DE-MFM-YOLO: A YOLO-Based Network with Detail-Enhanced Convolution and Modulated Feature Fusion for Multi-Scale Ship Detection in SAR Images

Yukai Zhang, Huijuan Zhao, Zhixin Wu
College of Information Technology, Shanghai Ocean University, Shanghai 201306, China
Corresponding author: Huijuan Zhao (e-mail: 15261081166@163.com).

ABSTRACT Synthetic Aperture Radar (SAR) plays an important role in maritime ship detection due to its all-weather and day-and-night imaging capability. However, reliable detection remains challenging because of strong speckle noise, complex sea clutter, and the weak representation of small ship targets. To address these issues, this paper proposes DE-MFM-YOLO, an efficient SAR ship detection framework based on YOLOv11. The proposed framework enhances local edge and texture representation through Detail-Enhanced Convolution (DEConv) and introduces a Modulation Fusion Module (MFM) to adaptively fuse multi-scale features and improve semantic interaction across different feature levels. Experimental results on the HRSID dataset demonstrate that DE-MFM-YOLO achieves an AP50 of 93.0% while maintaining real-time inference capability. The complementary integration of DEConv and MFM effectively improves feature representation, multi-scale feature aggregation, and the detection of small and densely distributed ship targets. These results indicate that the proposed method provides an effective balance between detection accuracy and computational efficiency for practical SAR maritime monitoring applications.

INDEX TERMS SAR ship detection; small targets; multi-scale feature fusion; YOLOv11

I. INTRODUCTION

Ship detection in Synthetic Aperture Radar (SAR) imagery has become an essential component of sustainable ocean governance and maritime monitoring systems, supporting applications such as maritime traffic management, search-and-rescue operations, maritime trade supervision, and maritime security enforcement [1][2][3][4]. Driven by the continuous expansion of global maritime activities, the demand for accurate and timely satellite-based ship detection has significantly increased in recent years. In this context, remote sensing technologies have been widely explored for large-scale maritime observation tasks. Among them, Synthetic Aperture Radar (SAR) has received growing attention due to its all-weather and day-and-night imaging capability, as well as its robustness against clouds, fog, and illumination variations, making it particularly suitable for reliable ship detection under diverse environmental conditions [5][6][7][8][9]. In addition, SAR provides rich backscattering information from sea surface targets, which is particularly beneficial for ship detection in low-visibility and high-clutter scenarios compared with optical imagery.

Traditional SAR ship detection methods mainly relied on statistical modeling and threshold-based decision strategies, among which the Constant False Alarm Rate (CFAR) algorithm has been extensively investigated. Li et al. [10] proposed a two-stage superpixel CFAR method based on a truncated kernel density estimation model to improve target detection accuracy by enhancing clutter modeling. Sun and Wang [11] further combined a steady CFAR detector with a knowledge-oriented GBDT classifier, introducing a post-classification strategy to reduce false alarms. To improve detection efficiency, Zhang et al. [12] developed a fast superpixel-based non-window CFAR detector that eliminates the conventional sliding-window mechanism while preserving detection performance. In addition, Li et al. [13] incorporated intensity and texture feature fusion with an attention contrast mechanism into the CFAR framework, enhancing the adaptability of threshold estimation under complex sea clutter. Despite these improvements, such methods remain fundamentally dependent on predefined statistical assumptions and manually designed thresholding strategies, making their performance highly sensitive to variations in sea clutter characteristics and target scales. Consequently, they often exhibit limited robustness in challenging scenarios involving heterogeneous backgrounds, coastal interference, and densely distributed ships, leading to increased false alarms and missed detections. To address these limitations, Zhao et al. [64] introduced a denoising-enhanced YOLO framework that combines statistical prior information with deep feature learning to improve detection robustness. Nevertheless, this category of methods still suffers from performance degradation under highly heterogeneous sea clutter conditions, indicating that robust feature representation remains a critical challenge for SAR ship detection.

With the rapid development of deep learning, numerous convolutional neural network (CNN)-based approaches have driven significant progress in SAR ship detection by enabling data-driven feature learning. Benefiting from their powerful nonlinear representation capability and hierarchical feature extraction, CNN-based methods have demonstrated clear advantages over traditional statistical approaches in terms of feature discrimination, detection accuracy, and robustness under complex maritime environments [42]. Representative deep learning detectors have further improved the capability of identifying ships with diverse scales and complex backgrounds, promoting the rapid development of SAR ship detection toward more accurate and intelligent solutions [14]. Building upon the success of CNNs, Transformer-based architectures have recently attracted increasing attention in object detection owing to their capability of modeling long-range dependencies through self-attention mechanisms and capturing richer global contextual information [20]. A representative framework is DETR [21], which adopts an end-to-end encoder – decoder architecture for joint object localization and classification, reducing the reliance on handcrafted components in conventional detection pipelines. More recently, Transformer-based methods have been extended to SAR ship detection. Yang et al. [65] proposed LPST-Det, which integrates a local-perception-enhanced Swin Transformer backbone with a cross-scale bidirectional feature pyramid network to strengthen multi-scale feature representation and contextual information modeling, thereby improving the detection performance of small and densely distributed ships in complex maritime environments. These advances demonstrate the effectiveness of Transformer-based architectures in further enhancing SAR ship detection performance.

Despite these advances, SAR ship detection still faces several challenges in practical applications. Owing to the coherent imaging mechanism of SAR, complex background clutter and speckle noise inevitably degrade image quality, particularly in coastal and port regions where waves, shore infrastructure, and surrounding objects often exhibit scattering characteristics similar to those of ships, resulting in increased false detections and reduced detection robustness [23]. Furthermore, large scale variation remains another critical challenge [22]. Ship targets in SAR imagery range from small fishing boats to large cargo vessels, while the inherent imaging characteristics of SAR often cause small ships to occupy only a few pixels with weak scattering responses. Consequently, fixed-receptive-field networks struggle to simultaneously capture discriminative features across different scales. Since small targets account for a considerable proportion of SAR ship datasets [24], their weak feature responses are more susceptible to background interference and speckle noise, further increasing the difficulty of accurate detection.

Motivated by these challenges, numerous representative strategies have been proposed to improve SAR ship detection performance. To improve the representation of ships with significant scale variation, multi-scale feature modeling techniques, such as Feature Pyramid Networks (FPN), have been widely adopted to enhance feature interaction across different resolutions [25]. To alleviate the influence of speckle noise and complex background clutter inherent in SAR imagery, attention mechanisms, including channel attention [26] and spatial attention or Transformer-inspired modules [27], have been introduced to emphasize discriminative ship features while suppressing irrelevant background responses. Considering the increasing demand for real-time maritime surveillance, lightweight network designs have also been developed to reduce computational complexity and facilitate deployment on resource-constrained platforms without sacrificing detection performance [28]. In addition, optimization strategies, such as IoU-based loss functions [29,47] and balanced label assignment methods [48 – 50], have been proposed to improve localization accuracy and training stability, particularly for densely distributed and small ship targets. Although these strategies have improved specific aspects of SAR ship detection, most existing methods still concentrate on optimizing individual components, such as feature extraction, feature fusion, or context modeling, rather than jointly exploiting their complementary strengths within a unified lightweight architecture. Consequently, achieving a desirable balance among detection accuracy, computational efficiency, and deployment capability remains a challenging task. Therefore, developing a lightweight detection framework that simultaneously enhances local detail representation, multi-scale feature learning, and small-target perception remains an important research direction for practical SAR maritime surveillance applications.

