Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Real-Time Water Surface Garbage Detection with Bi-directional Re-calibration and Structural Re-parameterization

Yehui Zheng
Guangdong University of Science and Technology
Jiaming Gu
Guangdong University of Science and Technology
Shu Gong
gongshu@gdust.edu.cn
Guangdong University of Science and Technology
Jiao Peng
Guangdong University of Science and Technology
Caimei Liu
Dongguan University of Technology
Sun Yu
Guangdong University of Science and Technology

Article

Keywords: Water surface debris detection, RT-DETR, BRM, Full-scale feature pyramid, Structural reparameterization, Small-object detection

Posted Date: April 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9438176/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Real-Time Water Surface Garbage Detection with Bi-directional Re-calibration and Structural Re-parameterization

Yehui Zheng [1], Jiaming Gu [1], Shu Gong [1]*, Jiao Peng[1], Caimei Liu [2], Sun Yu [1]

[1] Department of Computer Science, Guangdong University of Science and Technology, Dongguan, Guangdong 523000, China.

[2] Department of Computer Science, Dongguan University of Technology, Dongguan, Guangdong 523000, China.

*email: gongshu@gdust.edu.cn

Abstract

Floating debris detection is challenging for automated monitoring systems due to waves, lighting and the low textures of microscopic targets on resource-limited edge devices. To overcome this challenge, we propose BR-Trans, an end-to-end real-time detection framework for unmanned aquatic management. First, in order to prevent small debris from being obscured by ripples from aerial view, we introduce StarNet lightweight backbone based implicit mapping to project blurred textures to a robust high-dimensional space. Second, we reconstruct a Re-Calibration Feature Pyramid Network with parameter-free multi-scale alignment via SimFusion module. We design BR-Trans core with a dual stream mechanism: Deep Layer semantics suppress water reflections and high-frequency noise, and shallow Layer geometric details correct boundary localization of semi-submerged objects. Extensive experiments show that BR-Trans achieves 92.54% mAP@0.5 and 92.82% recall with only 71.7

million parameters, far surpassing YOLOv11 and RT-DETR-r50. Operating at ultra-real-time speed of 75.41 FPS, BR-Trans provides a highly efficient and flexible technology for autonomous aquatic ecology, pollution tracking and public health protection.

Keywords: Water surface debris detection; RT-DETR; BRM; Full-scale feature pyramid; Structural reparameterization; Small-object detection

Introduction

With urbanization and human activity increasing, huge amounts of waste (plastic bottles and foam) enter aquatic environments through littering by humans and surface runoff. According to the UN Environment Programme, between 75 and 199 million tons of plastic waste enter the oceans annually, which accounts for 85% of marine debris by weight[1]. This pollution degrades aquatic environments, destroys habitats, and severely threatens human health via the food chain[2] . Traditional surface debris detection relies on manual patrols, which lack efficiency, cost-effectiveness, and broad coverage. Recently, the integration of Unmanned Aerial Vehicle (UAV) remote sensing and deep learning has emerged as a powerful tool for large-scale, routine aquatic environmental monitoring[3].

However, due to the unstructured and highly dynamic nature of water surface environments—characterized by drastic changes in lighting, severe glare, and complex backgrounds—existing object detection algorithms often fail to perform optimally. Over the years, detection architectures have evolved significantly. Single-stage detectors like the YOLO series provide fast detection speeds suitable for real-time monitoring, with recent studies improving YOLOv11 and YOLOv5 to handle lighting interference in river channels[4] . Concurrently, the DETR series, particularly RT-DETR, has provided an end-to-end solution driven by efficient hybrid encoders [5].While recent benchmark tests have advanced UAV visual detection[6,7] and DETR architectures are rapidly evolving in matching mechanisms[8,9] , the application

of existing models in complex aquatic environments remains in its infancy. Furthermore, public datasets like TACO[10] and TrashCan[11] focus mainly on terrestrial scenes, lacking diverse aquatic samples.

Specifically, mainstream algorithms still face four severe technical bottlenecks when deployed in complex aquatic scenarios. First, from a drone's aerial perspective, floating debris accounts for an extremely low proportion of pixels with blurred boundaries. Traditional low-dimensional convolutional networks are highly prone to "feature vanishing" when limited semantic information is intertwined with high-frequency background noise[12,13] . Second, there is a "semantic disconnection" during cross-scale feature fusion. Existing models lack parameter-free global alignment mechanisms, resulting in severe information decay during feature pyramid propagation[14,15] . Third, existing models lack a dynamic attention recalibration mechanism tailored for water surfaces, making them highly prone to false-positive background detections caused by dynamic water ripples and strong reflections[16,17]. Finally, high-precision models incur excessive computational overhead, making it difficult to deploy them on resource-constrained embedded chips, whereas conventional lightweight models often sacrifice significant detection accuracy[18,19].

To address these problemful bottlenecks, we propose BR-Trans, a complete real time water surface garbage detection system based on a bidirectional recalibration method. We aim for high precision small target capture, high anti-interference robustness and high frame rate edge inference.Specifically, we redesign the RT-DETR architecture, and make four contributions:

First, we design feature-enriched backbone network using high-dimensional implicit mapping. To solve the feature vanishing problem of small debris features which are easily obscured by water waves, we introduce a lightweight backbone using StarNet. By integrating low-dimensional inputs into high-dimensional nonlinear

feature space using Star Operation, we improve our model's ability to capture hidden texture details of microscopic floating objects.

Second, we reconstruct a full-scale feature pyramid and introduce SimFusion global alignment. To reduce information decay of cross-scale fusion pyramids, we design pyramids P2-5. With adaptive global parameter-free alignment, semantic barrier between layers is broken down. Multi-scale features ranging from small fragments to large-area debris are covered.

Third, we propose the Bidirectional Re-calibration Module (BRM) to deal with dynamic backgrounds. We propose a complementary dual stream mechanism of semantic injection and boundary sharpening. Deep-layer semantic attention is used to reduce high-frequency noise like water ripples (reducing false positives) and shallow-layer spatial details are used to correct blurred boundary lines of semi-submerged targets (reducing false negatives).

Finally, ultra light weight deployment is achieved using structural reparameterization. By RepC3 and IFM modules we decouple multibranch enhancement during training from single path acceleration during inference, so that BR-Trans has only 71.7 million parameter count with ultra-real time inference speed 75.41 FPS, which is perfect for engineering deployment on unmanned vessels and river patrol drones.

Results

Ablation Study

The StarNet backbone, SimFusion feature fusion module and BRM improvement module were used to perform ablation experiments using RT-DETR-r18 as the baseline. Sequential strategies were applied and the results were summarized in **Table 1**. The original ResNet units were replaced by StarBlock. The mAP@0.5 increased from 91.97% to 92.08% with precision increase 3.43%. SimFusion module for P2-P5

multi scale architecture increased the mAP@0.5 to 92.42%. SimFusion parameter free design reduced the model parameters from 77.8M to 71.8M (7% decrease). BRM module increased the mAP@0.5 to 92.54%. Recall rate increased from 91.63% to 92.82% and F1 score reached maximum 93.08. The detailed classification performance for each debris category is further illustrated via the Precision-Recall (P-R) curves in **Supplementary Note 2 and Supplementary Fig. S1**

Comparative Evaluation

Fully improved BR-Trans model was tested against 6 standard object detection models on YRDG[20]. In **Table 2**, BR-Trans got mAP@0.5 (92.54%), which represented improvement of 5.63% and 11.45% over YOLOv11 and YOLO26 respectively. It beat the original RT-DETR-r18 (0.57%) and the parameter intensive RT-DETR-r50 (1.64%). For the stricter mAP@0.5:0.95, BR-Trans was 49.22%, highest among all the compared models. On the model complexity, BR-Trans used 71.7 million parameters (a 56.3% reduction on RT-DETR-r50), while maintaining 75.41 FPS inference rate above real time detection (36 FPS).

