Journal of
Marine Science and Engineering

MDPI

Article

Collaborative Framework for Underwater Object Detection via Joint Image Enhancement and Super-Resolution

Xun Ji [1], Guo-Peng Liu [1] and Cheng-Tao Cai [2,3,4,*]

[1] College of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China; jixun@dlmu.edu.cn (X.J.); 1120210179_lgp@dlmu.edu.cn (G.-P.L.)
[2] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
[3] Heilongjiang Provincial Key Laboratory of Environment Intelligent Perception, Harbin 150001, China
[4] Key Laboratory of Intelligent Technology and Application of Marine Equipment, Harbin Engineering University, Ministry of Education, Harbin 150001, China
* Correspondence: caichengtao@hrbeu.edu.cn

Abstract: Underwater object detection (UOD) has attracted widespread attention, being of great significance for marine resource management, underwater security and defense, underwater infrastructure inspection, etc. However, high-quality UOD tasks often encounter challenges such as image quality degradation, complex backgrounds, and occlusions between objects at different scales. This paper presents a collaborative framework for UOD via joint image enhancement and super-resolution to address the above problems. Specifically, a joint-oriented framework is constructed incorporating underwater image enhancement and super-resolution techniques. The proposed framework is capable of generating a detection-favoring appearance to provide more visual cues for UOD tasks. Furthermore, a plug-and-play self-attention mechanism, termed multihead blurpooling fusion network (MBFNet), is developed to capture sufficient contextual information by focusing on the dependencies between multiscale feature maps, so that the UOD performance of our proposed framework can be further facilitated. A comparative study on the popular URPC2020 and Brackish datasets demonstrates the superior performance of our proposed collaborative framework, and the ablation study also validates the effectiveness of each component within the framework.

Keywords: underwater object detection; underwater image enhancement; super-resolution; joint learning; deep learning

Citation: Ji, X.; Liu, G.-P.; Cai, C.-T. Collaborative Framework for Underwater Object Detection via Joint Image Enhancement and Super-Resolution. *J. Mar. Sci. Eng.* **2023**, *11*, 1733. https://doi.org/10.3390/jmse11091733

Academic Editor: Weicheng Cui

Received: 1 August 2023
Revised: 27 August 2023
Accepted: 28 August 2023
Published: 1 September 2023

1. Introduction

The exploration of marine environments has received tremendous attention due to the urgent demand for natural resource management and ecosystem monitoring [**?**]. In recent years, the use of remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs) has become increasingly prevalent in various ocean engineering-related applications. This trend further emphasizes the significant potential and value of the underwater imaging community. As an essential step in perceiving and understanding complex marine habitats, underwater object detection (UOD) technology plays a pivotal role in maritime target positioning [**?**], wreck salvage [**?**], underwater archaeology [**?**], and many other practical applications, providing an effective strategy to uncover the mysterious underwater world.

With the rapid development of deep learning, the great potential of object detection technology has been significantly stimulated [**?**]. At present, various schemes presented by convolutional neural networks (CNNs) have been developed to deliver state-of-the-art performance in complex UOD tasks [**?**]. However, as opposed to ordinary camera imaging, underwater optical imaging often encounters more challenges. The images captured in marine environments inevitably suffer from severe quality degradation due to light attenuation and scattering effects, resulting in undesired haziness, underexposure,

and color distortion [**?**]. Specifically, since red light with the longest wavelength can be strongly absorbed in water, underwater images typically present a blue or turquoise color palette. This poses challenges in accurately distinguishing different underwater objects. Additionally, the irregular congregations and movements of marine organisms can also lead to undesired image blur and clutter, significantly increasing the difficulty of precise UOD [**? ?**]. It should be noted that the low-end underwater imaging equipment will further exacerbate the aforementioned drawbacks due to associated cost and hardware constraints, leading to reduced imaging resolution and loss of detailed texture information [**?**].

In summary, the current challenges in UOD extend beyond determining the locations and categories of diverse underwater objects. It also involves the pursuit of non-degraded and high-resolution images to effectively highlight the distinctive features of these objects. This presents a significant obstacle that has yet to be adequately addressed in prior research [**?**]. As a straightforward solution, degraded underwater images can be pre-processed by underwater image enhancement (UIE) and super-resolution (SR) technologies, so that the UOD-related schemes can better capture the characteristics of diverse underwater objects, effectively improving the detection precision.

1.1. Underwater Object Detection

Object detection is considered a fundamental task in the computer vision community, which refers to determining the locations and corresponding categories of objects in a given image. Traditional methods typically require three stages to achieve object detection, including informative region selection, feature extraction, and classification [**?**]. Extensive experiments have demonstrated that the traditional methods for object detection are typically intuitive and easy to implement. However, these methods are extremely limited in feature representation, which may lead to a high sensitivity of object detection precision to factors such as illumination, scale, and occlusion. Furthermore, traditional methods often encounter challenges when addressing object detection in complex scenarios. The confusion between object edges and the background leads to an increased likelihood of both false positives and negatives in detection results.

The emergence of deep learning has significantly facilitated advances in object detection. Due to the powerful feature learning and representation capabilities, the utilization of CNNs has demonstrated superior performance and efficiency. It should be noted that UOD is regarded as an extension of general object detection, which can be broadly categorized into the two-stage and one-stage frameworks.

The procedure of the two-stage framework is somewhat similar to that of traditional schemes. It first generates candidate region proposals, and then classifies them into different object categories. Representative models based on two-stage framework include R-CNN series [**? ? ?**], spatial pyramid pooling network (SPP-Net) [**?**], R-FCN [**?**], feature pyramid network (FPN) [**?**], etc. Inspired by the two-stage framework, several well-received approaches have been proposed to address the specific challenges of UOD tasks. Zeng et al. [**?**] developed a novel framework for robust detection of underwater seafood by aggregating the adversarial occlusion network (AON) into the standard Faster R-CNN. Superior performance can be achieved through mutual competition and learning between the two networks. Xu et al. [**?**] designed a refined marine object detector, which improves the SPP-Net with an appropriate attention mechanism. The features from different depths are fused utilizing an innovative bidirectional feature fusion strategy, thereby alleviating the weakening of features and further enhancing the overall detection precision. Liu et al. [**?**] presented a novel two-stage UOD network, which adopts the popular Swin Transformer as the backbone and eliminates the quantization errors of region of interest (ROI), so that the performance can be effectively enhanced. Song et al. [**?**] designed a new region proposal network termed RetinaRPN. It fully considers the intersection over union (IoU) prediction for uncertainty, and the object prior probability can be accordingly modeled.

One-stage framework eschews the use of region proposals and directly extracts hierarchical features to predict the detection results, significantly highlighting the powerful

real-time processing capabilities. At present, five representative benchmark methods based on the one-stage framework have been widely used in various UOD tasks, namely Single Shot MultiBox Detector (SSD) [**?**], CenterNet [**?**], RetinaNet [**?**], You Only Look Once (YOLO) series [**? ? ? ?**], and DEtection TRansformer (DETR) series [**? ?**]. Li et al. [**?**] developed a YOLOv3-based network for zooplankton detection, which adopts densely connected structures to facilitate feature transmission. Hu et al. [**?**] focused on the real-time detection of uneaten feed pellets in underwater images. It combines the popular DenseNet with the YOLOv4 model to achieve simultaneous improvement of detection precision and efficiency. Wang et al. [**?**] proposed a lightweight underwater object detection network termed LUO-YOLOX using weighted ghost-CSPDarknet and simplified PANet. Additionally, the authors also presented an efficient self-supervised pre-training joint framework based on underwater auto-encoder transformation (UAET), which can address the problems of color distortion and unclear targets in underwater images. Zhang et al. [**?**] presented the Transformer-based bi-directional feature pyramid network (BiFPN) for object detection. It combines the multihead self-attention into the original CSPDarkNet to achieve effective cross-scale feature fusion. Zhang et al. [**?**] proposed an attractive one-stage network for UOD, which combines the MobileNetv2 and depth-wise separable convolution to effectively reduce the computational load. Zhao et al. [**?**] proposed an improved YOLOv4 network for more precise UOD, which consists of a symmetrical bottleneck-type structure, an enhanced FPN-Attention module, and a label smoothing training strategy. Sun et al. [**?**] presented a novel UOD network combining the MobileViT and YOLOX, further facilitating the feature extraction ability of the network.

1.2. Underwater Image Enhancement and Super-Resolution

Generally speaking, underwater images inevitably suffer from severe quality degradation due to the following three common reasons [**? ? ?**]: (1) Light attenuation and scattering often lead to color distortion, low contrast, and opacity in underwater images. (2) Diverse active floating particles and suspended bodies in marine environments further aggravate image degradation, resulting in haze-like and blurred effects. (3) Low-end optical cameras are unable to fully consider the problems inherent in diverse marine environments, which often result in random noise and regional ambiguity in the generated underwater images. In this context, the above deficiencies seriously impact the effectiveness and precision of the UOD tasks. Figure **??** shows some visualized object detection samples in underwater images of varying visual quality based on the popular YOLOv5 model. As observed, high-quality underwater images are more conducive to stimulating the potential of such context-aware object detection approaches, which indicates that the necessary pre-processing of underwater images can boost the performance of UOD tasks. To this end, we will focus on two common image pre-processing schemes in this paper, namely underwater image enhancement (UIE) and super-resolution (SR), to provide more effective visual cues for UOD tasks.