From this perspective, we develop DE-MFM-YOLO, an enhanced lightweight SAR ship detection framework based on YOLOv11 [30]. The proposed framework is designed to simultaneously improve feature representation and multi-scale information interaction while preserving the inherent efficiency of one-stage detectors. Unlike existing approaches that mainly enhance either feature extraction or feature fusion independently, DE-MFM-YOLO establishes a collaborative optimization mechanism between the backbone and neck, enabling complementary learning of local structural details and high-level semantic information for more robust SAR ship detection. Specifically, a Detail-

Enhanced Convolution (DEConv) module is embedded into the backbone to reinforce local edge, texture, and structural representations through differential convolution operations, thereby enhancing the discriminative capability of shallow features under speckle noise, scattering effects, and complex background clutter. This enables the network to better preserve weak responses and fine-grained characteristics of small ship targets during the early feature extraction stage. Furthermore, a Modulation Fusion Module (MFM) is incorporated into the neck to adaptively recalibrate and aggregate multi-scale semantic features through learnable feature modulation, promoting more effective information interaction across different feature levels while reducing redundant feature propagation. The adaptive fusion mechanism enables the network to more effectively exploit complementary information from low-and high-level features, thereby improving the localization and recognition of ships with significant scale variations and dense spatial distributions. By jointly optimizing detail-aware feature extraction and adaptive multi-scale feature fusion, the proposed framework establishes stronger collaboration between local structural representation and global semantic perception, resulting in improved robustness against background interference and enhanced detection performance under complex maritime environments. In addition, both DEConv and MFM are designed as lightweight plug-and-play modules that can be seamlessly integrated into the YOLOv11 architecture with only a modest increase in computational overhead, making the proposed framework suitable for real-time deployment in practical SAR maritime monitoring systems. Extensive experiments conducted on the HRSID dataset demonstrate that DE-MFM-YOLO consistently outperforms the YOLOv11 baseline and several representative lightweight detection frameworks while maintaining real-time inference capability. The main contributions of this work are summarized as follows:

(1) We develop a Detail-Enhanced Convolution (DEConv) module to strengthen edge, texture, and structural feature representation, thereby improving the detection of small ships and weak-boundary targets in complex SAR imagery.

(2) We design a Modulation Fusion Module (MFM) to adaptively aggregate multi-scale feature responses, enhancing the network's ability to integrate complementary semantic information. Ablation studies demonstrate that MFM effectively improves feature fusion, leading to more accurate localization and classification.

(3) We propose DE-MFM-YOLO, a lightweight SAR ship detection framework based on YOLOv11. The proposed method achieves an AP50 of 93.0% on the HRSID dataset while maintaining real-time inference capability. Experimental results demonstrate that DEConv and MFM complement each other in improving detail representation and multi-scale feature modeling, achieving a better balance between detection accuracy and computational efficiency.

II. METHODS

A. EVALUATION METRICS

Precision, Recall, AP50, and AP are adopted in this study to comprehensively evaluate the performance of SAR ship detection models.

In object detection tasks, Intersection over Union (IoU) is commonly used to measure the overlap between a predicted bounding box and its corresponding ground-truth bounding box. It is defined as the ratio between the intersection area and the union area of the two bounding boxes and serves as a fundamental criterion for evaluating localization accuracy. The calculation of IoU is expressed as follows:

$$\mathrm{IoU} = \frac{\text{Intersection Area}}{\text{Union Area}} \tag{1}$$

In ship detection tasks, prediction results can be categorized into four groups: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Based on these quantities, Precision and Recall can be defined as shown in Equations (2) and (3), respectively:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

Average Precision (AP) is calculated as the area under the Precision – Recall (PR) curve, which summarizes the relationship between Precision and Recall across different confidence thresholds. The corresponding formulation is given in Equation (4):

$$AP = \int_0^1 P(R)\, dR \tag{4}$$

The MS COCO evaluation protocol provides a comprehensive set of detection metrics [32]. In this study, AP50 and AP are employed as the primary evaluation indicators. AP50 denotes the Average Precision obtained when the IoU threshold is fixed at 0.5. In contrast, AP represents the mean Average Precision computed over multiple IoU thresholds ranging from 0.50 to 0.95 with a step size of 0.05. This metric provides a more comprehensive assessment of the localization and classification performance of object detection models.

B. OVERALL NETWORK ARCHITECTURE

DE-MFM-YOLO is a lightweight SAR ship detection framework built upon YOLOv11 [30], aiming to enhance detail representation and multi-scale feature interaction while preserving high detection efficiency. By integrating the Detail-Enhanced Convolution (DEConv) module and the Modulation Fusion Module (MFM), the framework improves the representation of small and densely distributed ship targets in complex SAR imagery with only a modest increase in computational overhead.

As illustrated in Fig.1, the proposed DE-MFM-YOLO architecture consists of three main components: (1) a Backbone network, where DEConv modules are incorporated to strengthen edge, texture, and structural feature extraction, thereby improving the representation of fine-grained details in SAR imagery; (2) a Neck network, in which the Modulation Fusion Module (MFM) adaptively recalibrates and fuses multi-scale features from different semantic levels to enhance feature interaction and contextual representation; and (3) a Head network, which retains the original three-scale prediction strategy for object classification and bounding-box regression across targets of different sizes. Fig. 1 illustrates the complete architecture of the proposed DE-MFM-YOLO framework, including the end-to-end processing pipeline and the locations of the DEConv and MFM modules within the backbone and neck networks.

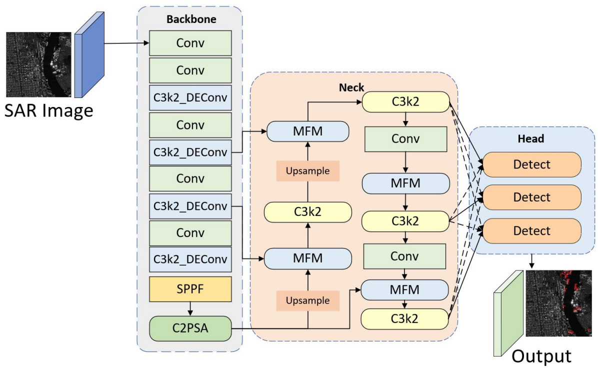

FIGURE 1. DE-MFM-YOLO's overall network architecture.

1) Backbone: Introduction of Detail-Enhanced Convolution (DEConv)

To better preserve local structural information during feature extraction, the Detail-Enhanced Convolution (DEConv) module [33] is incorporated into selected C3k2 blocks of the backbone. As illustrated in Fig. 2, the overall architecture of the DE module and the structural diagram of the proposed DEConv module are presented. By introducing differential convolution operations, DEConv enhances the extraction of edge, texture, and directional features, strengthening local detail representation during the early stages of feature learning. These enhanced representations provide a more informative feature foundation for subsequent multi-scale feature fusion and target localization.

Specifically, DEConv decomposes a conventional convolution into five parallel branches, including vanilla convolution (VC), center-difference convolution (CDC), angular-difference convolution (ADC), horizontal-difference convolution (HDC), and vertical-difference convolution (VDC). Unlike conventional convolution, the four difference-convolution branches capture complementary local gradient variations from different directions, enabling the network to learn richer edge, texture, and structural representations while the vanilla convolution preserves regional semantic information. By jointly exploiting these complementary representations, DEConv enhances local detail perception and strengthens feature representation during early-stage feature extraction. Furthermore, a structural re-parameterization strategy is employed to merge the multi-branch architecture into an equivalent convolution layer during inference, thereby maintaining an efficient inference structure.

The DEConv module is particularly suitable for SAR ship detection scenarios involving complex textures and strong background noise. By enhancing edge and texture responses, the module provides more discriminative local features for subsequent multi-scale feature fusion and target detection.

Given an input feature map F_{in}, DEConv generates the output feature map F_{out}. Through structural re-parameterization, the resulting equivalent convolution maintains the same computational complexity and inference efficiency as a standard convolution layer. The formulation is expressed as:

$$
\begin{aligned}
F_{\text{out}} &= \text{DEConv}(F_{\text{in}}) = \sum_{i=1}^{5} F_{\text{in}} * K_i \\
&= F_{\text{in}} * \left(\sum_{i=1}^{5} K_i\right) = F_{\text{in}} * K_{\text{cvt}}
\end{aligned}
\tag{5}
$$

Where DEConv $(\cdot)$ denotes the proposed detail-enhanced convolution operation, $K_{i=1:5}$ represent the convolution kernels of VC, CDC, ADC, HDC, and VDC, respectively, and $*$ denotes the convolution operation.