Comprehensive Performance Analysis

A performance radar graph **(Figure 1)** was plotted for mAP@0.5, mAP@0.5:0.95, recall, F1-score and GFLOPs. The polygon of BR-Trans model covers YOLOv8, YOLO11 and the baseline model along the accuracy-related axes, while on Recall, BR-Trans scored the highest (92.82%). In terms of cost, BR-Trans maintains high accuracy coverage and only 47% of the computation cost of RT-DETR-r50 is computation cost.

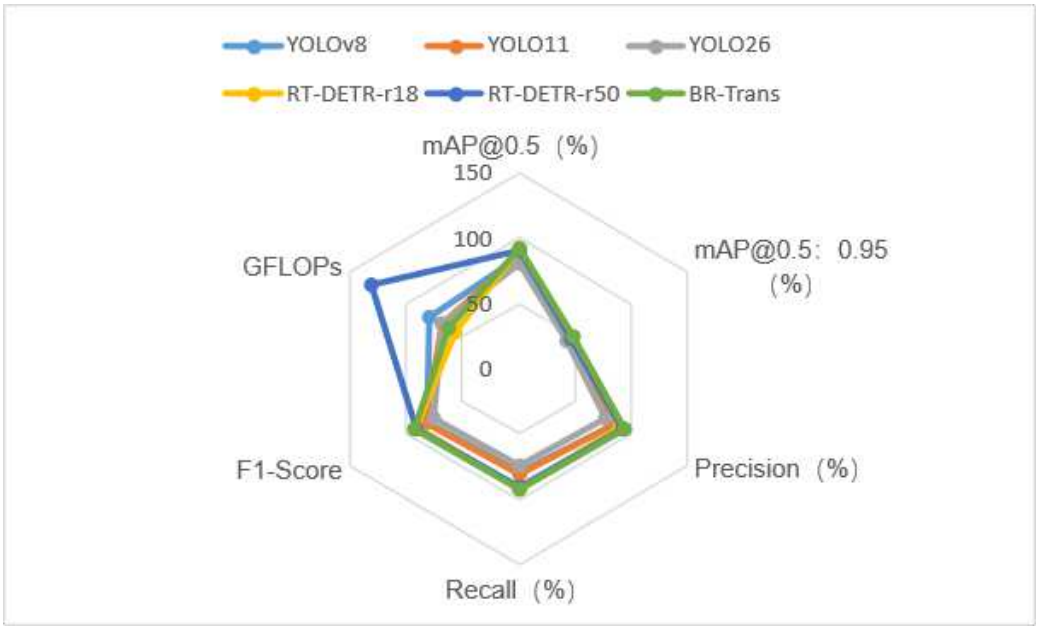


Figure 1

Training Stability and Convergence

Losses and accuracy were measured throughout training and validation **(Figure 2)**. Classification loss (L_{cls}) , bounding box regression loss (L_{L1}) and generalized intersection-over-union loss (L_{GIoU}) were significantly lower early. Around 100 epochs, the training loss and validation loss curves converged to low range without the validation loss showing a reverse increase. The mAP@0.5 and mAP@0.5:0.95 metrics reached steady levels around 150 epochs, and the learning behavior was

stable.

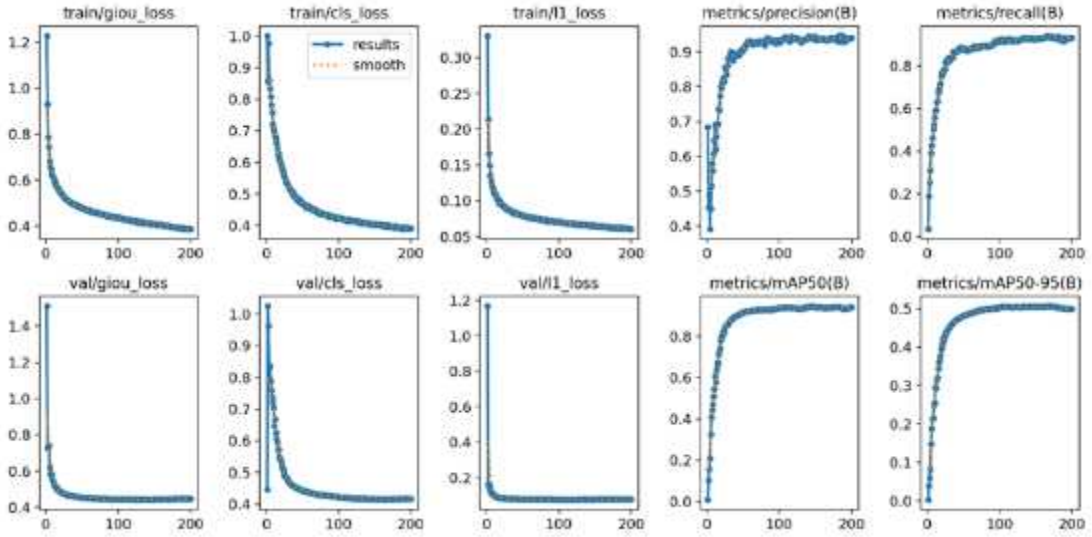


Figure 2

Qualitative Evaluation

We also generated a class activation heatmap **(Figure 3)** to compare the concentration on target features. RT-DETR showed false positive activation with incorrect labels for non-target background areas. YOLO26 and APM-YOLOv7 showed diffuse heatmap distribution with slight localization deviation for overlapped targets. High activation regions of BR-Trans focused strictly on floating debris targets, suppressing activation from complex shorelines and precise bounding box localization.

