Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 1, No. 3, 2026

LRS-RT-DETR: A Long-Range Floating Garbage Detector with Multi-Scale Feature Fusion

Chenglong Lu, Xiangguo Sun *

School of Mechanical Engineering, Sichuan University of Science & Engineering, Yibin 644000, China

Abstract: Real-time detection of floating water surface garbage is of great significance for water environment management. However, when the garbage target is more than 30 meters away from the camera, the target occupies only a small number of pixels in the image, and the complex water surface background also introduces detection interference; thus, general object detection methods struggle to achieve satisfactory performance. In this paper, we propose a long-range small object detection Transformer model named LRS-RT-DETR, which is improved based on the RT-DETR-R18 baseline, to adapt to the scenario of long-distance small garbage detection on water surfaces. First, we design a high-resolution feature fusion pyramid network called High-Resolution Feature Fusion Pyramid Network (HRFF-FPN). By employing the lossless Space-to-Depth (SPD) reorganization operation, we introduce the high-resolution P2 feature into the P3 detection node, significantly enhancing the model's spatial perception capability for extremely small distant targets without adding extra detection heads. After receiving the fused features from the P2 layer, we propose a multi-scale kernel branch module (MSKB), which combines skip connections with a multi-scale parallel receptive field branch to achieve multi-level refined modeling of the fused features. Experimental results on the targeted long-distance water surface garbage dataset Far-water-surface Garbage Dataset (FWSGD) show that LRS-RT-DETR achieves 91.5% mAP@0.5, an improvement of 2.7 percentage points over the baseline RT-DETR-R18, and 43.9% mAP@0.5-0.95, an improvement of 1.5 percentage points. Meanwhile, the model parameter count increases by only 0.6M, with controllable computational overhead, demonstrating good potential for real-time deployment on edge devices.

Keywords: Water Surface Garbage Detection, Small Object Detection, Feature Pyramid Network, Multi-Scale Feature Fusion.

1. Introduction

Prevention and control of water environment pollution and water cleanliness protection are important topics in global environmental sustainable development. Every year, a large amount of garbage enters marine waters, and inland waters are also affected. Common household waste generated by humans floats on the water surface for a long time and is difficult to degrade in natural waters. If the accumulation of floating garbage is not cleaned up in time, it will cause gradual deterioration of water quality, affecting navigation, aquatic life, and the safety of domestic water for humans. Traditional manual garbage cleaning mainly relies on simple tools and long hours of work on the water, which not only consumes substantial human and material resources but also fails to achieve long-term real-time monitoring. Intelligent cleaning robots and real-time monitoring devices based on computer vision provide a new approach for improving the water environment. Through the long-term development and progress of object detection, the emergence of various derivative models — whether anchor-based methods such as the YOLO series and Faster R-CNN series, or anchor-free end-to-end detectors such as the DETR series — has provided excellent technical foundations for research on water surface garbage detection and water environment governance [1]. A research team from Changzhou University and Hohai University proposed a PC-Net (Pyramid Classification Network) model based on an improved Faster R-CNN-R50, which effectively improves the detection accuracy of small water surface targets using pyramid anchor generation and classification map discrimination [2]. Another research team from the same universities proposed DENS-YOLOv6 (Detail Enhancement Noise Suppression YOLOv6) based on YOLOv6, designing a Detail Information Enhancement Module (DIEM) and an Adaptive Noise Suppression Module (ANSM) to optimize the detection ability of small water surface targets [3]. A research team from Guangzhou University proposed an efficient multi-scale hybrid model called EMSH-DETR (Efficient Multi-Scale and Hybrid Detection Transformer) based on the DETR model, which introduces multi-module collaboration and adaptive attention mechanisms to improve the modeling ability and operational efficiency for small water surface targets [4].