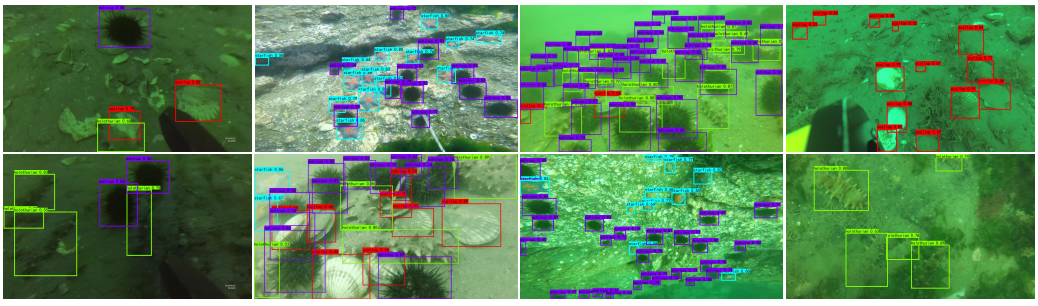


Figure 1. Object detection samples in underwater images of varying visual quality based on the popular YOLOv5 model.

UIE aims to correct the distorted colors and enhance the reduced contrast of raw underwater images, thereby alleviating the influence of light attenuation and scattering [**?**

?]. Existing UIE techniques can be broadly categorized into physical model-based [**? ?**], nonphysical model-based [**? ?**], and learning-based [**? ?**] methods. Physical-model-based methods estimate the background light and transmission map according to the underwater optical imaging model [**?**], but the precise enhancement of underwater images relies significantly on prior knowledge. Nonphysical-model-based methods directly modify image pixels to generate visually pleasing results without considering the degradation mechanism. Learning-based methods achieve end-to-end modeling of complex nonlinear systems by learning the mapping between paired raw and enhanced underwater images. On the other hand, SR is considered to address the inherent ill-posed problem, which aims to reconstruct a high-resolution image from its low-resolution observation, so that the image clarity can be improved [**? ?**]. Existing SR techniques can be broadly categorized into interpolation-based [**?**], reconstruction-based [**? ?**], and learning-based [**? ? ?**] methods. Interpolation-based methods treat each image pixel as an individual grid, and SR is performed by estimating pixel values between adjacent grids. Reconstruction-based methods involve deducing the inverse degradation process by utilizing prior information from the raw low-resolution image with necessary constraints. Learning-based methods infer the degradation by learning the mapping between paired low- and high-resolution images, enabling the exploration of the optimal SR process. To summarize, Table **??** presents an overview of the categories, definitions, advantages, and disadvantages of the UIE and SR techniques, respectively.

Table 1. Descriptions of the UIE and SR techniques.

Techniques	Categories	Definitions	Advantages	Disadvantages
UIE	Physical Model-Based [**? ?**]	Estimating background light and transmission map based on underwater optical imaging formulation	Sensitive to transparency, scattering effects, and absorption coefficients of water	Difficulty in parameter adjustment and optimization with limited generalization
	Nonphysical Model-Based [**? ?**]	Directly modifying image pixels	Simple and easy-to-implement	Lacking theoretical basis and guidance from physical models
	Learning-Based [**? ?**]	Learning the mapping functions between paired raw and enhanced images	Strong performance in constructing complex mapping with excellent generalization	Requiring sufficient training data
SR	Interpolation-Based [**?**]	Interpolating pixel values between known adjacent pixel grids	Simple and cost-effective	Limited improvement in high-frequency image details
	Reconstruction-Based [**? ?**]	Inferring inverse degradation process based on prior information	Promising preservation of enhanced image structures and details	Sensitive to noise and artifacts, and requiring additional prior information
	Learning-Based [**? ? ?**]	Learning the mapping functions between paired low- and high-resolution images	Superior performance in capturing and generating image details and textures	Requires sufficient training data

At present, continuous efforts have demonstrated that the learning-based methods can deliver state-of-the-art performance in both UIE and SR techniques. Furthermore, the learning-based methods can also avoid the construction and estimation of complex degradation models, which significantly increases the generalization ability to cope with diverse and complex underwater scenarios. Therefore, we adopt the learning-based mathematical model for both UIE and SR in this paper to construct complex mapping between degraded and high-quality image pairs. To summarize, our objective is to leverage the aforementioned salient properties and present the collaborative framework by jointly integrating the functions of UIE and SR, so that more detection-favoring visual cues can be revealed to facilitate the subsequent UOD tasks.

1.3. Main Novelties and Contributions

In this paper, a novel collaborative framework for UOD is proposed. The main novelties and contributions of our work can be summarized as follows:

- As opposed to the existing schemes, we present a collaborative framework via joint image enhancement and super-resolution. By employing a joint-oriented network training strategy, the proposed framework is more effective at generating a detection-favoring appearance to stimulate efficient and precise object detection of underwater images.
- A plug-and-play self-attention mechanism called multihead blurpooling fusion network (MBFNet) is developed, which enables our proposed framework to capture sufficient contextual information concerning the dependencies between feature maps from a broader and more focused viewpoint, thereby further enhancing UOD performance.
- A heuristic step-by-step training strategy is designed for our proposed collaborative framework. Compared with the conventional end-to-end training strategy, our designed step-by-step training strategy can effectively alleviate the potential gradient vanishing or exploding by dividing the whole training process into refined stages, so that the framework architecture can be better controlled.

2. Proposed Method

We first provide an overview of our proposed collaborative framework in Section **??**. The three primary components of our proposed framework, namely MBFNet, pre-processing module (PPM), and underwater object detection module (UODM), are then illustrated in Section **??**. Subsequently, Section **??** describes the loss function of our framework. Finally, the step-by-step training strategy of our framework is overviewed in Section **??**.

2.1. Overview of the Proposed Collaborative Framework

The architecture of our proposed collaborative framework is shown in Figure **??**, which mainly consists of three primary components including MBFNet, PPM, and UODM. The MBFNet is essentially considered a plug-and-play self-attention mechanism, which not only generates enhanced feature maps with coherent structures and rich details, but also enables our proposed collaborative network to better understand images from both global and local perspectives to capture more representative features. The PPM aims to pre-process the raw underwater images in terms of color correction and resolution improvement, thereby enhancing the availability of visual cues for better detection of diverse underwater objects. The UODM is defined as the central component for detecting fuzzy and difficult-to-find underwater objects. Note that the popular PANet structure [**?**] is utilized in the UODM for precise prediction of small-scale underwater objects, and the step-by-step training strategy is also operated to promote the high-quality integration and interaction between the PPM and UODM.

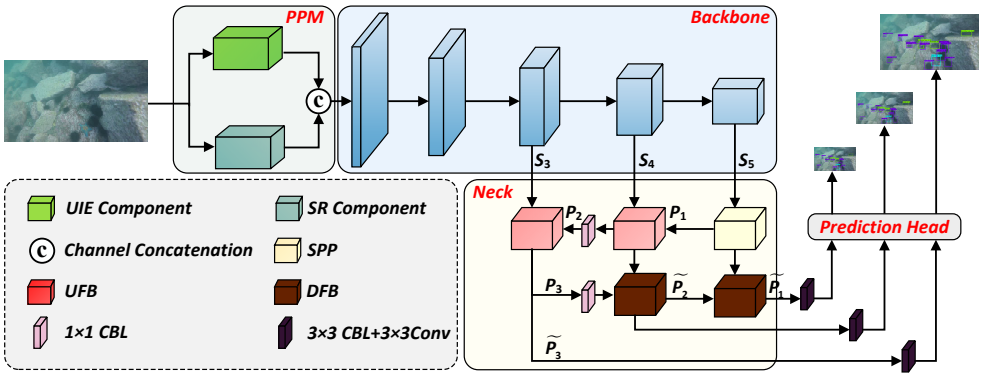


Figure 2. Architecture of our proposed collaborative framework.

2.2. Composition of the Proposed Collaborative Framework

2.2.1. Multihead Blurpooling Fusion Network

Inspired by the well-received multihead attention mechanism from Transformer-based models [**?**], we develop the effective MBFNet to facilitate the interaction of multiscale information, so that more reasonable weights can be generated. Our proposed MBFNet contains two initial versions, termed MBFNet-I and MBFNet-II, as shown in Figure **??**.

