Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 3, No. 3, 2025

Distant Water Surface Garbage Recognition Method in Complex Scenarios

Chenglong Lu, Xiangguo Sun *

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

Abstract: To improve the insufficient identification rate of remote garbage targets by water surface cleaning equipment, a model based on RT-DETR for remote garbage target detection called FRT-DETR was developed. First, CSPCAA-ADown was proposed as a modified backbone network, which better captures feature information of garbage targets beyond 30 meters at different scales while reducing computational burden, thus minimizing the impact of insufficient feature information caused by distance. Second, a CCGAF module was introduced to dynamically adjust the importance of features from different layers. Finally, an adaptive wavelet pooling module, WaveletPool, was employed for upsampling and downsampling feature maps, thereby reducing the impact of target scale variations and achieving good recognition rates even under conditions of blurry distant targets and complex backgrounds. The improved model achieves an mAP0.5 of 91.3% on the FWSG dataset with a reduced parameter count of 11.8M, providing a high-precision, high-efficiency, and low-computational-cost detection solution for water environment monitoring.

Keywords: Remote garbage detection on water surfaces, Dynamic feature adjustment, Scale adaptation.

1. Introduction

In aquatic environments, floating debris on water surfaces is a common pollutant. In the past, the collection of floating debris primarily relied on manual labor, which not only resulted in extremely heavy workload and high costs, but also posed unpredictable safety risks during water operations. With the rapid development of informatization and technology, various types of unmanned water surface cleaning equipment have emerged. In the research of unmanned water surface cleaning equipment, visual detection has become a critical component. The rapid advancement of deep learning technology in the field of object detection has provided new visual detection methods for water surface debris recognition.

In the historical development of visual tasks, convolutional neural networks have contributed indelible strength. Network models such as Faster R-CNN [1] and YOLO [2] can rapidly obtain target categories and bounding boxes, achieving good progress in water surface debris object detection. For instance, Zeng et al. proposed a lightweight SOE-YOLOv8 [3] model that modified the Neck section using Slim-Neck and modified the Backbone using lightweight convolution ODConv [4] to achieve lightweight deployment while ensuring detection performance. Du et al. proposed an improved YOLOv5s model using ghostbottleneck and incorporating ECA attention mechanism, effectively balancing detection accuracy and detection speed [5]. Zhou et al. proposed a YOLOv7-edge model [6] for small water surface targets, introducing the E-MP module and the BIformer [7] attention module, effectively addressing the edge blurring problem of floating debris. Although these studies have achieved good performance in water surface debris object detection, convolutional neural networks typically have strong local texture and structural information extraction capabilities but lack the ability to capture global image information of targets. Due to the small image size of debris targets at long distances on water surfaces (such as beyond 30 meters), the obtained information is relatively blurred, making it difficult to extract feature information, which limits their detection performance. The ability to capture global image information is also particularly important in actual visual detection tasks. While the receptive field can be expanded by adding convolutional layers, this brings the problem of model complexity and optimization difficulties. In transformers, the introduction of multi-head attention mechanisms enables them to clearly capture long-range dependencies in global vision perception, but they tend to overlook local features [8]. This is exactly opposite to convolutional neural networks. In some scholars' research, there has been active exploration of combining convolutional neural networks with transformers to possess the advantages of both. Carion et al. utilized transformer technology to propose a DETR (End-to-End Object Detection with Transformers) detector, which simplifies some detection processes, completely discarding anchor boxes generation and non-maximum suppression (NMS), adopting an end-to-end trainable encoder-decoder structure to replace region-based proposal methods, allowing all targets to be detected and processed simultaneously in parallel, better handling overlapping targets [9]. However, it has drawbacks such as long training time and slow convergence speed. Based on this, Zhu et al. proposed Deformable DETR (Deformable Transformers for End-to-End Object Detection) with a deformable attention mechanism that adaptively adjusts its sampling positions [10]. This significantly improves training speed and convergence speed. The RT-DETR (Real-Time Detection Transformer) developed by Baidu's team adopts a transformer structure, achieving end-to-end object detection without cumbersome post-processing steps, further improving the real-time performance of the detector [11]. RT-DETR detector has already shown excellent performance in many scenarios. In reality, water bodies such as reservoirs, scenic lakes, and ditches often require large-scale monitoring, and at long distances beyond 30 meters on water surfaces, targets have small sizes and irregular shapes, making their features blurred and unclear, bringing difficulties to target localization and recognition. Additionally, complex and variable backgrounds, water surface lighting and weather

changes, water surface reflection and ripples cause changes in image quality, leading to false detections and missed detections.

For the existing problems, this paper proposes an improved real-time detection model FRT-DETR to achieve fast and accurate detection of small debris targets at long distances on water surfaces, considering the limited hardware resources of unmanned cleaning equipment. The specific improvements are as follows:

Proposing CSPCAA-ADown as the modified network. Through nested attention pyramid structure, it adaptively fuses feature information from different scales. This captures detailed information of water surface debris targets at long distances beyond 30 meters at different scales, allowing the model to obtain more abundant feature information while reducing computational burden.

Proposing CCGAF adaptive feature fusion module for adaptively adjusting feature weights of different layers and weighted fusion, dynamically adjusting dependencies between different layer information to achieve efficient feature fusion.

