remote sensing

MDPI

Article

Maritime Distress Target Detection Based on Improved RT-DETR: For Robust Small Target Localization

Kun Liu [1,2], Xinbo Chang [3], Zhen Liu [3], Jian Xu [1], Yuhan Zhang [4] and Yang Liu [1,*]

[1] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China; kunliu@hrbeu.edu.cn (K.L.); xujian_bsa@hrbeu.edu.cn (J.X.)
[2] Qingdao Campus, PLA Naval Aviation University, Qingdao 266000, China
[3] School of Automation, Qingdao University, Qingdao 266000, China; changxinbo@qdu.edu.cn (X.C.); zliu@qdu.edu.cn (Z.L.)
[4] School of Future Technology, Harbin Institute of Technology, Harbin 150001, China; 2024111210@stu.hit.edu.cn
* Correspondence: y_liu@hrbeu.edu.cn

Highlights

What are the main findings?

- An improved RT-DETR-based maritime distress target detection method is proposed, integrating SFConv, SPE module, and Focaler-DIoU loss to significantly enhance small target detection and multi-scale feature representation.
- The proposed method achieves an mAP@50 of 0.8347, improving detection performance by 4.51% over the baseline while maintaining end-to-end real-time detection capability.

What is the implication of the main finding?

- The method provides a more accurate and robust solution for detecting small and complex maritime distress targets in dynamic ocean environments.
- It offers practical technical support for real-time UAV-based maritime monitoring and intelligent emergency rescue systems.

Abstract

With the rapid development of maritime transportation and resource development activities, maritime distress events are increasingly frequent, and efficient and accurate target recognition and rescue response methods are urgently needed. The traditional monitoring methods are limited by efficiency and real time, which is difficult to adapt to the complex and changeable marine environment. Therefore, based on the RT-DETR model of transformer architecture, an improved scheme for maritime distress target detection is proposed to improve the small target recognition ability and detection efficiency. Specific improvements include: a small target-focused convolution module (SFConv) is designed to enhance the efficiency of feature extraction and reasoning of small-scale targets; The cross-scale feature interaction optimization module (SPE) is further proposed to improve the ability of multi-scale perception and background suppression; The Focaler-DIoU loss function is introduced to enhance the discrimination performance of the model for difficult samples. On the basis of maintaining the end-to-end detection advantage of RT-DETR, the improvement is of 0.83474, which is 5.7% higher than the original model (0.78964). The accuracy and robustness of the model in complex marine environment is significantly improved, and technical support is provided for the construction of an efficient and intelligent marine monitoring and emergency response system.

Check for updates

Academic Editor: John Trinder

Received: 9 April 2026
Revised: 1 June 2026
Accepted: 2 June 2026
Published: 9 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: target detection; RT-DETR; small target convolution; shallow feature interaction; Focaler-IoU

1. Introduction

With the increasing intensity of maritime transportation and the continuous expansion of marine resource development activities, maritime distress incidents have become frequent. Rapid target identification and rescue response have thus emerged as critical challenges in safeguarding lives and property at sea.

Early studies mainly relied on motion modeling, saliency analysis, and handcrafted feature extraction methods to identify ships or rescue targets in sea-surface images. In addition to optical image-based approaches, researchers have explored various sensing modalities for maritime target detection. For example, radar-based methods have been developed to enhance target perception in complex maritime environments. Su et al. combined radar graph data with graph convolutional networks to improve target representation capability and detection performance [1]. Infrared imaging techniques have also attracted attention due to their robustness to illumination variations. Small maritime target detection using gradient vector field characterization of infrared images improved the perception of weak and small maritime targets by exploiting infrared image characteristics [2]. Furthermore, Maritime Target Detection of PCL System Based on Non-cooperative Pulse Radar utilized passive radar signals for maritime target detection, providing an alternative sensing solution under adverse conditions [3]. To further improve detection robustness, multimodal fusion methods have also been investigated. Maritime Target Detection Method Based on Feature Fusion of Visible and Infrared Images integrated complementary information from visible and infrared sensors, enhancing target discrimination capability in complex maritime scenes [4].

Although these methods achieved promising results in specific scenarios, they generally rely on handcrafted feature design or sensor-specific information, which limits their adaptability and generalization ability in diverse maritime rescue environments. Moreover, their performance is often susceptible to sea clutter, environmental interference, and target appearance variations, motivating the development of data-driven deep learning-based detection frameworks. With the emergence of deep learning, two-stage detection frameworks represented by Faster R-CNN were first introduced into maritime target detection tasks due to their strong feature extraction capability and high detection accuracy. Xu Ziyang improved Faster R-CNN to address image blur and weak small-target representation in UAV maritime images [5], significantly improving offshore target localization performance. Subsequently, one-stage detectors such as the YOLO series gained extensive attention because of their superior inference efficiency and suitability for real-time UAV deployment. Hou Ying et al. optimized YOLOv8 for UAV aerial imagery and enhanced the perception capability of small maritime objects [6]. Fu et al. proposed the DLSW-YOLOv8n network by incorporating lightweight design and maritime-scene-aware feature enhancement strategies, improving detection robustness under complex sea conditions [7]. In addition, researchers have attempted to integrate multi-scale feature fusion, attention mechanisms, and contextual information modeling into maritime object detection frameworks to alleviate challenges caused by small targets, occlusion, and dynamic ocean backgrounds. Xu Ming et al. introduced spatio-temporal and global contextual information to improve target recognition stability in complex maritime environments [8], while Zhang Yinsheng et al. enhanced local feature extraction to strengthen underwater and sea-surface target perception capability [9]. Although CNN-based detectors have

achieved remarkable success in maritime object detection through advances in feature extraction, attention mechanisms, and multi-scale fusion strategies, their receptive fields remain inherently local. As a result, capturing long-range contextual dependencies and distinguishing small targets from complex sea-surface backgrounds remain challenging, especially in UAV-based maritime rescue scenarios where targets are sparse, small, and easily confused with waves or reflections. These limitations have motivated researchers to explore Transformer-based architectures, which provide a more effective mechanism for modeling global contextual information and long-range feature interactions. Recently, Transformer-based detection frameworks have emerged as a new research trend due to their powerful global dependency modeling ability and end-to-end optimization mechanism. Compared with conventional CNN-based detectors, Transformer architectures can better capture long-range contextual information and improve robustness in cluttered maritime scenes. For example, Liu et al. proposed a Transformer-based maritime rescue detection framework that demonstrated improved performance for small and densely distributed rescue targets [10]. Among these methods, RT-DETR achieves efficient real-time detection by combining CNN feature extraction with Transformer encoding while eliminating the need for manually designed anchors and Non-Maximum Suppression (NMS) [11,12]. Its end-to-end architecture provides a favorable balance between detection accuracy and inference efficiency, making it particularly suitable for UAV-based maritime rescue applications. Nevertheless, despite its advantages, RT-DETR still faces challenges when applied to maritime distress target detection. Maritime distress targets are typically characterized by small object sizes, blurred boundaries, large scale variations, and complex sea-surface interference. Under such conditions, the feature representation capability of RT-DETR for small targets remains insufficient, while cross-scale feature interaction and localization accuracy still require further improvement.

Motivated by these challenges, this paper adopts RT-DETR as the baseline detector and develops a maritime distress target detection network specifically designed for UAV-based rescue scenarios. To address the aforementioned limitations, three dedicated modules are introduced from the perspectives of feature representation enhancement, cross-scale information interaction, and bounding-box regression optimization, namely Small-object Focused Convolution (SFConv), Shallow Path Enhancement for Cross-scale Interaction (SPE), and Focaler-DIoU loss reconstruction [13].

1. Small-object Focused Convolution (SFConv): To improve detection performance for small maritime distress targets, we designed a convolutional module specifically optimized for small-scale object detection. Combining three-layer convolution extraction, skip-connection augmentation, and structural reparameterization for inference acceleration, it effectively enhances small-object detection accuracy and model deployment performance.
2. Shallow Path Enhancement for Cross-scale Interaction (SPE): Addressing challenges such as the difficulty in detecting small maritime distress targets, significant scale variations, and strong background interference, this paper further proposes the SPE module. It focuses on optimizing the role of shallow-level features (P2) in cross-scale feature fusion. This module strengthens the upward propagation pathway of shallow semantic information, enhances the fine-grained expressive capability within the feature pyramid, and significantly boosts the model's perception of multi-scale targets in complex backgrounds.
3. Focaler-DIoU Loss Reconstruction: The Focaler-DIoU loss function is introduced, reconstructing the DIoU loss through a linear interval mapping mechanism. This enhances the model's attention and discrimination capabilities on challenging detection samples, thereby improving the overall robustness of the detection system.

