YOLO-MC: Small Target Detection Algorithm for Complex Water Environment

Zhaoming Wu, Mengjun An, Chengzhi Deng, Xiaowei Sun, Chenguang Xu

Abstract—**Detecting surface-floating objects is critical for applications like autonomous ship navigation and river sanitation. However, current methodologies have significant shortcomings in recall rate and detection precision due to the small size of floating objects and complex environmental interferences such as water surface oscillations and light variations. This study presents an enhanced model, YOLO-MC, derived from YOLOv11, to address these issues. To enhance the feature representation capability of the network model for small targets, a multiscale attention wavelet transform convolution (MSWTC) architecture is devised, and the C3k2 module is refined in the neck region. This structure enhances the feature extraction capability of small floating objects on the water surface, resulting in a significant increase in detection accuracy. A coordinate edge attention module (CEAM) is added before the detection head to increase the representation of edge information. Despite minor improvements in accuracy, it greatly improves recall rate and successfully augments the model's capacity to distinguish difficult targets. Experimental results indicate that the optimized model's recall rate R has grown by 2.1%, mAP@0.5 has increased by 2.3%, and mAP@0.5:0.95 has improved by 0.4%.**

Index Terms—**Detection of floating objects; minor targets; extraction of features; edge data**

I. Introduction

With rapid economic development and increased human activity, the problem of river water pollution is getting more serious, and floating waste on the water's surface has emerged as a major concealed threat to the natural environment and maritime safety. Traditional manual salvaging is inefficient, dangerous, and understaffed, making it difficult to satisfy actual needs. Recent breakthroughs in deep learning technology have made it possible to detect and clear surface trash automatically. Rapid identification and processing of floating items on the water's surface is possible because of multi-sensor fusion and fast target detection algorithms. However, due to the features of tiny target size, frequent fluctuations, and high reflections in water scene scenarios, existing detection algorithms frequently have issues with missed detection and false detection, making it difficult to adapt to the diversity of the natural environment. As a result, developing a small target detection model capable of detecting floating items on the water's surface has emerged as a critical technical challenge in environmental monitoring and water management.

Object detection is a key component of computer vision, and deep learning has accelerated progress in this area. The current dominant methods are classified into two types: CNN-based[1] and transformer-based[2].Transformer methods (such as RT-DETR[3], Deformable DETR[4], and Dynamic DETR[5]) use the self-attention mechanism to improve feature modeling capabilities and adapt to complex scenes, whereas CNN-based methods are classified into two-stage (such as R-CNN, Fast R-CNN, and Faster R-CNN[6]) and single-stage (such as SSD[7] and the YOLO series[8,9]) architectures. The two-stage method is more accurate but more computationally expensive, whereas the single-stage method has a lighter structure and is better suited to real-time detection. Nonetheless, in specific application scenarios such as surface rubbish identification, items are small and easily disturbed, and single-stage algorithms frequently struggle to attain optimal accuracy, whilst two-stage algorithms have large computational costs due to sophisticated processing procedures. Although Transformer can capture long-distance dependencies, its computing expense limits its usefulness. To meet the objectives of surface waste identification, it is very important to choose and refine appropriate detection algorithms in certain settings.

Because water surface variations, illumination changes, and reflections frequently produce spurious objects, resulting in misidentification and lower detection accuracy, this article introduces the C3k2MSWTC module in the neck structure based on YOLOv11. This module uses a multiscale wavelet transform to decompose the input signal into various frequency bands and scales. It not only captures the diverse characteristics of low-frequency long waves and high-frequency rapid disturbances in water surface fluctuations, but it also has excellent edge detection capabilities, allowing the model to more precisely perceive hidden small target features and improve fine-grained target detection performance. Furthermore, the CEAM introduced in front of the detection head combines coordinate attention and edge enhancement mechanisms, effectively enhancing

Manuscript received May 23, 2025; revised August 30, 2025.

This work was supported in part by Jiangxi Provincial Natural Science Foundation(20252BAC240112).

Zhaoming Wu is an associate professor at the School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang 330099, Jiangxi, China (e-mail: zmwunit@foxmail.com).

Mengjun An is a postgraduate student at the School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang 330099, Jiangxi, China (e-mail: anmengjun1025@163.com).

Chengzhi Deng is a professor at the School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang 330099, Jiangxi, China (corresponding author, e-mail: dengcz@nit.edu.cn).

Xiaowei Sun is a lecturer at the School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang 330099, Jiangxi, China (e-mail: 304652253@qq.com).

Chenguang Xu is a lecturer at the School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang 330099, Jiangxi, China (e-mail: xcg@nit.edu.cn).

the expression of target contours and local structural features, as well as improving the model's accuracy in identifying small targets, reflective blurred targets, and targets obscured by water waves. Overall, the addition of the aforesaid modules considerably improves the model's detection performance in complicated water surface situations.

The contributions of this work can be summarized in the following ways:

(1) To optimize C3k2, a multiscale attention wavelet transform convolution module (MSWTC) was developed, allowing the model to more sensitively capture information about small targets hidden in water surface fluctuations without significantly increasing the number of parameters, thereby improving small target recognition accuracy.

(2) At the same time, the CEAM module inserted before the detection head improves the model's ability to describe local features of small target edges, resulting in improved detection performance for drifting small objects.

(3) Experiments reveal that YOLO-MC has higher recall and detection accuracy on the public FloW-IMG and WSODD datasets.

II. Related Work

A. Small object detection

The fundamental issue of small object recognition is that its pixel fraction in the image is tiny[10], and the discriminant information that can be employed to identify targets is restricted. To solve the problems of insufficient feature extraction and low precision in small object detection, Xiao et al.[11] proposed a feature enhancement method, including refining the backbone network, increasing the number of small target queries, and optimizing the loss function to improve the performance of small object detection, and further improved the perception of small targets through hollow convolution and recursive prediction modules; to solve the problem of lack of semantic information in low-level networks in small object detection, Song et al.[12] proposed a small target detection algorithm based on multiscale feature fusion, which improved the detection performance by fusing shallow and deep features. In addition, a detector MSFYOLO based on a multiscale deep feature learning network was built, which combined global and local information and optimized the detection effect using a feature pyramid. At the same time, a novel feature extraction network, CourNet, was suggested, which can better represent the feature information of small objects. To tackle the limitations of small objects in complex background detection, Zhao et al.[13] proposed a highly efficient algorithm. The cross-scale feature fusion attention module (ECFA) employs the attention mechanism to effectively suppress noise and strengthen relevant features, thereby addressing feature redundancy and insufficient representation of small targets. SEConv, an efficient convolution module, aims to reduce computational redundancy and improve multiscale feature learning. Furthermore, a dynamic focus sample weighting function, DFSLoss, is provided to successfully address the issue of sample difficulty imbalance, and Wise-IoU is introduced to mitigate the influence of low-quality examples on model convergence. Deng et al.[14] proposed an extended feature pyramid network (EFPN), which introduces an ultra-high-resolution pyramid level specifically for small object detection, designs a feature texture transfer (FTT) module for super-resolution feature extraction and retains regional details, and adopts a cross-resolution distillation mechanism to transfer detailed information.

B. Floating object detection

Previous research on surface floating object identification has made substantial progress. Chen et al.[15] suggested a more advanced YOLOv5 model for real-time detection of small surface floating objects. The model significantly improved the detection effect of small objects by more effectively fusing shallow and deep features and alleviating the problem of missed detection; Shi et al.[16] proposed a surface floating object detection algorithm based on CDW-YOLOv8, which improved the C2f module (C2f-CA) by introducing a coordinate attention mechanism, replaced the Upsample with the DySample module, and added a small object detection layer to improve the perception of small floating debris; at the same time, the Focaler-WIoUv3 loss function was used for optimizing the positioning accuracy and reduce the impact of low-quality anchor frames; Chen et al.[17] proposed a detection and tracking method based on spatiotemporal information fusion, improved the SSD network by enhancing the high-resolution layer to adapt to small target detection, and introduced the fast directional gradient histogram (FHOG) and pyramid scale estimation to improve KCF tracker, and combined the detection and tracking results for spatiotemporal fusion, which effectively solved the problem of difficult detection of small floating objects in complex water surface environments; Zhang et al.[18] proposed a real-time water garbage detection model based on RefineDet. They upgraded the anchor refinement module, collected more detailed semantic information, and increased the model's detection accuracy. In addition, they introduced the focal loss function and tweaked its parameters to boost the model's detection performance.