Using adaptive pooling module WaveletPool (Wavelet Pooling for Neural Networks) for upsampling and downsampling feature maps of small blurred targets [12]. Through its wavelet transform, it adaptively adjusts target scale changes, ensuring detection accuracy even when small targets at long distances beyond 30 meters become blurred, thereby improving the model's processing capability to adapt to practical applications.

2. Rt-Detr

In 2023, technical personnel Wen et al. from Baidu's Visual Technology Department (VIS) developed a detector model RT-DETR that meets real-time detection performance requirements. This model possesses both the end-to-end detection performance of DETR and the lightweight design characteristics of the YOLO series. Its architecture consists of three core components: a lightweight backbone network, an IoU-aware encoder, and a parallel decoder. In the backbone network, CNN architecture multi-scale feature extractors ResNet or CSPNet are employed for multi-scale feature extraction and enhancement, integrating Feature Pyramid Network (FPN) structure for feature fusion [13]. This design achieves effective feature extraction for cross-scale targets while maintaining good computational efficiency. In the encoder, an improved DINO-style architecture is adopted while adding IoU-aware capability. This design improves the model's ability to capture target dependencies while ensuring real-time detection performance. Furthermore, RT-DETR employs an uncertainty minimal query selection strategy to efficiently select initial query objects, and finally generates categories and detection boxes through a decoder with auxiliary detection heads. It is worth noting that RT-DETR does not require NMS post-processing, surpassing many traditional detectors in both detection speed and accuracy, providing an extremely powerful and efficient solution for device deployment and practical applications.

Given the limited computational resources when deploying water surface cleaning equipment and the complex and variable water surface environment, this paper selects r18 as the base model for improvement from the multiple models (r18, r34, r50, L, X) provided by RT-DETR. The network structure of the r18 model is shown in Figure 1.

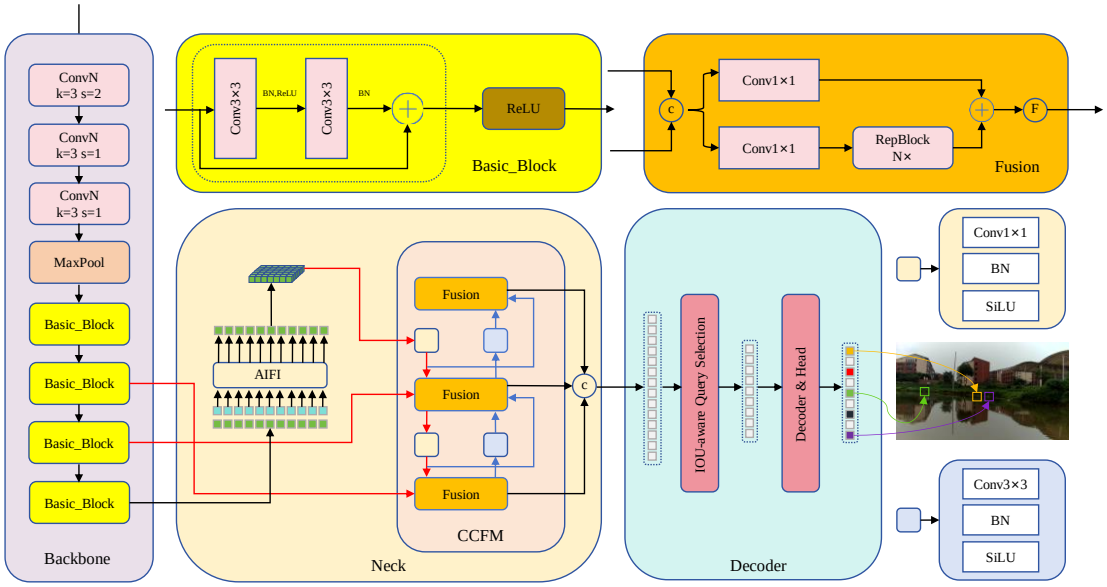

Figure 1. RT-DETR-r18 Structure Diagram

3. Frt-Detr

FRT-DETR (Far Real-Time Detection Transformer) is based on RT-DETR model with improvements in the following aspects: For the backbone network CSPCAA-ADown, the RepNCSPELAN4 [14] network is introduced, while adding the CSP-CAA module composed of the cross-scale attention interaction mechanism CAA module [15] to better capture intra-class differences of targets at different scales, distinguish between different types of debris at long distances beyond 30 meters on water surfaces and differences between background, lighting, etc., to better perform category distinction and improve false detection situations; For the Encoder, the CCGAF module composed of CGAF [16]+CAFM [17] is used for fusion and adaptive adjustment of features from different layers, further improving the lighting effects caused by water surface illumination changes and weather changes, enhancing anti-interference capability in practical applications; For upsampling and downsampling, the wavelet pooling module WaveletPool is used to adaptively adjust target scale changes, achieving good detection

performance even when debris targets at long distances beyond 30 meters on water surfaces have excessively small scales. The improved model structure is shown in Figure 2.