K_{cvt} represents the equivalent convolution kernel obtained after merging the parallel branches through re-parameterization.

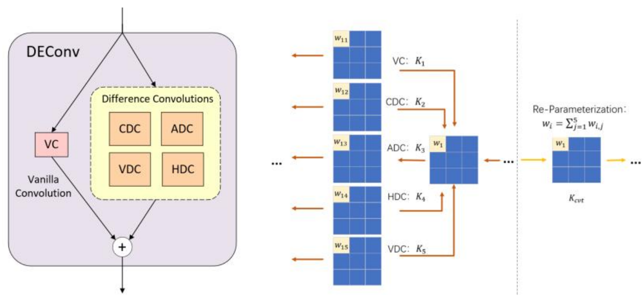

FIGURE 2. Architecture of the DE module and schematic illustration of the DEConv mechanism [33].

2) Neck: Modulated Feature Fusion Module (MFM)

Building upon the enhanced feature representations produced by the backbone network, the neck network further performs multi-scale feature aggregation to refine semantic information.

Multi-scale feature fusion in conventional FPN- and PANet-based architectures is typically implemented through top-down and bottom-up information propagation [32,49,57–59]. However, in SAR ship detection tasks, small targets often occupy only a limited number of pixels within shallow feature maps. During feature fusion, their weak responses may be overwhelmed by the strong activations of deep semantic features. This issue becomes more pronounced in complex maritime environments, where dominant high-level semantic information may suppress fine-grained structural details, thereby weakening the representation of small target boundaries and local features [3,8,23,57 – 59]. To alleviate this limitation, the Modulation Fusion Module (MFM) [34] is incorporated into the neck network to replace part of the conventional feature fusion structure, enabling more effective interaction and adaptive aggregation of multi-scale features.

As shown in Fig. 3, the proposed MFM performs adaptive multi-scale feature fusion through learnable modulation factors along both channel and scale dimensions. By dynamically adjusting feature responses, the module emphasizes informative feature channels while improving the integration of complementary information from different feature levels. Specifically, multi-scale feature maps are first processed through convolution layers to unify channel dimensions. Then, Global Average Pooling (GAP) and a shared Multi-Layer Perceptron (MLP) are applied to extract global contextual information and generate channel-wise feature importance descriptors, which are normalized via a Softmax function to produce the channel-weight matrix.

The generated weights are then employed to adaptively modulate feature responses before and after feature fusion, enhancing informative representations while suppressing less relevant information. Through this dynamic modulation mechanism, MFM effectively balances low-level structural details and high-level semantic information, helping to preserve fine-grained edge and texture features of small targets during feature aggregation. At the same time, the influence of dominant large-scale target responses is moderated, allowing the network to maintain robust target representations across different scales under complex sea-surface conditions.

Specifically, the module takes feature maps from different paths as inputs, including the feature map ($\widehat{F}^1_{\text{legm}}$) obtained after local–global feature fusion in the encoder and the shallow feature map (F^1_{rc}) processed by a 3 × 3 convolution. To estimate feature importance, GAP is first applied to extract global contextual information while preserving channel-wise statistical characteristics. The resulting feature descriptors are subsequently processed by a MLP, followed by a Softmax normalization operation to generate the weight matrix ($A^1_{r,c}$). The generated weight matrix explicitly characterizes the relative importance of individual feature channels and serves as an adaptive modulation factor during feature fusion, enabling the network to emphasize informative semantic responses while suppressing redundant or less discriminative features. Consequently, complementary information from different feature levels can be selectively aggregated according to their contributions, resulting in more discriminative multi-scale feature representations and improved robustness under complex SAR scenes.

Based on the generated weights, the input feature maps are adaptively reweighted through element-wise multiplication ($\odot$). This operation enhances informative features while suppressing less relevant responses, resulting in the following fused representation:

$$\widehat{F}^1_{rc} = A^1_{r,c} \odot \widehat{F}^1_{\text{legm}} + A^1_{r,c} \odot F^1_{rc} \tag{6}$$

Subsequently, feature concatenation and convolution operations are employed to further integrate multi-channel information and capture cross-channel contextual dependencies, producing the final fused feature representation.

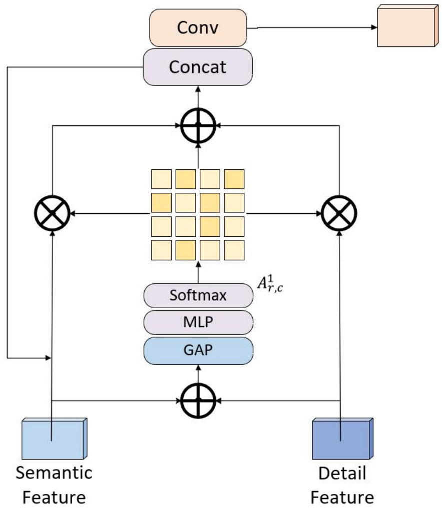

FIGURE 3. Schematic diagram of the MFM architecture [34].

3) Loss Function

The overall loss function of the proposed model consists of three components: confidence estimation loss, bounding box regression loss, and Distribution Focal Loss (DFL).

Given the complex maritime backgrounds, weak target responses, and substantial noise interference inherent in SAR imagery, this study adopts a composite loss function within the YOLOv11 detection framework to jointly optimize target discrimination and localization accuracy. The overall objective consists of confidence loss [51], bounding box regression loss [52], and Distribution Focal Loss (DFL) [53], which are jointly optimized to improve classification confidence, localization precision, and detection robustness under challenging maritime environments.

The confidence loss is implemented using the Binary Cross-Entropy (BCE) loss function [54], which distinguishes foreground targets from background regions by measuring the discrepancy between predicted confidence scores and ground-truth labels. As a fundamental supervision signal, it enhances the network's ability to identify ship targets in complex SAR scenes characterized by background clutter and noise interference. The formulation is given as:

$$L_{\mathrm{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \quad (7)$$

Furthermore, the bounding box regression loss adopts the Complete Intersection over Union (CIoU) loss [55] to introduce additional geometric constraints for more accurate spatial localization. Compared with conventional IoU-based losses, CIoU incorporates complementary geometric factors, including the distance between bounding-box centers and aspect-ratio consistency, thereby providing more informative optimization gradients. By jointly optimizing these geometric relationships, CIoU improves bounding-box alignment and localization precision, which is particularly beneficial for SAR imagery containing significant scale variations and a large proportion of small ship targets. The CIoU loss is defined as:

$$L_{\mathrm{box}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{\delta})}{c^2} + \alpha v \quad (8)$$

To further refine localization performance, especially for small ships, Distribution Focal Loss (DFL) [53] is introduced as a fine-grained regression refinement mechanism. Unlike direct coordinate regression, DFL models bounding-box coordinates as discrete probability distributions and leverages a focal learning strategy to emphasize informative localization positions during training. Through this probabilistic formulation, neighboring positions jointly contribute to boundary estimation, enabling more precise and stable object boundary prediction under low signal-to-noise ratio (SNR) conditions.

Finally, these components are jointly optimized in a unified objective function defined as the weighted sum of the three losses:

$$L = \lambda_{\mathrm{cls}} L_{\mathrm{cls}} + \lambda_{\mathrm{box}} L_{\mathrm{box}} + \lambda_{\mathrm{dfl}} L_{\mathrm{dfl}} \quad (9)$$

Where L_{cls}, L_{box}, and L_{dfl} denote the classification loss, bounding-box regression loss, and Distribution Focal Loss, respectively, while λ_{cls}, λ_{box}, and λ_{dfl} represent the corresponding weighting coefficients.