In the context of long-range water surface garbage detection, regardless of which detection model is used, a common challenge remains: when the garbage target is far from the water surface (beyond 30 meters), it occupies only a very small number of pixels in the input image, and the target scale may also change during detection. These factors are unique challenges in current water surface garbage detection scenarios, making it difficult for existing general object detection models to achieve satisfactory performance in long-range water surface garbage detection tasks. The RT-DETR (Real-Time Detection Transformer) detector, as a new-generation real-time Transformer model, greatly improves inference speed while retaining the advantages of the DETR series through its efficient hybrid encoding and uncertainty-minimized query selection mechanism, achieving a good balance between speed and accuracy [5]. In this study, we select the R18 variant of RT-DETR as the baseline and propose LRS-RT-DETR (Long-Range Small-target Detection Transformer). This backbone network has fewer parameters and faster inference speed, making it more suitable for real-time detection in resource-constrained scenarios. We design a High-Resolution Feature Fusion Pyramid Network (HRFF-FPN) to obtain high-resolution features from the P2 layer, solving the problem of insufficient perception of distant garbage targets in the baseline network without introducing

additional detection heads. Furthermore, we design a Multi-Scale Kernel Branch Module (MSKB) for multi-path feature fusion, improving the detection capability for long-range water surface garbage targets while keeping computational overhead under control.

2. Design of the LRS-RT-DETR Model

2.1. Overall Architecture

The overall architecture of the LRS-RT-DETR model follows the basic framework of the R18 network, consisting of a backbone, a feature fusion network, and a detection head. The R18 network constructs a three-level feature pyramid (P3-P5) in the feature fusion part and simultaneously sets three detection heads. The proposed HRFF-FPN (High-Resolution Feature Fusion Pyramid Network) does not set an independent detection head at the P2 layer. Instead, it uses an SPD (Space-to-Depth) module to extract features rich in small-target information from the P2 feature layer and fuses them with P3 features. The extracted fine-grained features are then concatenated with the P3 features and the upsampled P4 features in a three-way fusion, enabling the P3 layer to obtain richer spatial representations of small targets without adding any detection heads, thereby avoiding the increase in computational cost that would result from directly adding a P2 detection layer and thus preserving the model's inference speed. Finally, the MSKB (Multi-Scale Kernel Branch Module) performs multi-scale integration on the fused features. Its branch structure reduces computational redundancy while retaining more effective small-target information. This design fully exploits the high-resolution advantages of the P2 layer while avoiding the computational overhead of additional detection layers. The overall architecture is shown in Figure 1.

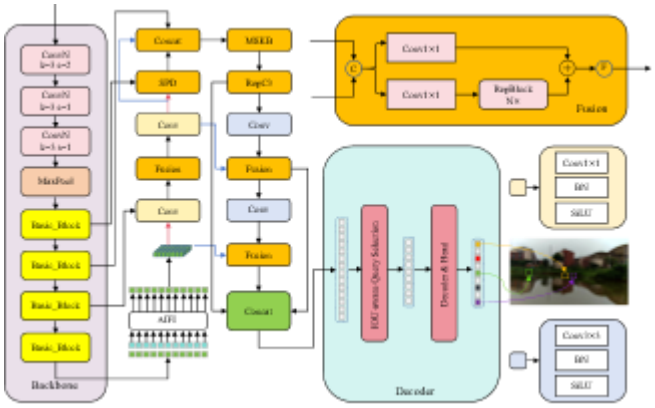

Figure 1. Architecture of LRS-RT-DETR

2.2. High-Resolution Feature Fusion Design

Small object detection is a technical challenge in many application scenarios. The most direct improvement method is to introduce high-resolution layers such as P2 or P1 to provide more detailed feature information. Taking the common approach of directly introducing the P2 feature layer in the R18 network as an example: given an input image size of 640×640 and a progressive halving of feature layer resolution, the spatial resolution of the P2 layer has four times as many pixels as that of the P3 layer. Consequently, the computational cost of subsequent convolutions and feature fusion increases exponentially, and even the detection head becomes more computationally complex. Although this method improves the detection accuracy of long-range water surface garbage targets, it also increases inference time, leading to excessive computational overhead and higher device deployment costs [6].

Therefore, HRFF-FPN is a method that utilizes the high-resolution feature layer without directly introducing a new detection layer, thereby avoiding excessive computational overhead. We place the SPD (Space-to-Depth) module in the feature fusion stage to receive the output features from the P2 layer and, through lossless feature information compression, fully integrate the information from the high-resolution layer into the P3 feature layer. This approach not only avoids the problem of excessive computational overhead caused by directly introducing a detection layer, but also mitigates the loss of small-target information that occurs in the original architecture during downsampling via standard pooling or strided convolution. It fundamentally improves the core bottleneck of insufficient P3 feature resolution when the original architecture handles long-range small water surface targets. The structure of HRFF-FPN is shown in Figure 2.