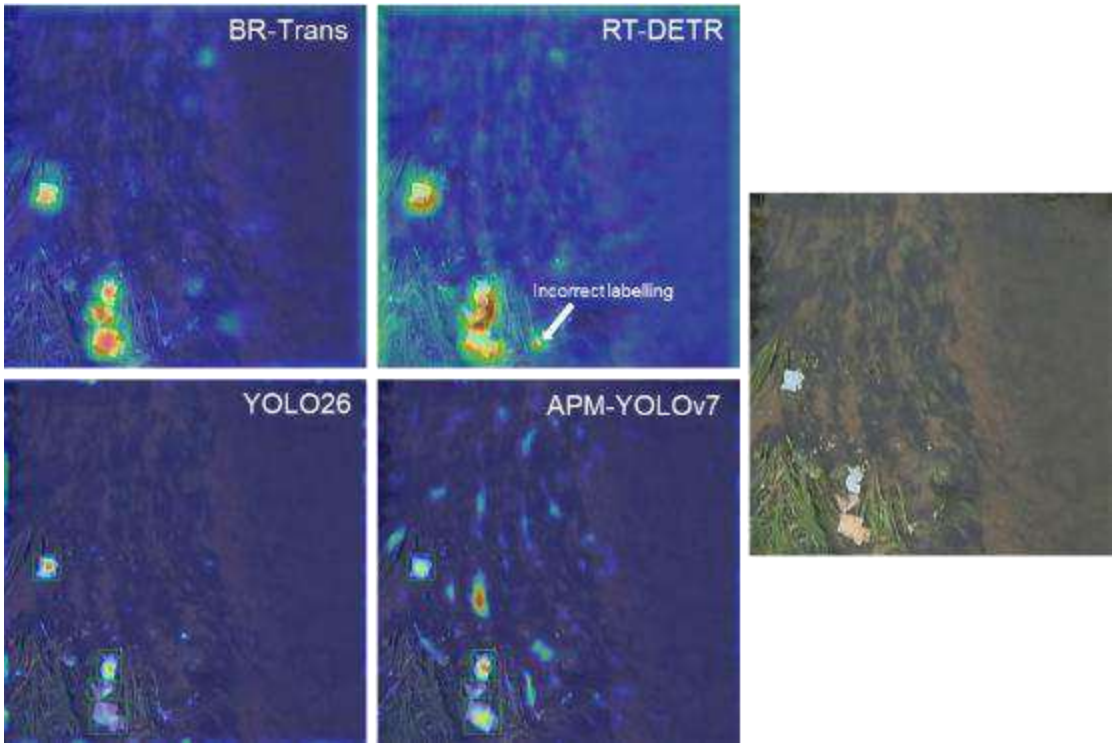


Figure 3

Discussion

The comprehensive evaluation of BR-Trans shows a successful balance among high-precision detection, a lightweight architecture, and real-time inference speed for water surface debris monitoring. By systematically restructuring the RT-DETR architecture, the model attained a peak mAP@0.5 of 92.54% and an inference speed of 75.41 FPS while using only 71.7 million parameters. These results indicate that the proposed mechanisms effectively reconcile the computational constraints of edge devices with the requirement for robust feature extraction in unstructured, dynamic aquatic environments.

The improvements reported in the ablation and comparision work are the result of our targeted architecture. Low-dimensional convolution networks often lose feature when sparse debris textures are merged by water ripples. StarBlock backbone using high-dimensional implicit mapping enhanced the capture of fine-grained texture features. SimFusion multi-scale architecture reduced semantic disconnection at cross-

scale feature fusion, with parameter-free global information aggregation reduced the number of parameter by 7.7%, and enhanced multi-scale feature interaction between target sizes.

The biggest improvement was the recall rate (92.82%) after BRM. In complex aquatic scenes, semisubmerged objects and strong surface reflections often give false negatives and false positives. BRM corrects these errors with a dual stream complement: deep semantic features suppress high frequency ripple noise and reduce false negatives and shallow geometric details refine localization of small objects and reduce false negatives. This advantage is supported by class activation heatmaps. Whereas conventional models such as RT-DETR or APM-YOLOv7 show diffuse background activations or misclassifies reflective waves as debris, tightly concentrated activation bounds around targets are retained.

Despite these advances, physical limitations persist in real-world applications. SimFusion improves global perception, but segmenting individual instances is difficult for dense accumulations (such as rivers after heavy rain occluded by debris), targets with extreme biomimetic camouflage or semi-transparent Objects remain difficult to detect. Future work will explore multi-modal data fusion (e.g., polarization imaging) to improve contour detection for transparent objects in reflective waters.

Finally, ultra-lightweight design and high frame rate performance of BR-Trans provide broad prospects for automated aquatic governance. Operating efficiently on limited embedded platforms via RepC3 structural reparameterization, the model is ready for deployment. It can serve as the “edge brain" of unmanned cleaning vessels executing automated patrol-and-clean operations using PID control algorithms. Furthermore, its ability to detect tiny targets makes it ideal for use on river patrol drones to establish air-ground collaborative governance system. In long-term deployments, BR-Trans can serve as reliable data probe for aquatic ecological big data, tracking debris flow and generating pollution heatmaps which provide precise

quantitative support for tracing pollution sources and watershed management policies.

Beyond the quantitative achievements, the deployment of BR-Trans carries profound implications for broader environmental and public health sectors. By effectively bridging the gap between high-precision deep learning and the limited computational capacity of edge devices, this model enables fully autonomous, large-scale deployment on unmanned surface vessels (USVs) and patrol drones. This 'air-ground collaborative' automated governance system significantly reduces the human and economic costs associated with traditional water quality monitoring. More importantly, the early and accurate detection of microscopic debris at the source prevents the long-term degradation of macro-plastics into microplastics, thereby protecting drinking water security and mitigating toxicological threats to the aquatic food chain and human health. Ultimately, BR-Trans serves not merely as an algorithmic improvement, but as a scalable technological probe for aquatic ecological big data, providing policymakers with precise, real-time quantitative evidence for global watershed management and sustainable environmental conservation.

Methods

Dataset and Environmental Context

All experiments were performed with Yellow River Drifting Garbage Dataset (YRDG). Unlike generic Earth object detection data sets such as COCO or TACO, the YRDG data set was collected and analyzed for the typical high unstructured and challenging Yellow River environment characterized by rapid and unpredictable river currents, high water turbidity resulting in low underwater visibility, complex and cluttered riverbank backgrounds and dynamic visual interference caused by high light (glare) and waves ripples.

Three types of water bodies are represented: major rivers, main yellow river stem and water bodies near riverside residential areas. Data collected for weather (sun, cloudy, rain) and time of day. Complex shapes and distribution of floating debris under different water and light conditions[21]. This data sets allow trained models to apply to real-world automated surface

debris inspection in rivers.3807 high resolution annotated aerial images (pixel resolution) are chosen randomly to train and test the model. The images are randomized into training set, validation set, and test set in 8:1:1. 84% of the bounding boxes contain “micro” small targets (usually smaller **than** 32×32 **pixels**) corresponding to the main detection challenge for operational unmanned river patrols, where detection early detection of small floating objects is critical for safe vessel route. The classification scheme contains 7 categories representing the most common types of inland waterway debris: plastic, paper, glass, metal, textiles/fibers, natural debris (e.g. dead branches) and other (irregulative or highly degraded fragments). This general set allows model generalisations of targets with extreme intra-class variation in shape, texture and reflectance.

Baseline Model and Problem Formulation

We selected Real-Time Detection TRansformer (RT-DETR) as the baseline algorithm for our study[22,23] . While traditional CNN single-stage detectors (like the YOLO series) rely heavily on anchor-box heuristics and NMS for post processing (which introduces additional inference latency and hyperparameter tuning problems in dense object cases)RT-DETR eliminates these bottlenecks. RT-DETR retains the global contextual perception advantages of the Transformer architecture while directly solves the $O(L^2)$ computational complexity problem (where L is the sequence length) typical of traditional Transformer encoders with a simple hybrid CNN-Transformer encoder design[24].