In summary, the proposed improvements retain the end-to-end detection advantages of RT-DETR while specifically addressing common challenges in maritime distress target detection—such as ambiguity, weak targets, and complex background interference—providing robust support for efficient and precise intelligent maritime monitoring systems.

2. Materials and Methods

2.1. Overview of the RT-DETR Network Model

Transformer-based object detectors (DETRs) have garnered significant attention in recent years. Their key advantage lies in eliminating manually designed components—such as non-maximum suppression (NMS)—commonly found in traditional detection methods. This simplifies the detection workflow and enables end-to-end modeling capabilities. Compared to anchor-based approaches, DETRs model object detection as a set prediction problem. Leveraging self-attention mechanisms, they achieve stronger global modeling capabilities and object relationship construction. However, these methods generally suffer from high inference latency and slow convergence. This stems primarily from the high computational complexity of their encoder and decoder modules, coupled with limited efficiency in multi-scale feature fusion. These factors severely limit DETR's applicability in speed-sensitive scenarios such as real-time object detection and edge deployment. To address these challenges, RT-DETR (Real-Time Detection Transformer) was proposed as an end-to-end object detection framework balancing accuracy and speed. It retains DETR's modeling advantages while significantly improving detection efficiency. The overall structure of RT-DETR is illustrated in Figure 1.

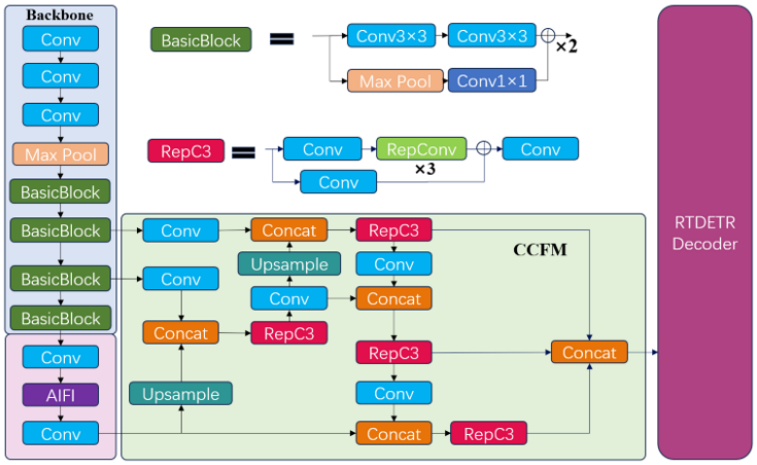


Figure 1. RT-DETR network model structure.

For clarity of the network architecture illustrated in Figures 1–3, the main abbreviations are defined as follows: Conv denotes the standard convolution layer for feature extraction; MaxPool represents the max pooling operation used for downsampling and receptive field expansion; SFConv refers to the proposed Small-object Focused Convolution module for enhancing fine-grained feature representation of small targets; AIFI denotes the Attention-based Intra-scale Feature Interaction module for strengthening contextual feature modeling; Concat indicates channel-wise feature concatenation for multi-scale feature fusion; Upsample refers to spatial resolution enhancement of feature maps; and RepC3 denotes the re-parameterized C3 module, which improves feature extraction efficiency while maintaining low inference cost.

The RT-DETR network model primarily consists of three core components: the backbone network, the hybrid encoder, and the transformer decoder. The backbone network

typically employs lightweight convolutional neural networks (such as ResNet or MobileNet) to extract multi-scale features [14]. The Hybrid Encoder achieves efficient intra- and inter-scale feature interaction by integrating self-attention mechanisms with local perception capabilities; the decoder introduces an IoU-aware Query Selection mechanism to intelligently filter initial queries with high regression quality from candidate regions, reducing redundant computations and enhancing target localization accuracy. Finally, by combining the output from the Auxiliary Head to generate precise bounding boxes and category information, the model achieves efficient and accurate detection performance.

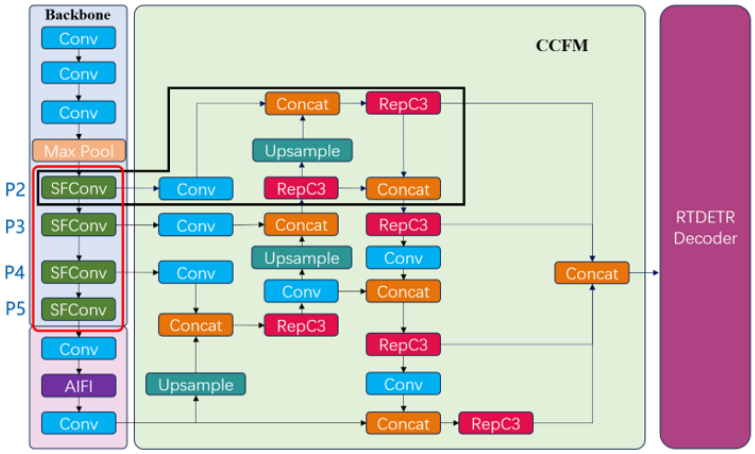


Figure 2. The whole improvement model based on RT-DETR.

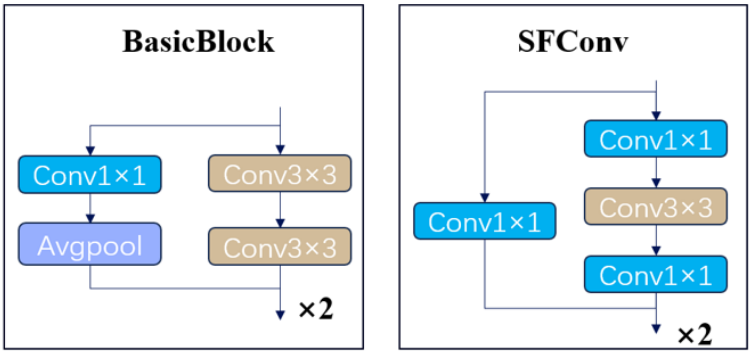


Figure 3. Structure comparison of SFConv and BasicBlock.

2.2. Overall Structure of the Improved Model Based on RT-DETR

To address common challenges in maritime distress target detection—such as small target scales, complex backgrounds, and low target contrast—this paper performs targeted optimizations on the RT-DETR network architecture to enhance detection performance in complex maritime environments. First, to strengthen the backbone network's feature extraction capabilities for small targets and complex backgrounds, an improved Small-Target Focused Convolution (SFConv) is proposed to replace the traditional BasicBlock residual structure. Second, to improve the model's perception of small-scale information, a shallow feature path P2 is introduced, enhancing its fusion capability with multi-scale features to better address detection challenges involving targets with significant size variations. Finally, considering the varying difficulty of detection samples—particularly hard-to-localize small objects—this paper introduces the Focaler-DIoU loss function. This guides the model to prioritize bounding box regression for challenging samples during training, thereby enhancing detection robustness and accuracy. These improvements systematically enhance the RT-DETR network's adaptability and detection performance for maritime distress target

detection by addressing three critical aspects: feature extraction, feature fusion, and loss function optimization. The improved model based on RT-DETR is illustrated in Figure 2.

2.3. Improved Convolution Module SFConv

2.3.1. Design Motivation and Overall Structure

Drawing inspiration from the "multi-layer convolution extraction + residual enhancement + reparameterization acceleration" approach in the Swift Parameter-free Attention Network (SPAN) [15], this paper proposes an improved small-object-focused convolution, SFConv, to replace the BasicBlock residual module in traditional backbone networks. This aims to enhance the model's feature extraction capabilities for small objects and complex background images. The two-layer convolutional structure of the traditional BasicBlock module limits its feature representation capability and suffers from insufficient receptive field when processing distant, small-scale objects. This issue is particularly pronounced against wave-like textures, where abundant high-frequency, repetitive noise features easily obscure the weak response signals of small targets. This paper designs SFConv's main branch as a "bottleneck-style" convolutional structure. This design offers flexibility in channel adjustment while aggregating spatial context through convolution. It effectively preserves fine-grained contours and texture features of small targets without excessive smoothing, significantly enhancing feature reconstruction and long-range perception capabilities. Simultaneously, the module introduces skip connections as parallel branches, improving information flow stability and gradient propagation efficiency. Furthermore, considering the stringent real-time and computational constraints of maritime UAV platforms, a structural reparameterization mechanism is incorporated into this module. This significantly reduces computational overhead while maintaining detection accuracy. A comparison diagram between SFConv and BasicBlock structures is shown in Figure 3.