III. RESULTS

A. DATASET

The High-Resolution SAR Images Dataset (HRSID) [31] is adopted as the primary dataset in this study. HRSID contains 5,604 high-resolution SAR images with spatial resolutions ranging from 0.5 m to 3 m, and each image has a size of 800 $\times$ 800 pixels. The dataset covers diverse maritime scenes, imaging geometries, and complex background conditions, providing a representative benchmark for ship detection tasks in practical scenarios.

The dataset includes ships of varying scales, orientations, and densities, making it suitable for evaluating multi-scale detection performance under complex environments.

To ensure a fair evaluation, the dataset is split into training and validation sets with an 80:20 ratio, resulting in 4,483 training images and 1,121 validation images. All images are resized and normalized before training to ensure consistent input dimensions and intensity distribution.

B. IMPLEMENTATION DETAILS

All experiments were conducted on a workstation equipped with an NVIDIA RTX 5090 GPU. The models were implemented using PyTorch and the Ultralytics YOLOv11 framework.

To ensure a fair comparison among different models, identical hyperparameter settings were applied across all experiments. The input image size was fixed at 800 $\times$ 800 pixels, and each model was trained for 100 epochs with a batch size of 16. Stochastic Gradient Descent (SGD) was used as the optimization algorithm, with an initial learning rate (lr_0) of 0.01 and a learning rate decay factor (lrf) of 0.01, resulting in a final learning rate of $\mathrm{lr}_0 \times$ lrf = 0.0001.

From an efficiency perspective, the observed reduction in inference speed from 769.23 FPS to 526.32 FPS is primarily attributed to the additional computational overhead introduced by the DEConv and MFM modules. Specifically, DEConv introduces multiple parallel differential convolution branches during feature extraction, while MFM performs channel-wise modulation and global context aggregation in the neck network. Although these operations increase computational cost, they are designed to be lightweight and are partially re-parameterized or simplified during inference.

Importantly, under a unified evaluation setting (batch size = 1, FP16 inference with NMS enabled), DE-MFM-YOLO still achieves an inference speed of over 500 FPS while maintaining competitive detection performance, demonstrating its suitability for real-time SAR ship detection. It should be noted that FPS values may vary across different hardware and deployment backends; therefore, the reported results are intended to reflect relative architectural efficiency under a unified experimental setting.

C. COMPARISON WITH MAINSTREAM METHODS

To evaluate the effectiveness of the proposed method, experiments were conducted on the HRSID dataset using several representative object detection frameworks, including the two-stage detector Faster R-CNN [15], the Transformer-based detector RT-DETR [43], and lightweight one-stage detectors from the YOLO family, including YOLOv8 – YOLOv12 [19][44][45][46]. Given the complex backgrounds and substantial scale variations of ship targets in SAR imagery, AP50 and mAP are particularly important for assessing localization and detection performance across different target sizes. Meanwhile, Precision and Recall provide complementary evaluations of classification reliability and detection completeness under challenging maritime conditions. Recent studies further demonstrate that multi-scale feature enhancement and attention mechanisms significantly improve robustness in cluttered SAR scenes, while transformer-based architectures enhance global context modeling and small target sensitivity [60][61]. In addition, lightweight YOLO-based detectors with improved feature fusion and re-parameterization strategies continue to achieve strong trade-offs between accuracy and efficiency for SAR ship detection tasks [62][63]. Furthermore, denoising-enhanced and sparse transformer-based methods have recently been proposed to mitigate speckle noise and improve feature discriminability in complex maritime environments [64]. The experimental results in terms of Precision, Recall, AP50, mAP, parameter count, and inference speed (FPS) are summarized in TABLE I.

Overall, the proposed method achieves a favorable balance between detection performance and computational efficiency. Compared with the YOLOv11 baseline, it attains an AP50 of 93.0%, along with a Precision of 91.7% and a Recall of 85.4%. These improvements are achieved while maintaining an inference speed of 526.32 FPS and a model size of 7M parameters. The results indicate that the proposed method improves detection accuracy without introducing noticeable computational overhead.

Compared with YOLOv8-s, the proposed method shows consistently better detection performance while maintaining a compact model scale. It achieves an AP50 of 93.0% and a Recall of 85.4%, while preserving high inference efficiency. These results suggest improved capability in small-target detection and multi-scale feature representation without increasing model complexity.

The proposed approach also outperforms Faster R-CNN with a ResNet-50 backbone, with an improvement of 6.5 percentage points in AP50, while also achieving significantly higher inference speed. This demonstrates its advantage for real-time SAR ship detection scenarios.

It is worth noting that the proposed method achieves performance comparable to Transformer-based detectors such as RT-DETR. Meanwhile, this performance is obtained with lower computational cost. The model maintains a compact parameter size (7.92M) and efficient inference speed, without incurring the heavy computational burden typically associated with Transformer-based architectures. These results indicate that a lightweight convolution-based design can achieve competitive accuracy while maintaining a more efficient balance between performance and computational cost in SAR ship detection tasks.

D. VISUALIZATION RESULTS

As shown in Fig. 4, representative visualization examples are presented to provide an intuitive comparison of detection performance under different scenarios. Visual comparisons of detection results were conducted among conventional object detection models, representative YOLO-series models, and the proposed DE-MFM-YOLO framework (see Fig. 1).

For a fair comparison, the best-performing baseline models and backbone variants were selected as reference methods when evaluating the effectiveness of DE-MFM-YOLO. The visual results highlight the differences in localization accuracy, target completeness, and small-object detection capability among the compared approaches.

During inference, the confidence threshold was set to 0.75 and the Non-Maximum Suppression (NMS) threshold was fixed at 0.6. These settings were chosen based on the detection characteristics of the proposed model. Specifically, DE-MFM-YOLO is able to generate relatively reliable confidence predictions for ship targets, allowing more valid detections to be retained while effectively suppressing redundant or low-confidence predictions. As a result, the selected parameter configuration provides a favorable balance between detection sensitivity and prediction reliability during evaluation.

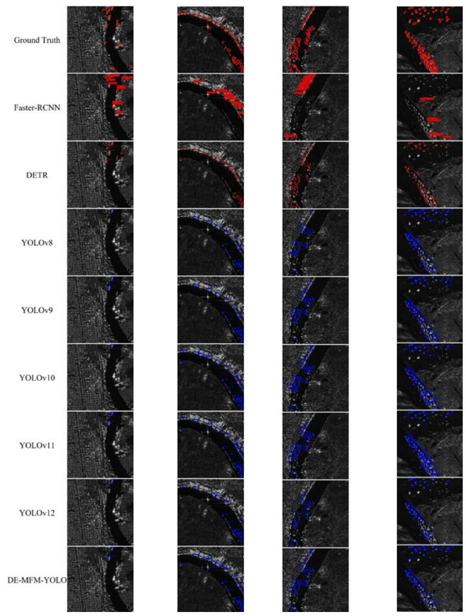

FIGURE 4. Comparison of the DE-MFM-YOLO model with traditional object detectors and YOLO-series models.

Fig. 4 presents a visual comparison of different object detection architectures for SAR ship detection. The conventional Faster R-CNN exhibits noticeable missed detections in densely distributed ship regions and shows relatively large localization errors in certain scenarios. Although DETR demonstrates improved performance in detecting continuously distributed vessels, challenges remain in accurately localizing targets located in high-density regions and near image boundaries. The YOLO-series models provide efficient detection performance; however, they may still suffer from overlapping bounding boxes and reduced recall in scenes containing small or densely distributed targets.

Compared with the reference methods, DE-MFM-YOLO demonstrates improved detection performance under complex maritime environments. In particular, the proposed model exhibits stronger robustness in challenging scenarios characterized by background clutter, including ports, coastal regions, and rough sea conditions. The predicted bounding boxes show better alignment with target boundaries, while the occurrence of false detections is reduced.

TABLE 1. Comparison results of different methods on the HRSID dataset.