III. YOLO-MC Network Model

A. Introduction to YOLO-MC

The overall structure of the YOLO-MC algorithm is shown in Figure 1. Based on the overall architecture of YOLOv11n, the algorithm designs the MSWTC module to optimize the C3k2 structure of the neck; in addition, we design CEAM in front of the detection head so that the detection head can focus more on small target floating objects. Through the above improvements, the YOLO-MC algorithm achieves a balance between performance and efficiency between computational complexity, number of parameters, and computational efficiency, thereby ensuring that it can provide high-performance detection capabilities when deployed on edge devices for small target floating objects[19,20] on the water surface.

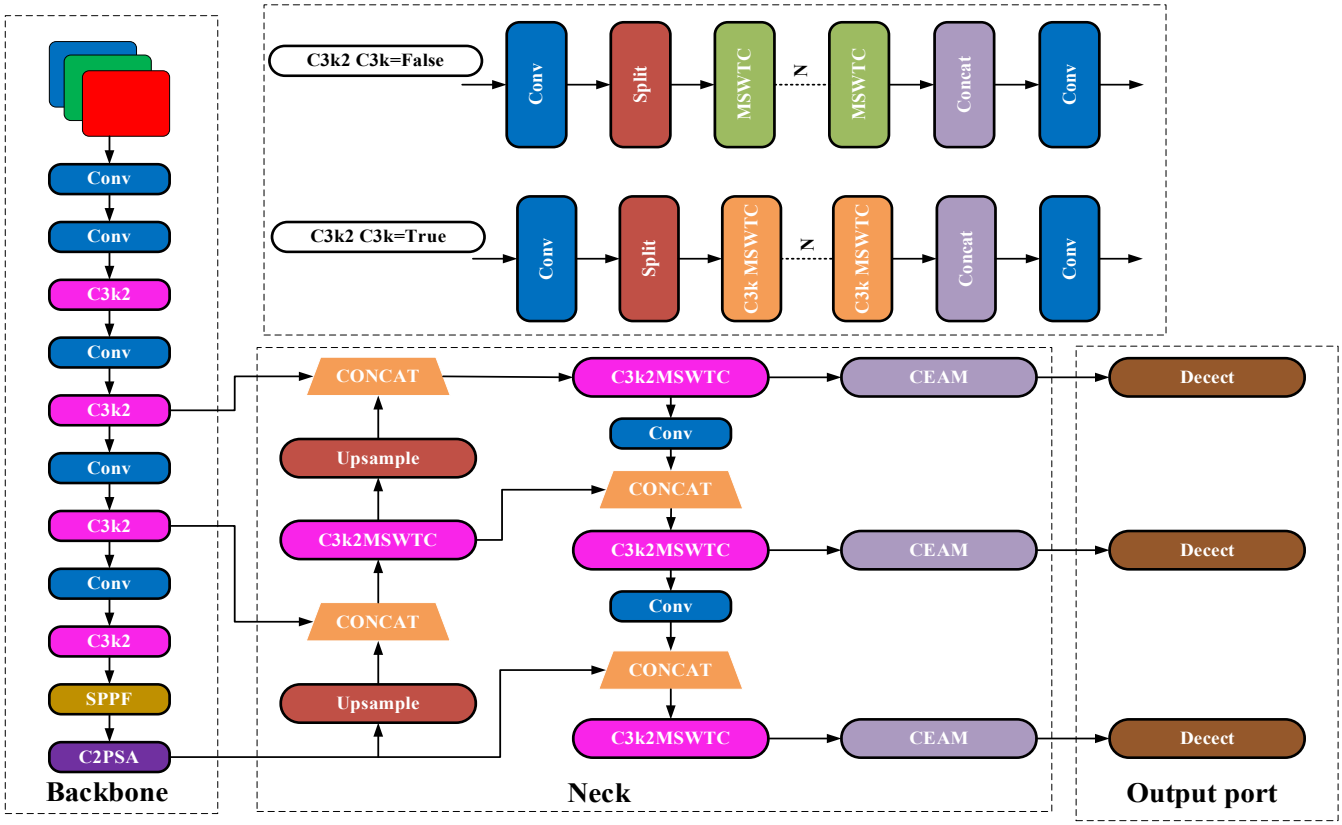


Fig. 1. YOLO-MC model. When C3k is False, the C3k2 module acts as a traditional C2f module and contains a conventional bottleneck structure; when C3k is True, the bottleneck module will be replaced by a more efficient C3 module.

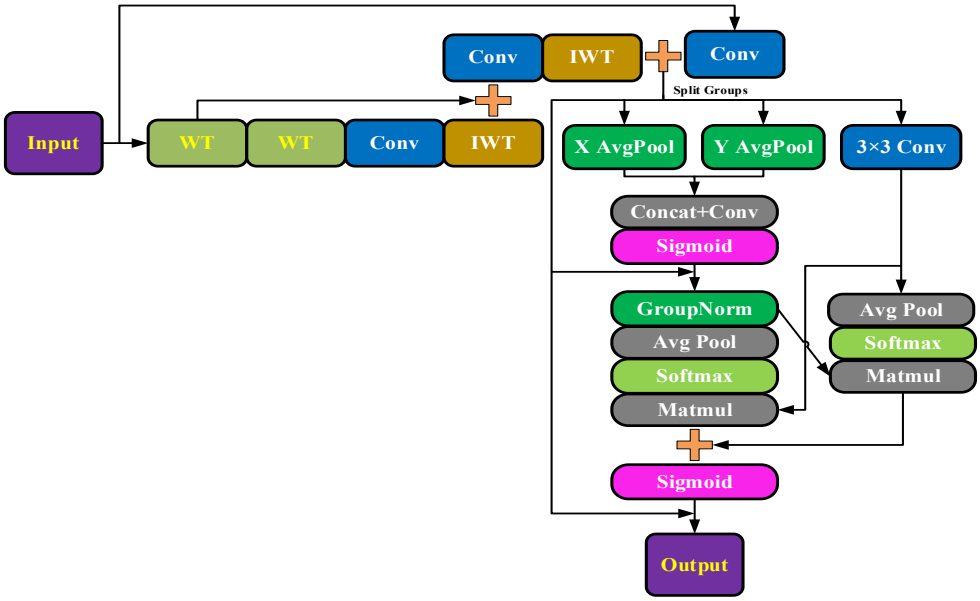


Fig. 2. MSWTC module

B. *MSWTC*

For solving feature interference induced by water surface fluctuations and lighting variations in the detection of small floating objects on the water surface, this study presents a new module based on multiscale attention wavelet transform convolution (MSWTC), as illustrated in Figure 2. The original architecture of this module was based on the Wavelet Transform Convolution (WTC) structure described by Finder et al.[21]. Its main idea is to employ the wavelet transform to broaden the receptive field[22] of a convolutional neural network (CNN) without considerably altering model parameters. Traditionally, large-size convolution kernels have been utilized to broaden the receptive field, increasing both the number of parameters and the computational complexity. The MSWTC module divides the input signal into various frequency bands using cascading layers of wavelet transforms, extracting low-frequency and high-frequency information at different scales. Specifically, the low-frequency section keeps the general contour and global structural information and is stable to brightness variations produced by illumination changes, whereas the high-frequency part can capture local detail changes caused by water surface oscillations. Although these high-frequency signals are easily disturbed by noise, independent convolution processing[23] allows them to better restore the

target's edge information. Following decomposition, the characteristics of each frequency band are processed using a separate convolution technique. This approach can include adaptive filters for various frequency domain variables, increasing the accuracy of information extraction. This method of layer-by-layer augmenting local features and global structures using a succession of modest convolution kernel operations enables the network to record subtle feature changes while retaining general context information, effectively widening the network's receptive field.

MSWTC also introduces a multiscale attention method to account for the varying significance of characteristics at different scales[24] to the detection job in real-world water scenes. This mechanism can dynamically allocate attention weights based on the feature responses extracted from each frequency band, allowing the network to adapt to changes in illumination or water surface, focusing on scale features[25] that are more discriminative for detecting small target drifting objects[26]. Specifically, features at key scales typically hold more stable structural information, whereas other scales may be influenced by external noise. Dynamic weighting not only reduces noise interference but also optimizes the overall feature representation.