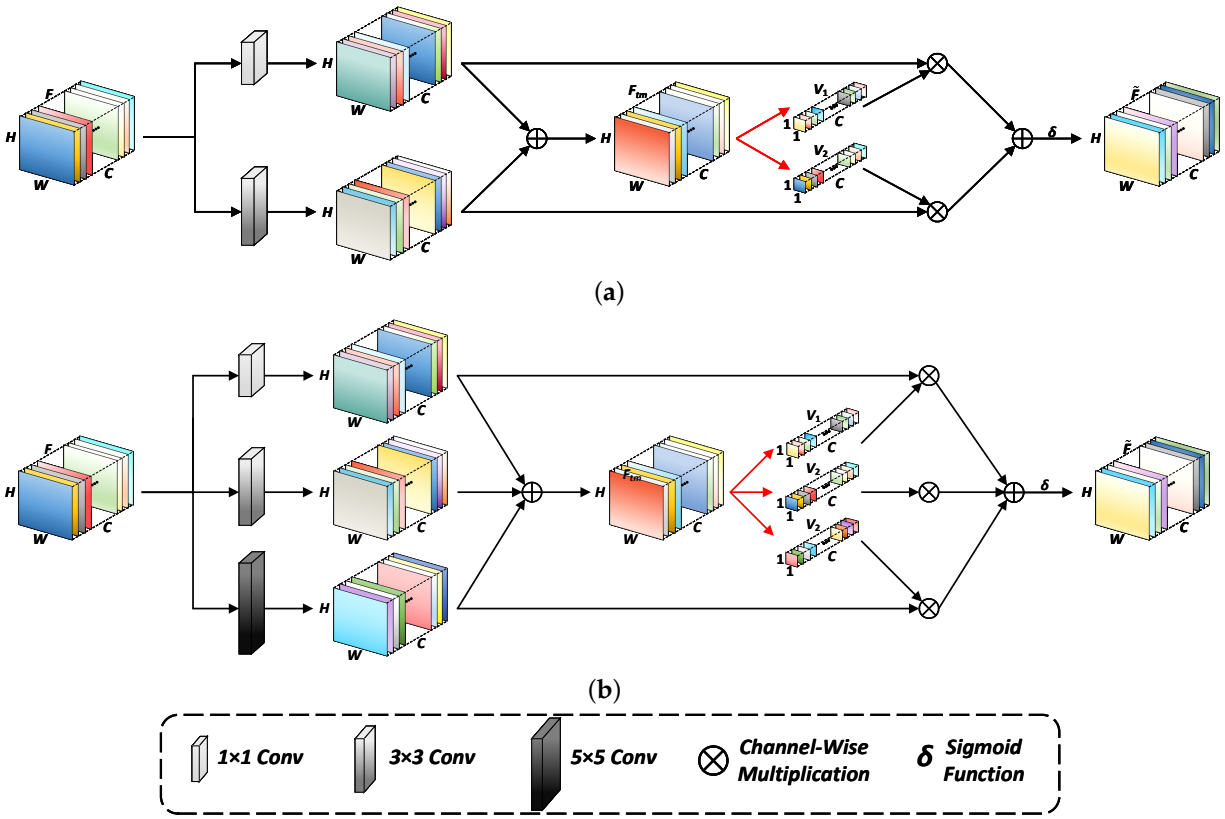


Figure 3. Architecture of the MBFNet. (**a**) MBFNet-I. (**b**) MBFNet-II.

For the implementation process of the MBFNet-I, the input feature map $F \in \mathbb{R}^{H\times W\times C}$ is firstly fed into the dual branches processed by a 1×1 convolutional layer and a 3×3 convolutional layer, respectively. Then, the transmissive feature map $F_{tm} \in \mathbb{R}^{H\times W\times C}$ can be generated by the element-wise addition operation, which can be expressed by

$$F_{tm} = Conv_{1\times 1}(F) + Conv_{3\times 3}(F) \tag{1}$$

Subsequently, the generated F_{tm} is further squeezed into two weight vectors by the blurpooling and full connection operations, which can be formulated by

$$\begin{cases} V_1 = FC(BP_{1\times 1}(F_{tm})) \\ V_2 = FC(BP_{3\times 3}(F_{tm})) \end{cases} \tag{2}$$

where $V_1 \in \mathbb{R}^{1\times 1\times C}$ and $V_2 \in \mathbb{R}^{1\times 1\times C}$ symbol the two generated weight vectors. $BP_{n\times n}$ denotes the blurpooling operation with the filter size of $n\times n$. FC represents the full connection. Finally, the output feature map $\tilde{F}$ can be computed as follows:

$$\tilde{F} = \delta(Conv_{1\times 1}(F)\circ V_1 + Conv_{3\times 3}(F)\circ V_2) \tag{3}$$

where $\circ$ denotes the channel-wise multiplication, and δ symbols the sigmoid activation function. The MBFNet-II is highly similar to the MBFNet-I in implementation process,

except that the former adopts a triple-branch network structure guided by the 1×1, 3×3, and 5×5 convolutional layers, respectively.

In summary, our proposed MBFNet consists of multiple parallel branches, which process the same input feature maps to compute weighted aggregation results. This mechanism effectively enhances the self-representation capability of the model, so that the features at different scales can be significantly prioritized. Furthermore, the MBFNet adopts the blurpooling operation instead of traditional maxpooling, which can effectively address the issue of shift-equal variance with minimal computational load, so that the features can be better represented. As a consequence, our proposed MBFNet can capture and associate contextual information from a large neighborhood, which is useful for learning spatial information between difficult-to-find objects and backgrounds. This self-attention mechanism can not only improve semantic discrimination but also significantly reduces confusion between object categories, so that the challenges of the UOD tasks in complex marine environments can be well addressed.

2.2.2. Pre-Processing Module

The proposed PPM aims to improve the visual quality of underwater images for better detection and prediction of diverse underwater objects, which is composed of the UIE component and the SR component processed in parallel.

The UIE component is developed to alleviate the visual deficiency in marine environments including color distortion and contrast reduction. As shown in the top left portion of Figure **??**, the UIE component adopts a symmetrical and lightweight CNN-based structure, which can be further abstracted into three procedures: (a) The feature extraction is conducted through a 3×3 CBL layer (In this paper, the $n \times n$ CBL layer is defined as the combination of a $n \times n$ convolutional layer, Batch Normalization, and Leaky ReLU activation function.) and a residual block. Note that the residual block consists of two 3×3 CBL layers and a shortcut connection with a 1×1 convolutional layer, as shown in the top right portion of Figure **??**. (b) The feature optimization is operated through the MBFNet-II to learn more representative and distinctive features. (c) The information recovery is performed through a residual block and a 3×3 CBL layer, which is exactly opposite to the feature extraction stage. Mathematically, the implementation process of the UIE component can be briefly formulated as follows:

$$Y^{UIE} = \mathscr{F}(X, \theta^{UIE}) \tag{4}$$

where $X \in \mathbb{R}^{H \times W \times C}$ and $Y^{UIE} \in \mathbb{R}^{H \times W \times C}$ denote the raw underwater image and its corresponding predicted UIE image, respectively. $\mathscr{F}(\cdot, \theta^{UIE})$ indicates the overall function of the UIE component parameterized by θ^{UIE}.

The SR component is designed to restore the clear underwater image with abundant high-frequency details based on low-resolution observation. As shown in the bottom portion of Figure **??**, the architecture of the SR component is similar to that of the UIE component, but the difference lies in the following three aspects: (a) For the feature extraction stage, the SR component adopts two residual blocks (For the first residual block here, all the CBL layers maintain a consistent number of output channels, which is slightly different from the parameters shown in the top right portion of Figure **??**). (b) The bicubic upsampling operation is conducted followed by the embedded MBFNet-I. Note that we set the upsampling factor to 2 as an example, but it is self-evident that the upsampling factor can be modified accordingly based on the actual situation. (c) For the information recovery stage, the SR component adopts two 3×3 CBL layers, and a shortcut connection with a 1×1 convolutional layer is operated to add the input feature maps to the output. Mathematically, the implementation process of the SR component can be briefly formulated as follows:

$$Y^{SR} = \mathscr{G}(X, \varphi^{SR}) \tag{5}$$

where $Y^{SR} \in \mathbb{R}^{H \times W \times C}$ denotes the predicted SR image. $\mathcal{G}(\cdot, \varphi^{SR})$ indicates the overall function of the SR component parameterized by θ^{SR}.

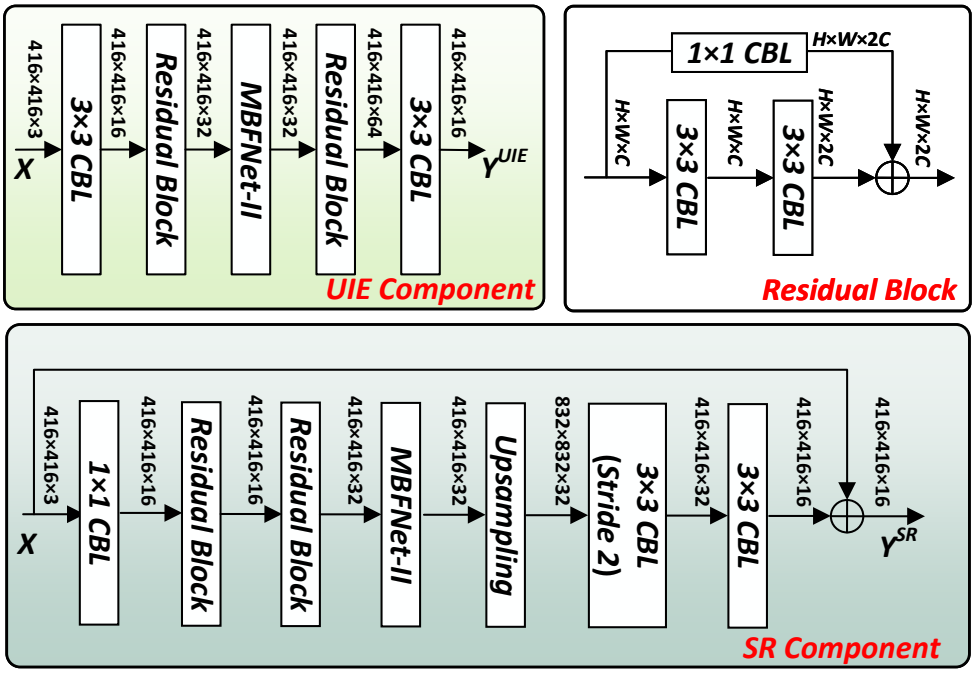


Figure 4. Architecture of the UIE component and the SR component.