Models	Backbone or size	P/%	R/%	AP50/%	AP/%	Params (M)	FPS (img/s)
YOLOv8	s	91.5	81.9	90.4	65.7	3.01	500.00
YOLOv9	s	90.9	84.5	91.5	67.5	7.28	312.50
YOLOv10	s	88.4	83.1	88.5	65.1	8.06	476.19
YOLO11	s	91.9	83.9	91.5	68.7	2.58	769.23
YOLO12	s	92.2	82.7	91.6	68.6	2.56	454.54
Faster R-CNN	R50	86.1	81.5	86.5	61.1	41.10	310.10
RT-DETR	–	91.4	85.7	92.0	60.6	32.81	232.56
DE-MFM-YOLO (ours)	–	91.7	85.4	93.0	70.1	7.92	526.32

The improvements can be attributed to the complementary effects of the proposed DEConv and MFM modules. By enhancing local edge and texture representations, DEConv strengthens the feature responses of small ship targets while suppressing background interference, thereby improving localization accuracy and reducing redundant detections. Meanwhile, the Modulation Fusion Module (MFM) facilitates adaptive integration of multi-scale information, enabling more effective feature aggregation across different semantic levels. This capability is particularly beneficial in densely distributed ship scenes, where accurate discrimination between adjacent targets is critical.

Overall, the proposed method achieves a favorable balance between detection accuracy and recall while maintaining robust performance across different target scales and background conditions. The qualitative results indicate that DE-MFM-YOLO provides a reliable solution for SAR ship detection, particularly in scenarios involving small targets, dense target distributions, and complex maritime environments.

E. ABLATION STUDIES

To evaluate the individual contributions and joint effectiveness of the proposed DEConv and MFM modules, ablation experiments were conducted by progressively introducing these components into the YOLOv11 baseline model and comparing different module combinations on the HRSID dataset. The effectiveness of each module was further analyzed by independently enabling DEConv and MFM to examine their respective roles and their complementarity. TABLE II reports the detection accuracy and inference efficiency of the baseline model, the single-module variants (DEConv only and MFM only), and the full model incorporating both modules.

According to the results presented in TABLE II, different module combinations exhibit distinct impacts on detection accuracy, recall, and inference efficiency. A detailed analysis is provided as follows.

(1) Introducing DEConv only

When the Detail-Enhanced Convolution (DEConv) module is incorporated into the backbone network, the detection performance improves noticeably, with AP50 increasing to 92.6%, while Precision and Recall reach 92.6% and 84.3%, respectively.

This improvement is primarily attributed to the enhanced local feature extraction capability introduced by DEConv. By integrating multiple differential convolution branches, including center-difference and directional-difference filters, DEConv strengthens the representation of edges, textures, and fine-grained structural cues. This allows the network to better capture weak ship signatures in SAR imagery, particularly for small vessels with low signal-to-clutter ratios.

From the ablation perspective, DEConv contributes most directly to detection accuracy and localization quality, indicating that improved feature representation at the backbone stage is critical for SAR ship detection. This effect is also clearly reflected in the visual results in Fig. 4, where models equipped with stronger feature extraction ability exhibit more accurate bounding-box alignment and fewer missed small targets compared with conventional detectors.

In addition, DEConv helps mitigate boundary degradation caused by speckle noise and sea clutter, leading to more stable feature responses in complex maritime environments. Although the multi-branch design introduces additional computational operations, the structural re-parameterization strategy ensures that the inference stage remains efficient, with FPS reduced from 769.23 to 555.55 while still satisfying real-time requirements. Overall, DEConv primarily enhances feature discriminability, which is consistently reflected in both quantitative improvements and visual localization quality.

(2) Introducing MFM only

When only the Modulation Fusion Module (MFM) is introduced into the neck network, AP50 slightly increases to 91.6%, while Precision and Recall decrease to 90.5% and 81.9%, respectively. Meanwhile, inference speed decreases from 769.23 FPS to 666.67 FPS.

From the ablation results, MFM plays a role in feature aggregation and multi-scale interaction, but its effectiveness is highly dependent on the quality of backbone features. Without sufficiently enhanced low-level representations, the adaptive modulation mechanism may partially suppress weak responses corresponding to small ship targets, leading to limited gains in Recall.

This phenomenon is also consistent with the visual results in Fig. 4, where models relying only on feature fusion mechanisms tend to show incomplete detection of small or densely distributed ships, especially in cluttered coastal or

TABLE 2. Ablation study results on the HRSID dataset. ✓ indicates that the corresponding module is adopted, while × indicates that it is not included.

Models	DEConv	MFM	P/%	R/%	AP50/%	AP/%	FPS (img/s)
YOLOV11 Baseline	×	×	91.9	83.9	91.5	68.7	769.23
+DEConv	✓	×	92.6	84.3	92.6	69.9	555.55
+MFM	×	✓	90.5	81.9	91.6	65.5	666.67
DEConv+MFM (ours)	✓	✓	91.7	85.4 (+1.50%)	93.0 (+1.64%)	70.1 (+2.04%)	526.32

port scenes. In such cases, some weak targets are either suppressed or merged into surrounding background responses.

Nevertheless, MFM improves the overall balance of feature contributions across scales by dynamically reweighting channel responses, which helps refine certain high-confidence detections. This explains the slight improvement in AP50 observed in the quantitative results. However, since MFM focuses on feature redistribution rather than feature enrichment, its standalone benefit remains limited, indicating that it functions more effectively as a complementary module rather than an independent enhancement.

(3) Joint Introduction of DEConv and MFM

When DEConv and MFM are simultaneously integrated into the network, the model achieves the best overall performance, with AP50 reaching 93.0%, while Precision and Recall increase to 91.7% and 85.4%, respectively.

The ablation results demonstrate that the two modules exhibit strong complementarity: DEConv enhances low-level feature representation by strengthening edge and texture information, while MFM improves high-level feature interaction through adaptive multi-scale fusion. Their combination enables the network to simultaneously improve feature quality and feature utilization efficiency.

This complementary effect is clearly reflected in Fig. 4. Compared with conventional detectors and baseline YOLO-series models, the proposed method produces more complete detections in densely distributed ship regions, with fewer missed targets and reduced false positives. In particular, small ship instances that are easily overlooked by other methods are more consistently detected, and bounding boxes show improved alignment with object boundaries.

Moreover, in complex backgrounds such as ports and coastal regions, the combined effect of DEConv and MFM significantly improves robustness. DEConv suppresses background interference at the feature extraction stage, while MFM refines multi-scale semantic interaction, reducing redundancy and improving target separability.

From the ablation perspective, the joint configuration achieves the most stable trade-off between detection accuracy and inference efficiency. Although the computational cost increases moderately, the model still maintains real-time performance, indicating that the two modules enhance performance in a mutually reinforcing manner rather than introducing redundant computation.

Overall, both quantitative ablation results and qualitative visual comparisons confirm that the integration of DEConv and MFM leads to consistent improvements across localization accuracy, recall, and robustness in complex SAR ship detection scenarios.

IV. DISCUSSION

As demonstrated by the experimental results, the proposed method achieves improved detection performance and inference efficiency compared with the baseline model. However, the effectiveness of the proposed modules may vary depending on their placement within the network architecture and their interaction with surrounding components. Therefore, this section further investigates the influence of different insertion positions of DEConv and MFM within the YOLOv11 framework on overall detection performance.

A. OPTIMAL PLACEMENT OF DECONV AND MFM MODULES

To investigate the optimal insertion positions of the DEConv and MFM modules within the YOLOv11 architecture, a series of placement experiments were conducted. Specifically, DEConv was introduced at different stages of the backbone network, denoted as DE-1, DE-2, and DE-3, while MFM was inserted at different locations within the neck network, denoted as M-1 and M-2. The corresponding module placement configurations are illustrated in Fig. 5 and Fig. 6, where the experiment identifiers are consistent with those reported in TABLE III. TABLE III summarizes the detection performance of all experimental configurations in terms of Precision, Recall, AP50, AP, and inference speed (FPS). These results provide a quantitative basis for evaluating the influence of different module insertion positions on both detection accuracy and computational efficiency.

