Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2024.056736

ARTICLE

SAR-LtYOLOv8: A Lightweight YOLOv8 Model for Small Object Detection in SAR Ship Images

Conghao Niu[1,*], Dezhi Han[1], Bing Han[2] and Zhongdai Wu[2]

[1]School of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China

[2]Shanghai Ship and Shipping Research Institute Co., Ltd., Shanghai, 200135, China

*Corresponding Author: Conghao Niu. Email: niuconghao8@gmail.com

Received: 29 July 2024 Accepted: 22 October 2024 Published: 22 November 2024

ABSTRACT

The high coverage and all-weather capabilities of Synthetic Aperture Radar (SAR) image ship detection make it a widely accepted method for maritime ship positioning and identification. However, SAR ship detection faces challenges such as indistinct ship contours, low resolution, multi-scale features, noise, and complex background interference. This paper proposes a lightweight YOLOv8 model for small object detection in SAR ship images, incorporating key structures to enhance performance. The YOLOv8 backbone is replaced by the Slim Backbone (SB), and the Delete Medium-sized Detection Head (DMDH) structure is eliminated to concentrate on shallow features. Dynamically adjusting the convolution kernel weights of the Omni-Dimensional Dynamic Convolution (ODConv) module can result in a reduction in computation and enhanced accuracy. Adjusting the model's receptive field is done by the Large Selective Kernel Network (LSKNet) module, which captures shallow features. Additionally, a Multi-scale Spatial-Channel Attention (MSCA) module addresses multi-scale ship feature differences, enhancing feature fusion and local region focus. Experimental results on the HRSID and SSDD datasets demonstrate the model's effectiveness, with a 67.8% reduction in parameters, a 3.4% improvement in AP (average precision) @0.5, and a 5.4% improvement in AP@0.5:0.95 on the HRSID dataset, and a 0.5% improvement in AP@0.5 and 1.7% in AP@0.5:0.95 on the SSDD dataset, surpassing other state-of-the-art methods.

KEYWORDS

SAR; ship detection; MSCA; deep learning

1 Introduction

The importance of ship safety and accurate positioning at sea has grown due to the increase in maritime cargo transactions. Conventional methods encounter obstacles such as adverse weather conditions and interference from water backgrounds when trying to detect ship targets, which has become a primary focus. The detection of small ship targets can be improved through further research [1]. Synthetic Aperture Radar (SAR) offers high-resolution imaging, overcoming weather obstacles and producing detailed images comparable to optical photographs [2]. SAR's ship target detection has advantages in maritime safety and marine resources development over other remote sensing methods [3].

With advancements in SAR imaging technology, a variety of novel ship detection methods have emerged, generally classified into traditional and deep learning-based approaches. Conventional techniques rely on preprocessing and feature extraction but are constrained by fixed thresholds and are highly susceptible to noise and clutter. For instance, Constant False Alarm Rate (CFAR) [4] detection requires empirical or statistical threshold adjustments, increasing computational complexity. These methods often struggle with environmental variability, lack robustness against noise, and have difficulties adapting to different scales. In contrast, deep learning has markedly improved SAR ship detection by addressing many of these limitations. Convolutional Neural Networks (CNNs) excel in detecting ships amidst complex backgrounds and varying scales by automatically extracting high-level features through multiple convolutional, pooling, and fully connected layers. Li et al. [5] introduced an optimized Faster Region-based Convolutional Neural Network (R-CNN) framework that enhances both the speed and accuracy of SAR ship detection through refined network structure and feature extraction. Similarly, Jiang et al. [6] developed a method that integrates Faster R-CNN with Fast Nonlocal Mean (FNLM) filtering and an optimized Chan-Vese model, effectively extracting ship contours, managing complex scenarios, and improving detection and monitoring in SAR images.

Despite the progress made in SAR ship detection, there are still many unresolved issues. Due to their limited pixel presence, small target ships often have weak signals that can easily be overwhelmed by background noise, making their detection more complicated. The effectiveness of traditional feature extraction methods and detection accuracy are reduced due to the small size, which also restricts feature information. Additionally, the complex marine environment, with background interferences like waves and sea ice, further hampers detection, especially for low-contrast, small targets. Therefore, there are still challenges in SAR ship detection, such as improving small target feature extraction, minimizing background noise, and enhancing model accuracy.

The introduction of SAR-LtYOLOv8 is a model for detecting small ship targets in SAR images despite complex backgrounds. SAR-LtYOLOv8 improves multi-scale detection accuracy and reliability by employing a deep learning network. Experiments have demonstrated that SAR-LtYOLOv8 is more effective in detecting the HRSID and SSDD datasets than other methods.

YOLOv8n is the foundational model used to optimize the backbone and neck components of the detection framework, as described in this paper. As follows are the main contributions summarized:

1. By replacing the YOLOv8n backbone with the Slim Backbone (SB) structure, model complexity, parameter count, and computational costs are reduced, leading to enhanced accuracy and robustness in detecting small ship targets. Furthermore, by utilizing the Delete Medium-sized Detection Head (DMDH) structure for the model head and removing a medium-sized detection head, the model's accuracy, parameters, and efficiency are improved.
2. In both the backbone and neck of YOLOv8n, the Conv module is replaced by the Omni-Dimensional Dynamic Convolution (ODConv) [7] module, which reduces computational complexity and speeds up operations. The neck is equipped with the Large Selective Kernel Network (LSKNet) [8] module to improve accuracy by facilitating interaction between shallow and deep features, which improves the model's ability to detect small targets.
3. The neck section of YOLOv8n has been equipped with the Multi-scale Spatial-Channel Attention (MSCA) module as a proposed solution. This module synthesizes the principles of Multi-Scale Class Activation Maps (MS-CAM) [9] and convolutional block attention module (CBAM) [10], combining multiscale channel attention with spatial-level attention. The MSCA module enhances the performance of ship target detection by enhancing the model's focus on information across different scales.

4. This paper's proposed method is proven effective in experiments on the HRSID and SSDD datasets.

The paper is structured as follows: Section 2 reviews relevant ship detection algorithms. Section 3 outlines the proposed method and improvement modules. Section 4 details the experimental setup, datasets, configurations, evaluation metrics, and results. Section 5 concludes and suggests future directions.

2 Related Works

Traditional remote sensing algorithms for ship target detection, such as CFAR, Gaussian model-based CFAR, template matching, spectral residuals, and wavelet-based methods, have demonstrated some capabilities but often fall short in terms of speed and accuracy for SAR ship detection. As a result, there is a growing trend toward using deep learning for more precise and timely detection. Deep learning, particularly through algorithms like YOLO, has emerged as a prominent approach in computer vision due to its superior accuracy and faster processing speeds. YOLO, a single-stage detection algorithm, directly predicts candidate frames from images, allowing for extremely rapid detection and making it well-suited for applications that require swift responses, such as maritime traffic safety monitoring.

Several studies have advanced the YOLO model for improved SAR ship detection. Zhao et al. [11] introduced Convolutional Block Attention, Receptive Fields, and Adaptive Spatial Feature Fusion Modules for YOLO (CRAS-YOLO), which integrates CBAM, Receptive Fields Block (RFB), and Adaptively Spatial Feature Fusion (ASFF) modules to enhance both precision and recall. Chen et al. [12] proposed a multi-scale ship detection model for complex scenes, named CSD-YOLO, based on YOLOv7. Their model introduces an Spatial Atrous Shuffle Feature Pyramid Network (SAS-FPN) module that integrates atrous spatial pyramid pooling and shuffle attention to enhance the model's ability to focus on crucial information while disregarding irrelevant data. This approach reduces feature loss for small ships and effectively fuses feature maps across different scales of SAR images, thereby improving detection accuracy. Sun et al. [13] developed Bi-directional Feature Fusion and Angular Classification for YOLO (BiFA-YOLO) with bidirectional feature fusion and angular classification for enhanced accuracy. Guo et al. [14] improved Mobilenet with a lightweight convolutional unit Depthwise Separable Convolution, Batch Normalization, and Activate or Not (ACON) Activation Function (DBA) module and introduced S-Mobilenet for enhanced feature extraction. Zhao et al. [15] proposed a novel SAR ship detection model named Swin-Transformer-Based YOLO Model with Attention Mechanism (ST-YOLOA), which integrates the Swin Transformer network and Coordinate Attention (CA) model within the STCNet backbone to enhance feature extraction and capture global information. They employed the Path Aggregation Network (PANet) path aggregation network with a residual structure for improved global feature extraction and introduced a novel up/down-sampling method to address local interference and semantic information loss. The decoupled detection head improves convergence speed and detection accuracy. Jiang et al. [16] addressed low signal-to-noise ratio and resolution issues with a high-speed and lightweight ship detection algorithm based on YOLOv4. Cai et al. [17] introduced Feature Fusion and Feature Enhancement for YOLO (FE YOLO), enhancing contextual information capturing with an Improved Extended Efficient Layer Aggregation Network (IELAN) module. Zhan et al. [18] presented Edge-Guided Infrared Ship Detection for YOLO (EGISD-YOLO), improving the Cross Stage Partial Network (CSP) module of YOLO for better feature information reusability and addressing image noise with a Deconvolutional Channel Attention (DCA) module. These approaches significantly

enhance accuracy and real-time performance in SAR ship detection. Zhang et al. [19] proposed a multi-scale fusion framework (Swin-PAFF) for SAR target detection, addressing challenges such as strong scattering, indistinct edge contours, multi-scale representation, sparsity, and severe background interference. Their approach leverages the Transformer's global context perception and the feature pyramid structure's multi-layer feature fusion. The proposed method features an end-to-end SAR target detection network with a Transformer backbone and incorporates a Swin Contextual Cross-information Network (Swin-CC) backbone network model that combines Spatial Depthwise Pooling (SDP) and self-attentive mechanisms. Additionally, they introduce a cross-layer fusion neck module (PAFF) to handle multi-scale variations and complex conditions. Wang et al. [20] proposed Fast Feature Pyramid Module (FastPFM), a novel ship detection model designed to address challenges in SAR imaging, such as blurred ship contours, complex backgrounds, and uneven scale distribution. Their approach uses FasterNet as the backbone to enhance computational efficiency and feature extraction. The Feature Bi-level Routing Transformation model (FBM) is employed to gather global feature information and improve focus on target regions. The PFM module collects multi-scale target information by connecting features across stages, and an additional feature fusion layer is introduced to boost small ship detection accuracy.