In the previous content, we introduced the basic principles and advantages of the MSWTC module. Next, we will use formulas to explain in detail the working mechanism of the module and how to optimize signal processing to improve the accuracy of small target floating object detection. For a given two-dimensional image X, a one-dimensional transformation is applied to the two spatial dimensions (width and height) of the image to obtain outputs in four different frequency bands:

$$\begin{aligned} f_{LL} &= \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad f_{LH} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \\ f_{HL} &= \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \quad f_{HH} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{aligned} \tag{1}$$

Among them, f_{LL} is a low-frequency filter to preserve the overall structure, f_{LH} , f_{HL} and f_{HH} are horizontal high-frequency filters to capture horizontal edges, vertical high-frequency filters to capture vertical edges, and diagonal high-frequency filters to capture corner details.

Convolving the input image X with these filter's results in four output channels:

$$\begin{bmatrix} X_{LL} & X_{LH} \\ X_{HL} & X_{HH} \end{bmatrix} = Conv([f_{LL}, f_{LH}, f_{HL}, f_{HH}], X) \tag{2}$$

Among them, X_{LL} is the low-frequency part, and the other three are high-frequency components, corresponding to the high-frequency information of horizontal, vertical and diagonal lines respectively.

To capture a wider range of contextual information, the cascaded wavelet decomposition is performed by recursively performing wavelet transform on the low-frequency components. The decomposition result of each layer contains low-frequency and high-frequency components, gradually reducing the spatial resolution and increasing the frequency resolution:

$$X_{LL}^{(i)}, X_{LH}^{(i)}, X_{HL}^{(i)}, X_{HH}^{(i)} = WT(X_{LL}^{(i-1)}) \tag{3}$$

Where $X_{LL}^{(0)} = X$ is the original input, and i represents the current wavelet decomposition level.

The cascaded wavelet transform is further combined with the small convolution kernel to recursively process the low-frequency part of each layer. The convolution process of each layer is as follows:

$$\begin{aligned} X_{LL}^{(i)}, X_H^{(i)} &= WT(X_{LL}^{(i-1)}), \\ Y_{LL}^{(i)}, Y_H^{(i)} &= Conv(W^{(i)}, (X_{LL}^{(i)}, X_H^{(i)})) \end{aligned} \tag{4}$$

Where $X_H^{(i)}$ represents all high-frequency components of the i layer.

Then, the results of different frequency levels are combined by inverse wavelet transform:

$$Z^{(i)} = IWT(Y_{LL}^{(i)} + Z^{(i+1)}, Y_H^{(i)}) \tag{5}$$

Among them, $Z^{(i)}$ is the aggregated output of all convolution results starting from the i layer.

After the wavelet transform convolution, the attention mechanism is introduced to weight the features of different scales. The goal of the attention mechanism is to dynamically adjust the weights of the features of each scale according to their relevance to the current task. The attention weight $\alpha_{att}(\alpha)$ is calculated by the feature $Z_\alpha^{(i)}$ of each scale α . The weight can be calculated by the following attention mechanism:

$$\begin{aligned} f_\alpha &= GAP(Z_\alpha^{(i)}) \cdot W_{att} \\ \alpha_{att}(\alpha) &= \frac{exp(f_\alpha)}{\sum_{\alpha'} exp(f_{\alpha'})} \end{aligned} \tag{6}$$

Among them, f_α is the correlation score of the scale feature calculated in some way, GAP is the global average pooling, W_{att} is the learnable weight matrix, and $\alpha_{att}(\alpha)$ reflects the importance of each scale feature in the current task.

After calculating the attention weight of each scale, we weight the multiscale features obtained by wavelet transform convolution. The weighted feature $\tilde{x}_\alpha$ of each scale can be expressed as:

$$\tilde{x}_\alpha(t) = \alpha_{att}(\alpha) \cdot Z_\alpha^{(i)} \tag{7}$$

Where $Z_\alpha^{(i)}$ is the signal feature at a certain scale α obtained by wavelet transform convolution, and $\alpha_{att}(\alpha)$ is the weighting coefficient calculated by the attention mechanism.

Finally, the weighted multiscale features are fused to obtain the final output feature $\tilde{x}(t)$, which can be expressed as:

$$\tilde{x}(t) = \sum_\alpha \tilde{x}_\alpha(t) \tag{8}$$

Combining the above parts, the overall formula of the MSWTC module can be expressed as:

$$\tilde{x}(t) = \sum_\alpha \alpha_{att}(\alpha) \cdot IWT(Y_{LL}^{(i)} + Z^{(i+1)}, Y_H^{(i)}) \tag{9}$$

From the perspective of signal processing, the improved C3k2MSWTC module effectively improves the signal-to-noise ratio through adaptive spectrum control[27], which significantly enhances the signal characteristics of small targets under complex water surface conditions. The C3k2MSWTC module implements fine filtering on each frequency component, effectively suppressing non-ideal responses in water surface fluctuations and false signals caused by reflections, while retaining and enhancing the spectral characteristics of key signals. Therefore, the entire model shows excellent robustness and adaptability when

processing subtle signal changes in a spatial dynamic water surface environment, providing a solid foundation for image signal extraction for the detection of small floating objects on the water surface.

C. CEAM

To improve the model's perception of small target edge information[28], this article offers a coordinate edge attention module (CEAM), as illustrated in Figure 3, which is embedded at the front of the detection head to improve responsiveness to fine-grained features. This module combines the coordinate attention mechanism[29] with the edge enhancement technique and is divided into three parts: First, edge features are extracted using a learnable deep separable convolution[30], which is then spliced with the original feature map in the channel dimension to add spatial semantic information. Following that, a direction-sensitive pooling operation (in the horizontal and vertical directions) is employed to extract contextual features in the spatial dimension, and feature compression and nonlinear augmentation are achieved using lightweight convolution[31] and activation functions. The channel attention branch generates channel weights using global average pooling and convolution to alter the relevance of distinct channels, while the spatial attention branch combines average pooling and maximum pooling to direct the model's attention to crucial spatial positions. Finally, the fused channel and spatial attention weights work on the input characteristics to significantly improve the edge area.

Let the input feature map be $X \in R^{N\times C\times H\times W}$, where N represents the batch size, C represents the number of channels, and H and W represent the height and width respectively. First, define an edge extraction operator $\mathcal{E}(\cdot)$ (composed of depthwise separable convolution, batch normalization, and ReLU activation) to perform edge enhancement on the input X:

$$E = \mathcal{E}(X) \in R^{N\times C\times H\times W} \tag{10}$$

To fully integrate the fine-grained edge features extracted by the edge detector with the original semantic features, the two are spliced in the channel dimension to form an enhanced feature map containing rich spatial edge information and semantic information:

$$X_{edge} = Concat(X, E) \in R^{N\times 2C\times H\times W} \tag{11}$$

Subsequently, in order to capture the directional information of features in different spatial dimensions, the enhanced feature map X_{edge} is adaptively averaged pooled in the horizontal and vertical directions, thereby extracting global context information in the height and width directions:

$$\begin{aligned} X_H &= Pool_H(X_{edge}) \in R^{N\times 2C\times H\times 1} \\ X_W &= Pool_W(X_{edge}) \in R^{N\times 2C\times W\times 1} \end{aligned} \tag{12}$$

To achieve unified modeling of features in spatial dimensions, X_W needs to be transposed in height and width dimensions to make it consistent with X_H in dimensions. Then the two are concatenated in the height direction, and the contextual information from the horizontal and vertical directions is integrated to construct a unified direction-aware representation feature map.

$$Y = Conv(Concat(X_H, X_W^{\top})) \in R^{N\times M\times (H+W)\times 1} \tag{13}$$

Where M is the number of intermediate channels, and nonlinear transformation is achieved here through convolution, batch normalization and h-swish activation.

Next, in the channel attention branch, global average pooling (GAP) is applied to the fused feature Y to obtain global context information in the channel dimension. This operation compresses the spatial dimension to 1×1 and generates a global response representation for each channel. Subsequently, two layers of consecutive 1×1 convolutions and activation functions are used for performing nonlinear mapping on the channel features to capture the dependencies between channels, and the final channel attention weights are obtained by normalization through the Sigmoid function.

$$A_c = Sigmoid(Conv_2(ReLU(Conv_1(GAP(Y))))) \in R^{N\times C\times 1\times 1} \tag{14}$$

Among them, $Conv_1$ and $Conv_2$ are 1×1 convolution operators for dimensionality reduction and dimensionality increase, respectively, which are used for compressing the amount of calculation and enhance the nonlinear expression ability. The attention weight will be used as a channel-level modulation factor in the subsequent steps to perform weighted enhancement on the feature map.