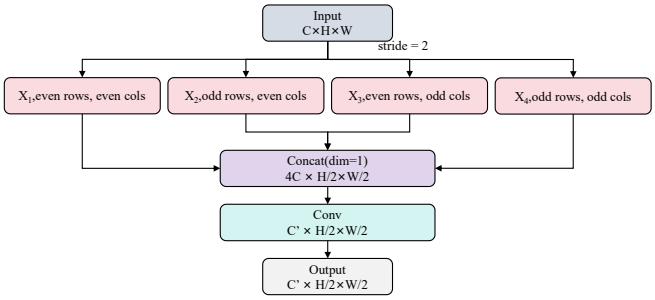

Figure 2. Structure of HRFF-FPN

The core idea of SPD (Space-to-Depth) originates from the space-to-depth rearrangement operation, with the main objective of losslessly compressing information from the high-resolution layer so that the feature channels can be aligned with the target feature layer [7]. After receiving the feature map information from the high-resolution P2 layer, SPD first performs a periodic sampling operation, decomposing the input feature map along the spatial dimension into four complementary sub-feature maps according to parity positions. The feature separation operation is shown in Equation (1).

$$X_1 = X[\ldots,::2,::2], X_2 = X[\ldots,1::2,::2]$$
$$X_3 = X[\ldots,::2,1::2], X_4 = X[\ldots,1::2,1::2] \quad (1)$$

Where, in the input feature map $X \in R^{C\times H\times W}$, C=64 is the number of channels, and H×W=160×160 is the spatial size of the feature map.

Each sub-feature map is compressed to half the size of the original feature map, corresponding to the sampling results of even-row-even-column, odd-row-even-column, even-row-odd-column, and odd-row-odd-column of the original feature map, respectively. The four sub-feature maps are then concatenated along the channel dimension to form an intermediate representation with four times the number of channels and half the spatial resolution. A standard convolution with batch normalization and an activation function is then applied to integrate and compress the channels, producing a target with 128 channels. The integration operation is shown in Equation (2) and (3).

$$X_{cat} = Concat([X_1, X_2, X_3, X_4], dim = 1) \in R^{4C\times\frac{H}{2}\times\frac{W}{2}} \quad (2)$$

$$Y = Conv_{3\times3}(X_{cat}) \in R^{C'\times\frac{H}{2}\times\frac{W}{2}} \quad (3)$$

Finally, the output features are channel-fused with the upsampled features from the P4 layer and the information from the P3 feature layer.

$$F_{cat} = Concat([F_{P2}, F_{P4}, F_{P3}], dim = 1) \quad (4)$$

After the operation, the output feature map has a spatial size of 80×80, which can be perfectly fused with the P3 layer, providing correct preprocessing for subsequent feature concatenation. Compared with standard convolution and max pooling, SPDConv does not discard any spatial information. Instead, it encodes the spatial relationships of adjacent pixels into different channels, laying the foundation for subsequent convolutions to perceive the original structural information along the channel dimension. The entire process achieves effective transfer of P2 feature information to P3, enhancing small target detection capability without adding extra detection layers.

2.3. Multi-Scale Kernel Branch Module

After concatenating the output P2 fine-grained features, the upsampled P4 semantic features, and the backbone P3 features, the resulting feature tensor contains information at different scales and semantic levels. If only ordinary convolution is used, the network model often struggles to simultaneously establish relationships between global features and local details. In this study, we design the Multi-Scale Kernel Branch Module (MSKB) for feature fusion, deeply integrating the fused features through parallel multi-scale receptive field modeling. Part of the features are directly retained via a skip branch, while the other part undergoes feature transformation through the Multi-Scale Kernel (MSK) branch, and the two paths are then fused. This approach reduces computational cost while obtaining more accurate feature information and also facilitates good gradient propagation [8]. Within the feature transformation branch, the MSK module employs a parallel combination of horizontal, vertical, isotropic large-receptive-field convolutions and point convolutions, covering multi-scale, large-range spatial relationships from single-pixel to multi-pixel details, allowing the model to perceive target features from different scales. The synergy between MSKB and HRFF-FPN endows the model with the ability to accurately extract high-resolution information and efficiently fuse features. The structure of MSKB is shown in Figure 3.