Finally, the quality-enhanced 32-channel feature map Y^{QE} output from the PPM can be obtained by performing the channel concatenation operation between Y^{UIE} and Y^{SR}, which is expressed as follows:

$$Y^{QE} = [Y^{UIE}, Y^{SR}] \tag{6}$$

2.2.3. Underwater Object Detection Module

The proposed UODM aims to achieve precise detection and prediction of diverse underwater objects with different scales. In line with the typical object detection framework, our proposed UODM also consists of three components including the backbone, the neck, and the prediction head [**?**].

The backbone is utilized to extract sufficient features from the quality-enhanced image Y^{QE}. To fully stimulate the network performance, we mainly employ the well-received ResNet-50 [**?**] as the backbone, which has been pre-trained on the VOC2007 dataset for network initialization. Note that our proposed UODM can also be implemented using any other widely-used benchmark model as the backbone.

The neck in our proposed UODM aims to enhance the feature aggregation by integrating both low- and high-level information, so that a multiscale feature pyramid map can be generated to capture underwater objects at different scales. Here, we develop a lightweight strategy based on the popular PANet structure [**?**], which comprises two efficient blocks termed upsampling fusion block (UFB) and downsampling fusion block (DFB) to reduce the redundancy of gradient information by employing the cross-stage operation. As shown in Figure **??**, the UFB is composed of an upsampling layer, the MBFNet-I, and a 1×1 CBL layer. The DFB adopts a highly similar structure to the UFB, but it replaces the upsampling layer in the UFB with a 3×3 CBL layer with stride 2. These two blocks are used to facilitate the fusion and transmission of multiscale contextual information, which benefits further improvements in the precision and efficiency of UOD. Note that the popular SPP-Net [**?**] is also employed at the beginning of the neck to enhance the robustness and precision of our model. This implementation is considered to alleviate the issue posed by the excessive

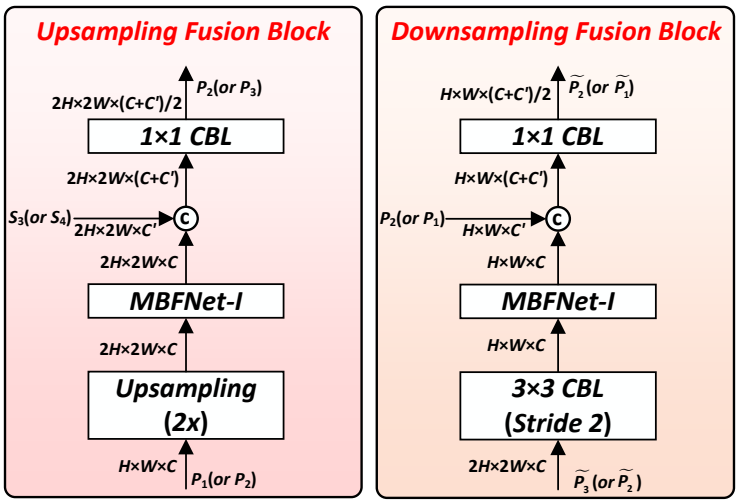
number of channels in the final stage of the backbone, which may potentially prolong inference times.

Following the general object detection framework, the prediction head of our proposed UODM leverages the multiscale features output from the neck to achieve the prediction at different scales, which enables the UODM to precisely locate and detect objects with distinctive sizes in underwater images. Finally, our proposed UODM can generate the outputs at three scales, and the corresponding bounding boxes and classification results are determined using non-maximum suppression.

Mathematically, the implementation process of the UODM can be briefly formulated as follows:

$$Y^{UOD} = \mathscr{H}(Y^{QE}, \rho^{UOD}) \tag{7}$$

where Y^{UOD} denotes the predicted UOD set consisting of the images with three scales. $\mathscr{H}(\cdot, \rho^{UOD})$ indicates the overall function of the UODM parameterized by ρ^{UOD}.



(**a**) (**b**)

Figure 5. Architecture of the upsampling fusion block and the downsampling fusion block. (**a**) Upsampling fusion block. (**b**) Downsampling fusion block.

2.3. Loss Function

In this paper, we formulate a composite loss function to align with the optimization objective of our model, which is composed of the UIE loss, the mean square error (MSE) loss, and the UOD loss.

2.3.1. UIE Loss

Since no ground truths are available for the UIE tasks in our collaborative framework, we employ the self-correlated total variation (TV) loss $\mathscr{L}_{TV}$ [**?**] to promote spatial smoothness while preserving the image style. Mathematically, the TV loss can be expressed by

$$\mathscr{L}_{TV} = \sum_{i}^{M} \sum_{j}^{N} \sqrt{(Y_{i+1,j}^{UIE} - Y_{i,j}^{UIE})^2 + (Y_{i,j+1}^{UIE} - Y_{i,j}^{UIE})^2} \tag{8}$$

where $Y_{i,j}^{UIE}$ symbols the pixel value at the location (i, j) in the predicted UIE image.

Moreover, the feature reconstruction loss $\mathscr{L}_{feat}$ [**?**] is adopted to further constrain the bias between the raw underwater image X and the corresponding predicted UIE image Y^{UIE}, which can be expressed by

$$\mathscr{L}_{feat} = \frac{1}{H \times W \times C} \| \zeta(X) - \zeta(Y^{UIE}) \|^2 \tag{9}$$

where $\zeta(\cdot)$ [**?**] represents the function used to extract the feature representation from the candidate image, which is pre-trained through the VGG-16 [**?**] network.

Finally, the overall UIE loss $\mathscr{L}_{UIE}$ can be computed by the linear combination of the above two loss functions, which is formulated as follows:

$$\mathscr{L}_{UIE} = \omega_{TV}\mathscr{L}_{TV} + \omega_{feat}\mathscr{L}_{feat} \tag{10}$$

where ω_{TV} and ω_{feat} are two weight coefficients for $\mathscr{L}_{TV}$ and $\mathscr{L}_{feat}$, respectively.

2.3.2. MSE Loss

The MSE loss $\mathscr{L}_{MSE}$ is considered a commonly used loss function for the SR task, as it can effectively preserve the sharpness of edges and refined details. Mathematically, the MSE loss can be expressed as follows:

$$\mathscr{L}_{MSE} = \frac{1}{N_s}\sum_{i=1}^{N}\|Y_i - Y_i^{SR}\|^2 \tag{11}$$

where N_s is the number of training samples in the SR component. Y_i and Y_i^{SR} symbol the predicted SR underwater image and its corresponding ground truth, respectively.

2.3.3. UOD Loss

To achieve precise detection of fuzzy and difficult-to-find underwater objects, we design the UOD loss inspired by [**?**], which is the combination of classification loss $\mathscr{L}_{cls}$, localization loss $\mathscr{L}_{loc}$, and IoU loss $\mathscr{L}_{iou}$. These three components of the loss functions can be formulated as follows:

$$\begin{cases} \mathscr{L}_{cls} = \frac{1}{N_u}\sum_{i}^{N} FL(p_i, \hat{p}_i) \\ \mathscr{L}_{loc} = \frac{1}{N_{Pos}}\sum_{i\in Pos}^{N} 1 - CIoU_i \\ \mathscr{L}_{iou} = \frac{1}{N_{Pos}}\sum_{i\in Pos}^{N} BCE\left(CIoU_i, \hat{CIoU_i}\right) \end{cases} \tag{12}$$

where N_u denotes the total grid number in the image, FL symbols the focal loss [**?**], N_{pos} denotes the prediction box, $CIoU_i$ and $\hat{CIoU_i}$ represents whether the N_{pos} box in the N_u grids contains objects or not. BCE denotes the cross entropy loss function [**?**]. Then, the UOD loss $\mathscr{L}_{UOD}$ is defined as the linear combination of the above three components, which can be expressed as follows:

$$\mathscr{L}_{UOD} = \mathscr{L}_{cls} + \mathscr{L}_{loc} + \mathscr{L}_{iou} \tag{13}$$

Considering that we have adopted a joint learning strategy to train both the PPM and UODM, the final loss function $\mathscr{L}$ of our proposed collaborative network can be formulated as follows:

$$\mathscr{L} = \lambda_{UIE}\mathscr{L}_{UIE} + \lambda_{SR}\mathscr{L}_{SR} + \lambda_{UOD}\mathscr{L}_{UOD} \tag{14}$$

where λ_{UIE}, λ_{SR}, and λ_{UOD} are three weighting coefficients. In this paper, we empirically set these coefficients as 1, 1, and 10, respectively.

2.4. Step-by-Step Training Strategy

Since our proposed collaborative framework consists of different modules with specific functions, we adopt a step-by-step training strategy for the model in this paper. Compared with the conventional end-to-end strategy, the step-by-step training strategy allows for each distinctive module to be trained individually, so that the overall framework architecture can be better controlled. In addition, the step-by-step training strategy can effectively alleviate the potential gradient vanishing or exploding by dividing the whole training process into refined stages, facilitating more reasonable controlled optimization and possible faster

convergence. The pseudo-code of the training process for our proposed collaborative framework is illustrated in Algorithm **??**. Note that the validation and testing procedures in our proposed collaborative framework follow the standard rules of general CNN-based models. The performance on the validation set is evaluated after each training epoch to determine the hyperparameter tuning or early stopping, and the performance on the testing set provides an unbiased estimate of the model's ability to detect diverse underwater objects.