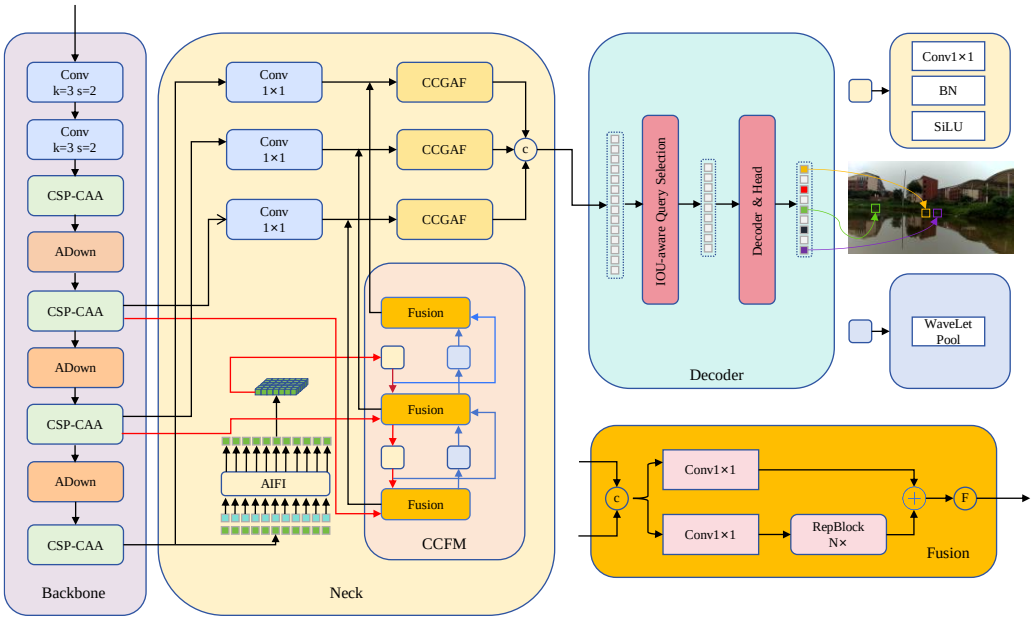

Figure 2. FRT-DETR-r18 Structure Diagram

3.1. Backbone Network Optimization

In the detection of small debris targets at distances beyond 30 meters on water surfaces, the ResNet [18] backbone network in the original model uses residual connection structures that can alleviate gradient vanishing problems in deep networks, but its simple skip connection approach struggles to effectively extract detailed features of long-distance small targets. Secondly, due to the small pixel proportion of long-distance water surface debris targets in images, the backbone network of the original model adopts fixed receptive field sizes and cannot adaptively perceive and adjust the feature extraction range. In this situation, target feature information may be confused and diluted by background, lighting, and other information, making it difficult to accurately capture debris target features beyond 30 meters in complex water surface environments. Finally, the original model backbone lacks explicit modeling of channel attention and spatial attention, making it difficult for the network to highlight feature representations of key regions.

To address the above problems, this paper proposes the CSPCAA-ADown backbone network. This network contains the CSP-CAA module, which achieves adaptive fusion of multi-scale features through the RepNCSPELAN4 module, while introducing the CAA module to adaptively adjust convolution kernel shapes, enhancing the ability to capture feature information of long-distance small targets, suppressing interference from unimportant information, and improving image quality. Additionally, the lightweight downsampling module ADown is adopted to ensure relatively small computational resource requirements while improving detection performance. This backbone network, through the combination of RepNCSP structure and channel attention modules, uses depthwise separable convolution to replace traditional convolution within the module, reducing parameter count while enhancing feature expression capability for long-distance water surface debris. It adaptively adjusts feature weights, effectively addressing different lighting conditions and complex scenarios, achieving fast inference and efficient detection. The overall structure of the CSPCAA-ADown backbone network is shown in Figure 3.

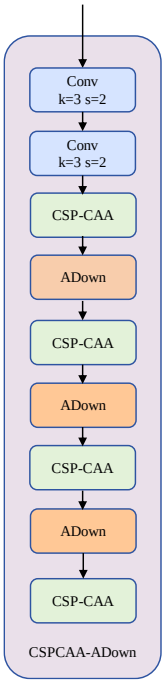

Figure 3. CSPCAA-ADown

3.1.1. CSP-CAA

The CSP-CAA backbone network module aims to enhance feature expression capability and semantic information. This structure adopts a multi-branch design during training to enhance feature expression, while during inference it can be equivalently converted to a single convolution operation, achieving a balance between model inference speed and performance. The CSP-CAA module structure is shown in Figure 4.