Despite advancements in SAR ship target detection, challenges remain in complex marine environments, especially in near-coastal areas where small and medium-sized ships are prevalent. The dense distribution and blurred boundaries of these vessels, coupled with background noise from onshore buildings, islands, and vehicles, complicate detection. Moreover, maritime images often suffer from additional noise due to waves, clouds, and sea reflections, further hindering ship detection. While existing algorithms offer some improvements, issues like blurred ship boundaries and dense distributions continue to pose significant challenges.

To address these issues, this paper utilizes the YOLOv8n single-stage target detection algorithm as the base model. YOLOv8n provides real-time performance and high accuracy, featuring a multi-scale detection head suitable for various ship sizes. Enhancements include replacing the model backbone with the SB structure, substituting the model head with the DMDH structure, and incorporating ODConv in both the backbone and neck. Additionally, the LSKNet and MSCA modules are introduced in the neck. Experimental evaluations demonstrate that the proposed model achieves high accuracy in SAR ship target detection and shows improved detection performance.

3 Main Methods

In this section, we first introduce the baseline YOLOv8 framework, followed by the ODConv module, the LSKNet module, and the MSCA module. Finally, we present the SAR-LtYOLOv8 framework, which integrates these modules into a converged network structure.

3.1 General Architecture of YOLOv8

YOLOv8, a single-stage target detection algorithm, is renowned within the YOLO series for its performance [21]. It encompasses four main network architectures: YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l. The overall network architecture of YOLOv8n, depicted in Fig. 1, comprises key components: inputs, backbone network, neck, YOLOv8 head, and outputs. Besides, the details of the modules in the network architecture are illustrated in Fig. 2.

Darknet-53 [22] is used to optimize the backbone network's features, which effectively extract features for high-performance target detection tasks. In YOLOv8, the neck component integrates features using techniques like spatial pyramid pooling and feature pyramid networks [23], feeding into

the YOLOv8 head for predicting bounding boxes, scores, and category probabilities. YOLOv8 employs either anchor-based or anchor-free detection techniques and outputs bounding boxes, category labels, and confidence scores through regression and classification. Optimization of the YOLOv8 architecture includes incorporating CSP [24] and Cross Stage Partial Network Fusion (C2F) [25] modules for enhanced feature fusion. By transitioning to a decoupled head structure, classification and detection can be separated, which allows for an anchorless point-based approach for better flexibility and performance, particularly in small target detection on SAR ships. Additionally, YOLOv8 improves loss calculation and label assignment strategies using VariFocal Loss (VFL Loss), Distribution Focal Loss (DFL Loss), and Complete Intersection Over Union Loss (CIOU Loss) [26], while employing the Task-Aligned Assigner [27] for efficient learning of target features.

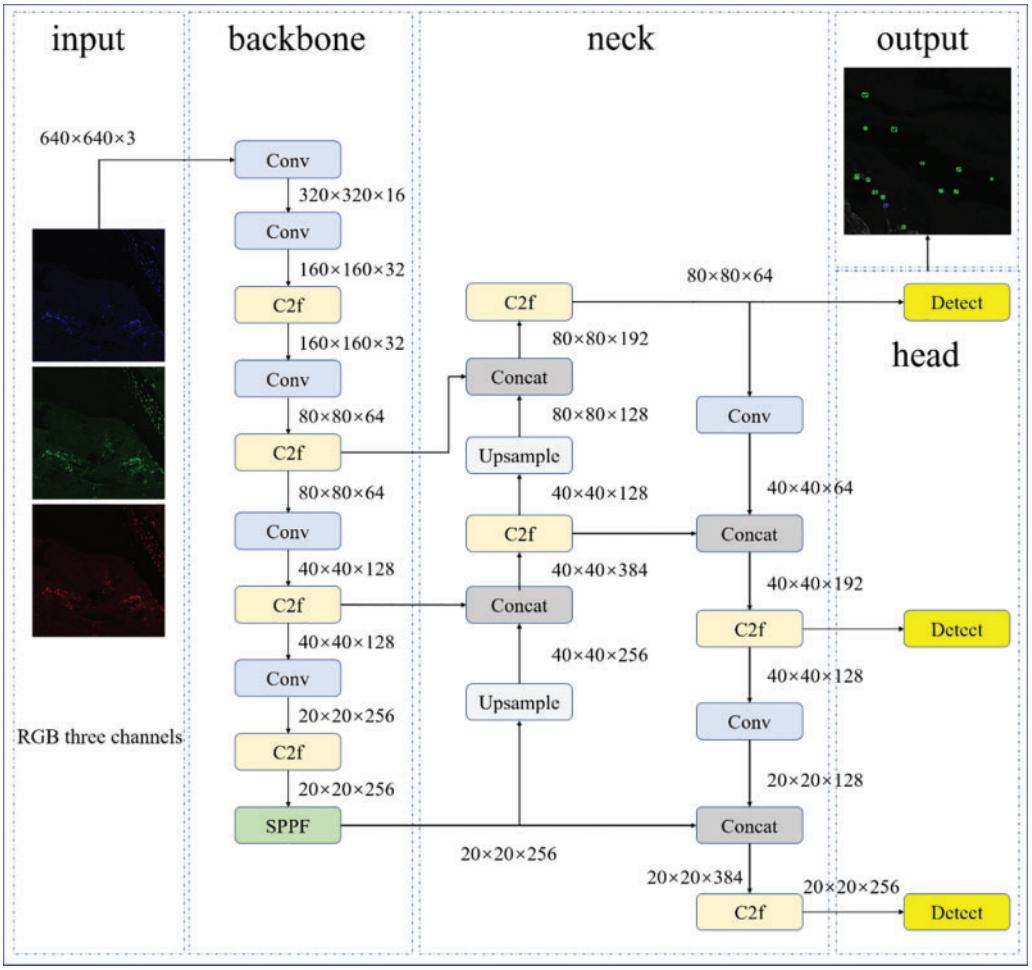

Figure 1: General framework of YOLOv8n

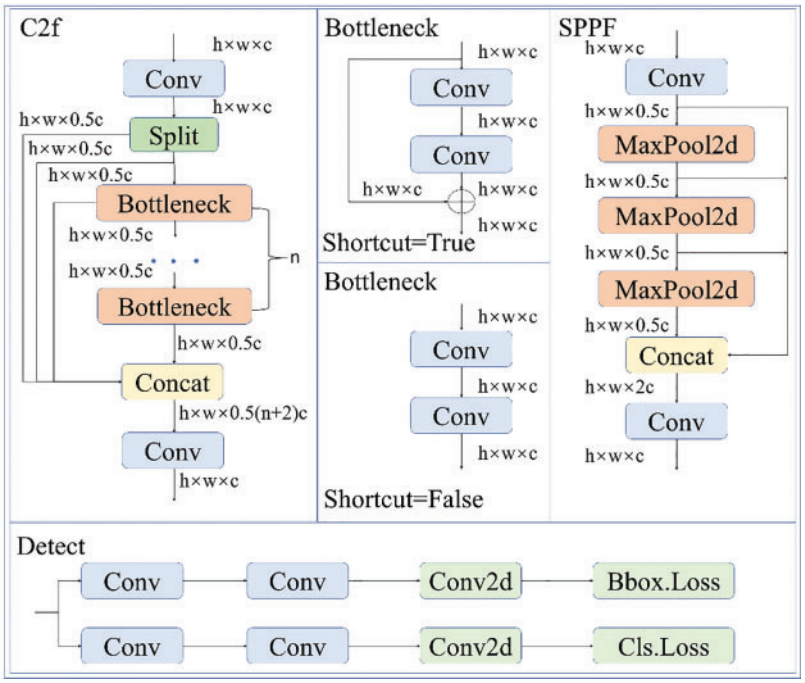

Figure 2: Models of YOLOv8n

3.2 General Architecture of SAR-LtYOLOv8

In this section, we present the SAR-LtYOLOv8 network architecture, an enhanced model based on YOLOv8n. Modifications include adopting the SB and DMDH structures, replacing Conv with ODConv in the trunk and neck, integrating LSKNet into the neck, and introducing the MSCA module. The specific architecture of SAR-LtYOLOv8n is shown in Fig. 3.

Luo et al. [28] showed that shallow features with higher spatial resolution effectively capture fine structures and local features, improving the detection of small and dense targets. The BiFA-YOLO [13], ST-YOLOA [15], and EGISD-YOLO [18] models all investigate modifications to the neck structure of YOLO to enhance feature extraction. These models alter the neck structure to obtain information at different scales, reduce local interference and semantic loss, and improve the overall ability to capture global information. Therefore, optimizing the neck structure of YOLOv8 can further enhance feature fusion and information transmission processes. Such improvements not only boost the detection capabilities for small ship targets and contour details but also improve the accuracy of multi-scale ship target recognition in nearshore scenarios. By finely tuning the neck structure, the model can better capture and leverage features from different layers. We redesigned the network by reducing its depth, removing redundant detection heads, and incorporating the SB and DMDH structures to replace the backbone and head of YOLOv8. SAR-LtYOLOv8 has only two detection heads and omits a layer of Conv and C2f modules in its trunk. Using the SB and DMDH structures reduces downsampling by one, doubling the feature map's width and height, which aids in small object detection.