In the application scenario of maritime target detection, complex sea surface environments commonly exhibit large-area wave texture features characterized by strong periodicity and high-frequency properties. Such background information tends to accumulate layer by layer within deep convolutional networks, causing aliasing with target features. This gradually drowns out the weak responses of small targets. The SFConv module employs a bottleneck-style convolutional structure to reorganize and compress features along the channel dimension. It extracts local spatial context information within a controlled receptive field, effectively limiting the cross-scale propagation of background texture features and enhancing the saliency of small targets' local discriminative features. Simultaneously, residual connections provide a stable information pathway for feature propagation, helping mitigate the gradual attenuation of small target features as layers increase in deep networks. Combined with a structural reparameterization mechanism, SFConv maintains strong feature modeling capabilities during training while achieving efficient computation in inference through a simplified, stable forward structure. This enhances the model's robustness against small targets in the context of sea wave interference at the mechanism level.

2.3.2. Implementation Details of SFConv

The main branch of the SFConv module designed in this paper employs a three-layer sequential convolution architecture specifically optimized for small object detection tasks. This design effectively reduces parameter count and computational complexity while preserving model expressiveness. The primary function of the first convolutional layer is to perform channel dimension expansion on the input feature map. This process not only enhances the network's nonlinear expressive capability but also provides richer dimensional support for subsequent spatial feature extraction. The intermediate convolutional layer

serves as the module's core computational unit, primarily sensing and extracting local spatial contextual information. Against wave backgrounds, repetitive high-frequency texture features easily overwhelm the weak response signals of small objects. A single 3×3 convolutional layer effectively preserves local contours and detailed texture information of small objects without excessive smoothing. Compared to two consecutive convolutions, a single layer reduces information compression and loss, preserving small targets' local texture features more completely. The third convolution layer compresses the number of channels, aligning the module's output feature map size with the desired dimension. This layer not only enhances computational efficiency but also suppresses redundant feature propagation to some extent, helping the network focus more on expressing critical regional information.

Although the main branch of this module employs a richer network architecture to enhance the model's ability to extract detailed features, it inevitably introduces additional computational overhead. To better adapt to the computational resource and response speed requirements of actual drone deployment environments, this paper utilizes Structural Reparameterization (SRP) to equivalently transform the training-time multi-branch structure into a single convolutional layer during inference. Specifically, SFConv consists of a learnable 1×1 skip convolution branch and a 1×1–3×3–1×1 sequential convolution branch. Different from RepVGG, SFConv does not adopt a parameter-free identity mapping branch; instead, all shortcut information is modeled via a learnable 1×1 convolution. During inference, Batch Normalization and all convolutional branches are algebraically fused into an equivalent single convolution kernel, and the skip branch is fully included in this fusion process. Therefore, no independent identity or auxiliary branches remain in the deployed model.

First proposed in architectures such as RepVGG and RefConv [16,17], structural reparameterization enables the use of complex multi-branch structures during training to enhance representation capacity, while merging them into a lightweight equivalent structure during inference. Compared with RepVGG, which primarily re-parameterizes parallel 3×3, 1×1, and identity branches, SFConv introduces a deeper transformation path (1×1–3×3–1×1), which improves feature modeling capability within a single fused kernel. Compared with RefConv, SFConv does not involve dynamic kernel modulation or feature-dependent re-weighting mechanisms; instead, it follows a purely structural equivalence formulation with deterministic kernel fusion. This approach significantly reduces parameter size and simplifies the forward propagation path without sacrificing detection accuracy, thereby improving inference speed and deployment efficiency. It is particularly well-suited for resource-constrained edge inference environments on maritime UAVs.

2.4. Cross-Scale Feature Interaction Optimization

To more effectively address challenges in maritime distress target detection—such as small target detection difficulties, large scale variations, and strong background interference—we optimized the role of shallow features (P2) in cross-scale feature fusion. This led to the SPE module: introducing a shallow feature path (P2) to enhance cross-scale feature fusion capabilities. The improved structure is shown in the red box in Figure 4.

2.4.1. Introduction of Shallow Features

The original RT-DETR model assumes that deeply extracted features contain the most useful information, thus performing intra-scale interactions only on the deepest feature layer (P5) and cross-scale interactions on the deeper layers P3 and P4, while neglecting the shallowest layer (P2). However, shallow features often contain more image details with higher resolution, and their smaller receptive fields are better suited for small object detection.

Considering maritime distress targets are typically small and easily overlooked during high-altitude or long-range detection, this paper incorporates shallow feature maps from the backbone network into the cross-scale interaction process. This enables higher-resolution, richer spatial detail to propagate to the detection head. Compared to traditional approaches that fuse only P3-P5 features, adding the P2 path significantly enhances the model's perception of small-scale targets, improving detection performance for small objects.

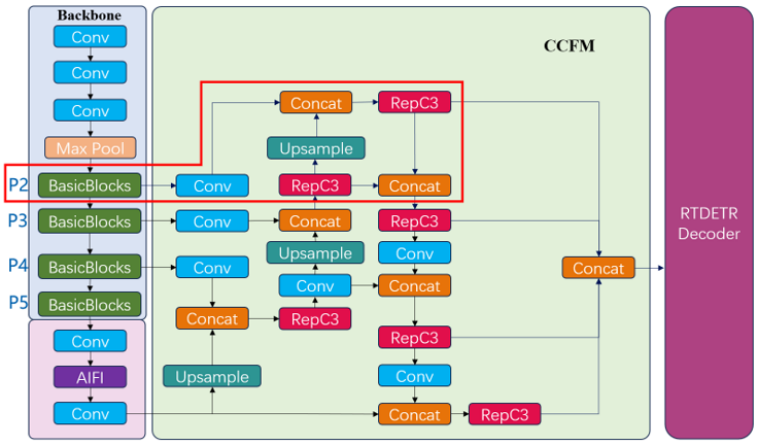


Figure 4. Structural diagram of shallow feature introduction and path enhancement.

2.4.2. Path-Enhanced Network

This paper constructs a multi-level downsampling and fusion path-enhanced network. As illustrated in Figure 3, building upon the original three-level scale fusion, it performs layer-by-layer downsampling starting from the P2 layer while integrating features with higher layers. The red block diagram represents the P2 layer feature fusion process. Specifically, P2 features are progressively downsampled using stride-2 convolution to match the spatial resolutions of P3, P4, and P5, while higher-level features are correspondingly upsampled via nearest-neighbor interpolation to ensure consistent feature alignment across scales. After alignment, a 1×1 convolution is applied to unify channel dimensions before feature fusion, ensuring that the decoder input dimensionality remains unchanged compared to the baseline architecture. Concurrently, multi-level RepC3 modules achieve deep feature integration. This network strengthens feature propagation across scales, significantly mitigating interference from complex sea backgrounds (e.g., wave patterns, light reflections) on detection accuracy. Furthermore, deeper feature interaction enhances the model's robustness toward targets with large scale variations (e.g., coexisting vessels and persons in distress).

In terms of computational complexity, although the introduction of the P2 branch increases the number of fusion operations, the additional cost is mainly attributed to lightweight operations such as downsampling, interpolation, and 1×1 convolutions. No heavy convolutional layers are introduced, and thus the overall increase in GFLOPs remains moderate, providing a favorable trade-off between accuracy improvement and computational efficiency.

Overall, the proposed multi-level fusion design effectively enhances shallow-to-deep feature propagation and significantly improves the model's ability to capture fine-grained maritime targets under complex and dynamic sea environments.

