ITC 4/53
Information Technology and Control
Vol. 53 / No. 4/ 2024
pp. 1204-1220
DOI 10.5755/j01.itc.53.4.37003

PMF-YOLOv8: Enhanced Ship Detection Model in Remote Sensing Images

Received 2024/04/16 | Accepted after revision 2024/06/25

HOW TO CITE: Chen, D., Zhao, H., Li, Y., Zhang, Z., Zhang, K. (2024). PMF-YOLOv8: Enhanced Ship Detection Model in Remote Sensing Images. *Information Technology and Control*, 53(4), 1204-1220. https://doi.org/10.5755/j01.itc.53.4.37003

PMF-YOLOv8: Enhanced Ship Detection Model in Remote Sensing Images

Dan Chen, Hongdong Zhao, Yanqi Li, Zhitian Zhang, Ke Zhang

School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, China;
Innovation and Research Institute of Hebei University of Technology in Shijiazhuang, Shijiazhuang 050299, China

Corresponding author: zhaohd@hebut.edu.cn

Satellite remote sensing technology plays a pivotal role in ship monitoring at sea, with ship detection by artificial intelligence becoming the primary means. However, due to the intricate marine environment and the similarity between classes of remote sensing ships, the detection of remote sensing ships still faces significant challenges. Existing detection models tend to overlook the loss of fine-grained features of remote sensing ships during the deepening of the network. To address this issue, we proposed an enhanced Pyramid for Multi-Scale Feature Fusion (PMF) to optimize the YOLOv8 algorithm. After incorporating a fusion of shallow-level features into the neck portion of YOLOv8, an adaptive spatial feature fusion approach coupled with a path aggregation network was employed to process the output features of the backbone network. This integration enhances the learning of fine-grained features and addresses the issue of feature loss, a common challenge in existing networks. Furthermore, to enhance feature extraction, we introduced an enhanced R-C2f module. Finally, Inner-MPDIoU was employed as the bounding box loss to address the issue of missed detections that may arise in the context of dense remote sensing ships. Experiments were conducted on FGSC-T, a dataset comprising 22 classes of ships, to assess the efficacy and viability of the algorithm. In comparison to the original YOLOv8, the mAP50, mAP50-95, Recall, and Precision increased by 3.7%, 4.1%, 5.7%, and 2.5%, respectively. Furthermore, the detection speed of PMF-YOLOv8 can reach 74 fps, which meets the requirements for real-time detection of remote sensing ships.

KEYWORDS: Remote sensing ships, YOLOv8, Feature fusion, Fine-grained detection, Bounding box loss.

1. Introduction

Satellite remote sensing technology has become the primary method for monitoring ships at sea due to its advantages, such as wider coverage and no need for ship communication. Especially with the rise of high-resolution remote sensing technology, recognizing the category and location of ships based on remote

sensing images has become a new research hotspot [32]. With the help of advanced radar technology, image processing algorithms and artificial intelligence technology, the efficient identification and accurate classification of maritime targets can be achieved. This field of research plays a crucial role in both military and civilian sectors, enhancing naval combat capability and safeguarding national security. Additionally, it has multiple applications in marine resource development and maritime traffic management. This research promotes scientific and technological progress, providing more effective means of monitoring and managing maritime activities.

Traditional ship detection algorithms typically use manual feature extraction and simple classifiers for detection. Compared to traditional means of object extraction, implementations using deep learning methods often provide better results. To improve the effectiveness of ship detection while increasing the robustness of the network, researchers in the field of deep learning in remote sensing image processing and object detection usually start from the following two key aspects.

On the one hand, remote sensing images are processed to exploit the color, texture, shape and other information in the images through preprocessing, enhancement and feature extraction to more accurately distinguish different target classes. In addition, data augmentation and model regularization techniques [42] are applied. By transforming, augmenting or adding noise to the training data, the variety and amount of data can be increased to improve the ability of the model to generalize. Meanwhile, regularization techniques such as dropout and L1 / L2 regularization can be used to effectively avoid model overfitting and improve the robustness and generalization performance of the network.

On the other hand, researchers are also trying different object detection algorithms, and research on object detection algorithms includes both one-stage and two-stage algorithmic models. Two-stage object detection algorithms tend to have higher detection accuracy. The so-called two-stage consists of two stages: the generation of suggestion frames and the classification of the frames in order to filter the prediction frames. Classical network models include SPPnet [11], Faster-RCNN [22], etc. One-stage object detection algorithms, on the other hand, omit the generation of suggestion frames and directly generate prediction frames, which greatly improves the detection speed while maintaining the accuracy as much as possible, and is more suitable for real-time monitoring requirements [39-40]. The classic single-stage object detection algorithms are SSD [19], YOLO [21, 27], etc.

While two-stage algorithms are slower due to the necessity of generating a large number of bounding boxes, they exhibit higher detection accuracy and recognition rates attributed to the increased precision of the bounding boxes. Conversely, one-stage algorithms operate at a faster pace but may compromise some accuracy by directly predicting the object's location. Considering the distinct algorithmic characteristics and the demands for precision and real-time performance in maritime detection, we have opted for the one-stage YOLOv8 model as our foundational framework.

The research on the detection of ships is currently encountering several challenges. Firstly, the dimensions of the ship occupy only a small proportion of the pixels in the optical remote sensing image, which is insufficient for accurate detection. Furthermore, the similarity between ships is often considerable, and the distinction between them is not always apparent, which frequently results in erroneous detection. Moreover, the dynamic nature of the maritime environment necessitates the real-time detection of ships to meet high standards. In order to address the aforementioned issues, researchers have employed a range of solution strategies.

Zhou et al. [10] proposed Oriented R-CNN, which introduces a rotating detection box in the process of feature learning. Li et al. [13] utilized Graph Neural Network (GNN) to incorporate ship trajectory information. Zhang et al. [35] introduced polarization fusion networks with geometric feature embedding. However, integrating new information entails increased computational load and inference slowdown, as well as potential introduction of noise during calculation, ultimately hindering model convergence. Ren et al. [23] proposed an efficient lightweight network called YOLO-Lite to effectively improve ship detection speed by reducing network layers, however, this reduction may compromise the model's ability to capture complex features and diminish its generalization capability.

After the successful application of Transformer models in natural language processing, they have also demonstrated strong performance in computer vision tasks, from introducing the Transformer architecture in ViT [5] to fusing multi-scale information in CrossViT [2] and improving computational efficiency and flexibility in AdaViT [20]. Various innovations have driven the utilization of Transformers for vision tasks. For example, Zheng et al. [43] applied ViT to traffic sign recognition, while Tummala et al. [25] used ViT for brain tumor classification. Additionally, Xu et al. [28] designed the LPSW backbone network and SAIEC network framework by incorporating the Swin Transformer, and Liu et al. [16] utilized the Transformer mechanism to enhance network feature extraction with TS2Anet, resulting in higher accuracy in remote sensing ship detection.

Despite their outstanding performance, Transformer models face challenges such as high computation and memory requirements, complex architecture design, and reliance on large-scale data, which hinder their efficient application in resource-constrained environments. Given these limitations and our specific focus on limited computing power resources and real-time ship detection requirements, it is evident that the Transformer model cannot meet our needs.

Based on this analysis of existing research shortcomings, we aim to address these issues by improving feature extraction methods, multi-scale feature fusion techniques, and optimizing loss function strategies to achieve real-time high-precision ship detection.

We proposed an enhanced Pyramid for Multi-Scale Feature Fusion (PMF) to optimize the YOLOv8 algorithm. The method presented not only effectively improves detection accuracy but also meets real-time speed requirements for detection.

The main contributions of this paper are as follows:

1. Integrating fusion into the shallow-level feature extraction layer allows the network to effectively leverage spatial information from higher-resolution feature maps, enhancing the learning of detailed features in small-sized targets.
2. We adopt FASFF, an enhanced adaptive feature fusion module that employs an adaptive spatial feature fusion approach.
3. We enhanced the R-C2f module by introducing the RFAConv module to address parameter sharing issues. This allows for tailored treatments of various receptive field sizes, improving the network's adaptability to different input features.
4. Inner-MPDIoU is utilized as the bounding box regression loss. MPDIoU excels in handling boundary information and dense targets, while Inner accurately evaluates the bounding box overlap degree.