In the spatial attention branch, the average and maximum values of the features X_h along the height direction and X_w along the width direction are calculated in the channel dimension to extract the spatial response information from different statistical perspectives. These two representations are then concatenated in the channel dimension to form a fused representation, and the features are integrated through a 7×7 convolution operator to finally obtain the spatial attention weight.

$$\begin{aligned} A_h &= Sigmoid(Conv_{7\times 7}([Avg(X_h), Max(X_h)])) \\ A_w &= Sigmoid(Conv_{7\times 7}([Avg(X_w), Max(X_w)])) \end{aligned} \tag{15}$$

Among them, $[\cdot,\cdot]$ represents concatenation in the channel dimension, $\mathrm{Conv}_{7\times 7}$ is a convolution operation with a convolution kernel size of 7×7, and $\mathrm{Avg}(\cdot)$ and $\max(\cdot)$ represent average pooling and maximum pooling operations in the channel dimension, respectively. The final output attention weights A_h and A_w are used to adjust the spatial response strength along the height and width directions, respectively.

To further enhance the directional sensitivity of the feature, after obtaining the channel and spatial attention weights, independent 1×1 convolutions (respectively denoted as $Conv_h$ and $Conv_w$) are applied to the directional pooled features X_h and X_w to adjust their channel expression capabilities. Subsequently, the convolution transformed features are weighted fused element by element with the channel attention weight A_c and the spatial attention weights A_h and A_w to obtain the final directional attention response:

$$\begin{aligned} a_h &= Sigmoid(Conv_h(X_h) \odot A_c \odot A_h) \\ a_w &= Sigmoid(Conv_w(X_w) \odot A_c \odot A_w) \end{aligned} \tag{16}$$

Among them, $\odot$ represents element-wise multiplication, and $Sigmoid(\cdot)$ is used to normalize the attention response values in each direction. The above operation effectively improves the model's responsiveness to key structural areas by fusing the attention information in the channel and spatial dimensions, especially when dealing with small targets with

blurred edges or weakened textures.

Finally, the original input features are weighted using the attention weights:

$$Y_{out} = X \odot a_h \odot a_w \tag{17}$$

CEAM enhances the network's capacity to represent target boundary information by combining edge enhancement and direction perception[32]. The module's learnable edge improvement branch extracts minor structural changes between the target and the background, improving the model's responsiveness in edge areas. Furthermore, CEAM does spatial direction-sensitive modeling, which involves compressing and fusing information in the horizontal and vertical directions to properly locate important positions. This structure is particularly well-suited for identifying small floating targets[33] with low contrast and weak edges, boosting feature representation discriminability and model perception, resulting in increased robustness and flexibility in complicated water surface situations.

IV. Experiment

A. Dataset

To completely assess the performance and robustness of the modified algorithm in the job of tiny surface target detection, this research used two representative public datasets: FloW-IMG[34] and WSODD[35]. The two complement each other in terms of target quantity, scene complexity, and environmental diversity, allowing the algorithm to be verified from many angles.

The FloW-IMG dataset was created by Oka Zhibo. It is the world's first floating item detection dataset created from the perspective of an unmanned ship, with an emphasis on rubbish detection jobs in real inland waters. The dataset consists of 2,000 high-resolution images with 5,271 accurately annotated target instances that cover difficult scenes such as complex lighting conditions, multi-angle perspectives, dense interweaving of multiple targets, and small targets at long distances. These photos accurately depict the distribution of floating items on the water's surface in various conditions, laying the groundwork for the adaptability and effectiveness of detection algorithms in real-world applications. The WSODD dataset broadens the test dimension by focusing on detection tasks for a variety of common tiny items on the water surface. WSODD contains 7,467 images, covering 14 types of typical floating objects on the water surface, with a total of 21,911 target instances annotated, indicating a significant increase in data scale and target type. The dataset includes three types of water settings: oceans, lakes, and small rivers, as well as a range of weather conditions such as sunny, cloudy, and foggy days, as well as three time periods: daytime, nighttime, and twilight, which considerably increases the environment's richness and complexity. At the same time, WSODD particularly includes demanding samples such as poor contrast, weak edges, severe occlusion, and long-distance tiny objects, offering a rigorous testing platform for assessing the algorithm's durability under harsh situations. Figure 4 depicts the distribution of large, medium, and small objects in the dataset.

In conclusion, the FloW-IMG and WSODD datasets complement each other in terms of scale, type, and scene diversity, providing comprehensive and credible data support for assessing the detection accuracy and environmental adaptability of the method suggested in this study.

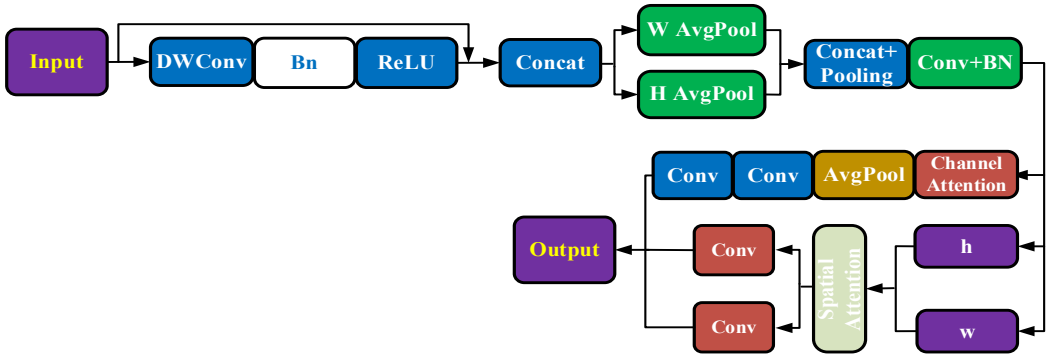


Fig. 3. CEAM module

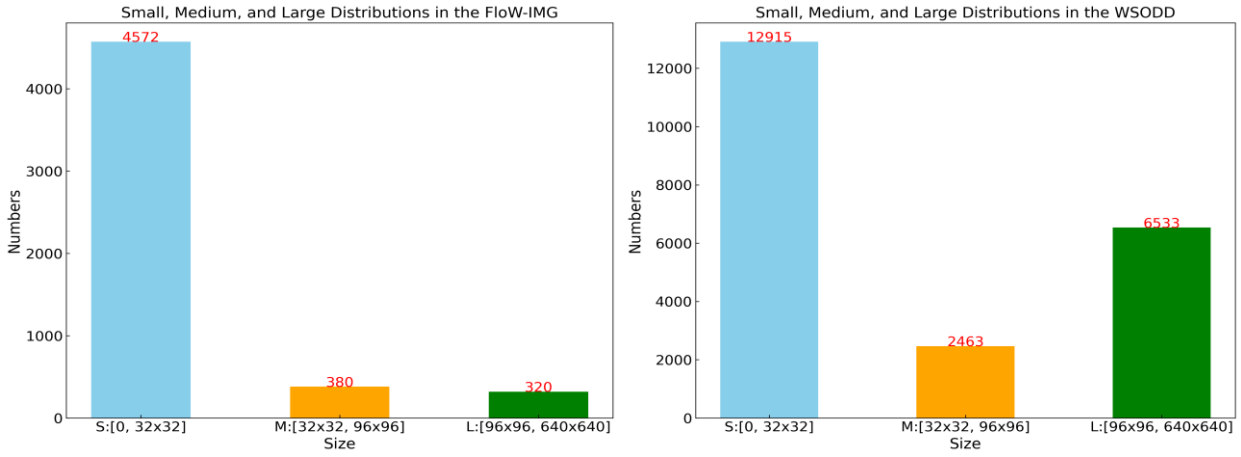


Fig. 4. Distribution of large, medium, and small objects in the FloW-IMG and WSODD datasets

B. Experimental parameters and evaluation indicators

This experiment was carried out using the Windows 11 operating system, with the deep learning framework Pytorch 2.4.1 and CUDA 12.4, and the hardware configuration was NVIDIA RTX 4060. The batch size was set to 16, the model was trained for 300 cycles, and the data was loaded using 4 threads. The FloW-img and WSODD datasets were split 8:1:1, respectively. Other hyperparameters remained unchanged and were set to their default values.

To effectively assess the improvement effect, numerous assessment markers were established, including recall (R) and mean average precision (mAP), parameters, and GFLOPs. The computation formula is given below:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{18}$$

$$mAP = \frac{\sum_{i=1}^{N} \int_0^1 PdR}{N} \times 100\% \tag{19}$$

Among them, TP represents true positives, FN represents false negatives, N is the number of categories, P is the precision, and $\int_0^1 PdR$ is the average precision (AP).