We replaced the Conv layers in the trunk and neck with ODConv to speed up the model and reduce computational complexity. We also added the LSKNet and MSCA modules to the neck, which combine local, global, and channel attention mechanisms to extract features at various spatial scales. Local attention focuses on the details within the target area, enhancing detection accuracy

by capturing the structure and texture of the vessel. Global attention, on the other hand, aids in understanding the broader context of the scene surrounding the target. Furthermore, the channel attention mechanism dynamically adjusts the channel weights to prioritize target-related features, such as high reflectivity or varying shapes, thereby improving recognition and robustness. The MSCA module is particularly important for small target detection in SAR images, significantly enhancing the model's feature extraction and selection capabilities and providing an effective solution for small target detection in SAR ship images.

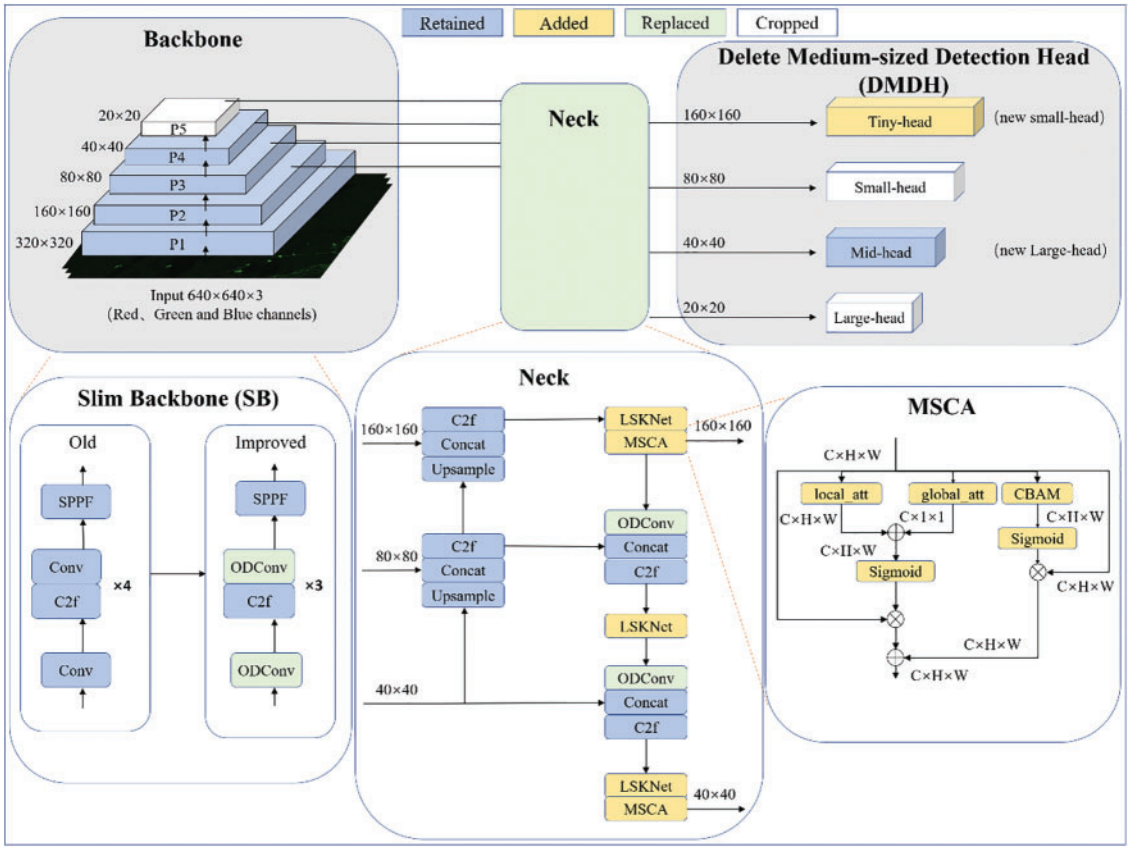

Figure 3: General framework of SAR-LtYOLOv8n

SAR images, typically captured by high-altitude satellites, depict vessels at very small scales. In the original YOLOv8 model, the backbone performs five downsampling operations, generating five feature maps (P1, P2, P3, P4, and P5), where Pi corresponds to a resolution of 1/2 i of the original image. The neck network uses both top-down and bottom-up paths to combine features of different sizes. However, detection mostly happens at the P3, P4, and P5 layers, which have 80 × 80, 40 × 40, and 20 × 20 feature map resolutions, respectively. However, for small vessels in SAR datasets, many of which are smaller than 20 × 20 pixels, significant feature information is lost during downsampling, making it challenging for the P3 layer to perform high-resolution detection. To address the aforementioned issue, we modified the input of the small detection head to use the P2 layer feature map. The new input feature map for the small detection head has a resolution of 160 × 160 pixels, capturing richer low-level feature information that significantly enhances the model's ability to detect small objects. In fact, based on the original model's detection head input scales, we only removed

the medium detection head while retaining the small and large detection heads. However, due to the changes in feature map sizes in the new model, the input sizes for the small and large detection heads have doubled. Essentially, the new model adds an extra tiny-detection head, retains the medium detection head, and removes the original small and large detection heads. Finally, the specific details of the newly added modules are provided in the following sections.

3.3 ODConv Module

ODConv [29] is a "full-dimensional" dynamic convolution method that enhances dynamic behavior across spatial, input channel, and output channel dimensions. In this study, the Conv module in both the trunk and neck of the YOLOv8 model is replaced with ODConv, differing from traditional convolution in several significant ways. The detailed computation process of the ODConv module is illustrated in Fig. 4.

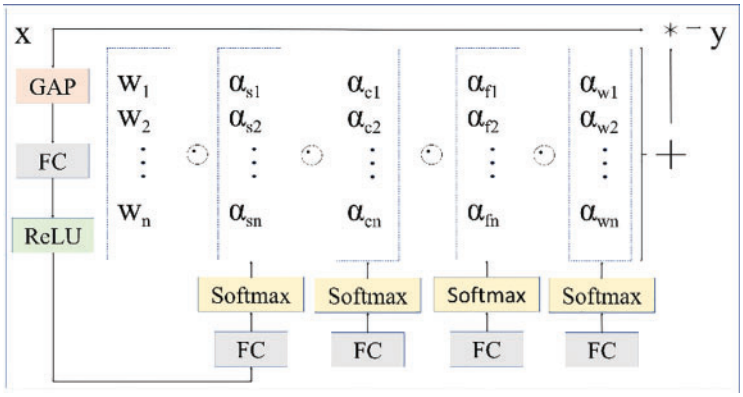

Figure 4: General framework of ODConv

ODConv adjusts convolution kernel weights dynamically based on input data features, which improves adaptability and efficiency across spatial, input, and output channel dimensions. This comprehensive feature capture boosts performance while reducing computation time, thus accelerating both training and inference. By utilizing a single kernel, ODConv achieves comparable or superior performance to multi-kernel dynamic convolutions, significantly cutting down on additional parameters and redundant operations. The formula for ODConv is provided in Eq. (1).

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \ldots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n\right) * x \tag{1}$$

In the formula, $\alpha_{wi} \in R$ represents the attention scalar of the convolution kernel W_i. The terms $\alpha_{si} \in R^{k \times k}$, $\alpha_{ci} \in R^{cin}$, and $\alpha_{fi} \in R^{cout}$ indicate the introduced attention points, which correspond to the spatial dimension of the kernel space, the input channel dimension, and the output channel dimension. The symbol $\odot$ denotes the element-wise multiplication operation along different dimensions of the kernel space.

3.4 LSKNet Module

LSKNet is a deep neural network model tailored for addressing challenges in image processing tasks, showing notable performance enhancements in tasks like image classification and semantic segmentation [30]. The overall network architecture is depicted in Fig. 5.

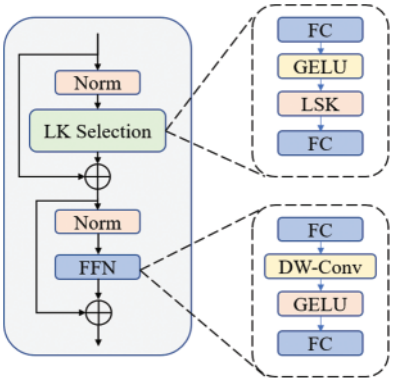

Figure 5: General framework of LSKNet

LSKNet is a backbone network comprising LSKNet blocks, each composed of a large kernel selection sub-block and a Feed-Forward Network (FFN) sub-block. The Large Kernel Selection (LK Selection) sub-block dynamically adjusts the network's receptive domain with large kernel convolutions and a spatial kernel selection mechanism, enhancing multi-scale feature extraction. The FFN sub-block refines features through channel mixing, integrating Fully Connected (FC) layers, Depthwise Convolution (DW-Conv), Gaussian Error Linear Unit (GELU) activation, and another fully connected layer.

3.5 MSCA Module

The MSCA module is a novel fusion module inspired by the MS-CAM [9] and the CBAM [31]. Firstly, MS-CAM is an improved category activation mapping technique that analyzes features at different scales to capture target information in images. Secondly, CBAM is an attention mechanism module for convolutional neural net works that enhances feature representation by processing channel attention and spatial attention in parallel.

By fusing the principles of MS-CAM and CBAM, MSCA module aims to realize an integrated feature recalibration mechanism that covers both channel-level and spatial-level attention. This innovative approach takes full advantage of the multi scale channel attention of MS-CAM and the spatial attention of CBAM to generate synergistic effects, improving feature representation at different scales. MSCA module provides a sophisticated solution for enhancing multi-scale feature representation and optimizing feature maps to improve the performance of various computer vision tasks. The overall MSCA network architecture is shown in Fig. 6.

