Journal of *Marine Science and Engineering*

MDPI

Article

RTDETR-MARD: A Multi-Scale Adaptive Real-Time Framework for Floating Waste Detection in Aquatic Environments

Baoshan Sun [1,2,*], Haolin Tang [1,2], Liqing Gao [1,2], Kaiyu Bi [1,2] and Jiabao Wen [3]

[1] School of Computer Science and Technology, Tiangong University, Tianjin 300387, China; 2331101236@tiangong.edu.cn (H.T.); lqgao@tiangong.edu.cn (L.G.); 2231081209@tiangong.edu.cn (K.B.)
[2] Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tiangong University, Tianjin 300387, China
[3] School of Electrical and Information Engineering, Tianjin University, Tianjin 300052, China; wen_jiabao@tju.edu.cn
* Correspondence: sunbaoshan@tiangong.edu.cn

Abstract: Accurate and efficient detection of floating waste is crucial for environmental protection and aquatic ecosystem preservation, yet remains challenging due to environmental interference and the prevalence of small targets. To address these limitations, we propose a Multi-scale Adaptive Real-time Detector (RTDETR-MARD) based on RT-DETR that introduces three key innovations for improved floating waste detection using unmanned surface vessels (USVs). First, our hierarchical multi-scale feature integration leverages the gather-and-distribute mechanism to enhance feature aggregation and cross-layer interaction. Second, we develop an advanced feature fusion module incorporating feature alignment, Information Fusion, information injection, and Scale Sequence Feature Fusion components to ensure precise spatial alignment and semantic consistency. Third, we implement the Wise-IoU loss function to optimize localization accuracy through high-quality anchor supervision. Extensive experiments demonstrate the framework's effectiveness, achieving state-of-the-art performance of 86.6% mAP50 at 96.8 FPS on the FloW dataset and 49.2% mAP50 at 107.5 FPS on our custom water surface waste dataset. These results confirm RTDETR-MARD's superior accuracy, real-time capability, and robustness across diverse environmental conditions, making it particularly suitable for practical deployment in ecological monitoring systems where both speed and precision are critical requirements.

Keywords: real-time object detection; floating waste monitoring; adaptive feature aggregation; multi-scale feature fusion

check for updates

Academic Editor: Rafael Morales

Received: 13 April 2025
Revised: 18 May 2025
Accepted: 19 May 2025
Published: 21 May 2025

Citation: Sun, B.; Tang, H.; Gao, L.; Bi, K.; Wen, J. RTDETR-MARD: A Multi-Scale Adaptive Real-Time Framework for Floating Waste Detection in Aquatic Environments. *J. Mar. Sci. Eng.* **2025**, *13*, 996. https://doi.org/10.3390/jmse13050996

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The detection of floating waste on water surfaces is crucial for addressing the global issue of marine pollution [1]. Marine waste (as illustrated in Figure 1a), particularly plastics, poses a severe threat to aquatic ecosystems by disrupting habitats and endangering marine life [2]. Additionally, plastic waste poses risks to human health by facilitating the transport and accumulation of toxic substances. Rivers serve as major conduits between land and ocean, acting as primary pathways for plastic waste to reach marine environments (as illustrated in Figure 1b). Studies estimate that approximately 275 million metric tons of plastic waste are produced annually in coastal regions, with 4.8 to 12.7 million metric tons ultimately entering the ocean [3]. Early detection and efficient removal of floating waste in inland water bodies, such as rivers, are essential to prevent its entry into larger water systems and mitigate its detrimental impacts on both marine and human health.



(a) Marine waste (b) River waste

(c) Cleaning manually (d) Cleaning autonomously

Figure 1. Severe water pollution and cleaning methods.

Existing methods for detecting floating waste can be broadly classified into traditional and modern technology-driven approaches. Traditional methods primarily involve manual inspections and physical sampling, in which human operators visually identify and collect floating waste using boats or nets (as illustrated in Figure 1c). Although straightforward, these methods are both inefficient and costly, rendering them impractical for large-scale or real-time monitoring. In contrast, modern technology-driven approaches have recently emerged, providing more efficient solutions through automation and advanced sensing technologies. For instance, advanced monitoring technologies (as illustrated in Figure 1d), including unmanned aerial vehicles (UAVs) and unmanned surface vehicles (USVs) equipped with high-resolution cameras and sensors, facilitate large-scale water body monitoring, enabling real-time detection and tracking of floating waste [4]. In typical applications, cameras are mounted on the front or top of USVs using stabilizing mechanisms to ensure image clarity during movement. As the USV moves along a predefined path, it continuously captures images of the water surface, which are then used for object detection. These systems are primarily used to identify common floating waste such as plastic bottles, foam debris, and plastic bags, which may be partially submerged or clustered. Moreover, intelligent water surface cleaning robots have been developed to autonomously collect floating waste, significantly improving efficiency and reducing the need for manual labor.

Despite significant advancements in floating waste detection, existing methods continue to encounter substantial challenges. The high costs associated with advanced technologies, including satellite remote sensing and drone-based monitoring, hinder their large-scale deployment, particularly in resource-limited regions. Moreover, many systems face difficulties in detecting small floating waste, such as plastic bottles and fragments, which are often less conspicuous yet pose significant environmental threats. Furthermore, these methods exhibit sensitivity to dynamic environmental factors, including varying lighting conditions and complex surface dynamics, which diminishes their reliability in real-world applications. These limitations underscore the critical need for cost-effective, adaptable, and highly precise solutions to enhance floating waste detection.

To address these limitations, we propose RTDETR-MARD (Multi-scale Adaptive Real-time Detector), a novel framework optimized for USV-based floating waste detection. Our contributions are threefold:

- We leverage the gather-and-distribute (GD) mechanism to enhance multi-scale feature aggregation and cross-layer information interaction, improving fine-grained spatial representation.
- Inspired by GD, we develop a feature fusion framework integrating the Feature Alignment Module (FAM), Information Fusion Module (IFM), Information Injection Module (IIM), and Scale Sequence Feature Fusion Module (S^2F^2M), which jointly refine multi-scale representations and enhance both spatial and semantic consistency. To further improve localization, we adopt the Wise-IoU (WIoU) loss for better anchor supervision.
- Extensive experiments on the FloW and our custom water surface waste datasets demonstrate that RTDETR-MARD achieves 86.6% mAP@0.5 at 96.8 FPS on FloW and 49.2% mAP@0.5 at 97.5 FPS on our dataset, surpassing the RT-DETR baseline in both accuracy and efficiency.

2. Related Work

2.1. Floating Waste Detection

- Datasets

  Recent advancements in deep learning have significantly propelled research in aquatic surface waste detection. To support this research, several benchmark datasets have been developed, providing annotated image collections that facilitate the training and evaluation of detection algorithms. The FloW dataset [5] is a specialized dataset focusing on a single category of floating waste, making it useful for fundamental research on waterborne waste detection. The WSODD dataset [6], on the other hand, provides a more diverse set of annotated images, including both floating waste and other objects commonly found in aquatic environments. Additionally, datasets such as Trash-ICRA19 [7] and AquaTrash [8] have contributed to the field, particularly in the context of underwater object detection. To further advance research in this area, we introduce our dataset, which contains a diverse collection of annotated images for floating waste detection. A summary of these datasets is presented in Table 1.

Table 1. Summary of floating and underwater waste detection datasets.