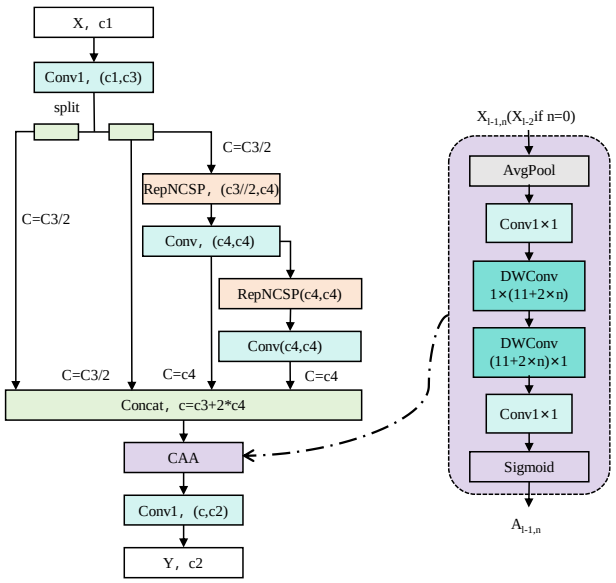

Figure 4. CSP-CAA

First, the input feature map undergoes channel dimension adjustment through a 1×1 convolutional layer. The resulting feature map is then split along the channel dimension into two branches, y1 and y2, each with c3/2 channels. y1 serves as an identity mapping to preserve the original feature information. Next, y2 passes through two cascaded RepNCSP and 3×3 convolutional layers for deep feature extraction, yielding enhanced feature maps y2' and y2". Finally, y1, y2', and y2" are concatenated along the channel dimension and processed through a CAA module for self-attention-based adjustment of weight strategies between different channels, thereby enhancing the information expression of important channels and suppressing interference from complex water surface environmental conditions. This results in high-quality output feature maps. The relevant formulas are as follows:

$$y1, y2 = Split\left[Conv1\times1(x)\right] \tag{1}$$

$$y2^{'} = Conv3\times3\left[RepNCSP(y2)\right] \tag{2}$$

$$y2^{''} = Conv3\times3\left[RepNCSP\left(y2^{'}\right)\right] \tag{3}$$

$$CAA = \sigma\left(Conv\left(Vh\left(Hh\left(Conv\left(AvgPool\left(x\right)\right)\right)\right)\right)\right) \tag{4}$$

$$output = conv1\times1\left(CAA\left(Concat\left(y1, y2', y2''\right)\right)\right) \tag{5}$$

Vh and Hh represent depth-wise separable convolutions in the vertical and horizontal directions respectively, σ denotes the sigmoid activation function, Split represents the channel separation operation, and Concat represents the concatenation operation along the channel dimension.

The CAA module is a cross-dimensional self-aware attention mechanism based on separable convolutions. When integrated with the RepNCSPELAN4 backbone, it enhances the directional features of distant marine debris targets on water surfaces. During the feature modeling process, the CAA module leverages the modeling capabilities of separable convolutions in both horizontal and vertical directions to achieve adaptive enhancement for distance perception and sparse distribution representation of distant marine debris images on water surfaces. Additionally, the CAA module can perform self-aware enhancement for different feature information, exhibiting stronger discriminative capability against variations in background, illumination, and other conditions.

3.1.2. ADown

In the backbone networks of baseline models, simple downsampling operations such as max pooling or strided convolutions are commonly adopted. However, such single operations are in adequate for handling the complex scenarios of detecting distant marine debris beyond 30 meters on water surfaces. The ADown module employs a dual-branch structure to process input feature information through parallel pathways, fully leveraging the advantages of different downsampling strategies. This module uniformly distributes features along the channel dimension, with each branch containing half the channel count of the original information, thereby reducing computational complexity while preserving critical feature information through different processing pathways. This design demonstrates strong environmental adaptability, enabling stable performance under varying illumination conditions and water surface states while reducing computational overhead.

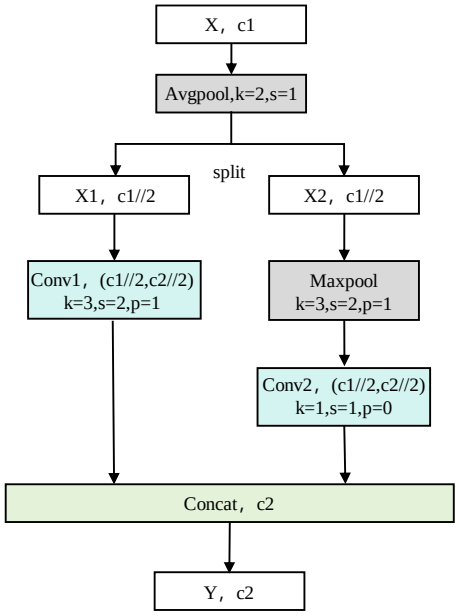

Figure 5. ADown

3.2. Feature Fusion Module

In practical applications, target appearance characteristics undergo dramatic changes due to factors such as water surface reflection, illumination variations, water ripples, different weather conditions, and minute target sizes, making it difficult for the RT-DETR model to effectively capture target detail features. We propose the CCGAF module to enhance its feature representation capability for small targets beyond 30 meters on water surfaces. First, to adaptively adjust feature weights across different scales, making target region features more prominent and enhancing the most discriminative feature expression for debris targets, a spatial attention mechanism is employed to highlight spatial location information of target regions. The CGAF module incorporating channel attention and spatial attention mechanisms from DEA-Net is introduced. Second, to obtain global contextual feature information and enhance feature representation capability, the CAFM module with cross-scale and self-attention mechanisms from HCANet is introduced to better extract and fuse different feature information and representational capabilities, adapting to recognition performance across various water surface scenarios.

Furthermore, the CCGAF module design considers computational efficiency by employing strategies such as feature reuse and parallel computation, maintaining low computational overhead while improving performance. This efficient feature enhancement approach makes it highly suitable for application in practical water surface debris detection systems. The overall structure is shown in Figure 6.

