TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fmars.2022.1056300

OPEN ACCESS

EDITED BY
Hong Song,
Zhejiang University, China

REVIEWED BY
Narayanamoorthi R,
SRM Institute of Science and Technology, India
Siquan Yu,
Shenyang Institute of Automation (CAS), China
Helton Maia,
Federal University of Rio Grande do Norte, Brazil
Chao Qi,
Nanjing Agricultural University, China

*CORRESPONDENCE
Zhuang Zhou
jonahzhou199526@gmail.com

SPECIALTY SECTION
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

RECEIVED 28 September 2022
ACCEPTED 25 October 2022
PUBLISHED 17 November 2022

CITATION
Yan J, Zhou Z, Zhou D, Su B, Xuanyuan Z, Tang J, Lai Y, Chen J and Liang W (2022) Underwater object detection algorithm based on attention mechanism and cross-stage partial fast spatial pyramidal pooling. *Front. Mar. Sci.* 9:1056300. doi: 10.3389/fmars.2022.1056300

Underwater object detection algorithm based on attention mechanism and cross-stage partial fast spatial pyramidal pooling

Jinghui Yan[1], Zhuang Zhou[1]*, Dujuan Zhou[1,2], Binghua Su[1], Zhe Xuanyuan[3], Jialin Tang[1], Yunting Lai[1], Jiongjiang Chen[1,2] and Wanxin Liang[1]

[1]Key Laboratory of Intelligent Detection in Complex Environment of Aerospace Land and Sea, Beijing Institute of Technology, Zhuhai, China, [2]Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China, [3]Faculty of Science and Technology, Hong Kong Baptist University United International College, Beijing Normal University, Zhuhai, China

For the routine target detection algorithm in the underwater complex environment to obtain the image of the existence of blurred images, complex background and other phenomena, leading to difficulties in model feature extraction, target miss detection and other problems. Meanwhile, an improved YOLOv7 model is proposed in order to improve the accuracy and real-time performance of the underwater target detection model. The improved model is based on the single-stage target detection model YOLOv7, incorporating the CBAM attention mechanism in the model, so that the feature information of the detection target is weighted and enhanced in the spatial dimension and the channel dimension, capturing the local relevance of feature information, making the model more focused on target feature information, improved detection accuracy, and using the SPPFCSPC module, reducing the computational effort of the model while keeping the model perceptual field unchanged, improved inference speed of the model. After a large number of comparison experiments and ablation experiments, it is proved that our proposed ACFP-YOLO algorithm model has higher detection accuracy compared with Efficientdet, Faster-RCNN, SSD, YOLOv3, YOLOv4, YOLOv5 models and the latest YOLOv7 model, and is more accurate for target detection tasks in complex underwater environments advantages.

KEYWORDS

Underwater Object detection, ACFP-YOLO, YOLOv7, attention, SPPFCSPC

Introduction

Underwater target detection refers to the localization and identification of a specific target in an underwater scene. The technology is widely used in underwater cable laying, oil exploration, salvage and rescue, marine fish detection, undersea aquaculture, underwater navigation, smart fishery farming, underwater target striking and other fields (Lin and Zhao, 2020) (YU, 2020) (Klausner and Azimi-Sadjadi, 2019). Although target detection algorithms for ground targets are relatively well established, the detection of underwater targets in the underwater environment still faces many challenges. The main reason is that the underwater medium, underwater light conditions, underwater submarine environment, etc. are more complex than the surface environment (Qiang et al., 2020) (Lei et al., 2022). Due to the differential attenuation of different wavelengths of light in water, scattering of light by plankton and suspended particles in water (Wei et al., 2021), making the target in underwater images and videos blurred and with severe color cast, it seriously affects the features of the target and creates serious obstacles for feature learning and recognition understanding of underwater targets. Therefore, underwater target detection continues to face a huge challenge (Jiang and Wang, 2020).

In order to improve the detection of fuzzy underwater targets and small underwater targets in the underwater environment, while maintaining the efficiency of the algorithm and ensuring good detection of the model. In this paper, based on the framework of YOLOv7 (Wang et al., 2022) we introduce the Convolutional Block Attention Module(CBAM) (Woo et al., 2018) attention mechanism and Cross-Stage Partial Fast Spatial Pyramid Pooling(SPPFCSPC) module, and propose a detection model that is suitable for target detection in underwater environment and has stronger feature extraction ability and better detection speed in underwater scenes.

Section 2 of this paper introduces the development process of underwater target detection and the current research problems in this field. Section 3 introduces the overall architecture of our proposed ACFP fusion model in detail, and explains the theory of CBAM attention mechanism and SPPFCSPC module. Section 4 introduces our experimental environment, experimental parameters, datasets and evaluation indicators. Section 5 is the result part. We conduct qualitative and quantitative analysis through the compatibility comparison of different attention mechanisms, ablation experiments, and comparison experiments with mainstream algorithms. Section 6 is a summary of the entire article, illustrating the advantages of our method for high-accuracy and real-time underwater object detection scenarios, as well as future research directions.

The contribution of this paper is shown below:

1. In this paper, we propose a target detection network model for underwater environment based on the improvement of YOLOv7. In which we fuse the CBAM attention mechanism and the SPPFCSPC module in the YOLOv7 model, this fusion idea effectively improves the detection accuracy of the model for underwater fuzzy targets and small targets, and provides an effective solution for the underwater target detection task.
2. On the URPC dataset, the mAP value of ACFP-YOLO is 80.62%, and the detection speed FPS value is 64.21. On the underwater garbage detection dataset, the mAP value is 74.92%, and the detection speed FPS value is 65.56. On both datasets, ACFP-YOLO achieves the highest detection accuracy and has better inference speed.
3. After a large number of comparative experiments and ablation experiments, we quantitatively and qualitatively verify that our model is superior to classical and state-of-the-art methods in the task of underwater environmental target detection from different perspectives. The detailed experiments provide a detailed idea and an important basis for other researchers to refer to our work.

Related work

In the early 21st century some researchers started to study underwater target detection algorithms, and with the development of artificial intelligence, new branches of research on underwater target detection algorithms have emerged. For the present the field is divided into two branches, one is the traditional algorithm based on the detailed feature analysis of the image, and the other is the neural network model algorithm based on the training of a large number of underwater images, by extracting features, analyzing the image features, extracting the target information and obtaining the final detection results (Tinghui et al., 2022).

Traditional target detection algorithms mainly digitize images and then apply mathematical theoretical knowledge to analyze and model them. However, the same target from different angles presents a different shape on the picture, and there are great difficulties in mathematical modeling of it, and the established model is difficult to be put into the application of realistic underwater scenes. QIU et al. (2019) proposed an underwater motion target detection algorithm based on surface feature ripples, this algorithm in photoelectric polarization imaging mode for underwater target detection, which became a mature masterpiece of traditional algorithms in underwater target detection. But, the traditional algorithm for modeling the features of a certain class or a certain target in underwater targets in different water quality environments is a major limitation, and the robustness for detecting underwater images with different complex backgrounds under different lighting is poor.

With the development of deep learning in recent years, target detection algorithms based on convolutional neural networks have been developed (Girshick et al., 2015) and have achieved better results in recent years, and this type of algorithm has significant advantages in various detection tasks. Villon et al. (2016) did a study on the performance of traditional and deep learning algorithms, which used a two-stage extraction of HOG features and the use of an SVM classifier to compare detection on a coral reef fish dataset, and the results showed that the deep learning algorithm has more advantages. The current mainstream methods can be divided into two-stage target detection algorithms and one-stage target detection algorithms. Among them, the two-stage target detection algorithm performs the detection task in two stages, generating the suggestion frame first in the first stage and then making predictions in the second stage. This type of algorithm has high detection accuracy but is slow, and its representative algorithms are the Faster R-CNN (Ren et al., 2015) series. The one-stage target detection algorithm treats the target detection task as a single regression problem, and although it is slightly lower in accuracy than the two-stage algorithm, it has a faster detection speed, represented by the YOLO family of algorithms, such as YOLO (Redmon et al., 2016), YOLO9000 (Redmon and Farhadi, 2017), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 (Bochkovskiy et al., 2020), YOLOv5 (Jocher, 2020)and SSD (Liu et al., 2016).

