

Article

Underwater Image Enhancement and Small Object Detection Method Based on RBE-CycleGAN and MSFDC-Net

Zongren Li ^{1,†}, Chundong Xu ^{2,†}, Wenjun Hui ¹, Rui Chen ¹ and Xiaofang Kong ^{1,*}

¹ National Key Laboratory of Transient Physics, Nanjing University of Science and Technology, Nanjing 210094, China; lizongren2026@163.com (Z.L.); wenjunviej@163.com (W.H.); chenerqi@njust.edu.cn (R.C.)

² School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; chundongxu@126.com

* Correspondence: xiaofangkong@njust.edu.cn

† These authors contributed equally to this work.

Abstract

Underwater object detection plays a vital role in marine exploration and resource exploitation. However, complex underwater environment leads to severe color deviation, blurring, and information loss of small targets, which greatly restrict detection performance. To address these problems, this paper integrates the Channel Attention and Spatial Attention Block (CASAB) attention mechanism into residual blocks based on generative adversarial networks to correct color distortion and improve the clarity of degraded underwater images. For underwater small object detection, MobileNetV2 is selected as the backbone network within the Faster R-CNN framework, and a multi-scale feature fusion strategy is adopted to reduce feature loss caused by repeated downsampling. In the detection head, coordinate attention and parallel dilated convolution are further integrated to suppress background noise and expand the receptive field of feature extraction. Experimental results on the Underwater Robot Professional Contest (URPC) dataset demonstrate that the proposed method yields gains of 10.06%, 9.43%, and 12.29% in three evaluation metrics: Underwater Image Quality Measure (UIQM), Underwater Colour Image Quality Evaluation (UCIQE) and Natural Image Quality Evaluator (NIQE), together with 7.81% in Mean Average Precision (mAP) and an 8.57% increase in Mean Recall (mRecall). These results demonstrate the effectiveness of all improvements.

Keywords: residual block; Channel Attention and Spatial Attention Block; MobileNetV2; Feature Pyramid Network; coordinate attention; dilated convolution



Academic Editor: Ana Cláudia Teodoro

Received: 6 May 2026

Revised: 12 June 2026

Accepted: 17 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Marine exploration is of great significance to humanity. It serves as an indispensable technical foundation for both marine resource exploitation and marine ecosystem conservation. With the rapid advancement of marine robotics technology, vision-based underwater object detection has emerged as a cost-effective and promising approach for marine exploration, drawing extensive attention from marine scientific research and engineering communities [1,2]. Nevertheless, the intricate underwater environment frequently leads to low image contrast and severe color distortion, which greatly hinders the in-depth utilization and analytical research of underwater images [3]; moreover, the small proportion of target pixels as well as complex background interference also poses considerable challenges to underwater object detection [4]. As illustrated in Figure 1 [5], the Jaffe–McGlamery theoretical model built upon linear superposition and aquatic medium modeling indicates

that underwater imaging consists of reflected and scattered components [6]. The scattering effect induced by suspended particles in water inevitably results in image blurring, obscuring critical details such as target edges and textures, and further degrading the recognition and localization accuracy of detection algorithms.

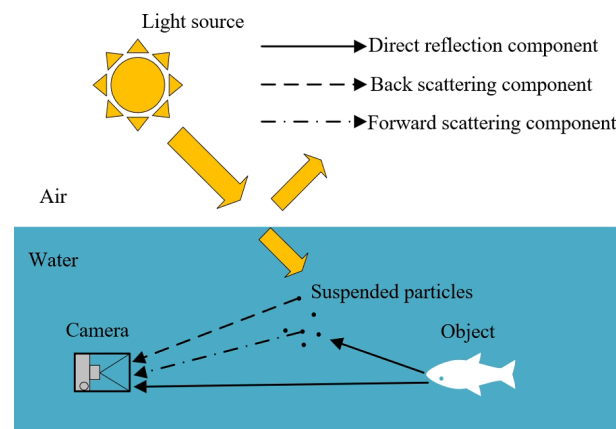


Figure 1. Diagram of an underwater optical imaging model. The three types of arrows in the figure represent the three components of underwater camera imaging, namely the direct reflection component, forward scattering component, and backward scattering component.