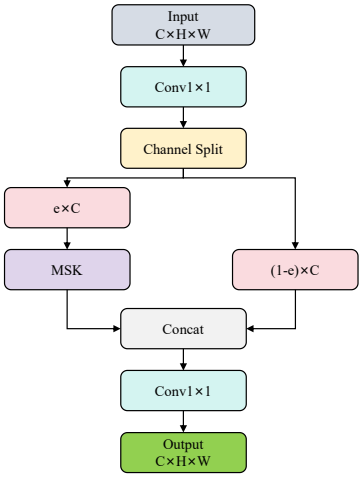

Figure 3. Structure of MSKB

For the input concatenated feature F_{cat}, a 1×1 convolution is first applied for channel adjustment. The result is then split into two channel groups: branch F_2 retains the original features through an identity mapping, while branch F_1 undergoes feature transformation. The outputs of the two branches are concatenated along the channel dimension and passed through a second 1×1 convolutional layer to perform channel integration and dimension alignment, producing the final output of MSKB. The separation operation is shown in Equations (5), (6), and (7).

$$F' = CV_1(F_{cat}) \in R^{C\times H\times W} \quad (5)$$

$$F_1, F_2 = Split(F', [(e\cdot C), C-(e\cdot C)], dim = 1) \quad (6)$$

$$Y = CV_2(Concat(OmniKernel(F_1), F_2)) \in R^{C\times H\times W} \quad (7)$$

Where CV_1 denotes the convolution operation, and e is the split ratio parameter. In this paper, e is set to 0.25, which preserves rich feature diversity while keeping the overall computational cost of the module within an acceptable range.

Inspired by the multi-scale receptive field design of OmniKernel, this paper introduces the Frequency Channel Attention (FCA) [9], Spatial-Channel Attention (SCA) [10], and Frequency Gating Module (FGM) [11] on its basis, and proposes an improved MSK module aimed at comprehensively modeling the input features after HRFF-FPN fusion across multiple levels: global frequency features, medium-range spatial features, and local detail features [12]. The structure of MSK is shown in Figure 4.

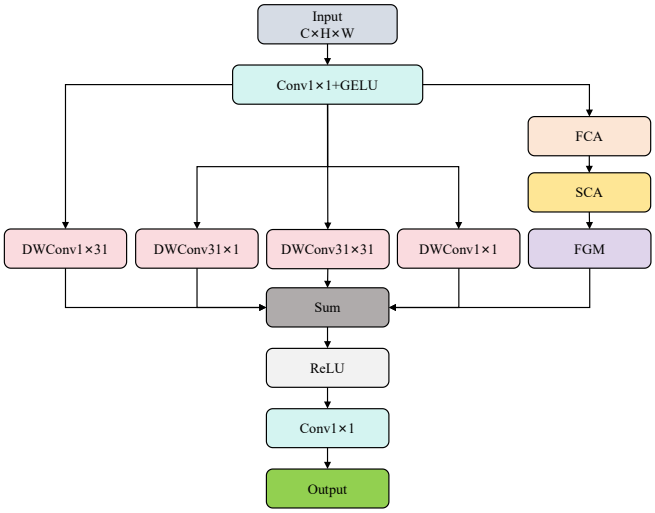

Figure 4. Structure of MSK

The input features first undergo preliminary processing through a transformation layer consisting of a 1×1 convolution and a GELU activation function. Compared with ReLU, the GELU activation function has smoother nonlinear characteristics, which helps preserve fine-grained feature differences in the initial transformation stage and is particularly important for weak activation signals of small targets. The preprocessing operation is shown in Equation (8).

$$F_{in} = GELU(Conv_{1\times1}(F_{cat})) \in R^{Ce\times H\times W} \quad (8)$$

In the frequency channel attention (FCA) branch, global average pooling is applied to F_{in}, followed by a 1×1 convolution to generate frequency domain attention weights, and feature modulation is then performed in the frequency domain. X_{fft} is a complex-domain spectral tensor, whose real and imaginary parts correspond to the magnitude and

phase information of different spatial frequency components, respectively. High-frequency components correspond to edges and texture details in the image, while low-frequency components correspond to the overall structure and background. By performing attention modulation in the frequency domain, FCA is able to adaptively weight different frequency components with a global receptive field. The attention weights are generated by global average pooling (GAP) followed by a 1×1 convolution. The operation of the FCA branch is shown in Equation (9).

$$X_{fft} = F(F_{in}), X_{fca} = \left| F^{-1}(Conv_{1\times1}(GAP(F_{in})) \cdot X_{fft}) \right| \quad (9)$$

Where $F(\cdot)$ and $F^{-1}(\cdot)$ denote the two-dimensional fast Fourier transform (FFT2) and its inverse (IFFT2), respectively.