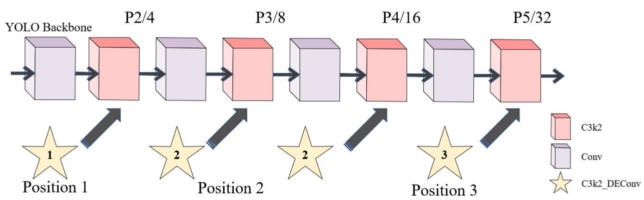

FIGURE 5. Alternative insertion positions of the DEConv module in the YOLOv11 backbone. PX/Y denotes the feature map at pyramid level X (X = 2, 3, 4, 5), where Y indicates the down-sampling factor. The positions indexed by i (i = 1, 2, 3) correspond to the three experimental groups (DE-i) listed in Table III.

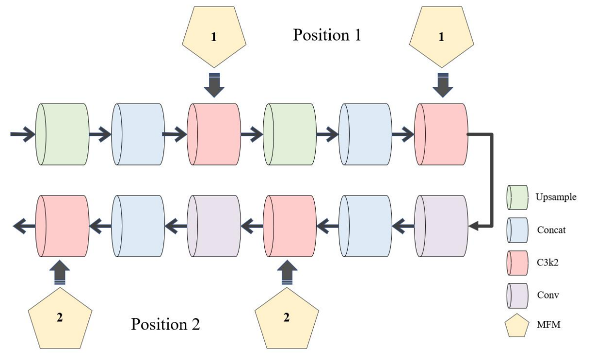

FIGURE 6. Alternative insertion positions of the MFM module in the YOLOv11 neck network. The positions indexed by i (i = 1, 2) correspond to the two experimental groups (M-i) in Table III.

TABLE 3. Impact of inserting DEConv and MFM at different positions

Model	P/%	R/%	AP50/%	AP/%	FPS (img/s)
Baseline	91.9	83.9	91.5	68.7	769.23
DE-1	91.4	83.3	92.1	68.3	555.55
DE-2	91.3	83.2	91.6	67.2	588.24
DE-3	90.8	82.8	90.9	65.9	625.00
M-1	91.7	84.9	92.8	69.1	500.00
M-2	90.8	86.7	93.0	70.6	454.55
ours	91.7	85.4	93.0	70.1	526.32

The experimental results indicate that introducing the DEConv module alone (DE-1/DE-2/DE-3) does not lead to substantial improvements in detection accuracy and, in some configurations, may even result in a slight performance decline. One possible reason is that the additional computational complexity introduced by DEConv increases the processing burden of the network, leading to a reduction in inference speed. Moreover, applying DEConv only to shallow or intermediate feature layers may not be sufficient to consistently enhance localization and classification performance. In some cases, it may also affect the balance between semantic and spatial information across different feature levels. These observations suggest that the effectiveness of local feature enhancement depends on the overall feature propagation process and may require complementary mechanisms to fully exploit its advantages.

Although the MFM module does not lead to substantial improvements across all evaluation metrics when deployed independently, it generally enhances Recall and promotes more effective multi-level feature interaction. A possible explanation is that MFM relies on sufficiently discriminative low-level features. Without the support of DEConv, the adaptive modulation mechanism may suppress weak small-target responses together with background noise, thereby limiting the overall performance gain under certain configurations. Nevertheless, MFM provides a more effective feature fusion strategy and establishes favorable conditions for subsequent collaborative optimization. In particular, a relatively conservative placement strategy (e.g., M-1) improves detection performance while maintaining moderate inference efficiency, whereas a deeper placement configuration (e.g., M-2) further enhances feature fusion capability and achieves the highest AP among the evaluated configurations. However, this improvement is accompanied by a noticeable reduction in inference speed. These results highlight the trade-off between feature fusion effectiveness and computational efficiency when selecting the insertion position of MFM. Although M-2 achieves slightly higher AP, it introduces a larger computational burden and lower inference speed. Therefore, the final configuration was selected by jointly considering detection accuracy, model complexity, and deployment efficiency.

These findings suggest that feature fusion alone may not fully exploit its advantages when the underlying feature representations remain insufficiently discriminative. Therefore, enhancing feature extraction and feature fusion simultaneously becomes a more effective design strategy.

Based on the above observations, DEConv and MFM were jointly incorporated into the overall network architecture. Under this integrated design, DEConv enhances the representation capability of key feature extraction layers, while MFM strengthens information interaction and coordination across different feature levels during feature fusion. Although the parameter increase introduced by DEConv and MFM is relatively small, the additional feature interaction, adaptive modulation, and feature aggregation operations increase computational complexity and reduce inference speed to some extent. This observation indicates that parameter count alone does not fully reflect the actual computational cost of a network. Nevertheless, the proposed architecture achieves a more balanced trade-off between detection performance and inference efficiency, maintaining favorable Recall and AP metrics while achieving higher inference speed than the M-2 configuration. These findings further demonstrate that collaborative optimization of feature extraction and feature fusion modules is an effective strategy for lightweight object detection networks. From an efficiency perspective, the observed reduction in inference speed from 769.23 FPS to 526.32 FPS is primarily attributed to the additional computational overhead introduced by the DEConv and MFM modules. Specifically, DEConv introduces multiple parallel differential convolution branches during feature extraction, while MFM performs channel-wise modulation and global context aggregation in the neck network. Although these operations increase computational cost, they are designed to be lightweight and are partially re-parameterized or simplified during inference. Importantly, under a unified evaluation setting (batch size = 1, FP16 inference with NMS enabled), the proposed model still achieves more than 500 FPS, which significantly exceeds the real-time requirements of practical SAR ship monitoring systems (typically 25 – 60 FPS). Therefore, the trade-off between detection accuracy and moderate computational overhead is considered acceptable for real-world deployment scenarios. To ensure fairness, the FPS of all compared models was measured under the same evaluation protocol. The observed decrease in inference speed is primarily attributed to the additional computational branches introduced by DEConv and the feature modulation operations in MFM. Therefore, the comparison reflects architectural differences rather than implementation bias.

B. COMPARISON WITH SIMILAR METHODS

To quantitatively evaluate the effectiveness of the proposed DEConv and MFM modules, several representative lightweight and attention-based modules were incorporated into the YOLOv11s baseline for comparative analysis. Specifically, StarNet [38], FSAS [39], GSConv [40], and SCSA [41] were selected as reference methods. This comparison aims to evaluate the effectiveness of the proposed modules in terms of feature representation, detection accuracy, and overall network efficiency. The experimental results are summarized in TABLE IV.

TABLE 4. Comparison of different modules (baseline: YOLOv11s).

Model	P/%	R/%	AP50/%	AP/%	FPS (img/s)
Baseline	**91.9**	83.9	91.5	68.7	**769.23**
+StarNet	90.3	78.4	87.8	59.3	212.77
+FSAS	91.8	83.3	91.4	66.3	166.67
+GSConv	90.6	83.7	92.1	67.8	112.36
+SCSA	91.0	82.7	90.4	65.2	128.21
DE-MFM-YOLO (ours)	91.7	**85.4 (+1.5%)**	**93.0 (+1.5%)**	**70.1 (+1.4%)**	526.32

performance and computational efficiency in distinct ways, reflecting varying trade-offs between accuracy and inference speed.

Among the compared methods, the integration of the StarNet module results in a slight decrease in detection accuracy. A possible explanation is that the additional feature aggregation operations introduce redundant contextual information, which may weaken the discriminative representation of small ship targets in SAR imagery. Similarly, although the SCSA module introduces only a limited increase in model complexity, its attention mechanism may emphasize background responses together with target features, thereby reducing the effectiveness of feature selection for small and weak ship targets. These observations suggest that directly introducing generic feature enhancement modules does not necessarily guarantee improved performance for SAR ship detection, where target characteristics and background distributions differ substantially from those of natural images.