Datasets	Year	Number of Images	Categories	Application Domain
FloW [5]	2021	2000	1	Floating waste detection
WSODD [6]	2021	7467	14	Floating waste detection
Trash-ICRA19 [7]	2020	5700	3	Underwater waste detection
AquaTrash [8]	2020	369	4	Underwater waste detection
Ours	2025	4000	11	Floating waste detection

- Methods

  In the field of object detection models, researchers have explored various deep learning architectures for aquatic surface waste detection, including the YOLO series [9–14], Faster R-CNN [15], and SSD [16]. These models have been rigorously validated across multiple datasets to evaluate their efficacy in identifying floating waste. For example, Yang et al. [17] introduced an improved version of YOLOv7 specifically adapted to the Chengdu Sha River environment by refining the network aggregation and downsampling modules, which led to notable gains in both detection accuracy and inference speed. Ma et al. [18] further advanced the field by integrating YOLOv5s with UNet segmentation and dark channel prior denoising, effectively enhancing detection

accuracy by mitigating wave and lighting interference. Additionally, Chen et al. [19] implemented a lightweight YOLOv5 variant to address computational constraints in unmanned surface vessel deployments.Deep learning-based object detection has also demonstrated its versatility across other domains, such as the automated identification of urban exterior materials from street view images [20], highlighting its adaptability to diverse visual recognition tasks.

- Application
  Recent advancements in sensing platforms have significantly improved operational capabilities for aquatic waste detection. Notable developments include the work by Luo et al. [21], who developed drone-mounted hyperspectral sensors enabling airborne pollutant mapping through spectral analysis. In parallel, Chang et al. [22] demonstrated an integrated USV system incorporating multi-sensor arrays and robotic manipulators, achieving simultaneous waste identification and automated collection in aquatic environments. Similarly, the UAV- and deep learning-based rebar detection method for concrete columns [23] demonstrates the relevance of UAV-based object detection in construction and aquatic waste applications. The study highlights the growing utility of UAVs in precise object identification, which is crucial for applications in various fields, including waste detection and environmental monitoring.

The significance of floating waste detection as a crucial component of environmental protection and resource management cannot be overstated. However, this field faces several key challenges: (1) Difficulty distinguishing backgrounds arises due to minimal color and texture differences between waste and the water surface, especially under changing lighting conditions, making separation difficult. (2) Small target detection is particularly challenging, as tiny objects like plastic bottles occupy only a few pixels in an image, reducing detection accuracy. (3) Environmental interference, including water surface fluctuations, lighting variations, reflections, and shadows, can further disrupt detection algorithms. (4) Algorithm robustness is essential, as detection models must adapt to diverse conditions such as varying water qualities and weather changes to maintain reliable performance across different scenarios.

2.2. Transformer for Object Detection

The Detection Transformer (DETR) [24] establishes an end-to-end object detection framework based on the Transformer architecture [25]. Its fundamental innovation lies in reformulating object detection as a set prediction task, utilizing a Transformer encoder–decoder structure to directly generate fixed-size detection sets. This approach eliminates traditional manual components, including Non-Maximum Suppression (NMS) and anchor generation, thereby simplifying conventional detection pipelines. While demonstrating strong performance on large-scale objects, DETR encounters two primary limitations: suboptimal detection capability for small targets and slow training convergence speed due to its global attention mechanism's computational complexity.

To address these challenges, researchers have developed multiple enhanced variants. Deformable DETR [26] improves small object detection through deformable attention mechanisms that adaptively sample sparse spatial locations. DINO-DETR [27] accelerates model convergence and enhances small target recognition by incorporating knowledge distillation strategies during training. DAB-DETR [28] further combines deformable convolution operations with attention mechanisms, enabling the dynamic adjustment of receptive fields to better handle densely arranged small objects in aquatic environments while maintaining computational efficiency.

RT-DETR (Real-Time Detection Transformer) [29], developed by Baidu researchers, establishes a novel real-time end-to-end detection framework that addresses the compu-

tationally prohibitive nature of conventional DETR architectures while preserving post-processing-free advantages. This Transformer-based solution introduces three key technical contributions: (1) an efficient hybrid encoder architecture for multi-scale feature processing, (2) an IoU-aware query selection mechanism optimizing initial query generation, and (3) a Dynamic decoder layer configuration enabling speed–accuracy tradeoffs without model retraining. Experimental validations demonstrate RT-DETR achieves state-of-the-art latency–accuracy balance across multiple hardware platforms, advancing real-world deployment potential for aquatic waste detection systems.

Compared to CNN-based detection models, Transformer architectures such as RT-DETR offer enhanced performance in complex scenarios where understanding global context is essential. For example, Wang [30] employed CNN-based detectors like Faster R-CNN and YOLOv8 with various backbones to perform window and window-state detection in structured building façade images, achieving reliable results. However, these CNN models primarily depend on local receptive fields, which can limit their performance in visually complex environments. In contrast, Hwang et al. [31] demonstrated that Transformer-based architectures, including ViT and PVT, outperformed CNNs when used as backbones for YOLOv10 in estimating building dimensions from street view images, particularly in tasks requiring richer spatial reasoning and precise boundary localization. Inspired by such findings, we adopt a Transformer-based approach for our aquatic waste detection task. Floating debris in aquatic environments often exhibits weak contrast, irregular shapes, and is subject to visual interference such as reflections, occlusions, and background clutter. In these challenging conditions, the global self-attention mechanism of Transformers allows for more effective distinction between targets and background, resulting in improved generalization and robustness compared to traditional CNN methods.

3. Methods

Figure 2 illustrates the architectural design of the proposed RTDETR-MARD model. Our key innovation lies in utilizing the GD mechanism [32] to propose a novel feature fusion module that enhances multi-scale feature aggregation and processing. The selection of the GD mechanism was primarily motivated by the specific challenges in surface floating object detection, where the targets vary significantly in scale, shape, and appearance. Traditional fusion structures like FPN often suffer from information loss and insufficient cross-scale interaction, which can hinder the accurate detection of small or occluded objects. The GD mechanism works by dynamically gathering and fusing information from all levels and subsequently distributing it across different layers. This design aligns well with the needs of our application scenario, as it enables more efficient integration of both low-level spatial features and high-level semantic context. This process improves cross-layer information flow, avoids information loss inherent in traditional structures like FPN, and strengthens the model's ability to capture both fine-grained spatial details and global context across multiple scales.

Building on the GD mechanism, we design a comprehensive feature fusion architecture that includes FAM, IFM, IIM, and S^2F^2M [33]. These modules work in tandem to enhance feature representation by aligning, fusing, and injecting information from different scales, while refining spatial alignment and semantic consistency. This modular approach ensures robust multi-scale feature integration, which is crucial for accurate detection of small, floating waste in aquatic environments. Additionally, to further improve localization accuracy, we incorporate the WIoU loss function [34], which dynamically adjusts gradient contributions based on the prediction quality. This prioritizes high-confidence detections and suppresses error-prone anchors, enabling more accurate localization of small objects in challenging environments.