The structure of the local and global attention modules is depicted in Fig. 7. The primary distinction between the global and local attention modules lies in their approach to feature extraction. The global attention module begins with a global average pooling operation to aggregate comprehensive global information from the entire feature map. In contrast, the local attention module focuses on capturing fine-grained spatial details within the input feature map. By combining these two attention mechanisms, the model effectively gains a nuanced understanding of features at both local and global scales. The local attention module enhances the ability to discern intricate spatial patterns, while the global attention module provides a broad contextual overview. This synergistic integration of local and global perspectives allows for a more robust representation of features across different regions and scales of the input image, thereby improving the model's overall performance and accuracy.

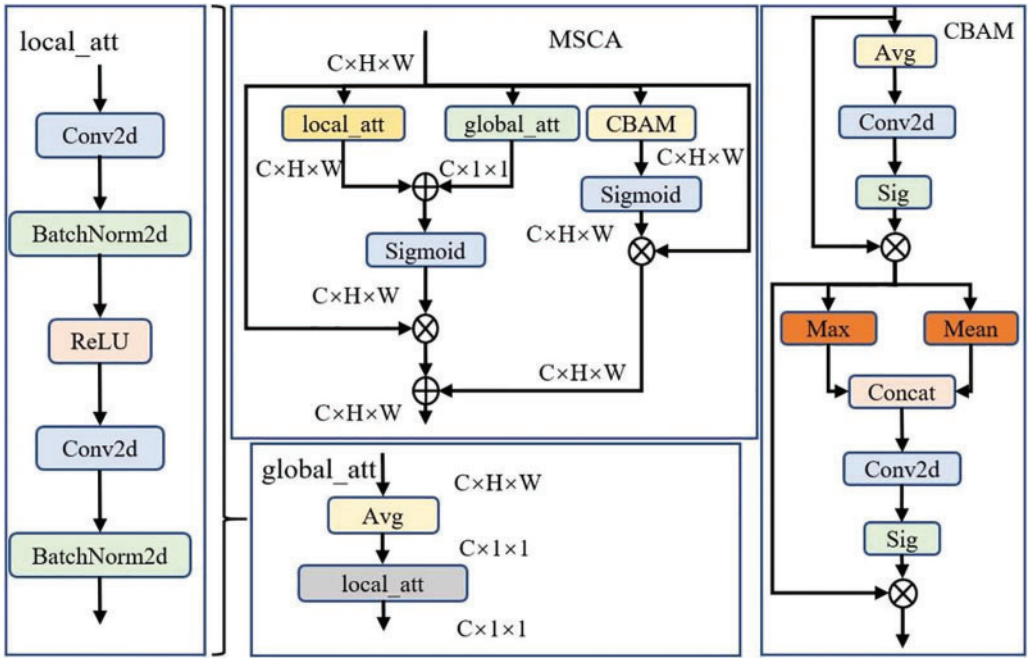

Figure 6: General framework of MSCA

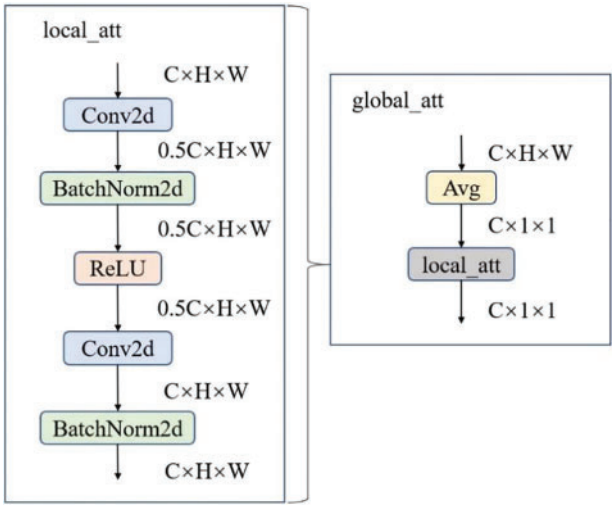

Figure 7: Structure of local attention module and global attention module

The local attention module utilizes pointwise convolution to aggregate local channel context, allowing interactions only within each spatial location. It employs convolution operations considering only the channels at each spatial position, promoting interactions among local features. The specific representation of $L(X)$ is as follows:

$$L(X) = BN\left(Conv_2\left(\sigma\left(BN\left(Conv_1(X)\right)\right)\right)\right) \tag{2}$$

Here, W represents the width, H represents the height, and C represents the number of channels. $Conv_1$ has a convolution kernel size of $C \times \frac{c}{r} \times 1 \times 1$, and $Conv_2$ has a convolution kernel size of $\frac{c}{r} \times C \times 1 \times 1$. Additionally, r denotes the channel scaling ratio. Finally, σ represents the Sigmoid activation function.

In contrast to the local attention module, the input to the global attention module first undergoes an average pooling operation. This global attention mechanism enhances the model's understanding of the context by combining the global information from the input feature map with local features. The specific representation of $G(X)$ is as follows:

$$G(X) = BN\left(Conv_2\left(\sigma\left(BN\left(Conv_1\left(AvgPool\left(X\right)\right)\right)\right)\right)\right) \tag{3}$$

The CBAM module combines channel attention and spatial attention. The specific structure of the CBAM module is shown in Fig. 8.

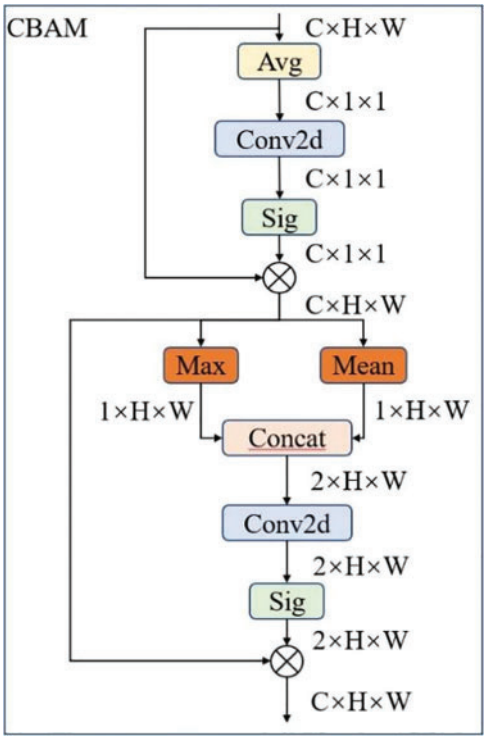

Figure 8: Structure of local attention module and global attention module

As shown in the figure, the structure of the model follows a sequence where channel attention comes first, followed by spatial attention. The formula for the channel attention module $C(X)$ is given in Eq. (4).

$$C(X) = \sigma\left(Conv_3\left(AvgPool\left(X\right)\right)\right) \tag{4}$$

Here, the convolution kernel size of $Conv_3$ is $C \times C \times 1 \times 1$, and σ represents the Sigmoid activation function. The specific formula for $CBAM(X)$ is given by Eq. (5).

$$CBAM(X) = X \times \sigma\left(Conv_3\left(Cat\left(Mean\left(C\left(X\right)\right), Max\left(C\left(X\right)\right)\right)\right)\right) \tag{5}$$

Here, *Cat* denotes the concatenation of tensors along the first dimension, while *Mean* and *Max* functions respectively compute the mean and maximum values of the tensor. The channel attention module operates by performing global average pooling and max pooling operations on the feature map, resulting in two channel descriptors representing global average features and maximum features, respectively. These descriptors are then passed through fully connected layers and activation functions to generate channel attention weights.

The final MSCA module integrates local information, global information, spatial information, and channel information of the feature map, enhancing attention to multi-scale feature information and improving the model's detection performance on multi-scale targets. The specific formula for $MSCA(X)$ is given by Eq. (6).

$$MSCA(X) = X \times \sigma(CBAM(X)) + X \times \sigma(G(X) + L(X)) \tag{6}$$

In SAR ship target detection, traditional feature extraction methods often falter due to challenges such as complex backgrounds, low contrast, and significant interference. To overcome these issues and achieve a thorough understanding of image features across different scales and perspectives, the MSCA module is introduced. This module incorporates local, global, and channel attention mechanisms to extract features at various spatial scales. Local attention zeroes in on details within the target region, enhancing detection accuracy by capturing ship structure and texture. Global attention, on the other hand, helps in understanding the scene's broader context surrounding the target. Additionally, the channel attention mechanism dynamically adjusts channel weights to prioritize features relevant to the target, such as high reflectivity or distinct shapes, thereby improving recognition rates and robustness. Essential for detecting small targets in SAR imagery, the MSCA module significantly enhances the model's ability to extract and select features, providing an effective solution for small target detection in SAR ship images.

3.6 Loss Function

The loss function measures the difference between the model's predictions and actual labels, guiding parameter adjustments. In YOLOv8, it includes classification loss using BCE Loss [32] and regression loss using DFL Loss [33] and CIOU Loss [34]. The bounding box regression loss, assessed with CIOU by default, measures the difference between predicted and true bounding boxes, as shown in Eq. (7).

$$L_{CIoU} = 1 - IoU - \frac{\rho^2\left(B_{gt}, B_{pre}\right)}{C^2} + \alpha v \tag{7}$$

Here, IoU represents the Intersection over Union between the predicted and ground truth boxes, while $\rho^2\left(B_{gt}, B_{pre}\right)$ denotes the Euclidean distance between their center points. C is the diagonal length of the smallest enclosing region for both boxes. α is a weight parameter balancing the impact of aspect ratios, with its formula shown in Eq. (8).

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{8}$$

Here, v measures aspect ratio consistency, with its formula shown in Eq. (9).

$$v = \frac{v}{\pi^2}(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w_{pre}}{h_{pre}}) \tag{9}$$

Here, w and h represent the width and height of the predicted box, while w_{gt} and h_{gt} denote the width and height of the ground truth box, respectively.