From the perspective of inference efficiency, the baseline model achieves an inference speed of 769.23 FPS, demonstrating strong real-time capability. As expected, the introduction of additional modules generally increases computational overhead and reduces inference speed. In particular, StarNet and FSAS introduce more complex feature interaction operations and attention mechanisms, resulting in FPS values of 212.77 and 166.67, respectively. In contrast, the grouped-convolution design adopted by GSConv effectively reduces redundant computations, enabling the model to maintain a relatively higher inference speed of 112.36 FPS while preserving competitive detection performance. These results indicate that lightweight architectural designs remain important for balancing detection accuracy and deployment efficiency in practical SAR ship detection scenarios.

Compared with the above methods, the proposed DEConv and MFM modules exhibit stronger complementarity. DEConv enhances local feature representation and strengthens the extraction of discriminative ship characteristics, while MFM promotes effective information interaction among multi-scale feature maps during feature fusion. The combination of these two modules enables the network to simultaneously improve feature extraction and feature fusion capabilities, thereby achieving superior overall detection performance. Experimental results demonstrate that the proposed DE-MFM-YOLO achieves a favorable balance between detection accuracy and inference efficiency, outperforming the baseline model and other enhanced variants in terms of overall effectiveness. These findings further indicate that collaborative optimization of feature extraction and feature fusion is a more effective strategy for lightweight SAR ship detection than introducing a single generic enhancement module.

V. CONCLUSION

Small targets, blurred appearances, and large variations in object scale are persistent challenges in SAR-based ship detection for practical maritime monitoring systems. These issues significantly degrade feature discriminability and hinder robust detection performance under complex sea clutter and imaging conditions. To address these issues, this study proposes an efficient ship detection framework based on an enhanced YOLOv11 architecture. A Detail-Enhanced Convolution (DEConv) module is introduced into the backbone to strengthen edge and texture representations, enabling more discriminative feature learning in complex SAR scenes. In addition, a Modulation Fusion Module (MFM) is embedded in the neck to adaptively reweight multi-scale features through learnable modulation factors, improving feature aggregation across different semantic levels. The resulting architecture enhances representation capability while maintaining real-time detection efficiency, which is essential for large-scale maritime monitoring applications.

Experimental results on the HRSID dataset demonstrate consistent improvements over the baseline and representative methods. In particular, AP50 increases from 91.5% to 93.0% while maintaining comparable model complexity and real-time inference capability. Qualitative results further show improved localization accuracy in challenging scenarios, including densely distributed vessels, low-contrast targets, and significant scale variations. These results indicate strong potential for reliable automated maritime perception in real-world environments.

Overall, this study provides an effective and efficient solution for SAR ship detection and contributes to sustainable maritime monitoring systems, including maritime traffic supervision, illegal fishing detection, and marine environmental protection. Future work will focus on improving robustness under complex maritime conditions such as dense ship distribution and strong speckle noise, as well as enhancing feature representation for more challenging SAR scenes.

REFERENCES

[1] F. S. Paolo, D. Kroodsma, J. Raynor, T. Hochberg, P. Davis, J. Cleary, L. Marsaglia, S. Orofino, C. Thomas, and P. Halpin, "Satellite mapping reveals extensive industrial activity at sea," *Nature*, vol. 625, pp. 85–91, 2024, doi: 10.1038/s41586-023-06825-8.

[2] L. Min, F. Dou, Y. Zhang, D. Shao, L. Li, and B. Wang, "CM-YOLO: Context Modulated Representation Learning for Ship Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, Art. no. 4202414, 2025, doi: 10.1109/TGRS.2025.3538848.

[3] C. Alexandre, R. Devillers, D. Mouillot, R. Seguin, and T. Catry, "Ship Detection with SAR C-Band Satellite Images: A Systematic Review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 14353–14367, 2024, doi: 10.1109/JSTARS.2024.3437187.

[4] M. J. Er, Y. Zhang, J. Chen, and W. Gao, "Ship detection with deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, pp. 11825–11865, 2023, doi: 10.1007/s10462-023-10455-x.

[5] Y. Li et al., "Ship Detection and Classification Based on Cascaded Detection of Hull and Wake from Optical Satellite Remote Sensing Imagery," *GISci. Remote Sens.*, vol. 60, Art. no. 2196159, 2023, doi: 10.1080/15481603.2023.2196159.

[6] Z. Liu, J. Lyu, and Y. Fu, "Scattering-Guided Class-Irrelevant Filtering for Adversarially Robust SAR Automatic Target Recognition," *Signal Process.*, vol. 239, Art. no. 110273, 2026, doi: 10.1016/j.sigpro.2025.110273.

[7] Z. Wang, R. Wang, J. Ai, F. Yang, and Y. Zhou, "Global and Local Context-Aware Ship Detector for High-Resolution SAR Images," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, pp. 4159–4167, 2023, doi: 10.1109/TAES.2023.3237520.

[8] J. Li et al., "Small Ship Detection Based on Improved Neural Network Algorithm and SAR Images," *Remote Sens.*, vol. 17, Art. no. 2586, 2025, doi: 10.3390/rs17152586.

[9] S. Baraha et al., "CNN-Augmented SAR Image Despeckling Using Modified Speckle Reducing Anisotropic Diffusion and Discrete Wavelet Transform," *Signal Process. Image Commun.*, vol. 138, Art. no. 117380, 2025, doi: 10.1016/j.image.2025.117380.

[10] S. Li, H. Wei, Y. Mao, and J. Fan, "A Novel Two-Stage Superpixel CFAR Method Based on Truncated KDE Model for Target Detection in SAR Images," *Electronics*, vol. 14, Art. no. 1327, 2025, doi: 10.3390/electronics14071327.

[11] S. Sun and J. Wang, "Ship Detection in SAR Images Based on Steady CFAR Detector and Knowledge-Oriented GBDT Classifier," *Electronics*, vol. 13, Art. no. 2692, 2024, doi: 10.3390/electronics13142692.

[12] L. Zhang, Z. Zhang, S. Lu, D. Xiang, and Y. Su, "Fast Superpixel-Based Non-Window CFAR Ship Detector for SAR Imagery," *Remote Sens.*, vol. 14, Art. no. 2092, 2022, doi: 10.3390/rs14092092.

[13] N. Li, X. Pan, L. Yang, Z. Huang, Z. Wu, and G. Zheng, "Adaptive CFAR Method for SAR Ship Detection Using Intensity and Texture Feature Fusion Attention Contrast Mechanism," *Sensors*, vol. 22, Art. no. 8116, 2022, doi: 10.3390/s22218116.

[14] J. Li, C. Xu, H. Su, L. Gao, and T. Wang, "Deep learning for SAR ship detection: Past, present and future," *Remote Sens.*, vol. 14, Art. no. 2712, 2022, doi: 10.3390/rs14112712.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[16] Q. Shuai and X. Wu, "Object detection system based on SSD algorithm," in Proc. Int. Conf. Culture-Oriented Sci. Technol., Beijing, China, 2020, pp. 141–144, doi: 10.1109/ICCST50977.2020.00033.

[17] J. Cheng et al., "A feature enhanced RetinaNet-based for instance-level ship recognition," *Eng. Appl. Artif. Intell.*, vol. 126, Art. no. 107133, 2023, doi: 10.1016/j.engappai.2023.107133.

[18] M. Zhu et al., "A novel anchor-free method based on FCOS + ATSS for ship detection in SAR images," *Remote Sens.*, vol. 14, Art. no. 2034, 2022, doi: 10.3390/rs14092034.

[19] Q. Cao et al., "LH-YOLO: A Lightweight and High-Precision SAR Ship Detection Model Based on the Improved YOLOv8n," *Remote Sens.*, vol. 16, Art. no. 4340, 2024, doi: 10.3390/rs16224340.