In the spatial-channel attention (SCA) branch, global average pooling and a 1×1 convolution are applied to the input features to generate channel attention, achieving joint spatial and channel modulation. The operation of the SCA branch is shown in Equation (10).

$$X_{sca} = Conv_{1\times1}(GAP(X_{fca})) \cdot X_{fca} \quad (10)$$

The output features are then refined by the Frequency Gating Module (FGM), which achieves adaptive feature enhancement through element-wise multiplication of the spatial convolution path and the frequency cross path. The computation of FGM is performed in two steps. First, the frequency-domain gating output X_a is computed via cross-path element-wise multiplication in the frequency domain. Second, the final output X_{gfm} is obtained by residual scaling fusion with learnable parameters α and β, where α and β are initialized to 0 and 1, respectively, ensuring the module approximates an identity mapping at the beginning of training. The operation of the FGM branch is shown in Equation (11) and Equation (12).

$$X_a = \left| F^{-1}(DWConv_{1\times1}(X_{sca}) \cdot F(DWConv_{1\times1}(X_{sca}))) \right| \quad (11)$$

$$X_{fgm} = X_a \cdot \alpha + X_{sca} \cdot \beta \quad (12)$$

In the multi-scale kernel branch, MSK runs four depthwise separable convolutions (DWConv) with different receptive fields in parallel (kernel size k=31): horizontal $DW_{1\times31}$, vertical $DW_{31\times1}$, isotropic large receptive field $DW_{31\times31}$, and point convolution $DW_{1\times1}$, thereby covering a spatial receptive field range from local to global. Finally, the outputs of all branches are added element-wise to the input residual, passed through a ReLU activation, and then sent to the output convolution. The operations of the multi-scale kernel branch are shown in Equations (13) and (14).

$$F_{out} = F_{in} + DW_{1\times31}(F_{in}) + DW_{31\times1}(F_{in}) + DW_{31\times31}(F_{in}) + DW_{1\times1}(F_{in}) + X_{fgm} \quad (13)$$

$$Y_{\text{mskb}} = Conv_{1\times1}(ReLU(F_{out})) \quad (14)$$

3. Experiments and Analysis

3.1. Dataset and Experimental Setup

The experiments use a self-built long-range water surface garbage dataset named FWSGD (Far water surface garbage dataset). Data were collected from scenic lakes, reservoirs, and ditch areas with complex backgrounds at distances beyond 30 meters. To improve the model's generalization ability, different lighting conditions and time points were chosen to capture three common types of water surface garbage: plastic bottles, plastic bags, and metal cans. Pixel-level transformations and weather simulation were applied as data augmentation methods to enhance the model's adaptability. The dataset consists of 3,219 training images, 920 validation images, and 460 test images.

Figure 5. Examples from the dataset

Training and validation were conducted on an Ubuntu 20.04 system with an E5-2680 V4 CPU and an RTX 3090 GPU, using PyTorch 1.13 and Python 3.8. The training hyperparameters were set as follows: batch size = 8, initial learning rate = 0.0001, AdamW optimizer, and 200 training epochs. Evaluation metrics include average precision at IoU=0.5 and IoU=0.5-0.95, F1-Score, as well as model parameters and computational cost, to comprehensively assess the overall performance of the model.

3.2. Ablation Study

This study conducts detailed ablation experiments to verify the effectiveness, advantages, disadvantages, and synergistic effects of each improved module. RT-DETR-R18 is used as the baseline model. HRFF-FPN denotes the configuration that adds only the SPD-processed P2 features during P3 fusion without using MSKB. MSKB denotes the configuration that replaces the standard convolution with a branch fusion architecture but does not use HRFF-FPN. LRS-RT-DETR is the complete improved model. The ablation experimental results are shown in Table 1.