C. Ablation experiment

In order to completely test the proposed algorithm's detection performance and validate the function of each modified module in increasing model performance, we performed ablation experiments on the FloW-IMG dataset. First, we carried out experiments on the C3k2MSWTC module, as shown in Table 1, to investigate the impacts of the number of wavelet decomposition layers (wt_levels) and the attention mechanism hyperparameters (groups). On this basis, we froze the wavelet filter layers (wt_filter and iwt_filter) to avoid training these fixed mathematical transformation layers, reducing the computational cost and allowing us to focus on other adjustable hyperparameters. Specifically, wt_levels determines the number of wavelet decomposition layers, which influences the model's capacity to extract features at various scales. We examined the impact of different decomposition layers on performance by modifying this hyperparameter; on the other hand, groups represent the number of channel groupings, which influences the effect of the attention mechanism. By changing groups, we adjusted attention allocation between channels and increased the model's performance.

TABLE I
C3K2MSWTC ABLATION EXPERIMENT

wt_levels	groups	mAP@0.5	mAP@0.5: 0.95	Params
1	4	0.870	0.507	2.526M
2	4	0.864	0.506	2.531M
3	4	0.876	0.504	2.536M
1	8	0.881	0.517	2.522M
2	8	0.869	0.499	2.527M
3	8	0.874	0.512	2.532M
1	16	0.865	0.497	2.521M
2	16	0.870	0.501	2.526M
3	16	0.877	0.501	2.531M

Based on the preliminary trial findings of the C3k2MSWTC module, we finally chose the optimal configuration, the performance obtained with wt_levels:1, groups:8. To investigate the effect of the attention module on the overall network of the C3k2MSWTC module, we carried out a comparative experiment on the attention module alone, as shown in Table 2, to determine the particular contribution of each attention module to model performance.

TABLE II
COMPARATIVE EXPERIMENT OF ATTENTION MECHANISM

Attention	R	mAP@0.5	mAP@0.5:0.95	Params	GFLOPs
YOLOv11n	0.814	0.867	0.507	2.582M	6.3
+ CAA	0.780	0.848	0.485	2.767M	6.7
+ LIA	0.805	0.873	0.504	2.661M	6.3
+ ELA	0.797	0.869	0.503	2.586M	6.3
+ CCA	0.780	0.848	0.485	2.767M	6.7
+ CPCA	0.797	0.871	0.508	2.762M	7.0
+ SEAM	0.810	0.873	0.493	2.686M	6.5
+ MSDA	0.806	0.873	0.502	2.928M	6.9
+ SEAttention	0.818	0.874	0.507	2.593M	6.3
+ FCAttention	0.797	0.873	0.499	2.669M	6.4
+ TripletAttention	0.770	0.870	0.506	2.583M	6.4
+ MOCAttention	0.784	0.871	0.505	2.628M	6.3
+ CEAM(our)	0.831	0.873	0.507	2.682M	6.5

As indicated in Table 2, we carried out an attention mechanism ablation comparative experiment to better demonstrate the benefits of the CEAM attention mechanism. However, there are some disparities in the performance of various attention modules in crucial variables such as average precision and recall. It is worth mentioning that, while these attention mechanism comparison studies enhanced model performance to some extent, their benefits were not as significant as the CEAM attention mechanism we presented. The CEAM mechanism considerably enhanced the model's average precision and recall rate, as well as its anti-interference performance and detection stability. These results from the attention comparison experiments conclusively demonstrate that the CEAM attention mechanism is indispensable and critical in improving detection performance in complex water environments, particularly when dealing with interference factors such as water surface fluctuations and reflections.

The results in Tables 1 and 2 provide a detailed examination of the individual contributions of each enhanced module to overall detection performance. The experimental results clearly show that the introduction of a single module improves detection performance, not only by increasing key indicators such as recall rate and average precision, but also by demonstrating the synergistic effect that occurs when different modules are used together. The experimental results are presented in Table 3. This series of ablation tests offers an empirical basis for the optimization design and confirms the importance of each module in improving the model's capacity to resist interference and extract important information when detecting small targets on the surface of water.

TABLE III
ABLATION EXPERIMENT RESULTS

Model	R	mAP@0.5	mAP@0.5:0.95	Params	GFLOPs
YOLOv11n	0.814	0.867	0.507	2.582M	6.3
C3k2MSWTC	0.803	0.881	0.517	2.522M	6.3
CEAM	0.831	0.873	0.507	2.682M	6.5
YOLO-MC	0.835	0.890	0.511	2.622M	6.5

When the C3k2MSWTC module is used alone, mAP@0.5 improves by 1.4%, and mAP@0.5:0.95 improves by 1%, implying that the positioning accuracy and overall accuracy of the detection frame have improved, but the recall rate R is reduced by 1.1%, indicating that some real targets are not detected. This phenomenon indicates that, while improving detection accuracy and anti-interference ability, the model may have adopted stricter judgment criteria for some edge or fuzzy targets, resulting in a degree of missed detection. When the CEAM module is used alone, the recall rate R improves by 1.7%, mAP@0.5 improves by 0.6%, and mAP@0.5:0.95 remains constant, indicating that the improved module has played a positive role in improving the robustness of small target detection, covering more targets, and improving rough detection performance, but there is still room for improvement in high-precision positioning capabilities. The YOLO-MC structure proposed in this article improves the recall rate R by 2.1%, mAP@0.5 by 2.3%, and mAP@0.5:0.95 by 0.4%. Although the number of parameters increases by 0.04M and the computational complexity by 0.2 GFLOPs, the model maintains excellent real-time detection performance and improves robustness in complex environments thanks to efficient module optimization.

D. Visual analysis

To easily demonstrate the difference in detection ability between YOLOv11n and YOLO-MC, we conducted a thorough discussion of the experimental data using visual analysis. Figures 5 and 6 depict the recall rate R and mAP@0.5 convergence curves obtained during the YOLO-MC training and verification processes, respectively. The horizontal axis displays the number of iterations, and the experiment runs for 300 rounds. Observing the validation set curve of YOLO-MC on the Flow-IMG dataset, it is clear that YOLO-MC outperforms YOLOv11n in mAP indicators, and the recall rate curve converges after 240 iterations, with the recall rate R value of YOLO-MC stabilizing at roughly 83%.