The selective absorption of light at different wavelengths by water bodies triggers obvious color deviation in underwater images. This distorts the color features of both targets and backgrounds, and further interferes with the accurate feature extraction of detection algorithms. The light attenuation characteristics in water are presented in Figure 2. The absorption effect of water medium on light is positively correlated with wavelength, and becomes more prominent as water depth increases. Red light with longer wavelength and lower frequency is preferentially absorbed by water, so most captured underwater images generally exhibit an overall blue-green tone [7].

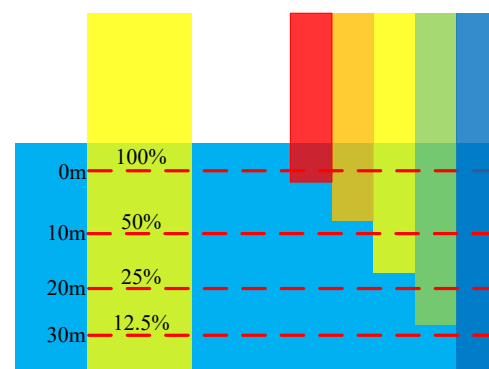


Figure 2. Diagram of light attenuation characteristics in water. The colored rectangular bars on the right correspond to sunlight of different hues.

Enhancing underwater images to correct color deviation can reduce feature matching errors caused by color distortion in object detection tasks. Improvements in image contrast, saturation and sharpness can strengthen the edge, texture and structural features of targets, making small and weak-textured objects more distinguishable and effectively mitigating the missed detection of tiny underwater targets. Preprocessing via image enhancement can lay a solid foundation for subsequent object detection. Traditional image enhancement methods such as histogram equalization [8] and wavelet transform [9] are derived from visual prior assumptions. Although these methods can improve image definition and contrast to some extent, they are incapable of effectively coping with illumination variations and severe color

distortion under complex underwater background conditions [10]. In recent years, deep learning-based image enhancement methods learn scene-adaptive features from ample training data. Such methods can restore fine details including image textures and edges, and effectively alleviate the color deviation problem of underwater images [11]. However, these methods pay little attention to key targets such as sea cucumbers and sea urchins in the image enhancement procedure and ignore the detailed edge features of objects, which is unfavorable to the implementation of downstream tasks [12]. This paper proposes a ResNet Block Enhanced-CycleGAN (RBE-CycleGAN). Built on the basic framework of CycleGAN, the proposed network embeds the Channel Attention and Spatial Attention Block (CASAB) attention mechanism into residual blocks. This design enables the model to concentrate on feature representation of key target regions, filter valid feature information and suppress redundant noise. By learning the mapping relationship between low-quality underwater images and high-quality clear ones from sufficient datasets, the method realizes end-to-end underwater image enhancement.

Most concerned objects in underwater detection tasks, such as sea cucumbers, sea urchins, scallops and starfish, are physically small in size and occupy only a tiny proportion of image pixels. In the image domain, targets with a resolution smaller than 32×32 pixels are generally defined as small objects [13]. When detecting small objects, repeated convolution and downsampling operations will gradually reduce the spatial size of feature maps and weaken the characteristic information of targets. Insufficient visual cues will further hinder the extraction of discriminative features for object classification and localization [14]. Complex underwater environment also creates severe interference for small object detection. For instance, similar tonal features between targets and backgrounds make small objects hard to identify, and analogous texture distributions lead to easy confusion with background regions. Meanwhile, occlusion in complicated underwater scenes further increases the difficulty of accurate object detection [15].

Due to the low proportion of effective pixels of small targets, single-stage detectors are prone to missed detection and false detection. Therefore, this paper adopts the two-stage detection algorithm Faster R-CNN with MobileNetV2 as the backbone network for underwater small object detection. To address the loss of small target information caused by multiple downsampling operations, a multi-scale feature fusion strategy is adopted. The original backbone network only feeds the final-layer feature map into the detection head for proposal generation. Nevertheless, the excessive spatial shrinkage of the last feature map seriously limits the effective feature extraction of small targets. Considering that the edge and texture features of small targets mainly reside in middle-level features, the proposed Multi-Scale Feature Dilated Convolution Network (MSFDC-Net) fuses the third, fourth and fifth inverted residual features extracted from the backbone network. These layers contain abundant edge and texture clues of targets, which can supplement feature information for tiny objects and alleviate feature loss. The fused multi-scale features are subsequently delivered to the detection head for final detection.