As shown in Figure 2, the inputs to the neck include feature maps B_2, B_3, B_4, and B_5 are extracted from the backbone, $B_i \in \mathbb{R}^{N \times C_{Bi} \times R_{Bi}}$. The batch is denoted by N, the channel by C, and the size by $R = H \times W$. The resolutions of $R_{B_2}, R_{B_3}, R_{B_4}, R_{B_5}$ are given as follows:

$$R_{B_2} = R, R_{B_3} = \frac{1}{2}R, R_{B_4} = \frac{1}{4}R, R_{B_5} = \frac{1}{8}R. \tag{1}$$

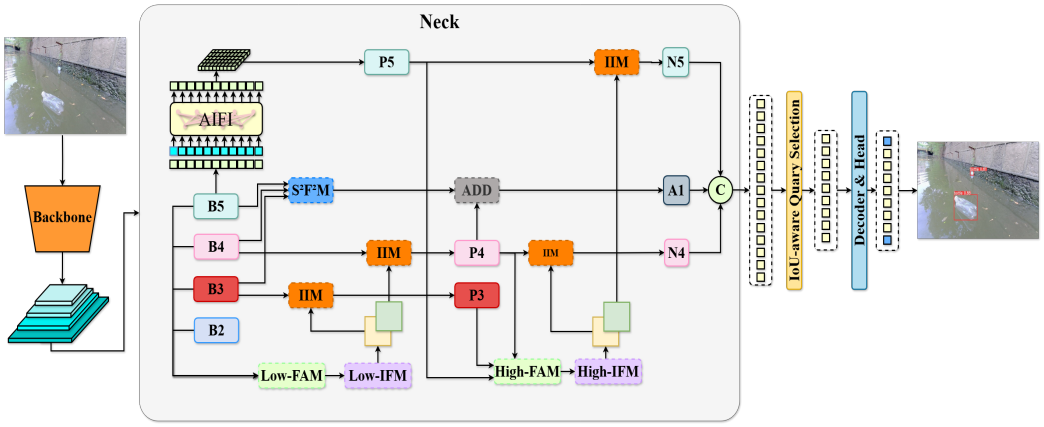


Figure 2. RTDETR-MARD model structure.

3.1. Feature Aggregation Module

The FAM is crucial for feature alignment, with distinct implementations at low levels and high levels to accommodate feature maps of varying resolutions. This dual-branch architecture enables effective integration of deep semantic features and shallow spatial features from the backbone network, ultimately generating comprehensive feature representations optimized for target detection.

The Low-FAM aims to standardize the multi-scale feature maps (e.g., B_2, B_3, B_4, and B_5) from the backbone to a uniform size for subsequent feature fusion operations. As shown in Figure 3, the Low-FAM processes four input feature maps: B_2, B_3, B_4, and B_5. To ensure feature alignment, B_2 and B_3 are downsampled via adaptive average pooling , while B_5 is upsampled using bilinear interpolation, unifying all feature sizes to match that of B_4 ($R_{B4} = \frac{1}{4}R$). By standardizing the feature dimensions to R_{B4}, we obtain the aligned feature map F_{align}, facilitating efficient feature aggregation and subsequent processing. The formula is as follows:

$$R'_{B2} = \text{AvgPool}(R_{B2}), \tag{2}$$

$$R'_{B3} = \text{AvgPool}(R_{B3}), \tag{3}$$

$$R'_{B5} = \text{Bilinear}(R_{B5}), \tag{4}$$

$$F_{align} = [R'_{B2}; R'_{B3}; R_{B4}; R'_{B5}]. \tag{5}$$

The choice of B_4 as the target alignment size was driven by the need to balance accuracy and efficiency. Aligning all feature maps to the resolution of B_4 ensures the preservation of sufficient low-level details while effectively reducing computational overhead.

The High-FAM aims to standardize multi-scale high-level feature maps (e.g., R_{P3}, R_{P4}, R_{P5}) to a uniform size, enabling efficient feature fusion and optimizing the processing capability of the Transformer module. As shown in Figure 4, the High-FAM processes three input feature maps: R_{P3}, R_{P4}, and R_{P5}. To ensure feature alignment, AvgPool is applied to downsample the input features to match the smallest size within the group $R_{P5} = \frac{1}{8}R$. This operation not only standardizes feature dimensions but also facilitates information aggregation while reducing computational complexity, thereby improving the computational efficiency of the Transformer module. The formula is as follows:

$$R'_{P3} = \text{AvgPool}(R_{P3}), \tag{6}$$

$$R'_{P4} = \text{AvgPool}(R_{P4}), \tag{7}$$

$$F_{align} = [R'_{P3}; R'_{P4}; R_{P5}]. \tag{8}$$

By aligning all feature sizes to R_{P5}, High-FAM enhances the model's ability to extract high-level semantic information while maintaining computational efficiency.

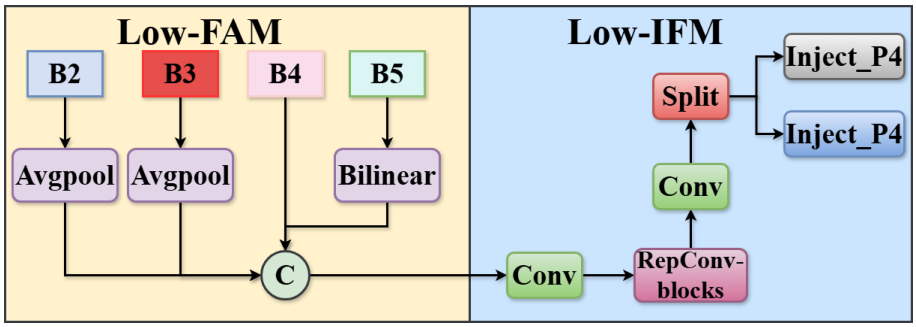


Figure 3. The low-level gather-and-distribute branch.

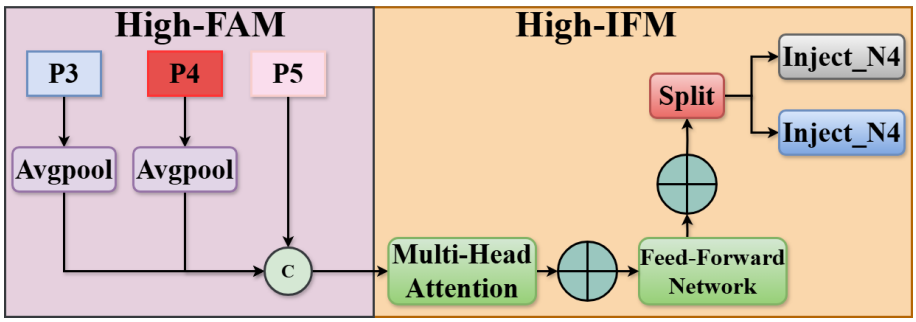


Figure 4. The high-level gather-and-distribute branch.

3.2. Information Fusion Module

The IFM plays a pivotal role in enhancing feature representation within the network architecture. By integrating feature maps from multiple semantic levels, it strengthens the model's capacity to recognize targets of varying scales. This module achieves effective feature fusion and information transmission through two distinct components: the Low-IFM, responsible for aggregating low-level details, and the High-IFM, which captures and refines high-level semantic information.

As illustrated in Figure 3, the Low-IFM implementation begins by receiving the feature map F_{align}, which has been aligned by Low-FAM. Specifically, F_{align} is constructed by summing the feature channels from multiple levels, formulated as

$$F_{align} = \text{Low-FAM}([B2, B3, B4, B5]), \tag{9}$$

where the number of input channels is calculated as $\text{sum}(C_{B2}, C_{B3}, C_{B4}, C_{B5})$. Then, F_{align} is processed by a multilayer reparameterized convolutional block (RepBlock) to produce the fused feature map:

$$F_{fuse} = \text{RepBlock}(F_{align}), \tag{10}$$

where the output channel of F_{fuse} is set to $C_{B4} + C_{B5}$, with an adjustable middle channel (e.g., 256) to accommodate varying model sizes. Next, F_{fuse} is segmented in the channel dimension into two feature maps, F_{inj_P3} and F_{inj_P4}, as follows:

$$F_{inj_P3}, F_{inj_P4} = \text{Split}(F_{fuse}). \tag{11}$$

These split feature maps are subsequently integrated with features from different hierarchical levels to strengthen multi-scale representation. Finally, an attention mechanism is employed to incorporate global context into the local feature space, resulting in enriched high-level feature representations.