[20] C. Yu and Y. Shin, "SMEP-DETR: Transformer-Based Ship Detection for SAR Imagery with Multi-Edge Enhancement and Parallel Dilated Convolutions," *Remote Sens.*, vol. 17, Art. no. 953, 2025, doi: 10.3390/rs17060953.

[21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," arXiv preprint arXiv:2010.04159, 2020.

[22] Z. Hong et al., "Multi-scale ship detection from SAR and optical imagery via a more accurate YOLOv3," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 6083–6101, 2021, doi: 10.1109/JSTARS.2021.3087555.

[23] S. Liu, P. Chen, and Y. Zhang, "A multiscale feature pyramid SAR ship detection network with robust background interference," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 9904–9915, 2023, doi: 10.1109/JSTARS.2023.3325376.

[24] J. Chen et al., "LKPF-YOLO: A Small Target Ship Detection Method for Marine Wide-Area Remote Sensing Images," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, pp. 8482–8496, 2024.

[25] J. Si et al., "Maritime ship detection method for satellite images based on multiscale feature fusion," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 6642- 6655, 2023, doi: 10.1109/JSTARS.2023.3296898.

[26] R. L. Rocha and F. A. P. de Figueiredo, "Enhancing YOLO-Based SAR Ship Detection with Attention Mechanisms," *Remote Sens.*, vol. 17, Art. no. 3170, 2025, doi: 10.3390/rs17183170.

[27] Y. Chen et al., "AJANet: SAR Ship Detection Network Based on Adaptive Channel Attention and Large Separable Kernel Adaptation," *Remote Sens.*, vol. 17, Art. no. 1745, 2025, doi: 10.3390/rs17101745.

[28] M. Zhang, X. Rong, and X. Yu, "Light-SDNet: A lightweight CNN architecture for ship detection," *IEEE Access*, vol. 10, pp. 86647–86662, 2022, doi: 10.1109/ACCESS.2022.3199352.

[29] P. Huang et al., "IA-CIoU: An improved IoU bounding box loss function for SAR ship target detection methods," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 10569–10582, 2024, doi: 10.1109/JSTARS.2024.3402540.

[30] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," arXiv preprint arXiv:2410.17725, 2024.

[31] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020, doi: 10.1109/ACCESS.2020.3005861.

[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 2117–2125, doi: 10.1109/CVPR.2017.106.

[33] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "DEA-Net: Single Image Dehazing Based on Detail-Enhanced Convolution and Content-Guided Attention," *in Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, 2024, pp. 1234–1242.

[34] Y. Liu, H. Zhao, X. Wang, J. Li, and Q. Zhang, "Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing," *IEEE Trans. Multimedia*, vol. 26, pp. 5842–5855, 2024.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.324.

[36] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 9627–9636, doi: 10.1109/ICCV.2019.00972.

[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[38] X. Ma, X. Dai, Y. Bai, Y. Wang, and Y. Fu, "Rewrite the Stars," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2024, pp. 5694–5703, doi: 10.1109/CVPR52733.2024.00544.

[39] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient Frequency Domain-Based Transformers for High-Quality Image Deblurring," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, BC, Canada, 2023, pp. 5886–5895, doi: 10.1109/CVPR52729.2023.00570.

[40] H. Li et al., "Slim-Neck by GSConv: A Lightweight Design for Real-Time Detector Architectures," *J. Real-Time Image Process.*, vol. 21, Art. no. 62, 2024, doi: 10.1007/s11554-024-01436-6.

[41] Y. Si et al., "SCSA: Exploring the synergistic effects between spatial and channel attention," *Neurocomputing*, vol. 634, Art. no. 129866, 2025, doi: 10.1016/j.neucom.2025.129866.

[42] T. Zhang et al., "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, Art. no. 3690, 2021, doi: 10.3390/rs13183690.

[43] Y. Zhao et al., "DETRs Beat YOLOs on Real-Time Object Detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2024, pp. 16965–16974, doi: 10.1109/CVPR52733.2024.01605.

[44] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2024, doi: 10.1007/978-3-031-72751-1_1.

[45] A. Wang et al., "YOLOv10: Real-Time End-to-End Object Detection," *in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 107984–108011, doi: 10.52202/079017-3429.

[46] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," arXiv preprint arXiv:2502.12524, 2025.

[47] P. Qin et al., "Multilayer feature extraction network for military ship detection from high-resolution optical remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 11058–11069, 2021, doi: 10.1109/JSTARS.2021.3123080.

[48] K. Zhou et al., "Ship detection in SAR images based on multi-scale feature extraction and adaptive feature fusion," *Remote Sens.*, vol. 14, Art. no. 755, 2022, doi: 10.3390/rs14030755.

[49] Y. Zhao et al., "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 2738–2756, 2020, doi: 10.1109/JSTARS.2020.2997081.

[50] Z. Wang et al., "Global and local context-aware ship detector for high-resolution SAR images," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, pp. 4159–4167, 2023, doi: 10.1109/TAES.2023.3237520.

[51] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," arXiv preprint arXiv:1702.05659, 2017, doi: 10.4467/20838476SI.16.004.6185.

[52] H. Rezatofighi et al., "Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.

[53] X. Li et al., "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[54] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, pp. 5393–5397, 2020, doi: 10.30534/ijatcse/2020/175942020.

[55] Z. Zheng et al., "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," *in Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 12993- 13000, doi: 10.1609/aaai.v34i07.6999.

[56] C. Niu et al., "SAR-LtYOLOv8: A Lightweight YOLOv8 Model for Small Object Detection in SAR Ship Images," *Comput. Syst. Sci. Eng.*, vol. 48, pp. 827–845, 2024, doi: 10.32604/csse.2024.056736.

[57] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "AugFPN: Improving Multi-scale Feature Learning for Object Detection," arXiv preprint arXiv:1912.05384, 2019, doi: 10.1109/CVPR42600.2020.01261.

[58] Z. Zhang, X. Qiu, and Y. Li, "SEFPN: Scale-Equalizing Feature Pyramid Network for Object Detection," *Sensors*, vol. 21, no. 21, Art. no. 7136, 2021, doi: 10.3390/s21217136.

[59] H.-J. Park, J.-W. Kang, and B.-G. Kim, "ssFPN: Scale Sequence Feature-Based Feature Pyramid Network for Object Detection," *Sensors*, vol. 23, no. 9, Art. no. 4432, 2023, doi: 10.3390/s23094432.

[60] Y. Huang, D. Wang, B. Wu, and D. An, "NST-YOLO11: ViT Merged Model with Neuron Attention for Arbitrary-Oriented Ship Detection in SAR Images," *Remote Sens.*, vol. 16, Art. no. 4760, 2024, doi: 10.3390/rs16244760.

[61] C. Yu and Y. Shin, "SMEP-DETR: Transformer-Based Ship Detection for SAR Imagery with Multi-Edge Enhancement," *Remote Sens.*, vol. 17, 2025, doi: 10.3390/rs17060953.

[62] Y. Xu, X. Xue, C. Li, S. Zhao, X. Xu, and C. Zeng, "RLE-YOLO: A Lightweight and Multiscale SAR Ship Detection Based on Improved YOLOv8," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3550541.

[63] J. Niu, M. Li, H. Lin, Y. Liu, Z. Liu, H. Li, and S. Niu, "LD-YOLO: A Lightweight Dynamic Convolution-Based YOLOv8n Framework for Robust Ship Detection in SAR Imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 23, pp. 1–5, 2026, doi: 10.1109/LGRS.2025.3630098.

[64] X. Zhao et al., "Denoising-Enhanced YOLO for Robust SAR Ship Detection," arXiv preprint arXiv:2602.23820, 2026.

[65] Z. Yang, X. Xia, Y. Liu, G. Wen, W. E. Zhang, and L. Guo, "LPST-Det: Local-Perception-Enhanced Swin Transformer for SAR Ship Detection," *Remote Sens.*, vol. 16, no. 3, Art. no. 483, 2024, doi: 10.3390/rs16030483.