RT-DETR employs Bipartite Matching to achieve NMS-free, end-to-end inference. Mathematically, let the ground-truth set of objects in an image be denoted as $y = \{y_i\}_{i=1}^{N}$, and the model’s predicted set of bounding boxes and class probabilities be denoted as $\hat{y} = \{\hat{y}_i\}_{i=1}^{M}$, where $M \geq N$. The model seeks to find an optimal bipartite matching between the predictions and the ground truth by searching for a permutation of N elements $\sigma \in \mathfrak{S}_N$ that minimizes the global matching cost:

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i=1}^{N} \mathcal{L}_{match}\left(y_i, \hat{y}_{\sigma(i)}\right)$$

where $\mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)})$ is a pair-wise matching cost that accounts for both the class

prediction accuracy and the bounding box localization similarity. The Hungarian Algorithm is utilized to efficiently solve this assignment problem and directly output the optimal prediction sequence.

Unlike standard DETR models that rely on static, content-independent object queries, RT-DETR introduces an Intersection-over-Union (IoU)-aware query selection mechanism. This mechanism explicitly selects the top-K features with the highest confidence scores directly from the encoder's output feature maps to serve as the initial object queries. By explicitly constraining these query features to align with the actual spatial distribution of real objects in the scene, the convergence speed of the Transformer decoder is drastically accelerated[25]. However, despite these architectural advantages, applying the original RT-DETR directly to water surface debris detection yields suboptimal results due to extreme target size variations, semantic disconnection in feature pyramids, and dynamic background noise. Consequently, we completely restructured its backbone, multi-scale feature fusion, and feature extraction stages to formulate the BR-Trans architecture.

Overall Architecture of BR-Trans

The proposed BR-Trans framework abandons the Bi-FPN (Bidirectional Feature Pyramid Network) architecture used in the original RT-DETR and instead adopts a customized real-time detection framework based on full-scale recalibration (Fig. 4). The pipeline is designed to balance high precision with edge-device inference speed and comprises three core, interconnected components:

1. A lightweight StarNet-based backbone: Designed for enhanced high-dimensional implicit feature mapping to capture the faint textures of microscopic floating objects.
2. A Full-scale Re-Calibration FPN utilizing a SimFusion module: Engineered for the parameter-free global spatial alignment and aggregation of multi-scale features, eliminating the semantic gaps between hierarchical layers.
3. A feature distribution and recalibration stage: Incorporating a novel Bidirectional Re-calibration Module (BRM) and a Structural Reparameterization Module (RepC3) to

enable mutual "semantic injection" and "boundary sharpening" without increasing the computational graph's depth during edge inference.

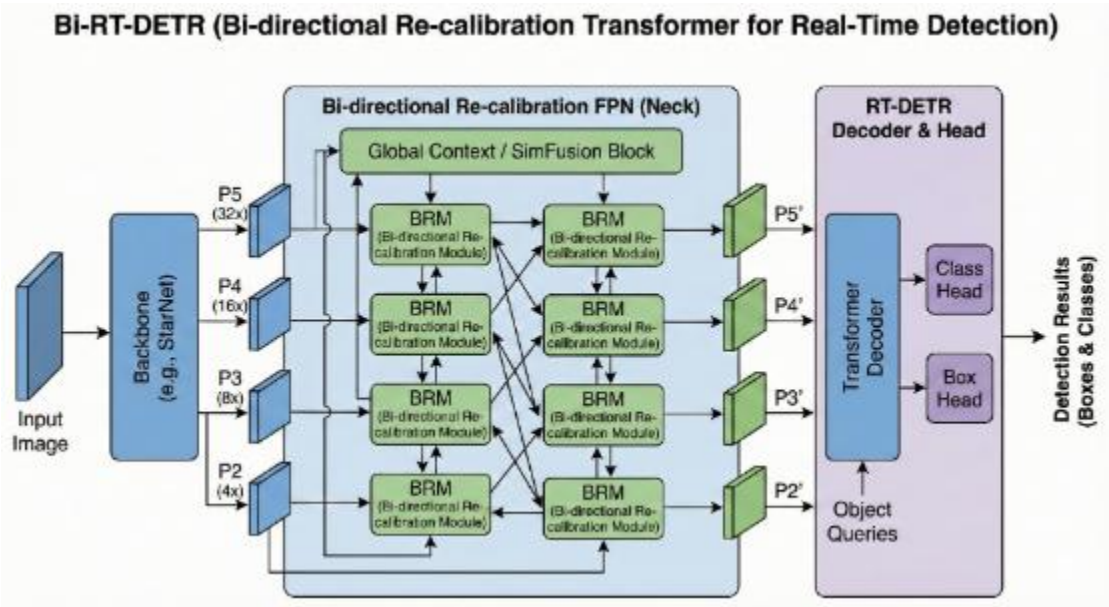


Figure 4

Feature-Enhanced Backbone Network via Implicit Mapping

From a drone's aerial perspective, small floating debris often occupies an extremely low proportion of pixels, resulting in blurred boundaries and sparse textural features. When these faint features are intertwined with the high-frequency background noise of dynamic water ripples, traditional low-dimensional convolutional networks (such as ResNet) are highly susceptible to the "feature vanishing" phenomenon. To address this critical challenge, we replaced the standard ResNet BasicBlocks in the original RT-DETR backbone with StarBlocks[26] **(Figure 5b)**.

Unlike traditional convolutional networks which rely on stacking layers of lower size or increasing channel width to increase representational capacity – and which result in a huge computational overhead – the StarBlock consists of simple element-wise multiplication to map low-dimensional input features to high-dimensional nonlinear latent features. If the input feature tensor to a StarBlock is $X \in \mathbb{R}^{H \times W \times C}$ then the internal feature fusion procedure is defined as:

$$\boldsymbol{S(X) = (W_1X + B_1) \odot (W_2X + B_2)}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication). $W_1, W_2 \in \mathbb{R}^{C\times(C/r)}$ and $B_1, B_2 \in \mathbb{R}^{C/r}$ represent the weight matrices and bias vectors of two independent linear projection branches, respectively, with $\boldsymbol{r}$ being the channel reduction ratio.

This operation is mathematically analogous to the polynomial kernel function widely used in Support Vector Machines (SVMs)[27]. Theoretical analysis demonstrates that a single-layer Star operation can implicitly rewrite the feature space into a higher-dimensional space containing approximately $(d/\sqrt{2})^2$ linearly independent dimensions, where d is the number of channels. By recursively stacking L layers of StarBlocks, the network amplifies the feature dimensions exponentially, approximating a nearly infinite-dimensional implicit feature space:

$$\boldsymbol{\mathcal{H} \approx \left(\frac{d}{\sqrt{2}}\right)^{2^L}}$$

This high-dimensional nonlinear mapping introduces a very strong induction bias into the network and improves the model's ability to detect very small textural differences between the contours of small debris fragments and the highly reflective water surface background. Importantly, high frequency details are retained at scales without significantly increasing FLOPs[28].

Bidirectional Re-calibration Module (BRM)

StarBlock extracts rich high-dimensional details, but these features can be semantically misaligned when propagated through the feature pyramid. Shallow-level features retain incredibly rich geometry and spatial resolution, but lack semantic discrimination power needed to distinguish debris from water ripples. Deep-level features (e.g.,) are semantically abstract but spatial localization is ambiguous due to repeated downsampling. To bridge this gap, the BRM is introduced as the central gating unit for cross-scale feature interaction **(Figure 5a)**.