2.5. Focaler-DIoU

During maritime distress target detection, varying levels of difficulty exist among targets. Smaller, harder-to-locate targets are typically classified as difficult detection samples. To address detection tasks of varying difficulty, the model must prioritize bounding box regression

for different levels of complexity. To enable the model to focus on more challenging detection samples, this paper introduces the Focaler-IoU loss function. This loss function reconstructs the IoU loss using a linear interval mapping approach, as shown in Equation (1).

$$\mathrm{IoU}_{\mathrm{focaler}} = \begin{cases} 0, & \mathrm{IoU} < d, \\ \frac{\mathrm{IoU}-d}{u-d}, & d < \mathrm{IoU} < u, \\ 1, & \mathrm{IoU} > u. \end{cases} \tag{1}$$

where $\mathrm{IoU}_{\mathrm{focaler}}$ represents Focaler-IoU, $d, u \in [0, 1]$. During the training sample screening process, this paper ignores samples with already satisfactory prediction performance by setting an upper threshold u. When IoU exceeds u, the sample prediction is highly accurate, and its gradient information is limited. Continuing to include it in training may cause gradient redundancy and dilute the learning signals from moderately difficult samples. Therefore, through experimentation, this paper treats samples with IoU greater than 0.95 as "converged samples" and excludes them from loss calculations. This guides the model to focus its training attention on areas that have not yet been sufficiently learned. Simultaneously, for challenging samples within the IoU range, this paper fixes the linear mapping method of d = 0 to prevent further weakening of their training weights. This ensures these high-information samples fully contribute to model optimization. The above strategy removes redundant simple samples while maximally preserving the training contribution of critical challenging samples, thereby enhancing the model's overall detection performance in complex maritime backgrounds. The loss function is defined as shown in Equation (2).

$$L_{\mathrm{Focaler-IoU}} = 1 - \mathrm{IoU}_{\mathrm{focaler}}. \tag{2}$$

Applying Focaler-IoU to the DIoU loss function yields Focaler-DIoU, whose loss function is expressed as shown in Equation (3) [18].

$$L_{\mathrm{Focaler-DIoU}} = L_{\mathrm{DIoU}} + \mathrm{IoU} - \mathrm{IoU}_{\mathrm{focaler}}. \tag{3}$$

where L_{DIoU} denotes the loss function for DIOU, where IoU is the ratio of the intersection to the union of the predicted box and the ground truth box areas.

3. Results

3.1. Experimental Dataset

To bridge the gap between terrestrial and marine visual systems, Varga et al. from the University of Tübingen, Germany, released the large-scale SeaDronesSee dataset for computer vision object detection and tracking at the 2022 WACV conference [19]. This dataset comprises over 54,000 annotated images containing 400,000 drone instances within a range of 5–260 m and 0–90 degrees. Detection targets are categorized into five classes: swimmers, boats, jet skis, lifesaving equipment, and buoys. Given the practical challenges of collecting real-world maritime distress targets, SeaDronesSee effectively simulates their states, making it a more realistic choice for this experiment's dataset. To accelerate experimentation and expedite practical maritime rescue applications, hierarchical sampling was employed to select a subset of images as a sub-dataset. Comparative experiments were conducted on the full SeaDronesSee dataset provided by the official source to validate the proposed method's generalization capability under complete data distribution. The sub-dataset comprises 893 training images and 155 validation images. The statistical summary of the sub-dataset is presented in Table 1.

Table 1. Introduction to the SeaDronesSee Sub-dataset.

Category	Training Set	Validation Set
Number of Sub-datasets	893	155
Number of Annotations (Buoy)	434	59
Number of Annotations (Boat)	1380	234
Number of Annotations (Swimmer)	3666	624
Number of Annotations (Jet Ski)	228	34
Number of Annotations (Life-saving Equipment)	88	36

3.2. Experimental Environment and Parameter Settings

The processor used in this experiment is: 13th Gen Intel® Core™ i7-13700KF (24 CPUs) 3.4 GHz, RAM: 64 GB, equipped with an NVIDIA GeForce RTX 4090D GPU. The operating environment is Python 3.9.18 running on the Windows 11 operating system. Based on this setup, PyTorch version 2.3.0 was built and accelerated using CUDA 11.8 and cudnn 8.700. Selected experimental parameters are shown in Table 2.

Table 2. Some Experimental Parameters.

Parameter	Value
Input Image Size	640 × 640
Initial Learning Rate	0.0001
Final Learning Rate	1.0
Momentum	0.9
Batch Size	16
Number of Epochs	300
Warm-up Epochs	16
Optimizer	AdamW

The loss function curve of RT-DETR-r18 shown in Figure 5 illustrates the trend of validation set loss values during the training process of the baseline model. The figure reveals that the model rapidly learned the primary features of the data in the early stages, with the loss curve declining sharply, indicating swift adaptation to the initial fitting. In the middle phase, the model exhibited a slower but still steady decline, signifying ongoing optimization. After 150 iterations, the loss value fluctuates minimally, indicating gradual convergence and stabilization. At 300 iterations, the loss function continues to decline, suggesting minor optimization potential remains—evidence that the model has not yet overfitted. Therefore, this experiment employs 300 training iterations to ensure thorough training without overfitting.

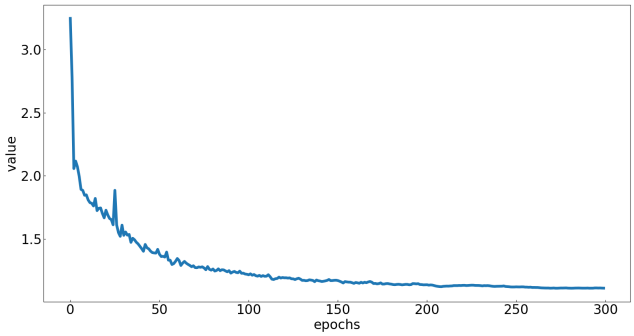


Figure 5. Loss function curve of RT-DETR-r18 verification set.

3.3. Model Performance Evaluation Metrics

This paper employs accuracy (P), recall (R), and mean average precision (mAP) as experimental performance evaluation metrics. Their calculation formulas are as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{4}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P(R)dR}, \tag{6}$$

$$\mathrm{mAP} = \frac{\sum \mathrm{AP}}{\mathrm{NC}} \times 100\%. \tag{7}$$

Among these, TP represents the number of correctly predicted positive samples, FP represents the number of incorrectly predicted positive samples, TN represents the number of correctly predicted negative samples, FN represents the number of incorrectly predicted negative samples, NC indicates the number of categories contained in the sample, AP denotes the data mean for each category of samples, and mAP denotes the average accuracy mean across all samples.

Furthermore, to more precisely evaluate the model's detection performance and ensure more accurate model localization, the experiment employs threshold mAP with a value of 0.5 as the primary metric, denoted as mAP@50.

3.4. Experimental Evaluation of Improved Modules

3.4.1. Comparative Experiment of SFConv Convolution Modules

1. Convolution Module Comparison Experiment
   To further enhance the feature extraction capability of the detection model, we systematically compared various improved convolution modules—including SFConv—by replacing the BasicBlock in the backbone network.
   Specifically, we evaluated the following mainstream convolution architectures: PConv (Perturbed Convolution) [20], DBB (Diverse Branch Block) [21], DEConv (Detail-enhanced Convolution) [22], DRB (Dilated Reparam Block) [23], DualConv [24], DySnake Conv [25], RFCBAMConv [26], WTConv [27], and our proposed SFConv. All modules were uniformly replaced based on the rtdetr-r18 main branch and evaluated under identical training configurations. Experimental results are shown in Table 3. Figure 6 presents a visual comparison of the data in Table 3.

Table 3. Comparison of Performance Metrics of Various Convolution Modules.

Method	P	R	mAP@50	mAP@50:95
rtdetr-r18	0.9221	0.7619	0.7896	0.4317
PConv	0.9373	0.7468	0.7804	0.4194
DBB	0.9101	0.7408	0.7665	0.4226
DEConv	0.9317	0.7549	0.7759	0.4186
DRB	0.9125	0.7220	0.7335	0.3861
DualConv	0.9179	0.7596	0.7669	0.4130
DySnakeConv	0.9252	0.7244	0.7496	0.4239
RFCBAMConv	0.9136	0.7131	0.7440	0.3969
WTConv	0.9386	0.7332	0.7615	0.4127
SFConv	0.9420	0.7493	0.7920	0.4433

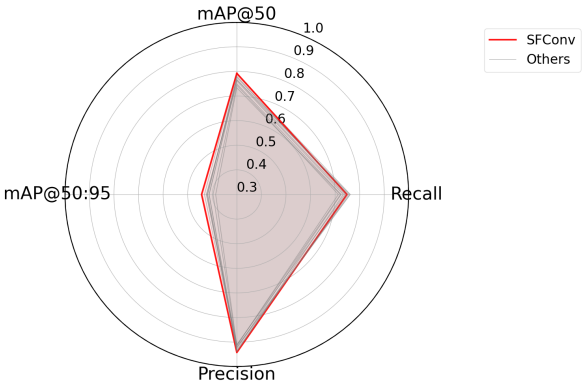


Figure 6. Comparison chart of detection performance between SFConv and other convolution modules.