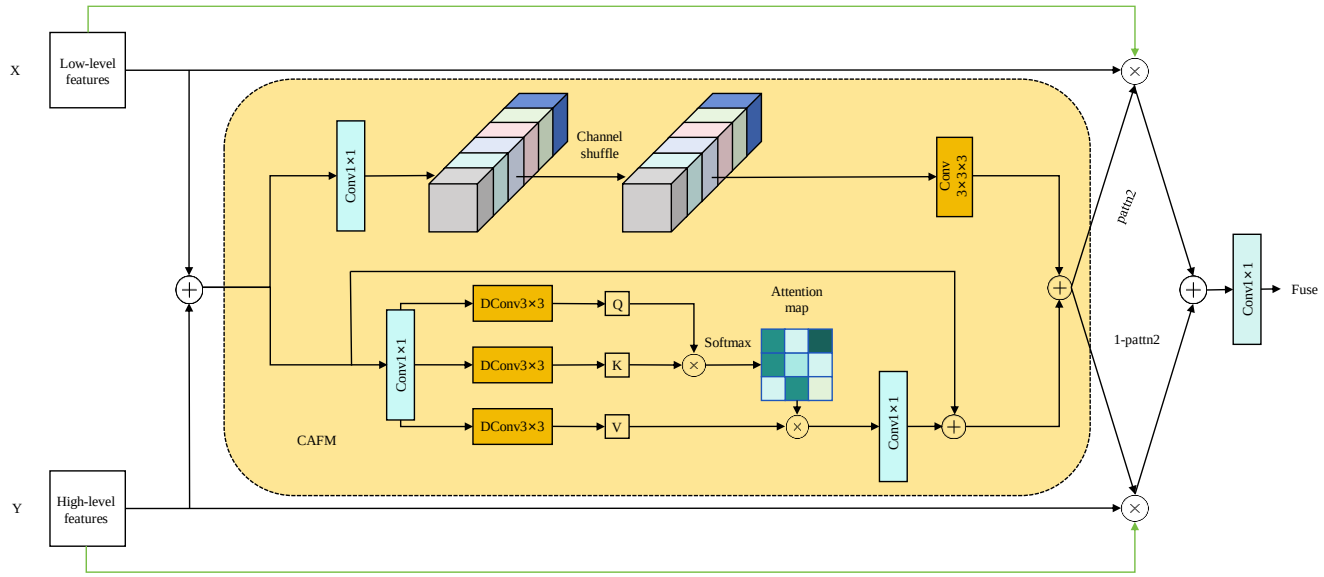

Figure 6. CCGAF

This module inputs the initial feature map obtained by adding two scale feature maps containing low-level features x and high-level features y into the CAFM module to obtain global and local features pattn1. The features are then output to the pixel attention module to obtain pixel-level attention weights pattn2. Finally, the features from both high and low scales are weighted and fused, and the fusion result of the CCGAF module is output through a 1×1 convolution.

$$pattn1 = CAFM\left(x+y\right) \tag{6}$$

$$pattn2 = \sigma\left[PixelAttention\left(x+y, pattn1\right)\right] \tag{7}$$

$$Fuse = Conv\left[pattn2 \otimes x + \left(1-pattn2\right) \otimes y + \left(x+y\right)\right] \tag{8}$$

Where $\otimes$ represents element-wise multiplication.

3.3. Upsampling and Downsampling Optimization

The original model employs max pooling downsampling and nearest neighbor interpolation upsampling, which may result in poor adaptability in complex water surface environments. The forced feature selection mechanism of max pooling downsampling only retains local maximum values, leading to severe feature loss due to spatial information loss when processing small targets after pooling. The simple copy-and-fill strategy of nearest neighbor interpolation cannot effectively reconstruct and recover lost information, lacking accurate localization capability at long distances. The WaveletPool module is based on the principles of inverse wavelet transform to precisely decompose and reconstruct feature information. By decomposing input features into low-frequency and high-frequency components, it simultaneously preserves both global structure and local details of targets, making it particularly suitable for processing distant small targets. Moreover, this module employs fixed wavelet filters without requiring additional parameter computation, achieving performance improvements for distant target recognition on water surfaces while occupying fewer computational resources. Through feature decomposition of input feature maps, followed by convolution and feature fusion, multi-dimensional output images are obtained. Its structure is shown in Figure 7.

$$Y_i = F_i * X_{in}, i \in \left\{LL, LH, HL, HH\right\} \tag{9}$$

$$X_{out} = Concat\left(Y_{LL}, Y_{LH}, Y_{HL}, Y_{HH}\right) \tag{10}$$

Where $*$ represents the convolution operation.

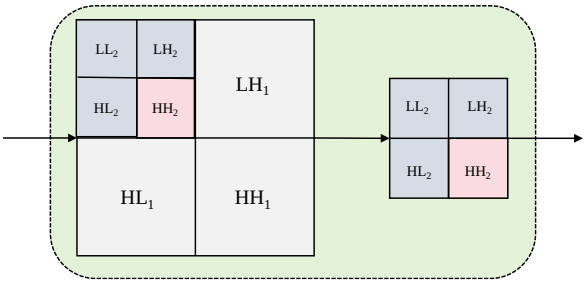

Figure 7. WaveletPool

4. Experiments and Results

4.1. Experimental Setup

The experiments are conducted for training and validation according to the following parameters, and model performance is compared. The specific experimental platform and settings are shown in Table 1. The key parameter settings during experimental training and validation are as follows: imgsz is 640×640, batch size is 8, epochs are 200, and optimizer is AdamW. All experiments do not employ pre-trained weights to ensure consistency during the model training process.