To tackle the occlusion and background interference faced by small targets, as well as the limited receptive field of standard convolution in feature extraction, this paper designs a parallel dilated convolution module integrated with coordinate attention. This module adopts convolutions with different dilation rates to extract multi-scale surrounding features in a parallel manner, which expands the receptive field coverage and establishes effective spatial dependency between targets and their surrounding context. Meanwhile, the Coordinate Attention (CA) [16] mechanism is introduced into the module. Through pooling operations along horizontal and vertical dimensions, CA performs cross-spatial information encoding on feature maps, and calculates the attention weight of each pixel in a channel-wise manner. This enables the network to adaptively focus on small target regions,

suppress background interference and redundant features, and alleviate the problem that small targets are easily occluded by complex surroundings and submerged in irrelevant background information. The main innovations of this paper are summarized as follows:

1. Aiming at the blurring degradation and color deviation of small objects induced by complex underwater environment, this paper proposes an underwater image enhancement algorithm named RBE-CycleGAN. On the basis of CycleGAN, the generator architecture is optimally redesigned. By embedding the CASAB attention mechanism into residual connection blocks, the proposed method achieves effective sharpness enhancement and color restoration for key target objects.
2. To solve the problem of small target information loss caused by repeated downsampling, this paper proposes an underwater object detection algorithm called MSFDC-Net. It optimizes the feature extraction structure of the backbone network and fuses partial middle-layer and deep-layer features. This approach effectively compensates for the lost detailed information of small targets and improves the model's perception capability for tiny underwater objects.
3. Aiming at the limited receptive field of single-scale convolution in the Region Proposal Network (RPN) Head, this paper proposes a parallel dilated convolution module integrated with coordinate attention. It enables effective capture of spatial context around targets and accurately models the spatial dependency between objects and their surrounding environment.

This paper is organized in accordance with the sequential order of two major tasks: underwater image enhancement and object detection. It first introduces the basic theories and improved strategies of the underwater image enhancement algorithm. Subsequently, the relevant theories and improvement schemes of underwater object detection are elaborated. Finally, comparative experiments and ablation experiments are conducted to fully verify the effectiveness of the proposed RBE-CycleGAN in correcting color deviation and improving image clarity, as well as the performance improvement effect of MSFDC-Net on underwater object detection.

2. Related Works

2.1. Underwater Image Enhancement Algorithm

Underwater image enhancement mainly aims to correct image blurring and color distortion caused by light scattering from suspended particles and selective light absorption in water. With preprocessing operations such as deblurring and enhancement of edge and texture details, the color distribution of underwater images can be restored to be closer to the normal imaging characteristics in the air environment. This method effectively improves image clarity and visual quality, facilitates the extraction of object feature information, and further boosts the performance of subsequent underwater object detection tasks [17].

Most traditional underwater image enhancement methods rely on fixed mathematical and physical models to directly correct and process image pixel values. Huang et al. [18] proposed a relative global histogram stretching algorithm with adaptive parameter updating. Aiming at the low contrast and color shift problems of underwater images, the method adaptively adjusts the stretching range of the histogram to improve visual visibility and restore scene details, while effectively avoiding artifacts and extra noise introduced during the enhancement process. Restricted by the processing logic of global transformation, this algorithm can hardly take into account the differences among local image regions, and presents limited refinement and enhancement effects on weak-texture areas and small-sized targets.

Drews et al. [19] put forward an improved underwater dark channel prior algorithm to solve the inaccurate transmittance estimation existing in the conventional dark channel

prior method for underwater scenes. This method removes the interference component of the red channel, and establishes the dark channel prior model merely using blue and green color channels. It greatly optimizes the calculation accuracy of light transmittance in underwater scenes. However, the algorithm highly depends on the inherent texture information of the captured scene, thus exhibiting unsatisfactory robustness in complex and dynamically changing underwater environment.

Chiang et al. [20] presented a wavelength compensation and dehazing method. According to the differential attenuation rule of light wavelengths underwater, the approach compensates and restores severely degraded color channels via wavelength difference correction. Combined with the atmospheric scattering dehazing model, it inversely deduces the scattering components in water, so as to rectify the image quality degradation caused by turbid underwater conditions.

Li et al. [21] proposed a GBR-based dehazing and red channel correction method. The scheme performs transmittance refinement and dehazing processing on blue and green channels respectively, and adopts independent color gain compensation for the red channel which suffers severe attenuation and distortion. This approach effectively alleviates low visibility and insufficient contrast in underwater images. Nevertheless, the method only conducts differentiated optimization at the channel dimension, while adopting a global unified correction strategy in the spatial dimension, which limits its enhancement performance for local image regions.