The remaining section of this paper is as follows: Section 2 presents some related works. Section 3 introduces the proposed methodology, while Section 4 gives the experimental results and analysis. Section 5 concludes our work as well as future research directions.

2. Related Work

In this section, we will focus on some research methods and existing problems proposed by researchers for ship detection at sea at this stage, especially the processing strategies for multi-scale features.

Typically, conventional ship detection algorithms employ manual feature extraction and simple classifiers for detection. Manual feature extraction techniques, such as Scale Invariant Feature Transform (SIFT) [7], Histogram of Oriented Gradients (HOG) [30], and Speeded Up Robust Features (SURF) [1], are used. Common simple classifiers include Support Vector Machines (SVMs), Adaptive Boosting, and Decision Trees.

However, traditional methods for object detection face several challenges and shortcomings. Firstly, these methods often rely on manually designed feature extractors that require domain expertise to select and adjust correctly. Secondly, the process of manual feature extraction is inefficient and requires constant updates and adjustments for different targets and environmental changes, which limits its applicability in complex scenarios. Deep learning-based object detection algorithms can automatically learn and extract key features from images through an end-to-end learning approach, resulting in stronger generalization and adaptability [31].

Feature Pyramid is an important method for multi-scale feature fusion [41, 33] in computer vision, which plays a key role in object detection tasks. Early Image Pyramid techniques generated image pyramids of different resolutions for multiscale analysis by pro-

gressively downsampling an image. However, the hand-designed filters and sampling strategies used often lacked flexibility and adaptability.

Spatial Pyramid Pooling Network (SPP-Net) proposed a spatial pyramid pooling layer, which enabled convolutional neural networks to process input images of any size and capture multi-scale features by pooling at different scales [11]. However, SPP-Net relied on a fixed pooling layer configuration, which could not fully adapt to the needs of different tasks and datasets. Sungyi Lin et al. [14] proposed FPN (Feature Pyramid Network), which fused deep and shallow features through top-down paths and lateral connections to construct high-resolution and low-resolution multi-scale feature pyramids, greatly improving object detection performance. Although FPN significantly improved the performance, it only performed fusion at partial levels of the network and might omit some useful information at intermediate layers.

Path Aggregation Network (PANet) [18] further improved on FPN by adding bottom-up paths to enhance feature fusion so that multi-scale information could be better propagated and fused, thus further improving the effect of object detection and instance segmentation. However, the complex path structure of PANet increased the computational overhead and implementation difficulty.

Neural Architecture Search Feature Pyramid Network (NAS-FPN) [8] used neural architecture search (NAS) to automatically design the feature pyramid structure and find the optimal feature fusion strategy, which improved the performance of object detection tasks. Although NAS-FPN was more automated and optimized in design, the NAS process was very time-consuming and had high computational resource requirements, making it difficult to generalize to resource-limited application scenarios.

BiFPN (Bi-directional Feature Pyramid Network) [24] further improved the expression ability and computational efficiency of the feature pyramid by introducing learnable weights to fuse features of different scales and using bidirectional feature fusion paths. Although BiFPN improved efficiency, it might introduce too many parameters in the feature fusion process, complicating the training and inference process.

AugFPN (Augmented FPN) [9] further improved the expression ability and detection effect of features by introducing a feature enhancement module on the basis of FPN. However, the enhancement module design of AugFPN was complex, increasing the complexity of the network structure and the computational cost.

In the task of ship detection at sea, Zhou et al. [44] proposed MSSDNet based on YOLOv5, adding FC-FPN and CSPMRes2 in the process of feature fusion to solve the problem of feature loss in the process of feature fusion. Chen et al. [4] designed a SAS-FPN to adapt to multi-scale ship detection. Yan et al. [29] adopted the feature fusion strategy of ReBiFPN to effectively capture and enrich multi-scale feature information. These methods improved the learning of multi-scale features of neural networks to a certain extent. However, these methods still had some shortcomings in generalization ability, computational cost, and real-time requirements in practical applications.

Aiming at the actual needs of ship detection, we improved the structure of YOLOv8n to enhance the effectiveness of ship detection in optical remote sensing images.

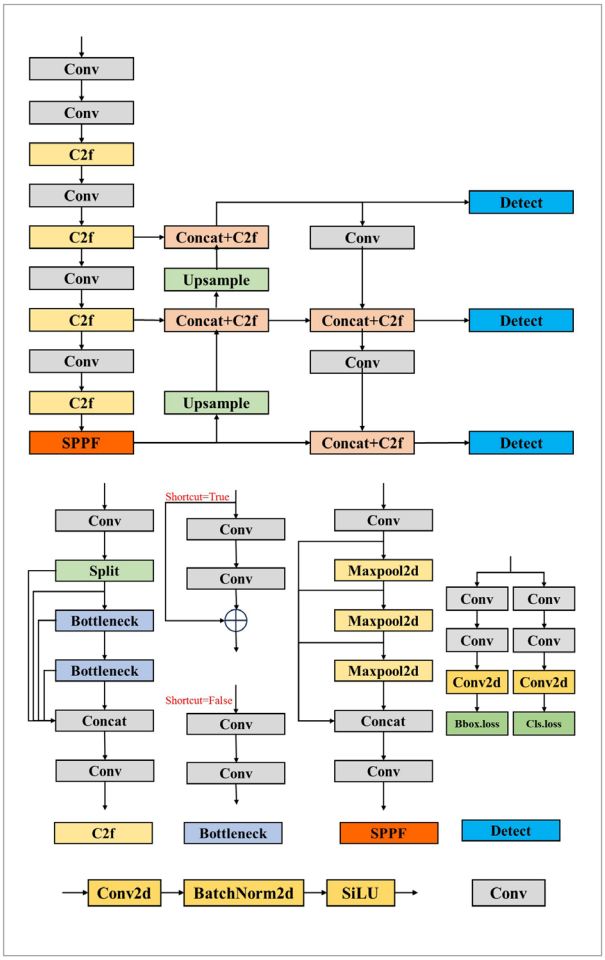
Firstly, in view of the large difference in the size of ships and the inter-class similarity of different categories of ships, we introduced a feature fusion module at the large-scale feature extraction position to better extract the detailed features of ships. Secondly, in solving the problem of feature loss, we used the improved FASFF, which effectively improved the detection accuracy by adding adaptive feature fusion in different network layers. Aiming at the parameter sharing problem existing in traditional convolution, we improved C2f by using RFAConv. Finally, to more accurately evaluate the degree of overlap between object detection boxes, we adopted MPDIoU instead of CIOU.

Compared with the baseline YOLOv8, our model improved the mAP50, MAP50-95, precision, and recall by 3.7%, 4.1%, 5.7%, and 2.5%, respectively. Taken together, our model performs much better in ship detection and provides a more reliable solution for practical applications.

3. Methods

This section focuses on the designed network structure. YOLOv8 is a significant update to YOLOv5, released on January 10, 2023 by Ultralytics. It introduces support for image classification, segmentation, detection, and keypoint detection. Compared to its pre-

Figure 1
YOLOv8 structure (at the top) includes C2f module, Bottleneck module, SPPF module, the Detect module (in the middle) and Conv module (at the bottom)

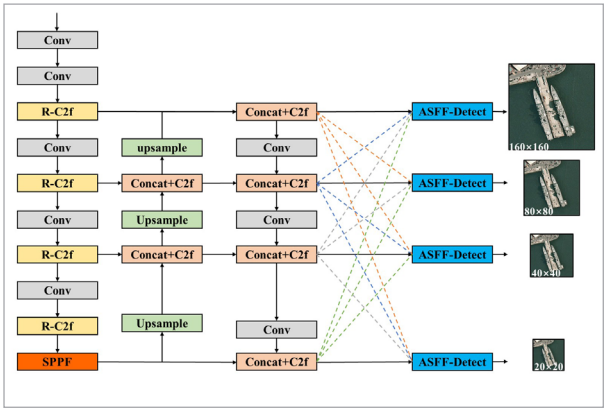


decessor, YOLOv8 offers faster detection speed and higher accuracy. The architecture of YOLOv8 consists of three components: backbone, Neck, and Head, with four versions available: n, s, m, l, and x. The "n" version is the lightest and fastest iteration of YOLOv8, while the "x" version is the most accurate but slowest [12]. In this study, we have selected the lightweight YOLOv8n for deployment on embedded devices in later stages. The structure of YOLOv8 is depicted in Figure 1.

