International Journal of Innovative Computing, Information and Control
Volume **22**, Number **2**, April **2026**

ICIC International ©2026 ISSN 1349-4198
pp. **457–474**

FLOATING OBJECT DETECTION METHOD FOR WATER SURFACE IN COMPLEX SCENES

Guangmiao Zeng[1,3], Rongjie Wang[2] and Wanneng Yu[2,*]

[1]School of Navigation
Xiamen Ocean Vocational College
No. 61, Sports Road, Siming District, Xiamen 361012, P. R. China
zengguangmiao@xmoc.edu.cn

[2]School of Marine Engineering
Jimei University
No. 176, Shigu Road, Jimei District, Xiamen 361021, P. R. China
wangrongjie@jmu.edu.cn; *Corresponding author: wnyu2007@jmu.edu.cn

[3]Applied Technology Engineering Center of Fujian Provincial Higher Education
for Marine Resource Protection and Ecological Governance
No. 4566, Hongzhong Avenue, Xiang'an District, Xiamen 361100, P. R. China

Received April 2025; revised August 2025

Abstract. *Plastic-based marine wastes threaten ecosystem health and disrupt port and inland-water operations. To enable scalable cleanup, unmanned surface vessels (USVs) require real-time onboard perception for interception and collection under low camera vantage, dynamic water surfaces, and limited onboard compute. An attention-enhanced water-surface floating-object recognition algorithm is introduced that strengthens responses to true targets, down-weights distractors, and achieves accurate recognition in complex scenes with light spot reflection, insufficient light and the presence of interfering objects. Real-world experiments show that the improved algorithm outperforms the original Yolov7 baseline by +1.9 mAP@[0.5:0.95] and +1.6 mAR@[0.5:0.95], and remains competitive with recent SOTA detectors, achieving +0.9 mAP@[0.5:0.95] and a 4.2× speedup over Cascade R-CNN.*

Keywords: Object recognition, Small targets, Floating objects, Attention mechanism, Deep learning

1. **Introduction.** Marine wastes pollution has become a global environmental problem, both in inland rivers and in the ocean, which seriously threatens the ecological balance. Using the northern South China Sea as a regional case, modeled estimates based on mismanaged plastic waste suggest an annual plastic input to the ocean of ∼2.56-7.08 million tonnes (model estimates referenced to the 2010 baseline year). In ship-based surveys of floating marine macro-litter (FMML; items $>$ 2.5 cm) conducted during boreal spring-summer of 2019-2021, plastics accounted for 72.0% of items (by count); the density of anthropogenic FMML was approximately 118.7 ± 86.2 items km^{-2}. This share lies in the upper-middle range of globally reported FMML composition (plastics 34.8%-99.0% by item count), suggesting that the South China Sea is among the global hotspots of FMML pollution [1-3]. The impact of plastic on marine ecology can be mitigated if as much of this plastic wastes are removed as possible, as shown in Figure 1 for some of the surface wastes floats and their impact on marine life.

DOI: 10.24507/ijicic.22.02.457

FIGURE 1. Some surface floating objects and their impact on marine life

For these marine wastes, many coastal countries have paid more attention and carried out some degree of clean-up work, but still mainly by manual cleaning. The inefficiency and high cost of manual cleaning is still an important issue in the face of the large extent and quantity of marine wastes. Therefore, automatic cleaning solutions incorporating multiple sensors have been designed for wastes cleanup, which first analyze remote sensing image information to determine the distribution of floating objects on the water surface and then give different levels of pollution on the water surface [6], and later search for polluted areas using drones and capture targets that are difficult to collect manually [7]. For areas with a larger number of wastes floating in the distribution, unmanned boats with cleaning robots can be used to go and perform autonomous cleaning operations [8]. For underwater wastes, unmanned underwater vehicles can also be used to collect them [9]. These unmanned devices are able to do their work efficiently without the deep learning technology that has developed rapidly in recent years. Nowadays, this technology has been better developed in the field of marine environment perception, and the intelligent technology of unmanned robots has become more and more perfect, and it has become a trend to use unmanned equipment instead of manual work for marine wastes cleaning [10].

However, whether it is an unmanned aircraft, an unmanned boat or an unmanned underwater vehicle, the prerequisite for target cleaning operations is the accurate identification of the objects to be cleaned, which requires a reliable automatic identification system to achieve the detection of the objects. Deep learning-based object recognition technology, which has been widely used in many fields, can provide several feasible and reliable solutions for accurate and real-time detection of objects to be cleaned.

Compared to large surface objects, it is more difficult to detect for small objects. These small floating objects also exist in different kinds such as wood blocks, aquatic plants, plastic bottles, and plastic bags. Among them, only plastic products are the floating waste objects that need to be salvaged by unmanned boats.

In this paper, we take floating objects on the water surface as the research object. Small floating objects, such as plastic bottles, are almost impossible to be distinguished from interfering objects by radar [11,12] because of their small size, transparent color, and easy integration into the background; instead, they can be captured by features using visual sensors [13,14]. Therefore, a target detection algorithm based on visible sensors combined with visual information is needed to design a recognition task for small targets of floating objects. The recognition algorithm should not only achieve high recognition accuracy, but also needs high operation speed to meet the unmanned boat for autonomous cleaning operations.

The complexity and variability of the water surface environment, such as the reflection of light spots on the water surface, insufficient light and interfering objects are shown in Figure 2, which has a large impact on the recognition system and leads to a decrease in the object recognition accuracy. Therefore, it is necessary to design a water surface floating small object recognition method to meet different complex scenes.

(a) (b) (c)

FIGURE 2. Example diagrams of floating objects on the water in complex scenes: (a) Light spot reflection; (b) insufficient light; (c) interfering objects

In addition, today's small object recognition algorithms are studied based on public datasets, such as the COCO dataset [15] and the PASCAL VOC dataset [16]. However, there are still differences between the styles of small objects in public datasets and floating object targets on the water surface, and many excellent small object recognition algorithms are not optimized for water surface application scene. In this paper, we use the FloW dataset [17] for training and testing, which is a dataset made for floating objects on water, and verify the advantages of the algorithm proposed in this paper by comparing and evaluating multiple algorithms.

This work has several contributions.

First, a water floating object recognition algorithm incorporating attention mechanism is proposed.

Second, recognition tests are conducted for different scenes of floating objects on the water, and the recognition accuracy and recall rates are improved.

Third, the test is conducted using an offline small portable platform, which meets the requirement of implementing real-time detection on a shipboard platform.