Based on the experimental results, SFConv demonstrated particularly outstanding performance on metric mAP@50, achieving a score of 0.79203—the highest among all models and significantly outperforming the baseline model. The accuracy (P) metric reached 0.94208, indicating robust stability in target discrimination. It also achieved 0.44332 on the mAP@50 : 95 metric, significantly surpassing most comparison modules (e.g., RFCBAMConv's 0.39629 and DRB's 0.38613).

These results demonstrate that while modules like PConv and WTConv also enhance model performance to some extent, SFConv more effectively captures and transmits multi-scale feature information while maintaining structural compactness, yielding superior detection accuracy and generalization capabilities.

2. Comparative Experiments on Structural Reparameterization Mechanism

To validate the effectiveness of SFConv's structural reparameterization mechanism, two sets of comparative experiments were designed (Table 4). Experiments were conducted under two scenarios: a baseline configuration (SFConv only) and an enhanced configuration (SFConv + SPE + Focaler-DloU). These scenarios compared the computational efficiency and complexity of models with and without structural reparameterization. Visual comparison results are shown in Figure 7.

Table 4. Performance Comparison of Models With or Without Structural Re-parameterization.

Model	Re-Parameterization	Layers	Parameters	GFLOPs	FPS
SFConv	No	379	2,346,056	67.1	43.8
SFConv	Yes	299	19,879,464	57.0	46.8
SFConv + SPE + FocalerDIoU	No	418	22,368,200	88.3	38.7
SFConv + SPE + FocalerDIoU	Yes	338	18,601,608	78.2	40.6

After introducing structural reparameterization in the baseline configuration, the number of model layers decreased from 379 to 299 (a reduction of 21.1%), the number of parameters decreased from 23.6 M to 19.9 M (a 15.7% reduction), computational complexity (GFLOPs) optimized from 67.1 to 57.0 (a 15.1% reduction), while inference speed (FPS) improved by 6.8% (from 43.8 to 46.8). This demonstrates that the mechanism significantly reduces model complexity while enhancing inference efficiency by dynamically merging network branches.

In the enhanced configuration, structural reparameterization continues to show stable advantages when combined with SFConv, SPE, and Focaler-DloU modules: the number of layers decreased by 19.1%, the number of parameters dropped from 22.4 M to 18.6 M, computational complexity decreased by 11.4%, and FPS increased by 4.9%. Although the overall computational demand of SFConv + SPE + Focaler-DloU increases, structural

reparameterization effectively controls computational cost growth while maintaining real-time performance advantages.

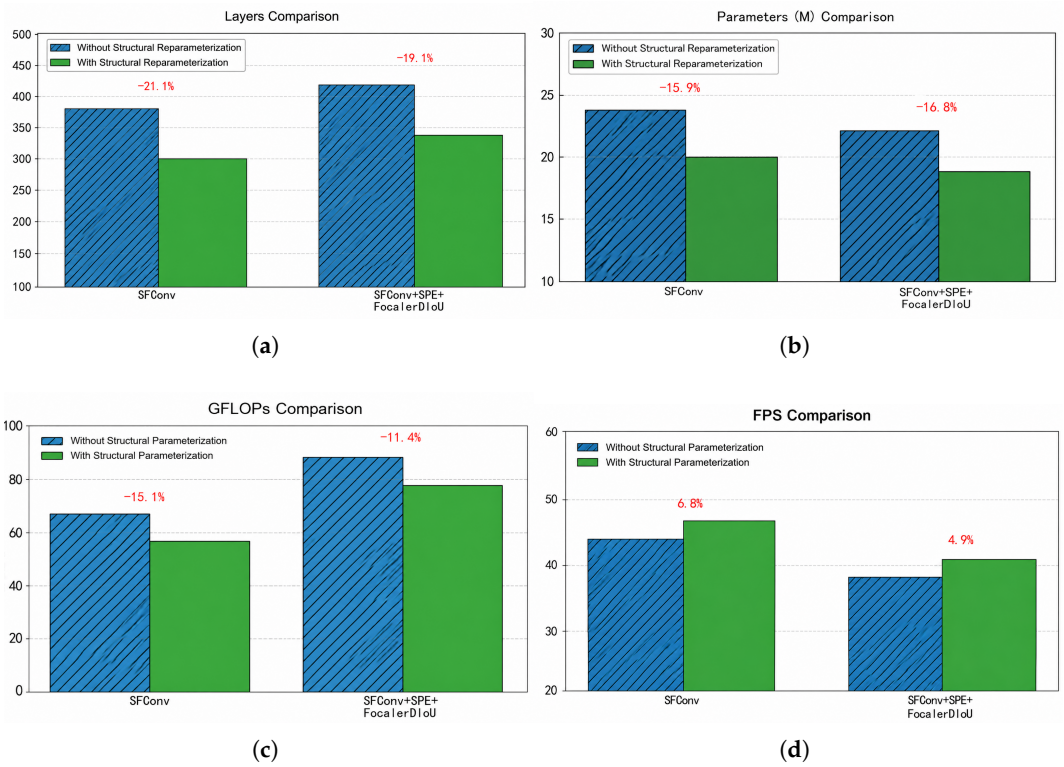


Figure 7. Performance comparison diagram of the model with or without structural reparameterization (**a**) Layers Comparison; (**b**) Parameters Comparison; (**c**) GFLOPs Comparison; (**d**) FPS Comparison.

Experimental results demonstrate that the structural reparameterization mechanism achieves a balance between model lightweighting and inference acceleration through multi-branch structure learning during training and parameter fusion during inference.

3.4.2. Parameter Sensitivity Analysis of Loss Function

To validate the rationality of the threshold parameter u in Focaler-DIoU and analyze its impact on small maritime target detection performance, this study conducted sensitivity analysis experiments with varying u values while keeping the network architecture, training strategy, and other hyperparameters constant. Only u was altered in these experiments, with all other settings matching the main experiments to ensure comparability of results.

Based on practical task requirements, $u \in [0.75, 0.80, 0.85, 0.90, 0.95, 1]$ were selected for comparative experiments.

As shown in Figure 8, as threshold u gradually increases, mAP@50 continues to improve within a certain range, reaching its optimal performance at $u = 0.95$. When u is further increased to 1.0, detection performance shows a significant decline. This indicates that excessively high thresholds cause premature exclusion of some moderately difficult small target samples, hindering the model's feature learning capabilities in complex sea surface backgrounds. Therefore, this paper ultimately selects $u = 0.95$ as the default value. This setting is also consistent with the original configuration in Focaler-IoU.

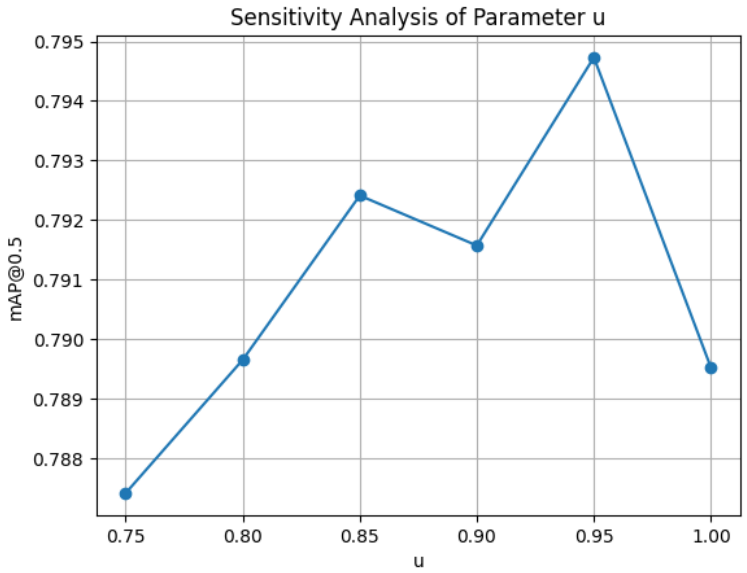


Figure 8. Sensitivity Analysis of Parameter u.