4 Experimentation and Analysis

In this section, we present an overview of the two datasets employed, detailing the experimental parameter configurations and evaluation metrics. Subsequently, we conduct a comprehensive series of ablation experiments and comparative analyses, presenting the results of the comparative experiments visually to demonstrate the effectiveness of the proposed model.

4.1 Dataset Introduction

The HRSID dataset is a high-resolution SAR ship dataset that has garnered significant attention since its release [35]. It features high-resolution SAR images from various geographic locations and marine environments, captured using different radar platforms and polarizations. Unlike traditional optical images, SAR images offer reliable ship detection across a range of weather and lighting conditions, making them particularly valuable for marine surveillance, rescue operations, and similar applications. The HRSID dataset includes ship targets of varying sizes and types, with small ships accounting for 54.5%, medium ships for 43.5%, and large ships for 2% of the dataset. For this study, the dataset is divided into three subsets: 70% for training, 10% for validation, and 20% for testing.

Additionally, the SSDD dataset [36] is employed in this study. The SSDD dataset is designed to provide ship images in a wide range of marine environments, from sunny weather to rough sea conditions, and from simple marine scenes to complex harbor areas. It includes various ship detection scenarios, featuring ships ranging from small fishing boats to large cargo vessels. This diversity makes the SSDD dataset particularly suited for evaluating the robustness and generalizability of ship detection algorithms. Detailed information about the datasets is provided in Table 1.

Table 1: Parameter information of HRSID and SSDD datasets

Parameters	HRSID	SSDD
Data source	Sentinel-1B, TerraSAR-X, TanDem	RadarSat-2, TerraSAR-X, Sentinel-1
Polarization method	Horizontal-Horizontal (HH), Vertical-Vertical (VV), Horizontal-Vertical (HV)	HH, VV, Vertical-Horizontal (VH), HV
Shooting locations	Houston, St. Paul, etc., USA	Yantai, China; Visakhapatnam, India
Resolution (m)	0.5–3	1–15
Image size (pixels)	800 × 800	Approx. 500 × 500
Number of training set images	4482	928
Number of test set images	1121	232
Total number of ships	16,951	2456

4.2 Experimental Parameters Introduction

In this study, we conducted deep learning experiments on a computer running the Ubuntu 18.04 operating system. Detailed configuration information is presented in Table 2.

Table 2: Configuration details for the experimental setup

Name	Configuration
Platform	PyTorch 1.8.1, Python 3.8, Cuda 11.1
GPU	RTX 3080 (10 GB) $*$ 1
Processor	12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz
Operating system	Ubuntu 18.04
RAM	40 G

4.3 Evaluation Metrics Introduction

In this study, we use the Common Objects in Context (COCO) evaluation metrics, a standardized set of metrics designed to assess the performance of object detection models on datasets [37]. The primary metrics include Average Precision (AP) and Average Recall (AR), which are calculated at different Intersection Over Union (IoU) thresholds and object scales. AP combines precision and recall to evaluate the model's detection accuracy across different categories. These metrics provide a comprehensive assessment of the model's performance in detecting objects of various sizes and levels of overlap with ground truth annotations.

We commonly use Precision and Recall in object detection tasks to evaluate the model's performance. Precision refers to the proportion of true positive samples among all detected positive samples, while Recall refers to the proportion of detected positive samples among all actual positive samples. The formulas for Precision and Recall are shown in Eqs. (10) and (11), respectively.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

When calculating AP, we first determine the Precision-Recall curve for each category and average these curves. Specifically, we compute the area under each category's Precision-Recall curve, referred to as the Area Under Curve (AUC), and then average the AUC values across all categories to obtain the AP. Different IoU thresholds correspond to different AP metrics, including AP_{50}, AP_{75}, AP_{S}, AP_{M}, and AP_{L}. AP is calculated as the average AP across IoU thresholds ranging from 0.5 to 0.95, with a step size of 0.05. This metric is referred to as Mean Average Precision (MAP). The MAP formula is shown below:

$$\mathrm{MAP} = \frac{1}{N} \int_{i=1}^{N} AP_i \tag{12}$$

AP_{50} and AP_{75} represent the AP values calculated with IoU thresholds of 0.5 and 0.75, respectively. AP_S, AP_M, and AP_L are used to measure the average precision for detecting small, medium, and large targets.

4.4 Ablation Experiments

4.4.1 The Influence of SB on the Experimental Evaluation Index

This paper focuses on enhancing the detection of small ship targets in the HRSID SAR ship dataset. Shallow feature maps, with their smaller receptive fields, are particularly sensitive to local details and excel at capturing the characteristics of small targets. In object detection tasks, shallow feature maps are commonly utilized for locating and identifying small objects. Consequently, this paper emphasizes improving the processing of shallow feature maps within the network.

To achieve this goal, we reduced one layer of ODConv and the C2f module in the backbone, thereby halving the number of channels and doubling the image's height and width. We named this new backbone structure SB, and the resulting model is called YOLOv8n-SB.

We conducted experiments using both YOLOv8n-SB and YOLOv8n on the HRSID dataset. The experimental results are shown in Table 3.

Table 3: The experimental results of YOLOv8n-SB

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
Baseline	65.1	89.3	74.3	54.6	**79.0**	28.2	300,5843
YOLOv8n-SB	**67.1**	**91.8**	**76.8**	**56.9**	77.5	**45.0**	**907,155**

Note: The bold values represent the best results in the comparison data.

During the experiments, we kept all other components unchanged and only reduced one layer each of ODConv and C2f in the backbone. We then compared the modified model, YOLOv8n-SB, with the original YOLOv8n. As shown in Table 3, the results indicate a significant reduction in the number of parameters for YOLOv8n-SB. Additionally, there were increases of 2% in AP, AP_{50}, AP_{75}, and AP_S, and a remarkable increase of 16.8% in AP_L. However, AP_m decreased by 1.5%, which is primarily due to the increase in feature map size to twice its original dimensions after reducing the number of layers in the backbone. Although the higher resolution of the feature maps helps capture local information for small targets, thereby improving the detection accuracy for small targets, the reduced receptive field makes feature extraction for medium-sized targets less comprehensive. Moreover, medium-sized ship targets have relatively fewer contours and details, so the higher resolution offers limited improvement for medium-sized target detection. Thus, the negative impact of the smaller receptive field outweighs the positive impact of higher resolution, leading to decreased detection accuracy for medium-sized targets. In contrast, for large targets, which have a larger size and richer contours and details, the higher resolution still contributes to better detection accuracy despite the reduced receptive field. The experimental results indicate that the simplified backbone module positively affects the model's performance. These results demonstrate that the simplified backbone module positively enhances the model's performance.

4.4.2 The Influence of DMDH on the Experimental Evaluation Index

In this study, we kept other structures unchanged while reducing the detection heads in the YOLOv8-SB architecture. We successfully decreased the model's parameter count by removing unnecessary detection heads. This paper aims to improve the detection accuracy of small SAR ship targets. Therefore, we experimented with removing the medium-sized and large-sized detection heads. The head structure with the medium-sized detection head removed is named DMDH, and the head structure with the large-sized detection head removed is named Delete Large-sized Detection Head (DLDH). The experimental results are shown in Table 4.

Table 4: Experimental results of different head

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
YOLOv8n-SB	67.1	91.8	76.8	56.9	77.5	45.0	907,155
YOLOv8n-SB-DMDH	**68.7**	**92.3**	**79.3**	**58.7**	**78.7**	**48.7**	801,394
YOLOv8n-SB-DLDH	67.9	92.0	78.5	58.0	78.1	39.8	**622,834**

Note: The bold values represent the best results in the comparison data.

Table 5: Experimental results of different attention modules

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
Baseline	65.1	89.3	74.3	54.6	79.0	28.2	**300,5843**
YOLOv8n-cbam	65.8	89.5	75.1	55.2	**79.5**	26.9	307,5991
YOLOv8n-ca	64.6	89.1	72.6	54.1	78.8	24.9	305,8947
YOLOv8n-se	66.0	89.9	75.2	**55.2**	78.9	32.1	304,0659
YOLOv8n-ms_cam	66.0	89.7	**75.3**	55.0	79.3	30.3	307,7875
YOLOv8n-MSCA	**66.5**	**90.5**	75.2	55.0	79.0	**39.7**	321,8135

Note: The bold values represent the best results in the comparison data.

Based on the results presented in Table 5, removing the large detection head resulted in a decrease in AP_L. This outcome is attributed to the fact that the large detection head is specifically designed to process deep feature maps, which have a larger receptive field and provide more comprehensive global information. This design enables it to effectively capture the overall features of larger objects within the image. By leveraging these deep feature maps, the large detection head enhances the model's ability to detect large targets. In YOLOv8, Non-Maximum Suppression (NMS) is employed to consolidate detection results by eliminating overlapping bounding boxes and retaining the most probable targets. In the HRSID dataset, images of small ships are predominant, with large ships being relatively rare and medium-sized targets also infrequent. Additionally, the sizes of medium-sized and small ships are quite similar. As a result, the effectiveness of the medium detection head is limited, and its receptive field overlaps with that of small targets, potentially leading to redundant detections. Removing the medium detection head reduces such redundancy, optimizing the NMS process and

enhancing detection accuracy. Furthermore, by eliminating the medium detection head, the model can better focus on small and large targets, improving the detection performance for small targets and ultimately leading to an increase in the AP_m metric. We plan to further optimize the results by adjusting the detection head parameters, enhancing the dataset, and reducing redundant detections.

4.4.3 The Influence of MSCA on the Experimental Evaluation Index

We conducted individual ablation experiments on the HRSID dataset to validate the effectiveness of each module. Specifically, we investigated the impact of the MSCA, ODConv, and LSKNet modules and their combinations on the model's accuracy.