Table 1. Ablation Experiment

Models	mAP@0.5 (%)	mAP@0.5-0.95 (%)	F1-Score (%)	FLOPs (G)	Params (M)	Size (MB)
R18	88.8	42.4	88.8	56.9	19.8	76.9
HRFF-FPN	89.8	43.8	89.7	60.3	20.1	77.9
MSKB	89.4	41.8	89.7	66.2	20.6	79.9
LRS-RT-DETR	91.5	43.9	90.8	65.2	20.4	79.3

When only the HRFF-FPN module is introduced, the model's mAP@0.5 increases from 88.8% (baseline) to 89.8%, mAP@0.5-0.95 increases from 42.4% to 43.8%, and the F1-Score increases from 88.8% to 89.7%, indicating a certain degree of improvement in detection accuracy. This demonstrates that the strategy of HRFF-FPN, which losslessly extracts high-resolution P2 features via SPD (Space-to-Depth) and fuses them into the P3 detection node, is effective. The introduction of high-resolution fine-grained features indeed enhances the model's spatial perception of long-range small water surface targets, verifying the effectiveness of HRFF-FPN in alleviating the bottleneck of

insufficient P3 feature resolution. Meanwhile, FLOPs increase slightly from 56.9G to 60.3G, and the number of parameters increases from 19.8M to 20.1M, indicating a small increase in computational overhead. This suggests that HRFF-FPN improves detection accuracy while maintaining high computational efficiency.

When only the MSKB module is introduced, the model achieves 89.4% mAP@0.5 and the F1-Score improves to 89.7%, but mAP@0.5-0.95 is only 41.8%, slightly lower than the level achieved by HRFF-FPN alone. This indicates that, through the synergistic effect of multi-scale parallel receptive fields and frequency-domain attention mechanisms, MSKB improves overall target localization and classification discrimination. In particular, its F1-Score matches that of the HRFF-FPN scheme, demonstrating MSKB's independent contribution to multi-scale feature integration. However, when MSKB is used alone, FLOPs increase to 66.2G and the number of parameters increases to 20.6M, which represents slightly higher computational overhead compared with the HRFF-FPN scheme. Moreover, its mAP@0.5-0.95 is lower than that of HRFF-FPN, indicating that without the support of high-resolution fine-grained features, relying solely on multi-scale receptive field expansion provides limited improvement in fine localization accuracy, and the full potential of MSKB's multi-scale modeling remains underutilized.

In the complete scheme LRS-RT-DETR (HRFF-FPN + MSKB), all evaluation metrics reach their optimal levels. mAP@0.5 improves to 91.5%, an increase of 2.7 percentage points over the baseline; mAP@0.5-0.95 reaches 43.9%, an increase of 1.5 percentage points; and the F1-Score improves to 90.8%, an increase of 2.0 percentage points. Although LRS-RT-DETR has FLOPs of 65.2G and 20.4M parameters, which are slightly higher than those of the baseline model, the overall computational overhead remains modest. By halving the number of channels inside the modules, the issue of a sharp increase in computational cost caused by simply stacking the two modules is effectively avoided. Overall experiments demonstrate that the high-resolution fine-grained features provided by HRFF-FPN offer a richer and more discriminative input foundation for MSKB's multi-scale frequency-domain modeling. The two modules form a complementary synergistic mechanism of precise feature extraction and efficient multi-scale integration, enabling the complete scheme to surpass both the individual module schemes and the baseline model in terms of both accuracy and efficiency. This fully validates the overall design rationality of our proposed improvements and their practical effectiveness in the task of long-range small-target garbage detection on water surfaces.

3.3. Comparative Experiment

To further evaluate the comprehensive performance of the proposed LRS-RT-DETR method in the task of long-range small-target garbage detection on water surfaces, this study selects three recent representative improvement schemes in the fields of feature fusion and multi-scale perception as baselines for horizontal comparison experiments. MAFPN (Multi-Branch Auxiliary Fusion Feature Pyramid Network) is a structure highly similar to our module, which enhances multi-scale feature representation through multi-branch auxiliary fusion and heterogeneous reparameterized convolutions [13]. CSFCN (Context and Spatial Feature Calibration) is a module that improves feature spatial perception and semantic alignment through context and spatial feature calibration mechanisms, and like our module, is applied to alleviate the small target problem [14]. GLSA (Global-Local Spatial Attention) captures multi-scale spatial dependencies via a global-to-local spatial aggregation mechanism and is comparable to our approach [15]. All three methods were adapted and trained under the same RT-DETR-R18 baseline framework to ensure fairness and consistency of comparison. The comparative experimental results are shown in Table 2.