The BRM constructs a bidirectional complementary recalibration mechanism using an asymmetric attention generation operator. For the low-level spatial detail stream ($F_{low} \in \mathbb{R}^{C' \times H \times W}$), we introduce a 3×3 Depthwise Convolution (DWConv)[29] followed by Batch Normalization (BN) and a 1×1 pointwise convolution to construct a highly sensitive spatial attention mask

M_{spat}. This mask accurately captures the geometric contours of small objects while suppressing flat background regions:

$$M_{spat} = \sigma(\text{Conv}_{1\times1}(\text{BN}(\text{DWConv}_{3\times3}(F_{low}))))$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, constraining the mask values to the range $(0, 1)$.Simultaneously, for the high-level semantic stream (F_{high}), we introduce Global Average Pooling (GAP) to aggregate the global contextual information of the water surface. This is followed by a Multi-Layer Perceptron (MLP) consisting of two 1×1 convolutions separated by a ReLU activation (forming an SE-like block) [30]to construct the channel attention weights M_{chan}. This mask identifies which semantic channels are critical for suppressing the periodic background noise generated by water ripples:

$$M_{chan} = \sigma\left(\text{MLP}\left(\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{high}(i,j)\right)\right)$$

Based on these meticulously generated attention masks, we reformulated the standard feature fusion formula, replacing simple addition or concatenation with a cross-residual calibration scheme:

$$\tilde{F}_{low} = F_{low} + (F_{low} \otimes \text{Upsample}(M_{chan}))\text{(Semantic Injection)}$$

$$\tilde{F}_{high} = F_{high} + \left(F_{high} \otimes \text{Downsample}(M_{spat})\right)\text{(Boundary Injection)}$$

Using residual connections ($+F$), BRM provides direct and stable gradient propagation during backpropagation. Tensor multiplication ($\otimes$) is an adaptive perturbation term. Deep-layer semantic knowledge is used to "purify" shallow layers, filtering out high-frequency ripple noise to reduce false positives. Deep-layer spatial knowledge is used to "sharpen" deep layers,

correcting blurred localization bounds of semi-submerged targets.

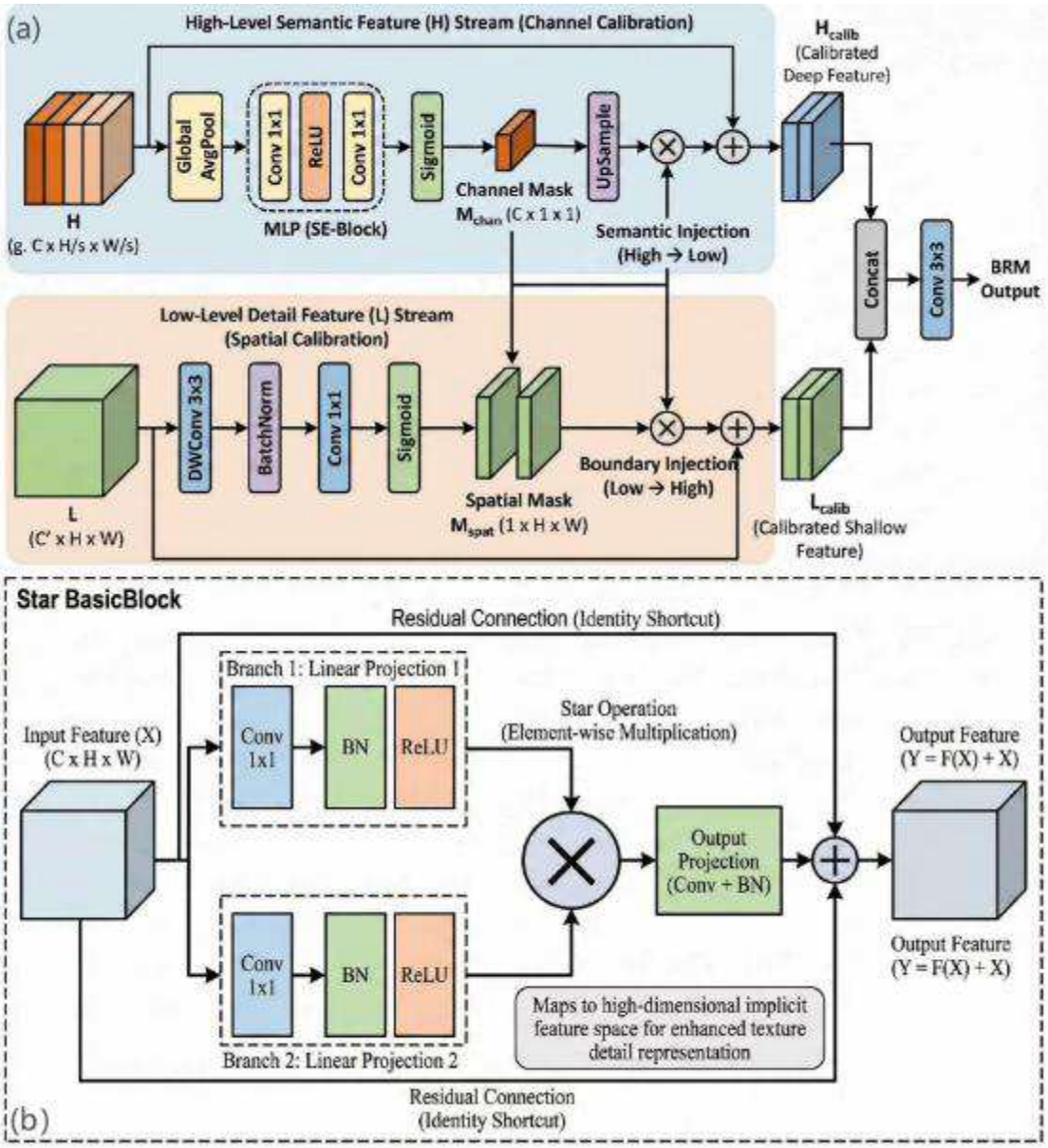


Figure 5

Full-Scale Feature Pyramid Construction and SimFusion

To compensate for the physical limitations of traditional RT-DETR for small floating objects we rebuilt the topology of the feature pyramid. In typical FPNs the lowest feature map level is P_3 (downsampled by a factor of 8) , hence floating objects of less than 8×8 pixels lose almost all spatial information before entering the feature fusion network, and will thus be

missed. To avoid this we explicitly extended the P_2 branch (downsampled by a factor of 4). Let the hierarchical feature set output by the backbone network be $P = \{P_2, P_3, P_4, P_5\}$, where each feature map P_i has a spatial resolution of $\frac{H}{2^i} \times \frac{W}{2^i}$. The P_2 layer allows lossless geometric edges for very small objects.