Deep learning-based image enhancement methods learn feature representations of diverse scenes from massive datasets. They can restore texture, edge and other fine details, correct color deviation of underwater images, and further improve the visual quality and practical applicability of underwater images for subsequent object detection tasks. Li et al. [22] proposed the UWCNN algorithm for underwater image enhancement to address severe light scattering, overall greenish tone deviation and reduced clarity in underwater images. Constructed on the prior imaging of underwater scenes, the convolutional processing module of this method can effectively suppress color distortion and improve the sharpness of underwater images. However, the model is trained mainly with synthetic datasets, which leads to pixel over-compensation when dealing with real underwater images under extreme and severe attenuation conditions.

Wang et al. [23] proposed UIEC²-Net, a convolutional network with joint optimization of dual color spaces RGB and HSV. With a multi-branch convolutional structure, the network separately fulfills the tasks of image denoising, color deviation correction, as well as brightness and saturation adjustment. Among them, the RGB module is responsible for noise suppression and color bias removal; the global HSV adjustment module tunes image attributes through neural curve layers; and the attention mapping module fuses RGB and HSV outputs to generate visually clear images. Even so, this method delivers unsatisfactory enhancement results for underwater images with yellowish and bluish tones. When coping with severely color-shifted underwater images under extreme conditions, the H channel is prone to local color distortion.

Zhu et al. [24] put forward the Cycle-Consistent Generative Adversarial Network (CycleGAN). Adopting a dual-generator and dual-discriminator architecture, the model leverages bidirectional adversarial learning together with cycle consistency constraints. It is capable of modeling both image degradation and enhancement processes with unpaired data, providing a viable unsupervised learning paradigm for underwater image enhancement tasks. Li et al. [25] designed an unsupervised generative adversarial network named WaterGAN. The generator of this method is built upon the physical propagation model of underwater light. By modeling three key processes including light attenuation, scattering

and camera imaging effect, it achieves color correction for underwater images under the condition of unpaired training data.

Liu et al. [26] proposed a multi-scale feature fusion algorithm named MLFcGAN to solve the unsatisfactory enhancement performance caused by image enhancement networks merely relying on local features. In the encoder stage, the algorithm adopts residual structures to extract multi-scale local features and acquires global semantic information through gradual downsampling operations. By fusing global semantics with multi-scale local features, the method effectively improves the accuracy of color deviation correction and preserves texture details better for underwater images.

Synthesizing the aforementioned underwater image enhancement algorithms, two common limitations still remain: insufficient feature extraction capability and lack of effective attention to key target regions. Most network structures designed merely based on underwater scene priors do not introduce in-depth feature mining modules, making it difficult to accurately explore the deep semantic features of underwater images. Although UIEC²-Net incorporates an attention mechanism, it only applies attention to the global adjustment of color, brightness and saturation, failing to focus on salient target areas or enhance object details in a targeted manner, which is unfavorable to subsequent downstream detection tasks. To enhance the feature extraction ability of the enhancement network and strengthen its focus on key target regions, this paper constructs the model on the basis of cycle-consistent generative adversarial networks. The CASAB attention mechanism is embedded into residual connection blocks. Through joint modeling of channel and spatial attention, the semantic feature channels related to targets are enhanced, and key object regions can be automatically located while background interference is suppressed. Meanwhile, global consistency constraints are maintained to achieve high-quality underwater image enhancement.

2.2. Underwater Target Detection Algorithm

Compared with traditional methods that manually design and extract low-level visual features, deep learning maps raw image pixel information into high-level and abstract semantic features layer by layer. It exhibits superior feature representation ability and strong generalization robustness in underwater object detection tasks.