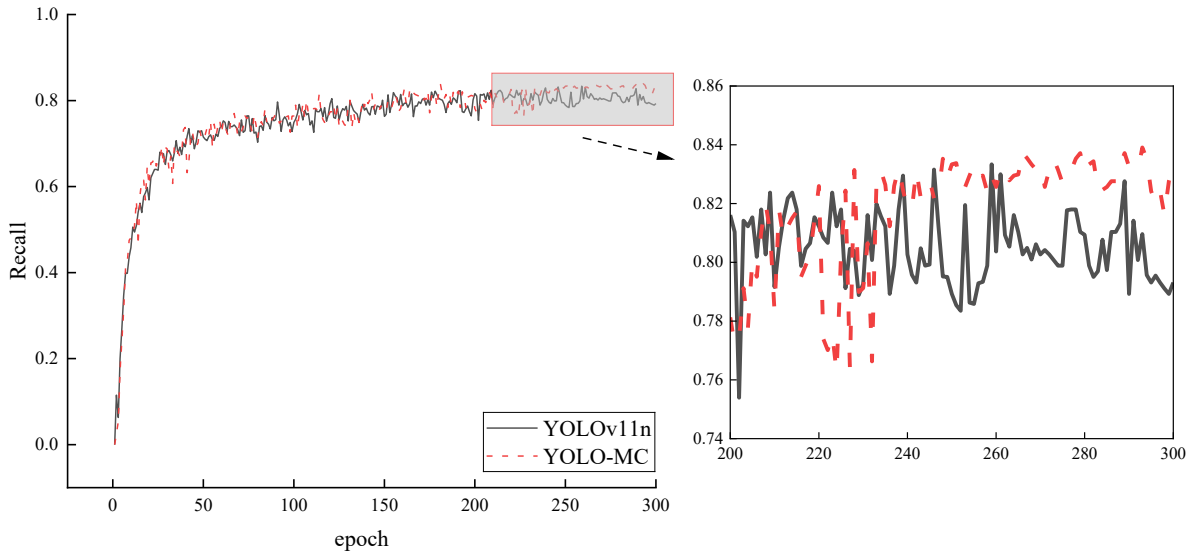


Fig. 5. Recall rate

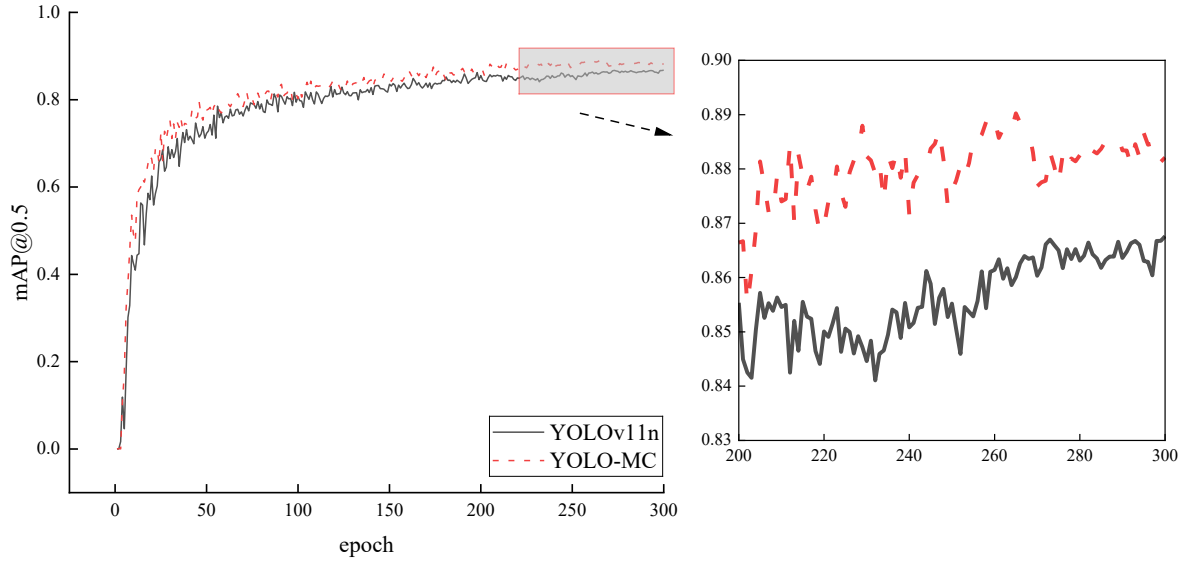
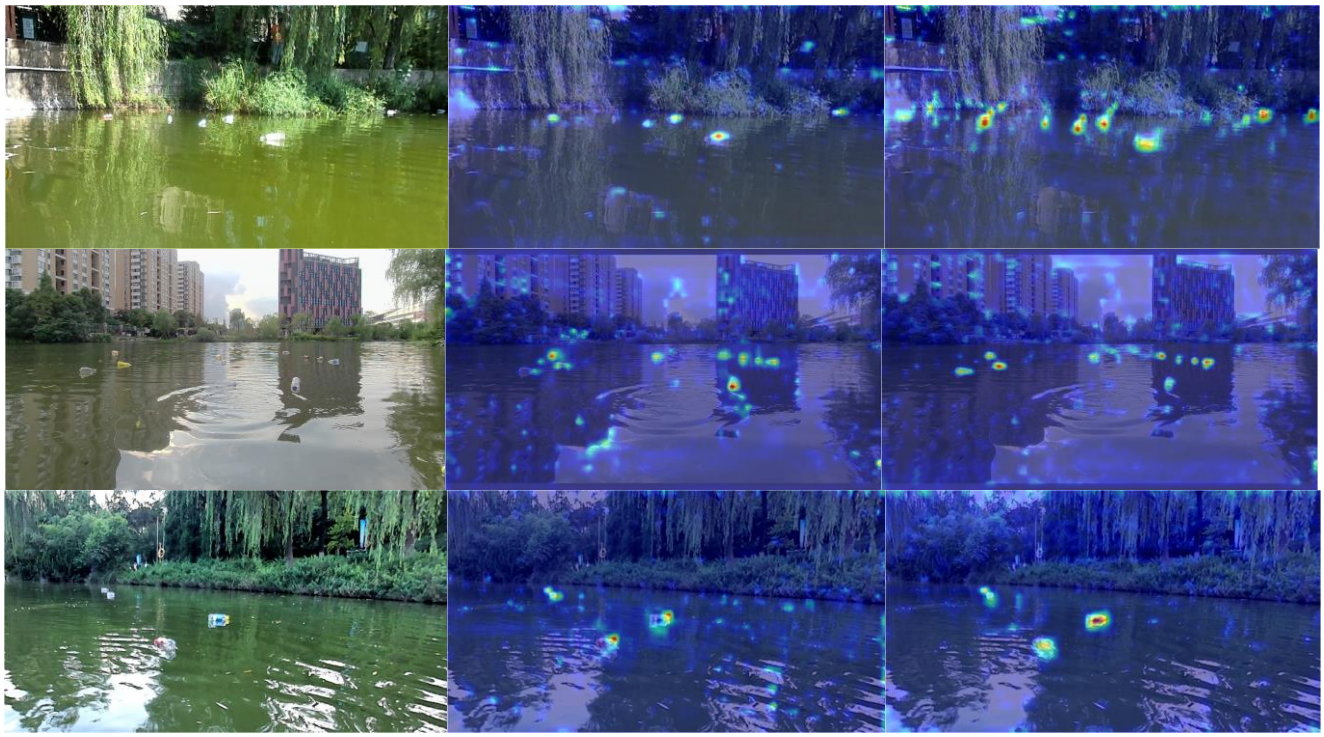


Fig. 6. mAP@0.5



Original image YOLOv11 YOLO-MC

Fig. 7. Comparison of heat maps of different algorithms

Furthermore, we depict the detection results as a heat map to more intuitively highlight how the algorithm described in this research optimizes target attention throughout the detection process, as shown in Figure 7. The heat map clearly shows the algorithm's level of attention to different locations throughout processing, particularly the recognition and positioning of significant objects. The heat map visually represents the model's attention intensity in various areas through color changes, with blue and green indicating lower attention or a weaker activation value, and yellow and red indicating higher attention or a stronger activation value. This color-level contrast can draw the model's attention to the target area, effectively measuring the algorithm's detection ability and accuracy in complex backdrops.

E. Comparative experiment

To evaluate the efficacy of YOLO-MC in identifying diminutive drifting objects on intricate horizontal surfaces, we chose several established and popular target detection algorithms, including YOLOv5n, YOLOv6n, YOLOv8n, YOLOv9-tiny, YOLOv10n, YOLOv10s, YOLOv11n, YOLOv11s, YOLOv12n, and YOLOv12s, for comparative analysis while maintaining a consistent experimental environment and the FloW-IMG dataset. Table 4 summarizes the experimental data.

TABLE IV
COMPARATIVE EXPERIMENTAL RESULTS OF DIFFERENT ALGORITHMS

Model	R	mAP@0.5	mAP@0.5:0.95	Params	GFLOPs
YOLOv5n	0.799	0.870	0.506	2.503M	7.1
YOLOv6n	0.828	0.871	0.500	4.234M	11.8
YOLOv8n	0.826	0.881	0.508	3.006M	8.1
YOLOv9-tiny	0.793	0.870	0.508	1.971M	7.6
YOLOv10n	0.805	0.868	0.501	2.695M	8.2
YOLOv10s	0.785	0.883	0.516	8.036M	24.4
YOLOv11n	0.814	0.867	0.507	2.582M	6.3
YOLOv11s	0.824	0.884	0.528	9.413M	21.3
YOLOv12n	0.785	0.861	0.488	2.557M	6.3
YOLOv12s	0.816	0.881	0.513	9.231M	21.2
YOLO-MC(our)	0.835	0.890	0.511	2.622M	6.5

Table 4 shows that, in the detection experiment using the FloW-IMG small target floating object dataset, the YOLO-MC algorithm presented in this study achieved a detection accuracy of 0.890 in the mAP@0.5 metric, outperforming many prominent comparative algorithms.In the recall rate R comparison, YOLO-MC outperformed the other ten models tested, with a significantly lower missed detection rate, proving its robustness and efficacy in small target identification tasks. While YOLOv10s, YOLOv11s, and YOLOv12s outperform YOLO-MC in the more strict mAP@0.5:0.95 assessment criteria, their larger parameter scale and computational complexity significantly limit their practical application in resource-constrained environments. In comparison to earlier lightweight models such as YOLOv5n, YOLOv6n, YOLOv8n, YOLOv9-tiny, YOLOv10n, YOLOv11n, and YOLOv12n, YOLO-MC achieves greater detection accuracy while maintaining reduced model complexity, indicating a successful performance balance. In conclusion, YOLO-MC has struck an optimum balance between detection accuracy, computing efficiency, and deployment feasibility, demonstrating its application potential and practical utility in micro surface target identification.