Table 2. Comparative Experiment

Models	mAP@0.5 (%)	mAP@0.5-0.95 (%)	F1-Score (%)	FLOPs (G)	Params (M)	Size (MB)
MAFPN	89.2	42.7	88.8	56.3	22.9	88.8
CSFCN	89.3	42.5	88.5	65.5	21.1	81.7
GLSA	89.7	43.6	89.3	63.7	21.9	85.0
LRS-RT-DETR	91.5	43.9	90.8	65.2	20.4	79.3

In terms of detection accuracy, LRS-RT-DETR achieves the highest mAP@0.5 and mAP@0.5-0.95 among all compared methods, indicating that our method yields superior bounding box regression quality for long-range small targets under high-precision localization requirements. This advantage mainly benefits from the lossless introduction of high-resolution fine-grained P2 features by HRFF-FPN, which provides sufficient spatial information support for precise localization. In terms of F1-Score, LRS-RT-DETR leads the other three methods by a significant margin (90.8%), demonstrating the overall advantage of our method in balancing precision and recall. This shows that the MSKB module, through its multi-scale frequency-domain attention mechanism, effectively suppresses interference from complex water backgrounds and plays a significant role in reducing both missed detections and false positives.

Regarding computational efficiency and model lightweighting, LRS-RT-DETR has fewer parameters than MAFPN and GLSA, and is comparable to CSFCN, reflecting its advantage in parameter utilization efficiency. In terms of model size, LRS-RT-DETR is also the smallest, offering better storage occupation and edge-deployment suitability. Although the FLOPs of LRS-RT-DETR (65.2G) are much higher than those of MAFPN (56.3G), considering that LRS-RT-DETR significantly outperforms MAFPN in accuracy across all metrics, this difference in computational cost is entirely reasonable. Moreover, 65.2G FLOPs remains acceptable on mainstream edge inference hardware. The above comparison fully validates the effectiveness of the cooperative design of the proposed HRFF-FPN and MSKB modules. LRS-RT-DETR achieves the optimal overall balance among accuracy, parameter efficiency, and deployment suitability, providing a more practical technical solution for the task of long-range garbage detection on water surfaces.

3.4. Visualization Analysis

To visually present the performance differences between our improved method and the baseline model in real water surface scenarios, Figure 6 shows a comparison of detection results on two typical water scenes. The left column presents the detection results of the RT-DETR-R18 baseline model, while the right column shows the results of our LRS-RT-DETR. The positions where the baseline model missed detections are marked with black circles.

Figure 6. Comparative analysis of detection results

The first scene in the upper row is a near-circular water area with vegetation occlusion and multiple distributed targets. The baseline model (R18) clearly misses a target in the area marked by the black circle in the upper-right corner of the image. In this region, the target occupies very few pixels, its color is similar to the surrounding vegetation, and background interference is strong. In contrast, LRS-RT-DETR successfully detects the target at the same location with a confidence of 0.72, indicating that HRFF-FPN effectively compensates for the insufficient spatial representation of extremely small targets in the baseline's P3 feature layer by introducing high-resolution fine-grained features from the P2 layer, thereby endowing the model with the ability to perceive such low-pixel targets under strong background interference. It can also be observed that the detection bounding boxes and confidence scores of both models are relatively close for the remaining targets, suggesting that the improved method achieves a targeted increase in small-target recall while maintaining the original detection capability, without introducing significant additional false positives.

The second scene in the lower row is a natural river channel with a complex shoreline background. This scene contains both "bottle" and "package" targets, some of which are located near the image edges. The background includes various interfering elements such as people, vegetation, and water surface reflections. The baseline model (R18) misses a target at the lower-left corner marked by the black circle. This target lies at the image edge, is extremely small, and exhibits weak feature responses against the complex background. LRS-RT-DETR successfully detects the target at this location and outputs a valid confidence prediction, demonstrating that the MSKB module, through the synergistic modeling of multi-scale parallel receptive fields and frequency-domain attention mechanisms, significantly enhances the model's activation response to small targets in edge regions. Comparing the detection boxes of the two columns shows that LRS-RT-DETR achieves a notably higher confidence score for "package" targets than R18, indicating that expanding the multi-scale receptive field also provides positive gains for non-rigid targets with irregular shapes and complex textures.