The High-IFM focuses on fusing high-level feature maps generated by Low-IFM to capture information about large-sized targets. As shown in Figure 4, High-IFM begins by receiving the aligned feature maps F_{align}, which are generated by High-FAM by processing features from $[P3, P4, P5]$, formulated as

$$F_{align} = \text{High-FAM}([P3, P4, P5]). \tag{12}$$

Then, F_{align} is processed through a Transformer block to capture long-range dependencies using the multi-head attention mechanism and a Feed-Forward Network (FFN), generating the global fusion features:

$$F_{fuse} = \text{Transformer}(F_{align}). \tag{13}$$

Next, the number of channels in F_{fuse} is reduced using a 1×1 convolution operation to $\text{sum}(C_{P4}, C_{P5})$, allowing for better feature adaptation. The reduced feature map is then partitioned into F_{inj_N4} and F_{inj_N5} along the channel dimension as follows:

$$F_{inj_N4}, F_{inj_N5} = \text{Split}(\text{Conv}1 \times 1(F_{fuse})). \tag{14}$$

These split feature maps are subsequently employed for fusion with the current level feature to enhance hierarchical representation. In the final stage, global context is incorporated into local features through an attention-based strategy, producing refined high-level fusion features.

3.3. Information Injection Module

The IIM employs the attention mechanism to integrate global information, fused by the IFM, into the local features. As illustrated in Figure 5, this strategy enhances the semantic information of the feature map and significantly improves the model's ability to detect the target's location. This fusion strategy substantially enhances the model's capability to accurately detect targets across various scales and excels particularly in small object detection.

To describe the mechanism of IIM, we take the low-level operation as an example. In this level, the local features $F_{local} = B_i$ are directly obtained from the current layer's input. Meanwhile, the global injected features F_{inj_Pi} are generated by an external module (such as IFM), which integrates multi-level, multi-scale contextual information to extract global representations.

The global injected features F_{inj_Pi} are processed through a convolutional layer and a Sigmoid function to generate the global activation weight map $F_{global_act_Pi}$, and through another convolutional layer to generate the global embedding feature $F_{global_embed_Pi}$, both of which are resized to match the size of the local features. The local features are processed through a convolutional layer to extract the local embedding feature $F_{local_embed_Pi}$.

Subsequently, the local embedding features are element-wise multiplied with the global activation weight map for dynamic weighting, and then element-wise added to the global embedding features, completing the feature fusion to obtain the fused feature $F_{att_fuse_Pi}$. Finally, the fused features are passed through a RepBlock for further processing, outputting the enhanced feature P_i.

$$F_{global_act_Pi} = \text{Resize}(\text{Sigmoid}(\text{Conv_act}(F_{inj_Pi}))). \tag{15}$$

$$F_{global_embed_Pi} = \text{Resize}(\text{Conv}_{global_embed_Pi}(F_{inj_Pi})). \tag{16}$$

$$F_{att_fuse_Pi} = \text{Conv}_{local_embed_Pi}(B_i) \times F_{ing_act_Pi} + F_{global_embed_Pi}. \tag{17}$$

$$P_i = \text{RepBlock}(F_{att_fuse_Pi}). \tag{18}$$

The same mechanism is applied in the high-level operation of IIM, with necessary adaptations to process high-level features. At this level, F_{local} corresponds to P_i, ensuring consistency in the information injection process across different feature levels.

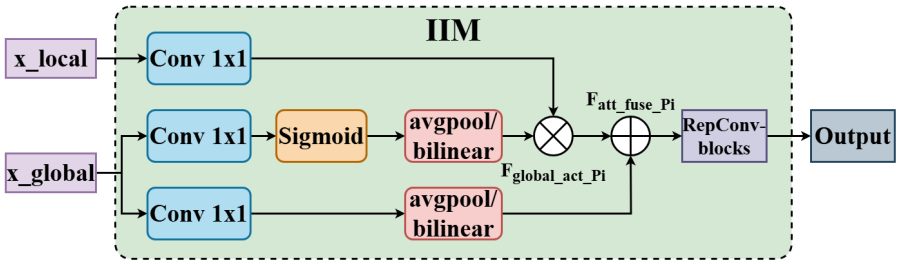


Figure 5. Information injection module.

3.4. Scale Sequence Feature Fusion

S^{2}F^{2}M enhances the model's ability to detect floating debris of various sizes by effectively integrating multi-scale features. As illustrated in Figure 6, S^{2}F^{2}M adopts a sequential multi-scale feature fusion strategy. It first extracts feature maps from the backbone layers B3, B4, and B5, which are then convolved with 2D Gaussian kernels of increasing standard deviation to smooth the feature maps and extract scale-related information.

In the specific context of surface floating object detection, targets are typically small in size, exhibit low contrast with the water surface, and often appear in dense clusters, leading to frequent occlusion and misdetection. To address these challenges, S^{2}F^{2}M generates multiple scale-aware representations by applying Gaussian smoothing with different kernel sizes, which helps the model retain both fine-grained details and global semantic context.

Subsequently, these multi-scale feature maps are horizontally stacked, and 3D convolution is employed to extract scale sequence features, enabling the model to better capture spatial variations in object appearance. Since the feature maps generated from Gaussian smoothing have different resolutions, S^{2}F^{2}M uses the nearest neighbor interpolation method to align the B4 and B5 layers to match the resolution of B3. The B3 layer serves as the reference because it contains rich information about small objects.

$$F_\sigma(i,j) = \sum_u \sum_v f(i-u, j-v) \times G_\sigma(u,v). \tag{19}$$

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{20}$$

Building upon this, S^{2}F^{2}M applies the Unsqueeze method to transform 3D tensors into 4D tensors, changing their shape from $[H, W, C]$ to $[D, H, W, C]$ for more effective feature

fusion. Then, these 4D feature maps are concatenated along the depth dimension to form a unified 3D feature map:

$$F_{4D} = \text{Concat}(F_{3D}^{P3}, F_{3D}^{P4}, F_{3D}^{P5}). \tag{21}$$

Finally, 3D convolution, 3D batch normalization, and the SiLU activation function are applied to further refine the feature representation, ensuring better discriminability of the fused features:

$$F_{out} = \text{SiLU}(\text{BN}(\text{Conv3D}(F_{4D}))). \tag{22}$$

This method significantly enhances the robustness of multi-scale feature fusion under challenging conditions such as small object size, dense object distribution, and background interference caused by water reflections and ripples.

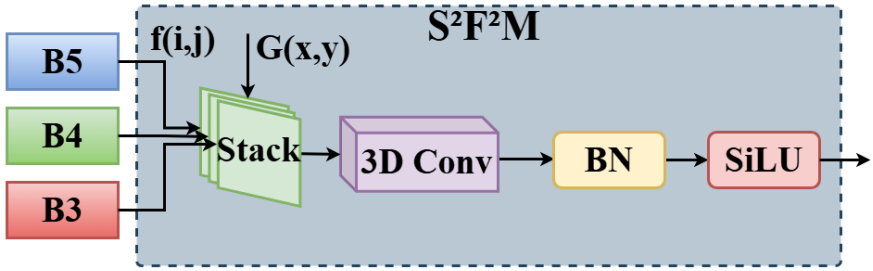


Figure 6. The structure of S^2F^2M.

