y., Chu, P., Tong, M., Kong, Q.: Small object detection via pixel level balancing with applications to blood cell detection. Frontiers in Physiology 13, 911297 (2022)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
19. Hu, Q., Hu, S., Liu, S.: Banet: A balance attention network for anchor-free ship detection in sar images. IEEE Transactions on Geoscience and Remote Sensing 60, 1–12 (2022)
20. Ji, P., Xing, S., Dai, D., Pang, B.: Deceptive targets generation simulation against multichannel sar. Electronics 9(4), 597 (2020)
21. Jiang, Z., Zhao, L., Li, S., Jia, Y.: Real-time object detection method based on improved yolov4-tiny. arXiv preprint arXiv:2011.04244 (2020)
22. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
23. Li, Z., Chen, J., Xiong, Y., Yu, H., Zhang, H., Gao, B.: A ship detection and imagery scheme for airborne single-channel sar in coastal regions. Remote Sensing 14(18), 4670 (2022)
24. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
25. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
26. Ma, X., Hou, S., Wang, Y., Wang, J., Wang, H.: Multiscale and dense ship detection in sar images based on key-point estimation and attention mechanism. IEEE Transactions on Geoscience and Remote Sensing 60, 1–11 (2022)
27. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra r-cnn: Towards balanced learning for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 821–830 (2019)
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
29. Shen, X., Chen, C., Han, D., Xu, Y., Wang, X., Zhou, H.: A triple-branch hybrid dynamic-static alignment strategy for vision-language tasks. Neural Networks p. 107871 (2025)
30. Shen, X., Han, D., Chang, C.C., Oad, A., Wu, H.: Gfsnet: Gaussian fourier with sparse attention network for visual question answering. Artificial Intelligence Review 58(6), 1–30 (2025)
31. Shen, X., Han, D., Chang, C.C., Xu, Y., Chen, C.: Multimodal context-aware consistency alignment for vision-language tasks. Expert Systems with Applications 295, 128857 (2026)
32. Tang, X., Zhang, J., Xia, Y., Cao, K., Zhang, C.: Pegnet: An enhanced ship detection model for dense scenes and multi-scale targets. IEEE Geoscience and Remote Sensing Letters (2025)
33. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems 37, 107984–108011 (2024)
34. 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 IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023)
35. Wang, X., Wang, A., Yi, J., Song, Y., Chehri, A.: Small object detection based on deep learning for remote sensing: A comprehensive review. Remote Sensing 15(13), 3265 (2023)
36. Wang, X., Li, G., Zhang, X.P., He, Y.: A fast cfar algorithm based on density-censoring operation for ship detection in sar images. IEEE Signal Processing Letters 28, 1085–1089 (2021)
37. Xu, H.Y., Xu, F., Jin, Y.Q.: Optimal sensing principle of synthetic aperture radar. IEEE Transactions on Geoscience and Remote Sensing 62, 1–14 (2023)
38. Xu, X., Zhao, J., Li, Y., Gao, H., Wang, X.: Banet: A balanced atrous net improved from ssd for autonomous driving in smart transportation. IEEE Sensors Journal 21(22), 25018–25026 (2020)

39. Yaseen, M.: What is yolov8: An in-depth exploration of the internal features of the next-generation object detector (2024)
40. 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17, 7242–7259 (2024)
41. Zhang, M., Ouyang, Y., Yang, M., Guo, J., Li, Y.: Orpsd: Outer rectangular projection-based representation for oriented ship detection in sar images. Remote Sensing 17(9), 1511 (2025)
42. Zhang, X., Yan, M., Zhu, D., Guan, Y.: Marine ship detection and classification based on yolov5 model. In: Journal of Physics: Conference Series. vol. 2181, p. 012025. IOP Publishing (2022)
43. Zhao, W., Syafrudin, M., Fitriyani, N.L.: Cras-yolo: A novel multi-category vessel detection and classification model based on yolov5s algorithm. IEEE Access 11, 11463–11478 (2023)
44. Zhou, S., Zhang, M., Wu, L., Yu, D., Li, J., Fan, F., Liu, Y., Zhang, L.: Sar ship detection network based on global context and multi-scale feature enhancement. Signal, Image and Video Processing 18(3), 2951–2964 (2024)
45. Zhou, Z., Cui, Z., Zang, Z., Meng, X., Cao, Z., Yang, J.: Ultrahi-prnet: An ultra-high precision deep learning network for dense multi-scale target detection in sar images. Remote Sensing 14(21), 5596 (2022)
46. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE 111(3), 257–276 (2023)

Fangyuan Xiong is currently pursuing her postgraduate degree at the College of Computer Science and Information Engineering, Shanghai Maritime University. Her research focuses on computer vision and intelligent perception, with particular emphasis on lightweight object detection algorithms, ship detection and tracking, and the application of deep learning in marine monitoring. Her recent work primarily centers on improving the YOLO series of models.

Dezhi Han (Senior Member, IEEE) received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China,in 2005.He is currently a Professor of Computer Science and Engineering with Shanghai Mar itime University, Shanghai, China. His research interests include cloud computing, mobile net working, wireless communication, and cloud security.

Xiang Shen is currently pursuing his Ph.D. at Shanghai Maritime University and is also a joint Ph.D. student at The University of Sydney, supported by the China Scholarship Council (CSC). His research focuses on multimodal learning, with particular emphasis on visual question answering, multimodal fusion strategies, and the development of large-scale multimodal models. His current work also explores the integration of AI agents for adaptive perception and reasoning in complex environments, aiming to bridge the gap between visual understanding and natural language processing.

Manlin Zhu is currently pursuing his master's degree at Shanghai Maritime University. His research focuses on computer vision and ship detection, with particular emphasis on image processing, object detection, and perception techniques in marine environments. He is committed to applying deep learning to marine monitoring scenarios.

Received: October 18, 2025; Accepted: January 5, 2026.