F. Verification experiment

To improve the reliability of the improved algorithm proposed in this paper, it is verified on the WSODD dataset. There are 14 categories of labels in this dataset, namely boat, ship, ball, bridge, rock, person, rubbish, mast, buoy, platform, harbor, tree, grass, and animal, which covers all common target types on the water surface and is a common training and evaluation dataset for small target detection.

As shown in Table 5, when compared to the original YOLOv11n model, the upgraded model has a 3.1% improvement in recall rate and a 1.1% rise in mAP@0.5 on the WSODD dataset. At the same time, the suggested model has a superior recognition effect on all detection targets in the dataset, demonstrating that this strategy increases recall while simultaneously improving the model's capacity to detect small targets.

TABLE V
COMPARATIVE EXPERIMENTAL RESULTS ON THE WSODD DATASET

Class	YOLOv11n			YOLO-MC		
	R	mAP@0.5	mAP@0.5:0.95	R	mAP@0.5	mAP@0.5:0.95
boat	0.824	0.903	0.527	0.826	0.903	0.533
ship	0.894	0.930	0.669	0.889	0.923	0.665
ball	0.530	0.677	0.253	0.545	0.697	0.250
bridge	0.941	0.972	0.714	0.956	0.971	0.710
rock	0.653	0.728	0.327	0.711	0.743	0.348
person	0.520	0.574	0.288	0.553	0.598	0.258
rubbish	0.623	0.726	0.404	0.689	0.739	0.405

CONTINUED TABLE V
COMPARATIVE EXPERIMENTAL RESULTS ON THE WSODD DATASET

Class	YOLOv11n			YOLO-MC		
	R	mAP@0.5	mAP@0.5:0.95	R	mAP@0.5	mAP@0.5:0.95
mast	0.667	0.663	0.354	0.574	0.639	0.312
buoy	0.786	0.872	0.552	0.852	0.867	0.531
platform	0.786	0.859	0.543	0.857	0.878	0.586
harbor	0.855	0.917	0.560	0.837	0.904	0.558
tree	1.000	0.983	0.608	0.950	0.941	0.589
grass	0.500	0.636	0.470	0.500	0.514	0.359
animal	0.240	0.310	0.073	0.508	0.594	0.210
all	0.701	0.768	0.453	0.732	0.779	0.451

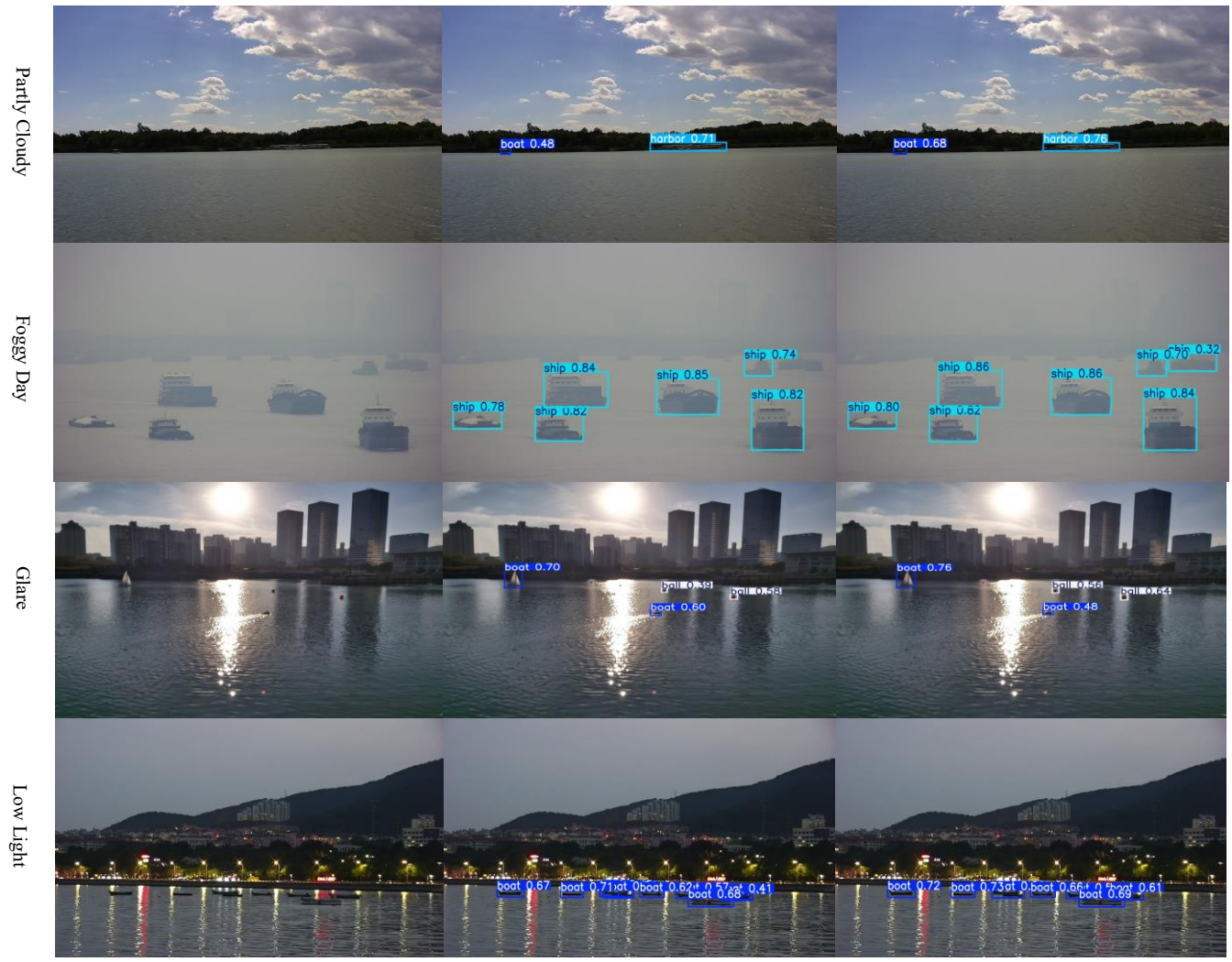


Fig. 8. Detection comparison of different algorithms

To completely examine the practical usefulness of the proposed YOLO-MC algorithm on the WSODD water surface small target detection dataset, we did a thorough evaluation of the detection performance in realistic complex

conditions. We chose a varied selection of authentic real water surface landscapes as test examples. These sceneries, as seen in Figure 8, include detailed circumstances such as partly cloudy, foggy days, glare, and low light and are intended to meet a range of normal visual interference settings. In such cases, target visibility is typically reduced and edge information is easily occluded, providing substantial challenges for the detecting system. As a result, these test scenarios not only meet real-world application requirements but also give a full evaluation of the model's robustness and generalization capabilities in complex conditions.

The figure 8 displays the comparison of target identification outcomes between the YOLOv11n basic model and the model developed in this paper under varied environmental settings. YOLO-MC displays improved resilience and adaptability compared to the original model, accurately identifying and finding small floating objects in adverse settings, including blurring target edges, harsh reflections on water surfaces, and poor lighting. This result further verifies the effectiveness of the introduced C3k2MSWTC module and CEAM attention mechanism in dealing with complex lighting and dynamic water surface background interference.

V. Conclusion

This paper proposes an improved YOLO-MC algorithm to significantly enhance the detection performance of small floating objects on the water, addressing the problem of insufficient detection accuracy caused by factors such as changes in illumination and fluctuations of the water surface in complex environments. The algorithm first designs the C3k2MSWTC module, which decomposes the input image in the frequency domain through wavelet transform and multiscale convolution strategy, fully extracts the low-frequency structural information and high-frequency detail features of the target, and thereby optimizes the feature expression; at the same time, the CEAM module is used to further enhance the model's ability to capture the target's texture and edge information. Through the integrated optimization of the above modules, the proposed YOLO-MC algorithm shows obvious advantages in the detection of small floating objects on the water.

Experimental results on the FloW-IMG and WSODD datasets show that, when compared to existing mainstream algorithms, the YOLO-MC algorithm has significantly improved detection accuracy and robustness, particularly when dealing with small floating objects on the water. The model also performs exceptionally well in capturing detailed information and resisting environmental interference. These findings present an efficient and robust approach for detecting small targets in complex aquatic environments, as well as a solid theoretical and practical framework for future technology enhancement and implementation.