Table 1. Experimental platform and settings

Configuration Environment	Version Model
Deep Learning Framework	Pytorch 1.13.1
Computing Framework	CUDA 11.7
Language	Python 3.8
Computing System	ubuntu20.04
CPU	E5-2680v4
GPU	RTX 3090

4.2. FWSGD Dataset

The dataset was captured using a high-definition industrial camera model JY-USB 5MP equipped with a lens focal length range of 10-100mm. The image construction for the FWSGD dataset was conducted across three different water bodies during morning hours (9:00-12:00) and afternoon hours (16:00-20:00) under varying temporal and lighting conditions, while repeatedly capturing scenes under different weather conditions such as sunny, cloudy, and rainy weather. Images of three common types of garbage—plastic, metal can, and bolt—located beyond 30 meters were collected at different focal lengths and orientations to simulate various scenarios for model training. Through data augmentation, the collected dataset underwent multiple condition simulations including random weather variations, random noise variations, perspective transformations, and grid distortions to expand dataset diversity and enhance model generalization capability. LabelImg was employed for manual annotation of the collected images to generate corresponding txt label files, ultimately yielding 4,599 image data samples, which were divided into a training set, validation set, and test set to ensure objective model evaluation.

Figure 8. Different Lighting Conditions and Different Focal Lengths

Figure 9. Different Complex Backgrounds

Figure 10. Different Enhancement Methods

4.3. Ablation Experiment

To demonstrate that the model improvements have enhancement effects on distant water surface targets while simultaneously reducing computational burden for device deployment, the model validation results are compared as shown in Table 2. Compared to the baseline model r18, the CSPCAA-ADown improved backbone network model achieved a 1.1 percentage point increase in mAP0.5, while Params and GFLOPs were reduced by 9.6M and 24.5G respectively, indicating that this improved backbone network can enhance semantic information capture capability for small distant garbage targets on water surfaces, discard redundant information and strengthen key features, while also reducing computational parameters and improving detection efficiency as well as target detection rate. Although the CCGAF feature fusion improved model shows slightly increased computational load and modest accuracy improvement compared to the baseline model, this module extracts and fuses different feature information, enabling the model to have stronger adaptability and discriminative power under complex water surface lighting variation conditions, which is crucial for improving false detection and misdetection situations under complex lighting conditions. The WaveletPool up-sampling and down-sampling improved model achieves accuracy improvement while reducing some computational load, providing enhanced adaptability for target localization in situations with blurred water surface backgrounds and features at distant ranges. After overall improvements, FRT-DETR achieved 2.5 and 0.9 percentage point improvements in mAP0.5 and mAP0.5-0.95 respectively compared to the baseline model, while Params and GFLOPs were also reduced by 8.1M and 14.7G respectively. The optimized model achieved an FPS value of 25.7f/s, meeting engineering application requirements. Data analysis demonstrates the effectiveness of model improvements for recognizing distant garbage targets beyond 30 meters on water surfaces, not only optimizing the model's detection accuracy but also ensuring appropriate model parameters and computational load for better and more efficient application in practical detection tasks.

Table 2. Ablation Experiment

Model	mAP0.5/%	mAP0.5-0.95/%	P/%	R/%	FLOPs/G	Params/M	FPSbs=1
r18	88.8	42.4	87.2	93	56.9	19.9	42.8
CSPCAA-ADown	89.9	42.3	84.5	93	32.4	10.3	25.3
CCGAF	88.9	43.5	86.7	92	65.1	21.4	32
WaveletPool	89	43	86.3	92	44.8	19.1	41
CSPCAA-ADown+ CCGAF	90.2	44.1	86.3	93	44.8	12.7	26
CSPCAA-ADown+ WaveletPool	90.9	42.8	85.3	94	30	9.3	24.9
CCGAF+ WaveletPool	88.6	42.6	86.8	93	62.8	10.5	33.4
FRT-DETR	91.3	43.3	85.8	94	42.2	11.8	25.7

Note: All parameters in the table are validated on the test set.

4.4. Comparative Experiment

To validate the superior performance of FRT-DETR in detecting distant garbage targets on water surfaces, several mainstream detection models were selected for performance comparison, including YOLOv5m, v8m, v11m, YOLO-DETR hybrid improved models YOLOv5-detr and v8-detr, as well as Mamba-YOLO-B [19] and RT-DETR r34 models. Evaluation was conducted based on core metrics including mAP0.5, mAP0.5-0.95, Params, and GFLOPs.

Based on the validation results shown in Table 3, FRT-DETR achieved the best performance on both mAP@0.5 and mAP@0.5–0.95 metrics, reaching 91.3% and 43.3%, respectively. Compared to other models, this approach demonstrates a significant improvement in accuracy. While FRT-DETR may not match certain lightweight versions of the YOLO series in terms of speed, it offers a substantial advantage in detection accuracy. Additionally, in comparison with the RT-DETR r34 model, FRT-DETR reduces computational workload by more than 50% and decreases the number of parameters by approximately 62%. This remarkable increase in efficiency enables the model to be much lighter for real-world deployment, greatly lowering the demand for computational resources.