Zhang et al. [27] proposed a lightweight underwater object detection algorithm based on YOLOv4 and multi-scale attention feature fusion. The algorithm adopts MobileNetV2 as the backbone network, and reconstructs the Neck and Head modules with depthwise separable convolution to reduce network parameters. Meanwhile, Mish activation function is used to replace ReLU6, which effectively enhances the nonlinear representation capability of the model. Li et al. [28] presented a novel detection model named CME-YOLOv5. By replacing the original C3 module with the improved C3CA module and adding an extra detection branch, the model strengthens the feature extraction capability for small targets and achieves better detection performance. In addition, EIOU loss is adopted to replace GIOU loss, which optimizes the bounding box regression speed and positioning accuracy. Nevertheless, this algorithm is only optimized for dense small fish targets, and thus shows poor adaptability when transferred to other underwater detection scenarios. Wang et al. [29] developed a lightweight underwater object detection algorithm named B-YOLOX-S. At the network input, Poisson blending is deployed to balance the quantity of training samples, and wavelet transform is introduced to restore image quality. In the Neck part, BiFPN-S is embedded to realize multi-scale features fusion and raise the utilization rate of multi-scale information, which further improves the detection performance for tiny underwater targets. Based on YOLOv5s, Hua et al. [30] proposed an underwater object detection method with feature enhancement and progressive dynamic aggregation. This method

designs a feature enhancement gating module. Combined with the attention mechanism, it can selectively strengthen valid feature information and suppress background noise. Meanwhile, dynamic weighted fusion is conducted between adjacent network layers to avoid the loss of feature details of small targets. Yang et al. [31] proposed a novel underwater detection network based on structural reparameterization. The method designs a cone-rod module that integrates depthwise convolution and standard convolution to simulate the visual perception mechanism of the human eye. It combines channel and spatial attention to suppress the interference of underwater optical noise, and adopts WIoU loss to optimize the regression accuracy of bounding boxes. Ouyang et al. [32] put forward a lightweight underwater object detection method with deformable upsampling. This work constructs a hybrid backbone that integrates CNN and Transformer structures. It can effectively strike a balance between global feature representation extraction and network lightweight optimization. A deformable upsampling module is newly designed to substitute the traditional fixed spatial upsampling scheme. This design effectively lowers deviation in multi-scale feature fusion and better accommodates geometric deformation characteristics of underwater targets. On this basis, the proposed method achieves obvious performance gains in the detection of tiny marine objects.

By summarizing existing research on small underwater target detection, two major challenges remain to be solved. Frequent downsampling operations easily erase feature details of tiny targets. It becomes difficult for networks to retain their fine-grained information. Standard convolutions also have narrow receptive fields. They cannot build effective spatial location modeling for small targets. As a result, models fail to lock key target areas and filter background noise. Localization accuracy drops significantly under such constraints. To address these problems, this work takes MobileNetV2 as the backbone and introduces a multi-scale feature fusion scheme. It compensates missing critical features of small targets from diverse scale dimensions. We further design an attention-embedded multi-branch parallel convolution module for the detection head. The module precisely locates underwater small targets, suppresses complicated background disturbances, and remedies the weak spatial representation ability of miniature underwater objects.

2.3. Underwater Image Feature Analysis

Suspended sediment and planktonic impurities in water damage the quality of underwater images. Backscattering from suspended particles blurs image details, while the selective light absorption of water bodies causes severe color distortion. As water depth increases, the overall image tone gradually changes from yellow to green and finally to blue. The complex image degradation restricts the performance of subsequent object detection tasks. Image blurring weakens the network's ability to extract edge and texture features of targets. Color deviation makes foreground objects visually similar to the background, which makes it hard for the detection model to distinguish targets from background interference and ultimately causes a high missed detection rate of underwater objects.

Classical underwater image enhancement methods have limitations. Histogram equalization only adjusts the gray distribution of image pixels; wavelet transform is mainly used for simple noise reduction; and dark channel prior achieves image defogging based on physical imaging rules. Although these methods can improve image quality to some degree, they cannot effectively solve the inherent color deviation problem of underwater images. Moreover, all of them adopt global pixel correction schemes, without targeted optimization for edge contours and texture details of detection targets.

Most conventional algorithms for underwater small target detection still adopt a single-path feature extraction architecture, including Faster R-CNN, YOLOv4 and YOLOv5. Such methods cannot address the inherent problems of small targets, such as low proportion

of pixels and feature loss induced by repeated downsampling operations. In addition, standard convolution has a constrained receptive field. When detecting small targets, it fails to fully capture spatial context information around target regions, making it difficult to separate foreground objects from complex background interference.