From the visualization results of the two scenes, it can be concluded that LRS-RT-DETR significantly enhances the recall ability for distant targets occupying very few pixels compared with the baseline RT-DETR-R18, effectively reducing the miss detection rate. Moreover, the model's robustness for target detection in complex backgrounds and edge regions is improved, and the interference from vegetation, water surface reflections, and people in the background is notably reduced. The model also provides more accurate confidence scores for detected targets, resulting in a more reasonable overall confidence distribution. The above visualization results, together with the quantitative indicator analyses in Tables 1 and 2, cross-validate each other, further verifying the effectiveness and practical value of the proposed improvement scheme in real-world long range water surface garbage detection scenarios.

4. Conclusion

In this paper, to address the technical challenges of extremely small pixel occupation and complex, variable backgrounds in long-range small-target garbage detection on water surfaces, we propose the LRS-RT-DETR detection method based on the RT-DETR-R18 baseline. In the feature fusion stage, we design the HRFF-FPN module to introduce high-resolution P2 features, enhancing the perception of small-target features while avoiding the substantial computational overhead caused by introducing a new feature layer. Furthermore, we design the MSKB module, which combines skip connections with the multi-scale parallel receptive fields of MSK to achieve multi-level fine-grained modeling of the fused features. Experimental results on the self-built FWSGD show that LRS-RT-DETR achieves improvements in accuracy over the baseline RT-DETR-R18, reaching 91.5% mAP@0.5 and 43.9% mAP@0.5:0.95, which correspond to increases of 2.7 and 1.5 percentage points, respectively. Meanwhile, the model parameter count increases by only 0.6M, with a slight rise in computational cost, which remains within a reasonable range overall. Ablation studies and visualization analyses further validate the synergistic effectiveness and design rationality of the HRFF-FPN and MSKB modules.

Acknowledgements

Supported by The Innovation Fund of Postgraduate, Sichuan University of Science & Engineering. (Y2024029).

References

[1] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C]//Proceedings of the European Conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[2] Li N, Huang H, Wang X, et al. Detection of floating garbage on water surface based on PC-Net [J]. Sustainability, 2022, 14(18): 11729.

[3] Li N, Wang M, Yang G, et al. DENS-YOLOv6: A small object detection model for garbage detection on water surface [J]. Multimedia Tools and Applications, 2024, 83(18): 55751-55771.

[4] Liu C, Li J, Ke Z, et al. EMSH-DETR: An efficient multi-scale and hybrid DETR for floating garbage detection [J]. Measurement Science and Technology, 2026, 37(1): 015407.

[5] Zhao Y, Lv W, Xu S, et al. DETRs beat YOLOs on real-time object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.

[6] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[7] Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects [C]//Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland, 2022: 443-459.

[8] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.

[9] Qin Z, Zhang P, Wu F, et al. FcaNet: Frequency channel attention networks [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 783-792.

[10] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module [C]//Proceedings of the European Conference on Computer Vision. 2018: 3-19.

[11] Zihan C, Hongyun Z, Duoqian M, et al. Multigranularity Dynamic Scene Image Deblurring Network Based on Deep Fusion of Frequency Domain and Spatial Domain Features [J]. Pattern Recognition and Artificial Intelligence, 2024, 37(6): 557-569.

[12] Cui Y, Ren W, Knoll A. Omni-kernel network for image restoration [C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(2): 1426-1434.

[13] Yang Z, Guan Q, Zhao K, et al. Multi-branch auxiliary fusion YOLO with re-parameterization heterogeneous convolutional for accurate object detection [C]//Chinese conference on pattern recognition and computer vision (PRCV). Singapore: Springer Nature Singapore, 2024: 492-505.

[14] Li K, Geng Q, Wan M, et al. Context and spatial feature calibration for real-time semantic segmentation [J]. IEEE Transactions on Image Processing, 2023, 32: 5465-5477.

[15] Tang F, Xu Z, Huang Q, et al. DuAT: Dual-aggregation transformer network for medical image segmentation [C]//Chinese conference on pattern recognition and computer vision (PRCV). Singapore: Springer Nature Singapore, 2023: 343-356.