The proposed model exhibits outstanding overall performance on long-range water surface trash detection tasks, with notable improvements in detection precision, while maintaining moderate computational cost and runtime efficiency. This balance between computation and model size makes FRT-DETR a highly practical and ideal solution for field applications in water surface trash detection scenarios.

Table 3. Comparative Experiment

Model	mAP0.5/%	mAP0.5-0.95/%	FLOPs/G	Params/M	FPSbs=1
Yolo v5m	75	37.9	64	25	45.3
Yolo v5-detr	88.9	41.2	10.7	5.5	38.4
Yolo v8m	77.5	39.8	78.7	25.8	51.6
Yolo v8-detr	87.3	40.6	11.7	6	49.8
Yolo v11m	78.9	40	67.7	20	28.5
Mamba-YOLO-B	77.8	40.3	44.5	20.4	27.6
RT-DETR r34	90.1	40.5	88.8	31.1	39
FRT-DETR	91.3	43.3	42.2	11.8	25.7

Note: All parameters in the table are validated on the test set.

5. Visual Analysis

This section conducts a comparative analysis of the key parameters and detection results before and after the model enhancement, providing an in-depth assessment of the improved model's performance gains. The experimental results displayed in Figure 11 indicate that the upgraded model, FRT-DETR, outperforms the baseline model, RT-DETR, across all core metrics, including precision (P), recall (R), mAP@0.5, and mAP@0.5–0.95. These findings strongly validate the effectiveness of the proposed improvements.

From the perspective of specific metrics, in terms of accuracy, both models ultimately achieved approximately 90% performance levels, validating the significant effectiveness of the proposed CSPCAA-ADown network structure and CCGAF feature fusion module in improving detection accuracy and stability. Regarding recall rate, FRT-DETR achieved excellent performance of approximately 90% in the later training stages, showing notable improvement compared to RT-DETR's level of about 85%. This improvement is primarily attributed to the WaveletPool adaptive pooling module's effective capture of multi-scale target features. In terms of mAP0.5 metrics, FRT-DETR achieved approximately 5 percentage points improvement over RT-DETR, fully demonstrating the improvement effectiveness of the CCGAF module in feature fusion. Particularly noteworthy is that under the more stringent mAP0.5-0.95 evaluation standard, FRT-DETR still maintained significant advantages, reaching a performance level of approximately 0.42 in the later training stages, showing clear improvement compared to RT-DETR's 0.40. This result indicates that the improved model maintains stable performance advantages even under high-threshold detection standards, further validating the effectiveness of the improvement scheme.