Algorithm 1 Step-by-Step Training Strategy

Input: The overall collaborative framework Φ with parameter group:
A set of outputs from different modules $Y = \{Y^{UIE}, Y^{SR}, Y^{UOD}\}$;
Training set: τ;
Mini-batch $(x_s, y_s) \in \tau$;
Convergence threshold: γ;
Loss function: $\mathscr{L} = \{\mathscr{L}_{UIE}, \mathscr{L}_{SR}, \mathscr{L}_{UOD}\}$;
Actual value of the loss function: l
Output: Untrained network: $\Phi(X; Y^{UOD})$;
// Step 1: Train the PPM individually.
procedure TRAIN(Φ, τ)
  repeat
    $l \leftarrow \mathscr{L}_{UIE}(\mathscr{F}(x_s, \theta^{UIE}), y_s); \mathscr{L}_{SR}(\mathscr{G}(x_s, \varphi^{SR}), y_s)$
  until $l < \gamma$
end procedure
$Y \leftarrow Y \setminus \{Y^{UIE}, Y^{SR}\}$ // Freeze the PPM.
// Step 2: Train the UODM individually.
procedure TRAIN(Φ, τ)
  repeat
    $l \leftarrow \mathscr{L}_{UOD}(\mathscr{H}(x_s, \rho^{UOD}), y_s)$
  until $l < \gamma$
end procedure
$Y \leftarrow Y \cup \{Y^{UIE}, Y^{SR}\}$ // Activate the PPM.
// Step 3: Train the PPM and UODM jointly.
procedure TRAIN(Φ, τ)
  repeat
    $l \leftarrow \mathscr{L}_{UIE}(\mathscr{F}(x_s, \theta^{UIE}), y_s); \mathscr{L}_{SR}(\mathscr{G}(x_s, \varphi^{SR}), y_s); \mathscr{L}_{UOD}(\mathscr{H}(x_s, \rho^{UOD}), y_s)$
  until $l < \gamma$
end procedure
return Trained network: $\Phi\left(X; Y^{UOD}\right)$ // Regress bounding box, the object location, and classification result.

3. Experiments

3.1. Data Processing

To demonstrate the effectiveness of our proposed collaborative network, all experiments have been performed on two popular datasets, namely URPC2020 (Available online: https://www.urpc.org.cn/index.html (accessed on 17 October 2022)) and Brackish (Available online: https://www.kaggle.com/aalborguniversity/brackish-dataset (accessed on 18 May 2023). URPC2020 is a well-received dataset specifically used for the UOD task. It contains a total of 7543 images with four categories including echinus, holothurian, scallop, and starfish. The URPC2020 dataset can be considered a long-tailed dataset due to significant differences in the number of samples across different categories, which poses a great challenge in achieving accurate UOD performance. Brackish is an open-source underwater dataset, which is annotated based on real filmed underwater videos. It contains a total of 14,518 video occurrences with six categories including large fish, crabs, jellyfish, shrimps, small fish, and starfish. For each dataset, we randomly allocate 80% of the data for training, 10% for validation, and 10% for testing. Note that we have refrained from utilizing data augmentation means to thoroughly demonstrate the superior performance of our proposed collaborative framework. Instead, we have implemented label smoothing to reduce the discrepancies between different bounding boxes. The detailed descriptions of the two datasets are shown in Table **??**.

Table 2. Descriptions of the URPC2020 and Brackish datasets.

Dataset	Species Category	Annotations	Data Type
URPC2020	Holothurian	5537	Jpg images
	Echinus	22,343	
	Star fish	6841	
	Scallop	6720	
Brackish	Big fish	3241	Video occurrences
	Crab	6538	
	Jelly fish	637	
	Shrimp	548	
	Small fish	9556	
	Star fish	5093	

3.2. Experimental Setup

The proposed collaborative framework is implemented using the PyTorch platform with the version 1.10.0. The parameters are tuned using the Adam optimizer. All experiments are performed on a server with a 2.10 GHz Inter(R) Xeon(R) Gold 6130 CPU, an NVIDIA RTX 3090Ti GPU, 24 GB RAM, and Ubuntu 21.04 operating system. During the training process, the batch size is set to 16 with a total of 200 training epochs. We employ the cosine learning rate decay strategy, where the initial learning rate is 0.01, and a momentum of 0.937 is also applied to accelerate convergence. Additionally, the weight decay coefficient is set to 0.0005 to prevent data overfitting.

For quantitative evaluation, we adopt the commonly used average precision (AP) metrics at different levels of IoU thresholds, including $AP_{0.5:0.95}$ (the AP at IoU thresholds ranging from 0.5 to 0.95 with increments of 0.05), $AP_{0.5}$ (the AP at the fixed IoU threshold of 0.5), and $AP_{0.75}$ (the AP at the fixed IoU threshold of 0.5). In addition, we evaluate the computational complexity of the model by employing the metrics including the size of parameters, float point operations (FLOPs), inference time, and frames per second (FPS).

3.3. Comparative Study

We compared our proposed collaborative framework with several well-received object detection methods including Faster R-CNN [**?**], SSD [**?**],CenterNet [**?**], RetinaNet [**?**], YOLOv3 [**?**], YOLOv4 [**?**], YOLOv5m, YOLOv5l, YOLOx [**?**], YOLOv7 [**?**], DETR [**?**], YOLOv4-AFFM [**?**], and YOlOx-DCA [**?**].

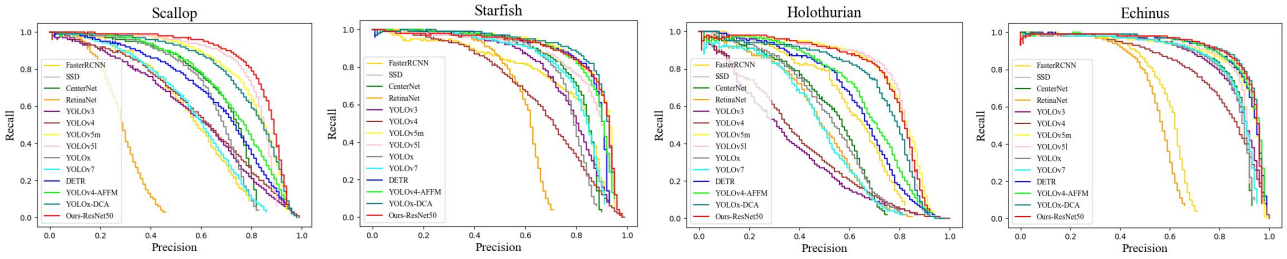
Table **??** reports the comparative study results for different methods in detection precision on the URPC2020 dataset, where the numbers in red, blue, and green, respectively, indicate the best, the second-best, and the third-best results. As observed, our proposed collaborative framework consistently yields the most promising overall AP values, and the AP values in all categories are competitively strong among the candidate methods. In addition, we also provide further experiments that apply our proposed PPM to several well-received general object detection networks including YOLOv5m, YOLOv5l, and DETR. The experimental results reveal that the performance of these networks is still inferior to our proposed collaborative framework, even if they have been input with quality-enhanced underwater images by the PPM. Figure **??** shows the Precision–Recall (PR) Curves of different methods for object categories on the URPC2020 dataset. It can be seen that our proposed collaborative network generally exhibits the most promising performance across different categories in terms of precision and recall.

Table **??** reports the comparative study results for different methods in detection precision on the Brackish dataset, which includes more challenging and complex underwater scenarios. As observed, all candidate methods exhibit a significant decrease in detection performance on the Brackish dataset compared to URPC2020. However, our proposed collaborative framework still outperforms other comparison methods with the best AP performance. In addition, it can also be observed that when our proposed PPM is applied to the competitive YOLOv5m, YOLOv5l, and DETR, their performance is still far inferior to our proposed collaborative framework, which is consistent with the conclusions obtained

from Table **??**. Figure **??** shows the PR Curves of different methods for object categories on the Brackish dataset, and the results can also demonstrate the outstanding advantages of our collaborative framework in detecting fuzzy and difficult-to-find underwater objects.

Table 3. Results of comparative study for different methods in detection precision on the URPC2020 dataset.

Method	Backbone	AP(%)				$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$
		Scallop	Starfish	Holothurian	Echinus			
Faster R-CNN [**?**]	VGG-16	31.25	53.90	48.33	60.93	24.34	48.73	20.01
SSD [**?**]	VGG-16	55.19	75.75	47.34	79.73	27.76	64.50	17.78
CenterNet [**?**]	ResNet-50	68.57	79.96	51.81	85.22	31.82	71.39	22.23
RetinaNet [**?**]	ResNet-50	28.76	59.20	47.75	54.80	21.45	47.63	15.68
YOLOv3 [**?**]	DarkNet-53	67.31	74.87	55.21	78.28	31.20	68.92	29.01
YOLOv4 [**?**]	CSPDarkNet-53	61.49	69.89	59.10	79.53	31.28	67.54	28.75
YOLOv5m	CSPDarkNet-53-M	70.82	77.87	77.79	86.17	43.21	78.16	38.53
YOLOv5l	CSPDarkNet-53-L	76.17	79.17	73.60	88.30	44.28	79.13	42.13
YOLOx [**?**]	ResNet-50	65.28	76.68	51.58	84.40	31.92	69.49	24.56
YOLOv7 [**?**]	ELAN-Net-L	57.78	79.49	46.95	85.58	29.81	67.45	20.13
DETR [**?**]	ResNet-50	69.93	84.28	62.07	87.22	42.16	75.87	40.07
YOLOv4-AFFM[**?**]	MobileNetv2	73.06	86.00	66.85	90.14	36.61	79.01	29.78
YOLOx-DCA [**?**]	MobileVIT	79.22	86.77	72.28	88.73	41.82	81.75	37.63
YOLOv5m-PPM	CSPDarkNet-53-M	66.62	80.08	51.98	84.53	32.30	70.08	24.06
YOLOv5l-PPM	CSPDarkNet-53-L	71.89	83.31	59.19	86.19	36.21	75.14	29.40
DETR-PPM	ResNet-50	78.51	86.55	71.39	87.91	37.91	81.09	27.03
Ours	ResNet-50	80.95	81.38	76.72	90.04	44.51	82.27	40.73



(**a**) (**b**) (**c**) (**d**)

Figure 6. PR Curves of different methods for object categories on the URPC2020 dataset (IoU = 0.5). (**a**) Scallop. (**b**) Starfish. (**c**) Holothurian. (**d**) Echinus.