3.5. Loss Function

In target detection, the accuracy of bounding box regression is a crucial factor affecting model performance. However, traditional bounding box regression methods, including IoU, GIoU [35], CIoU, SIoU [36], EIoU [37], and DIoU [38], face limitations in addressing the imbalance problem. Specifically, a small number of bounding boxes that overlap with the target box can dominate the optimization process, leading to suboptimal performance for detecting small targets, such as water surface floaters. WIoU incorporates a Dynamic Non-Monotonic Focusing Mechanism (DNFM), which adaptively modifies the gradient gain distribution in response to the quality of anchor boxes, thereby improving detection accuracy, particularly for small objects. In this paper, we replace the GIoU loss function in the original algorithm with the WIoUv1 loss function. WIoUv1 modifies the loss function by incorporating Distance Attention and a two-layer attention mechanism, enabling the model to focus on anchor frames of average quality. The formula of WIoUv1 is shown below:

$$L_{WIoUv1} = R_{WIoU} L_{IoU}, \tag{23}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{24}$$

where $R_{WIoU} \in [1, e)$ enlarges L_{IoU} for normal sample anchor frames; $L_{IoU} \in [1, e)$ reduces R_{WIoU} for high-quality sample anchor frames, and focuses on the distance between the centers when the anchor frames overlap well with the target frames. WIoU focuses on the distance between the centroids when the anchor frame overlaps the target frame better.

4. Experiments and Results

4.1. Dataset

As shown in Figure 7, the dataset contains floating bottle samples collected from water surfaces. Figure 8 reveals that small targets constitute more than half of the total

markers. Thus, this dataset is well suited to evaluate the model's ability to detect small water floating waste.



Figure 7. Dataset example.

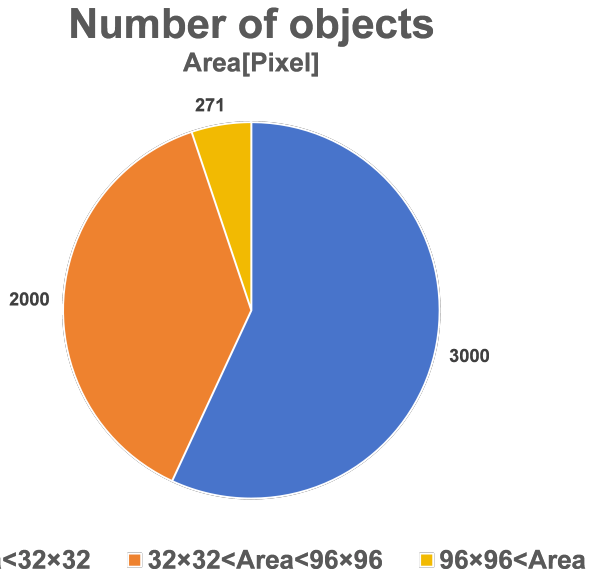


Figure 8. The data distribution of FloW-Img.

4.2. Implementation

All experiments were conducted on a high-performance workstation equipped with an NVIDIA RTX 3090 GPU, providing adequate computational capacity for both training and evaluation. The software environment was based on Python 3.8 and PyTorch 1.9.1, running on Windows 11 with CUDA 11.1 to enable efficient GPU acceleration.

To ensure fairness and comparability in evaluating the model effects, all ablation and comparative experiments were conducted without using pre-trained weights. Details of the experimental environments and parameter configurations are provided in Table 2.

Table 2. Training hyperparameters for RTDETR- MARD.

Parameters	Values
Epoch	300
Lr0	0.001
Lrf	1.0
Image Size	640
Batch Size	4
Workers	4
Momentum	0.9
Weight_Decay	0.0001

4.3. Evaluation Metrics

4.3.1. Precision

Precision is a measure of the proportion of true positives in the positive predictions of a model. In the context of target detection, Precision can be defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{25}$$

where True Positives (TPs) are the number of samples that the model correctly predicts as the positive class. False Positives (FPs) are the number of samples that the model incorrectly predicts as the positive class.

4.3.2. Recall

Recall, or True Positive Rate, is a key metric in object detection, reflecting the proportion of actual positives correctly identified by the model. It is computed in the algorithm as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{26}$$

where False Negative (FN) is the number of actual positive samples that the model incorrectly predicts as negative.

4.3.3. mAP

Mean Average Precision (mAP) is a crucial evaluation metric in target detection tasks, used to assess a model's overall performance across multiple categories. The metric is obtained by taking the mean AP over all categories, thus summarizing the model's effectiveness in handling diverse object classes. The mAP is calculated using the formula

$$\text{mAP} = \frac{1}{N} \sum_{j=1}^{N} (\text{AP})_j. \tag{27}$$

$$AP = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}). \tag{28}$$

4.4. Analysis of Results

4.4.1. Comparison of Computed Parameter Quantities of RTDETR Model Versions

RTDETR offers several variants by utilizing ResNet [39] backbone networks with varying depths and complexities, such as RTDETR-R18, RTDETR-R34, RTDETR-R50, and RTDETR-R101. Figure 9 illustrates that the number of parameters and computational requirements of these models increase significantly with greater network depth. Models with a higher number of parameters and computational demands may not be optimal for float-

ing waste detection tasks with limited computational resources. Considering the model's overall performance and the balance between parameter count and computational burden, we selected RTDETR-R18 as the benchmark model and conducted further optimization and enhancement based on it.

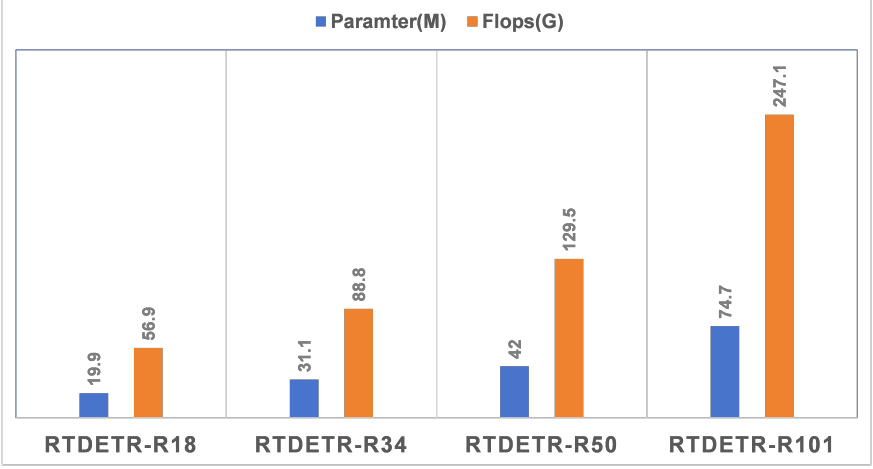


Figure 9. Parameters and FLOPs for each version of RTDETR.

4.4.2. Comparison with Other Models

The experimental results presented in Table 3 and Figure 10 demonstrate that the RTDETR model surpasses the YOLO series in floating waste detection. Under identical environmental conditions, the mAP of the RTDETR-R18 model improves by 9.8%, 4.4%, and 6.7% over YOLOv5m, YOLOv8m, and YOLOv10m, respectively, for models of the same size. Furthermore, the enhanced RTDETR-MARD model surpasses RTDETR-R18 in small target detection, achieving improvements of 11.6%, 6.2%, and 8.5% in mAP over YOLOv5m, YOLOv8m, and YOLOv10m, respectively, without a significant increase in parameters and computational complexity. Although the YOLOv5l and YOLOv8l models exhibit substantial increases in parameters and computational requirements, their performance remains inferior to that of the RTDETR-R18 and RTDETR-MARD models.

The training process illustrated in Figure 11 reveals that RTDETR-MARD significantly surpasses RTDETR in mAP performance, despite a comparable number of convergence rounds between the two models. This finding confirms the efficacy of our enhanced approach. Specifically, RTDETR-MARD demonstrates superior detection accuracy in small target detection for water-floating waste, indicating that the improved strategy significantly boosts the model's performance in this area. This outcome not only highlights the practical effectiveness of our method but also strongly supports further model optimization for water-floating waste detection tasks.