2. Related Works.

2.1. Sea surface object detection method.

At present, among the sea surface object detection methods, there are more methods for ship object detection and they are mainly improved based on two types of classical algorithms, namely, the two-stage object detection algorithm Faster R-CNN [18] and the one-stage object detection algorithm Yolo-v3 [19] mainly. For example, generative adversarial networks are applied to training data augmentation [20], anchor frame and localization uncertainty of bounding boxes are optimized [21], redesigned for data augmentation methods [22]. These methods have improved the recognition accuracy of sea surface objects to some extent by improving the algorithmic network structure and the data enhancement of the input images.

These sea surface object detection methods are not only implemented based on unmanned boat platforms, but also there exists research on the use of UAVs for sea surface object identification. For example, the use of thermal imaging technology combined with machine vision to achieve detection and tracking of sea surface targets [23], the use of satellite sensing data combined with recursive area clustering to achieve sea surface object detection and trajectory prediction [7], and so on, have improved the accuracy of recognition of distant ship objects. Although there is less research on small objects such as floating objects in the sea surface object recognition methods, there are abundant detection methods for ship objects that are in the same environment, which has some significance for the research on floating object recognition in the marine background environment.

2.2. **Small object detection method.** Accurate recognition of small objects has long been a difficult problem in computer vision and a major focus of researchers. At present, small object recognition mostly focuses on improving the accuracy of recognition as the research focus.

In pedestrian small object detection, the pedestrian detection accuracy is improved by combining dense cropping strategy and local attention mechanism for the difficult problem of dense and few features of pedestrian objects in high altitude view [24]. In vehicle small object detection, the vehicle detection accuracy is improved by combining the direction-aware rotating bounding box with a multi-parameter regression model for the difficulty that the vehicle targets are in shadow and obscured from each other [25]. In terms of ship small object detection, the accuracy of ship detection was improved by combining image reconstruction with classification attention to address the difficulties of motion blur and fog interference in sea surface scenes [26]. A deep learning–based algorithm that is robust to clutter, occlusion, and illumination variation has been proposed, which is relevant to small object detection under challenging water surface conditions [27].

Whether in land scenes or water scenes, in order to improve the object detection accuracy, the algorithm is often designed to be more complex, so it does not gain advantages in the calculation speed. In the water scene, the detection method for small object of ship should not only focus on improving the detection accuracy, but also need some consideration for the detection speed.

In terms of object detection, most of the surface objects use small targets of ships as the research object, and the recognition algorithm for small targets of floating objects is less studied. The use of the Yolov3 algorithm for the recognition of floating objects on the water surface was proposed in [8] and deployed on an unmanned surface robot to achieve high-speed, high-precision object detection, but it did not take account of the presence of complex backgrounds (e.g., spot reflections, insufficient light, and interfering objects) on outdoor water surfaces.

In [17], the complex background condition of the presence of disturbances on the water surface was considered and a data set of surface floaters in multiple scenes was established, but no target detection algorithm was proposed that could achieve high-speed and high-precision target detection in complex scenes. Therefore, it is necessary to design a surface floating object detection algorithm that can be deployed on small unmanned boats and meet both fast and high accuracy requirements.

3. **Water Surface Floating Object Detection Method.**

3.1. **Network model structure design.** The Yolov7 algorithm is an object detection algorithm proposed in 2022, and the test results on the COCO dataset show that it outperforms various object detectors such as YOLOR, YOLOX, Scaled-YOLOv4, and YOLOv5 in terms of speed and accuracy. We further analyze the problem of detecting floating objects on the water surface and propose an object detection algorithm Yolo-FC that incorporates an attention mechanism. The high-level workflow is provided in Figure 3, where the input image is first pre-processed, features are extracted by the backbone and fused across scales by the FPN (feature pyramid network), the fused features are refined by the proposed attention-based Yolo-FC algorithm, and the detection head with NMS (non-maximum suppression) outputs the final floating-object predictions. The network structure of Yolo-FC is shown in Figure 4.

The upper half of Figure 4 shows the overall framework structure, which contains Backbone, Attention, FPN and Head layers, representing the framework structure of the network structure from shallow to deep in different stages of computation, which will be