3.1. PMF-YOLOv8

The network structure of PMF-YOLOv8, as illustrated in Figure 2, comprises four distinct components. These include the following: an input module for preprocessing images, a backbone feature extraction module for capturing image features, a neck module for further refining fused features, and a detection header module for conducting regression-based object localization.

Figure 2
Structure of PMF-YOLOv8: In the backbone part, the C2f module is replaced by the R-C2f module. In the neck part, we add a feature fusion module for shallow features and use the PAN-FASFF feature fusion method



The improvements made to PMF-YOLOv8 are as follows:

1. The addition of feature fusion to the shallow-level feature extraction layer allows for full utilisation of the spatial information provided by higher resolution feature maps. This enables the network to effectively learn the detailed features of small-sized targets. Additionally, the shallow-level network typically contains more fine-grained features, effectively addressing the issue of inter-class similarity of ships.
2. We introduce the FASFF, an improved adaptive feature fusion module. FASFF uses an adaptive spatial feature fusion method, which enables the model to select the most useful features at each spatial location, filter out conflicting information, enhance scale invariance, and improve the accuracy of detecting ship targets.
3. The R-C2f module is introduced to solve the parameter sharing problem that exists in traditional convolution. The network employs the RFAConv module to provide different treatments for various regions and sizes of receptive fields, thereby enhancing the network's expressive power. This makes the network more flexible and better able to adapt to different input features, further improving the detection of ships.

4 InnerMPDIoU is introduced. Compared to CIOU, MPDIoU performs better in handling boundary information and dense targets. InnerMPDIoU can more accurately evaluate the degree of overlap of object detection frames, effectively solving the problem of leakage that may occur during ship detection.

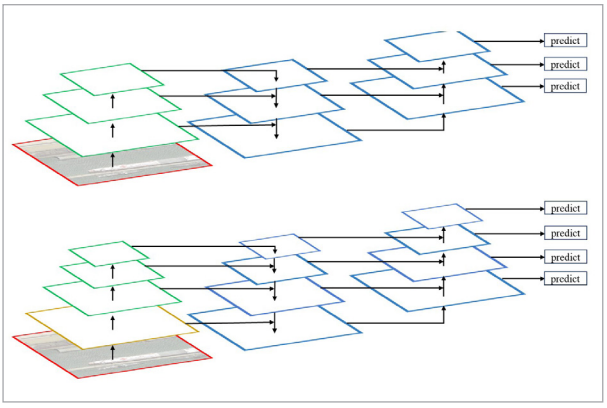
3.2. Feature Pyramid

SSD is a one-stage target detector that first attempted to extract information from multiple feature scales in a bottom-up manner in order to simultaneously predict the target category and location [19]. The Feature Pyramid Network (FPN) is a deep neural network architecture designed to solve multi-scale object detection. Its core objective is to obtain rich semantic information at different scales by constructing bottom-up and top-down feature propagation paths [14]. FPN aims to build a multi-scale feature pyramid to achieve this goal. The bottom-up path extracts low-level features, while the top-down path transfers high-level semantic information to the low-level through up-sampling and feature fusion. This design enables FPN to be more accurate and effective in dealing with multi-scale targets. FPN has become an important technique in the field of object detection and semantic segmentation. By processing features at different scales effectively, FPN enhances the model's capability to adjust to scale changes and multi-scale targets, thereby improving the detection task's performance [15].

While FPN addresses the issue of feature loss at various scales, the top-down feature propagation process, particularly during multilevel feature fusion, may still result in information loss or blurring. As a result, the Path Aggregation Network (PAN) was developed. PAN is a deep neural network structure that aims to solve the problem of multiscale feature fusion. It is mainly used for object detection and semantic segmentation tasks. The core goal of PAN is to integrate feature information at different levels effectively. This is achieved by introducing a path aggregation module to improve the model's ability to perceive the target. Specifically, PAN combines bottom-up and top-down feature propagation paths, and utilizes lateral connectivity and up-sampling operations to achieve effective feature fusion and retention. This design enables PAN to generate feature pyramids with richer semantic information, leading to better performance in object detection and semantic segmentation tasks [18]. Furthermore, PAN incorporates a path aggregation module that facilitates the fusion and retention of features at various levels through bottom-up and top-down feature propagation paths, as well as lateral connectivity and up-sampling operations. This design allows PAN to handle multi-scale features more flexibly and enhances the model's ability to perceive the target. The Neck component of YOLOv8 maintains the style of YOLOv5 and also employs the concept of PAN.

Based on the significant size difference between various types of ships and the abundance of small target ships, we introduce a feature fusion layer for small targets into the original YOLOv8 structure, based on PAN. The shallow-level feature extraction layer retains a relatively high image resolution, allowing for more spatial information in the higher-resolution feature map to facilitate better learning of detailed features of small-size targets. Furthermore, the network's shallow-level layers tend to capture more fine-grained features, which are essential for understanding the shape and texture of small targets. Therefore, we have constructed a feature pyramid PAN-4 that excels in extracting features from small targets, effectively enhancing their detection. This structure also enables the network to better differentiate between ship classes with high similarity. The structure of the feature pyramid is shown in Figure 3.

Figure 3
The original PAN structure of YOLOv8 (at the top), and the PAN-4 structure with a shallow feature extraction layer added in PMF-YOLOv8 (at the bottom)



3.3. FASFF Module

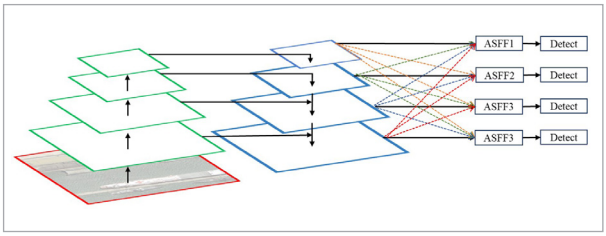
The FASFF model, an improved version of ASFF (Adaptively Spatial Feature Fusion), is illustrated in Figure 4. It employs an adaptive spatial feature fusion

technique to effectively filter out conflicting information and enhance scale invariance, thereby enhancing the accuracy of ship detection. This approach enables the model to dynamically select the most relevant features at each spatial location and determine the most important feature hierarchies for final prediction based on contextual information at each feature location and scale. The core concept of FASFF lies in adaptively learning the fusion spatial weight of each scale feature map, which involves two main steps: feature resizing and adaptive feature fusion [17].

Feature resizing

To begin, we denote the feature map of level l resolution as x^l, with $l \in [1, 4]$ in Figure 4. For each level l, it is necessary to adjust the feature maps of different resolutions to match the shape of x^l. This is because the features of each level of YOLOv8 have varying resolutions and channel numbers. Therefore, we must utilize up-sampling and down-sampling strategies for the feature maps of different levels.

Figure 4
Structure of the FASFF



Adaptive feature fusion

The process of feature fusion is shown in Equation (1), where $x_{ij}^{r\to l}$ denotes the feature map adjusted from the feature map with r-level resolution to the feature map with l-level resolution, and i and j denote the positions corresponding to the feature map with l-level resolution. y_{ij}^l represents the mapping y^l of the feature vector at position (i, j) between the channels. The weights of the four different feature layers mapped to the l-levels, $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l$, and δ_{ij}^l, are adaptively learned by the network. To improve computational efficiency, it is stipulated that $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l + \delta_{ij}^l = 1$, $\alpha_{ij}^l, \beta_{ii}^l, \gamma_{ij}^l, \delta_{ij}^l \in [0.1]$. Equation (2) defines the expression for α_{ij}^l.

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1\to l} + \beta_{ij}^l \cdot x_{ij}^{2\to l} + \gamma_{ij}^l \cdot x_{ij}^{3\to l} + \delta_{ij}^l \cdot x_{ij}^{4\to l}, \tag{1}$$

$$\alpha_{ij}^l = \frac{e^{\lambda_I^l}}{e^{\lambda_\alpha^l} + e^{\lambda_\beta^l} + e^{\lambda_\gamma^l} + e^{\lambda_\delta^l}}. \tag{2}$$

The Softmax function is utilized to define $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l$, and δ_{ij}^l with λ_α^l, λ_β^l, λ_γ^l, and λ_δ^l as the control parameters, respectively. These parameters can be learned through standard backpropagation. The outputs of the FASFF module, y^1, y^2, y^3, and y^4, are utilized for ship detection in the detection head.