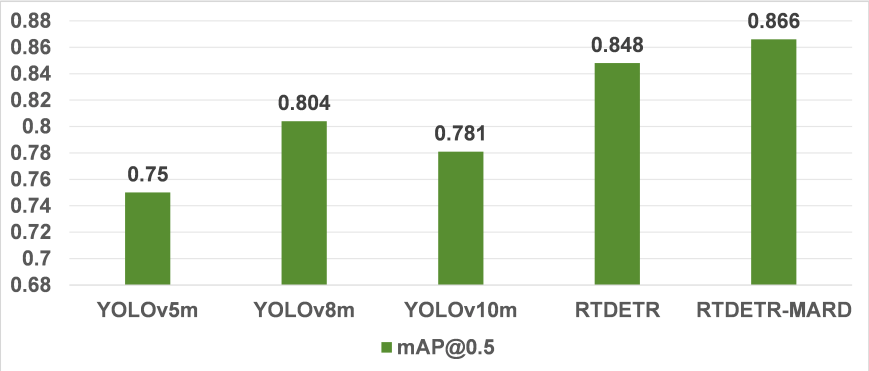


Figure 10. mAP comparison of similar-parameter models on the FloW dataset.

Table 3. Comprehensive comparison of object detectors on the FloW dataset.

Method	P (%)	R (%)	mAP50 (%)	#Param (M)	FLOPs (G)
YOLOv5s	82.3	69.2	75.1	7.0	15.8
YOLOv5m	82.2	68.7	75.0	20.9	47.9
YOLOv5l	82.2	71.6	74.4	46.1	107.6
YOLOv5n	78.7	66.8	70.8	1.8	4.1
YOLOv8n	80.7	69.2	75.0	3.0	8.1
YOLOv8m	82.7	74.5	80.4	25.9	78.7
YOLOv8l	83.1	75.1	81.1	43.6	164.8
YOLOv8s	82.6	71.8	77.7	11.1	28.4
YOLOv10m	81.6	70.1	78.1	16.4	63.4
RTDETR-R18	85.1	81.0	84.8	19.9	56.9
Our model	85.3	83.2	86.6	22.5	63.2

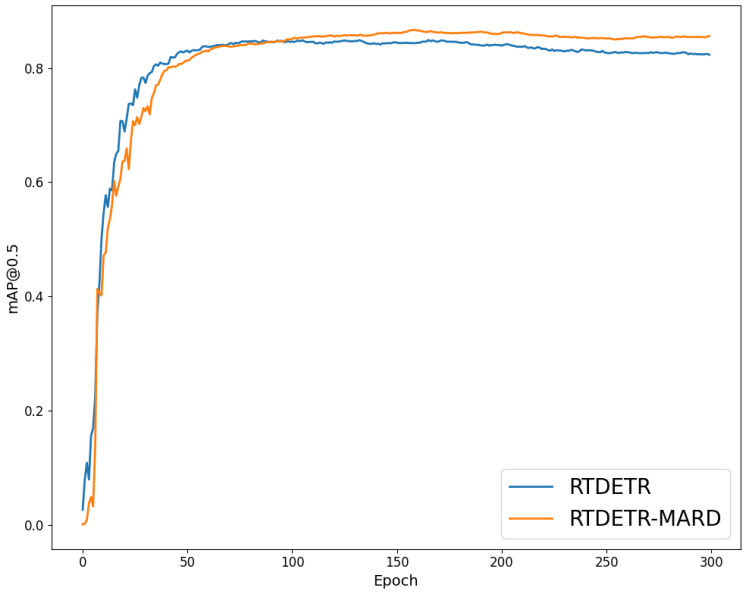


Figure 11. mAP of RTDETR versus RTDETR-MARD.

4.4.3. Comparison of Different IoU Loss Functions on Various Aspects of Performance

In order to investigate the role of different IoU loss functions in enhancing detection accuracy, we performed ablation experiments within the RTDETR-MARD framework. The results are detailed in Table 4.

Table 4. Experimental results of RTDETR-MARD using various types of IoUs on the FloW dataset.

IoU Types	P	R	mAP50 (%)
GIoU	0.855	0.817	86.0
CIoU	0.842	0.817	85.1
SIoU	0.864	0.797	85.5
EIoU	0.851	0.798	85.1
DIoU	0.852	0.807	85.2
Wise-IoUv1	0.853	0.832	86.6
Wise-IoUv2	0.841	0.825	85.5
Wise-IoUv3	0.843	0.811	84.9

To thoroughly explore the optimal performance configuration of the RTDETR-MARD model for floating waste detection, this paper systematically evaluates several popular IoU loss function variants—namely GIoU, CIoU, SIoU, EIoU, DIoU, and three versions of Wise-

IoU—using the FloW dataset. This experimental design aimed to identify the loss function most suitable for enhancing both the accuracy and efficiency of the RTDETR-MARD model in floating waste detection. As shown in Table 4, since the number of parameters (22.5 M) and GFLOPs (63.2) remain constant across all configurations, they were omitted from the table to avoid redundancy and improve readability. However, Wise-IoUv1 outperforms the other loss functions in terms of detection effectiveness.

4.4.4. Ablation Experiment

To evaluate the contribution of each proposed improvement, we conducted a series of comparative experiments based on the RTDETR-R18 architecture on the FloW dataset. The corresponding results are presented in Table 5. As illustrated in Table 5, incorporating the Feature Fusion Mechanism (FFM) into the network yields a 0.4% gain in mAP50 over the baseline, suggesting that this module contributes to improved detection performance, particularly for small objects. FFM comprises FAM, IFM, and IIM, each playing a distinct role in facilitating the network's feature fusion process. This allows the model to function effectively even without additional fusion strategies such as S^2F^2M.

Furthermore, when FFM and S^2F^2M are jointly integrated, the mAP50 increases by an additional 1.2%, indicating that S^2F^2M plays a positive role in further enhancing cross-scale feature representation. When the FFM and S^2F^2M are further combined with the WIoU loss function, the model's parameter count increases from 19.8 M to 22.5 M, and FLOPs rise from 56.9 G to 63.2 G. Despite this moderate increase, the mAP50 improves by 1.8%, demonstrating the superior performance of the final improved model in detecting small floating waste targets on water surfaces.

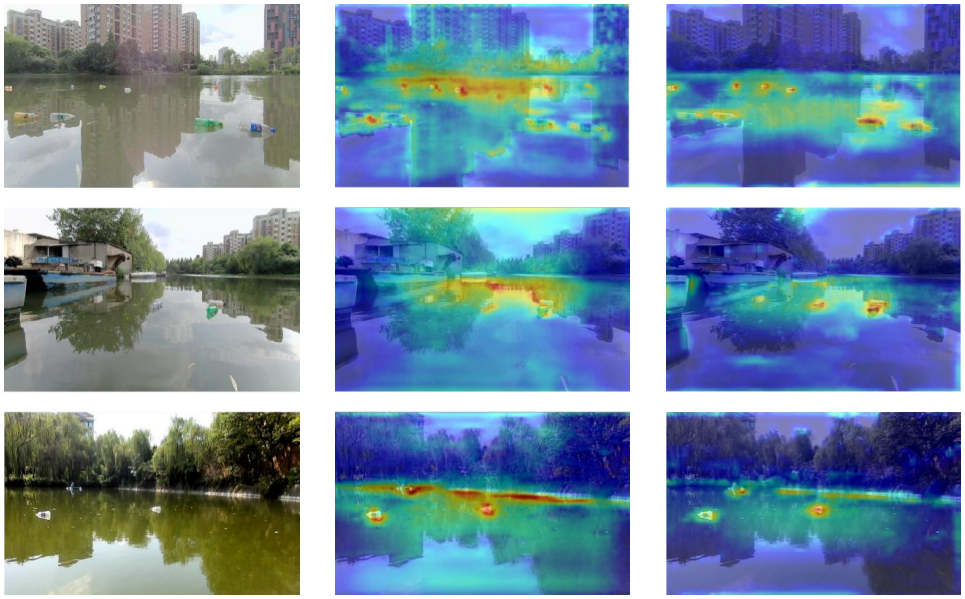
Additionally, the FPS results indicate that, while improving detection performance, the RTDETR-MARD model maintains a high frame rate, showcasing its excellent real-time capability, further proving its advantages in practical applications.