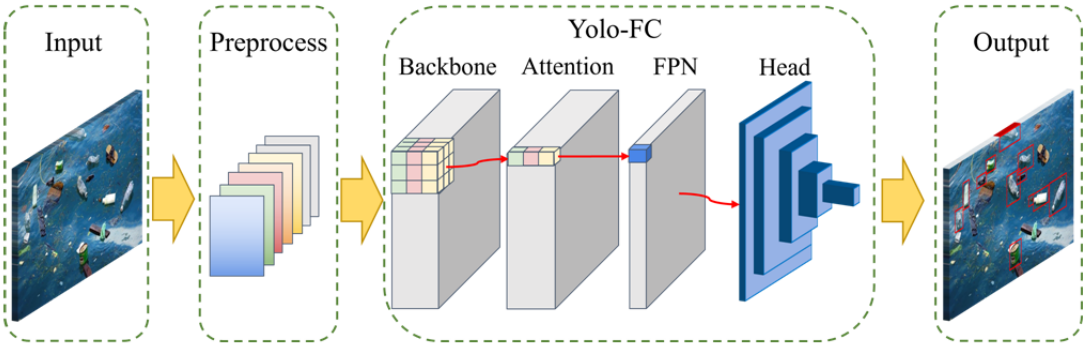

FIGURE 3. The high-level workflow of Yolo-FC algorithm

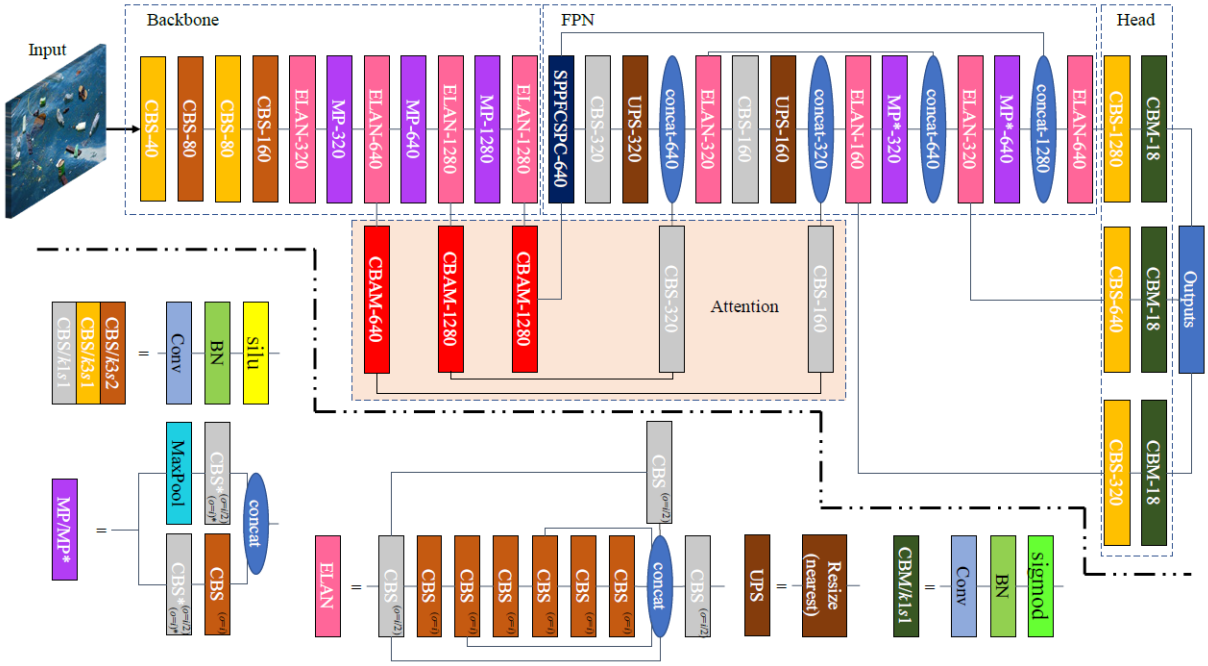

FIGURE 4. Network structure diagram of Yolo-FC algorithm

divided into four parts in turn for detailed elaboration later. The bottom half of Figure 4 shows how the different structural blocks are composed, where the CBS indicates the convolutional blocks, which all consist of a convolutional layer (Conv), a batch normalization layer (BN), and an activation function (Silu). The different colors of the CBS indicate the different sizes of the convolutional kernel k and step s of the convolutional layer. CBM also represents the convolutional block, unlike CBS, the activation function in CBM uses the Sigmod activation function, and the two activation functions are calculated as shown in Equations (1) and (2).

$$Silu(x) = x \cdot \frac{1}{1 - e^{-x}} \tag{1}$$

$$Sigmoid(x) = \frac{1}{1 - e^{-x}} \tag{2}$$

UPS denotes the upper sampling layer, which is computed using nearest neighbor interpolation (nearest). ELAN denotes the multi-branch stacking module, where concat

denotes the combined connection computation (concat), $o = i$ means the number of output channels is equal to the number of input channels, and $o = i/2$ means the number of output channels is equal to half of the number of input channels.

3.1.1. *Backbone.* In the backbone network, Yolo-FC uses the ELAN module for feature extraction and then uses the transition module for downsampling to obtain three effective feature layers for the next step of network construction.

In the ELAN module, the network divides the input features into 5 branches for computation, in the order of 1 convolutional block, 1 convolutional block, 3 convolutional blocks, 5 convolutional blocks and 7 convolutional blocks of branches, which are computed in concat and then passed through 1 convolutional block for output. The dense residual structure enables the fusion of features of 5 different depths and attenuates the effect of the gradient disappearance problem due to the increasing depth of the network by using jump-connected residual blocks.

In the MP module, the network divides the input features into 2 branches for computation, the first one is the maximum pooling with convolutional block, the second one is two convolutional kernels and convolutional blocks with different step sizes, and the subsequent output results are obtained by concatenating the two branches for computation.

3.1.2. *Attention.* After the feature extraction of the input image by the backbone network, the network will use the attention mechanism to improve the attention to the effective features, generate the attention information in both channel and space dimensions by the convolutional block attention module (CBAM), and combine them to produce a new feature map, and the network structure diagram of CBAM is shown in Figure 5. Channel attention determines which channels are emphasized, and spatial attention localizes informative positions. This combination suppresses water-surface reflections and ripples.

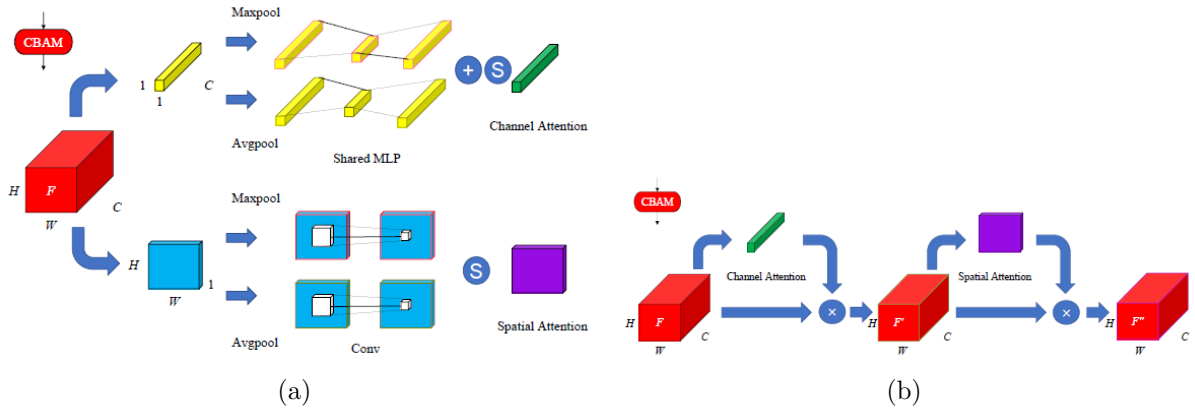

FIGURE 5. Flowchart of the algorithm for the CBAM attention mechanism: (a) The way to generate channel attention feature maps and spatial attention feature maps; (b) the operation process of the input feature map F

From Figure 5(a), it can be seen that for the network input feature map F, it is divided into two processes, which generate the channel attention feature map M_c and the spatial attention feature map M_s, respectively, as shown in Equations (3)-(5).

$$F \in \mathbb{R}^{C \times H \times W} \tag{3}$$

$$M_c \in \mathbb{R}^{C \times 1 \times 1} \tag{4}$$

$$M_s \in \mathbb{R}^{1 \times H \times W} \tag{5}$$

where C, H and W denote the number of channels, height and width of the feature map, respectively.

Figure 5(b) gives the process of performing operations on the feature map F. The feature map F' is first generated using the channel attention algorithm, and then the feature F'' is generated using the spatial attention algorithm map, as shown in Equations (6) and (7).

$$F' = M_c(F) \otimes F \tag{6}$$

$$F'' = M_s(F') \otimes F' \tag{7}$$

where $\otimes$ denotes the corresponding multiplication of congruent elements. This order first filters channels affected by reflections and then enforces spatial consistency.