Table 4. Results of comparative study for different methods in detection precision on the Brackish dataset.

Method	Backbone	AP(%)						$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$
		Crab	Fish	Jellyfish	Shrimp	Smallfish	Starfish			
Faster R-CNN [**?**]	VGG-16	54.81	64.75	7.25	21.48	17.50	85.74	12.34	41.92	11.01
SSD [**?**]	VGG-16	24.01	87.28	21.25	22.36	11.38	79.08	15.56	40.94	14.31
CenterNet [**?**]	ResNet-50	75.25	89.55	27.97	28.27	11.39	91.08	22.86	53.92	18.31
RetinaNet [**?**]	ResNet-50	66.71	83.75	18.43	20.30	7.30	86.72	18.21	47.20	15.34
YOLOv3 [**?**]	DarkNet-53	70.92	83.71	21.96	24.02	7.70	86.01	26.17	49.05	23.49
YOLOv4 [**?**]	CSPDarkNet-53	65.53	87.91	0.01	8.71	26.15	91.73	17.71	46.67	13.22
YOLOv5m	CSPDarkNet-53-M	87.08	77.47	45.00	62.32	42.73	94.99	39.12	68.27	38.03
YOLOv5l	CSPDarkNet-53-L	87.06	95.95	65.00	75.38	32.55	94.56	44.33	75.08	42.14
YOLOx [**?**]	ResNet-50	77.71	82.44	28.01	33.11	36.06	94.43	30.13	58.63	29.04
YOLOv7 [**?**]	ELAN-Net-L	78.82	93.30	35.29	44.63	18.27	93.21	31.25	60.59	27.65
DETR [**?**]	ResNet-50	88.04	94.50	67.39	74.00	27.17	94.04	43.18	74.19	41.02
YOLOv4-AFFM [**?**]	MobileNetv2	86.86	95.66	64.33	75.47	32.40	94.57	44.31	74.88	41.33
YOLOx-DCA [**?**]	MobileVIT	85.61	92.11	36.79	56.65	41.50	94.30	35.21	67.83	33.97
YOLOv5m-PPM	CSPDarkNet-53-M	75.56	85.02	25.49	28.16	9.02	87.90	21.03	51.86	20.65
YOLOv5l-PPM	CSPDarkNet-53-L	79.99	88.20	36.19	40.69	11.58	90.64	27.06	57.13	24.79
DETR-PPM	ResNet-50	75.11	92.59	46.00	54.97	20.68	91.31	33.19	63.44	30.87
Ours	ResNet-50	89.07	96.52	75.56	81.40	37.71	95.04	47.32	79.21	45.21

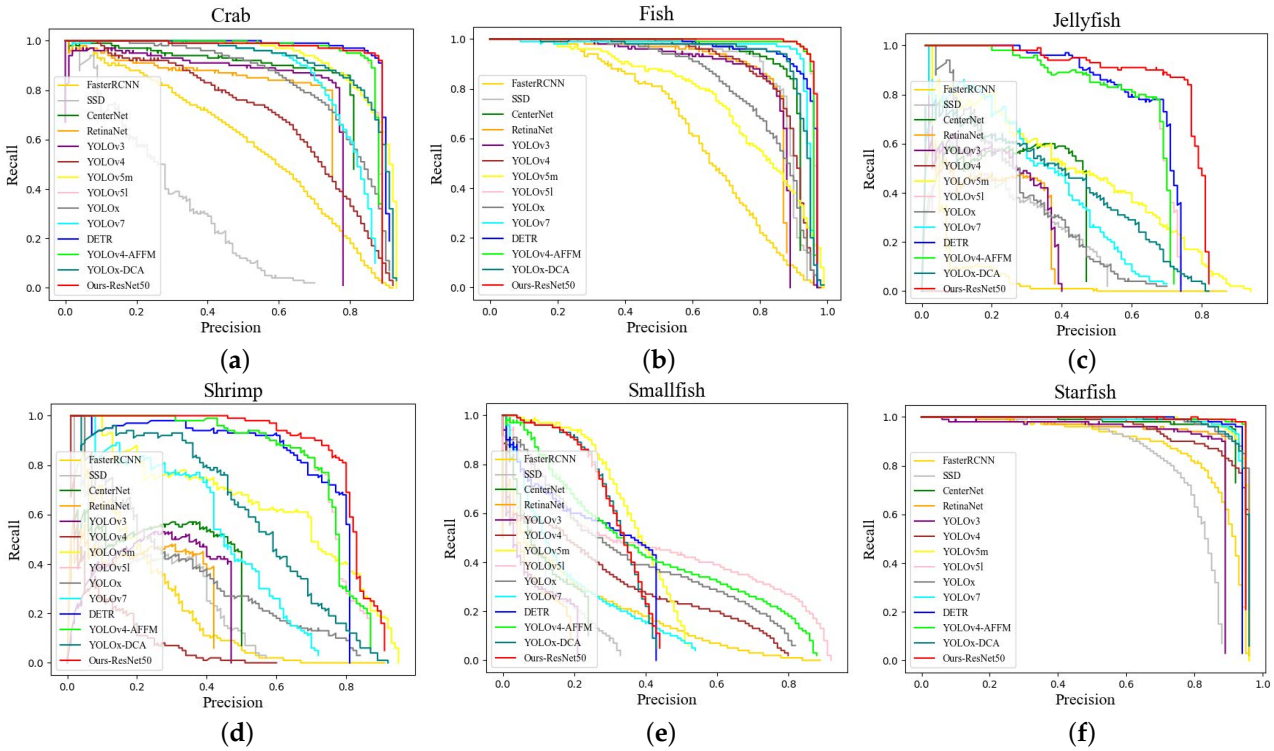
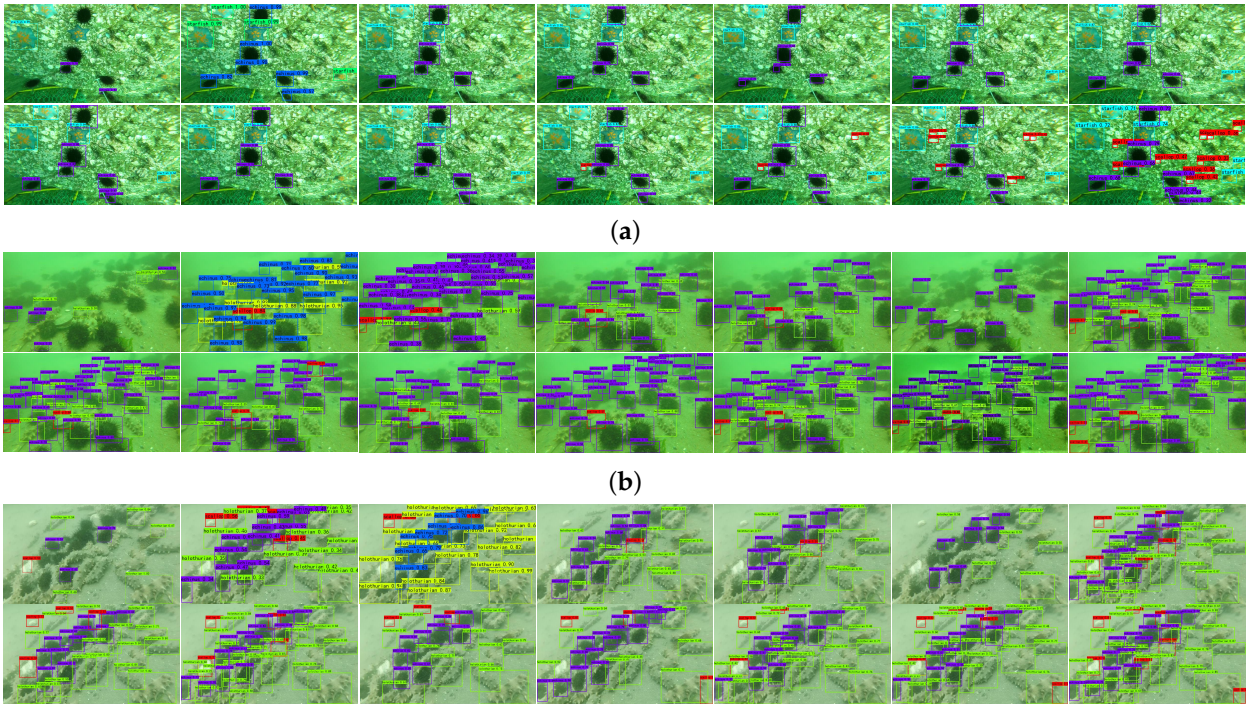


Figure 7. PR Curves of different methods for object categories on the Brackish dataset (IoU = 0.5). (**a**) Starfish. (**b**) Crab. (**c**) Fish. (**d**) Smallfish. (**e**) Shrimp. (**f**) Jellyfish.