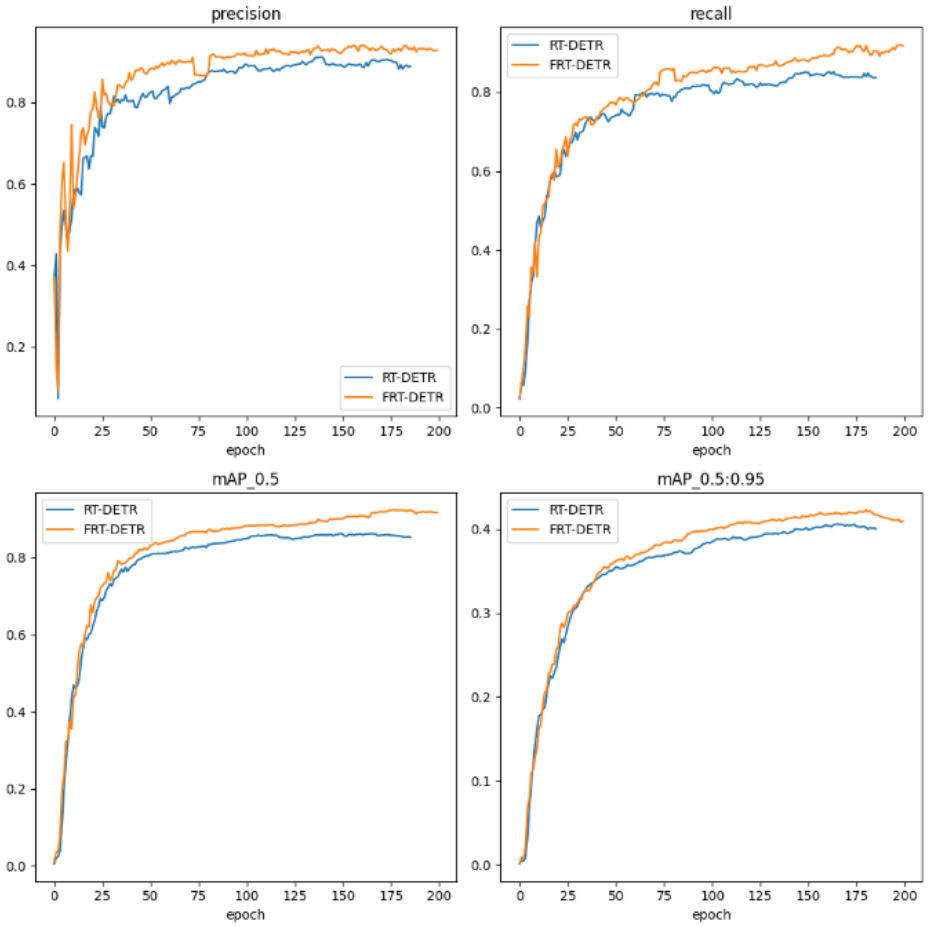

Figure 11. Comparison Curve

As shown in Figures 12 and 13, the improved model significantly ameliorated the missed detection and false detection issues of multiple floating garbage targets in distant water surface areas (regions marked with white circles). Through comparison, it can be observed that compared to the prediction results of the baseline model RT-DETR on the left, the improved FRT-DETR model on the right demonstrates superior detection performance. This improvement is primarily attributed to the CCGAF module's adaptive adjustment capability for feature map channel weights across different channels, significantly enhancing the model's feature discrimination ability under weak water surface lighting conditions. Meanwhile, the spatial attention mechanism effectively highlights the spatial location information of target regions, while the self-attention mechanism accurately captures the dependency relationships between targets and their surrounding environment under complex background conditions, thereby substantially improving missed detection issues and enhancing overall detection accuracy.

Furthermore, this paper's CSPCAA-ADown network, constructed based on the RepNCSPELAN4 backbone network, innovatively integrates the Coordinate Attention Adaptive (CAA) mechanism, significantly enhancing the capability to capture spatial relationships and contextual relationships. This improvement enables the model to more accurately understand the spatial distribution of features and their interrelationships when facing challenges such as uneven lighting and tiny detection targets, effectively reducing false detection rates in complex scenarios.

Figure 12. Improvement In the Situation of Missed Detections

Figure 13. Improvement In the Situation of False Detections

6. Conclusions

Addressing the recognition challenges of tiny garbage on distant water surfaces beyond 30 meters under different lighting and weather conditions in complex backgrounds, this paper proposes the FRT-DETR model based on improved RT-DETR. The model significantly enhances detection performance through multi-level feature extraction and fusion mechanisms, fully considering the special requirements of distant water surface garbage recognition. The research innovation lies in the organic combination of multiple advanced attention mechanisms and feature fusion methods, with the introduction of adaptive pooling strategies to construct a detection framework more suitable for distant water surface garbage recognition tasks. Experimental results

demonstrate that the model exhibits excellent performance in distant detection, effectively reducing false detection and missed detection rates while maintaining stable detection effects under various complex environments. Meanwhile, through optimizing the model structure, computational load and parameter count are significantly reduced, making it more convenient for practical engineering applications and providing reliable target detection support for water surface cleaning equipment, which has important practical significance.

References

[1] Ren S, He K, Girshick R, et al. Towards real-time object detection with region proposal networks, Adv [J]. Neural Inf. Process, 2015, 28.

[2] Redmon J. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Zeng Z C, Xu Y, Wang J Y, et al. A lightweight water surface object detection algorithm based on SOE-YOLO [J]. Journal of Graphics, 2024, 45(04): 736-744.

[4] Li C, Zhou A, Yao A. Omni-dimensional dynamic convolution [J]. arXiv preprint arXiv:2209.07947, 2022.

[5] Du A Q, Guo F L, Zhang Z L, et al. Surface Garbage Detection Method Based on Improved YOLOv5s [J]. Journal of Wuhan Polytechnic University, 2023, 42(05): 98-105+113.

[6] Zhou H P, Li Y H, Dang A P. Small Target Detection of Floating Garbage on Water Surface Based on YOLOv7 with Edge Enhancement [J]. Journal of Langfang Normal University (Natural Science Edition), 2024, 24(02): 45-51.

[7] Zhu L, Wang X, Ke Z, et al. Biformer: Vision transformer with bi-level routing attention [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333.

[8] Vaswani A. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017.

[9] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C] //European conference on computer vision, 2020: 213-229.

[10] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection [J]. arXiv preprint arXiv:2010.04159, 2020.

[11] 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.

[12] Williams T, Li R. Wavelet pooling for convolutional neural networks [C]//International conference on learning representations. 2018.

[13] 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.

[14] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information [C]//European Conference on Computer Vision. Springer, Cham, 2025: 1-21.

[15] Huang Y, Jia W, He X, et al. CAA: Channelized axial attention for semantic segmentation [J]. arXiv preprint arXiv: 2101.07434, 2021.

[16] Chen Z, He Z, Lu Z M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention [J]. IEEE Transactions on Image Processing, 2024.

[17] Hu S, Gao F, Zhou X, et al. Hybrid Convolutional and Attention Network for Hyperspectral Image Denoising [J]. IEEE Geoscience and Remote Sensing Letters, 2024.

[18] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[19] Wang Z, Li C, Xu H, et al. Mamba YOLO: SSMs-Based YOLO For Object Detection [J]. arXiv preprint arXiv: 2406.05835, 2024.

[20] Hu J L, Zhou M, Shen F. Improved RTDETR Detection Algorithm for Small Targets in UAV Applications [J]. Computer Engineering and Applications, 2024, 60(20): 198-206.

[21] Chen L, Zhu J. Water surface garbage detection based on lightweight YOLOv5 [J]. Scientific Reports, 2024, 14(1): 6133.

[22] Redmon J. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[23] Junzhe Z, Fuqiang J, Yupeng C, et al. A water surface garbage recognition method based on transfer learning and image enhancement [J]. Results in Engineering, 2023, 19: 101340.

[24] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-58.