Wang et al. [1] proposed the Neural Architecture Search-based YOLOX (NAS-YOLOX) model, and Zhao et al. [11] proposed the CRAS-YOLO model. Both studies explore the role of attention mechanism modules within feature pyramid networks, modifying these modules to obtain richer feature information and achieve multi-scale feature fusion. In the experiments, we integrated the MSCA module into the neck component of the baseline model while keeping other parts unchanged. We then compared these results with experiments utilizing various other mainstream attention mechanisms. The experiment results are presented in Table 5.

Clearly, integrating the MSCA module leads to an overall improvement in model accuracy. Additionally, there is a notable 11.5% enhancement in AP_L. These results demonstrate that the MSCA module actively enhances model performance. Due to the integration of multi-scale channel attention and spatial attention in the MSCA module, the YOLOv8n-MSCA model can capture the fine details and features of maritime targets with greater precision. Specifically, the MSCA module performs fine-grained recalibration of features across different scales, significantly enhancing the model's accuracy in target recognition and localization. This improvement substantially boosts the YOLOv8n-MSCA's performance in detecting maritime targets in complex environments and across various scales, aiding in more accurate identification and classification of these targets, and significantly increasing the model's reliability and effectiveness in practical applications.

4.4.4 The Influence of ODConv on the Experimental Evaluation Index

In the experiments, we replaced the Conv modules in the backbone and neck components of the YOLOv8-SB model with the ODConv module while keeping other parts unchanged. Additionally, we conducted experiments where both the Conv modules in the backbone and neck components were replaced with ODConv. Subsequently, we compared the performance of the modified models in the experiments, and the results are presented in Table 6.

Table 6: Experimental results of ODConv

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
Baseline	65.1	89.3	74.3	54.6	**79.0**	28.2	3,005,843
YOLOv8n-SB	67.1	91.8	76.8	56.9	77.5	45.0	**907,155**
YOLOv8n-SB-backbone	68.5	91.9	79.3	59.3	78.8	38.4	918,222
YOLOv8n-SB-neck	67.6	92.0	78.6	58.1	77.9	38.6	914,119
YOLOv8n-SB-all	**69.1**	**92.5**	**79.7**	**59.8**	78.9	**43.5**	925,186

Note: The bold values represent the best results in the comparison data.

The model achieves the highest AP values when both the Conv modules in the backbone and head components are replaced with ODConv. Furthermore, although the AP_L of YOLOv8n-SB-all decreased by 1.5% compared to YOLOv8n-SB, other metrics showed improvements. These results indicate that replacing the Conv modules in the backbone and neck components of the YOLOv8n-SB model with ODConv can maximize the model's performance.

4.4.5 The Influence of LSKNet on the Experimental Evaluation Index

In the experiments, we integrated the LSKNet module into the neck component of the baseline model while keeping other parts unchanged. Then, we compared these results with experiments using the YOLOv8n model. The experiment results are presented in Table 7.

Table 7: Experimental results of LSKNet

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
Baseline	65.1	89.3	74.3	**54.6**	79.0	28.2	**300,5843**
YOLOv8n-LSKNet	**65.6**	**89.6**	**74.6**	54.5	**79.5**	**29.0**	334,3333

Note: The bold values represent the best results in the comparison data.

We incorporate the LSKNet module, which leads to an overall improvement in model accuracy. These results demonstrate that the LSKNet module actively contributes to enhancing model performance. The LSKNet model improves the identification of specific ship features and shapes by optimizing the extraction of spatial features. It enhances the model's ability to recognize ship contours, sizes, and orientations, particularly in complex environments. Additionally, LSKNet advances feature representation and multi-scale information fusion, leading to more accurate detection of ships across various scales.

4.4.6 The Influence of Overall Improvement on the Experimental Evaluation Index

We conducted ablation experiments to evaluate the impact of different module combinations on model accuracy. As detailed in Table 8, while keeping the YOLOv8n architecture unchanged, we carried out the following experiments: In the first group, we replaced the model's backbone with the SB structure. In the second group, we swapped the model's neck for the DMDH structure. The third group involved substituting Conv with ODConv in both the backbone and head of the model. In the fourth group, we incorporated the LSKNet module into the model's neck. Finally, in the fifth group, we added the MSCA module to the neck.

Table 8: The ablation experiments

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
YOLOv8n	65.1	89.3	74.3	54.6	79.0	28.2	3,005,843
SB;	67.1	91.8	76.8	56.9	77.5	45.0	907,155
SB; DMDH	68.7	92.3	79.3	58.7	78.7	48.7	**801,394**
SB; DMDH ODConv	69.2	92.3	79.8	59.1	79.4	**51.1**	819,425

(Continued)

Table 8 (continued)

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
SB; DMDH ODConv; LSKNet	69.2	92.4	79.3	59.4	**79.5**	49.1	913,203
SB; DMDH ODConv; LSKNet MSCA	**69.5**	**92.7**	**80.2**	**60.0**	79.1	49.1	967,223

Note: The bold values represent the best results in the comparison data.

Integrating all modules into the SAR-LtYOLOv8 model achieved significant performance gains on the HRSID dataset. The parameter count reduced by 67.8%, AP_{50} increased by 3.4%, $AP_{50–95}$ increased by 4.4%, AP_S increased by 5.4%, AP_M increased by 0.1%, and AP_L increased by 20.9%.

4.5 Comparative Experiments

The research's experiments were conducted on the HRSID and SSDD datasets, comparing our proposed method against prominent SAR ship detection techniques such as YOLOv8 [38], Task-aligned One-stage Object Detection (TOOD) [39], Adaptive Training Sample Selection (ATSS) [40], VarifocalNet (VFNet) [41], YOLOv5n [42], YOLOv6n [43], Swin-PAFF [19], NAS-YOLOX [1], and FastPFM [20]. The results are presented in Tables 9 and 10.

Table 9: Comparison of quantitative evaluation indexes on HRSID

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline (2023)	65.1	89.3	74.3	54.6	79.0	28.2
YOLOv3 (2018)	56.1	78.9	63.8	36.4	75.7	47.4
TOOD (2021)	64.6	88.8	73.1	51.4	78.4	40.2
ATSS (2020)	58.7	84.8	66.4	42.2	75.4	38.8
VFNet (2021)	61.7	84.5	69.8	45.0	78.8	45.1
Retinanet (2018)	48.6	75.2	54.9	28.1	69.1	28.4
PAFPN (2019)	62.5	83.5	72.1	47.5	77.9	45.3
YOLOv5n (2020)	66.0	89.8	75.3	54.6	79.0	33.8
YOLOv6n (2023)	62.1	86.3	69.4	50.3	78.7	19.4
Swin-PAFF (2023)	64.6	91.3	73.3	65.7	67.9	–
NAS-YOLOX (2023)	63.9	91.1	71.9	65.2	68.6	34.1
FastPFM (2024)	66.0	92.1	74.8	**68.6**	69.3	31.5
SAR-LtYOLOv8 (ours)	**69.5**	**92.7**	**80.2**	60.0	**79.1**	**49.1**

Note: The bold values represent the best results in the comparison data.

Table 10: Comparison of quantitative evaluation indexes on SSDD

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline (2023)	68.5	96.6	81.1	65.3	74.8	65.7
YOLOv3 (2018)	65.2	94.3	77.9	60.4	73.1	**72.3**
TOOD (2021)	62.5	94.2	73.7	62.5	65.2	36.8
ATSS (2020)	62.5	95.7	73.7	61.5	66.8	50.1
VFNet (2021)	60.1	92.3	73.7	60.2	62.9	46.0
Retinanet (2018)	43.6	82.6	42.8	45.0	43.7	18.5
PAFPN (2019)	65.7	94.9	77.7	64.8	68.4	57.2
YOLOv5n (2020)	68.2	96.7	82.4	65.7	73.2	66.0
YOLOv6n (2023)	68.2	96.8	**84.1**	65.9	73.6	61.8
Swin-PAFF (2023)	46.4	80.3	47.6	39.7	39.1	–
NAS-YOLOX (2023)	–	–	–	–	–	–
FastPFM (2024)	50.2	83.1	52.0	52.1	44.3	7.3
SAR-LtYOLOv8 (ours)	**68.9**	**97.1**	83.0	**67.0**	**72.4**	65.7

Note: The bold values represent the best results in the comparison data.

The experimental results unequivocally demonstrate that SAR-LtYOLOv8 outperforms others, achieving the highest AP values on both the HRSID and SSDD datasets. This highlights the model's exceptional performance, robustness, and remarkable generalization capabilities across diverse test image datasets.

This research further divided the SAR ship dataset SSDD into offshore and inshore subsets to verify the model's detection capabilities in different scenarios. The results are presented in Table 11. The results indicate that the model achieves higher precision in detecting ship targets in offshore scenarios due to the relatively simple background and fewer marine background interference factors in offshore scenes. Conversely, inshore scenes are characterized by more complex backgrounds with the presence of coastlines and buildings, resulting in stronger interference, thus leading to lower precision in ship target detection. The experimental results demonstrate that SAR-LtYOLOv8 achieves higher AP_S and AP_L values than other models. Its AP values are comparable to those of YOLOv6n, which performs well. With the AP metrics largely consistent, SAR-LtYOLOv8 achieves AP_S of 59.0% and 70.3% and AP_L of 62.6% and 70.7% in inshore and offshore scenarios, respectively. Therefore, SAR-LtYOLOv8, as developed in this paper, demonstrates good generalization performance on the SSDD dataset, enhancing the model's ability to detect small SAR ship targets.