To intuitively demonstrate the superior detection performance of our proposed collaborative framework compared to other comparison methods, some qualitative results of representative underwater images are provided, as shown in Figure **??**. As observed, underwater images typically suffer from color distortion, resolution reduction, and texture blurring, which pose significant challenges for accurate object detection. However, our proposed collaborative framework can always achieve robust detection for a more diverse range of underwater objects with the highest precision. Particularly, when there are occlusions between multiple underwater objects, our framework still produces promising and robust detection results. In addition, Figure **??** shows the visualization results of the individual UIE and SR components, including the intermediate outputs and the corresponding heatmaps generated by the Grad-CAM technique [**?**]. As observed, the PPM has been demonstrated to generate more visually pleasing intermediate outputs in the proposed collaborative network. Moreover, the results of the Grad-CAM heatmaps further reveal that our PPM enables the UODM to focus more on the underwater object-related regions, thereby facilitating a significant improvement in UOD performance.

Furthermore, the computational complexity of our proposed collaborative framework is fully concerned. As shown in Table **??**, although our framework is not inferior to some extremely lightweight object detection models in computational complexity, it still performs similarly to other popular methods including RetinaNet, YOLOv5m, YOLOx, YOLOv7, and DETR, which indicates that our framework is also capable of addressing real-time UOD problems.

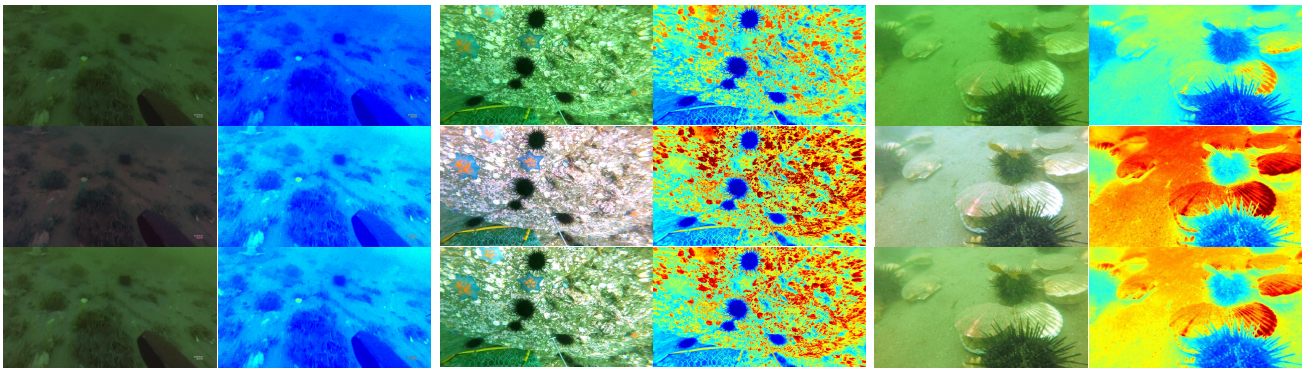


(**a**)

(**b**)

(**c**)

Figure 8. Qualitative detection results of representative underwater images. Note that for each representative underwater image, the detection results in the first row are provided by Faster R-CNN [**?**], SSD [**?**], CenterNet [**?**], RetinaNet [**?**] YOLOv3 [**?**], YOLOv4 [**?**], while the detection results in the second row are provided by YOLOv5m, YOLOv5l, YOLOx [**?**], YOLOv7 [**?**], DETR [**?**], YOLOv4-AFFM [**?**], YOlOx-DCA [**?**] and ours. (**a**) Representative underwater image 1. (**b**) Representative underwater image 2. (**c**) Representative underwater image 3.



(**a**) (**b**) (**c**)

Figure 9. Visualization results of the individual UIE and SR components. Note that for each representative image, the left column from top to bottom represents the raw underwater image, the intermediate output of UIE, and the intermediate output of SR, respectively. The right column shows the corresponding results of the Grad-CAM heatmaps. (**a**) Representative underwater image 1. (**b**) Representative underwater image 2. (**c**) Representative underwater image 3.

Table 5. The comparative study results for different methods in computational complexity on the URPC2020 dataset.

Method	Backbone	#param (M)	FLOPs (G)	Inference Time (ms)	FPS
Faster R-CNN [**?**]	VGG-16	28.30	909.57	58.12	17.20
SSD [**?**]	VGG-16	3.941	2.653	6.22	160.27
CenterNet [**?**]	ResNet-50	33.67	70.21	7.08	141.21
RetinaNet [**?**]	ResNet-50	36.39	69.71	20.81	48.03
YOLOv3 [**?**]	DarkNet-53	61.54	65.62	10.69	94.49
YOLOv4 [**?**]	CSPDarkNet-53	63.95	59.98	13.28	75.33
YOLOv5m	CSPDarkNet-53-M	21.01	21.39	18.95	52.75
YOLOv5l	CSPDarkNet-53-L	46.65	48.42	11.07	90.31
YOLOx [**?**]	ResNet-50	54.15	65.77	22.75	43.95
YOLOv7 [**?**]	ELAN-Net-L	37.62	44.98	20.21	49.49
DETR [**?**]	ResNet-50	36.74	31.924	21.64	46.28
YOLOv4-AFFM [**?**]	MobileNetv2	10.73	63.22	21.47	44.18
YOLOx-DCA [**?**]	MobileVIT	4.51	25.35	27.28	56.73
Ours	ResNet-50	26.63	48.94	24.97	40.05

3.4. Ablation Study

In this section, an ablation study is performed to further demonstrate the effectiveness of each component in our collaborative framework, which involves the following experiments:

(1) *-w/o UIEC:* Removing the UIE component so that only the SR component remains operational in the PPM.
(2) *-w/o SRC:* Removing the SR component so that only the UIE component remains operational in the PPM.
(3) *-w/o PPM:* Removing both UIE and SR components so that the PPM is completely disabled.
(4) *-rp EWA-PPM:* Replacing the original channel concatenation with the element-wise addition in the PPM.
(5) *-rp MBFNet-I-PPM:* Replacing the MBFNet-II with the MBFNet-I in the PPM.
(6) *-rp MBFNet-III-PPM:* Replacing the MBFNet-II with the MBFNet-III (MBFNet-III is defined as a four-branch structure, which further adds a similar branch guided by a 7×7 convolutional layer on basis of the MBFNet-II) in the PPM.
(7) *-rp MBFNet-II-UODM:* Replacing the MBFNet-II with the MBFNet-I in the UODM.
(8) *-w/o MBFNet:* Removing all the MBFNets embedded in the PPM and UODM.
(9) *-rp ETE:* Replacing the step-by-step training strategy with the end-to-end training strategy to train the collaborative framework.

For a fair comparison, all candidate methods adopt the collaborative framework with ResNet-50 as the baseline, and the experiments employ the same experimental setup. Table **??** reports the results of ablation study about network composition. Some crucial conclusions can be revealed as follows:

- The effectiveness of our proposed PPM has been demonstrated. When either component in the PPM is removed, the corresponding AP values experience a decrease of approximately 11% to 50%. In addition, the channel concatenation has been shown to be an operation that significantly outperforms the element-wise addition, where the latter demonstrates a substantial decrease in different kinds of AP values to 14.5%, 13.6%, and 5.1%, respectively.
- The effectiveness of our proposed MBFNet has been demonstrated. We attempt to modify different versions of the MBFNet, but the detection performance with such changes is significantly inferior to the existing framework. Furthermore, when all the MBFNets are removed, the UOD performance of the framework experiences a drastic decline, with a decrease in different kinds of AP values to 19.1%, 18.9%, and 16.5%, respectively.

- The effectiveness of the step-by-step training strategy in our framework has been demonstrated. When our proposed collaborative framework is trained by the conventional end-to-end strategy, the corresponding AP values have even dropped by more than 30%, indicating a significant decrease in UOD performance.

Table 6. Results of ablation study regarding network composition. ↓ denotes the number after the arrow represents the change compared to ours.

Method	$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$
-w/o UIEC	35.91 (↓ 19.3%)	72.70 (↓ 11.6%)	19.20 (↓ 52.8%)
-w/o SRC	37.43(↓ 15.9%)	69.83 (↓ 15.1%)	33.89(↓ 16.6%)
-w/o PPM	37.91(↓ 14.8%)	69.76(↓ 15.2%)	35.89(↓ 11.8%)
-rp EWA-PPM	38.06(↓ 14.5%)	71.05 (↓ 13.6%)	38.67(↓ 5.1%)
-rp MBFNet-I-PPM	37.46 (↓ 15.8%)	77.31 (↓ 6.1%)	35.16(↓ 13.7%)
-rp MBFNet-III-PPM	40.07 (↓ 9.9%)	71.31 (↓ 13.3%)	39.05(↓ 9.9%)
-rp MBFNet-II-UODM	35.89(↓ 19.4%)	66.52 (↓ 19.1%)	33.12(↓ 18.6%)
-w/o MBFNet	36.01(↓ 19.1%)	66.76 (↓ 18.9%)	34.00 (↓ 16.5%)
-rp ETE	27.18(↓ 38.9%)	56.87 (↓ 30.9%)	25.13 (↓ 38.3%)
Ours	44.51	82.27	40.73

4. Conclusions

This paper presents a novel collaborative framework via joint image enhancement and super-resolution, which aims to address UOD tasks in complex marine environments. The proposed framework mainly consists of the PPM and the UODM, where the former can achieve pre-processing of underwater images to provide more effective visual cues for UOD tasks, and the UODM can effectively detect various fuzzy and difficult-to-find underwater objects. Moreover, a convenient self-attention mechanism termed MBFNet is developed, which can capture and associate scene information from a large neighborhood, so that the confusion between different object categories can be significantly reduced. Extensive experiments based on the URPC2020 and Brackish datasets reveal that our proposed collaborative framework outperforms other well-received competitors in terms of both quantitative evaluation metrics and qualitative detection effects. Additionally, results from the ablation study also demonstrate the effectiveness of each component in our framework.