However, propagating features at such a wide range of scales (from $P_2$to P_5) tends to increase information attenuation and semantic misalignment. To eliminate semantic problems, we took inspiration from Gather-and-Distribute[31] for the SimFusion module **(Figure 6a)** . SimFusion is a global information bridge that performs parameter-free spatial alignment of all multi-scale features. It uses adaptive average pooling (for downsampling higher-resolution features like P_2) and bilinear interpolation (for upsampling lower-resolution features like P_5) to mathematically unify all features from different levels to a target resolution R (which leads to a dense, globally gathered context tensor F_{gather}:

$$F_{gather} = \text{Concat}\left(\bigcup_{i=2}^{5} \mathcal{T}(P_i, R)\right)$$

where $\mathcal{T}(\cdot, R)$ is the deterministic spatial alignment operator.Then deep fusion on F_{gather}. is performed. Internally, IFM creates a global information flow by introducing RepBlocks to generate global information flow. By establishing global information flow at top of the network and propagating it down layer by layer through BRM modules even the lowest level microscopic features of P_2 can perceive global water body context. This approach minimizes conditional probability of false positive background detections $P(\text{False} \mid \text{Background})$. Calculating P_2 layer continuously throughout pyramid increases theoretical computational overhead, which is compensated by SimFusion's channel compression and structural reparameterization acceleration.

Model Lightweighting via Structural Reparameterization

High-precision deep learning models are required for edge computing devices of USVs and river patrol drones because edge devices typically lack memory bandwidth and parallel computing power of cloud servers. To satisfy ultra-real-time detection requirements without

loss of feature extraction capabilities, we developed a "training-inference decoupled" lightweight architecture using RepC3 **(Figure. 6b)**.

The basic principle of structural reparameterization is to use algebraic equivalence transformations to separate training architecture from inference. At training time RepC3 intentionally constructs a heavy multi-branch parallel topology[32], consisting of 3×3, 1×1 convolution branch and identity mapping branch (shortcut). This multi-branch design enriches gradient flow and allows the network to learn very diverse feature representations (e.g., different texture of plastic bottles vs sharp edges of metal cans). Assuming input feature tensor is X and output feature is Y, the training-time forward pass is defined mathematically as follows:

$$Y_{train} = (\mathrm{BN}_{3\times3}(W_{3\times3} * X)) + (\mathrm{BN}_{1\times1}(W_{1\times1} * X)) + \mathrm{BN}_{id}(X)$$

where $*$ denotes the convolution operation, and BN denotes the Batch Normalization layer associated with each branch.During the deployment and inference stage, executing multiple branches concurrently incurs severe memory access cost (MAC) overhead and slows down inference speed due to repeated tensor allocations. Based on the linear additivity property of convolution operations, we losslessly merge all parallel branches into a single-pass 3×3 convolution. First, we fuse the Batch Normalization parameters (scaling factor γ, shifting factor β, running mean μ, and running variance σ^2) directly into the convolutional weights W and biases B for each branch:

$$W_{fused} = \frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} W, B_{fused} = \beta - \frac{\gamma\mu}{\sqrt{\sigma^2 + \epsilon}}$$

After fusing the BN layers, the 1×1convolution kernel is zero-padded to match the spatial dimensions of the 3×3 kernel, denoted as $\mathrm{Pad}(W'_{1\times1})$. The identity mapping is equivalent to a 3×3Dirac delta convolution kernel I, where the central weight is 1 and all other weights are 0. Finally, the equivalent weights $W_{eq}$and biases $B_{eq}$for the unified single-path inference structure are calculated by summing the fused parameters:

$$W_{eq} = W'_{3\times3} + \mathrm{Pad}(W'_{1\times1}) + \mathrm{I}$$

$$B_{eq} = B'_{3\times3} + B'_{1\times1} + B'_{id}$$

This mathematically equivalent transformation ensures that . However, for edge devices inference on edge devices only executes one convolution operation. This eliminates memory access and structure fragmentation caused by multiple branches. This reduces the number of trainable parameters and FLOPs, while keeping high accuracy learned from multi-branch training.

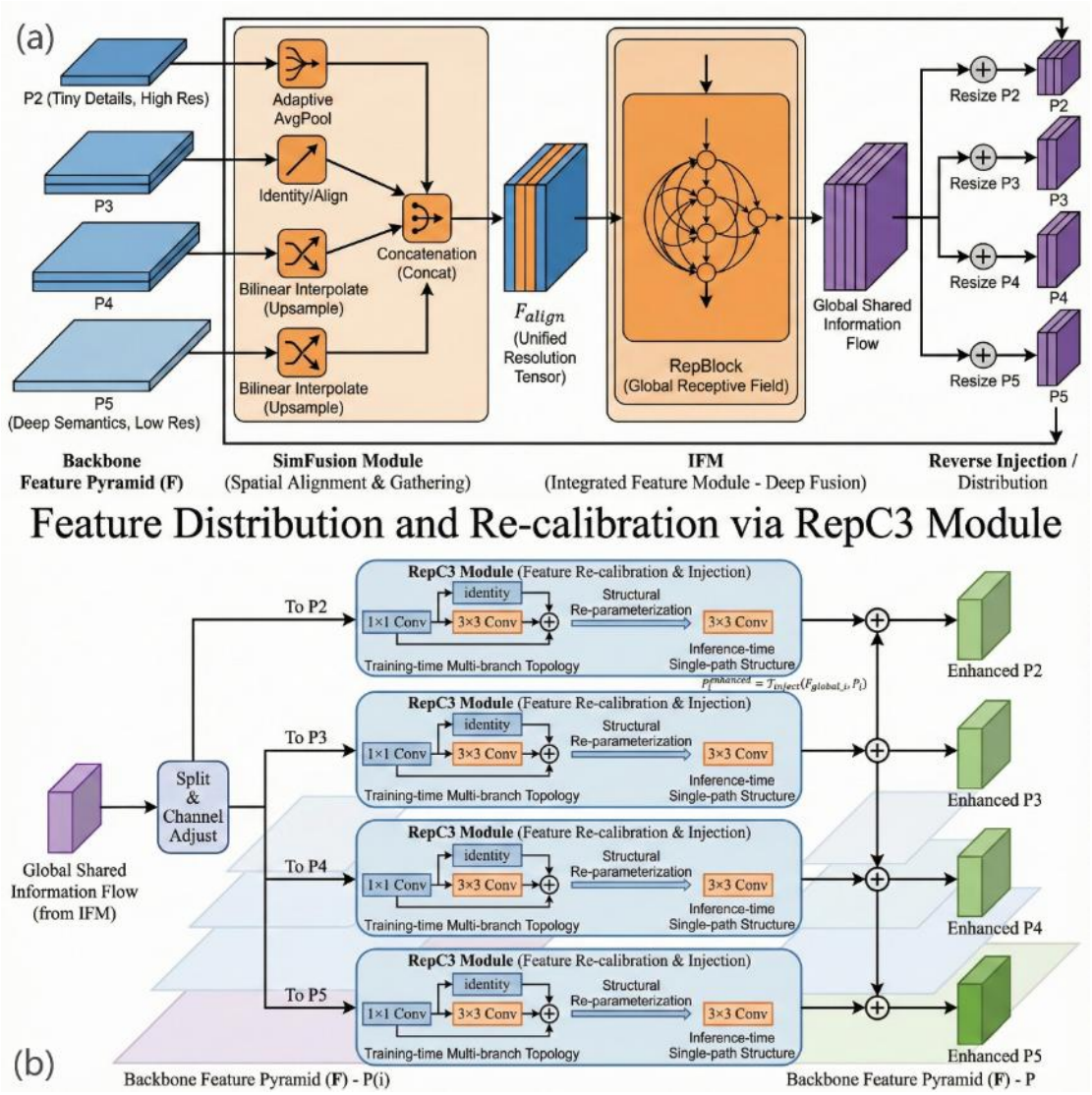


Figure 6

Loss Function Design

To supervise the end-to-end training of the BR-Trans model effectively, a composite loss

function was employed. The total loss $\mathcal{L}_{total}$ is computed as a weighted sum of three distinct components: Classification Loss ($\mathcal{L}_{cls}$), Bounding Box Regression Loss ($\mathcal{L}_{L1}$), and Generalized Intersection over Union Loss ($\mathcal{L}_{GIoU}$).