Table 5. Ablation experiment results.

FFM	S^2F^2M	WIoU	mAP50 (%)	Params (M)	FLOPs (G)	FPS
-	-	-	84.8	19.9	56.9	123.7
✓	-	-	85.2	22.3	59.9	107.1
-	-	✓	85.1	19.9	56.9	137.9
✓	-	✓	85.1	22.3	59.9	106
✓	✓	-	86.0	22.5	63.2	98.7
✓	✓	✓	86.6	22.5	63.2	96.8

4.4.5. Experiment Results Visualization

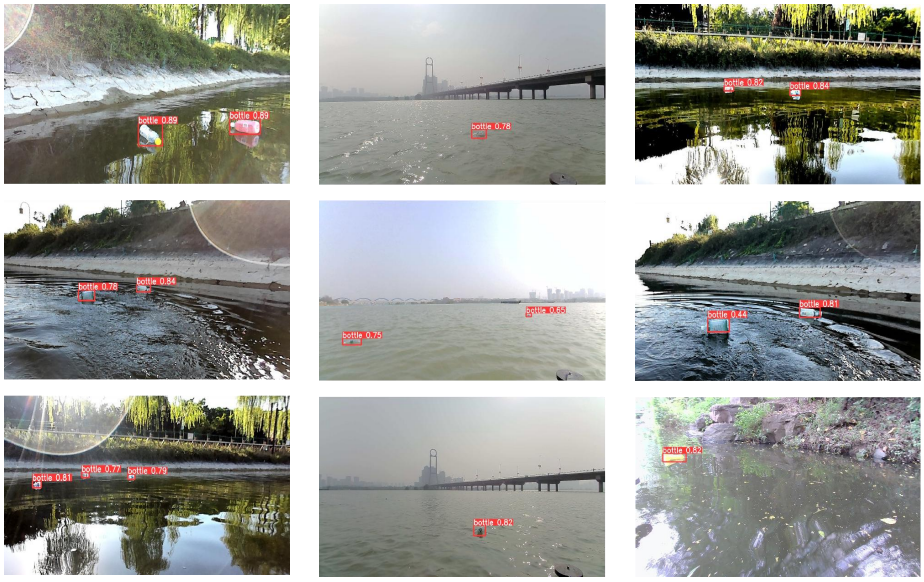
In order to visually interpret the feature representations acquired during network training, we performed a heat map-based analysis on two representative models: RTDETR and RTDETR-MARD. The results are presented in Figure 12, where (a) shows the original image, and (b) and (c) illustrate the heat maps generated by RTDETR and RTDETR-MARD, respectively. As shown in the figures, the heat maps produced by the baseline model exhibit high responses in numerous uncorrelated regions, indicating that the RTDETR model tends to be overly sensitive to water surface backgrounds. In contrast, the heat map generated by the RTDETR-MARD model shows high-response regions that accurately encompass target objects such as plastic bottles and cans. This suggests that the RTDETR-MARD model effectively mitigates the over-sensitivity issue observed in the baseline model and enhances its ability to accurately detect floating litter on the water surface.



(a) Original image (b) Heat map of RTDETR (c) Heat map of RTDETR-MARD

Figure 12. Heat map analysis results of the model on the FloW dataset.

Our RTDETR-MARD model demonstrates exceptional performance across various environmental conditions. We selected three representative water surface environments for testing, encompassing strong sunlight, fog, and complex water surface conditions such as surface fluctuations, numerous floating leaves, and intricate reflections, as depicted in Figure 13. The performance of the RTDETR-MARD model in these scenarios not only confirms its exceptional adaptability but also underscores its robustness in complex environments. This demonstrates the potential of our RTDETR-MARD model for real-world applications, particularly in situations where precise detection is crucial under varying natural conditions.



(a) Water surface under strong sunlight (b) Water surface in heavy fog (c) Complex water suface conditions

Figure 13. Test results in different water surface environments.

While the overall results validate the model's strong generalization capabilities, we further investigated its limitations through failure case analysis. As shown in Figure 14,

the left two images depict false negatives, where floating debris was not successfully detected. These failures are primarily caused by visual interference from environmental elements such as the reflection of large structures (e.g., water wheels), dark-colored plastic bottles blending into the background, and low-contrast conditions like overcast lighting. The right two images illustrate false positives, where the model mistakenly identifies background reflections—such as those of trees and bridges—as floating objects due to their visual similarity in shape and texture. These observations reveal that although RTDETR-MARD demonstrates high robustness, its performance can still be compromised under certain visual disturbances. Future improvements may include integrating boundary-aware attention mechanisms and multi-frame temporal information to mitigate such challenges.



(a) Missed detections (b) False detections

Figure 14. Failure cases of RTDETR-MARD in complex water surface environments. (**a**) Missed detections: some floating debris was not detected due to occlusion, low contrast, or background blending. (**b**) False detections: background elements such as reflections of trees and bridges were incorrectly identified as floating objects.

4.4.6. Generalization Experiment

To comprehensively evaluate the effectiveness of our model, we constructed a custom water surface waste dataset, which includes diverse samples collected from various real-world environments. As shown in Table 6, the experimental results on this dataset clearly indicate that our proposed RTDETR-MARD achieves the highest mAP50 (49.2%) among all the compared methods. Despite having a moderate model size of 22.5 million parameters and a computational load of 63.2 GFLOPs, RTDETR-MARD outperforms the other models in detection accuracy. This performance proves that our model strikes an excellent balance between detection performance and computational efficiency. Furthermore, the high mAP50 achieved on both the FloW dataset and our private dataset demonstrates the outstanding generalizability and robustness of RTDETR-MARD under diverse environmental conditions.

Table 6. Experimental results on the self-constructed dataset.

Method	P (%)	R (%)	mAP50 (%)	Params (M)	FLOPs (G)
RTDETR-MARD	61.0	48.9	49.2	22.5	63.2
YOLOv8s	59.9	43.3	47.7	11.1	28.5
YOLOv5s	58.8	46.1	47.3	9.1	24.1
YOLOv8m	59.4	47.3	47.3	25.8	78.7
YOLOv8n	48.1	45.7	46.8	3.0	8.1
YOLOv5n	51.9	47.0	46.8	11.1	28.5
RTDETR-R18	52.3	48.8	46.3	19.9	57.0
YOLOv5m	52.0	48.2	46.2	25.1	64.4
YOLOv11m	57.9	42.9	45.8	20.1	68.2
YOLOv10m	56.4	41.0	44.0	16.5	63.5

5. Conclusions

This paper presents a novel approach for floating waste detection, named RTDETR-MARD, which significantly improves upon the RT-DETR model. We thoroughly analyze the challenges of floating waste detection and enhance the model by exploiting RT-DETR's strengths. By incorporating the GD mechanism, the model achieves improved multi-scale feature aggregation and cross-layer interaction. We also propose a new feature fusion module that integrates FAM, IFM, IIM, and S^2F^2M to ensure better spatial and semantic consistency across scales. Furthermore, the WIoU loss function enhances localization by emphasizing high-quality anchor supervision. Our approach delivers improved small object detection accuracy and retains rapid inference capability, offering practical value for real-time use in dynamic water scenarios. Extensive experiments validate the model's superior performance and robustness. Heat map visualizations further highlight its enhanced feature extraction capability.