Table 11: Ship detection in offshore and inshore scenarios of SSDD

Method	Scenarios	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline (2023)	Inshore	59.0	90.0	64.2	56.0	**65.8**	57.3
	Offshore	72.1	98.4	87.7	69.3	77.0	76.6
YOLOv3 (2018)	Inshore	53.7	83	61.7	47.2	64.1	**64.2**
	Offshore	69.4	97.8	84.0	65.9	74.6	**80.8**

(Continued)

Table 11 (continued)

Method	Scenarios	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
TOOD (2021)	Inshore	47.8	80.1	50.4	48.1	51.1	35.8
	Offshore	67.7	98.2	81.8	67.9	69.6	42.0
ATSS (2020)	Inshore	48.4	83.9	48.9	48.6	50.1	43.7
	Offshore	67.6	**98.5**	83.3	66.2	72.0	58.5
VFNet (2021)	Inshore	44.1	75.6	46.0	44.5	46.2	35.4
	Offshore	66.0	97.5	84.1	66.0	68.5	56.5
Retinanet (2018)	Inshore	23.5	52.8	19.3	26.3	22.7	18.4
	Offshore	52.0	93.6	53.2	52.6	53.0	21.9
YOLOv5n (2020)	Inshore	**59.6**	90.9	69.6	57.9	64.1	62.0
	Offshore	71.3	98.4	86.9	68.9	75.3	71.1
YOLOv6n (2023)	Inshore	59.2	**90.9**	**70.8**	57.7	64.4	53.6
	Offshore	71.0	98.4	88.1	68.8	75.6	72.4
Swin-PAFF (2023)	Inshore	37.0	60.3	42.6	40.9	31.1	–
	Offshore	50.7	87.7	50.8	52.3	47.6	–
FastPFM (2024)	Inshore	43.2	68.1	48.1	49.7	32.2	12.9
	Offshore	–	–	–	–	–	–
SAR-LtYOLOv8	Inshore	58.9	90.5	69.8	**59.0**	59.7	62.6
	Offshore	**72.7**	98.4	**88.3**	**70.3**	**77.1**	70.7

Note: The bold values represent the best results in the comparison data.

4.6 Visualization Result Verification and Analysis

We performed image predictions on the HRSID dataset and visually compared the results. Our method demonstrated a significant improvement over other approaches. We randomly selected three images from both inshore and offshore areas for the prediction box annotations. As illustrated in Fig. 9, the SAR-LtYOLOv8 model markedly enhanced the detection of small ship targets across various scenarios.

As shown in Fig. 10, we present the detection results of several popular models on the HRSID dataset for ship targets. Blue boxes indicate missed detection targets, red boxes indicate false detection targets, and green boxes indicate correctly detected targets. The ground truth boxes and the predicted boxes by the SAR-LtYOLOv8 model are shown in Fig. 10.

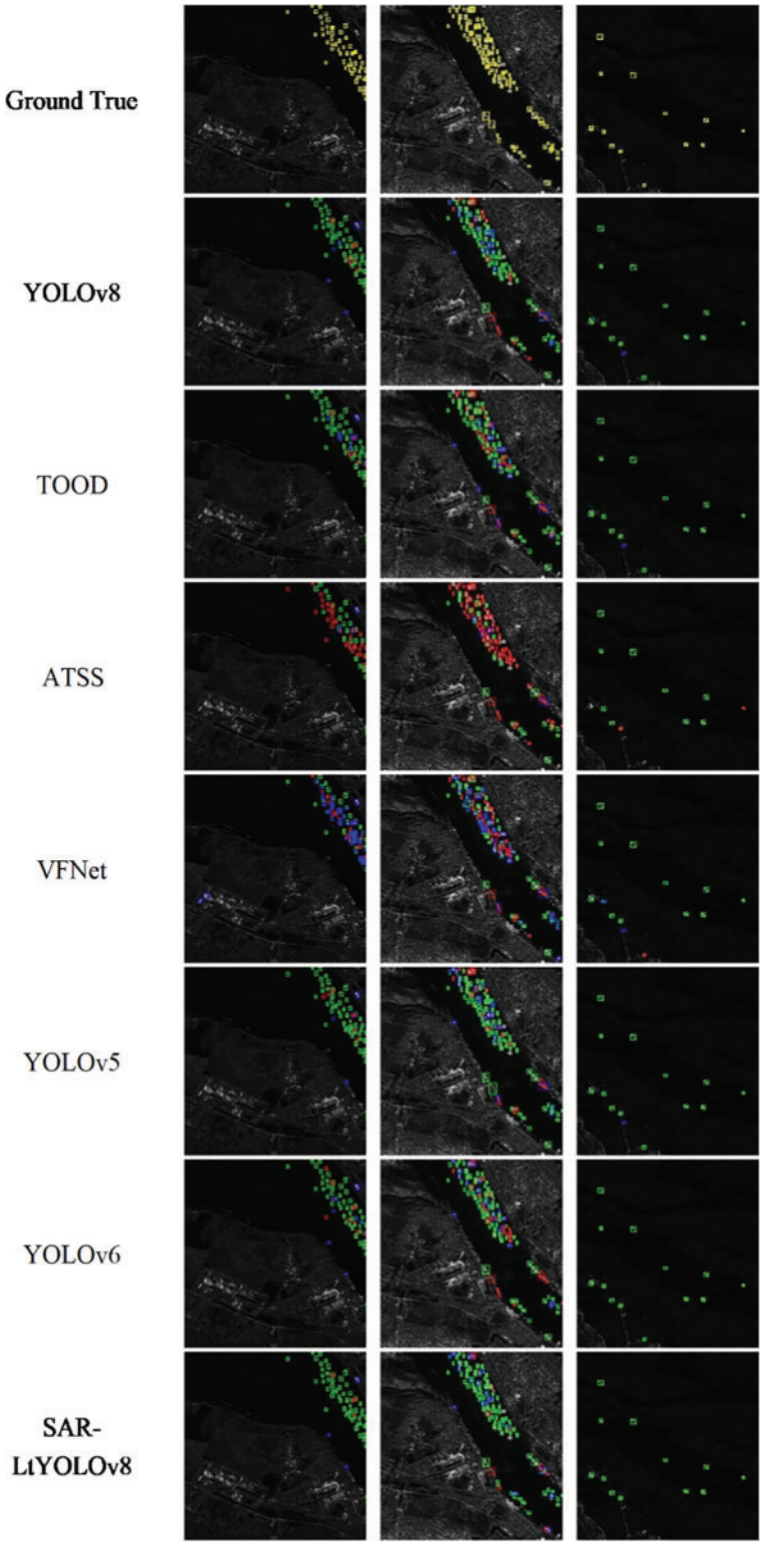

Figure 9: HRSID SAR ship detection results

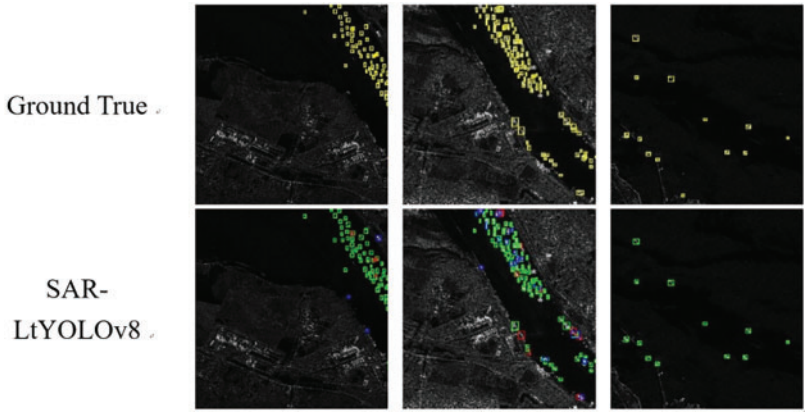

Figure 10: HRSID SAR ship detection results

5 Conclusion

To address the challenges associated with SAR ship images, such as indistinct outlines, low resolution, high noise, small target focus, and background interference, we propose an enhanced YOLOv8 model, termed SAR-LtYOLOv8. This model replaces YOLOv8's backbone and neck structures with SB and DMDH structures, respectively, and integrates the ODConv, LSKNet, and MSCA modules in the neck section. On the HRSID dataset, our model achieves a 67.8% reduction in parameters compared to YOLOv8n, with an AP_{50} of 92.7%, representing a 3.4% improvement, and an AP_{50-95} of 60%, a 5.4% improvement. Additionally, it improves AP_S, AP_M, and AP_L by 5.4%, 0.1%, and 20.9%, respectively. On the SSDD dataset, it reaches an AP_{50} of 97.1%, a 0.5% increase, and an AP_{50-95} of 67%. Comparative experiments with various general SAR ship models demonstrate the superiority of our proposed model. However, the model has limitations, particularly in detecting large vessels. To address this, we plan to augment the SAR image dataset of large vessels within the HRSID dataset. We will select large ship images from the Dior Ship dataset and the Ships-Google-Earth dataset, convert these images to grayscale, and introduce significant or simulated noise to mimic the characteristics of SAR images. The modified images will then be predicted using the SAR-LtYOLOv8 model. Finally, we will manually review and correct the predicted bounding boxes, classes, and confidence scores compared to the original images to create an expanded dataset. By mitigating the current deficiency of large vessel image data in existing SAR datasets, we aim better to meet the model's requirements for large vessel detection and improve its overall applicability in maritime vessel detection. Moreover, SAR-LtYOLOv8's advanced architecture also makes it highly suitable for detecting small, densely packed targets. For instance, in scenarios like detecting vehicles on congested roadways, the model's improved resolution and focus capabilities can enhance accuracy in distinguishing individual vehicles amid heavy traffic. Similarly, in crowded scenes such as public gatherings or events, the model's sensitivity to small objects and ability to handle high noise levels enable it to effectively identify and track individuals or objects despite the dense background. This versatility underscores the model's broader applicability beyond maritime contexts, potentially offering significant advancements in urban monitoring and event security.

Acknowledgement: None.