3.4.3. Loss Function Comparison Experiment

Table 5 summarizes the performance metrics of the RT-DETR-r18 model using different loss functions. It can be observed that the mAP@50 value reached 0.78964. Meanwhile, Focaler-DIoU—introduced specifically for maritime distress target detection—improved Precision and Recall while elevating mAP@50 to 0.79473, representing a 0.51% increase over the baseline model. Compared to other loss functions, Focaler-DIoU simultaneously strengthens both classification and regression branches during training. This not only significantly improves mAP@50 but also concurrently enhances the model's accuracy (P) and recall (R), resulting in detection performance that comprehensively outperforms traditional methods across all metrics.

Figure 9 shows the variation curve of mAP@50 when training RT-DETR-r18 using various loss functions. The blue line represents the baseline RT-DETR-r18 model, the red line represents the RT-DETR-r18 model using Focaler-DIoU, and models using other loss functions are indicated by black lines. The green box in the figure zooms in on the details between 240 and 300 iterations. It is clearly observable that the Focaler-DIoU curve consistently outperforms the baseline and other methods, ultimately stabilizing between 0.79 and 0.80. This demonstrates that employing Focaler-DIoU effectively enhances the model's detection accuracy.

Table 5. Comparison of Performance Metrics of RT-DETR-r18 with Different Loss Functions.

Method	P	R	mAP@50	mAP@50:95
rtdetr-r18	0.9221	0.7619	0.7896	0.4317
+ MPDIoU [28]	0.9355	0.7569	0.7832	0.4271
+ InnerDIoU [29]	0.9257	0.7429	0.7892	0.4180
+ Focaler-GIoU	0.9321	0.7205	0.7695	0.4207
+ Focaler-EIoU	0.9361	0.7230	0.7430	0.4117
+ Inner-MPDIoU	0.9200	0.7432	0.7848	0.4323
+ Focaler-MPDIoU	0.9378	0.7366	0.7658	0.4112
+ Proposed	0.9231	0.7602	0.7974	0.4244

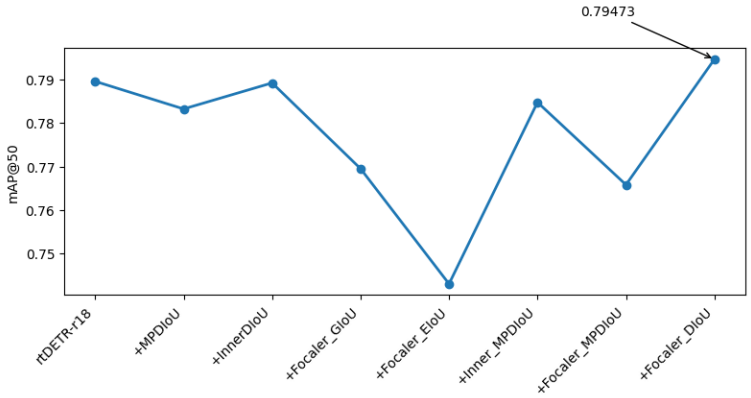


Figure 9. The comparison curve of mAP@50 when RT-DETR-r18 is trained with each loss function.

3.5. Ablation Studies

In this ablation study, we evaluated the detection performance of the baseline model RT-DETR-r18 alongside three key modules (SFConv, SPE, Focaler-DIoU) and their various combinations, focusing on the maritime distress target detection task. The performance of each experimental module across four metrics (P, R, mAP@50, mAP@50:95) is shown in Table 6. And the visualization effect of the ablation experiment is shown in Figure 10.

When SFConv, SPE, and Focaler-DIoU were added individually, increased from the baseline model's 0.7896 to 0.79203, 0.79697, and 0.79474, respectively. Combining SFConv and SPE with the baseline model further improved to 0.80546. Finally, the combination of SFConv + SPE + Focaler-DIoU achieved an value of 0.83474, representing a 4.51% improvement over the baseline model. As shown in Figure 10, exhibits a continuous upward trend with the progressive addition of modules, reaching its peak when all three are integrated. Each individual module provides relatively limited improvement over the baseline, while the full integration of SFConv, SPE, and Focaler-DIoU yields a more significant performance gain. This is mainly because the three modules contribute to different yet complementary aspects of the detection pipeline. Specifically, SFConv enhances multi-scale feature representation and strengthens the extraction of fine-grained semantic information, which is crucial for small and low-contrast maritime targets. SPE further improves feature aggregation by enhancing spatial perception and suppressing background interference in complex sea environments. Meanwhile, Focaler-DIoU optimizes the bounding box regression process by focusing more on hard-to-regress samples and improving localization accuracy. When these modules are combined, SFConv provides stronger feature representations, SPE improves spatial discrimination, and Focaler-DIoU refines localization quality, forming a mutually reinforcing mechanism across feature extraction, feature fusion, and box regression stages. This complementary interaction leads to a more substantial overall improvement in detection performance, especially in challenging small-object scenarios.

Table 6. Ablation Study on Detection Performance of Each Module.

Method	P	R	mAP@50	mAP@50:95
RT-DETR-r18	0.9221	0.7619	0.7896	0.4317
+ SFConv	0.9420	0.7479	0.7920	0.4433
+ SPE	0.9381	0.7735	0.7969	0.4554
+ Focaler-DIoU	0.9239	0.7690	0.7947	0.4244
+ SFConv + SPE	0.9007	0.7695	0.8054	0.4445
+ SFConv + SPE + Focaler-DIoU	0.9177	0.7955	0.8347	0.4719

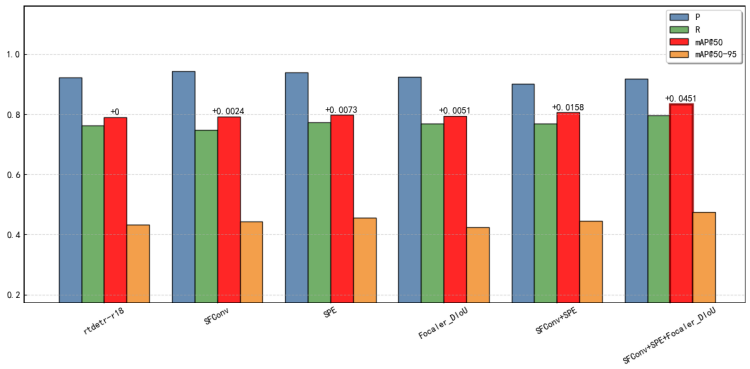


Figure 10. Gain histogram of mAP@50 in Ablation Experiment of each module.

3.6. Full Dataset Validation Experiment

To further validate the generalization capability of the proposed method under the full data distribution, this paper conducted comparative experiments on the official SeaDronesSee full dataset for both the pre- and post-improvement models. This experiment strictly followed the same training strategy and evaluation metrics as the previous experiments, with the sole modification being the expansion of the training data from a subset to the full dataset to ensure comparability of results. In addition, several recent advanced lightweight detectors, including EfficientViT, YOLOv10s, YOLO11s, YOLOv12s and YOLOv13s, which have demonstrated strong performance in aerial or maritime small-target detection tasks, were also introduced for comparison.

The experimental results are shown in Table 7. It can be observed that under the full dataset conditions, the proposed RT-DETR-SSF (RT-DETR-r18 + SFConv + SPE + Focaler-DIoU) model still achieved the best detection performance in terms of the core evaluation metric mAP@50, outperforming methods based on EfficientViT, YOLOv10s, etc. The model incorporating the SFConv module maintained stable performance gains consistent with the conclusions drawn from the subset ablation experiments, demonstrating strong robustness and generalization capability under more complex and comprehensive data distributions.

Table 7. Performance Comparison on the Full SeaDronesSee Dataset.