Although the evaluation leverages both public and private datasets, this study is limited to offline testing and has not yet been deployed on fully autonomous platforms such as USVs or low-power edge devices, which are essential for real-world environmental monitoring applications. In future work, we aim to address these limitations by integrating RTDETR-MARD into embedded systems to enable real-time detection, onboard inference, and active intervention. Moreover, the proposed framework holds promise for broader applications in smart environmental governance, sustainable water management, and AI-assisted ecological restoration.

Author Contributions: Conceptualization, B.S., H.T. and L.G.; methodology, H.T. and J.W.; software, K.B. and J.W.; validation, B.S., H.T. and L.G.; formal analysis, H.T. and J.W.; investigation, B.S. and K.B.; resources, L.G.; data curation, J.W. and K.B.; writing—original draft preparation, B.S. and H.T.; writing—review and editing, L.G. and J.W.; visualization, K.B.; supervision, L.G.; project administration, L.G.; funding acquisition, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by Natural Science Foundation of China Grants (61972456 and 61173032) and by the Tianjin Natural Science Foundation (20JCYBJC00140).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We would like to thank the School of Computer Science and Technology, Tiangong University, for supporting our work.

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

References

1. Zhang, M.; Zheng, X.; Wang, J.; Pan, Z.; Che, W.; Wang, H. Trajectory Planning for Cooperative Double Unmanned Surface Vehicles Connected with a Floating Rope for Floating Garbage Cleaning. *J. Mar. Sci. Eng.* **2024**, *12*, 739. [CrossRef]
2. Chen, X.; Zhang, P.; Lu, J.; Chen, Y.; Zhang, J. Hydrology Modulates the Microplastics Composition and Transport Flux Across the River–Sea Interface in Zhanjiang Bay, China. *J. Mar. Sci. Eng.* **2025**, *13*, 428. [CrossRef]
3. Haward, M. Plastic Pollution of the World's Seas and Oceans as a Contemporary Challenge in Ocean Governance. *Nat. Commun.* **2018**, *9*, 667. [CrossRef] [PubMed]
4. Kong, S.; Tian, M.; Qiu, C.; Wu, Z.; Yu, J. IWSCR: An Intelligent Water Surface Cleaner Robot for Collecting Floating Garbage. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**, *51*, 6358–6368. [CrossRef]
5. Cheng, Y.; Zhu, J.; Jiang, M.; Fu, J.; Pang, C.; Wang, P.; Sankaran, K.; Onabola, O.; Liu, Y.; Liu, D.; et al. FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10953–10962.
6. Zhou, Z.; Sun, J.; Yu, J.; Liu, K.; Duan, J.; Chen, L.; Chen, C.P. An Image-Based Benchmark Dataset and a Novel Object Detector for Water Surface Object Detection. *Front. Neurorobot.* **2021**, *15*, 723336. [CrossRef]
7. Fulton, M.S.; Hong, J.; Sattar, J. Trash-ICRA19: A Bounding Box Labeled Dataset of Underwater Trash. The Data Repository for the University of Minnesota (DRUM). 2020. Available online: https://conservancy.umn.edu/items/c34b2945-4052-48fa-b7e7-ce0fba2fe649 (accessed on 1 March 2025).
8. Panwar, H.; Gupta, P.K.; Siddiqui, M.K.; Morales-Menendez, R.; Bhardwaj, P.; Sharma, S.; Sarker, I.H. AquaVision: Automating the Detection of Waste in Water Bodies Using Deep Transfer Learning. *Case Stud. Chem. Environ. Eng.* **2020**, *2*, 100026. [CrossRef]
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
10. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
11. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
12. Khokhar, S.; Kedia, D. Integrating YOLOv8 and CSPBottleneck Based CNN for Enhanced License Plate Character Recognition. *J. Real-Time Image Process.* **2024**, *21*, 168. [CrossRef]
13. Luo, S.; Dong, C.; Dong, G.; Chen, R.; Zheng, B.; Xiang, M.; Zhang, P.; Li, Z. YOLO-DAFS: A Composite-Enhanced Underwater Object Detection Algorithm. *J. Mar. Sci. Eng.* **2025**, *13*, 947. [CrossRef]
14. Zhu, J.; Li, H.; Liu, M.; Zhai, G.; Bian, S.; Peng, Y.; Liu, L. Underwater Side-Scan Sonar Target Detection: An Enhanced YOLOv11 Framework Integrating Attention Mechanisms and a Bi-Directional Feature Pyramid Network. *J. Mar. Sci. Eng.* **2025**, *13*, 926. [CrossRef]
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
17. Yang, M.; Wang, H. Real-Time Water Surface Target Detection Based on Improved YOLOv7 for Chengdu Sand River. *J. Real-Time Image Process.* **2024**, *21*, 127. [CrossRef]
18. Ma, L.; Wu, B.; Deng, J.; Lian, J. Small-Target Water-Floating Garbage Detection and Recognition Based on UNet-YOLOv5s. In Proceedings of the 2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China, 14–16 April 2023; pp. 391–395.
19. Chen, L.; Zhu, J. Water Surface Garbage Detection Based on Lightweight YOLOv5. *Sci. Rep.* **2024**, *14*, 6133. [CrossRef] [PubMed]
20. Wang, S.; Han, J. Automated Detection of Exterior Cladding Material in Urban Area from Street View Images Using Deep Learning. *J. Build. Eng.* **2024**, *96*, 110466. [CrossRef]
21. Luo, W.; Han, W.; Fu, P.; Wang, H.; Zhao, Y.; Liu, K.; Wei, G. A Water Surface Contaminants Monitoring Method Based on Airborne Depth Reasoning. *Processes* **2022**, *10*, 131. [CrossRef]
22. Chang, H.C.; Hsu, Y.L.; Hung, S.S.; Ou, G.R.; Wu, J.R.; Hsu, C. Autonomous Water Quality Monitoring and Water Surface Cleaning for Unmanned Surface Vehicle. *Sensors* **2021**, *21*, 1102. [CrossRef]
23. Wang, S.; Kim, M.; Hae, H.; Cao, M.; Kim, J. The Development of a Rebar-Counting Model for Reinforced Concrete Columns: Using an Unmanned Aerial Vehicle and Deep-Learning Approach. *J. Constr. Eng. Manag.* **2023**, *149*, 13. [CrossRef]
24. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

26. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
27. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Shum, H.Y. DINO: DETR with Improved Denoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605.
28. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhang, L. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. *arXiv* **2022**, arXiv:2201.12329.
29. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
30. Wang, S. Evaluation of Impact of Image Augmentation Techniques on Two Tasks: Window Detection and Window States Detection. *Results Eng.* **2024**, *24*, 103571. [CrossRef]
31. Hwang, D.; Kim, J.-J.; Moon, S.; Wang, S. Image Augmentation Approaches for Building Dimension Estimation in Street View Images Using Object Detection and Instance Segmentation Based on Deep Learning. *Appl. Sci.* **2025**, *15*, 2525. [CrossRef]
32. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; pp. 51094–51112.
33. Kang, M.; Ting, C.M.; Ting, F.F.; Phan, R.C.W. ASF-YOLO: A Novel YOLO Model with Attentional Scale Sequence Fusion for Cell Instance Segmentation. *Image Vis. Comput.* **2024**, *147*, 105057. [CrossRef]
34. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
35. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 658–666.
36. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
37. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
38. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

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.