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{L1} + \lambda_3 \mathcal{L}_{GIoU}$$

$\lambda_1, \lambda_2,$ and λ_3 are hyperparameter weights, respectively, balancing each loss term contribution. $\mathcal{L}_{cls}$ uses Varifocal Loss to correct class imbalance and classify the debris. $\mathcal{L}_{L1}$ computes Mean Absolute Error between predicted and ground truth bounding box coordinates. In order to improve the geometrical alignment of bounding boxes (especially irregular shapes) $\mathcal{L}_{GIoU}$ uses the GIoU loss. GIoU loss penalizes bounding boxes far apart by taking the area of the smallest enclosing convex hull C (containing both predicted box A and ground truth box B)[33]:

$$\mathcal{L}_{GIoU} = 1 - \left(\frac{| A \cap B |}{| A \cup B |} - \frac{| C \setminus (A \cup B) |}{| C |} \right)$$

This composite loss strategy ensures robust optimization, particularly during the early stages of training where target localization is ambiguous.

Experimental Setup and Implementation Details

To systematically validate the performance of the proposed BR-Trans framework, we designed two rigorous sets of experiments. Detailed hardware configurations and hyperparameter settings are provided in **Supplementary Note 1** and S**upplementary Tables S1-S3 online**.

1. **Ablation Studies:** The study comprises five experimental groups with progressively increasing complexity, designed to thoroughly validate the contributions of the StarBlock backbone, the SimFusion module and the BRM mechanism, both individually and in combination.

2. **Comparative Experiments:** We compare the performance of the comprehensively improved BR-Trans model against a range of advanced object detectors widely used in industrial applications. The selected baseline models include single-stage

algorithms (YOLOv8m, YOLO11m, YOLO26) and the DETR series of algorithms (the original RT-DETR and RT-DETR-r50).

Model training and hyperparameter tuning were run on NVIDIA L20 Server GPU cluster for Transformer architectures. All inference tests and qualitative evaluations on BR-Trans deployed on USVs and river patrol drones were run on NVIDIA Jetson Orin Nano (8GB) edge module running Ubuntu 22.04 LTS. The platform has 6 Core Arm Cortex® A78AE CPU and 1024 Core NVIDIA Ampere GPU with 32 Tensor Cores. These are typical hardware requirements of edge robotic systems. All algorithms were run Python 3.11 with PyTorch 2.9.1 and TorchVision 0.24.1. All models were trained for 200 epochs. YRDG data was overfitted using AdamW optimizer. Initial learning rate is set to 1×10^{-3} with linear warmup for the first 3 epochs and cosine-annealing scheduler to decay the learning rate to a minimum of 1×10^{-5} . Batch size 16 . Input images are resized to 640×640 pixels and 4 parallel dataloader workers are used to maximize GPU utilization.

Evaluation Metrics

Model performance was comprehensively and objectively evaluated using standard COCO object detection metrics, covering both detection accuracy and model complexity limits.

Mean Average Precision (mAP) was used as the main detection metric, integrating precision and recall over confidence ranges. We reported mAP at fixed Intersection over Union (IoU) value 0.5 (mAP@0.5) for general localization performance, and mAP at fixed IoU value 0.50 (mAP@0.95) for high accuracy bounding box regression.Precision ($P = \frac{TP}{TP+FP}$) and Recall ($R = \frac{TP}{TP+FN}$) were explicitly calculated for false positive rate (susceptible to background noise) and false negative rate (missed detections of small objects) respectively. The F1-score ($F1 = \frac{2 \cdot P \cdot R}{P+R}$) provided harmonic mean of detection performance.

We assessed model complexity to ensure that it is suitable for use on resource-limited edge hardware: the number of trainable Parameters (Params) measures persistent storage and RAM requirement; FLOPs measure computational complexity (and directly correlate with real-time

FPS and power consumption on platforms such as NVIDIA Jetson series).

Data availability

The YRDG analysed in this paper, including the high resolution aerial images and bounding boxes for all debris types from the training, validation and testing sets, is available at https://github.com/jingcodejing/YRDG-dataset/tree/main, and the source code of the proposed BR-Trans framework generated in this paper is available on https://github.com/userzyh666/BR-Trans. In this code repository we implement the basic architecture (StarNet backbone, SimFusion, BRM and RepC3), training scripts, inference codes and evaluation procedures needed to reproduce experimental results and comparative results presented in this paper.

References

1. MacLeod, M., Arp, H. P. H., Tekman, M. B. & Jahnke, A. The global threat from plastic pollution. *Science* **373**, 61–65 (2021).
2. Vethaak, A. D. & Legler, J. Microplastics and human health. *Science* **371**, 672–674 (2021).
3. Li, J. *et al.* Deep learning for visual recognition and detection of aquatic animals: a review. *Rev. Aquac.* **15**, 409–433 (2023).
4. Junzhe, Z., Fuqiang, J., Yupeng, C., Weiyi, W. & Qing, W. A water surface garbage recognition method based on transfer learning and image enhancement. *Results Eng.* **19**, 101340 (2023).
5. Zhao, Y. *et al.* DETRs beat YOLOs on real-time object detection. in 16965–16974 (2024).
6. Kiefer, B. *et al.* 1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge

Results. in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* 265–302 (IEEE, Waikoloa, HI, USA, 2023). doi:10.1109/WACVW58289.2023.00033.

7. Zhang, R. *et al.* Survey on deep learning-based marine object detection. *J. Adv. Transp.* **2021**, 5808206 (2021).
8. Zhu, X. *et al.* Deformable DETR: deformable transformers for end-to-end object detection. Preprint at https://doi.org/10.48550/arXiv.2010.04159 (2021).
9. Huang, S. *et al.* DEIM: DETR with improved matching for fast convergence. 1–10 doi:10.1109/CVPR52734.2025.01412.
10. Proença, P. F. & Simões, P. TACO: Trash Annotations in Context for Litter Detection. Preprint at https://doi.org/10.48550/arXiv.2003.06975 (2020).
11. Hong, J., Fulton, M. & Sattar, J. TrashCan: A Semantically-Segmented Dataset towards Visual Detection of Marine Debris. Preprint at https://doi.org/10.48550/arXiv.2007.08097 (2020).
12. Tong, Y., Ye, H., Yang, J. & Yang, X. ACD-DETR: adaptive cross-scale detection transformer for small object detection in UAV imagery. *Sensors* **25**, (2025).
13. Niu, Y., Xu, R., Lin, Z. & Liu, W. STD-net: spatio-temporal decomposition network for video demoiréing with sparse transformers. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 8562–8575 (2024).
14. Lin, T.-Y. *et al.* Feature Pyramid Networks for Object Detection. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 936–944 (IEEE, Honolulu, HI, 2017). doi:10.1109/CVPR.2017.106.