3.4. Improved Strategy of Convolutional Module

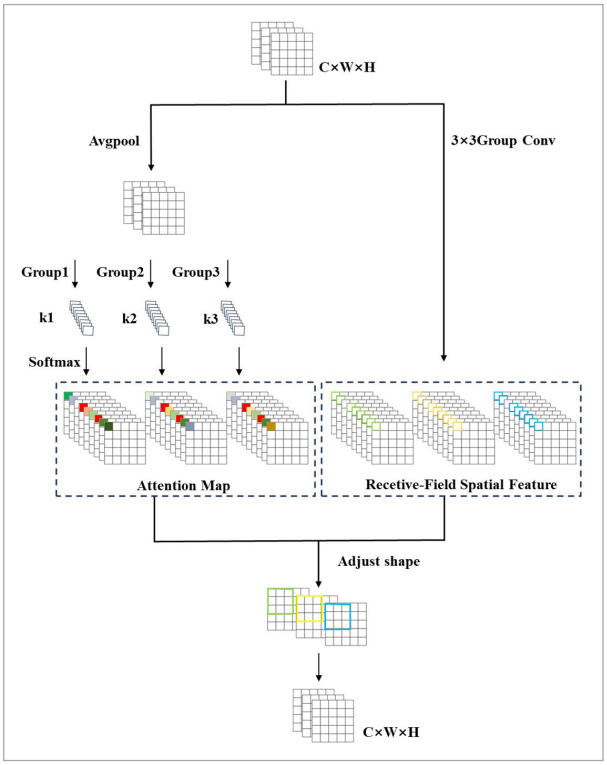
The convolution operation is a fundamental component in deep learning neural networks, crucial for feature learning and pattern recognition. However, the traditional convolution structure is constrained by the fixed size of the convolution kernel, limiting its ability to fully capture diverse feature information across different positions in the receptive field area. This constraint hinders the network's expressive power and generalization performance. To address this issue, researchers have proposed innovative solutions such as dilated convolution, which introduces an adjustable receptive field by incorporating holes between convolution kernels [34]. Additionally, the pyramid convolution structure enhances multi-scale feature perception through fusion of multi-scale convolution kernels [6]. Furthermore, deep learning models based on self-attention mechanisms have achieved significant success in capturing internal data or feature correlations while reducing dependence on external information [26] and improving network performance. However, despite its performance benefits, self-attention mechanisms also incur substantial computational costs.

In this paper, we improve the YOLOv8n model by introducing RFAConv [36] as an alternative to self-attention with lower computational requirements. RFAConv is integrated with C2f to form the R-C2f structure and focuses more on spatial characteristics of the receptive field using Rceptive-Filed Attention Convolution (RFA), addressing parameter sharing issues associated with traditional convolution kernels at a minimal increase in operational cost.

The specific process of RFAConv is shown in Figure 5. Firstly, Group Conv is utilized for spatial feature extraction of the receptive field. It is assumed that a 3×3

convolution kernel is employed to extract features, where each 3×3 window represents a receptive field slider. After extracting the receptive field features, the original features are mapped into a new feature space. It has been demonstrated that network performance can be enhanced by learning attention maps to interact with receptive field feature information. However, interacting with each receptive field feature may introduce additional computational overhead. To address this issue, AvgPool is employed to aggregate the global information of each receptive field feature.

Figure 5
Structure of the RFAConv



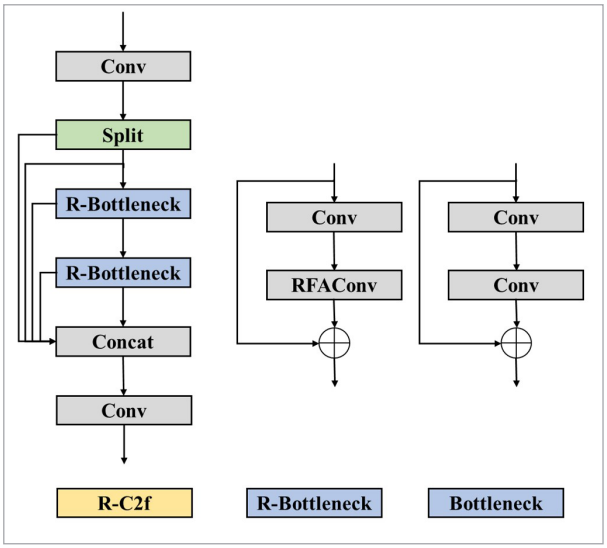
Subsequently, the 1×1 group convolution operation is used for information interaction. Finally, the softmax function emphasizes the importance of each feature in the receptive field features. The calculation process of RFA Refers to Equation (3).

$$F=Softmax\left(g^{i\times i}\left(AvgPool(X)\right)\right)\times ReLU\left(Norm\left(g^{k\times k}(X)\right)\right)=A_{rf}\times F_{rf}. \quad (3)$$

In this context, $g^{i\times i}$ denotes a grouping convolution with a size of $i \times i$, where k represents the size of the convolution kernel, *Norm* refers to normalization, X represents the input feature maps, and F is obtained by multiplying the attention map A_{rf} with the transformed receptive-field spatial feature F_{rf}.

RFAConv has the capability to enhance the detection of intricate features of ships by modulating the weight distribution across various receptive fields. Moreover, it can dynamically generate spatial features of the receptive field and flexibly adjust its shape and range based on the convolution kernel size. This adaptability allows RFAConv to accommodate diverse sizes of ships effectively. The RFAConv technique is utilized to enhance the C2f module, resulting in the creation of R-C2f as depicted in Figure 6.

Figure 6
Structure of the R-C2f: Compared with the original Bottleneck, the improved R-Bottleneck introduces the RFAConv



3.5. Loss Function

The YOLOv8 loss function comprises a categorization loss (VFL Loss) and a regression loss (CIOU Loss + DFL). The VFL Loss utilizes cross-entropy loss to predict the target category and rapidly direct the network's focus to the target location. The regression loss is composed of two components: CIOU Loss and DFL. The CIoU Loss function optimizes the prediction of the ground truth by considering factors such as

location, size, and shape. Meanwhile, the DFL function enhances detection accuracy and efficiency by optimizing the probability of predicted locations to the two locations closest to the label. This approach helps the network focus on the target location more quickly. These loss functions aim to improve the performance of the YOLOv8 model in object detection tasks while maintaining computational efficiency.

When the predicted bounding box and the ground truth bounding box do not intersect, Intersection over Union (IoU) cannot accurately reflect the distance relationship between the two. To address this issue, Distance IoU (DIoU) takes into account the distance between the the predicted bounding box and the ground truth bounding box, the overlap area, and the scale relationship. DIoU is calculated using Equation (4).

$$L_{DIoU}=1\text{-}IoU+\frac{\rho^2(b,b^{gt})}{c^2}. \tag{4}$$

The centroids of the the predicted bounding box and the ground truth bounding box are denoted as b and b^{gt}, respectively. The Euclidean distance between the two centroids is represented by ρ, and c represents the diagonal length of the smallest outer rectangle of predicted bounding box and the ground truth bounding box. CIoU takes into account the aspect ratio of predicted bounding box and the ground truth bounding box based on DIoU. The computation of CIOU is shown in Equations (5)-(7).

$$L_{CIoU}=1\text{-}IoU+\frac{\rho^2(b,\ b^{gt})}{c^2}+\alpha v, \tag{5}$$

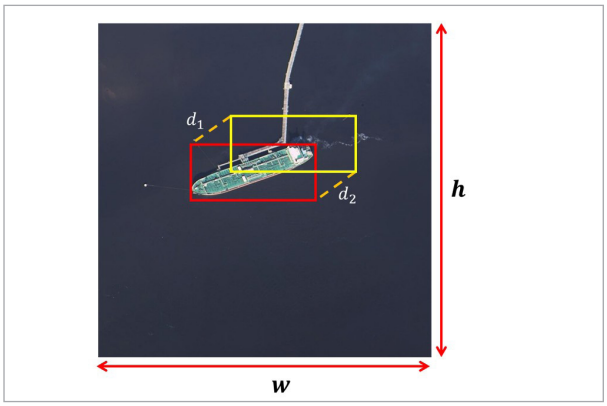
$$\alpha=\frac{v}{(1\text{-IoU})+v}\ , \tag{6}$$

$$v=\frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}}\text{-}arctan\frac{w^{p}}{h^{p}}\right)^2. \tag{7}$$

α is the weight parameter, v is used to measure the consistency of the aspect ratio. The width and height of the ground truth bounding box are represented by w^{gt} and h^{gt}, respectively, while the width and height of the predicted bounding box are represented by w^p and h^p.

When the center point of the predicted bounding box and the ground truth coincide and the aspect ratio is the same, the CIOU loses its effect. Therefore, we adopt the more advanced MPDIoU, as shown in Figure 7, to replace the CIOU. The MPDIoU takes the geometric characteristics of the bounding box into full consideration and achieves regression on the bounding box

Figure 7
MPDIoU process



by minimizing the distances between the predicted bounding box and the ground truth bounding box's upper-left corner-to-upper-left corner and lower-right corner-to-bottom-right corner points. The computation of MPDIoU is shown in Equations (8)-(10).

$$L_{MPDIoU}=1\text{-}IOU+\frac{d_1^2}{h^2+h^2}+\frac{d_2^2}{h^2+w^2}\ , \tag{8}$$

$$d_1^2=(x_1^B\text{-}x_1^A)^2+(y_1^B\text{-}y_1^A)^2\ , \tag{9}$$

$$d_2^2=(x_2^B\text{-}x_2^A)^2+(y_2^B\text{-}y_2^A)^2, \tag{10}$$

where d_1 denotes the distance between the upper left corner of the predicted bounding box and the ground truth, and d_2 denotes the distance between the lower right corner of the predicted box and the ground truth. h and w denote the height and width of the feature map, respectively.

We introduce InnerMPDIoU, which is an improvement over CIOU. MPDIoU performs better in dealing with boundary information and dense targets, while Inner can more flexibly reflect the overlapping region between two boxes. Overall, InnerMPDIoU can more accurately assess the degree of overlap between object detection boxes, reducing the risk of missed detections.

4. Experiments

4.1. Experimental Environment and Datasets

The hardware CPU used in this experiment is 12th Gen Intel (R) Core (TM) i9-12900HX 2.30 GHz, and the GPU is NVIDIA GeForce RTX 3080 Ti with 32GB

RAM. The software environment is CUDA12.3, torch version 1.12.0+cu113, python version 3.9.18, and the operating system is Windows 11.

In the experiment, the batchsize is 64, the epoch is 200, the lr0 is 0.01, the lrf is 0.01 and the weight decay coefficient is 0.0005. The mosaic data augmentation technology is turned on during the training process, and the mosaic data augmentation is turned off in the last ten rounds.

In the experiment, the ship images used for model training are from the public datasets: FGSC [37], FGSD [3] and ShipRSImageNet [38]. Based on FGSC, due to its extremely unbalanced data volume between categories, some data sets were screened in the other two data sets to supplement and delete FGSC. The resulting dataset FGSC-T, as shown in Table 1., contains 22 categories with a total of 3867 images and 4416 ships. MakeSense was used to annotate the dataset. Split into training, validation, and test sets in an 8:1:1 ratio.

4.2. Metric

To assess the effectiveness of the improved model, we utilize accuracy (P), recall (R), Average Precision (AP), mean Average Precision (mAP), and the parameter count as evaluation metrics. Precision (P) and recall (R) are calculated using Equations (11)-(12).

$$P=\frac{TP}{TP+FP}, \tag{11}$$

$$R=\frac{TP}{TP+FN}, \tag{12}$$

where *TP*, *TN*, *FP* and *FN* stand for true positive, true negative, false positive and false negative, respectively.

The expression formula of average precision (AP) and mean average precision (mAP) is shown in Equations (13)-(14).

$$AP=\int_0^1 P(R)dR, \tag{13}$$

$$mAP=\frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{14}$$

where, *AP* represents the area under the curve with Precision as the ordinate and Recall as the abscissa; mAP is the average of *AP* values for all categories.

4.3. Comparative Experiments of Modules

In order to validate the efficacy of the modules, we conducted comparative experiments on various multi-scale feature processing methods, improvements in the Conv module, and the addition of improved convolutions at different positions.

To verify the effectiveness of the FASFF method for multi-scale features, we compared it with BiFPN, ASF (currently mainstream in target detection), as well as RepGFPN and CCFM. These four methods were evaluated on dataset FGSC-T. The results in Table 2 show that while BiFPN and CCFM have fewer parameters, our model outperforms them significantly across all metrics except parameter count. RepGFPN and ASF-YOLO improve model effectiveness to some extent but fall short of our model in terms of mAP50, mAP50-95, Precision and Recall.

In order to assess the efficacy of RFAConv, we conducted comparative experiments with four convolution methods on the FGSC-T dataset: Omni-Dimensional Dynamic Convolution (ODConv), proposed in 2022; Switchable Atrous Convolution (SAConv), proposed in 2021; Deformable Convolution v4 (DCNv4), proposed in 2024; and Spatial and Channel Reconstruction Convolution (SCConv), proposed in 2023. As illustrated in Table 3, our model outperforms several other convolutional methods in terms of mean average precision (mAP50), mean average precision (mAP50-95), precision, and recall on the FGSC-T dataset, despite a slight increase in the number of parameters.

Furthermore, we have attempted to integrate RFAConv modules at various points in the network. The Bottleneck module comprises two Conv modules, which we have designated as Conv1 and Conv2, respectively. As illustrated in Table 4, replacing the Conv1 module of the Bottleneck with RFAConv has a negligible impact on the model's performance, comparable to that of a model with only a replacement of Conv2. However, replacing both Conv modules with RFAConv has a detrimental effect on the model's detection capabilities. This is due to the fact that RFAConv assigns a specific weight to each sensation by introducing an attentional mechanism. However, excessive repetition of RFAConv result in the repeated processing of information, which may lead to the model learning noisy rather than useful features, and ultimately reduce the overall performance of the model.

Table 1
The target counts per category.

Category	Number	Category	Number	Category	Number	Category	Number
Aircraft	170	Combat boat	279	Tarawa-class	88	Oil tanker	184
Destroyer	564	Auxiliary ship	287	Assault ship	158	Fishing boat	130
Landing	109	Container ship	101	Command	90	Passenger ship	110
Frigate	322	Car carrier	72	Submarine	279	Gas ship	94
Transport dock	92	Hovercraft	120	Medical ship	315		
Cruiser	310	Bulk carrier	394	Barge	54		

Table 2
Comparative experiments of different feature fusion methods for multi-scale features

model	mAP50	mAP50-95	Precision	Recall	Parameters
BiFPN	0.804	0.713	0.706	0.755	2792887
RepGFPN	0.827	0.730	0.733	0.841	3295339
CCFM	0.786	0.680	0.763	0.744	1973803
ASF-YOLO	0.825	0.722	0.718	0.830	3060854
ours	0.833	0.742	0.749	0.832	4387652

Table 3
Comparison experiments of different convolution modules

model	mAP50	mAP50-95	Precision	Recall	Parameters
ODConv	0.817	0.721	0.754	0.807	6862210
SAConv	0.838	0.743	0.795	0.761	6885914
DCNv4	0.843	0.749	0.764	0.810	6684439
SCConv	0.831	0.742	0.781	0.776	6698609
ours	0.850	0.760	0.813	0.802	6870928

Table 4
Comparison experiments of adding convolutions at different positions in the Bottleneck module

Location	mAP50	mAP50-95	Precision	Recall	Parameters
Conv1	0.849	0.758	0.817	0.794	6870928
Conv1+Conv2	0.840	0.752	0.811	0.777	6907216
Conv2(ours)	0.850	0.760	0.813	0.802	6870928

Table 5
Ablation results

yolov8n	PAN-4	FASFF	R-C2f	InnerMPDIoU	mAP50	mAP50-95	Precision	Recall	Parameters
√	×	×	×	×	0.827	0.720	0.779	0.769	3015138
√	√	×	×	×	0.832	0.737	0.773	0.79	4055416
√	√	√	×	×	0.842	0.748	0.818	0.769	6834640
√	√	√	√	×	0.850	0.760	0.813	0.802	6870928
√	√	√	√	√	0.864	0.761	0.804	0.826	6870928
√	×	√	×	×	0.833	0.742	0.749	0.832	4387652
√	×	×	√	×	0.829	0.730	0.789	0.801	3051426
√	×	×	×	√	0.831	0.724	0.81	0.762	3015138

4.4. Ablation Experiments

To further verify the effectiveness of the proposed method, four sets of ablation experiments, as shown in Table 5, are designed to analyze different improved methods. mAP50, MAP50-95, P and R are used as experimental evaluation indicators to comprehensively evaluate the performance of the model under different improved conditions, and to deeply understand the advantages and limitations of the model, so as to provide guidance for further optimization of the model.

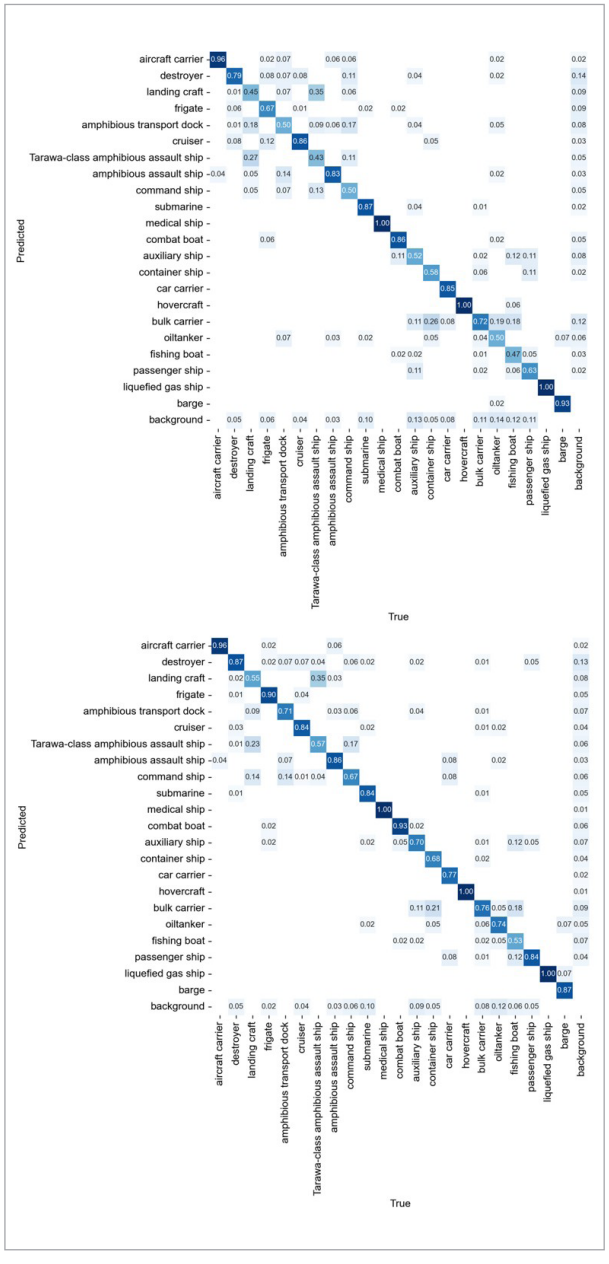
As is shown in Table 5, firstly, we add a feature fusion for the shallow-level feature extraction layer. Since the shallow-level feature extraction layer relatively retains the high resolution of the image, the feature map with higher resolution can provide more spatial information, so that the network can better learn the detailed features of small-size objects. In addition, the shallow-level network tends to have more fine-grained features, which are crucial for understanding the shape and texture of small objects. Therefore, according to the data in Table 5, after adding this layer, the mAP50 of the model increased from the original 82.7% to 83.2%, and the MAP50-75 increased from the original 72% to 73.7%.

On this basis, before detection head, we introduce FASFF. FASFF adopts an adaptive spatial feature fusion method, which can effectively filter out conflict information and enhance scale invariance to improve the accuracy of ship detection. This method allows the model to adaptively select the most useful features at each spatial location, and flexibly determine which feature hierarchies are most important for the final prediction according to the context of each feature location and scale. The experimental results show that after the introduction of the FASFF module, the precision of ship detection is improved from 77.3% to 81.8%, and the improvement is about 4.5%. In addition, the mAP50 increased from 83.2% to 84.2%, and the MAP50-95 increased from 73.7% to 74.8%. These results show that the introduction of this module significantly improves the detection performance of the model.

We introduce the R-C2f module to address the issue of parameter sharing in traditional convolutions. The module includes RFAConv, which allows for different processing of various regions and sizes of receptive fields during the feature extraction stage. This improves the network's expression ability and flexibility to adapt to different input features, ultimately enhancing the detection of ships. The experimental results indicate that the addition of the RF-C2f module increased the recall rate of ship detection from 76.9% to 80.2%, resulting in an improvement of about 3.2%, without significantly reducing the accuracy of ship detection. Additionally, mAP50 increased from 84.2% to 85.0%, and MAP50-95 increased from 74.8% to 76.0%, with an increase of 0.8% and 1.2%, respectively. The data demonstrates that the implementation of the R-C2f module significantly enhances the performance index for detecting ships.

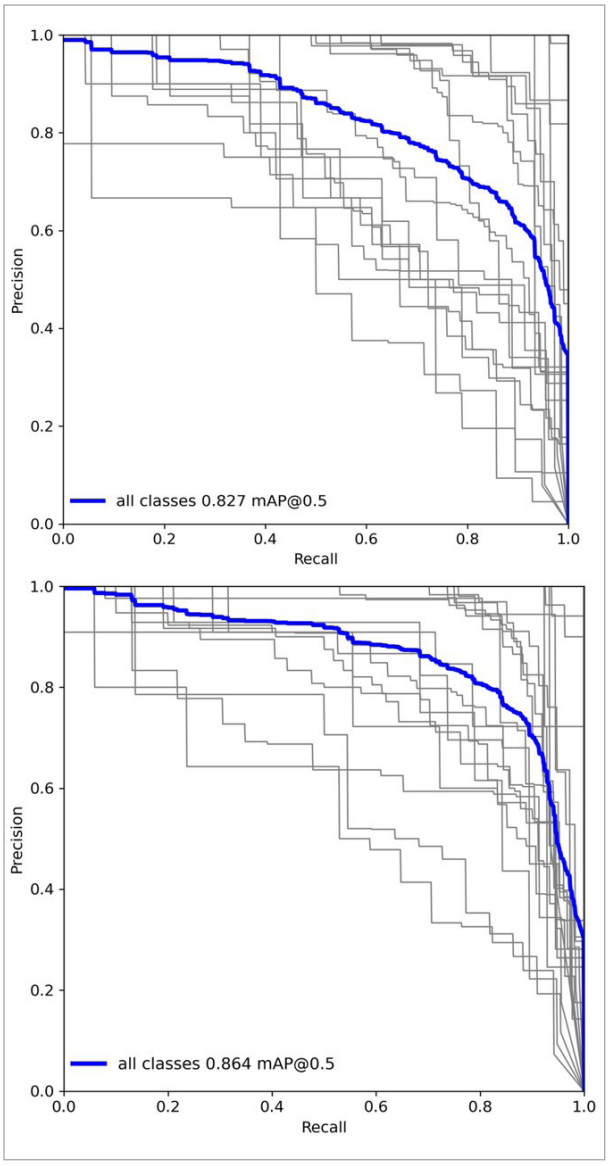
Finally, we introduce Inner-MPDIoU. Compared to CIOU, MPDIoU performs better in processing boundary information and dense targets, and can more accurately evaluate the degree of overlap between object detection boxes. Inner allows for more flexibility in reflecting overlapping areas between two boxes.

Figure 8
YOLOv8 confusion matrix (at the top) and the improved PMF-YOLOv8 confusion matrix (at the bottom)



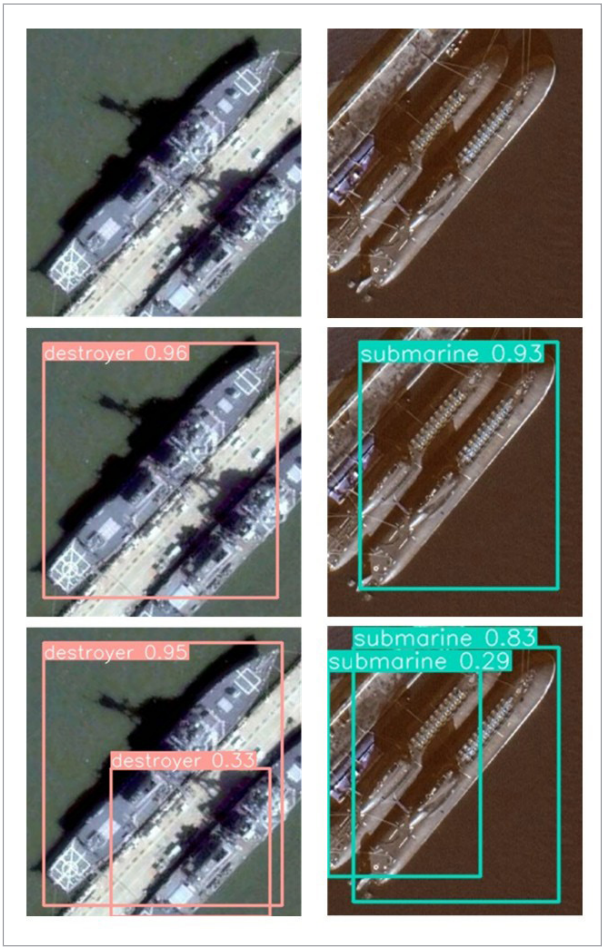
The results show that after the introduction of Inner-MPDIoU, despite a slight decrease in accuracy, the recall rate increased from 80.2% to 82.6%, an improvement of approximately 2.4%. Moreover, the mAP50 increased from 85.0% to 86.4%, which is a 1.4% increase.

Figure 9
P-R curve of YOLOv8(at the top) and P-R curve of improved PMF-YOLOv8(at the bottom)



Compared to the baseline YOLOv8n, the mAP50 improved by 3.7%, MAP50-95 improved by 4.1%, recall improved by 5.7%, and accuracy improved by 2.5%. By comparing the confusion matrix shown in Figure 8 with the P-R curve shown in Figure 9, it is evident that our model achieves better detection results. To provide a more comprehensive illustration of the superiority of our model, as shown in Figure 10, we selected the detection results of some scenes for comparison.

Figure 10
Visualization of detection effect. original image (at the top), YOLOv8 detection (in the middle), PMF-YOLOv8 detection (at the bottom)



In addition, we conducted ablation experiments for individual modules. As is shown in Table 5, the detection accuracy is improved to some extent by incorporating different modules. In practical applications, appropriate models can be selected based on different task scenarios and requirements.

4.5. Comparison with Other Advanced Models

To evaluate the performance of our model, we conducted comparative experiments on the FGSC-T dataset using classic object detection algorithms such as Faster R-CNN, SSD, mainstream detection models YOLOv5s and YOLOv7-tiny, as well as the recently proposed models YOLOR-W6, RTDETR-R18, and Gold-YOLOn. In addition to mAP50, mAP50-95, Precision, and Recall, we included Parameters and the size of the model's weight file as evaluation metrics.

From the data in Table 6, it is evident that our model outperforms the traditional SSD and Faster R-CNN models. Compared to the YOLOv5s model, our model improves mAP50, mAP50-95, Precision, and Recall by 1.5%, 2.5%, 3.0%, and 0.4%, respectively, while also having fewer parameters and a smaller weight file size. Although YOLOv7-tiny shows good performance in terms of lightweight characteristics, its performance in maritime vessel detection is significantly inferior to our model.

For the YOLOR-W6 model, mAP50 and mAP50-95 are lower by 1.3% and 1.8%, respectively, compared to our model, and it has a substantially larger number of parameters. Additionally, when compared to the recently proposed model RTDETR-R18 (2023), which

Table 6
Comparison experiments with Faster R-CNN, SSD, YOLOv5s, YOLOv7tiny, and YOLOR-W6

Models	Precision	Recall	mAP50	mAP 50-95	Parameters (/10^6)	Weight (MB)
Faster R-CNN	\	\	76.5	68.3	72	108
SSD	\	\	72.3	\	24.0	92.1
YOLOv5s	77.4	82.2	84.9	73.6	7.1	13.6
YOLOv7tiny	69.8	72.1	76.3	65.2	6.1	11.7
YOLOv8n	77.9	76.9	82.7	72.0	3.0	5.95
YOLOR-W6	\	\	85.1	74.3	79.2	151.8
Rtdetr-r18	82.1	75.6	85.7	73.2	20.1	48.6
Gold-YOLOn	78.3	80.1	82.2	72.4	5.6	11.2
Ours	80.4	82.6	86.4	76.1	6.8	13.4

uses ResNet18 as its backbone, our model's Precision is 1.7% lower, but our Recall is 5.0% higher, with mAP50 and mAP50-95 being 0.7% and 2.7% higher, respectively. Furthermore, our model has significantly fewer parameters and a smaller weight file size than RTDETR-R18.

Compared to the model Gold-YOLOn (2023), although Gold-YOLOn has fewer parameters and a smaller weight file size, our model surpasses it in mAP50, mAP50-95, Precision, and Recall by 4.2%, 3.7%, 2.1%, and 2.5%, respectively.

The comparative analysis above demonstrates that our model exhibits superior performance in the task of maritime vessel detection.

5. Conclusion

We present the enhanced PMF-YOLOv8 model, which utilizes a feature fusion method that combines an improved path aggregation network with an adaptive spatial feature fusion approach. Furthermore, we introduce the R-C2f module and the Inner-MPDIoU loss function. These enhancements effectively tackle the problem of feature loss during the learning process in the existing model, rectify the limitations of previous research that does not fully exploit the shallow features of remote sensing ships, and enable more accurate detection of remote sensing ships.

Compared to the baseline YOLOv8, our model improves mAP50, MAP50-95, precision, and recall by 3.7%, 4.1%, 5.7%, and 2.5%, respectively. This indicates that our improved method enhances model performance.

We also conducted comparative experiments with other advanced models, including Faster R-CNN, SSD, YOLOv5s, YOLOv7tiny, and YOLOR-W6. Our model outperforms these models with an mAP50 improvement of 9.9%, 14.1%, 1.5%, 10.1%, and 1.3%, respectively. Additionally, mAP50 increased by 7.8%, 12.3%, 2.5%, 10.9%, and 1.8%, respectively. The data above shows that our model has superior performance in detecting ships.

Furthermore, the PMF is capable of achieving a detection speed of 84 fps, which is sufficient for real-time detection.

In the future, we will further optimize our network structure for lightweight performance and reduce the number of parameters through techniques such as pruning, quantization, and knowledge distillation. The convolutional layer structure will also be enhanced with depthwise separable convolutions, group convolutions, etc. to minimize computational complexity.

Acknowledgement

This work was supported by the National Key Laboratory of Electromagnetic Space Safety Foundation Program of China (Grant No. 2021JCJQLB055008) and the Innovation and Research Institute of Hebei University of Technology in Shijiazhuang, Science and Technology Cooperation Special Project of Shijiazhuang (SJZ ZXC23001).

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, 2008, 110(3), 346-359. https://doi.org/10.1016/j.cviu.2007.09.014
2. Chen, C.-F. R., Fan, Q., Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. https://doi.org/10.1109/ICCV48922.2021.00041
3. Chen, K., Wu, M., Liu, J., Zhang, C. FGSD: A Dataset for Fine-Grained Ship Detection in High-Resolution Satellite Images. arXiv, March 15, 2020. http://arxiv.org/abs/2003.06832
4. Chen, Z., Liu, C., Filaretov, V. F., Yukhimets, D. A. Multi-Scale Ship Detection Algorithm Based on YOLOv7 for Complex Scene SAR Images. Remote Sensing, 2023, 15(8), 2071. https://doi.org/10.3390/rs15082071
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv, June 3, 2021. https://doi.org/10.48550/arXiv.2010.11929.
6. Duta, I. C., Liu, L., Zhu, F., Shao, L. Pyramidal Convolution: Rethinking Convolutional Neural Networks for Visual Recognition. arXiv, June 20, 2020. http://arxiv.org/abs/2006.11538

7. Fritz, G., Seifert, C., Kumar, M., Paletta. Building Detection from Mobile Imagery Using Informative SIFT Descriptors. Lecture Notes in Computer Science, Springer Berlin Heidelberg: Berlin, Heidelberg, 2005, 3540, 629-638. https://doi.org/10.1007/11499145_64
8. Ghiasi, G., Lin, T.-Y., Pang, R., Le, Q. V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. arXiv, April 15, 2019. https://doi.org/10.1109/CVPR.2019.00720
9. Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C. AugFPN: Improving Multi-Scale Feature Learning for Object Detection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE: Seattle, WA, USA, 2020, 12592-12601. https://doi.org/10.1109/CVPR42600.2020.01261
10. Guoqing, Z., Huang, L., Sun, Q. Improved Oriented R-CNN for Fine-Grained Detection of Remote Sensing Ship Targets. Computer Engineering and Applications, 2021, 1-15.
11. He, K., Zhang, X., Ren, S., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824
12. Hu, D., Yu, M., Wu, X., Hu, J., Sheng, Y., Jiang, Y., Huang, C., Zheng, Y. DGW-YOLOv8: A Small Insulator Target Detection Algorithm Based on Deformable Attention Backbone and WIoU Loss Function. IET Image Processing, 2024, 18(4), 1096-1108. https://doi.org/10.1049/ipr2.13009
13. Li, T., Xu, H., Zeng, W. Ship Classification Method for Massive AIS Trajectories Based on GNN. ournal of Physics: Conference Series, 2021, 2025(1), 012024. https://doi.org/10.1088/1742-6596/2025/1/012024
14. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. Computer Vision and Pattern Recognition, 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106
15. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. Focal Loss for Dense Object Detection. arXiv, February 7, 2018. https://doi.org/10.1109/ICCV.2017.324
16. Liu, D. TS2Anet: Ship Detection Network Based on Transformer. Journal of Sea Research, 2023, 195, 102415. https://doi.org/10.1016/j.seares.2023.102415
17. Liu, S., Huang, D., Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. arXiv, November 24, 2019. http://arxiv.org/abs/1911.09516
18. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. Path Aggregation Network for Instance Segmentation. arXiv, September 18, 2018. https://doi.org/10.1109/CVPR.2018.00913
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C. SSD: Single Shot MultiBox Detector. ECCV 2016. Lecture Notes in Computer Science, 2016, 9905, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2
20. Meng, L., Li, H., Chen, B.-C., Lan, S., Wu, Z., Jiang, Y.-G., Lim, S.-N. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New Orleans, LA, USA, 2022, 12299-12308. https://doi.org/10.1109/CVPR52688.2022.01199
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Computer Vision and Pattern Recognition, 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91
22. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
23. Ren, X., Bai, Y., Liu, G., Zhang, P. YOLO-Lite: An Efficient Lightweight Network for SAR Ship Detection. Remote Sensing, 2023, 15(15), 3771. https://doi.org/10.3390/rs15153771
24. Tan, M., Pang, R., Le, Q. V. EfficientDet: Scalable and Efficient Object Detection. arXiv, July 27, 2020. https://doi.org/10.1109/CVPR42600.2020.01079
25. Tummala, S., Kadry, S., Bukhari, S. A. C., Rauf, H. T. Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling. Current Oncology, 2022, 29(10), 7498-7511. https://doi.org/10.3390/curroncol29100590
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention Is All You Need. arXiv, August 1, 2023. https://doi.org/10.48550/arXiv.1706.03762
27. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. Computer Vision and Pattern Recognition, 2023, 7464-7475. https://doi.org/10.1109/CVPR52729.2023.00721
28. Xu, X., Feng, Z., Cao, C., Li, M., Wu, J., Wu, Z., Shang, Y., Ye, S. An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Seg-

mentation. Remote Sensing, 2021, 13(23), 4779. https://doi.org/10.3390/rs13234779

29. Yan, Z., Li, Z., Xie, Y., Li, C., Li, S., Sun, F. ReBiDet: An Enhanced Ship Detection Model Utilizing ReDet and Bi-Directional Feature Fusion. Applied Sciences, 2023, 13(12), 7080. https://doi.org/10.3390/app13127080
30. Yan, Z., Zemin, Z., Rong, D., Bo, L. Inspection Algorithm of Bottle Defects Based on Improved HOG Characteristics. Modern Manufacturing Engineering, 2019, 460(1), 126.
31. Yin, S. Object Detection Based on Deep Learning: A Brief Review. IJLAI Transactions on Science and Engineering, 2023, 1(02), 1-6.
32. Yin, S., Wang, L., Shafiq, M., Teng, L., Laghari, A. A., Khan, M. F. G2Grad-CAMRL: An Object Detection and Interpretation Model Based on Gradient-Weighted Class Activation Mapping and Reinforcement Learning in Remote Sensing Images. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, 2023, 16, 3583-3598. https://doi.org/10.1109/JSTARS.2023.3241405
33. Yin, S., Wang, L., Wang, Q., Ivanovic, M., Yang, J. M2F2-RCNN: Multi-Functional Faster RCNN Based on Multi-Scale Feature Fusion for Region Search in Remote Sensing Images.
34. Yu, F., Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv, April 30, 2016. https://doi.org/10.48550/arXiv.1511.07122
35. Zhang, T., Zhang, X. A Polarization Fusion Network with Geometric Feature Embedding for SAR Ship Classification. Pattern Recognition, 2022, 123, 108365. https://doi.org/10.1016/j.patcog.2021.108365
36. Zhang, X., Liu, C., Yang, D., Song, T., Ye, Y., Li, K., Song, Y. RFAConv: Innovating Spatial Attention and Standard Convolutional Operation. arXiv, October 12, 2023. http://arxiv.org/abs/2304.03198
37. Zhang, X., Lv, Y., Yao, L., Xiong, W., Fu, C. A New Benchmark and an Attribute-Guided Multilevel Feature Representation Network for Fine-Grained Ship Classification in Optical Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13, 1271-1285. https://doi.org/10.1109/JSTARS.2020.2981686
38. Zhang, Z., Zhang, L., Wang, Y., Feng, P., He, R. ShipRSImageNet: A Large-Scale Fine-Grained Dataset for Ship Detection in High-Resolution Optical Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14, 8458-8472. https://doi.org/10.1109/JSTARS.2021.3104230
39. Zheng, Q., Saponara, S., Tian, X., Yu, Z., Elhanashi, A., Yu, R. A Real-Time Constellation Image Classification Method of Wireless Communication Signals Based on the Lightweight Network MobileViT. Cognitive Neurodynamics, 2024, 18(2), 659-671. https://doi.org/10.1007/s11571-023-10015-7
40. Zheng, Q., Tian, X., Yu, Z., Ding, Y., Elhanashi, A., Saponara, S., Kpalma, K. MobileRaT: A Lightweight Radio Transformer Method for Automatic Modulation Classification in Drone Communication Systems. Drones, 2023, 7(10), 596. https://doi.org/10.3390/drones7100596
41. Zheng, Q., Zhao, P., Wang, H., Elhanashi, A., Saponara, S. Fine-Grained Modulation Classification Using Multi-Scale Radio Transformer with Dual-Channel Representation. IEEE Communications Letters, 2022, 26(6), 1298-1302. https://doi.org/10.1109/LCOMM.2022.3145647
42. Zheng, Q., Zhao, P., Zhang, D., Wang, H. MR-DCAE: Manifold Regularization-Based Deep Convolutional Autoencoder for Unauthorized Broadcasting Identification. International Journal of Intelligent Systems, 2021, 36(12), 7204-7238. https://doi.org/10.1002/int.22586
43. Zheng, Y., Jiang, W. Evaluation of Vision Transformers for Traffic Sign Classification. Wireless Communications and Mobile Computing, 2022(1), 3041117. https://doi.org/10.1155/2022/3041117
44. Zhou, K., Zhang, M., Wang, H., Tan, J. Ship Detection in SAR Images Based on Multi-Scale Feature Extraction and Adaptive Feature Fusion. Remote Sensing, 2022, 14(3), 755. https://doi.org/10.3390/rs14030755