Method	P	R	mAP@50	mAP@50:95
RT-DETR-r18	0.8409	0.8337	0.8559	0.5809
EfficientViT	0.9021	0.6814	0.7166	0.4032
YOLOv10s	0.8730	0.5908	0.6197	0.4032
YOLO11s	0.9018	0.5757	0.5975	0.3394
YOLOv12s	0.8969	0.5450	0.5750	0.3227
YOLOv13s	0.8980	0.5252	0.5123	0.5690
RT-DETR-SSF	0.8500	0.8392	0.8599	0.5857

The ablation experiments conducted on a subset of data in the preceding section primarily analyzed how different module designs influence the trend of model performance changes, thereby reducing experimental costs and enhancing analytical efficiency. In contrast, the comparative experiments based on the full dataset in this section validate the effectiveness and superiority of the proposed method in practical maritime application scenarios from a more comprehensive perspective. These two types of experiments complement each other in terms of objectives and analytical depth, collectively supporting the rationality and practical value of the proposed method.

3.7. Visualization Effect Comparison

To further verify the practical performance of the improved model in maritime distress target detection tasks, several typical aerial imaging scenarios were selected, and the detection results of the original RT-DETR model and the improved RT-DETR-SSF model were compared under the same input images. As shown from Figures 11–16, The subfigure (a) displays the detection results of the original RT-DETR model, while the subfigure (b) presents the detection results of the improved RT-DETR-SSF model.



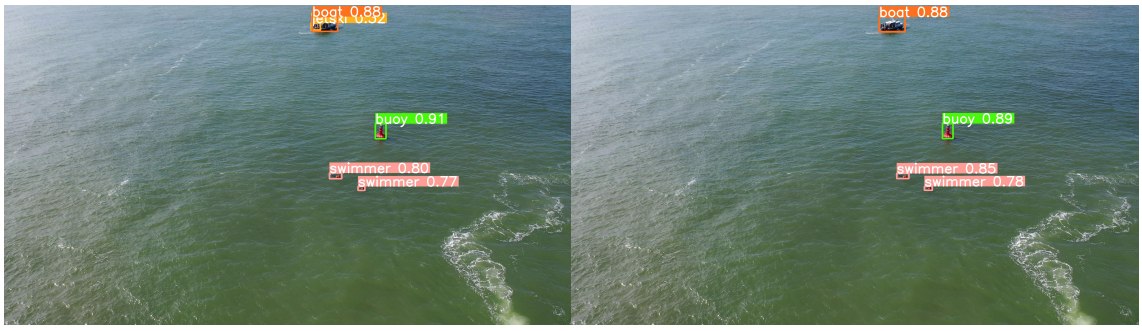
(**a**) (**b**)

Figure 11. Effect comparison of object detection before and after improvement: RT-DETR-r18 (**a**); RT-DETR-SSF (**b**).



(**a**) (**b**)

Figure 12. Effect comparison of object detection before and after improvement: RT-DETR-r18 (**a**); RT-DETR-SSF (**b**).



(**a**) (**b**)

Figure 13. Effect comparison of object detection before and after improvement: RT-DETR-r18 (**a**); RT-DETR-SSF (**b**).



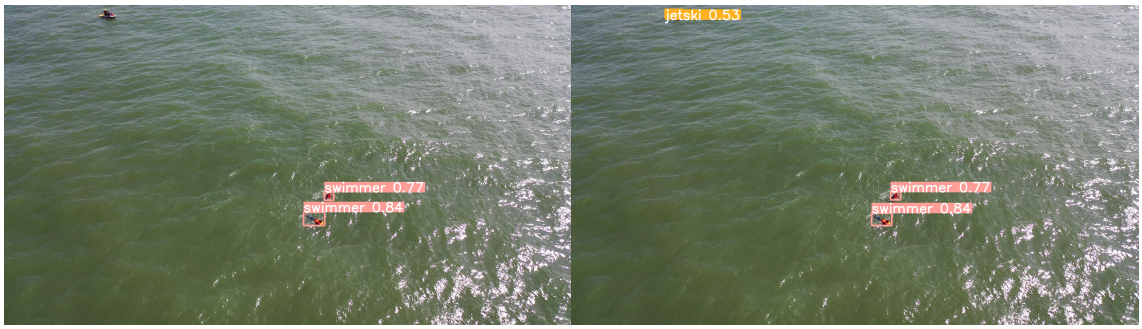
(**a**) (**b**)

Figure 14. Effect comparison of object detection before and after improvement: RT-DETR-r18 (**a**); RT-DETR-SSF (**b**).



(**a**) (**b**)

Figure 15. Effect comparison of object detection before and after improvement: RT-DETR-r18 (**a**); RT-DETR-SSF (**b**).



(**a**) (**b**)

Figure 16. Effect comparison of object detection before and after improvement: RT-DETR-r18 (**a**); RT-DETR-SSF (**b**).

The visual comparison results clearly reveal that the original model exhibits certain false detections and missed detections when handling distant and small-sized targets. For instance, some distant "boat" and "jet ski" targets were not correctly detected, while in areas with densely clustered or overlapping objects, issues such as duplicate bounding boxes and multiple detections—leading to redundant annotations—were observed.

In comparison, the improved model performs better in terms of detection accuracy and object recognition integrity. Specifically, it includes the following aspects:

1. Enhanced small object detection capability: The improved model can effectively detect smaller scale objects in the image;

2. Improved multi-target separation capability: As shown in Figures 13 and 14, for cases where there is target overlap or close proximity, the improved model can effectively suppress redundant boxes and reduce the occurrence of duplicate detections;
3. Enhanced detection of distant small targets: As shown in Figures 12, 15, and 16, the proposed method demonstrates improved detection performance for small targets at long distances, achieving more accurate localization and reducing missed detections in complex maritime scenes.

To further objectively evaluate the limitations of the proposed upgraded model, additional failure-case visualization results are provided in Figure 17. Compared with the original model, the upgraded model alleviates the missed-detection problem for distant blurry small targets to a certain extent, demonstrating improved sensitivity and confidence for small-object detection in long-range maritime scenes. As shown in Figure 17, several distant boat targets that are difficult to detect in the baseline model can now be successfully recognized by the upgraded model.



(**a**) (**b**)

Figure 17. Effect comparison of object detection before and after improvement: RT-DETR-r18 (**a**); RT-DETR-SSF (**b**).

However, despite these improvements, missed detections and classification instability may still occur for extremely distant and severely blurred small targets. In complex sea-surface environments, swimmer targets occupy only a very limited number of pixels and contain weak semantic and texture information, making them highly susceptible to interference from water ripples and specular reflections. As a result, a small number of ultra-small targets may still be overlooked.

These results indicate that, while the proposed model effectively enhances the feature representation capability for small targets and improves detection robustness under long-range conditions, the detection of extremely small and heavily degraded objects in maritime scenarios remains inherently challenging due to limited fine-grained features during deep feature extraction and insufficient contextual semantic information. Overall, the improved RT-DETR-SSF model exhibits better robustness and accuracy in detecting distressed targets in complex marine environments, validating its effectiveness for maritime small-target detection tasks.

4. Discussion

Maritime distress target detection remains a challenging task due to the high proportion of small targets, complex sea-surface interference, illumination variations, and strict real-time requirements in UAV-based rescue scenarios. To address these challenges, this paper improves the RT-DETR framework by introducing the Small Object Focused Convolution (SFConv) module, the Scale-Preserving Feature Interaction (SPE) module, and the Focaler-DIoU loss function. Experimental results demonstrate that the proposed RT-DETR-

SSF effectively enhances small-target representation and detection robustness in complex maritime environments.

The quantitative results verify the effectiveness of the proposed improvements. Compared with the baseline RT-DETR-r18, RT-DETR-SSF improves mAP@50 from 78.96% to 83.47% and achieves higher recall while maintaining comparable precision. These improvements indicate that the proposed modules effectively enhance both target localization and recognition capability. Specifically, SFConv strengthens fine-grained feature extraction for small maritime targets, SPE improves shallow-feature utilization and cross-scale information interaction, and Focaler-DIoU enhances bounding-box regression by focusing more on difficult samples during optimization. The combined effect of these modules enables the detector to better distinguish distress targets from complex sea-surface backgrounds.

Compared with other state-of-the-art methods, RT-DETR-SSF exhibits a more balanced detection performance. As shown in Table 7, although other methods achieve relatively high precision values, their recall rates decrease significantly, indicating that a considerable number of distress targets remain undetected. In contrast, RT-DETR-SSF achieves the highest mAP@50 (85.99%), the highest mAP@50:95 (58.57%), and the highest recall (83.92%) among all compared methods. Since missed detections may directly affect rescue efficiency and even endanger lives in practical maritime search-and-rescue missions, recall and overall localization accuracy are particularly important evaluation metrics. These results suggest that the proposed method provides a more favorable balance between detection accuracy and target coverage than existing approaches.