References

[1] V. Viswanatha, R. K. Chandana, A. C. Ramachandra, “Real time object detection system with YOLO and CNN models: A review,” Journal of Xi'an University of Architecture and Technology, vol. 14, no. 7, pp. 144-151, 2022.

[2] T. Shehzadi, D. Stricker, M. Z. Afzal, “Semi-Supervised Object Detection: A Survey on Progress from CNN to Transformer,” arXiv preprint arXiv:2407.08460, 2024.

[3] Y. Zhao, W. Y. Lv, L. Xu, J. M. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, “Detrs beat yolos on real-time object detection,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16965-16974, 2024.

[4] X. Zhu, W. J. Su, L. W. Lu, B. Li, X. G. Wang, J. F. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” arXiv preprint arXiv:2010.04159, 2020.

[5] X. Y. Dai, Y. P. Chen, J. W. Yang, P. Zhang, L. Yuan, L. Zhang, “Dynamic detr: End-to-end object detection with dynamic attention,” Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2988-2997, 2021.

[6] W. Sheng, X. F. Yu, J. Y. Lin, X. Chen, “Faster rcnn target detection algorithm integrating cbam and fpn,” Applied Sciences, vol. 13, no. 12, pp. 6913-6930, 2023.

[7] H. L. Wang, H. M. Qian, S. Feng, W. N. Wang, “L-SSD: lightweight SSD target detection based on depth-separable convolution,” Journal of Real-Time Image Processing, vol. 21, no. 2, pp. 33-47, 2024.

[8] A. Wang, H. Chen, L. H. Liu, K. Chen, Z. J. Lin, J. G. Han, G. Ding, “Yolov10: Real-time end-to-end object detection,” Advances in Neural Information Processing Systems, vol. 37, pp. 107984-108011, 2024.

[9] R. Khanam, M. Hussain, “Yolov11: An overview of the key architectural enhancements,” arXiv preprint arXiv:2410.17725, 2024.

[10] J. Luo, Z. Liu, Y. B. Wang, A. Tang, H. Zuo, P. Han, “Efficient Small Object Detection You Only Look Once: A Small Object Detection Algorithm for Aerial Images,” Sensors, vol. 24, no. 21, pp. 7067-7088, 2024.

[11] X. Xiao, X. R. Xue, Zh. Zhao, Y. Fan, “A recursive prediction-based feature enhancement for small object detection,” Sensors, vol. 24, no. 12, pp. 3856-3871, 2024.

[12] Z. Y. Song, Y. Zhang, Y. Liu, K. H. Yang, M. L. Sun, “MSFYOLO: Feature fusion-based detection for small objects,” IEEE Latin America Transactions, vol. 20, no. 5, pp. 823-830, 2022.

[13] F. Zhao, J. Zhang, G. Zhang, “FFEDet: fine-grained feature enhancement for small object detection,” Remote Sensing, vol. 16, no. 11, pp. 2003-2024, 2024.

[14] C. Deng, M. Wang, L. Liu, Y. Liu, “Extended feature pyramid network for small object detection,” IEEE Transactions on Multimedia, vol. 24, pp. 1968-1970, 2021.

[15] F. X. Chen, L. X. Zhang, S. Y. Kang, L. T. Chen, H. Dong, D. Li, X. Wu, “Soft-NMS-enabled YOLOv5 with SIOU for small water surface floater detection in UAV-captured images,” Sustainability, vol. 15, no. 14, pp. 10751-10768, 2023.

[16] C. Shi, M. Lei, W. Q. You, H. T. Ye, H. Sun, “Enhanced floating debris detection algorithm based on CDW-YOLOv8,” Physica Scripta, vol. 99, no. 7, pp. 076019-076033, 2024.

[17] R. Chen, J. Wu, Y. Peng, Z. Li, H. Shang, “Detection and tracking of floating objects based on spatial-temporal information fusion,” Expert Systems with Applications, vol. 225, pp. 120185-120209, 2023.

[18] L. Zhang, Y. X. Wei, H. B. Wang, Y. H. Shao, J. Shen, “Real-time detection of river surface floating object based on improved refinedet,” IEEE Access, vol. 9, pp. 81147-81160, 2021.

[19] X. Yang, Y. Song, L. He, H. Xue, Z. Dong, and Q. Zhang, “USV-YOLO: An Algorithm for Detecting Floating Objects on the Surface of an Environmentally Friendly Unmanned Vessel,” IAENG International Journal of Computer Science, vol. 52, no. 3, pp. 579-588, 2025.

[20] W. Du, X. Ouyang, N. Zhao, and Y. Ouyang, “BCS-YOLOv8s: A Detecting Method for Dense Small Targets in Remote Sensing Images Based on Improved YOLOv8s,” IAENG International Journal of Computer Science, vol. 52, no. 2, pp. 417-426, 2025.

[21] S. E. Finder, R. Amoyal, E. Treister, O. Freifeld, “Wavelet convolutions for large receptive fields,” European Conference on Computer Vision, pp. 363-380, 2024.

[22] M. Elsayed, M. Reda, A. S. Mashaly, A. S. Amein, “LERFNet: an enlarged effective receptive field backbone network for enhancing visual drone detection,” The Visual Computer, vol. 41, no. 4, pp. 2219-2232, 2025.

[23] Z. Lin, B. Leng, “SSN: Scale Selection Network for Multi-Scale Object Detection in Remote Sensing Images,” Remote Sensing, vol. 16, no. 19, pp. 3697-3719, 2024.

[24] C. Zhang, L. J. Liu, X. Zang, F. Liu, H. Zhang, X. Y. Song, J. D. Chen, “Detr++: Taming your multi-scale detection transformer,” arXiv preprint arXiv:2206.02977, 2022.

[25] Z. Jiang, B. J. Wu, L. Ma, H. W. Zhang, J. Lian, “APM-YOLOv7 for Small-Target Water-Floating Garbage Detection Based on Multi-Scale Feature Adaptive Weighted Fusion,” Sensors, vol. 24, no. 1, pp. 50-71, 2023.

[26] L. Dang, G. Liu, Y. Hou, and H. Han, "YOLO-FNC: An Improved Method for Small Object Detection in Remote Sensing Images Based on YOLOv7," IAENG International Journal of Computer Science, vol. 51, no. 9, pp. 1281-1290, 2024.
[27] Y. Su, W. X. Tan, Y. F. Dong, W. Xu, P. Huang, J. X. Zhang, D. K. Zhang, "Enhancing concealed object detection in Active Millimeter Wave Images using wavelet transform," Signal Processing, vol. 216, pp. 109303-109315, 2024.
[28] J. Pan, and Y. Zhang, "Small Object Detection in Aerial Drone Imagery based on YOLOv8," IAENG International Journal of Computer Science, vol. 51, no. 9, pp. 1346-1354, 2024.
[29] Q. Hou, D. Zhou, J. Feng, "Coordinate attention for efficient mobile network design," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713-13722, 2021.
[30] L. B. Li, R. P. Wang, M. J. Zou, F. S. Guo, Y. H. Ren, "Enhanced ResNet-50 for garbage classification: Feature fusion and depth-separable convolutions," Plos One, vol. 20, no. 1, pp. e0317999-e0318019, 2025.
[31] L. Li, B. Li, H. Zhou, "Lightweight multi-scale network for small object detection," PeerJ Computer Science, vol. 8, pp. e1145-e1170, 2022.
[32] R. Dong, S. Yin, L. Jiao, J. G. An, W. J. Wu, "ASIPNet: Orientation-Aware Learning Object Detection for Remote Sensing Images," Remote Sensing, vol. 16, no. 16, pp. 2992-3014, 2024.
[33] F. Lin, T. Hou, Q. N. Jin, A. J. You, "Improved YOLO based detection algorithm for floating debris in waterway, Entropy," vol. 23, no. 9, pp. 1111-1124, 2021.
[34] Y. W. Cheng, J. N. Zhu, M. X. Jiang, J. Fu, C. Pang, P. D. Wang, "Flow: A dataset and benchmark for floating waste detection in inland waters," Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10953-10962, 2021.
[35] Z. Zhou, J. Sun, J. B. Yu, K. Y. Liu, J. W. Duan, L. Chen, C. Chen, "An image-based benchmark dataset and a novel object detector for water surface object detection," Frontiers in Neurorobotics, vol. 15, pp. 723336-723349, 2021.