Most existing methods fail to address the defects in underwater image enhancement and object detection. To tackle the problems of color distortion and untargeted enhancement, this paper proposes RBE-CycleGAN, which embeds the CASAB mechanism into residual blocks. The model realizes end-to-end color correction of degraded underwater images by means of unpaired training data. Aiming at the low proportion of pixels of small targets and feature loss caused by repeated downsampling, this paper improves the classic Faster R-CNN framework. A multi-scale feature fusion strategy is introduced into the backbone network to make up for the lost detail information of small targets. Furthermore, the RPN detection head is optimized by adopting parallel dilated convolution combined with a coordinate attention module, which effectively solves the limited receptive field limitation of traditional 3×3 -pixel convolution mask. The overall technical framework of this paper is illustrated in Figure 3.

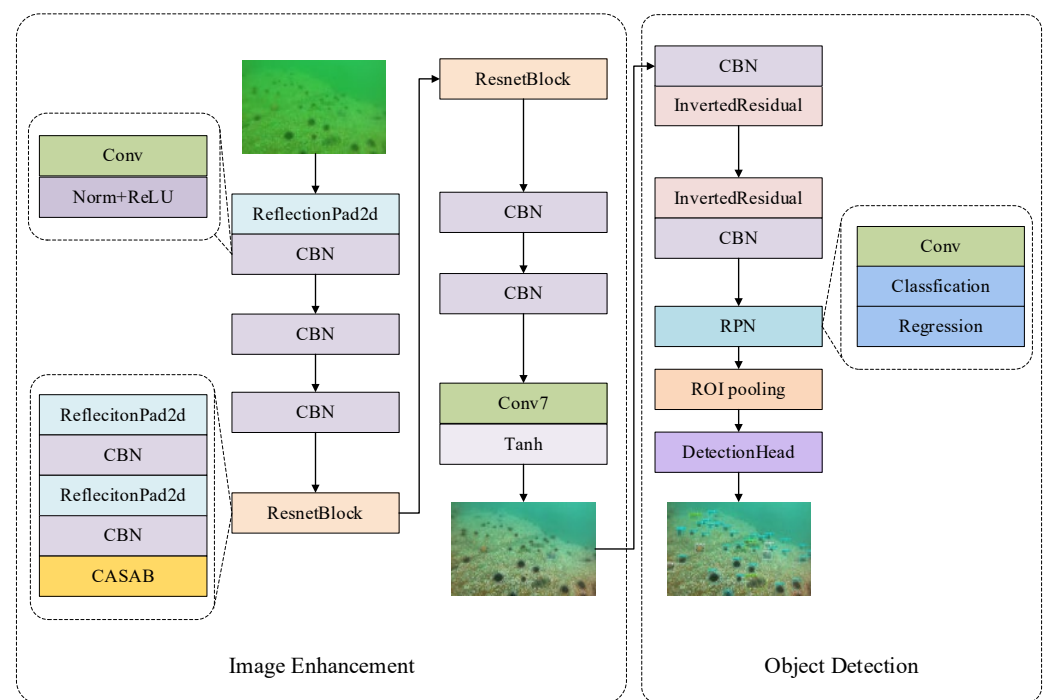


Figure 3. Overall network architecture diagram of image enhancement and object detection.

3. Underwater Image Enhancement and Small Target Detection Algorithm Based on RBE-CycleGAN and MSFDC-Net

3.1. Underwater Image Enhancement Algorithm Based on RBE-CycleGAN

Aiming at the pervasive color distortion problem in underwater object detection tasks, this paper proposes the RBE-CycleGAN algorithm, which enhances the residual block by introducing the CASAB attention mechanism. In the channel dimension, different weights are assigned to each feature channel to suppress noise interference and strengthen the network's ability to extract features from key image regions. In the spatial dimension, the feature receptive field is expanded to capture richer contextual information. By learning the bidirectional nonlinear mapping between the low-quality domain and high-quality domain with sufficient unpaired underwater data, the proposed method achieves end-to-end underwater image enhancement.

3.1.1. Image Enhancement Basic Theory

Generative adversarial networks mainly consist of three core modules: generator, discriminator and loss function [24]. The generator learns the degradation mechanism of underwater images from massive image data, and constructs a nonlinear mapping model from the low-quality domain to the high-quality target domain. It takes original low-quality underwater images as input and outputs enhanced high-quality results.

The discriminator is responsible for judging the domain attribution of samples, so as to distinguish real high-quality underwater images from synthetic samples generated by the generator. During the training process, the generator and discriminator form an adversarial iterative relationship. As the classification criteria of the discriminator become increasingly stringent, it reversely drives the generator to optimize the mapping model, making the generated samples continuously approximate the expected distribution of high-quality images, and finally realizing effective enhancement of low-quality underwater images.