In the channel attention module, the network focuses on the discriminative features of the image in the channel dimension by compressing the spatial dimension, and uses two methods of average pooling and maximum pooling for feature extraction, obtaining the overall features of the feature region by average pooling F^c_{avg} and combining with maximum pooling $F^c_{\max}$ to obtain the salient features of the feature region. After that, the final channel attention feature map M_c is obtained by using a weight-sharing multilayer perceptron (MLP) network for feature fusion, and the calculation process is shown in Equation (8). Average pooling encodes global context and max pooling captures peak activations. The shared MLP in Equation (8) suppresses spurious high-response channels from specular reflections and enhances object-relevant channels.

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))\right) \\ M_c &\in \mathbb{R}^{C/r\times 1\times 1},\ W_0 \in \mathbb{R}^{C/r\times C},\ W_1 \in \mathbb{R}^{C\times C/r},\ F^c_{avg} \in \mathbb{R}^{C\times 1\times 1},\ F^c_{\max} \in \mathbb{R}^{C\times 1\times 1} \end{aligned} \tag{8}$$

where σ denotes the Sigmoid activation function, W_0 and W_1 are the weights of the MLP, and r is the number of dimensionality reduction coefficients used in the MLP.

In the spatial attention module, the network focuses on the orientation features of the image in the spatial dimension by compressing the channel dimension, and also uses two methods of average pooling and maximum pooling for feature extraction, obtaining the overall features of the feature region by average pooling F^s_{avg} and combining with maximum pooling $F^s_{\max}$ to obtain the salient features of the feature region. After that, feature fusion is performed using convolutional layers to obtain the final spatial attention feature map M_s. The calculation process is shown in Equation (9).

$$\begin{aligned} M_s(F') &= \sigma\left(f^{7\times 7}([AvgPool(F'); MaxPool(F')])\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \\ M_s &\in \mathbb{R}^{H,W},\ F^s_{avg} \in \mathbb{R}^{1\times H\times W},\ F^s_{\max} \in \mathbb{R}^{1\times H\times W} \end{aligned} \tag{9}$$

where $f^{7\times 7}$ indicates that a convolution operation with a convolution kernel size of 7×7 is performed. The 7×7 convolution aggregates local context to suppress specular and emphasize spatially coherent object regions.

3.1.3. *FPN.* After feature enhancement by the attention mechanism network, the feature map F'' enters the feature pyramid stage for processing, and feature enhancement is performed by the improved spatial pyramid pooling optimization (ISPPCSPC) module. Compared with the spatial pyramidal pooling optimization (SPPCSPC) module, the ISPPCSPC module makes the network structure more efficient by using the same size convolutional kernel for the convolutional module instead of three different scale convolutional kernels and reusing the same convolutional module structure by tandem. The

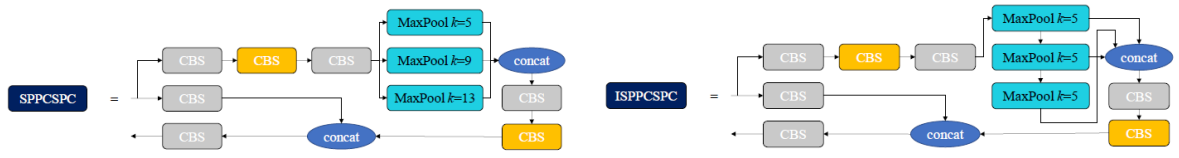

FIGURE 6. Schematic diagram of the network structure before and after the improvement of the SPPCSPC module: (a) The module before improvement; (b) the module after improvement

network structure before and after the improvement of the ISPPCSPC module is shown in Figure 6.

The network structure in the rest of the feature pyramid network (FPN) is shown in Figure 3, where the feature layers of different scales are fused after down-sampling and up-sampling operations of multiple convolutional blocks, so that the shallow and deep network features are mixed and better feature values are extracted.

3.1.4. *Head.* After the feature pyramid module three enhanced feature maps can be obtained, each feature layer has width, height and number of channels, and the network views the feature map as a collection of each feature point and uses three prior frames of different sizes to judge these feature points. After that, the prior frames contained in itself are adjusted according to the judgment feedback, and the non-maximum suppression method is used to identify and detect targets of different sizes in the original map, which improves the overall detection capability of the neural network for multi-scale targets.

3.2. **Network optimization methods.**

3.2.1. *Loss function.* The loss function of the Yolov7-FC algorithm contains three components: regression loss of the target ($Loss_{\text{reg}}$), classification loss ($Loss_{\text{cls}}$), and location loss ($Loss_{\text{loc}}$), as shown in Equations (10)-(14).

$$Loss = \lambda_{\text{reg}} \cdot Loss_{\text{reg}} + \lambda_{\text{cls}} \cdot Loss_{\text{cls}} + \lambda_{\text{loc}} \cdot Loss_{\text{loc}} \tag{10}$$

$$Loss_{\text{reg}} = \sum_{i=0}^{K\times K} \sum_{j=0}^{M} I_{ij}^{obj} Loss_{\text{BCE}} \left(\hat{C}_i, C_i\right) - \sum_{i=0}^{K\times K} \sum_{j=0}^{M} I_{ij}^{noobj} Loss_{\text{BCE}} \left(\hat{C}_i, C_i\right) \tag{11}$$

$$Loss_{\text{cls}} = \sum_{i=0}^{K\times K} I_{ij}^{obj} \sum_{k\in classes}^{K\times K} Loss_{\text{BCE}}(\hat{p}_i(k), p_i(k)) \tag{12}$$

$$Loss_{\text{BCE}} \left(\hat{N}, N\right) = \hat{N} \log(N) + \left(1 - \hat{N}\right) \log(1 - N) \tag{13}$$

$$Loss_{\text{loc}} = \sum_{i=0}^{K\times K} \sum_{j=0}^{M} I_{ij}^{obj} \cdot Loss_{\text{CIoU}} \tag{14}$$

where λ_{reg}, λ_{cls} and λ_{loc} represent the weights of three different categories of loss in the loss function, respectively, the Yolov7-FC network divides each input image into $K \times K$ cells first, and each grid produces M anchor boxes. After each anchor is subjected to the network's antecedent computation, an adjusted bounding box is obtained, and the total number of anchors is $K \times K \times M$. I_{ij}^{obj} and I_{ij}^{noobj} are used to determine whether the center coordinates of the target are in the jth anchor box in the ith grid, if yes the former is equal to 1 and the latter is equal to 0, otherwise the opposite. C_i is the confidence of the true box in the ith cell and $\hat{C}_i$ is the confidence of the prediction box in the ith cell. $p_i(k)$

denotes the conditional probability that the true box in the ith cell contains the kth type of target and $\hat{p}_i(k)$ denotes the conditional probability that the prediction box in the ith cell contains the kth type of target.