Funding Statement: This work was supported by the Open Research Fund Program of State Key Laboratory of Maritime Technology and Safety in 2024. This research also received partial funding from the National Natural Science Foundation of China (Grant No. 52331012) and the Natural Science Foundation of Shanghai (Grant No. 21ZR1426500).

Author Contributions: Conceptualization: Conghao Niu; methodology: Conghao Niu and Dezhi Han; software: Conghao Niu and Bing Han; validation: Conghao Niu, Dezhi Han and Bing Han; formal analysis: Conghao Niu and Zhongdai Wu; investigation: Conghao Niu; resources: Conghao Niu; data curation: Conghao Niu; writing and original draft preparation: Conghao Niu, Dezhi Han and Zhongdai Wu; writing—review and editing: Conghao Niu, Dezhi Han and Bing Han; visualization: Conghao Niu; supervision: Dezhi Han and Bing Han; project administration: Dezhi Han; funding acquisition: Dezhi Han and Bing Han. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are available on request from the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Wang, D. Han, M. Cui, and C. Chen, "NAS-YOLOX: A SAR ship detection using neural architecture search and multi-scale attention," *Connect. Sci.*, vol. 35, no. 1, pp. 1–32, 2023. doi: 10.1080/09540091.2023.2257399.

[2] S. Gao, J. M. Liu, Y. H. Miao, and Z. J. He, "A high-effective implementation of ship detector for SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, no. 6, pp. 1–5, 2022. doi: 10.1109/LGRS.2021.3115121.

[3] M. Amrani, A. Bey, and A. Amamra, "New SAR target recognition based on YOLO and very deep multi-canonical correlation analysis," *Int. J. Remote Sens.*, vol. 43, no. 15–16, pp. 5800–5819, 2022. doi: 10.1080/01431161.2021.1953719.

[4] J. Zhou and J. Xie, "Robust CFAR detector based on KLQ estimator for multiple-target scenario," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023. doi: 10.1109/TGRS.2023.3336053.

[5] Y. Li, S. Zhang, and W. -Q. Wang, "A lightweight faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022. doi: 10.1109/LGRS.2020.3038901.

[6] M. Jiang, L. Gu, X. Li, F. Gao, and T. Jiang, "Ship contour extraction from SAR images based on faster R-CNN and chan-vese model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023. doi: 10.1109/TGRS.2023.3247800.

[7] J. Qian, J. Lin, D. Bai, R. Xu, and H. Lin, "Omni-dimensional dynamic convolution meets bottleneck transformer: A novel improved high accuracy forest fire smoke detection model," *Forests*, vol. 14, no. 4, 2023, Art. no. 838. doi: 10.3390/f14040838.

[8] Y. Li *et al.*, "LSKNet: A foundation lightweight backbone for remote sensing," 2024, *arXiv:2403.11735*.

[9] X. Ma, Z. Ji, S. Niu, T. Leng, D. L. Rubin and Q. Chen, "MS-CAM: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in SD-OCT images," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3443–3455, 2020. doi: 10.1109/JBHI.2020.2999588.

[10] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 6817–6825, 2022. doi: 10.1109/JSTARS.2022.3198517.

[11] W. Zhao, M. Syafrudin, and N. L. Fitriyani, "CRAS-YOLO: A novel multi-category vessel detection and classification model based on YOLOv5s algorithm," *IEEE Access*, vol. 11, no. 20, pp. 11463–11478, 2023. doi: 10.1109/ACCESS.2023.3241630.

[12] Z. Chen, C. Liu, V. F. Filaretov, and D. A. Yukhimets, "Multi-scale ship detection algorithm based on YOLOv7 for complex scene SAR images," *Remote Sens.*, vol. 15, no. 8, 2023, Art. no. 2071. doi: 10.3390/rs15082071.

[13] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji and G. Kuang, "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4209. doi: 10.3390/rs13214209.

[14] Y. Guo, S. Chen, R. Zhan, W. Wang, and J. Zhang, "LMSD-YOLO: A lightweight YOLO algorithm for multi-scale SAR ship detection," *Remote Sens.*, vol. 14, no. 19, 2022, Art. no. 4801. doi: 10.3390/rs14194801.

[15] K. Zhao, R. Lu, S. Wang, X. Yang, Q. Li and J. Fan, "ST-YOLOA: A Swin-transformer-based YOLO model with an attention mechanism for SAR ship detection under complex background," *Front. Neurorobot.*, vol. 17, 2023, Art. no. 1170163. doi: 10.3389/fnbot.2023.1170163.

[16] J. Jiang, X. Fu, R. Qin, X. Wang, and Z. Ma, "High-speed lightweight ship detection algorithm based on YOLO-v4 for three-channels RGB SAR image," *Remote Sens.*, vol. 13, no. 10, 2021, Art. no. 1909. doi: 10.3390/rs13101909.

[17] S. Cai, H. Meng, and J. Wu, "FE-YOLO: YOLO ship detection algorithm based on feature fusion and feature enhancement," *J. Real-Time Image Process*, vol. 21, no. 2, pp. 1–13, 2024. doi: 10.1007/s11554-024-01445-5.

[18] W. Zhan, C. Zhang, S. Guo, J. Guo, and M. Shi, "EGISD-YOLO: Edge guidance network for infrared ship target detection," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 17, no. 3, pp. 10097–10107, 2024. doi: 10.1109/JSTARS.2024.3389958.

[19] Y. Zhang and D. Han, "Swin-PAFF: A SAR ship detection network with contextual cross-information fusion," *Comput. Mater. Contin.*, vol. 77, no. 2, 2023. doi: 10.32604/cmc.2023.042311.

[20] W. Wang, D. Han, C. Chen, and Z. Wu, "FastPFM: A multi-scale ship detection algorithm for complex scenes based on SAR images," *Connect Sci.*, vol. 36, no. 1, 2024, Art. no. 2313854. doi: 10.1080/09540091.2024.2313854.

[21] M. Hussain, "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection," *Machines*, vol. 11, no. 7, 2023, Art. no. 677. doi: 10.3390/machines11070677.

[22] D. Pathak and U. S. N. Raju, "Shuffled-Xception-DarkNet-53: A content-based image retrieval model based on deep learning algorithm," *Comput. Electr. Eng.*, vol. 107, no. 2, 2023, Art. no. 108647. doi: 10.1016/j.compeleceng.2023.108647.

[23] F. Chen, M. Deng, H. Gao, X. Yang, and D. Zhang, "NHD-YOLO: Improved YOLOv8 using optimized neck and head for product surface defect detection with data augmentation," *IET Image Process*, vol. 18, no. 7, pp. 1915–1926, 2024. doi: 10.1049/ipr2.13073.

[24] C. Wang, A. Bochkovskiy, and H. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/Cvf Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13029–13038.

[25] D. Singhania, R. Rahaman, and A. Yao, "C2F-TCN: A framework for semi-and fully-supervised temporal action segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11484–11501, 2023. doi: 10.1109/TPAMI.2023.3284080.

[26] B. Xiao, M. Nguyen, and W. Yan, "Fruit ripeness identification using YOLOv8 model," *Multimed. Tools Appl.*, vol. 83, no. 9, pp. 28039–28056, 2024. doi: 10.1007/s11042-023-16570-9.

[27] C. Feng, Y. Zhong, Y. Gao, M. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV). IEEE Comput. Soc.*, 2021, pp. 3490–3499.

[28] H. Luo, P. Wang, H. Chen, and M. Xu, "Object detection method based on shallow feature fusion and semantic information enhancement," *IEEE Sens. J.*, vol. 21, no. 19, pp. 21839–21851, 2021. doi: 10.1109/JSEN.2021.3103612.

[29] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," 2022, *arXiv:2209.07947*.

[30] Y. Li, Q. Hou, Z. Zheng, M. Cheng, J. Yang and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16794–16805.
[31] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211, pp. 3–19, 2018. doi: 10.1007/978-3-030-01234-2_1.
[32] Q. Guo, C. Wang, D. Xiao, and Q. Huang, "A novel multi-label pest image classifier using the modified Swin Transformer and soft binary cross entropy loss," *Eng. Appl. Artif. Intell.*, vol. 126, no. 3, 2023, Art. no. 107060. doi: 10.1016/j.engappai.2023.107060.
[33] A. S. Dina, A. B. Siddique, and D. Manivannan, "A deep learning approach for intrusion detection in Internet of Things using focal loss function," *Internet Things*, vol. 22, no. 1, 2023, Art. no. 100699. doi: 10.1016/j.iot.2023.100699.
[34] X. Wang and J. Song, "ICIoU: Improved loss based on complete intersection over union for bounding box regression," *IEEE Access*, vol. 9, pp. 105686–105695, 2021. doi: 10.1109/ACCESS.2021.3100414.
[35] 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.
[36] T. Zhang *et al.*, "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3690. doi: 10.3390/rs13183690.
[37] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312*.
[38] M. Safaldin, N. Zaghden, and M. Mejdoub, "An improved YOLOv8 to detect moving objects," *IEEE Access*, vol. 12, pp. 59782–59806, 2024. doi: 10.1109/ACCESS.2024.3393835.
[39] K. Ou *et al.*, "Drone-TOOD: A lightweight task-aligned object detection algorithm for vehicle detection in UAV images," *IEEE Access*, vol. 12, pp. 41999–42016, 2024. doi: 10.1109/ACCESS.2024.3378248.
[40] D. Patel and P. S. Sastry, "Adaptive sample selection for robust learning under label noise," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3932–3942.
[41] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "VarifocalNet: An IoU-aware dense object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8514–8523.
[42] S. Li, Y. Li, Y. Li, M. Li, and X. Xu, "YOLO-FIRI: Improved YOLOv5 for infrared image object detection," *IEEE Access*, vol. 9, pp. 141861–141875, 2021. doi: 10.1109/ACCESS.2021.3120870.
[43] C. Li *et al.*, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.