Chen et al. (2020) proposed a new sample-weighted super network (SWIPENET) and a robust training paradigm–curriculum Multi Class Adaboost (CMA), for underwater images with small blurred samples. Shi et al. (2021) proposed an improved Faster-RCNN based underwater detection algorithm to improve the detection accuracy of Faster-RCNN in underwater scenes for problems such as low quality of underwater images, overlapping or occluded targets, and different sizes or shapes. Zeng et al. (2021) proposed a Faster R-CNN-AON network for the complex and variable underwater environment with limited acquired sample images, and introduced an adversarial network to improve the overall detection performance of the model. However, the optimization algorithm based on two-stage algorithm still has the problems of low efficiency and poor real-time performance. To address this problem above, Tinghui et al. (2022) proposed an improved YOLOv5 underwater target detection network model, the model has unique performance in underwater target detection and maintains with real-time and accuracy. Lei et al. (2022) used Swin Transformer as the base backbone of YOLOv5 for underwater image blurring, making the network suitable for underwater images with blurred targets.

Attention mechanism and cross-stage partial fast spatial pyramid pooling (ACFP)

We present our underwater target detection model in detail in Section 3. We based on the YOLOv7 framework, incorporating the CBAM attention mechanism, Using the channel attention mechanism and spatial attention mechanism, the channel weight of the detection target is increased, while the perceptual field of the target to the original image is expanded, allowing the model to pay more attention to the feature information of the detection target. We improve the original SPPCSPC module to SPPFCSPC module, which reduces the computation of the model and improves the inference speed of the model while keeping the perceptual field unchanged.

Network model for underwater target detection

Before feeding the network with images, we first perform a distortion-free affine transformation of the original input image to a 640×640 size image, which is then used as the input to the model. The overall network framework model after our fusion is shown in Figure 1, where the modules marked with red boxes are the parts of the model fusion. The input images are first passed through the backbone network for feature extraction. In order to retain multi-scale information, the backbone network provides a variety of different scales and outputs the multi-scale feature maps to the neck network as the input of the neck network. After the neck network, the fusion of feature maps containing shallow fine-grained information and deep semantic information is combined, thus enhancing the expressive power of the network and assigning the multi-scale learning task to multiple detection networks of different sizes. Finally, the feature information is integrated and transformed into detection prediction output.

Backbone network is used for feature extraction of images, such as texture, color and shape of images. It can provide multiple scales, multiple combinations of sense field sizes and center steps, thus meeting the requirements of different scales and categories. The extraction process of backbone network is shown in Figure 1A, firstly, it goes through 4 CBS modules for convolution, normalization and activation, and then it is after the E-ELAN module and MP module to extract features alternately, leading to the output of the last 3 E-ELAN modules as the input of neck. Among them, the E-ELAN module is composed of multiple convolutional layers, and the MP module is composed of MaxPool and CBS modules as shown in Figure 2C, D.

In order to enable the model to learn diverse information and improve the performance of target detection, the role of neck network is to disperse the multi-scale output learning provided by backbone network to multiple feature maps and fuse the learned multi-scale information together, which improves the perceptual wildness of the model while effectively separating the most important contextual features and avoiding the image distortion problem to some extent. As in Figure 1B, the neck network is a PAFPN structure, consisting of a modified

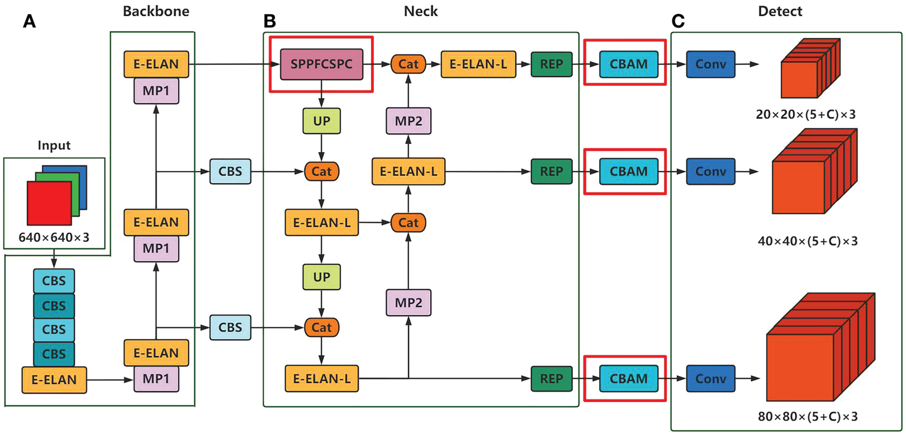

FIGURE 1
Overall architecture diagram of the model. Where **(A)** denotes the backbone network for feature extraction, **(B)** denotes the neck network for feature fusion, **(C)** denotes the detect network used to obtain the model prediction results, and *C* denotes the number of categories in the dataset.

FPN (Lin et al., 2017) and PANet (Liu et al., 2018) structure, for extracting features and fusing them. The PAFPN structure is basically the same as that of YOLOV5, except that the PAFPN structure of YOLOv7 uses the E-ELAN-L module for feature extraction and fusion, and the MP module for down sampling, this makes this structure more capable of strengthening features than previous PAFPN structures. Finally, after two feature extractions of the input image by backbone network and neck network, the 1×1 convolution is used to integrate the feature information to convert it into the final prediction information, as shown in Figure 1C, to obtain the prediction results of the model.

SPPFCSPC for ACFP

Spatial pyramid pooling is more effective than simply using maximum pooling to increase the received range of backbone features, significantly separating the most important contextual features, and this structure outputs fixed-size feature vectors after multi-scale feature extraction to increase the perceptual field of the network.

The SPPCSPC structure and the SPPFCSPC structure are shown in Figure 3. The SPPCSPC structure in YOLOv7 uses three independent pooling layers with different sizes of pooling kernels to compute a spatial pyramid pooling structure. The

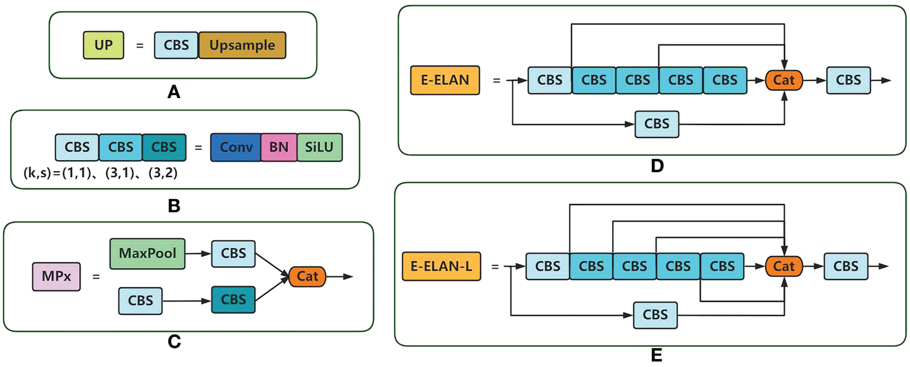

FIGURE 2
Structure diagram of the model part of the module. Where **(A)** denotes the upsampling module, **(B)** denotes the composition of different convolution modules, where *k* denotes the size of the convolution kernel and *s* denotes the convolution step size, **(C)** denotes the basic structure of the MP module, **(D)** denotes the basic structure of the E-ELAN module, and **(E)** denotes the basic structure of the E-ELAN-L module.

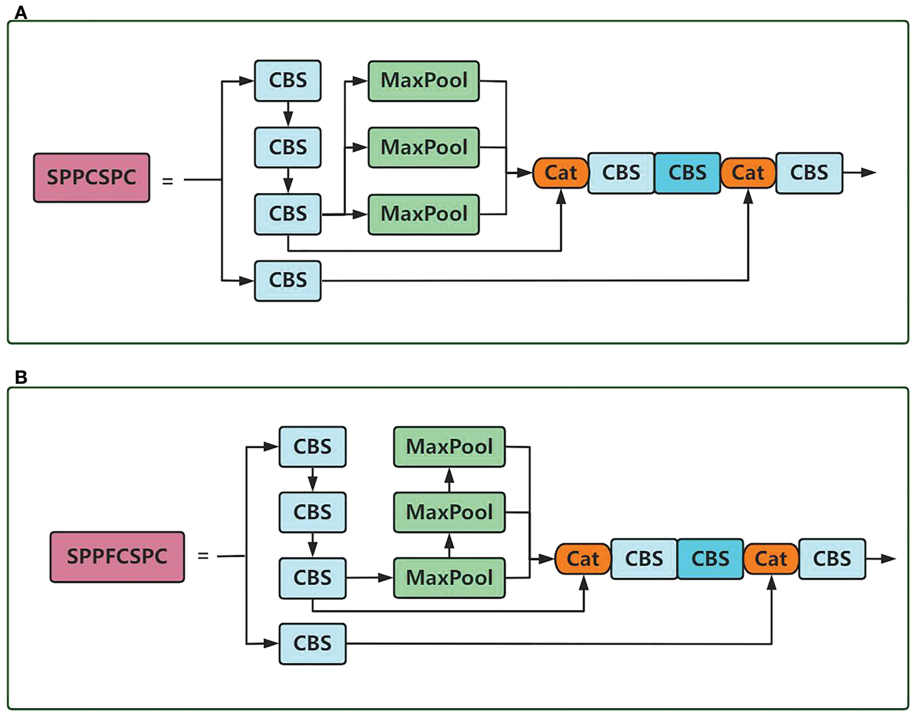

FIGURE 3
(A) indicates the SPPCSPC structure diagram, **(B)** indicates the SPPFCSPC structure diagram.

relevant pooling part of the equation is shown in equation (1), but the three pooling have the same input, and the results of the larger pooling kernel can be calculated on the computational results of the smaller output results of the pooling kernel, reducing the computational effort without changing the perceptual field of the module.

$$\mathbf{R(F)} = \mathbf{MaxPool}^{p=2}_{k=5}(\mathbf{F}) \circledast \mathbf{MaxPool}^{p=4}_{k=9}(\mathbf{F}) \circledast \mathbf{MaxPool}^{p=6}_{k=13}(\mathbf{F}) \tag{1}$$

Where R denotes the output result, $\circledast$ denotes tensor stitching, and F denotes the input feature layer.

The SPPFCSPC structure is optimized for the SPPCSPC structure, and the pooling part is calculated as shown in equations (2), (3), (4), and (5), linking three separate pooling uses less computation on the output results of the pooling layer of the smaller pooling kernel, yielding the pooling layer results of the larger pooling kernel, gaining speedup while keeping the perceptual field constant.

$$R_1(\mathbf{F}) = \mathbf{MaxPool}^{p=2}_{k=5}(\mathbf{F}) \tag{2}$$

$$R_2(R_1) = \mathbf{MaxPool}^{p=2}_{k=5}(R_1) \tag{3}$$

$$R_3(R_2) = \mathbf{MaxPool}^{p=2}_{k=5}(R_2) \tag{4}$$

$$\mathbf{R_4} = \mathbf{R_1} \circledast \mathbf{R_2} \circledast \mathbf{R_3} \tag{5}$$

Where R_1 denotes the pooling layer result for the minimum pooling kernel, R_2 denotes the pooling layer result for the medium pooling kernel, R_3 denotes the pooling layer result for the maximum pooling kernel, and R_4 denotes the final output result, $\circledast$ denotes tensor stitching.

Attention for ACFP

CBAM is an attention mechanism module that incorporates two dimensions of feature channel information and feature space information. As shown in Figure 4, CBAM processes the incoming feature layers by the channel attention mechanism and the spatial attention mechanism, respectively, and automatically obtains the importance level for each feature channel and feature space by learning, and uses the obtained importance level to enhance features and suppress features that are not important for the current task. The overall equation of CBAM is summarized as shown in equation (6)(7):

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

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

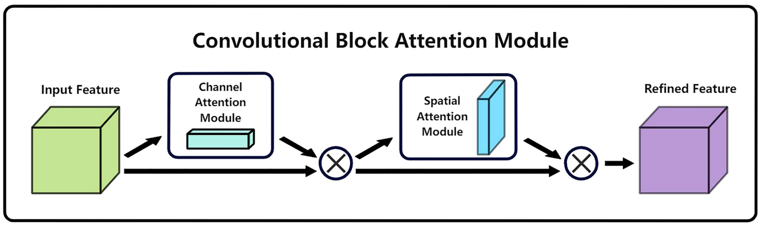

FIGURE 4
Overall structure of CBAM attention mechanism.

The working process of CBAM is to first multiply the input feature layer F through the channel attention mechanism and the obtained M_c with the input feature layer F to obtain the output of strengthening and suppression on the channel F', and then use F' as the input of the spatial attention mechanism, the obtained M_s is multiplied with F' to obtain the final output F'' of reinforcement and suppression in the channel content and spatial location.

The module used for the channel attention mechanism is shown in Figure 5A, which consists of MaxPool, AvgPool, and Shared MLP. The related equation is shown in equation (8):

$$\begin{aligned}\mathbf{M_c}(F) &= \sum(\mathbf{MLP}(\mathbf{AvgPool}(\mathbf{F})) + \mathbf{MLP}(\mathbf{MaxPool}(\mathbf{F}))) \\ &= \sum\left(W_1(\mathbf{W}_0(\mathbf{F}^c_{\mathrm{avg}})) + W_1(\mathbf{W_0}(\mathbf{F}^c_{\mathbf{max}}))\right)\end{aligned} \tag{8}$$

Where σ denotes the sigmoid function and W_0 W_1 denote the two shared fully connected layers that make up the *MLP*.

For the input single feature layer $F(H{\times}W{\times}C)$, MaxPool and AvgPool are performed in the $H{\times}W$ dimension respectively to compress the feature layer to 1×1×C. The one-dimensional parameters after MaxPool compression retain the feature texture of the original feature layer and converge the important information to distinguish the object features. The one-dimensional parameters compressed by AvgPool retain the global visual information of $H{\times}W$ before compression and have a larger perceptual area. After that, the results of MaxPool and AvgPool are fed into Shared *MLP* network for processing, and then the two processed results are summed to obtain the feature map channel weights.

The module used for the spatial attention mechanism is shown in Figure 5B, which consists of MaxPool, AvgPool and conv layer. The related equation is as in (9):

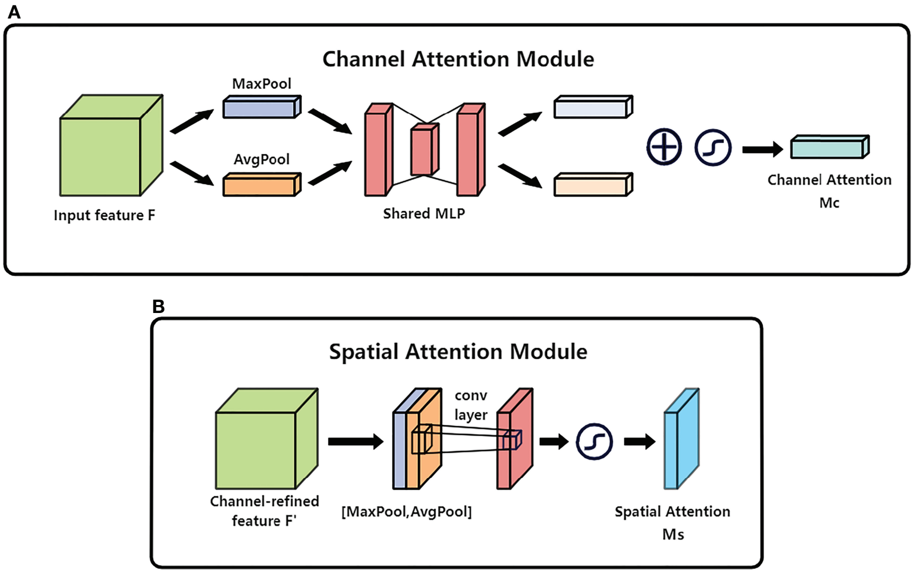

FIGURE 5
(A) denotes the specific structure of the channel attention mechanism branch and **(B)** denotes the specific structure of the spatial attention mechanism branch.

$$\mathbf{M_s(F)} = \sum(\mathbf{f}^{7\times7}([\mathbf{AvgPool(F)};\mathbf{MaxPool(F)}])) \\ = \sum\left(\mathbf{f}^{7\times7}([\mathbf{F}^{s}_{avg};\mathbf{F}^{s}_{\max}])\right) \tag{9}$$

Where σ denotes the sigmoid function and $f^{7\times7}$denotes the convolution operation with a convolution kernel size of 7 × 7.

For a single feature layer $F(H\times W\times C)$ input in, MaxPool and AvgPool are performed in the channel dimension respectively to compress the feature layer to $H\times W\times1$. The compressed feature layer focuses the effective information of the region in the space and is used to extract the efficient information region along the channel, after which the results of both are concat and then convolutional dimensionality reduction is performed to obtain the feature map space weights, thus capturing the local relevance of the feature information.

Loss function

All the experiments in this paper use training without the auxiliary training head, therefore, we only describe the loss function when training without the auxiliary training head in the following, and the overall loss calculation formula is shown in equation (10):

$$\mathcal{L}_{\mathrm{v7}}(\mathrm{t_p},\mathrm{t_{gt}}) = \sum_{\mathrm{m=0}}^{\mathrm{M}} \left[\alpha_{\mathrm{m}}^{\mathrm{bal}}\alpha_{\mathrm{box}}\sum_{\mathrm{i=0}}^{\mathrm{S^2}}\sum_{\mathrm{j=0}}^{\mathrm{B}}I_{\mathrm{kij}}^{\mathrm{obj}}\,\mathcal{L}_{\mathrm{CIoU}} + \alpha_{\mathrm{obj}}\sum_{\mathrm{i=0}}^{\mathrm{S^2}}\sum_{\mathrm{j=0}}^{\mathrm{B}}\mathrm{I}_{\mathrm{kij}}^{\mathrm{obj}}\mathcal{L}_{\mathrm{obj}} + \alpha_{\mathrm{cls}}\sum_{\mathrm{i=0}}^{\mathrm{s^2}}\sum_{\mathrm{j=0}}^{\mathrm{B}}I_{\mathrm{kij}}^{\mathrm{obj}}\mathcal{L}_{\mathrm{cls}}\right] \tag{10}$$

Where M denotes the output feature layer, S^2 denotes the cell, B denotes the number of anchors on each cell. a_{box}, a_{obj}, a_{cls} denote the weights of the corresponding terms, and the values taken in the experiment are a_{box}=0.05, a_{obj}=1.0, a_{cls}=0.03. I_{kij}^{obj} is the control function, which indicates whether the mth output feature map, the i-th cell, the j-th anchor box is a positive sample, if it is a positive sample, it is 1, and vice versa, it is 0. t_{p},t_{gt} is the prediction vector and ground-truth vector. a_m^{bal} is used to balance the weights of the output feature map of each scale, and the values are [4.0,1.0,0.4], which correspond to 80×80, 40×40, 20 ×20 for the output feature maps.

The localization coordinate loss expressed by the Bounding Box boundary regression loss function, using CIoU loss, calculates the localization loss of positive samples only, as shown in (11)(12):

$$\mathcal{L}_{\mathrm{CIoU}}(\mathrm{b},b_{\mathbf{gt}}) = 1 - \mathrm{CIOU} \\ = 1 - \left(\mathrm{IOU} - \frac{d_o^2}{d_c^2} - \frac{v^2}{1-\mathrm{IOU}+\mathrm{u}}\right) \tag{11}$$

$$\mathbf{u} = \frac{\mathbf{4}}{\pi^2}\left(\mathbf{arctan}\ \frac{b_{gt-w}}{b_{gt-h}} - \mathbf{arctan}\ \frac{b_W}{b_h}\right)^2 \tag{12}$$

Where b,b_{gt} denotes the prediction frame vector b_{x},b_{y},b_{w},b_{h} and the ground-truth vector $b_{\mathrm{gt\text{-}x}}$, $b_{\mathrm{gt\text{-}y}}$,$b_{\mathrm{gt\text{-}w}}$, $b_{\mathrm{gt\text{-}h}}$.*IOU* denotes the intersection ratio between the prediction frame and the ground-truth. d_o denotes the Euclidean distance between the prediction frame and the center point of the ground-truth. d_c denotes the diagonal distance between the prediction frame and the smallest outer rectangle of the ground-truth v denotes the impact factor measuring the aspect ratio.

The target confidence loss function adopts BCE loss, which only calculates the objective loss of the samples obtained from positive sample matching, and the specific formula (13):

$$\mathcal{L}_{\mathrm{obj}}(\mathrm{p_o},\mathrm{p_{iou}}) = \mathrm{BCE}_{\mathrm{obj;}}^{\mathrm{sig}}(\mathrm{p_o},\mathrm{p_{iou}};\mathrm{w_{obj}}) \tag{13}$$

Where p_o denotes the target confidence score in the prediction frame, p_{iou} denotes the prediction frame and the *IOU* value of the ground-truth corresponding to it.

The classification loss function, using BCE loss, calculates the classification loss of positive samples only, and the specific calculation formula is as (14):

$$\mathcal{L}_{\mathrm{cls}}(c_{\mathrm{p}},c_{\mathrm{gt}}) = \mathrm{BCE}_{\mathrm{cls}}^{\mathrm{sig}}(c_{\mathrm{p}},c_{\mathrm{gt}};w_{\mathrm{cls}}) \tag{14}$$

Where c_p denotes the probability of the target category in the prediction frame, c_{gt} denotes the probability of the category of the ground-truth to which the prediction frame corresponds.

Experiments

Experimental environment and hyperparameter settings

All experimental data in this paper are measured in the same environment. The hardware environment uses Intel(R) Core (TM) i7-12700KF@3.61 GHz CPU, 16GB RAM, NVIDIA GeForce RTX 3080 Ti graphics card. The system environment is Windows 10 Professional Edition. Python version 3.8, PyTorch version 1.12.0, CUDA version 11.6.

The relevant parameters in the experiment are shown in the Table 1. The gradient descent optimizer used to update the convolution kernel parameters is Adam, and the optimizer Momentum is 0.937, the learning rate update method during the training process is step, the maximum learning rate is 0.001, the frozen training batch size is 8. The epoch of freezing training is 50, the batch size of unfreezing training is 4, the epoch of unfreezing training is 50, all experiments only load the pre-training weights of the backbone network part, and other parts

TABLE 1 Experiment-related hyperparameter settings.

Hyperparameter	Freeze_train	Epoch	Batch_size	Max_learning_rate	Optimizer	Momentum	Lr_decay
Value	True	1-50	8	0.001	Adam	0.937	Step
	False	51-100	4				

are trained from scratch, and the total training epoch is 100. The frozen training model only trains other parts except the backbone network, and the entire network model is trained when unfreezing training.

Dataset

There are 2 datasets used in the experiments in this paper. The main experiments are performed on the URPC (Lab, 2018) dataset, and the auxiliary verification experiments are performed on the underwater garbage detection dataset (Fulton et al., 2019). For the URPC dataset, in order to enable the model to learn more features, more pictures with the same category are added, and the category of waterweeds is added. In order to make the category distribution of the training set and test set more reasonable, we replace Some pictures of the test set. The dataset consists of 4571 images, including 3771 training images and 800 testing images, covering 5 target categories: scallop, holothurian, starfish, echinus, and waterweeds. The underwater garbage detection data set has a total of 7337 pictures, including 6206 training pictures, 1461 test pictures, and 13 categories marked, namely timestamp, Paper, Wood, Bio, Metal, Rov, Plastic, Unknown, Papper, Platstic, Rubber, Cloth and Fishing. The pictures of the dataset were taken in the real marine environment, and the pictures have problems such as color distortion, low contrast, blurred feature information, etc., and there are occlusions, dense targets, and uneven distribution of the number of targets in different categories, which gives underwater problems. Object detection brings great challenges.

Evaluation indicators

There are seven main indicators used in this study to test the performance of the model. Precision(P) represents the proportion of the positive class that the model considers to be a positive class, and the calculation formula is in Equation 15. Recall(R) represents the proportion of the positive class divided by the model to the total positive class, and the calculation formula is in Equation 16. Average Precision (AP) means that each class is composed of Precision and Recall taking different thresholds The area under the curve, the larger the value, the better the recognition accuracy of the class, the formula for calculation is in Equation 17. The mean Average Precision (mAP) represents the average AP of all classes, and the larger the value, the better the model The better the accuracy of identifying the target, the calculation formula is in Equation 18. Frame Per Second (FPS) (Liu et al., 2009) represents the number of frames processed by the model per second, reflecting the speed of the model inference, the larger the value, the faster the inference speed of the model, and the better the model performance. Billions of floating point operations per second (GFLOPS) is the number of computations required by the model and measures the complexity of the model. Number of parameters(params) is the sum of the parameters in the model and is used to evaluate the model size.

$$\mathbf{P(Precision)} = \frac{\mathbf{TP}}{\mathbf{TP + FN}} \tag{15}$$

$$\mathbf{R(Recall)} = \frac{\mathbf{TP}}{\mathbf{TP + FN}} \tag{16}$$

$$\mathbf{AP} = \int_0^1 \mathbf{P(r)dr} \tag{17}$$

$$\mathbf{mAP} = \frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{AP_n} \tag{18}$$

Where TP represents the number of positive samples predicted by the model correctly, and FP represents the number of positive samples predicted by the model that are actually negative samples. FN represents the number of positive samples predicted by the model to be negative. P represents the precision of this class, r represents the recall of this class, N represents the number of all classes, and AP_n represents the average precision of class n.

Results

Compatibility of attention mechanisms

SENet (Hu et al., 2018) is a typical implementation method of channel attention mechanism, which focuses on obtaining the enhanced weights of the input feature layer on the channel, but ignores the weight information of the target spatial position. ECA (Wang et al., 2020) is also an implementation form of the channel attention mechanism, which obtains the enhancement weight of each feature layer by obtaining cross-channel information. Although it has better cross-channel information, it also ignores the spatial information of the target. The CA (Hou et al., 2021) attention mechanism embeds the location information into the channel attention, decomposes the channel attention into two feature encoding processes, aggregates the features along two spatial directions respectively, and obtains the weight of the fusion channel information and spatial information. CBAM combines the channel attention mechanism and the spatial attention mechanism to deal with the channel weight and the spatial weight respectively, that is, it pays attention to both the channel information and the spatial information. Different attention mechanisms focus on different information directions, and different models have different compatibility with different attention mechanisms.

We chose to use the CBAM attention mechanism in our improved model, and to verify the compatibility of the CBAM attention mechanism with the model, we compared it with the models without fused attention mechanism, fused SENet attention mechanism, fused ECA attention mechanism, and fused CA attention mechanism in separate experiments.

A visualization method in deep learning was used in the experiments for qualitative analysis, which is the Gradient Weighted Class Activation Mapping (Grad-CAM) (Zhou et al., 2016), used to show the differences in the regions of interest for the different attention mechanisms introduced by the model, reflecting the degree of influence of different regions on the results. In this case, the feature importance increases sequentially from blue to red light.

As shown in Figure 6, compared with the visualization results of other attention mechanisms, the overall coverage area of the heat map of the CBAM attention mechanism is larger, indicating that the model focuses on a larger learning area at locations with targets, and the overall feature extraction of the targets is more adequate, which is beneficial to the detection of small targets, and the red area also becomes larger, indicating that the effective target feature information is enhanced and the model is more focus on the target information that should have been focused on. From the experimental results, it can be seen that the introduction of CBAM attention mechanism makes the model pay more attention to the feature information of the target to be recognized, and suppresses the effect of target features that are not obvious due to the complex underwater background, and shows better results compared with other attention mechanisms.

The experiments were quantitatively analyzed with mAP assessment criteria. We changed only the attention mechanism module, and then measured the mAP values of each model, and compared the mAP values of different models to assess the compatibility of different attention mechanisms with the models, and the data of the comparison experiments are shown in Table 2.

The experimental data show that the model incorporating the CBAM attention mechanism has higher detection accuracy compared to the models incorporating the SE attention mechanism, incorporating the ECA attention mechanism, incorporating the CA attention mechanism, and not incorporating the attention mechanism. The detection accuracy of the model with fused SE attention mechanism and ECA attention mechanism decreased by 0.75% and 0.95%,

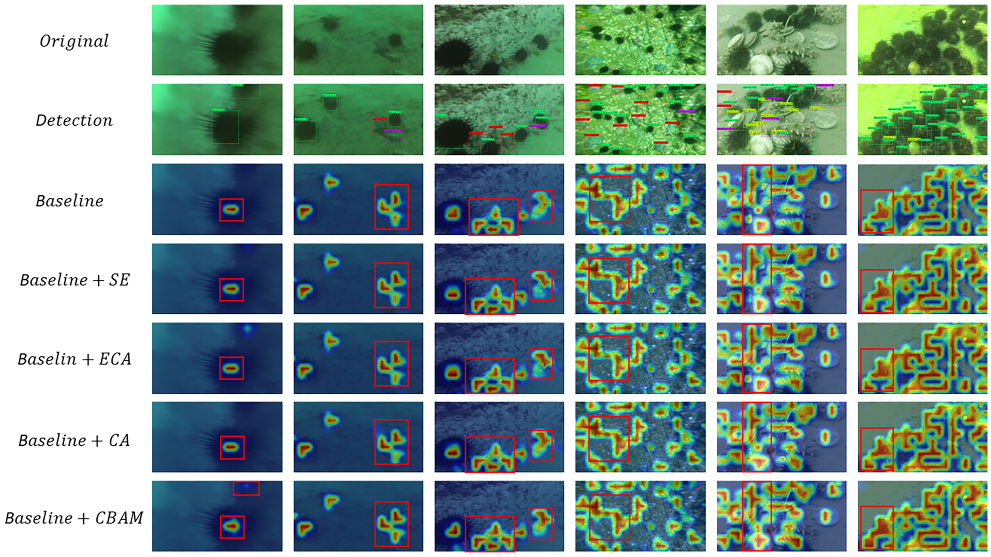

FIGURE 6
Heatmap of different attention mechanisms. Original shows the data set image. Detection shows the detection results of Baseline+CBAM, and the Baseline shows the heat map of YOLOv7. Baseline+XX shows the heat map of YOLOv7 integrated into the XX attention mechanism (XX is SE, ECA, CA, CBAM).

TABLE 2 mAP measurements for different attention mechanisms.

Model	Attention	Input shape	mAP(%)
YOLOv7	–	640×640	78.98
YOLOv7	SENet	640×640	78.23
YOLOv7	ECA	640×640	78.03
YOLOv7	CA	640×640	79.39
YOLOv7	CBAM	640×640	**80.67**

The bold value is the best value in the comparison.

respectively. The detection accuracy of the model with fused CA attention mechanism improved by 0.41%. The detection accuracy of the model with fused CBAM attention mechanism improved by 1.69%. Compared with the original YOLOv7 model, which is more adaptable to the underwater scenario and has better model compatibility.

Ablation experiments

As described in Section3, we introduced the CBAM attention mechanism in the model, as well as the improved SPPCSPC module. To verify the effectiveness of the improved YOLOV7 model for underwater target detection, we controlled a variable by the control variable method and quantitatively analyzed the experimental results. In the experiments, we measured the mAP and FPS values of each model and compared them by metrics to verify the importance of the improved module for the model. Three models were designed for comparison with the improved model in this experiment, where experiment 1 represents the original YOLOV7 model, experiment 2 incorporates the CBAM attention mechanism based on experiment 1, experiment 3 replaces the SPPCSPC module with the SPPFCSPC module based on experiment 1, and experiment 4 is the improved model. The experimental data are shown in Table 3.

Comparing the data from Exp.1 and Exp.2, the model with the introduction of the CBAM attention mechanism improves the average detection accuracy (mAP) by 1.69% and slightly reduces the model inference speed, indicating that the CBAM attention mechanism uses channel attention to establish the correlation between channels, thus suppressing the non-essential feature information, while using the spatial attention mechanism to extract the spatial location of the target more effectively. Through the parallel action of both, the model pays more attention to the feature information of the detection target, thus improving the quality of the feature mapping and significantly improving the overall accuracy of the model, but the CBAM attention mechanism increases the complexity of the model and reduces the inference speed of the network; comparing the data of Exp.1 and Exp.3, the model inference speed (FPS) is improved by 0.85%, indicating that replacing the SPPCSPC module is replaced with the SPPFCSPC module, the model inference speed is improved while keeping the perceptual field unchanged; comparing the data of Exp.1 with Exp.4, the average detection accuracy (mAP) on the model is improved by 1.64% and the inference speed is slightly reduced, indicating that the YOLOv7 model that incorporates the CBAM attention mechanism and replaces the SPPCSPC module sacrifices a small portion of speed in exchange for higher detection accuracy, balancing the one-sided performance degradation brought by using either one alone, and making the overall performance of the model more superior.

Comparison with mainstream algorithms

Our proposed ACFP-YOLO algorithm has good feature extraction ability in complex underwater scenes, and has a faster detection speed, and has better performance in underwater target detection. In order to verify the superiority of the ACFP-YOLO algorithm in this paper in underwater detection, we compare the algorithm in this paper with Efficientdet (Tan et al., 2020), Faster-RCNN (F-RCNN), SSD, YOLOv3, YOLOv4, YOLOv5, YOLOv7 target detection mainstream algorithms, in On the same data set URPC, the same training method is used for network model training, and the superiority of different algorithm models is compared through qualitative analysis and quantitative analysis.

In the experiment, we qualitatively analyze the performance of the algorithm through the detection renderings of different models, and we select the model with better detection effect for analysis. Figure 7 shows Faster-RCNN-ResNet50, YOLOv3, YOLOv4, YOLOv5-l, YOLOv7 detection renderings. From the intuitive renderings, it can be concluded that the detection effect of F-RCNN is better than that of YOLOv3 and YOLOv4, and is comparable to that of YOLOv5, but the detected target

TABLE 3 The impact of the fusion of different modules of the model on the metrics.

Index	CBAM	SPPFCSPC	Input shape	mAP (%)	FPS
Exp.1	×	×	640x640	78.98	65.94
Exp.2	✓	×	640x640	80.67	62.29
Exp.3	×	✓	640x640	78.45	66.50
Exp.4	✓	✓	640x640	80.62	64.21

"✓" indicates that the module is incorporated in the model, and "x" indicates that the module is not incorporated in the model.

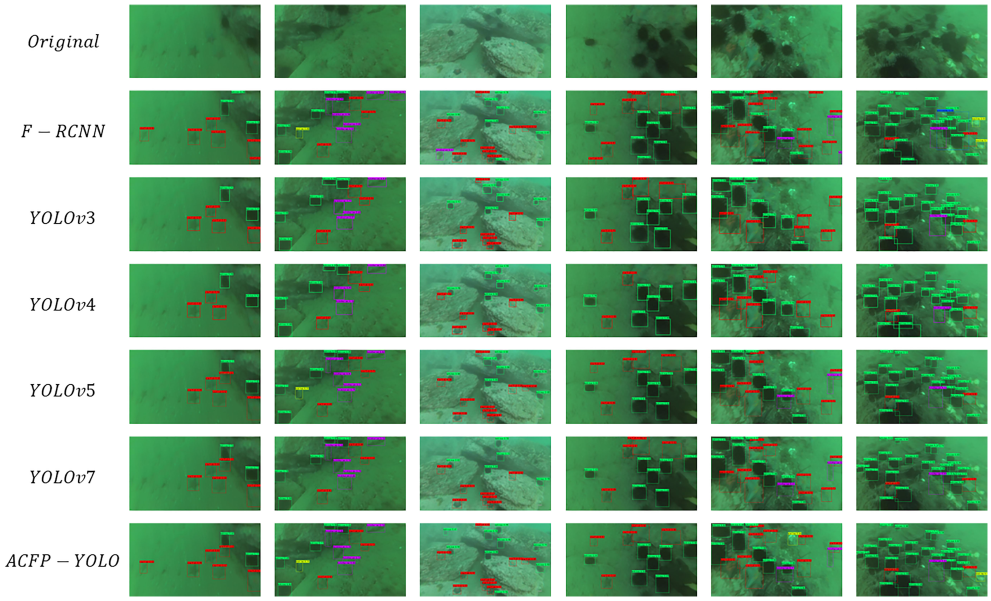

FIGURE 7
The first row *Original* represents the original image of the dataset, the second row represents the F-RCNN detection image, the third row represents the YOLOv3 detection image, the fourth row represents the YOLOv4 detection image, the fifth row represents the YOLOv5 detection image, and the sixth row represents the YOLOv7 detection image Figure, the seventh row represents the ACFP-YOLO detection map.

probability is generally lower than that of YOLOv5, and there is a target misjudgment. The ACFP-YOLO algorithm in this paper has better target recognition effect than YOLOv3, YOLOv4 and YOLOv5, and has fewer misjudgment and recognition compared with F-RCNN.

In order to better observe and compare the detection effect of the ACFP-YOLO algorithm, the ACFP-YOLO algorithm is compared with YOLOv7, and YOLOv7 has the best detection effect among other algorithms. Figure 8 shows the detection results of the ACFP-YOLO algorithm and the YOLOv7 algorithm. The target marked by the yellow box in the figure has blurred edges and distorted colors, which makes it difficult to identify the features similar to the background. The ACFP-YOLO algorithm is not obvious for such target features, the target edge is blurred, the detection effect of small targets is better, and the detection ability of edge feature information, overlapping and blurred targets is stronger. From the perspective of detection effect, the ACFP-YOLO algorithm has better feature extraction ability in complex underwater scenes, and improves the detection ability of small targets and targets with indistinct edge features. At the same time, the fusion of the CBAM attention mechanism enhances the spatial feature information of the model for small targets, improves the model's detection ability for small targets, and improves the detection ability of underwater targets at various scales.

In the experiment, we carried out quantitative comparative analysis of each model by measuring the mAP value, params value, GFLOPS value and FPS value of each model. Efficientdet used the D1 model for the experiment, and Faster-RCNN used the VGG backbone network and the ResNet50 backbone network for the experiment. All model comparison experimental measurement results are shown in Table 4.

Observing the experimental data, from the perspective of detection accuracy, the mAP value of the ACFP-YOLO algorithm is 80.62%, which is much higher than other mainstream target detection algorithms at present. The experimental data show that the ACFP-YOLO algorithm has more advantages in the underwater target detection task. From the unilateral point of view of detection speed, compared with models of the same scale, the improved model maintains a medium-to-high level of detection speed and has good real-time performance. Compared with the mainstream two-stage target detection algorithm Faster-RCNN (ResNet50), the improved YOLOv7 algorithm is 18.0% higher in accuracy and 152.97% faster in speed. The model size has increased by 9.245M. Compared with YOLOv5-l, the most widely used one-stage target detection algorithm in industry, mAP has increased by 3.45%, the detection speed has decreased by 1.29%, and the model size has decreased by 9.092M. Compared with the YOLOv7 algorithm, the improved ACFP-YOLO algorithm mAP increased by 1.64%, the detection speed decreased by 2.62%, and the model size increased by 0.345M. It shows that the improved ACFP-YOLO algorithm in this paper is an algorithm with high detection accuracy. While improving the

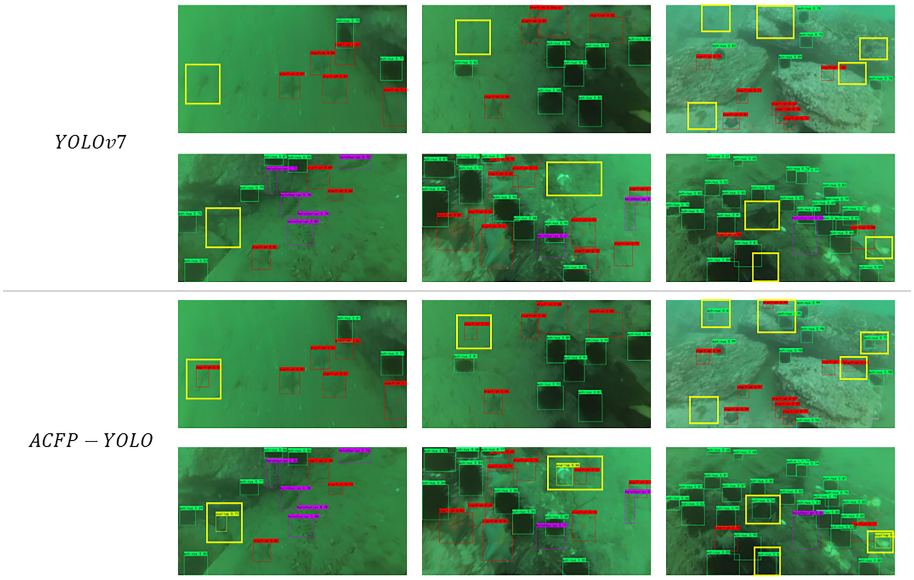

FIGURE 8
The detection effect of YOLOv7 and the detection effect of ACFP-YOLO.

detection accuracy of the model, it loses a small part of the detection speed.

In order to prove the superiority of the ACFP-YOLO algorithm in underwater scene performance, we use the same method to conduct a comparative experiment again on the underwater garbage dataset. We chose the detection effect of YOLOv7 and ACFP-YOLO, which has the best detection effect among other mainstream algorithms, for comparison. The Figure 9 shows the detection effect of ACFP-YOLO and YOLOv7 on the underwater garbage dataset. The original data set images are obtained at different water quality, different depths, and using different cameras. There are situations where the target features are not obvious, the features are attenuated, and the image background is complex. For this kind of picture, YOLOv7 shows a good detection effect, but ACFP-YOLO shows a better effect, we marked a part of the target objects with black boxes in Figure 9. Some of them were considered to be backgrounds by YOLOv7 because of the inconspicuous target features and small targets, and they were not detected. However, ACFP-YOLO has better detection ability for targets in this situation, and there is also a situation where the targets are occluded by each other. ACFP-YOLO can still detect it, while the detection effect of YOLOv7 is not as good as that of ACFP-YOLO. Through qualitative analysis, it can be seen that

TABLE 4 Performance metric values of mainstream target detection algorithms on URPC dataset.

Model	Input shape	mAP (%)	FPS	GFLOPS (G)	params (M)
SSD	416×416	57.73	100.93	116.159	24.146
YOLOv3	416x416	65.40	93.08	65.626	61.545
YOLOv4-tiny	416x416	52.19	215.96	6.84	5.883
YOLOv4	416x416	61.96	68.91	59.982	63.959
ACFP-YOLO	416x416	78.22	67.35	44.439	37.216
Efficientdet-D1	640x640	51.23	32.47	11.544	6.558
Faster-RCNN-VGG	600x600	59.41	36.51	369.817	136.771
Faster-RCNN-ResNet	600x600	62.62	25.40	940.972	28.316
YOLOv5-s	640x640	72.47	101.13	16.511	7.074
YOLOv5-l	640x640	77.17	65.05	114.627	46.653
YOLOv7-tiny	640x640	72.73	132.18	13.215	6.025
YOLOv7	640x640	78.98	65.94	105.182	37.216
ACFP-YOLO	640x640	80.62	64.21	105.191	37.561

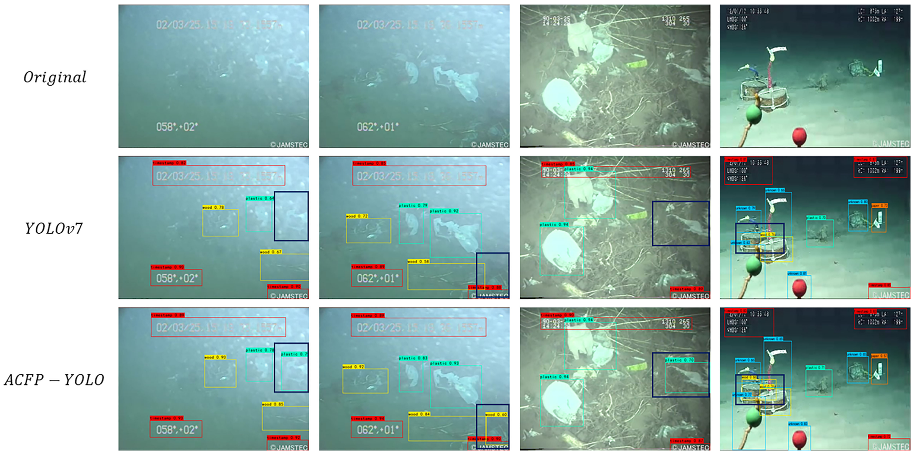

FIGURE 9
The first line Original is the original image of the dataset, the second line is the detection effect map of YOLOv7, and the third line is the detection effect map of ACFP-YOLO.

the ACFP-YOLO algorithm has better feature extraction ability in complex underwater scenes with different water quality and different depths, and has better detection ability for different color attenuation and occluded objects.

In the experiment of underwater garbage detection dataset, we selected consistent models for comparison, measured mAP value, GFLOPS value, params value and FPS of different models, and quantitatively analyzed the performance of each model through different indicators. All the comparative experimental measurement results in the experiment are shown in Table 5.

According to the data analysis in Table 5, from the perspective of detection accuracy, the mAP value of ACFP-YOLO algorithm on the underwater garbage dataset is 74.92%, which is still higher than other current mainstream target detection algorithms, the superior performance reflected in the underwater scene. The mAP value of ACFP-YOLO is 1.07% higher than that of YOLOv7, 2.60% higher than that of YOLOv5-l, and 4.38% higher than that of Faster-RCNN (ResNet). In terms of detection speed, the FPS value of the ACFP-YOLO algorithm is 65.56. Compared with models of the same scale, the detection speed remains at an upper-middle level and maintains good real-time performance. In terms of overall performance, the ACFP-YOLO algorithm improves mAP while losing a small part of the detection speed. It has better ability to extract features in complex underwater scenes. Our model is more suitable for efficient and accurate, and real-time requirements. Underwater missions.

TABLE 5 Performance Index Values of Mainstream Object Detection Algorithms on Underwater Garbage Detection Datasets.

Model	Input shape	mAP (%)	FPS	GFLOPS (G)	params (M)
SSD	461x416	60.21	143.85	61.951	25.216
YOLOv3	416x416	62.68	93.96	65.684	61.588
YOLOv4-tiny	416x416	56.03	217.79	6.853	5.902
YOLOv4	416x416	62.33	71.82	60.04	64.002
ACFP-YOLO	416x416	63.97	67.40	44.498	37.259
Efficientdet-D1	640x640	43.69	33.65	11.652	6.564
Faster-RCNN-VGG	600x600	69.89	43.81	370.014	136.935
Faster-RCNN-ResNet	600x600	70.54	26.89	941.071	28.398
YOLOv5-s	640x640	71.40	112.7	16.58	7.096
YOLOv5-l	640x640	72.32	67.30	114.765	46.696
YOLOv7-tiny	640x640	64.47	140.53	13.284	6.046
YOLOv7	640x640	73.85	66.83	105.32	37.259
ACFP-YOLO	640x640	74.92	65.56	105.328	37.604

In summary, the ACFP-YOLO algorithm, which integrates the CBAM attention mechanism and the SPPFCSPC module, achieves the highest detection accuracy compared with other mainstream algorithms in engineering applications, while maintaining a moderate level of detection and reasoning speed. Accuracy has a significant advantage in real-time underwater tasks.

Conclusion

In this paper, the ACFP-YOLO target detection model is proposed to address the problems of blurring and color deviation of images under the underwater map that make the extraction of object feature information difficult due to poor image quality. The model introduces the CBAM attention mechanism to enhance the extracted features in channel and spatial dimensions, which reduces the information loss in the feature extraction process and improves the overall feature extraction capability of the network, making the YOLOv7 model incorporating the attention mechanism have higher detection accuracy in underwater target scenes. The replacement of the SPPFCSPC module links the original three independent pooling layers together, reducing the model computation and obtaining faster model inference while keeping the perceptual field unchanged. By fusing the above two parts on YOLOv7, the improved ACFP-YOLO model has better performance in underwater target detection, and to a certain extent, solves the difficulties caused by the overlapping targets and complex background of underwater scenes to underwater target detection.

Our Future Work: Artificial intelligence has developed very rapidly, with numerous achievements in language translation, anomaly detection, target detection, and semantic segmentation, but few applications in intelligent exploitation of marine resources and underwater operations. In our future work, we will continue to study network models for underwater target recognition to improve the accuracy and speed of target recognition, and expand and enrich the dataset so that the models can be applied to more underwater scenarios with different conditions, and promote the application of AI in special underwater scenarios.

Data availability statement

Publicly available datasets were analyzed in this study, provided by 2022 China Underwater robot professional contest. This data can be found here: http://www.urpc.org.cn/index.html. Some experiments used the underwater garbage dataset, which can be found here: https://www.godac.jamstec.go.jp/dsdebris/e/index.html.

Author contributions

JY completed the main work of this paper. ZZ, ZX and BS guided JY in the research of this work. All authors contributed to the article and approved the submitted version.

Funding

This research was funded the Key Platform and Scientific Research Project for General Universities, Guangdong, 2022 (grant number 2022ZDZX4061), and the Science and Technology Program of Social Development, Zhuhai, 2022 (grant number 2220004000195). Our work was supported in part by the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, project code 2022B1212010006 and in part by Guangdong Higher Education Upgrading Plan (2021-2025) with UIC research grant R0400001-22 and R201902.

Acknowledgments

Throughout the process of conducting the experiment and writing the paper. I received great support and help, and received many valuable comments. I would like to express my sincere gratitude to all those who have helped me throughout the process. First of all, I would like to thank my supervisor, Zhou Zhuang, who gave me valuable advice on academic matters. This thesis would not have been possible without his patient guidance, insightful criticism and professional instruction. Secondly, I would like to thank Prof. Bing-Hua Su, who provided a good platform for me to complete my thesis writing. I would like to thank the professors for providing me with valuable advice. Last but not least, I would like to thank my parents, classmates and friends for their constant support and encouragement.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv*, 10934. doi: 10.48550/arXiv.2004.10934

Chen, L., Zhou, F., Wang, S., Dong, J., Li, N., Ma, H., et al. (2020). Swipenet: Object detection in noisy underwater images. *arXiv*, 10006. doi: 10.48550/arXiv.2010.10006

Fulton, M., Hong, J., Islam, M. J., and Sattar, J. (2019). "Robotic detection of marine litter using deep visual detection models," in *2019 international conference on robotics and automation (ICRA)* (New York: IEEE), 5752–5758.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 142–158. doi: 10.1109/TPAMI.2015.2437384

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* (New York: IEEE), 13713–13722.

Hu, J., Shen, L., and Sun, G. (2018). *Squeeze-and-excitation networks. computer vision and pattern recognition* (New York: IEEE).

Jiang, Z., and Wang, R. (2020). "Underwater object detection based on improved single shot multibox detector," in *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence.* (NY: ACM), 1–7.

Jocher, G. (2020) *Yolov5.* Available at: https://github.com/ultralytics/yolov5.

Klausner, N. H., and Azimi-Sadjadi, M. R. (2019). Performance prediction and estimation for underwater target detection using multichannel sonar. *IEEE J. Ocean. Eng.* 45, 534–546. doi: 10.1109/JOE.2018.2881527

Lab, P. (2018) *Underwater robot professional contest.* Available at: http://www.urpc.org.cn/index.html.

Lei, F., Tang, F., and Li, S. (2022). Underwater target detection algorithm based on improved yolov5. *J. Mar. Sci. Eng.* 10, 310. doi: 10.3390/jmse10030310

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. NY: IEEE, 2117–2125.

Lin, S., and Zhao, Y. (2020). Review on key technologies of target exploration in underwater optical images. *Laser Optoelectron. Prog.* 57, 060002. doi: 10.3788/LOP57.060002

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "Ssd: Single shot multibox detector," in *European Conference on computer vision* (Berlin: Springer), 21–37.

Liu, T.-Y., and Tie, YL. (2009). Learning to rank for information retrieval. *Found. Trends® Inf. Retr.* 3, 225–331. doi: 10.1561/1500000016

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (NY: IEEE), 8759–8768.

Qiang, W., He, Y., Guo, Y., Li, B., and He, L. (2020). Exploring underwater target detection algorithm based on improved ssd. *Xibei Gongye Daxue Xuebao/J. Northwest. Polytech. Univ.* 38, 747–754. doi: 10.1051/jnwpu/20203840747

Qiu, S., Xu, M., Jin, W., Yang, J., and Guo, H. (2019). Radon transform detection method for underwater moving target based on water surface characteristic wave. *Acta Optica Sin.* 39, 25–37. doi: 10.3788/AOS201939.1001003

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (NY: IEEE), 779–788.

Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (NY: IEEE), 7263–7271.

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv*, 02767. doi: 10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 28, 91–99. doi: 10.48550/arXiv.1506.01497

Shi, P., Xu, X., Ni, J., Xin, Y., Huang, W., and Han, S. (2021). Underwater biological detection algorithm based on improved faster-rcnn. *Water* 13, 2420. doi: 10.3390/w13172420

Tan, M., Pang, R., and Le, Q. V. (2020). "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* (NY: IEEE), 10781–10790.

Tinghui, H., Xinyu, G., Chunde, H., and Yueping, H. (2022). Research on underwater target detection algorithm based on fattention-yolov5. *Microelectron. Comput.* 39, 60–68. doi: 10.19304/J.ISSN1000-7180.2021.1261

Villon, S., Chaumont, M., Subsol, G., Villéger, S., Claverie, T., and Mouillot, D. (2016). "Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and hog+ svm methods," in *Advanced Concepts for Intelligent Vision Systems- 17th International Conference, (ACIVS) 2016.* (Italy: Lecce), 160–171.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv*, 02696. doi: 10.48550/arXiv.2207.02696

Wang, Q., Wu, B., Zhu, P., Li, P., and Hu, Q. (2020). "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* NY: IEEE.

Wei, X., Yu, L., Tian, S., Feng, P., and Ning, X. (2021). Underwater target detection with an attention mechanism and improved scale. *Multim. Tools Appl.* 80, 33747–33761. doi: 10.1007/s11042-021-11230-2

Woo, S., Park, J., Lee, J.-Y., and Kweon, S. (2018). Cbam: convolutional block attention module. *Eur. conf Comput. Vision* 10, 978–973. doi: 10.1007/978-3-030-01234-2_1

YU, H. (2020). Research progresson object detection and tracking techniques utilization in aquaculture: a review. *J. Dalian Ocean Univ.* 35, 793–804. doi: 10.16535/j.cnki.dlhyxb.2020-263

Zeng, L., Sun, B., and Zhu, D. (2021). Underwater target detection based on faster r-cnn and adversarial occlusion network. *Eng. Appl. Artif. Intell.* 100, 104190. doi: 10.1016/j.engappai.2021.104190

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (NY: IEEE), 2921–2929.