Author Contributions: Conceptualization, X.J. and G.-P.L.; methodology, X.J. and G.-P.L.; software, G.-P.L.; validation, X.J. and G.-P.L.; formal analysis, X.J.; investigation, G.-P.L.; resources, C.-T.C.; data curation, G.-P.L.; writing—original draft preparation, X.J.; writing—review and editing, X.J. and C.-T.C.; visualization, G.-P.L.; supervision, X.J. and C.-T.C.; project administration, X.J. and C.-T.C.; funding acquisition, X.J. and C.-T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 52171332), and the Key Projects of Heilongjiang Provincial Natural Science Foundation (Grant No. ZD2022F001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

. Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. https://doi.org/10.1109/LRA.2020.2974710.

. Wang, X.; Yang, L.T.; Meng, D.; Dong, M.; Ota, K.; Wang, H. Multi-UAV Cooperative Localization for Marine Targets Based on Weighted Subspace Fitting in SAGIN Environment. *IEEE Internet Things J.* **2022**, *9*, 5708–5718. https://doi.org/10.1109/JIOT.2021.3066504.

. Wright, A.E.; Conlin, D.L.; Shope, S.M. PAssessing the Accuracy of Underwater Photogrammetry for Archaeology: A Comparison of Structure from Motion Photogrammetry and Real Time Kinematic Survey at the East Key Construction Wreck. *J. Mar. Sci. Eng.* **2020**, *8*, 849. https://doi.org/10.3390/jmse8110849.
. Zhong, Y.; Chen, Y.; Wang, C.; Wang, Q.; Yang, J. Research on Target Tracking for Robotic Fish Based on Low-Cost Scarce Sensing Information Fusion. *IEEE Robot. Autom. Lett.* **2022**, *7*, 6044–6051. https://doi.org/10.1109/LRA.2022.3163.6.
. Zhang, D.; Han, J.; Cheng, G.; Yang, M.H. Weakly Supervised Object Localization and Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 5866–5885. https://doi.org/10.1109/TPAMI.2021.3074313.
. Liu, C.; Wang, Z.; Wang, S.; Tang, T.; Tao, Y.; Yang, C.; Li, H.; Liu, X.; Fan, X. A New Dataset, Poisson GAN and AquaNet for Underwater Object Grabbing. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 2831–2844. https://doi:10.1109/TCSVT.2021.3100059..
. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. https://doi.org/10.1109/TCSVT.2019.2963772.
. Ancuti, C.O.; Ancuti, C.; De Vleeschouwer, C. Color balance and fusion for underwater image enhancement. *IEEE Trans. Image Process.* **2017**, *27*, 379–393.
. Yeh, C.; Lin, C.H.; Kang, L.W; Huang, C.H.; Lin, M.H.; Chang, C.Y.; Wang, C.C. Lightweight deep neural network for joint learning of underwater object detection and color conversion. *IEEE Trans. Neural Netw. Learn Syst.* **2021**, *33*, 6129–6143.
. Chen, X.; Li, H.; Wu, Q.; Ngan, K.N.; Xu, L. High-quality R-CNN object detection using multi-path detection calibration network. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 715–727.
. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance Problems in Object Detection: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3388–3415.
. Zhao, Z.; Zheng, P.; Xu, S.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn Syst.* **2019**, *30*, 3212–3232.
. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.
. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 386–397.
. 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.* **2016**, *39*, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.
. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916.
. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 37, pp. 379–387.
. Lin, T. Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
. Zeng, L.; Sun, B.; Zhu, D. Underwater target detection based on Faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104190.
. Xu, F.; Wang, H.; Sun, X.; Fu, X. Refined marine object detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy. *Neural. Comput. Appl.* **2022**, *34*, 14881–14894.
. Zhou. Liu, J.; Liu, S.; Xu, S.; Zhou, C. Two-Stage Underwater Object Detection Network Using Swin Transformer. *IEEE Access* **2022**, *10*, 117235–117247. http://doi:10.1109/ACCESS.2022.3219592..
. Song, P.; Li, P.; Dai, L.; Wang, T.; Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing* **2023**, *530*, 150–164, https://doi.org/10.1016/j.neucom.2023.01.088..
. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD : Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*; Springer: Cham, Switzerland, 2016; pp. 21–37.
. Zhou. X.; Dequan, W.; Philipp, K. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
. 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 (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324.
. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
. 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* **2022**, arXiv:2207.02696
. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872.
. Kim, B.; Mun, J.; On, K.W.; Shin, M.; Lee, J.; Kim, E.S. MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19556–19565. https://doi.org/10.1109/CVPR52688.2022.01897.

. Li, Y.; Guo, J.; Guo, X.; Zhao, J.; Yang, Y.; Hu, Z.; Jin, W.; Tian, Y. Toward in situ zooplankton detection with a densely connected YOLOV3 model. *Appl. Ocean Res.* **2021**, *114*, 1879–15.
. Hu, J.; Zhao, D.; Zhang, Y.; Zhou, C.; Chen, W. Real-time nondestructive fish behavior detecting in mixed polyculture system using deep-learning and low-cost devices. *Expert Syst. Appl.* **2021**, *178*, 115051.
. Wang, Z.; Chen, H.; Qin, H.; Chen, Q. Self-Supervised Pre-Training Joint Framework: Assisting Lightweight Detection Network for Underwater Object Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 604. https://doi.org/10.3390/jmse11030604.
. Zhang, Z.; Lu, X.; Cao, G.; Yang, Y.; Jiao, L.; Liu, F. ViT-YOLO:Transformer-Based YOLO for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 Ocotber 2021; pp. 2799–2808. https://doi.org/10.1109/ICCVW54120.2021.00314.
. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight Underwater Object Detection Based on YOLO v4 and Multi-Scale Attentional Feature Fusion. *Remote Sens.* **2021**, *13*, 4706. https://doi.org/10.3390/rs13224706.
. Zhao, S.; Zheng, J.; Sun, S.; Zhang, L. An Improved YOLO Algorithm for Fast and Accurate Underwater Object Detection. *Symmetry* **2022**, *14*, 1669. https://doi.org/10.3390/sym14081669.
. Sun, Y.; Zheng, W.; Du, X.; Yan, Z. Underwater Small Target Detection Based on YOLOX Combined with MobileViT and Double Coordinate Attention. *Mar. Sci. Eng.* **2023**, *11*, 1178. https://doi.org/10.3390/jmse11061178.
. Jian, M.W.; Liu, X.Y.; Luo, H.J.; Lu, X.; Yu, H.; Dong, J. Underwater image processing and analysis: A review. *Signal Process. Image Commun.* **2021**, *91*, 116088.
. Qi, Q.; Zhang, Y.; Tian, F.; Wu, Q.J.; Li, K.; Luan, X.; Song, D. Underwater image co-enhancement with correlation feature matching and joint learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1133–1147.
. Jobson, D.J.; Rahman, Z.; Woodell, G.A. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **1997**, *6*, 965–976.
. Ancuti, C.; Ancuti, C.O.; Haber, T.; Bekaert, P. Enhancing underwater images and videos by fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 81–88.
. Peng, Y.T.; Cosman, P.C. Underwater image restoration based on image blurriness and light absorption. *IEEE Trans. Image Process.* **2017**, *26*, 1579–1594.
. Peng, Y.; Cao, K.; Cosman, P.C. Generalization of the Dark Channel Prior for Single Image Restoration. *IEEE Trans. Image Process.* **2018**, *27*, 2856–2868.
. Cheng, Z.; Yang, Q.; Sheng, B. Deep colorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 415–423.
. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robot. Autom. Lett.* **2017**, *3*, 387–394.
. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.H.; Liao, Q. Deep Learning for Single Image Super-Resolution: A Brief Review. *IEEE Trans. Multimedia* **2019**, *21*, 3106–3121. https://doi.org/10.1109/TMM.2019.2919431.
. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307.
. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2599–2613. https://doi.org/10.1109/TPAMI.2018.2865304.
. Chang, H.; Yeung, D.Y.; Xiong, Y. Super-resolution through neighbor embedding. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; Volume 1.
. Park, S.C.; Park, M.K.; Kang, M.G. Super-resolution image reconstruction: A technical overview. *IEEE Signal Process. Mag.* **2003**, *20*, 21–36.
. Passarella, L.S.; Mahajan, S.; Pal, A.; Norman, M.R. Reconstructing high resolution ESM data through a novel fast super resolution convolutional neural network (FSRCNN). *Geophys. Res. Lett.* **2022**, *49*, e2021GL097571.
. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4467–4480.
. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.

. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*; Springer International Publishing: Cham, Switzerland, 2016; pp. 694–711.
. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
. Wei, J.; Wang, S.; Huang, Q. F3Net: Fusion, feedback and focus for salient object detection. In Proceedings of the American Association for Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12321–12328.
. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.