From the perspective of computational efficiency and deployment practicality, the proposed model maintains moderate computational complexity while achieving superior detection performance. The model requires 78.2 GFLOPs and the trained weight file occupies only 43.5 MB. During inference, the overall GPU memory consumption is approximately 2–3 GB, which is significantly lower than the memory capacity available on NVIDIA Jetson Orin NX platforms. Therefore, the proposed model demonstrates promising potential for deployment in resource-constrained UAV-based maritime rescue systems.

Visualization results further support the quantitative findings. The proposed method exhibits stronger robustness in challenging maritime environments, particularly in scenarios involving distant targets, sea clutter interference, and complex background textures. Compared with the baseline detector, RT-DETR-SSF reduces missed detections and improves localization quality for small and blurred distress targets, demonstrating the effectiveness of the proposed feature enhancement and cross-scale interaction mechanisms.

Nevertheless, several limitations still exist. Although the proposed method improves small-target detection performance, missed detections may still occur when targets are extremely distant, heavily occluded, or severely blurred. In addition, the current study focuses on image-based object detection and does not explicitly exploit temporal information available in UAV video sequences. Future research will investigate lightweight temporal-spatial feature modeling, adaptive multi-scale representation learning, and more efficient Transformer architectures to further enhance robustness and deployment efficiency in complex maritime environments.

5. Conclusions

To address the challenges of maritime distress target detection, including small target scale, complex sea-surface interference, and real-time requirements, this paper proposes an improved RT-DETR-based model termed RT-DETR-SSF. By introducing the Small Object Focused Convolution (SFConv) module, the Scale-Preserving Feature Interaction (SPE) module, and the Focaler-DIoU loss function, the proposed method effectively enhances small-target feature representation and localization accuracy in complex maritime environ-

ments. Experimental results demonstrate that the proposed model improves the mAP@50 from 78.96% to 83.47%, while maintaining relatively low computational complexity and a lightweight model size of 43.5 MB. In addition, the proposed method exhibits promising deployment potential on embedded edge AI platforms, such as the NVIDIA Jetson Orin NX and NVIDIA Jetson Xavier NX, making it suitable for UAV-based maritime rescue applications. Future work will focus on improving detection robustness for extremely distant and severely blurred targets under more challenging maritime conditions.

Author Contributions: Conceptualization, K.L. and X.C.; methodology, K.L. and X.C.; software, Z.L.; validation, J.X. and Y.L.; data curation, Y.Z.; writing—original draft preparation, X.C.; writing—review and editing, K.L., Z.L. and J.X.; visualization, X.C.; supervision, K.L. and Y.L.; project administration, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key R&D Program of Shandong Province, China grant number 2024CXPT041 and National Natural Science Foundation of China grant number 61803217, 62003231 and Shandong Natural Science Foundation grant number ZR2023MF029 and Heilongjiang Natural Science Foundation grant number LH2023F025 and Hainan Natural Science Foundation grant number KYZ20250020 and Shandong Provincial Outstanding Young Innovation Team Support Program for Higher Education Institutions grant number 2022KJ142 and Shandong Provincial Mount Taishan Scholar Support Program Project grant number TSQN202408163.

Data Availability Statement: GitHub Repository: https://github.com/changxinbo-jpg/RTDETR-SSF.git (accessed on 20 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Su, N.; Chen, X.; Guan, J.; Huang, Y. Maritime target detection based on radar graph data and graph convolutional network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4019705. [CrossRef]
2. Yang, P.; Dong, L.; Xu, W. Small maritime target detection using gradient vector field characterization of infrared image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1827–1841. [CrossRef]
3. Zhang, C.S.; Liu, Y.; Song, J.; Sun, S.; Qian, S. Maritime target detection of PCL system based on non-cooperative pulse radar. In Proceedings of the 2024 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Zhuhai, China, 22–24 November 2024; pp. 1–6.
4. Liu, W.; Zhu, C.; Liu, Y.; Li, Z. Maritime target detection method based on feature fusion of visible and infrared images. In Proceedings of the 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE), Shanghai, China, 21–23 March 2025; pp. 2487–2490.
5. Xu, Z.Y. Aerial small target detection method based on improved Faster R-CNN. *Shipbuild. Stand. Qual.* **2024**, *4*, 30–37.
6. Hou, Y.; Wu, Y.; Kou, X.R.; Huang, J.C.; Tuo, J.D.; Wang, Y.Q.; Huang, X.J. Small object detection algorithm for UAV images based on improved YOLOv8. *Comput. Eng. Appl.* **2025**, *61*, 83–92.
7. Fu, Z.; Xiao, Y.; Tao, F.; Si, P.; Zhu, L. DLSW-YOLOv8n: A Novel Small Maritime Search and Rescue Object Detection Framework for UAV Images with Deformable Large Kernel Net. *Drones* **2024**, *8*, 310. [CrossRef]
8. Xu, M.; Ma, L.; Jiang, Y. Spatio-temporal and global context information fusion based vehicle re-identification algorithm. *J. Highw. Transp. Res. Dev.* **2025**, *42*, 21–28.
9. Zhang, Y.; Chen, G.; Zhang, P.; Tong, J.Y.; Shan, M.J.; Shan, H.L. Underwater target detection based on enhanced local features. *China Meas. Test* **2025**, *51*, 151–158.
10. Liu, K.; Qi, Y.; Xu, G.; Li, J. YOLOv5s maritime distress target detection method based on swin transformer. *IET Image Process.* **2024**, *18*, 1258–1267. [CrossRef]
11. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
12. Neubeck, A.; van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Los Alamitos, CA, USA, 20–24 August 2006; pp. 850–855.
13. Zhang, H.; Zhang, S. Focaler-IoU: More focused intersection over union loss. *arXiv* **2024**, arXiv:2401.10525.

14. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Wan, C.; Yu, H.; Li, Z.; Chen, Y.; Zou, Y.; Liu, Y.; Yin, X.; Zuo, K. Swift Parameter-free Attention Network for Efficient Super-Resolution. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 16–22 June 2024; pp. 6246–6256.
16. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 13728–13737.
17. Cai, Z.; Ding, X.; Shen, Q.; Cao, X. Refconv: Reparameterized refocusing convolution for powerful convnets. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 11617–11631. [CrossRef] [PubMed]
18. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
19. Varga, L.A.; Kiefer, B.; Messmer, M.; Zell, A. SeaDronesSee: A Maritime Benchmark for Detecting Humans in Open Water. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 3686–3696.
20. Park, S.; Yeo, Y.J.; Shin, Y.G. PConv: Simple yet effective convolutional layer for generative adversarial network. *Neural Comput. Appl.* **2021**, *34*, 7113–7124. [CrossRef]
21. Ding, X.; Zhang, X.; Han, J.; Ding, G. Diverse Branch Block: Building a Convolution as an Inception-like Unit. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 19–25 June 2021; pp. 10881–10890.
22. Chen, Z.X.; He, Z.W.; Lu, Z.M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Trans. Image Process.* **2024**, *33*, 1002–1015. [CrossRef] [PubMed]
23. Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; Shan, Y. Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 5513–5524.
24. Zhong, J.C.; Chen, J.Y.; Mian, A. DualConv: Dual convolutional kernels for lightweight deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 9528–9535. [CrossRef] [PubMed]
25. Qi, Y.; He, Y.; Qi, X.; Zhang, Y.; Yang, G. Dynamic Snake Convolution based on Topological Geometric Constraints for Tubular Structure Segmentation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 6047–6056.
26. Zhang, X.; Liu, C.; Yang, D.; Song, T.; Ye, Y.; Li, K.; Song, Y. RFAConv: Innovating spatial attention and standard convolutional operation. *arXiv* **2023**, arXiv:2304.03198.
27. Finder, S.E.; Amoyal, R.; Treister, E.; Freifeld, O. Wavelet Convolutions for Large Receptive Fields. In *Computer Vision—ECCV 2024*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2025; Volume 15112.
28. Ma, S.; Xu, Y. MPDIoU: A loss for efficient and accurate bounding box regression. *arXiv* **2023**, arXiv:2307.07662. [CrossRef]
29. Zhang, H.; Xu, C.; Zhang, S. Inner-IoU: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.