Meanwhile, the complete cross-merge ratio loss (CIoU) is used in the calculation of the location loss function, instead of the dichotomous cross-entropy loss used in the regression loss and classification loss, which can describe the location information more accurately. The complete cross-merge ratio loss is calculated as shown in Equations (15)-(19).

$$Loss_{\mathrm{CIoU}} = 1 - IoU + R_{\mathrm{CIoU}}\left(B, B^{gt}\right) \tag{15}$$

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{16}$$

$$R_{\mathrm{CIoU}}\left(B, B^{gt}\right) = \frac{\rho\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{17}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{18}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{19}$$

where IoU is the intersection ratio, the prediction box $B = (x, y, w, h)$, and the true box $B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$, which consist of x, y coordinates indicating the location of the center point and w, h coordinates indicating the width and height length. $R_{\mathrm{CIoU}}(B, B^{gt})$ is the penalty term between the prediction box B and the real box B^{gt}, b and b^{gt} represent the centroids of B and B^{gt}, $\rho(\bullet)$ denotes the Euclidean distance, and c is the diagonal distance of the smallest box that can contain both the prediction box and the real box. α is a positive trade-off parameter and v is a parameter that measures the consistency of the aspect ratio, which gives a higher priority to factors in the region where the predicted box overlaps with the true box relative to the non-overlapping part in the regression calculation.

In determining whether the prediction frame is a positive or negative sample, the sim-OTA algorithm is used to make the decision. The cost matrix *Cost* is obtained by combining the IoU loss $Loss_{\mathrm{reg}}$ between the prediction frame and the true frame with the category loss $Loss_{\mathrm{cls}}$ of the prediction frame and the true frame, as shown in Equation (20).

$$Cost = Loss_{\mathrm{cls}} + \varphi Loss_{\mathrm{reg}} \tag{20}$$

where the balance factor φ is set to 3 to balance the identification difficulty of the two losses.

From the cost matrix *Cost*, it can be seen that the higher the overlap between the real frame and the prediction frame the lower the cost, and the more accurate the classification the lower the cost, thus adaptively finding a few prediction frames with the best fit to the real frame.

After that, the N candidate frames with the largest $IoUs$ are selected according to the value of *Cost*, and a suitable number M of positive samples are assigned to different targets to be identified, as shown in Equation (21).

$$M = \left\lfloor \sum_{n=1}^{N} Cost(n) \times IoU(n) \right\rfloor \tag{21}$$

When there are multiple real boxes matched by the same candidate box, the real box with smaller *Cost* is selected as the only matching target.

In the early stage of training, using a large learning rate can make the network converge quickly, while in the later stage of training, using a small learning rate is more also beneficial for the network to converge to the optimal value. Therefore, the exponential decay strategy of learning rate is utilized for training, and the learning rate γ is calculated as shown in Equation (22).

$$\gamma = \varepsilon^{\tau}\gamma_0 \tag{22}$$

where γ_0 denotes the initial learning rate, ε is the decay rate, and τ is the number of iterations of the training network.

3.2.2. *Image pre-processing.* The mosaic image enhancement method (Mosaic) is a new data enhancement algorithm generated by expanding on the hybrid image enhancement (Mixup) method. Unlike the cut-and-mix method where two images are overlaid and fused, it instead uses four images to be cut and stitched to form a new image. This method can better enrich the background of the target and prevent the network from generalization degradation due to the similar background of the training set. However, since the distribution of its generated training images differs significantly from that of the natural images, it needs to be prohibited after z iterations to make the network deepen its understanding of the natural images.

In this paper, a combination of Mosaic and Mixup is used for image data enhancement, and the formula for determining whether data enhancement methods are used in each iteration is shown in Equation (23).

$$\begin{aligned} mosaic &= boll[rand(0,1) > \theta_1 \ \&\ epoch > z \cdot epoch] \\ mixup &= boll[rand(0,1) > \theta_2] \end{aligned} \tag{23}$$

where $boll(\ )$ denotes Boolean operation, & denotes with operation, the values of θ_1 and θ_2 are 0.5, and the value of z is 0.7.

4. **Experimental Results and Analysis.** The dataset used for training and testing in this paper is the FloW dataset, which contains a total of 2000 images with 1280×720 resolution and 5271 floaters, and the training and testing sets are divided into a $3:2$ manner, with 1200 images as the training set and 800 as the testing set. This follows the public dataset (FloW dataset) [17] to ensure direct comparability with its benchmark.

Small objects with pixel points smaller than 32×32 pixels have 2996 images in the dataset, which is more than 50% of the total number of objects. There are also 1974 medium objects with pixel points between 96×96 pixels and 32×32 pixels, which are more than 90% of the total number of small objects, and even the medium objects account for no more than 1% of the overall number of pixel points in the images, which is challenging for the recognition requirements of the objects detection algorithm. Given this small-object-dominated distribution, the larger test portion provides a stricter and more stable evaluation. Therefore, the absolute AP/AR reported here are conservative estimates. The distribution of small and large objects in the data is shown in Figure 7.

The algorithms in this paper are implemented on the open source neural network framework Pytorch (1.10.1). The computational workstation configuration contains 1 GPU (GeForce RTX 3090), CPU (AMD Ryzen 9 3950x 16 Core/ 3.5 GHz/72 M), and 128 G RAM. The small portable test platform is built on the NVIDIA Jetson AGX Orin development board.

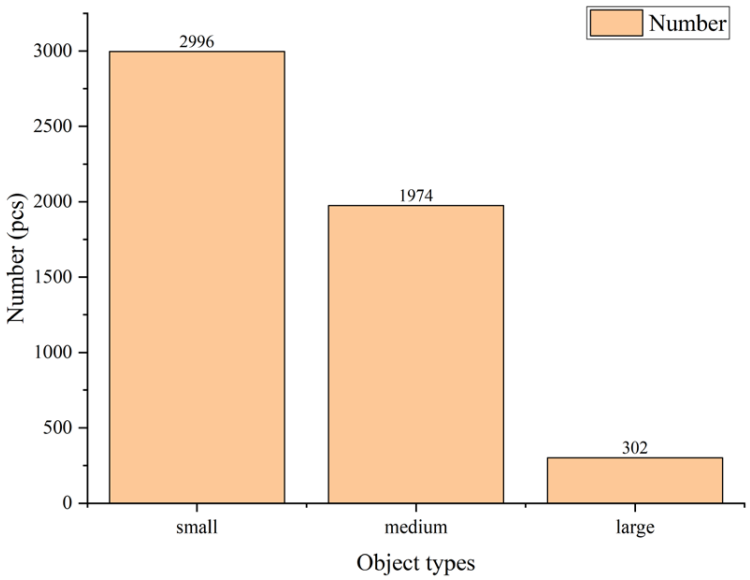

FIGURE 7. Distribution of objects of different types in the dataset

4.1. **Experimental data analysis.**

4.1.1. *The impact of attention mechanisms on detection accuracy.* First, the attention enhancement algorithm was incorporated into the network structure of the Yolov7 algorithm to improve the attention to the floating target, and the test results of the Yolov7 algorithm before and after the improvement are shown in Table 1. In Table 1, the CBAM algorithm and the ECA algorithm were incorporated into the Yolov7 algorithm structure, respectively, and the applicability of the neural network at the shallow A location and the deep B location for the attention mechanism was compared, and their locations are shown in Figure 8. In addition, the mAP boosting rate and mAR boosting rate in Table 1 are based on the test results of the Yolov7 algorithm optimized on the SGD optimizer without incorporating the attention mechanism.

TABLE 1. Comparison of test results of Yolov7 algorithm before and after combining attention enhancement methods

Algorithm	Location	SGD	ADAM	mAP:0.5:0.95(%)	mAR:0.5:0.95(%)
Yolov7	/	1		42.4	48.9
	/		1	39.4	45.7
+CBAM (Yolo-C)	A	1		42.7	48.8
	A		1	44.0	50.2
	B	1		42.2	48.7
	B		1	42.0	48.7
+ECA (Yolo-E)	A	1		42.7	49.0
	A		1	43.1	49.2
	B	1		43.3	49.7
	B		1	42.4	49.2

From Table 1, it can be seen that the best network structure for detection is built by combining the CBAM attention mechanism at position A and using the ADAM optimizer for training, followed by combining the ECA attention mechanism at position B and using the SGD optimizer for training, and the comparative test results are shown in Figure 9.

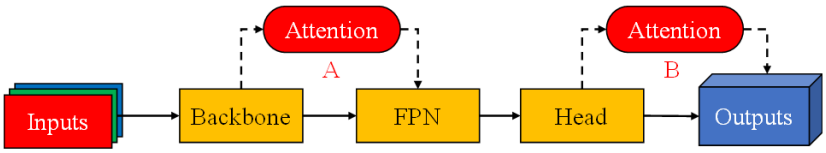

FIGURE 8. Schematic diagram of the combined position of attentional mechanisms

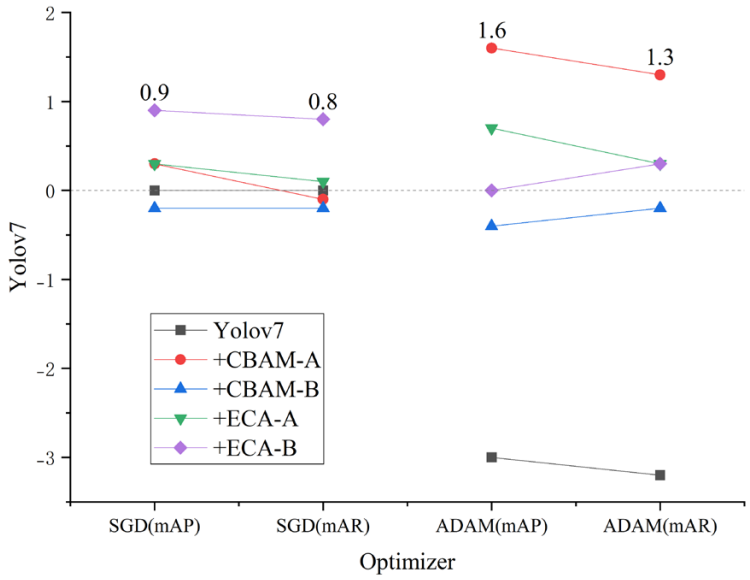

FIGURE 9. Comparison of Yolov7 algorithm test results under different optimizers

It can be seen from Figure 9 that among the different combinations, the incorporation of the attention mechanism in position A results in a better performance improvement, as well as the average test results of the network after training with the ADAM optimizer, which is reflected in the accuracy and recall metrics. However, for the original algorithm without combining the attention mechanism, the ADAM optimizer is much less effective than the SGD optimizer.

4.1.2. *The impact of ISPPCSPC on detection accuracy.* To further improve the performance of the algorithm, the FPN layer was optimized using the ISPPCSPC structure while combining the attention mechanism, and the test results of the Yolov7-F algorithm before and after the improvement are shown in Table 2. Similarly, the CBAM algorithm and the ECA algorithm are combined in the Yolov7-F algorithm structure respectively, and the applicability of the neural network shallow layer A position and deep layer B position for this structure is compared. Unlike Table 1, the mAP boosting rate and mAR boosting rate in Table 2 are based on the test results of the Yolov7 algorithm optimized on the SGD optimizer in Table 1 to verify the advantages of the ISPPCSPC structure.

From Table 2, it can be seen that after using the ISPPCSPC structure, the previously described optimal combinations have an additional boost in terms of precision rate and recall, and the comparison graph of test results for all combinations is shown in Figure 10. As can be seen in Figure 10, the networks trained using the ADAM optimizer are all boosted, and on the contrary, the networks trained using the SGD optimizer are all degraded to different degrees in terms of precision rate, though they are boosted in terms of recall rate. Thus, the combination of the ISPPCSPC structure with the ADAM optimizer is more effective, which can also be reflected on the network that does not pass

TABLE 2. Comparison of test results of Yolov7-F algorithm before and after combining attention enhancement methods

Algorithm	Location	SGD	ADAM	mAP:0.5:0.95(%)	mAR:0.5:0.95(%)
Yolov7 +ISPP-CSPC (Yolo-F)	/	1		42.3	49.0
	/		1	42.9	49.3
+CBAM +ISPPCSPC (Yolo-FC)	A	1		42.4	48.9
	A		1	44.3	50.5
	B	1		42.4	49.1
	B		1	43.1	49.2
+ECA +ISPPCSPC (Yolo-FE)	A	1		42.4	49.3
	A		1	43.4	49.4
	B	1		42.1	48.6
	B		1	42.6	49.3

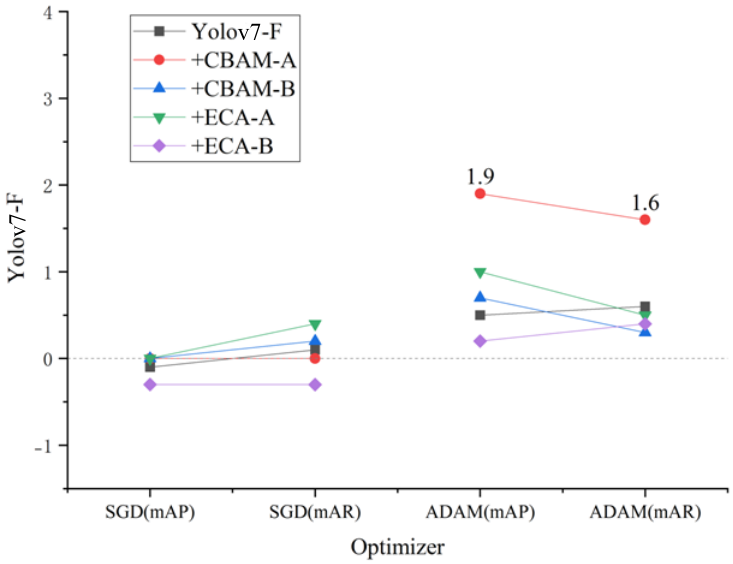

FIGURE 10. Comparison of Yolov7-F algorithm test results under different optimizers

the unincorporated attention mechanism. This is consistent with the role of ISPPCSPC in strengthening cross-scale aggregation. It increases gradient heterogeneity across the neck and heads, where ADAM's variance-normalized adaptive updates stabilize training and balance recall-precision. By contrast, SGD tends to raise recall rather than precision unless the schedule is substantially extended.

In addition, the variation of the test results at this time is more stable regardless of whether the attention mechanism is incorporated at position A or B. Therefore, the ISPPCSPC structure does not have a decisive effect on the position of the attention mechanism fusion. In Table 2, the entries highlighted in bold indicate the Yolo-FC network, which achieved the best recognition results. It realized a 1.9% improvement in mAP and an additional 1.6% increase in mAR.

4.1.3. *Comparison of the Yolo-FC method with other algorithms.* Table 3 gives the test results of the Yolov7-FC algorithm compared to other algorithms while Table 4 shows the comparison of execution speeds of the algorithms on small portable devices.

From Table 3, it can be seen that the Yolov7-FC algorithm has an advantage in both accuracy and recall, and still improves the accuracy by 0.9% compared to the Cascade

TABLE 3. Test results of Yolov7-FC algorithm compared with other algorithms

Algorithm	mAP:0.5:0.95(%)	mAR:0.5:0.95(%)
Yolov7-FC	44.3	50.5
Yolov7	42.4	48.9
Yolov5	42.1	48.8
YoloX	42.9	48.3
Yolov3	33.5	44.0
Cascade R-CNN [17]	43.4	/
FPN [17]	33.4	/
RetinaNet [17]	24.9	/
DSSD [17]	27.5	/

TABLE 4. Speed comparison of running algorithms on small portable devices

Algorithm	Computing speed on small portable devices(s)	Relative ratio(%)
Yolov7-FC	9.23	100.0
Yolov7	9.49	102.8
Yolov5	9.04	97.9
YoloX	7.68	83.2
Yolov3	13.10	141.9
Cascade R-CNN [17]	2.20	23.8
FPN [17]	4.17	45.2
RetinaNet [17]	4.29	46.5
DSSD [17]	16.13	174.8

R-CNN algorithm that performed well on this dataset [17]. Table 4 gives a comparison of the running speed of the algorithm on small portable devices, from which it can be seen that the improved algorithm increases the structural complexity in general compared to the original algorithm, which reduces the computing speed by 2.8%, but still meets the needs of offline real-time detection of small portable devices on board. The Cascade R-CNN algorithm [17], which performs better in terms of accuracy rate, is only 23.8% of the detection speed of the Yolov7-FC algorithm, which basically cannot meet the demand of real-time computing. For the faster detection speed of DSSD algorithm [17] and Yolov3 algorithm, its recognition accuracy and recall rate are low, which cannot meet the demand of floating object detection well.

4.2. **Analysis of visualization results.** Figure 11 displays the object detection performance of various algorithms under different weather and environmental conditions, providing a comprehensive perspective to compare and evaluate the performance advantages of the Yolo-FC algorithm in various scenes. The algorithmic results are summarized in Table 5.

Therefore, by comparing the recognition effect in a variety of backgrounds, it can be learned that the flare has the most serious effect on the recognition algorithm. While in the interference with not serious dim background, a better detection effect can be obtained. Although the small distant objects will have a certain impact on the algorithm, and can still be recognized more accurately in the case of high contrast with the background, once the target color is similar to the background, it will substantially increase the recognition difficulty and reduce the recognition accuracy and recall rate.

FIGURE 11. Comparison of recognition performance of different algorithms in various scenes: (a) Multiple distractions around the object; (b) object color similar to water surface; (c) object in a strong glare environment; (d) dense distribution of objects; (e) object in a dim environment; (f) object color similar to water surface in a dim environment

TABLE 5. Summary of algorithmic results in Figure 11

Scene (Figure 11)	Main challenge	Best model(s)	Typical failure(others)
(a) multi-distractions, long distance	Background clutter	Yolo-FC	Misses by Yolov3/Yolov7/ YoloX; Yolov7 FP on radar
(b) color similar to water	Low color contrast	Yolo-FC	Yolov3/Yolov5 miss the green floater
(c) strong glare	Specular highlights	Yolo-FC (YoloX detects more but with FP)	All degrade; YoloX FP at glare spot
(d) dense targets	Many small instances	Yolo-FC & YoloX	–
(e) dim scene	Low illumination	Yolo-FC	Yolov3 FP (distant ship) Yolov5 FP (radar)
(f) distant weak + color-similar	Small, low-contrast	Yolo-FC (Yolov5 runner-up with FP)	Most models miss

5. **Conclusions.** In this paper, we propose a water surface object recognition algorithm combining attention mechanism, which focuses the algorithm on the discriminative features for images in the channel dimension by compressing the spatial dimension, and improves the computational speed of the model by using fast spatial pyramid pooling. In the test experiment, the improved algorithm improved 1.9% in recognition accuracy and 1.6% in recall, and had more excellent recognition effect under different backgrounds of light spot reflection, insufficient light and the presence of interfering objects, and met the demand of detection in real time on small portable devices. Nevertheless, performance still degrades under strong glare or low target-background contrast (e.g., Figure 11(c)), occasionally causing missed detections or false positives. In future work, we will explore multi-sensor fusion (e.g., polarization or thermal with RGB) and reflection-aware preprocessing or augmentation to suppress specular highlights, together with lightweight temporal modeling to stabilize detections on dynamic water surfaces and cross-dataset evaluation or domain adaptation to broaden applicability.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant No. 52171308 and No. 51879118, in part by the Natural Science Foundation of Fujian Province No. 2025H0022, in part by Special Fund Project for Promoting High-Quality Development of the Marine and Fisheries Industries of Fujian Province No. FJHYF-L-2025-07-018, in part by the Xiamen Natural Science Foundation No. 3502Z202572054, in part by the Applied Technology Engineering Center of Fujian Provincial Higher Education for Marine Resource Protection and Ecological Governance Director's Fund No.202501.

REFERENCES

[1] M. Liu, M. Lin, X. Huang et al., Floating macro-litter pollution in the northern South China Sea, *Environmental Pollution*, vol.316, 120527, 2023.
[2] P. T. Harris, J. Tamelander, Y. Lyons et al., Taking a mass-balance approach to assess marine plastics in the South China Sea, *Marine Pollution Bulletin*, vol.171, 112708, 2021.
[3] O. Garcia-Garin, A. Aguilar, A. Borrell et al., Who's better at spotting? A comparison between aerial photography and observer-based methods to monitor floating marine litter and marine megafauna, *Environmental Pollution*, vol.258, 113680, 2020.
[4] G. Cesarini, R. Crosti, S. Secco et al., From city to sea: Spatiotemporal dynamics of floating macrolitter in the Tiber River, *Science of the Total Environment*, vol.857, 159713, 2023.
[5] T. I. M. Van Emmerik and A. Schwarz, Plastic debris in rivers, *Wiley Interdisciplinary Reviews: Water*, vol.7, no.1, e1398, 2020.
[6] S. Sannigrahi, B. Basu, A. S. Basu et al., Development of automated marine floating plastic detection system using Sentinel-2 imagery and machine learning models, *Marine Pollution Bulletin*, vol.178, 113527, 2022.
[7] M. Boulares and A. Barnawi, A novel UAV path planning algorithm to search for floating objects on the ocean surface based on object's trajectory prediction by regression, *Robotics and Autonomous Systems*, vol.135, 103673, 2021.
[8] S. Kong, M. Tian, C. Qiu et al., IWSCR: An intelligent water surface cleaner robot for collecting floating garbage, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol.51, no.10, pp.6358-6368, 2020.
[9] B. Xue, B. Huang, W. Wei et al., An efficient deep-sea debris detection method using deep neural networks, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.14, pp.12348-12360, 2021.
[10] J. Lin, P. Diekmann, C.-E. Framing et al., Maritime environment perception based on deep learning, *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.9, pp.15487-15497, 2022.
[11] S. Xu, J. Zhu, J. Jiang et al., Sea-surface floating small target detection by multifeature detector based on isolation forest, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.14, pp.704-715, 2020.

[12] L. Wen, J. Ding and Z. Xu, Multiframe detection of sea-surface small target using deep convolutional neural network, *IEEE Transactions on Geoscience and Remote Sensing*, vol.60, pp.1-16, 2021.
[13] T. Liu, B. Pang, L. Zhang et al., Sea surface object detection algorithm based on YOLO v4 fused with reverse depthwise separable convolution (RDSC) for USV, *Journal of Marine Science and Engineering*, vol.9, no.7, 753, 2021.
[14] Z. Yang, Y. Li, B. Wang et al., A lightweight sea surface object detection network for unmanned surface vehicles, *Journal of Marine Science and Engineering*, vol.10, no.7, 965, 2022.
[15] T.-Y. Lin, M. Maire, S. Belongie et al., Microsoft COCO: Common objects in context, *Computer Vision – ECCV 2014: The 13th European Conference*, Zurich, Switzerland, pp.740-755, 2014.
[16] M. Everingham, L. Van Gool, C. K. I. Williams et al., The PASCAL visual object classes (VOC) challenge, *International Journal of Computer Vision*, vol.88, pp.303-338, 2010.
[17] Y. Cheng, J. Zhu, M. Jiang et al., FLoW: A dataset and benchmark for floating waste detection in inland waters, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.10953-10962, 2021.
[18] R. Girshick, Fast R-CNN, *Proc. of the IEEE International Conference on Computer Vision*, pp.1440-1448, 2015.
[19] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement, *arXiv Preprint*, arXiv: 1804.02767, 2018.
[20] Z. Chen, D. Chen, Y. Zhang et al., Deep learning for autonomous ship-oriented small ship detection, *Safety Science*, vol.130, 104812, 2020.
[21] R. W. Liu, W. Yuan, X. Chen et al., An enhanced CNN-enabled learning method for promoting ship detection in maritime surveillance system, *Ocean Engineering*, vol.235, 109435, 2021.
[22] G. Zeng, W. Yu, R. Wang et al., Research on mosaic image data enhancement for overlapping ship targets, *arXiv Preprint*, arXiv: 2105.05090, 2021.
[23] F. S. Leira, H. H. Helgesen, T. A. Johansen et al., Object detection, recognition, and tracking from UAVs using a thermal camera, *Journal of Field Robotics*, vol.38, no.2, pp.242-267, 2021.
[24] X. Zhang, Y. Feng, S. Zhang et al., Finding nonrigid tiny person with densely cropped and local attention object detector networks in low-altitude aerial images, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.15, pp.4371-4385, 2022.
[25] Y. Guo, Y. Xu and S. Li, Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network, *Automation in Construction*, vol.112, 103124, 2020.
[26] J. Guo, H. Feng, H. Xu et al., D3-Net: Integrated multi-task convolutional neural network for water surface deblurring, dehazing and object detection, *Engineering Applications of Artificial Intelligence*, vol.117, 105558, 2023.
[27] S. Li, J. Lin, Y. Lv et al., Deep learning-based algorithm for complex small target detection in UAV aerial images, *International Journal of Innovative Computing, Information and Control*, vol.21, no.1, pp.135-152, 2025.

Author Biography

Guangmiao Zeng received the Ph.D. degree in Naval Architecture and Marine Engineering from Jimei University, China, in 2023. He is currently a Lecturer at School of Navigation, Xiamen Ocean Vocational College, China, and a member of Applied Technology Engineering Center of Fujian Provincial Higher Education for Marine Resource Protection and Ecological Governance, China. His research interests include intelligent information processing and computer vision.

Rongjie Wang received the Ph.D. degree in Electrical and Electronic Engineering from Sun Yat-sen University, China, in 2012. He is currently a Professor and Supervisor of Ph.D. students at School of Marine Engineering, Jimei University, China. His research interests include intelligent information processing, blind source separation and fault diagnosis of novel power system.

Wanneng Yu received the Ph.D. degree in Power Electronics and Power Transmission from Shanghai Maritime University, China, in 2010. He is currently a Professor and Supervisor of Ph.D. students at School of Marine Engineering, Jimei University, China. His research interests include intelligent ship control and electric propulsion.