15. Tan, M., Pang, R. & Le, Q. V. EfficientDet: scalable and efficient object detection. in 10781–10790 (2020).
16. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional Block Attention Module. in *Computer vision – ECCV 2018* (eds Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y.) vol. 11211 3–19 (Springer International Publishing, Cham, 2018).
17. Wang, Z. *et al.* Visual attention-based target detection and discrimination for high-resolution SAR images in complex scenes. *IEEE Trans. Geosci. Remote Sens.* **56**, 1855–1872 (2018).
18. Wang, H., Qian, H., Feng, S. & Wang, W. L-SSD: lightweight SSD target detection based on depth-separable convolution. *J. Real-Time Image Process.* **21**, 33–47 (2024).
19. Zhu, J., Yang, Y. & Cheng, Y. SMURF: a fully autonomous water surface cleaning robot with a novel coverage path planning method. *J. Mar. Sci. Eng.* **10**, (2022).
20. Jiang, Z., Wu, B., Ma, L., Zhang, H. & Lian, J. APM-YOLOv7 for Small-Target Water-Floating Garbage Detection Based on Multi-Scale Feature Adaptive Weighted Fusion. *Sensors* **24**, 50 (2023).
21. Liu, B., Jin, J., Zhang, Y. & Sun, C. WRRT-DETR: Weather-Robust RT-DETR for Drone-View Object Detection in Adverse Weather. *Drones* **9**, 369–389 (2025).
22. Lv, W. *et al.* RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer. Preprint at https://doi.org/10.48550/arXiv.2407.17140 (2024).
23. Wang, S., Xia, C., Lv, F. & Shi, Y. RT-DETRv3: Real-Time End-to-End Object

Detection with Hierarchical Dense Positive Supervision. 1–9 doi:10.1109/WACV61041.2025.00166.

24. Carion, N. *et al.* End-to-end object detection with transformers. in *Computer Vision – ECCV 2020* (eds Vedaldi, A., Bischof, H., Brox, T. & Frahm, J.-M.) 213–229 (Springer International Publishing, Cham, 2020). doi:10.1007/978-3-030-58452-8_13.
25. Zhang, H. *et al.* DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. Preprint at https://doi.org/10.48550/arXiv.2203.03605 (2022).
26. Ma, X., Dai, X., Bai, Y., Wang, Y. & Fu, Y. Rewrite the Stars. 1–10 doi:10.1109/CVPR52733.2024.00544.
27. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **10**, 1299–1319 (1998).
28. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, Las Vegas, NV, USA, 2016). doi:10.1109/CVPR.2016.90.
29. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1800–1807 (IEEE, Honolulu, HI, 2017). doi:10.1109/CVPR.2017.195.
30. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. in 7132–7141 (2018). doi:10.1109/CVPR.2018.00745.
31. Wang, C. *et al.* Gold-YOLO: efficient object detector via gather-and-distribute mechanism. *Adv. Neural Inf. Process. Syst.* **36**, 51094–51112 (2023).

32. Ding, X. *et al.* RepVGG: Making VGG-style ConvNets Great Again. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 13728–13737 (IEEE, Nashville, TN, USA, 2021). doi:10.1109/CVPR46437.2021.01352.
33. Rezatofighi, H. *et al.* Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 658–666 (IEEE, Long Beach, CA, USA, 2019). doi:10.1109/CVPR.2019.00075.

Funding

No funding was received for this work.

Acknowledgements

2026 Dongguan City Basic Research Joint Fund Project: Research on Unmanned Aerial Vehicle Assembly Recognition and Precise Delivery Technology in Rejection Environment in Low-altitude Logistics Scenarios; 2024 Guangdong Provincial Key Discipline Research Capacity Enhancement Project (Project No.: 2024ZDJS067); 2023 Guangdong Provincial University Innovation Team Project (Project No.: 2023KCXTD046)

Author Contributions

Conceptualization, Y.Z., S.G., and Y.S.; methodology, Y.Z.; software, Y.Z.; validation, Y.Z., J.G. and J.P.; investigation, Y.Z. and S.G.; resources, S.G. and Y.S.; data curation, Y.Z. and J.G.; writ-ing—original draft preparation, Y.Z.; writing—review and editing, Y.Z., S.G., and C.L.; visuali-zation, Y.Z.; supervision, S.G.; project administration, S.G.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Figure legends

Figure 1: Radar chart of comprehensive performance for each evaluated model.

Figure 2: Overview of the model training and validation loss curves and accuracy metrics.

Figure 3: Comparative visualization of class activation maps. The heatmaps demonstrate the focus regions of BR-Trans compared to baseline models in complex aquatic backgrounds.

Figure 4: Overall architecture of the proposed BR-Trans model. The system integrates a StarNet-based backbone, a SimFusion-enhanced Full-scale Re-Calibration FPN, and Bidirectional Re-calibration Modules (BRM) to effectively process multi-scale floating debris.

Figure 5: Detailed schematics of feature extraction modules. **(a)** High-Level and Low-Level calibration streams in the Bidirectional Re-calibration Module (BRM). **(b)** The Star BasicBlock utilizing high-dimensional implicit mapping.

Figure 6: Schematics of feature fusion and lightweighting modules. **(a)** The SimFusion Module for spatial alignment and gathering. **(b)** Feature Distribution via the RepC3 Module, illustrating the training-inference decoupled structure.

Labels

Table 1: Ablation Experiment Results

Star block	sim3bin/4bin	BRM	mAP@0.5 (%)	Precision (%)	Recall (%)	F1-score	Params	GFLOPs
			91.97	90.36	91.81	90.36	77.0 M	57
✓			92.08	93.79	90.27	91.96	77.8 M	58.2
	✓		92.26	93.98	90.82	92.35	71.7M	61.3
		✓	91.62	93.99	90.52	91.62	79.3 M	59.6
	✓	✓	91.63	93.01	90.66	91.79	70.9M	60.1
✓		✓	92.31	93.06	90.35	92.14	76.7 M	85.6
✓	✓		92.36	92.8	91.08	91.89	86.4M	60
✓	✓	✓	92.54	93.36	92.82	93.08	71.7 M	61.3

Table 2: Performance Comparison of Different Models

Model name	mAP@0.5 (%)	mAP@0.5: 0.95 (%)	Precision (%)	Recall (%)	F1-Score	Params	GFLOPs
YOLOv8	86.42	45.93	84.88	79.47	81.75	49.6 MB	78.7
YOLO11	86.91	45.95	84.78	81.21	82.79	38.7 MB	67.7
YOLO26	81.09	42.64	77.36	73.54	74.79	42.0 MB	67.9
RT-DETR-r18	91.97	48.55	90.36	91.81	90.36	77.0 MB	57
RT-DETR-r50	90.9	47.6	92.88	91.45	92.13	163.9 MB	129.6
Ours	92.54	49.22	93.36	92.82	93.08	71.7 MB	61.3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf