Received: 14 April 2023 | Revised: 23 March 2024 | Accepted: 8 April 2024

IET Image Processing

DOI: 10.1049/ipr2.13112

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

YOLO-UOD: An underwater small object detector via improved efficient layer aggregation network

Weiwen Chen[1,2] | **Tingting Zhuang**[3] | **Yuanfang Zhang**[4] | **Teng Mei**[1,2] | **Xiaoyu Tang**[1,2,3]

[1]School of Physics, South China Normal University, Guangzhou, China

[2]National Demonstration Center for Experimental Physics Education, Guangzhou, China

[3]School of Data Science and Engineering, Xingzhi College, South China Normal University, Shanwei, China

[4]The Autocity (Shenzhen) Autonomous Driving Co., Ltd., Shenzhen, China

Correspondence
Xiaoyu Tang, School of Physics, South China Normal University, Guangzhou, China.
Email: tangxy@scnu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 62001173; Climbing Program for College Students of Guangdong Province, Grant/Award Numbers: pdjh2022a0131, pdjh2023b0141

Abstract

Accurate detection of underwater objects is a key indicator technology to effectively enhance the field of marine development and application, and is of great importance to various fields including marine military defense and seafood aquaculture. Efficient and rapid detection of underwater targets is a crucial technological challenge in this field. To meet the challenges posed by these issues, this study applies the convolutional omni-efficient layer aggregation network (CO-ELAN) module to the detector backbone to improve the ability of the network structure to acquire underwater objects from image information. The module improves the feature representation of gradient branching through a multi-dimensional dynamic convolution and attention mechanism. In terms of loss calculation, the optimized normalized Wasserstein distance approach is used to predict the box distribution probabilistic modelling method to determine comparable distances to the ground box and obtain better samples of small target labels. Here, an underwater image enhancement algorithm based on white balance and underwater blur fusion is used to obtain clear images that enable improved detector performance. After the verification experiment on the URPC2018 dataset, it is found that the detector has better underwater detection ability compared with other detectors in the complex underwater environment. The proposed method achieves a 2.4% improvement over the YOLOv7 baseline model, while reducing computation costs by 5%.

1 | INTRODUCTION

Underwater object exploration is essential for various applications, including marine conservation, oceanography, and national defense [1, 2]. Accurately detecting and classifying these targets is one of the key steps in monitoring the health of marine ecosystems and identifying potential threats to underwater infrastructure [3–5]. At present, underwater target detection technology mainly consists of optical imaging, underwater sonar, and LIDAR detection methods. While underwater optical imaging has higher resolution and richer information, it has outstanding advantages in short-range underwater target detection tasks, but it is easily affected by the light in the water. The main reason is that light will be scattered and absorbed by water when propagating in underwater environment, resulting in attenuation and blurring of the image content captured by the final camera to a considerable extent, which further increases the difficulty of detecting underwater objects. In addition, the complexity of lighting conditions further limits the visibility of underwater environments. Underwater objects come in various sizes and shapes, encompassing various types such as marine organisms and marine debris, and their structures are also very complex, posing a challenge for target recognition and detection. Another reason is that there is a lot of background interference in the underwater environment, such as seaweed and rocks, which may be confused with the target and increase the difficulty of detection. Therefore, it is a challenging task to accurately and quickly detect objects in complex underwater environments with poor image quality.

At present, optical image target detectors can be broadly classified into two categories: underwater object detection based on traditional features, and underwater target detection using

deep learning networks. The majority of underwater object recognition algorithms that employ conventional features are designed manually and are well suited for objects with distinct characteristics. Sun et al. [6] proposed a multi-level wavelet transform method for underwater images to enhance the edge information of targets. This algorithm, which involves region comparison after preprocessing, demonstrates excellent performance in detecting targets in large-size underwater images. Feng et al. [7], considering the rapid decay of the R-channel in underwater optical imaging, introduced an R-channel image correction detection algorithm. This algorithm performs effectively for salient targets.

In recent years, the public have been enthusiastic about learning computer vision and machine learning techniques to detect underwater targets [8, 9]. One of the most promising approaches to solve this problem is deep learning-based object detection because of its generalization to feature extraction for complex-shaped objects. [10–12]. Sun et al. [13] innovatively introduced an underwater target detection model that utilizes a mobile vision transformer as its core component. By employing MobileViT as the central skeleton network algorithm, the model effectively enhances global feature extraction capabilities while decreasing the number of required parameters for the algorithm. However, this enhancement results in a significant increase in training time and computational resource consumption. To address issues of low target detection accuracy and poor real-time performance in complex underwater environments, Qiang et al. [14] proposed a fast object detection algorithm based on an optimized lightweight deep learning network. While this method improves these issues to some extent, it still falls short in detecting small targets. On the other hand, Li et al. [15] improved the underwater fish detection algorithm using transfer learning techniques to provide it with strong detection performance for tiny and hidden fish targets. Nonetheless, it is noteworthy that the algorithm's target detection effectiveness is not satisfactory in ambiguous underwater scenes.

However, detector acquisition of object features in underwater images is hampered by various environmental factors, and these difficult factors pose a challenge for underwater object detection tasks. Therefore, it is of practical significance to design an effective detector based on the characteristics of the underwater environment. In most mainstream deep learning detectors, the backbone network is an important structure for extracting object features. With the continuous development of detectors, various backbone network structures have been developed, such as VGGNet [16], ResNet [17], ResNetXT [18], ShuffleNet [19], and YOLOX [20]. However, underwater objects come in all shapes and sizes. And different sizes of receptive field of view will also affect the feature extraction ability of the convolutional module on objects of different scales [21–23]. General-purpose convolutional modules are typically employed to extract image features using a single, fixed-size convolutional kernel. However, the inherent limitation of a fixed perceptual domain associated with such kernels becomes evident when encountering underwater objects of diverse and substantial size variations. The fixed nature of the convolutional kernel's dimensions constrains its receptive field, thereby impeding the efficient extraction of meaningful features for both larger and smaller objects. This limitation hampers the thorough capture of critical object attributes, such as fine-grained textures, intricate shapes, and spatial configurations. Consequently, this design constraint undermines the model's capacity to accurately recognize and interpret underwater objects across a wide size spectrum, ultimately compromising the model's overall performance and versatility in scenarios involving underwater scenes characterized by varying object scales. Therefore, it is an important topic to design convolutional modules that can effectively capture and learn features of underwater targets of different scales to improve network performance. Furthermore, in complex and variable underwater environments, small marine organisms tend to intermingle and co-exist with various types of occlusions, resulting in a very limited number of effective pixels occupied by a single object during the imaging process. During underwater target detection, a slight deviation in the position of an object with fewer pixels often leads to a significant intersection over union (IoU) drop, which further leads to more false-negative samples in target detection [24], thus reducing the detector's ability to detect small underwater targets. On the other hand, the imaging process in underwater environments is affected by the significant light absorption and scattering effects of the water medium, resulting in severe image quality impairments in the form of blurred images, colour distortions (often characterized by a pronounced greenish bias), and reduced contrast. This series of degradation phenomena result in a large degree of distortion or loss of biometric information captured by the underwater camera, leaving less effective information for the detector to utilize, which greatly increases the difficulty and challenge of the underwater target recognition task.

To address the aforementioned challenges, this article proposes the YOLO-UOD detection model, an improvement of the first-order object detector based on YOLOv7 [25] for detecting small underwater objects. Figure 1 depicts the general framework of YOLO-UOD. In the underwater scene there are many objects of different sizes, even for the same species of organisms their interclass differences in size are also very large, resulting in their extremely different sizes presented in the image, coupled with the fact that they are often present in complex underwater backgrounds, these factors lead to the difficulty of the task of detecting multi-size targets in underwater. The current standard convolution builds a deep network by stacking and utilizes the feature maps of each layer to capture information about objects of different scales. However, the limited and fixed perceptual field of view of the ordinary convolution kernel makes it not sensitive enough to underwater targets of different sizes, and it is difficult to extract the overall features as well as the local details of these targets, which leads to the poor detection effect of the network model. In contrast, the multi-dimensional dynamic convolution mechanism can extract features of underwater objects at different scales by using different sizes of convolution kernels in different branches, where the small kernel convolution can capture the local details of subtle underwater objects, while the large kernel convolution has a larger field of view for recognizing large objects as well as

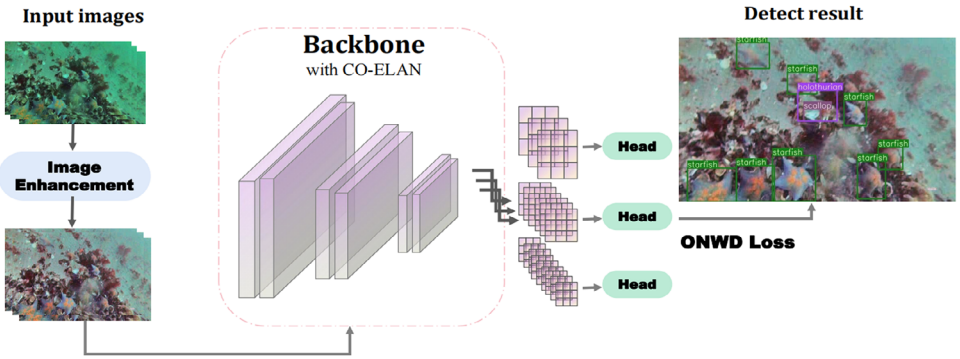

FIGURE 1 The YOLO-UOD detector.

obtaining more global information of the underwater image. In addition, learnable weight parameters are introduced in different convolutional kernel branches to dynamically correct the convolutional kernel size according to the input data. As a result, multi-dimensional dynamic convolution can flexibly adjust the perceptual field of view according to the image input target size in complex underwater scenes, to the extent that the detailed local features of multi-scale underwater targets as well as the global information (the information association between targets and backgrounds) can be effectively extracted from the multi-dimensional branches, thus improving the model's recognition accuracy of multi-scale underwater objects living in complex underwater backgrounds. To this end, we design a convolutional omni-efficient layer aggregation network (CO-ELAN) module based on a multi-dimensional dynamic convolution mechanism, which can flexibly adapt to objects of multiple sizes and efficiently extract object features. It is particularly suitable for underwater scenarios with complex and diverse objects, and can improve the model's ability to extract and learn multi-scale underwater target features. The current IoU method determines positive and negative samples by calculating the degree of overlap between two bboxes [26], but for small underwater targets, due to the scarcity of pixels, small positional deviations can lead to dramatic fluctuations in the IoU value, increasing the missed detection rate and degrading the model performance. The optimized normalized Wasserstein distance (ONWD) label evaluation method, on the other hand, determines the similarity between two bboxes by calculating the distance of their probability distributions (instead of using a simple overlapping region as the traditional IoU method), which can effectively measure the similarity even in the case of low overlap, and effectively increase the number of positive samples of the small targets, thus improve the network's ability to recognize dense small underwater objects. Consequently, this experiment adopts an optimized evaluation strategy, the ONWD label evaluation method, which aims to reduce the localization accuracy of the IoU for small target sensitivity, effectively classifies small targets into positive and negative samples, and optimizes its metrics for medium- and large-sized targets, which achieves a stable detection of small underwater targets while also ensuring the effective detection of normal-sized targets. Addressing the challenges posed by image blurring is critical as the degradation of the underwater environment results in limited valid image information available to computers. We hope to recover the characteristics of underwater targets, such as their shape, size, colour, and texture. These features are often blurred or distorted due to the detrimental effects of water absorption, scattering, and ambient light fluctuations. We would like to increase the available image information for underwater detectors by rendering these features more clearly in the image through image enhancement techniques. Here, we propose an underwater image enhancement scheme that uses defogging and colour reduction techniques to achieve this goal. The improvement of this paper can be summarized into the following three parts:

1. The YOLO-UOD detection model is specifically developed for complex underwater environments, and combined with the CO-ELAN architecture, it flexibly adapts to various object sizes and effectively extracts object features to cope with underwater scenarios with complex and diverse object shapes and sizes.
2. The model utilizes a labelling evaluation strategy that combines IOU and the ONWD to enhance its ability to detect small underwater objects, thereby reducing the likelihood of losing them.
3. An image enhancement algorithm based on the fusion of white balance and underwater dehazing has been proposed to reduce image blur caused by light and improved object detection performance in real-world underwater environments.

2 | RELATED WORK

Detecting underwater objects is a challenging computer vision task because of the particularity and complexity of underwater environment. With the continuous development and optimization of deep learning technology in the field of computer vision, and because deep learning-based detection algorithms extract complex object features with strong generalization, detection in real time, compared with traditional target detection algorithms have a huge improvement in accuracy, more and more

excellent detection algorithms have been widely used in underwater exploration task scenarios. Many excellent algorithms are introduced into deep learning object detection model to improve the performance of detectors [27].

One of the main research directions of current underwater target detection algorithms is to enhance the poor image quality to improve the detection effect. The second is to optimize the feature extraction structure of the detector to obtain more feature information of occlusion and small objects and to make the network structure lightweight to make the detector real time.

2.1 | Underwater image enhancement

General image enhancement techniques such as histogram equalization and contrast stretching have good results in relatively simple environments. However, these methods are sensitive to environmental factors and often fail to detect small- or low-contrast objects, resulting in poor image quality. To overcome these challenges, researchers have applied various image restoration methods to the original images collected underwater, aiming to restore the original features of the environment and enrich the image information. Specifically, Shi et al. [28] and Yang et al. [29] use contourlet transform, multi-scale retinex method, and contrast constraint adaptive histogram equalization method to enhance underwater images. Fan et al. [30] used generative adversarial networks [31] for image enhancement and restoration.

Other techniques used for underwater image enhancement include colour correction, image restoration, and fusion of multiple images. Colour correction aims to adjust the colour balance of the image, while image restoration aims to remove the blur and noise caused by the underwater environment. Fusion of multiple images involves combining several images taken at different exposures or under different lighting conditions to produce a high-quality image. Overall, many researchers are conducting underwater image enhancement research and have proposed many technologies to address the challenges of underwater environmental imaging.

2.2 | Efficient lightweight object detector

Most object detection applications primarily employ either faster single-stage target detectors or higher-accuracy two-stage target algorithms, with their performance varying due to the different methods and network structures utilized. While two-stage detectors excel in detection and classification accuracy, they fall short in real-time detection capabilities. Representative of the two-stage detection model is R-CNN [32], which has since inspired continuous algorithm optimization and development, leading to excellent two-stage algorithms such as Fast-RCNN [33] and Faster-RCNN [34]. On the other hand, the single-stage model boasts a unique advantage in detection speed but lacks detection accuracy. Notable among the one-stage detection models are the YOLO series [35] and the SSD model [36].

In recent years, the YOLO series has undergone continuous optimization, resulting in its detection accuracy surpassing that of most two-stage detectors. Moreover, it is deployable on mobile devices, making it a mainstream detector for real-time mission scenarios. In the realm of underwater detection, numerous researchers have undertaken research on underwater target detection using the YOLO algorithm. Liu et al. [37] employed a novel network structure in the YOLOv3 backbone to extract more effective image information, but this also significantly complicated the network architecture. Shi et al. [38] reduced the number of parameters in YOLOv4 and improved the network structure for deployment in underwater environments, albeit with room for improvement in detection accuracy. Huang et al. [39] introduced attention into YOLOv5, enhancing the network's spatial feature extraction ability while also increasing the model inference time.

The aforementioned research efforts significantly increase the number and complexity of model parameters while enhancing the backbone network. Besides, there is a need to further improve the accuracy of the detection model. To tackle these challenges, we will also investigate and refine an efficient and lightweight YOLO algorithm that achieves the dual effect of improving model detection accuracy while reducing resource consumption without excessively increasing the number of model parameters.

3 | PROPOSED METHODS

3.1 | Overview of the proposed methodology

YOLO-UOD, a novel algorithm for target detection systems, is introduced here. Built upon the YOLOv7 framework, it particularly addresses the challenge of detecting targets in underwater environments. The network structure of YOLO-UOD comprises a backbone, neck layers, and detection heads. The methodological advancements discussed here primarily focus on the backbone. We have designed a new module that enables the detector to efficiently learn local features of underwater targets of different shapes as well as sizes. Second, we introduce a loss computation method tailored for small- and medium-sized target bounding boxes, aiming to enhance detection accuracy further. Finally, we present an underwater image enhancement algorithm that can effectively augment the scene information of underwater images to improve the detector performance. In the following sections, we will provide detailed descriptions of these proposed improvements.

3.2 | Backbone layer with CO-ELAN

Due to the great variability in the size of underwater organisms and the fact that they live in complex underwater environments, these factors reduce the detection performance of most models. Here, we propose the CO-ELAN module to constitute the backbone of the detector, which can effectively improve the backbone's ability to recognize underwater multi-scale objects

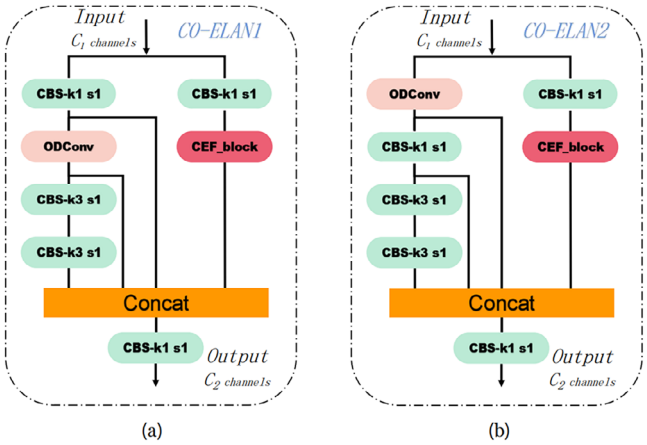

FIGURE 2 Detailed structure of the convolutional omni-efficient layer aggregation networks (CO-ELAN). Where (a) represents the CO-ELAN1 structure and (b) represents the CO-ELAN2 structure.

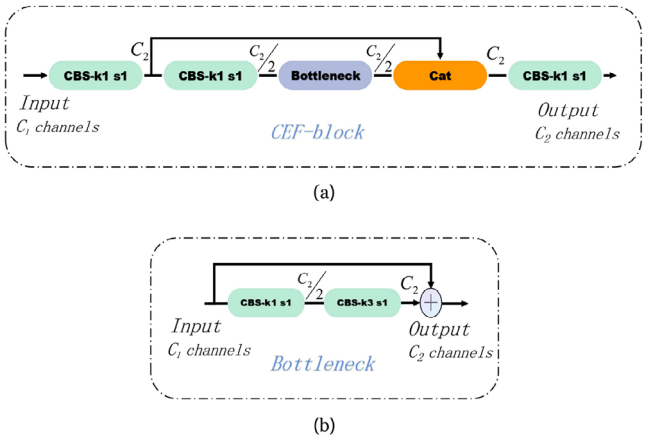

FIGURE 3 Detailed structure of the CEF block and bottleneck. Where (a) represents the CEF-block structure and (b) represents the bottleneck structure.

in complex underwater backgrounds. In the CO-ELAN module, in order to enable the convolution module to dynamically adjust the perceptual field of view according to the input information and capture the overall and detailed features of the multi-scale underwater target, we introduce multi-dimensional dynamic convolution in the main branch part in order to obtain the detailed local and overall multi-scale underwater target as well as the information between the target and the background. Finally, the features captured by the multiple superimposed convolution are separated and fused with the branch to obtain rich information about the local and overall features of the multi-scale target, as well as global information about the correlation between the multi-scale target and the background. The structure of CO-ELAN is illustrated in Figure 2, where (a) and (b) show two different forms of CO-ELAN. In the following, we will first introduce the cross-stage effective feature fusion block (CEF-block), a branch component of the CO-ELAN module.

Here, we introduced the CEF-block module and integrated it into the efficient layer aggregation network (ELAN) module to achieve the effect of branching structure to obtain more effective feature information. Our objective enables the network to obtain effective feature information of various objects in complex underwater scenes and detect small underwater targets more effectively. Here, the CEF-block module is added after the 1×1 convolution branch in ELAN to strengthen the channel information extraction capability of the branch part. Figure 3a shows the structure of CEF block in detail, where C_1 represents the input module channel and C_2 represents the final output module channel. The input image feature vector is represented as $\boldsymbol{X} \in R^{C1\times H1\times W1}$. The input feature vector in the module will first be processed by two 1×1 CBS (standard convolution, batch normalization, and silu activation functions) convolutions. The first convolution kernel outputs channel number C_2 and generates two branch feature maps y_1 and y_2. Then y_1 is input into the second convolution kernel, the number of output channels is reduced to $C_2/2$, and input the output tensor into the bottleneck structure for processing. The bottleneck output is connected with y_2, and 1×1 convolution kernel is input to adjust the module output channel number to C_2. In the convolutional neural network, the processing flow of general standard convolutional units is as follows: First, the input feature vectors will undergo convolution operation, and then batch normalization processing will be performed to prevent overfitting and achieve the effect of accelerated training. Finally, the normalized feature vector is activated by SiLU function and the result is output.

$$f_{\mathrm{CBS}}(x_{\mathrm{i}}, c_{\mathrm{o}}) = s(bn(C(x_{\mathrm{i}}, c_{\mathrm{o}}))) \tag{1}$$

The input feature vector diagram is represented by x_{i}, c_{o} represents the output channel after module processing, s represents the SiLU function, which will activate the feature information, bn represents the normalization operation, and C represents the general convolution operation.

$$Bottleneck(y_i) = y_i + f_{\mathrm{CBS}}\left(f_{\mathrm{CBS}}\left(y_i, \frac{C_2}{4}\right), \frac{C_2}{2}\right) \tag{2}$$

Following the dimensionality reduction and expansion through convolutional kernels, the resulting feature map will have an output channel of $C_2/2$. This feature map will be added to y_1, and the final output will be represented by *Bottleneck*, which is the bottleneck's output result.

$$CEF = f\left(concat\left(f(X_i, C_2), B\left(f\left(f(X_i, C_2), \frac{C_2}{2}\right)\right)\right), C_2\right) \tag{3}$$

The output result of the CEF-block module is represented by CEF. The input feature vector diagram is represented by X_i, C_1 represents the channel of the input module, and C_2 represents the output channel processed by the module. B indicates the final output of bottleneck structures, and *concat* indicates that you fuse the characteristics of different channel numbers. The CEF-block module can integrate channel information from

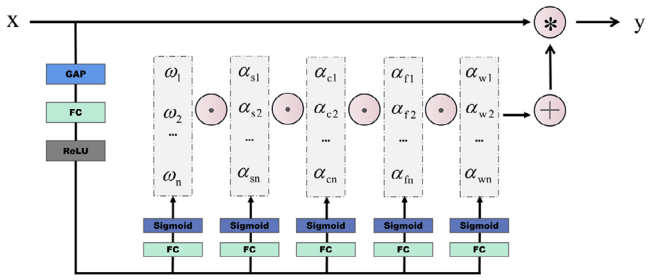

FIGURE 4 The structure of ODConv.

different channels and enhance the ability of branch structure to obtain valid information in the image.

The above is a detailed introduction of the CEF block in the CO-ELAN branch. In the main branch of CO-ELAN, we want to obtain more multi-dimensional spatial feature information of the image and use the appropriate range of sensory fields to deal with the feature information of different scales, so this paper introduces the ODConv convolutional structure [40] instead of some of the original basic convolutional units. ODConv is a new dynamic convolutional structure that utilizes a multi-dimensional attention mechanism (the dimensions of the number of input and output channels of the convolutional kernel, the dimension of the convolutional kernel's own receptive field, and the dimension of the number of convolutional kernels) to generate multi-dimensional spatial convolutional kernels. This structure can significantly improve the ability of the detector to acquire multi-dimensional feature information and reduce the computational cost and latency. The ODConv structure is shown in Figure 4. The use of ODConv in this method improves the multi-dimensional feature extraction capability and facilitates underwater target detection.Specifically, CO-ELAN uses ODConv to replace the convolutional kernel of size 3 and S, respectively, using CO-ELAN1 and CO-ELAN2 for the corresponding representations, as shown in Figure 2a,b.

Ultimately, the backbone layer portion of the detector consists of one CO-ELAN module without dynamic convolution and two CO-ELAN1 modules and one CO-ELAN2 module. In order to further improve the object localization in the network model and reduce the attention to the interference background, this experiment adds the coordinate attention (CA) [41] attention mechanism after the CO-ELAN of the network backbone to enhance the feature perception and position information of small underwater objects over long distances.

3.3 | ONWD applied to loss function

The optimal transport assignment (OTA) label assignment strategy [42] is employed in YOLOv7, which treats label assignment as a global optimal transport problem. However, this method mainly measures the similarity between two bounding boxes by calculating their IoU values. For small underwater targets, their limited pixel information and slight changes in scale can result in drastic changes in IoU values. Moreover, it is difficult to find stable and effective thresholds as evaluation metrics for general loss functions, making it challenging to provide high-quality sample labels for model training. Figure 5 shows the results of the IoU sensitivity analysis for both tiny and normal-sized targets, and the comparison illustrates the ability of the ONWD label evaluation method to perform an effective similarity assessment despite the relatively large offset displacements of small targets. ONWD label evaluation method changes the image pixels in the bbox into the form of a two-dimensional Gaussian probability distribution by assigning weights to the image pixels in the bbox, the center pixel in the bounding box has the highest proportion of weights, and the value of the weights decreases from the center to the boundary. Finally, calculate the Wasserstein distance between the two and normalize it to better evaluate the similarity of the two small target objects. With this approach, the similarity between the true labelled box and the predicted box of an underwater microminiature target can be effectively mapped to the Gaussian distance between them. The similarity between the two can also be measured when their overlap is relatively low, which effectively increases the number of positive samples required for model training and also performs well in similarity measurements of normal-sized objects. Among them, the simplified formula for the distribution distance is as follows:

$$W_2^2(N_a, N_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[cx_b, cy_b, \frac{w_b}{2} \cdot \frac{h_b}{2} \right]^T \right) \right\|_2^2 \tag{4}$$

N_a and N_b represent two-dimensional Gaussian distributions of prediction and target bboxes, respectively. cx, cy represent the center point, the width of the bboxes is indicated by w, and the height by h.

To further reflect the correlation between the two bboxes, map the distribution distance to a probability interval of 0 to 1, the original non-linear normalization function of NWD was optimized here. ONWD loss formula is shown below,

$$L_{\mathrm{ONWD}} = 1 - \left(\alpha * e^{-\frac{\sqrt{W_2^2(N_a, N_b)}}{c}} + \beta * e^{-\frac{1+\log\left(W_2^2(N_a, N_b)\right)}{c}} \right) \tag{5}$$

where α and β are proportionality coefficients, set at 0.5 in this case. c is a constant, which is set to the average size in the dataset.

The default loss function used in YOLOv7 is CIOU loss. Since the horizontal and vertical ratio of bbox is considered in the calculation formula, the regression speed can be better accelerated, and the regression accuracy is good and stable for detection targets of general size. Its calculation format is as follows:

$$L_{\mathrm{CIOU}} = 1 - \left(\mathrm{IOU} - \frac{d_o^2}{d_c^2} - \frac{v^2}{1 - \mathrm{IOU} + v} \right) \tag{6}$$

Where, the distance between the center of the actual target box and the prediction box is represented by d_o. d_c is the diagonal

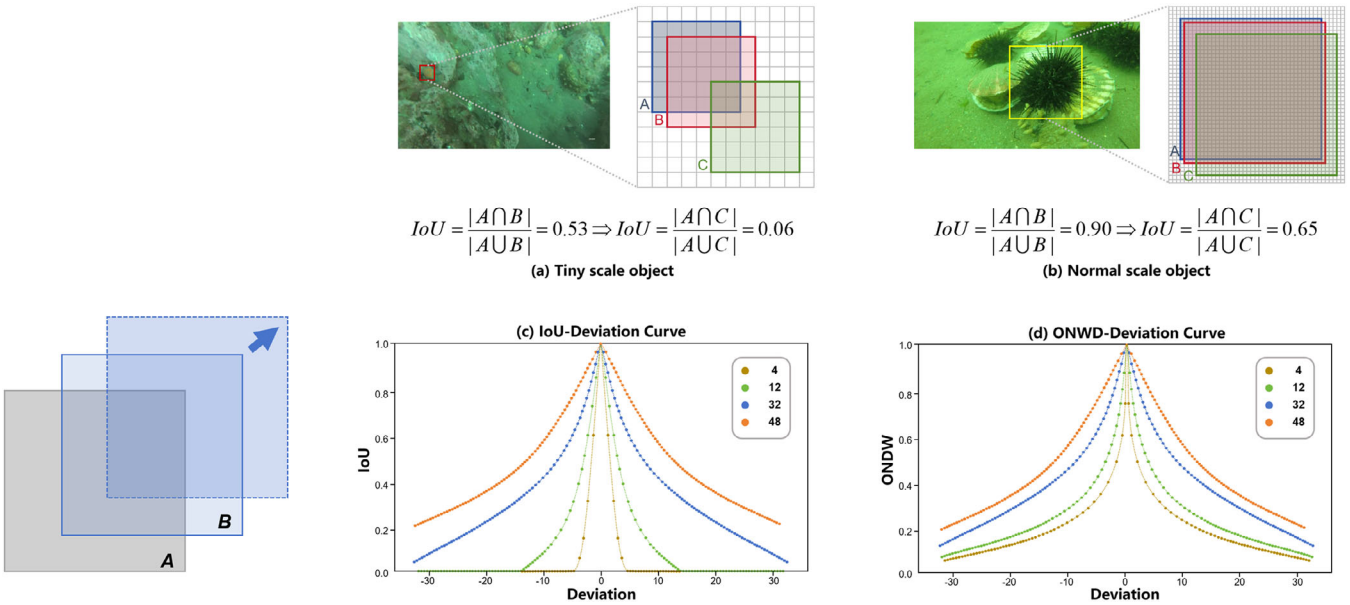

FIGURE 5 IoU sensitivity analysis for tiny as well as normal-sized targets, where A represents the real frame and B and C represent the predicted frames. (a) Demonstrates the drastic change in IoU for small targets in fine positional deviation (88.7% reduction), whereas the IoU for normal-sized targets demonstrated in (b) does not change so drastically (27.7% reduction). (c, d) Show the IoU as well as ONWD offset curves for four different sized targets, with the horizontal coordinates denoting the diagonal offset distance between the centers of the A and B frames, and the vertical coordinates denoting the similarity of the two metrics. It can be clearly seen that the smaller the target size, the faster the similarity curve of the IoU method decreases when offset by the same pixel distance, in contrast to the ONWD method, which is smoother and still capable of similarity evaluation when offset by a larger distance.

distance of the actual target box; v is used to confirm the ratio correlation between the target box and the prediction box.

Since CIOU does not consider the allocation of small target samples, ONWD can better solve such problems. Therefore, here, we integrated the optimized wd strategy into the loss calculation of the detector. We combined the evaluation results of ONWD and CIoU in a certain proportion as a better evaluation index of positive and negative label allocations. Below is the mathematical expression of the final loss function:

$$\text{LOSS} = k \times L_{\text{CIOU}} + (1 - k) \times L_{\text{ONWD}} \tag{7}$$

Here, the fusion loss function employs a scale coefficient k set at 0.7. This strategy enhances the precision and stability of tag allocation, rendering it less susceptible to small target scale variations, and more suitable for measuring similarity among densely populated underwater objects. As a result, the detector's performance is further improved.

3.4 | Underwater image enhancement

Here, how to change the poor quality images into better quality images to be fed into the detector is also one of our concerns and we hope to improve the performance of the detector in this aspect. In this underwater inspection mission, underwater visual signals are often affected by natural physical phenomena, resulting in low contrast, strong colour distortion, and high blurriness, thus posing significant challenges to underwater image processing.

Here, we thought about how to solve the problems such as the small amount of information that can be effectively processed by underwater images and the fuzzy image, we combine the white balance colour correction and defogging algorithm as a better underwater image enhancement method. The proposed method initially smoothes the image using a filter, followed by colour correction and white balancing, which effectively removes the interference caused by blue-green light. Subsequently, edge features of the image are enhanced by contrast stretching, and the channels are separated for defogging, thereby eliminating the blurring effect and resulting in clear and informative underwater images. The enhanced underwater image is illustrated in Figure 6. In our study, we applied the above image enhancement strategy to URPC2018 dataset, and re-divided the enhanced image into training set and test set according to the original proportion. The improved images were then evaluated using the YOLO-UOD detector proposed here.

4 | EXPERIMENTS

The following experimental platforms were used in this verification experiment: Ubuntu 18.04 system, Intel Core i7-9700K, NVIDIA RTX3060 graphics card, and CUDA framework version 11.4. In our experiment study, the training iteration process of each network model was 300 times. The SGD optimizer was used in the research. All images are scaled to a 640 × 640 scale. In order to ensure stable batch normalization and prevent overfitting, the training batch size was 8. In addition, Mosaic and

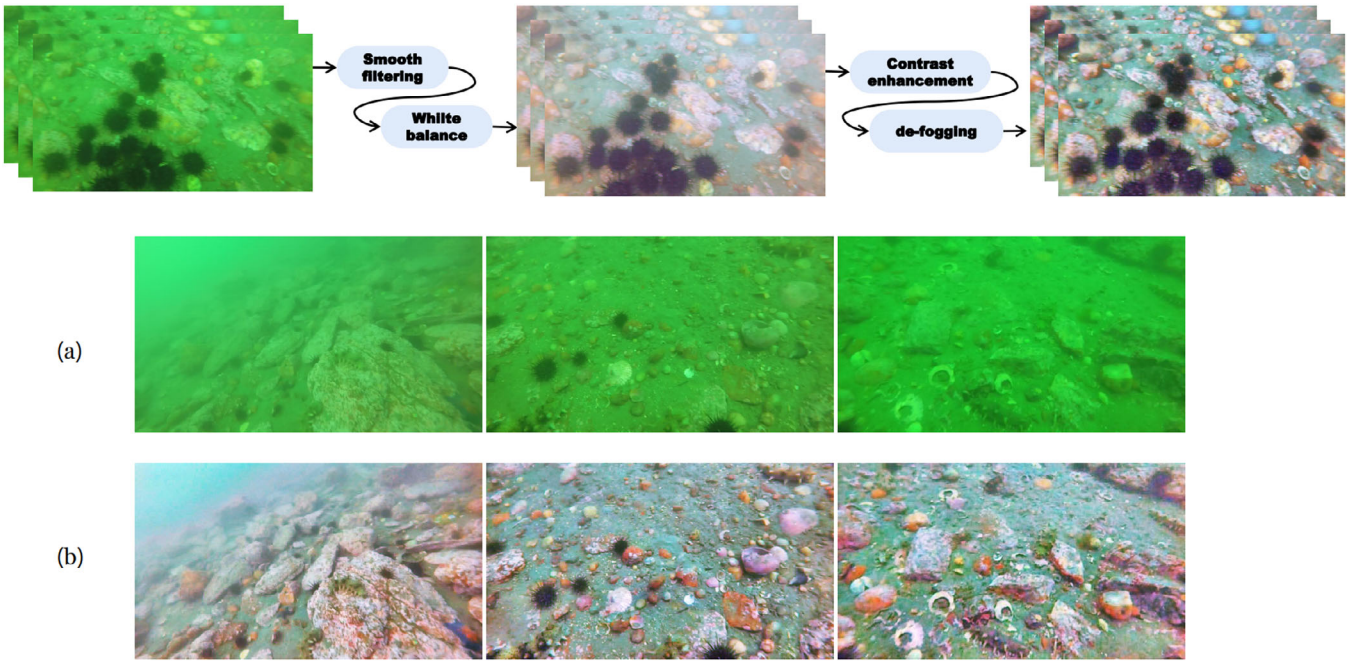

FIGURE 6 The figure below depicts the image enhancement process, wherein (a) represents the original input image and (b) shows the image after underwater enhancement. The enhanced image exhibits a greater amount of informative details, thus validating the effectiveness of our proposed image enhancement technique.

data scaling techniques are used to enhance the stability of the model in different situations.

4.1 | Dataset

In order to optimize the model for complex underwater environment, the underwater target detection dataset URPC2018 [43] of the National Underwater robot Grasping Competition was selected as the training and testing dataset in this experimental study. The dataset consists of underwater images captured from the natural seabed of Zhangzi Island, Dalian, China, and includes four distinct types of underwater creatures: echinus, holothurian, scallops, and starfish. The number and percentage of each type of label in the dataset are illustrated in Figure 7. In the underwater object detection dataset URPC2018, the training set contains 2901 image photos with marked information, while the test set contains 800 images without marked information.

The distribution of label sizes in URPC2018 dataset is shown in Figure 8. Most target labels in this dataset are small-size targets relative to the original image size, and the size scale is concentrated between 0 and 0.3. Additionally, the targets are distributed more dispersedly in the dataset. Furthermore, due to the underwater environment, the dataset contains images that are degraded with serious blurring. This degradation poses challenges in detecting small and occluded objects.

4.2 | Experimental evaluation index

During the experimental validation, we evaluated the performance of the experiment, including precision, recall rate, mean

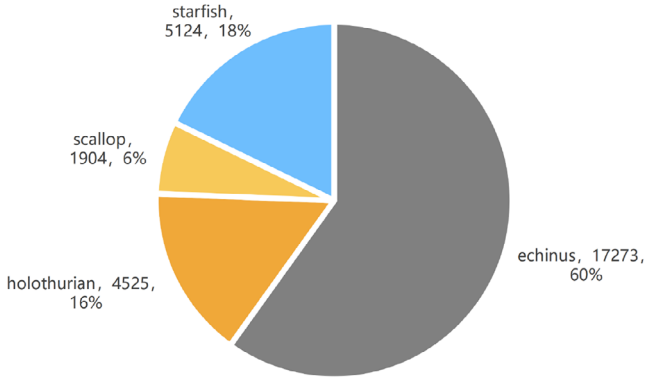

FIGURE 7 The URPC dataset contains multiple categories of underwater creatures, including echinus, holothurian, scallops, and starfish. The figure displays the number and proportion of each label category in the dataset.

average precision (mAP) under different IOU thresholds, computational resource consumption cost, and model parameters. In the target detection task domain, TP represents the true-positive samples in the predicted samples whose original labels are true being correctly categorized, FP represents the false-positive samples whose original categories are false but are predicted to be true, and FN represents the false-negative samples whose original categories are true but are judged to be false. The precision rate is defined as TP/(TP+FP), which represents the proportion of correctly predicted positive target labels in all correctly predicted categories. The recall rate is defined as TP/(TP+FN) and represents the proportion of correctly predicted labels to all real labels. By calculating the corresponding precision and recall values based on different detection thresholds, a corresponding set of precision and recall values can be obtained, which are plotted in a two-dimensional coordinate

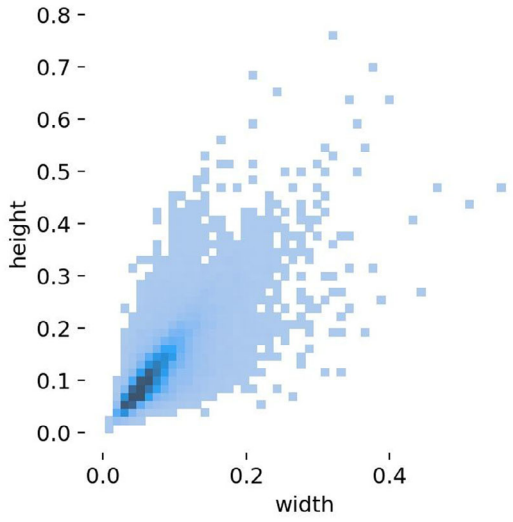

FIGURE 8 Object label size distribution in URPC2018 dataset.

system to form a precision-recall curve. This curve reflects how the accuracy of the model changes as recall increases. The average precision (AP) is the integral of the area formed between this PR curve and the recall axis and reflects the average precision of the model at all possible levels of recall. IOU is the ratio of the intersection and concatenation of the predicted results of a category in the data sample and the true values labelled in its sample.

mAP is an object detection performance index in the field of machine vision, which is used to evaluate the overall detection accuracy of object detection model for some data. It is obtained by calculating the AP value of all object classes that exist in the dataset. Generally, mAP is evaluated based on two evaluation criteria: $mAP_{0.5}$ and $mAP_{0.5:0.95}$, where $mAP_{0.5}$ represents the mean value of AP for each category with an IOU threshold of 0.5. $mAP_{0.5:0.95}$ indicates the average AP value of the IOU threshold between 0.5 and 0.95. How to evaluate the cost of computing resources consumed by the detection model can be calculated by calculating GFLOPs required by the detector to process the image. In general, the model size of the detector is evaluated and confirmed by the parameters to be calculated in the training of the model, which reflects the spatial complexity of the model.

4.3 | Comparison with other models

YOLOv7 is one of the most advanced stage target detectors available today, known for its excellent performance and versatility in target detection missions. Therefore, we chose it as the baseline to verify whether our proposed YOLO-UOD model is really effective. We also compared our approach to baseline models on the underwater detection dataset URPC2018 and other excellent probes for a more diverse experimental evaluation. See Table 1 for detailed comparative experimental data.

In comparison with the baseline detector YOLOv7, the new target detector we proposed achieved an impressive result in the experiment, reducing the computational cost by 5% after reducing 100.2G while only increasing the parameter count by 0.05%. Furthermore, precision increased by 1.6% and recall rate increased by 0.2%. In terms of the evaluation criteria $mAP_{0.5}$ and $mAP_{0.5:0.95}$, our model showed an improvement of 1.6% and 2.4%, respectively, indicating that it enhances detection performance in underwater environments with occlusion and dense small objects, while also reducing computational costs. In addition, we evaluate several advanced two-stage detectors, including the faster and classical detection model faster R-CNN, and further algorithm improvement cascade R-CNN [54] and sparse R-CNN [44] detection network. The experimental results revealed that these second-stage detectors performed poorly in underwater scene applications compared to our model, which not only has fewer parameters and lower computational costs than these detectors but also exhibits superior performance.

Compared to other one-stage detectors, our model exhibited a significant 10% improvement over both YOLOv3 and YOLOv4. Moreover, we achieved a 3% improvement over the widely used YOLOv5-l model and a 2% improvement over the high-performing YOLOX model. The performance of the detector proposed in this experiment is better than that of the above detection model in terms of detection accuracy and saving computing resource cost.

We also compared YOLO-UOD with YOLOv8-l, PPYOLOE+l, and RT-DETR, which are top performers in real-time detection tasks and demonstrate strong

TABLE 1 Model performance on the URPC2018 dataset.

Detectors	$mAP_{0.5}$ (%)	$mAP_{0.5:0.95}$ (%)	Param. (M)
Faster R-CNN	69.8	35.8	33.6
SSD	64.7	32.4	24.2
Sparse R-CNN [44]	61.2	30.9	106.1
Libra R-CNN [45]	68.8	36.0	41.4
FCOS [46]	69.7	34.7	32.0
ATSS [47]	62.1	29.3	32.1
RetinaNet [48]	65.3	32.2	36.2
Dynamic R-CNN [49]	66.9	35.8	41.5
Cascade R-CNN	69.2	37.0	68.9
HTDet [8]	76.3	38.5	7.7
YOLOv3	70.6	33.8	61.5
YOLOv4	76.9	39.2	52.5
YOLOv5-l	79.5	45.5	46.5
YOLOX	81.2	46.7	54.2
YOLOv7	81.5	46.1	36.9
YOLOv8-l	80.1	47.0	43.6
Deformable DETR [50]	69.9	33.8	40
DINO [51]	48.5	23.9	47
RT-DETR [52]	80.5	47.7	32
PPYOLOE+l [53]	82.9	48.6	52.2
YOLO-UOD(ours)	**83.1**	**48.5**	**37.3**

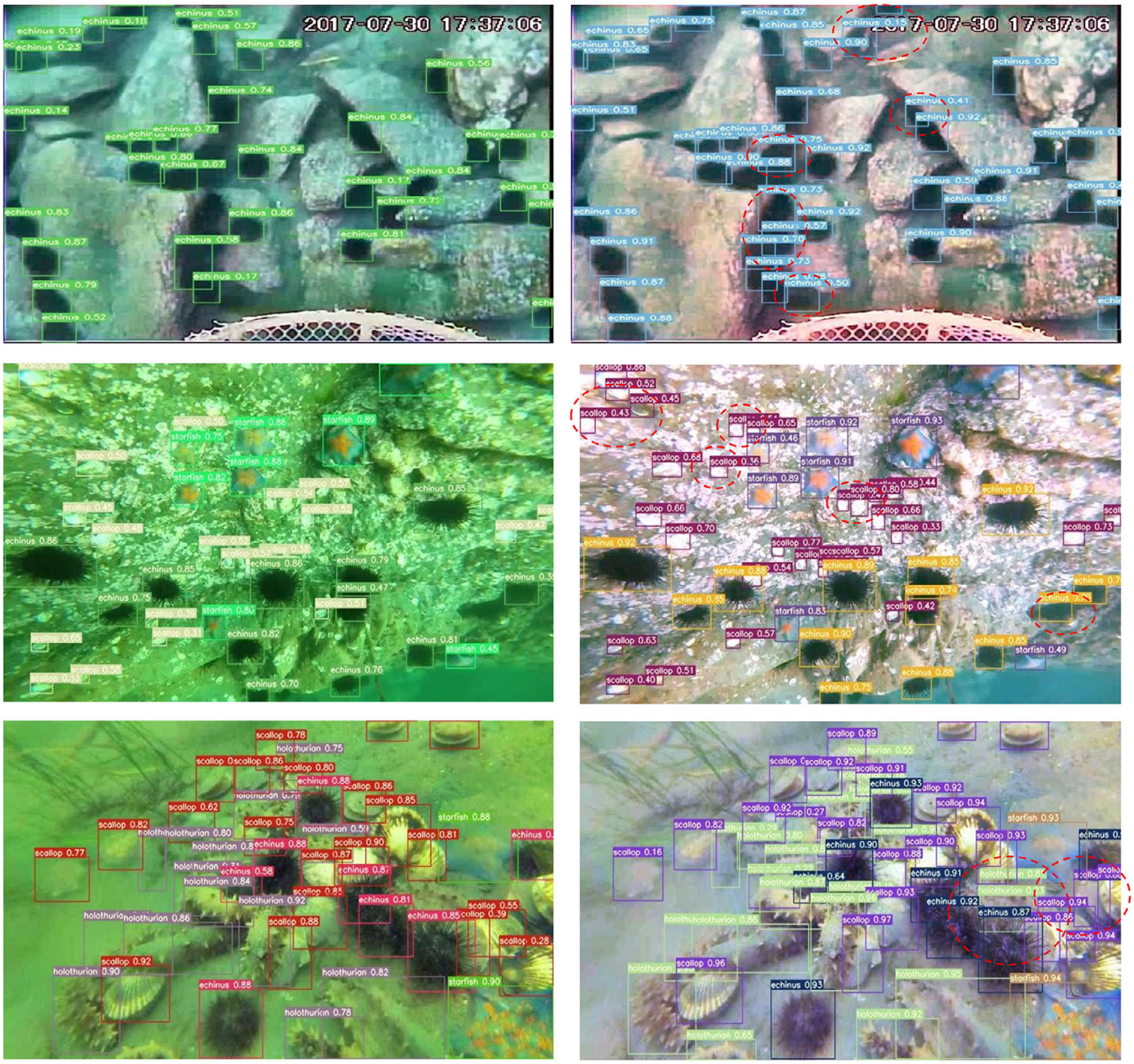

FIGURE 9 The figure includes the detection results of YOLOv7 baseline model and YOLO-UOD on the underwater dataset URPC2018. The differences between YOLOv7 and YOLO-UOD identification results are highlighted in red. The identification boxes in different colours correspond to the four categories in the URPC2018 dataset.

performance in underwater object detection. In this underwater detection task, our model outperforms YOLOv8-l significantly. The leading real-time single-stage detector is RT-DETR, a real-time fast visual transformer-based detector, which has excellent performance on the underwater task. In comparison, our model improves by 0.8% on $mAP_{0.5:0.95}$ and nearly 2.5% on $mAP_{0.5}$. The increase in the number of parameters is within acceptable limits and the overall performance has improved nicely. PPYOLOE+l exhibits surprisingly good performance on underwater small target detection. Our model reaches the same level of *mAP* as PPYOLOE+l on this task while requiring a significantly smaller number of parameters. Additionally, we compared YOLO-UOD with DINO and deformable DETR, both based on visual transformers. Our model outperformed them in detecting objects in complex underwater environments. We speculate that the difference in performance is due to the complexity of underwater environments and their reliance on richer data for further improvement.

As shown in Figure 9, YOLO-UOD detected fuzzy objects and small objects that YOLOv7 did not detect, and made

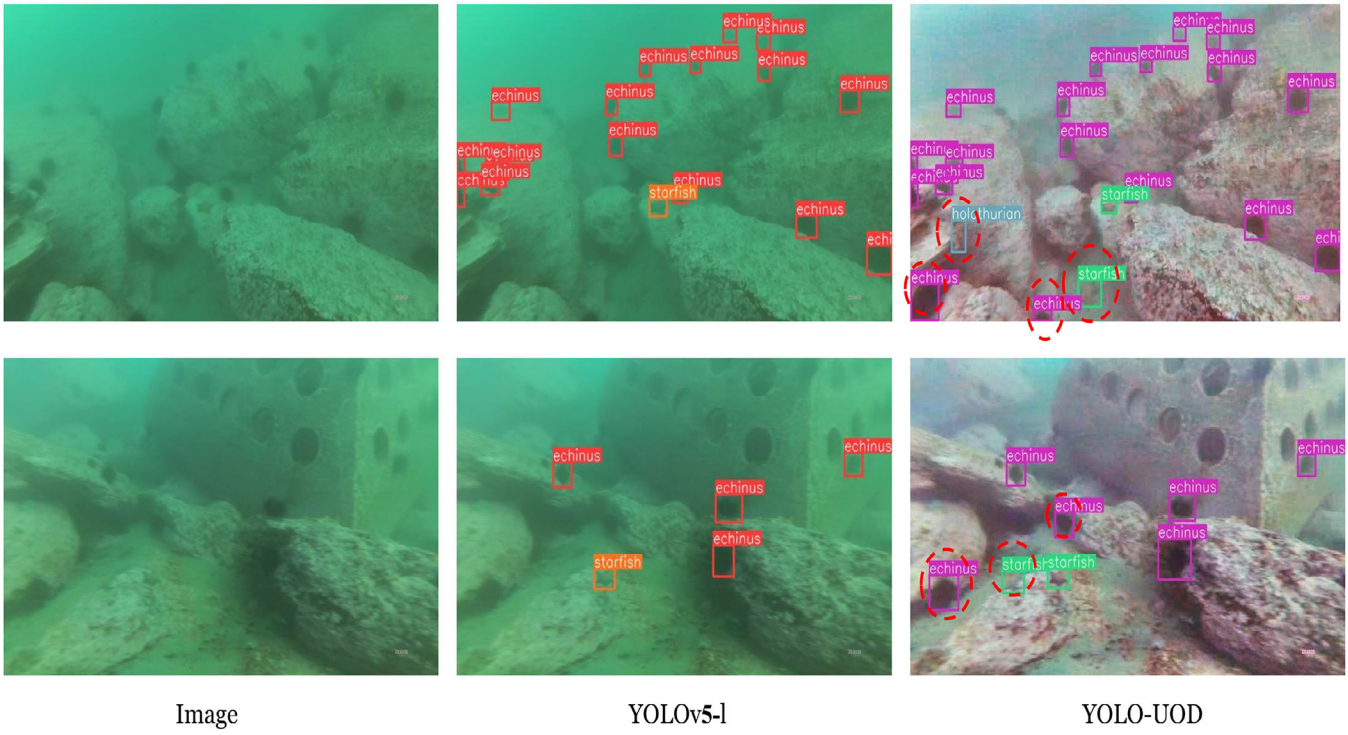

FIGURE 10 The figure shows the results of a comparative experiment between the more classical detector YOLOv5-l and YOLO-UOD models on the URPC2018 dataset. The differences between the YOLOv5-l and YOLO-UOD identification results are highlighted in red. The identification boxes in different colours correspond to the four categories in the URPC2018 dataset.

TABLE 2 Comparison of different models with CEF block.

Models	Precision (%)	Recall (%)	$mAP_{0.5}$ (%)	Param. (M)	FLOPs (G)
YOLOv5-l	80.5	74.9	79.5	46.5	109.1
YOLOv5-l + CEF block	81.4	72.0	80.1	46.4	150.1
YOLOv7	81.2	75.6	81.5	36.9	105.2
YOLOv7 + CEF block	80.6	76.8	82.2	38.3	111.7

accurate classification. For example, the obscured echinus and the scallop with a similar colour to the background can both have better detection effect, which also shows the superiority of YOLO-UOD compared with the baseline.

Upon inspection of Figure 10, it is evident that our proposed model is capable of detecting obscured and blurred objects in complex underwater environments, demonstrating the practical effectiveness of our method when compared to YOLOv5.

4.4 | Ablation experiment

In Table 2, we inserted the CEF-block module into both YOLOv7 and YOLOv5-l to verify its effectiveness. The validation method was to insert the CEF-block module into YOLOv7 and replace the C3 module with the CEF-block module in YOLOv5-l. The experimental results show that the CEF-block module increases a small amount of parameter volume but brings performance improvement.

We conducted comparative experiments on YOLOv7 and added several attention mechanism modules commonly used and effective in machine vision at present. Table 3 shows the experimental results of adding different attention mechanism modules into the baseline model. We find that introducing CA module into the baseline model can improve the identification accuracy of target detectors and reduce the model parameters to a certain extent. In contrast, other attention-mechanism modules we tested did not produce the same effect, suggesting that the coordinated attention module is effective in object detection tasks. The coordination attention module helps the model understand the image information better by coordinating the feature interactions of different scales, so as to improve the detection accuracy. Therefore, it is an effective attention mechanism to introduce coordinated attention module into target detection task.

Here, the model method was further ablated. As shown in Table 4, applying the CO-ELAN module in the baseline model can increase the detection result mAP by 1.3% and can improve

TABLE 3 The performance of various attention modules.

Attentions	Precision (%)	Recall (%)	$mAP_{0.5}$ (%)	Param. (M)	FLOPs (G)
YOLOv7(baseline)	81.2	75.6	81.5	36.9	105.2
YOLOv7 + CA	82.8	75.7	81.8	34.4	99.4
YOLOv7 + CBAM [55]	81.3	73.5	81.6	38.9	108.5
YOLOv7 + SE [56]	81.5	75.8	81.4	34.5	99.2
YOLOv7 + GAM [57]	81.3	71.5	79.5	49.9	149.1

TABLE 4 The ablation studies were conducted on three methods: CO-ELAN, ONWD, and white balance & defog(W&D).

Methods	CO-ELAN	ONWD	W&D	Precision (%)	Recall (%)	$mAP_{0.5}$ (%)	$mAP_{0.5:0.95}$ (%)
1	×	×	×	81.2	75.6	81.5	46.1
2	✓	×	×	81.1	76.5	82.8	47.4
3	×	✓	×	79.2	76.9	82.5	47.4
4	×	×	✓	81.6	76.1	82.2	48.1
5	✓	✓	×	82.0	71.6	82.9	46.4
6	✓	×	✓	81.0	75.8	82.7	47.7
7	×	✓	✓	81.8	75.5	82.4	47.9
8	✓	✓	✓	82.8	75.8	83.1	48.5

the recall rate by 0.9%. These results indicate that the feature extraction network equipped with CO-ELAN is capable of better capturing the details of underwater objects.With the application of the ONWD loss, the accuracy slightly decreased but the recall rate increased significantly, resulting in a 1.3% increase in mAP. This suggests that the new loss calculation method can reduce the impact of IoU perturbation on small objects, thereby improving performance. By enhancing the image with white balance and defogging operations, both precision and recall were improved, and mAP increased by approximately 2%, indicating that the increased image clarity and contrast allowed the detector to capture more effective information and reduce missed detections. When all of the above-mentioned methods were applied to the detector, both precision and recall were improved, and mAP increased by approximately 2.4%, reflecting the effectiveness of the methods.

The recognition accuracy of underwater objects and the distribution of concerns in images by the YOLO-UOD network model are displayed in Figure 11. Experimental results demonstrate the model's excellent accuracy in recognizing underwater objects and its effective attention to small objects in the thermal map.

In the complex underwater scenario, the proposed detection method can be trained normally and effectively in the dataset used in this experiment, so that the training loss values can be gradually decreased. Figure 12 shows the loss results of bbox, the loss results of object detection and classification results loss during the training process, where different colour curves represent different loss functions.

4.5 | Classification result analysis

We conduct a detailed analysis of the classifier performance on the URPC2018 underwater target detection dataset and provide a comprehensive confounding matrix, Figure 13 shows the classification results generated by our proposed model. From this graph, you can clearly see that the horizontal axis represents the actual labels of the URPC dataset, while the vertical axis represents the model's predicted values for each category. It is worth noting that the data in the figure uses a normalized value of 0 to 1 to show the results more clearly. The confusion matrix data provides a comprehensive analysis, and we can have a deeper understanding of the classification performance of the model and the accuracy of the classification results. Detailed detection data can be obtained from the graphs shown below. The classifiers have predicted echinus, holothurians, and starfish with an accuracy of over 80%, with the highest accuracy of 91% for sea urchins. However, the classifier still faces significant challenges in accurately classifying small-sized objects. The classification accuracy for scallops is relatively low due to their small size, and both scallops and holothurians are frequently misclassified as background because of their colour similarity with the underwater environment. As a result, the classification accuracy for these two categories is lower compared to other targets.

4.6 | Detection result analysis

We analysed the combined precision and recall curves for the URPC2018 dataset, whose the results are shown in Figure 14.

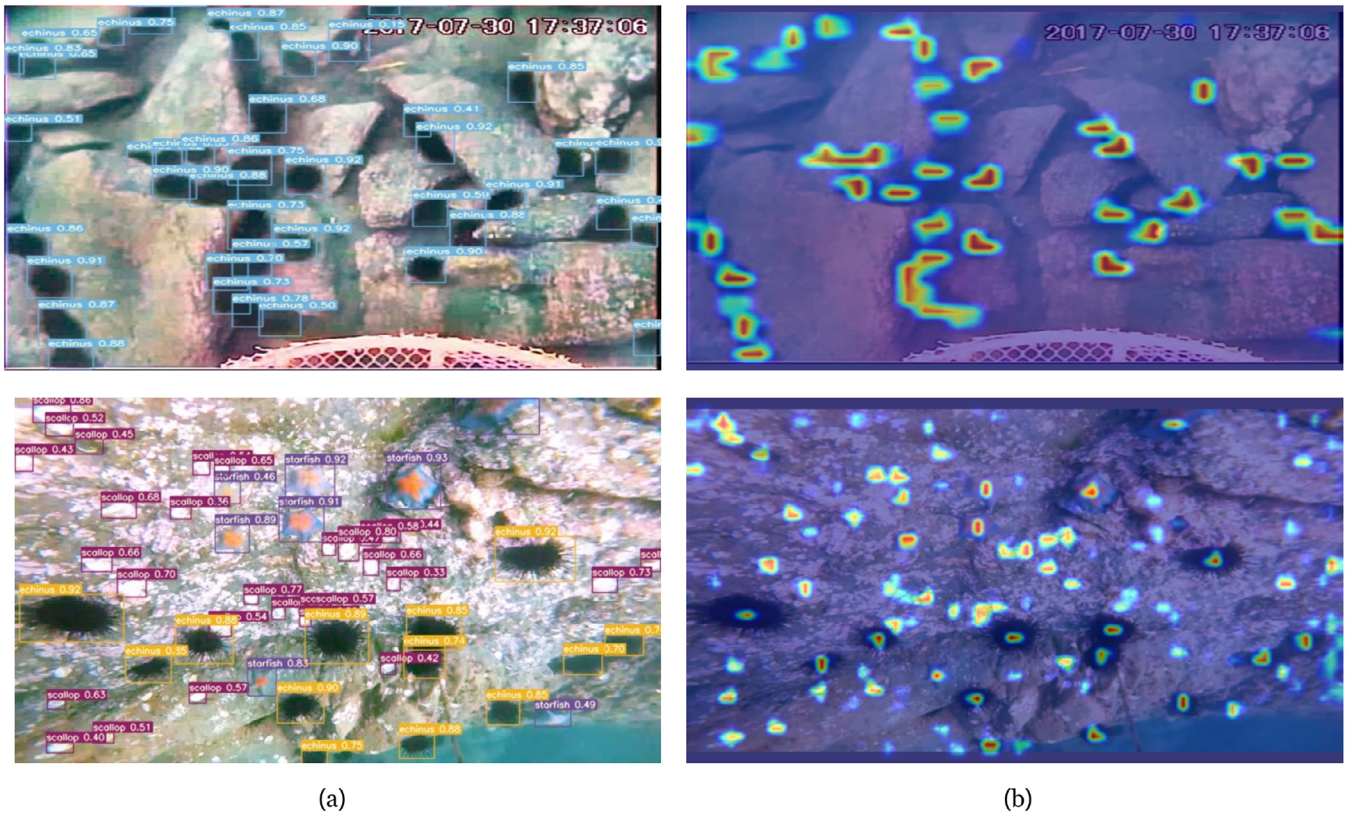

FIGURE 11 Figure includes (a) the detection result of YOLO-UOD, and (b) thermal maps generated by Grad-CAM for YOLO-UOD. In the thermal map, the darker colour indicates higher attention paid by the surface network model to the object.

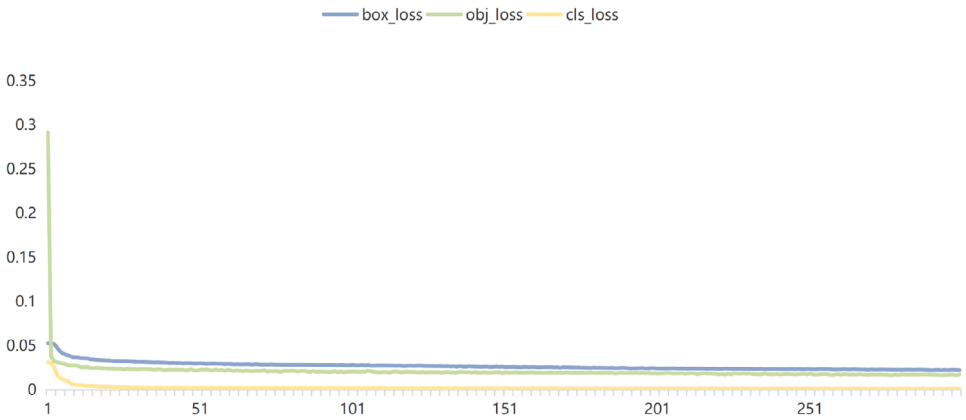

FIGURE 12 The loss function of YOLO-UOD detector training set.

AP represents the enclosed area between the precision curve and the recall curve for each category and the horizontal and vertical axes, while mAP calculates the average of all categories with respect to the curve area. According to the experimental results, our detector performed very well in detecting spiny snakes, probably because of the abundance of spiny snake tags and their unique characteristics in the dataset. In contrast, scallops, with fewer labels and smaller target pixels, blended in with the underwater background, exhibit the poorest detection performance among the four categories.

5 | CONCLUSION

We are thrilled to introduce our latest neural network architecture, YOLO-UOD, which effectively addresses the unique challenges of processing underwater images. By utilizing gradient branch superposition in combination with CO-ELAN modules, we are able to significantly enhance branch feature extraction. We further employ ODConv and coordinated attention modules to reduce parameter numbers, while our ONWD loss calculation strategy effectively improves small

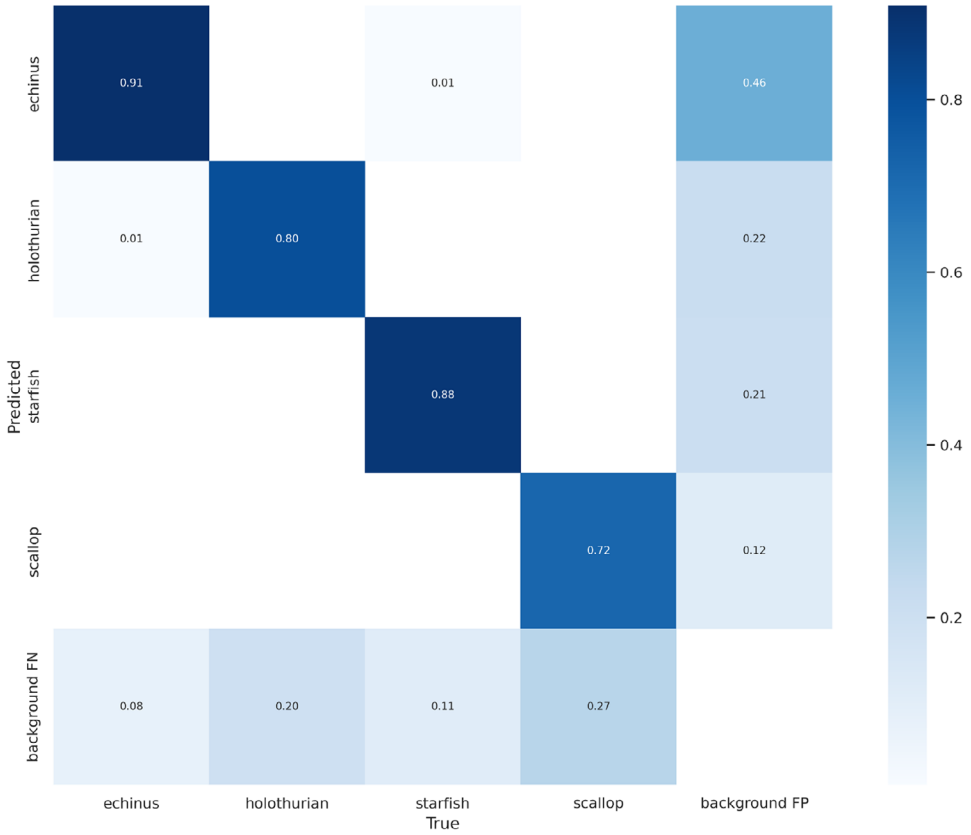

FIGURE 13 The presented confusion matrix is related to the classification results and the values within the confusion matrix vary from 0 to 1. The shades of colour in the graph reflect the magnitude of the values. Since the detector did not set the background class in the validation experiments on the URPC dataset, its value in the matrix was not set.

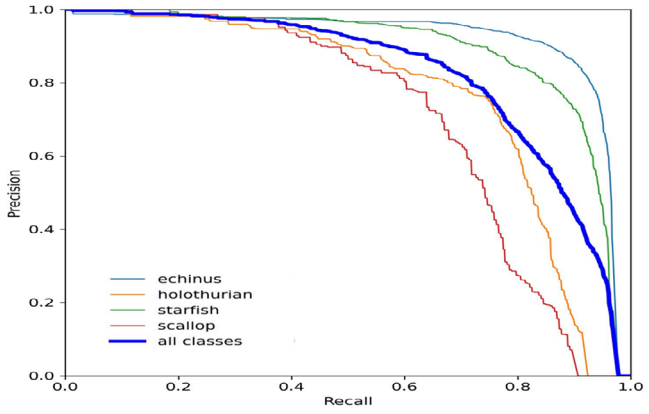

FIGURE 14 In the upper panel, plot precision and recall curves for each of the four different categories when the IOU threshold is 0.5. The different colour curves represent the levels of detection for different categories in the dataset, but the blue curve reflects the average detection results for all categories. The closed area of horizontal axis and vertical axis curve reflects the detection performance of the detector. Among these four categories, echinus has the best detection performance, possibly due to its rich labelling and distinctive features in the dataset. On the other hand, scallop has fewer labels, smaller target pixels, and a similar underwater background, resulting in the poorest detection performance.

target sample labelling. Additionally, we leverage colour balance and defogging algorithms to improve image sharpness and provide more effective information. Our experiments on the URPC2018 dataset demonstrate that our approach outperforms current detectors.

As we conducted our research, we discovered that the quality of underwater camera images has a significant impact on prediction performance. In light of this, we are committed to improving the quality of input images to maximize the effectiveness of our algorithm. Looking ahead, we aim to design faster and more accurate underwater identification algorithms, and to apply the feature extraction capabilities of our detector to create smaller, more efficient detectors that will greatly benefit underwater deployments.

AUTHOR CONTRIBUTIONS

Weiwen Chen: Methodology; writing—original draft; writing—review and editing. **Tingting Zhuang**: Data curation; writing—review and editing. **Yuanfang Zhang**: Software; writing—review and editing. **Teng Mei**: Data curation; writing—review and editing. **Xiaoyu Tang**: Conceptualization; supervision; validation.

ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (No. 62001173) and the Climbing Program for College Students of Guangdong Province (pdjh2022a0131 and pdjh2023b0141).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data openly available in a public repository.

ORCID

Xiaoyu Tang https://orcid.org/0000-0002-6038-9623

REFERENCES

1. Pang, Y., Wu, C., Wu, H., Yu, X.: Over-sampling strategy-based class-imbalanced salient object detection and its application in underwater scene. Visual Comput. 39, 1–16 (2022)
2. Hong, Y.: Research progress of aquatic animal target detection and tracking technology and its application. J. Dalian Ocean Univ. 35(6), 793–804 (2020)
3. Sen, L., Ying, Z.: A review of key techniques of target detection in underwater optical images. Adv. Laser Optoelectron. 57(6), 060002 (2020)
4. Zhang, A., Sun, G., Ma, P., Jia, X., Zhang, X.: Coastal wetland mapping with sentinel-2 MSI imagery based on gravitational optimized multilayer perceptron and morphological attribute profiles. Remote Sens. 11(8), 952 (2019)
5. Sharbain, H.A., Osman, A., El-Hag, A.: Detection and identification of ferroresonance. In: 2017 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), pp. 1–4 (2017)
6. Xin, S., Qing, L.: Wavelet transform-based fast detection algorithm for underwater salient targets. Comput. Eng. 49(4), 249–255 (2023)
7. Feng, H., XU, Z., Yin, X.: Underwater salient object detection based on red channel correction. In: 2021 2nd International Conference on Big Data and Artificial Intelligence and Software Engineering (ICBASE), pp. 446–449 (2021)
8. Chen, G., Mao, Z., Wang, K., Shen, J.: Htdet: A hybrid transformer-based approach for underwater small object detection. Remote Sens. 15(4), 1076 (2023)
9. Jia, J., Fu, M., Liu, X., Zheng, B.: Underwater object detection based on improved EfficientDet. Remote Sens. 14(18), 4487 (2022)
10. He, X., Zhaoqiong, H., Chen, L., Yonghong, Y.: Progress of deep learning in passive underwater target recognition. Signal Process. 35(9), 16 (2019)
11. Jun, X., Jianglei, D., Yuwen, Q.: Deep learning in underwater imaging technology (invited). Acta Photonica Sinica 51(11), 48 (2022)
12. Qiang, W., Xiangyang, Z.: Deep learning method and its application in underwater target recognition. Technical Acoust. 34(2), 140–143 (2015)
13. Sun, Y., Zheng, W., Du, X., Yan, Z.: Underwater small target detection based on YOLOX combined with MobileViT and bouble coordinate attention. J. Mar. Sci. Eng. 11(6), 1178 (2023)
14. Wei, Q.: Research on underwater target detection algorithm based on improved SSD. J. Northwest. Polytech. Univ. 38(4), 747–754 (2020)
15. Qing, L., Yi, L., Jiong, N.: Real-time underwater fish target detection based on improved YOLO and migration learning. Pattern Recognit. Artif. Intell. 32(3), 193–203 (2019)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
18. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
19. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
20. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding YOLO series in 2021. arXiv:2107.08430 (2021)
21. Christian, S., Wei, L., Yangqing, J.: Going deeper with convolutions. arXiv:1409.4842 (2014)
22. Xiang, L., Wenhai, W., Xiaolin, H.: Selective kernel networks. arXiv:1903.06586 (2019)
23. Brandon, Y., Gabriel, B., Jiquan, N.: CondConv: Conditionally parameterized convolutions for efficient inference. arXiv:1904.04971 (2020)
24. Wang, J., Xu, C., Yang, W., Yu, L.: A normalized Gaussian Wasserstein distance for tiny object detection. arXiv:2110.13389 (2021)
25. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOV7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696 (2022)
26. Zhaohui, Z., Ping, W., Wei, L.: Distance-IoU loss: Faster and better learning for bounding box regression. arXiv:1911.08287 (2019)
27. Amarasinghe, C., Ratnaweera, A., Maitripala, S.: Hybrid feature enhance filter (HFEF) for underwater vision navigation. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), pp. 420–426 (2020)
28. Dan, S., Qingwu, L., Xinnan, F., Guanying, H.: Underwater image enhancement algorithm based on contourlet transform and multi-scale rentinex. Adv. Laser Optoelectron. 47(4), 41–45 (2010)
29. Weizhong, Y., Yinli, X., Xi, Q., Wei, R., Daoliang, L., Zhenbo, L.: Underwater sea cucumber image enhancement method based on contrast limited histogram equalization. Trans. AES 32(6), 7 (2016)
30. Xinnan, F., Xin, Y., Pengfei, S., Song, H., Yuanxue, X.: Feature fusion to generate underwater image enhancement for adversarial networks. J. Comput. Aided Des. Graph 34(2), 9 (2022)
31. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE Signal Process Mag. 35(1), 53–65 (2017)
32. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
33. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer vision, pp. 1440–1448 (2015)
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017)
35. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
36. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., et al.: SSD: Single shot multibox detector. In: Proceedings of 14th European Conference on Computer Vision–ECCV 2016, pp. 21–37 (2016)
37. Ping, L., Hongbo, Y., Yang, S.: Improved marine biometric identification algorithm for YOLOV3 network. Comput Appl. Res. (S01), 4 (2020)
38. Xianpeng, S., Honggui, W.: Lightweight underwater target detection network improved based on YOLOV4. J. Harbin Eng. Univ. 44(1), 7 (2023)
39. Tinghui, H., Xinyu, G., Chunde, H., Yueping, H.: Research on underwater target detection algorithm based on fattention-YOLOV5. Microelectronics Comput. 39(6), 60–68 (2022)
40. Li, C., Zhou, A., Yao, A.: Omni-dimensional dynamic convolution. arXiv:2209.07947 (2022)
41. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)
42. Ge, Z., Liu, S., Li, Z., Yoshie, O., Sun, J.: Ota: Optimal transport assignment for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 303–312 (2021)
43. Chen, L., Liu, Z., Tong, L., Jiang, Z., Wang, S., Dong, J., et al.: Underwater object detection using invert multi-class adaboost with deep learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2020)
44. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., et al.: Sparse R-CNN: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14454–14463 (2021)
45. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra R-CNN: Towards balanced learning for object detection. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2019)
46. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636 (2019)
47. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9759–9768 (2020)
48. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
49. Zhang, H., Chang, H., Ma, B., Wang, N., Chen, X.: Dynamic R-CNN: Towards high quality object detection via dynamic training. In: Proceedings of 16th European Conference on Computer Vision–ECCV 2020, pp. 260–275 (2020)
50. Xi, Z., Wei, S., Le, L.: Deformable DETR: Deformable transformers for end-to-end object detection. arXiv:2010.04159 (2021)
51. Hao, Z., Feng, L., Shi, L.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv:2203.03605 (2022)
52. Wen, L., Yian, Z., Shang, X.: Global attention mechanism: DETRs beat YOLOs on real-time object detection. arXiv:2304.08069 (2023)
53. Shang, X., Xin, W., Wen, L.: PP-YOLOE: An evolved version of YOLO. arXiv:2203.16250 (2022)
54. Cai, Z., Vasconcelos, N.: Cascade R-CNN: High quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43(5), 1483–1498 (2019)
55. Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., et al.: Attention mechanisms in computer vision: A survey. Comput. Vis. Media 8(3), 331–368 (2022)
56. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
57. Liu, Y., Shao, Z., Hoffmann, N.: Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv:2112.05561 (2021)

How to cite this article: Chen, W., Zhuang, T., Zhang, Y., Mei, T., Tang, X.: YOLO-UOD: An underwater small object detector via improved efficient layer aggregation network. IET Image Process. 18, 2490–2505 (2024). https://doi.org/10.1049/ipr2.13112