The loss function measures the distribution difference between generated data and real data to evaluate the quality of synthesized images. It updates the parameters of the generator and discriminator through the back propagation algorithm, serves as the evaluation criterion for their performance, and guides the two modules to achieve continuous iterative optimization in the adversarial training process.

3.1.2. CASAB Attention Mechanism

The attention mechanism can focus on key information and filter redundant information during image feature extraction, thereby improving the efficiency of feature representation. Reasonable structural design of attention modules is of great significance for image generation tasks. Inspired by the Convolutional Block Attention Module (CBAM) [33], this paper introduces a dual-branch attention mechanism integrating channel attention and spatial attention to strengthen critical feature channels and important spatial regions.

Considering that the attention mechanism is mainly used for feature selection, and residual connections should keep the identity features intact to prevent the loss of original information, the CASAB module is embedded into each residual block of the generator. It is placed after two convolutional layers and before the addition of residual skip connections, acting as a selective gate for the learned residual features. This attention mechanism enhances informative feature channels via the channel attention module and highlights critical spatial regions through the spatial attention module. The resulting attention-weighted features are then incorporated into the residual connection. During feature propagation, the skip connection forms a direct gradient highway in backpropagation, while the CASAB branch delivers attention-weighted gradients. Such a layout of the attention mechanism enables effective screening of specific channels and regions, and meanwhile guarantees stable model training.

In the CASAB channel attention branch, global average pooling is adopted to capture contextual information of feature maps, and global maximum pooling is used to acquire salient regional features. The pooled features are processed through fully connected (FC) layers and the Sigmoid activation function to generate channel attention weights.

For spatial attention, mean pooling, maximum pooling, minimum pooling and sum pooling are performed on each channel of the feature map to extract spatial feature information. All pooled results are concatenated along the channel dimension. Large-kernel depthwise convolution is applied to reduce computational complexity while expanding the receptive field, so as to capture richer spatial context information. Spatial feature weights are finally obtained through the activation function.

The original feature map is multiplied with channel attention weights and spatial attention weights respectively to generate channel-weighted and spatially weighted feature

maps. The two weighted feature maps are then added element-wise to obtain the final output of the CASAB attention mechanism. The network structure of CASAB is shown in Figure 4 [34].

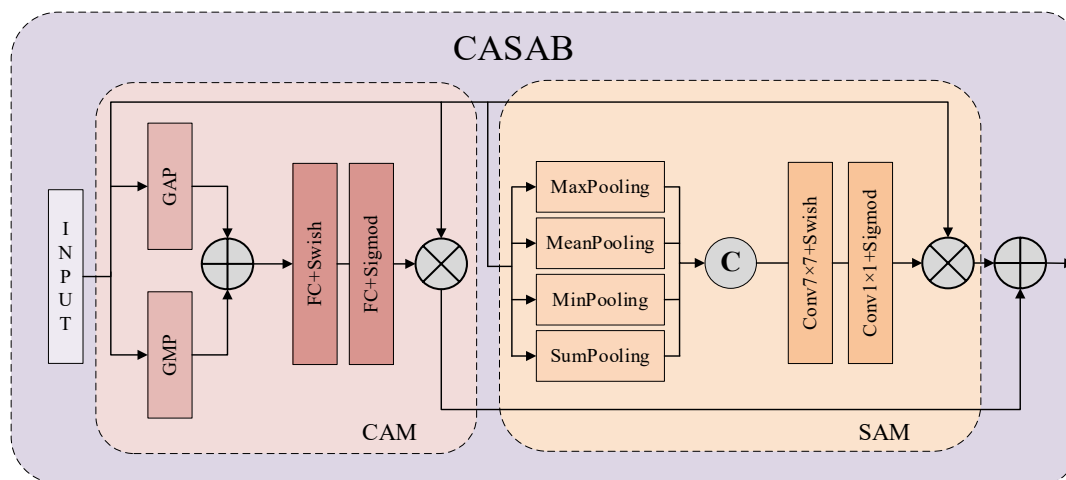


Figure 4. CASAB attention mechanism structure.

3.1.3. RBE-CycleGAN Network Structure

As shown in Figure 5, the generator adopts a ResNet-based residual structure.

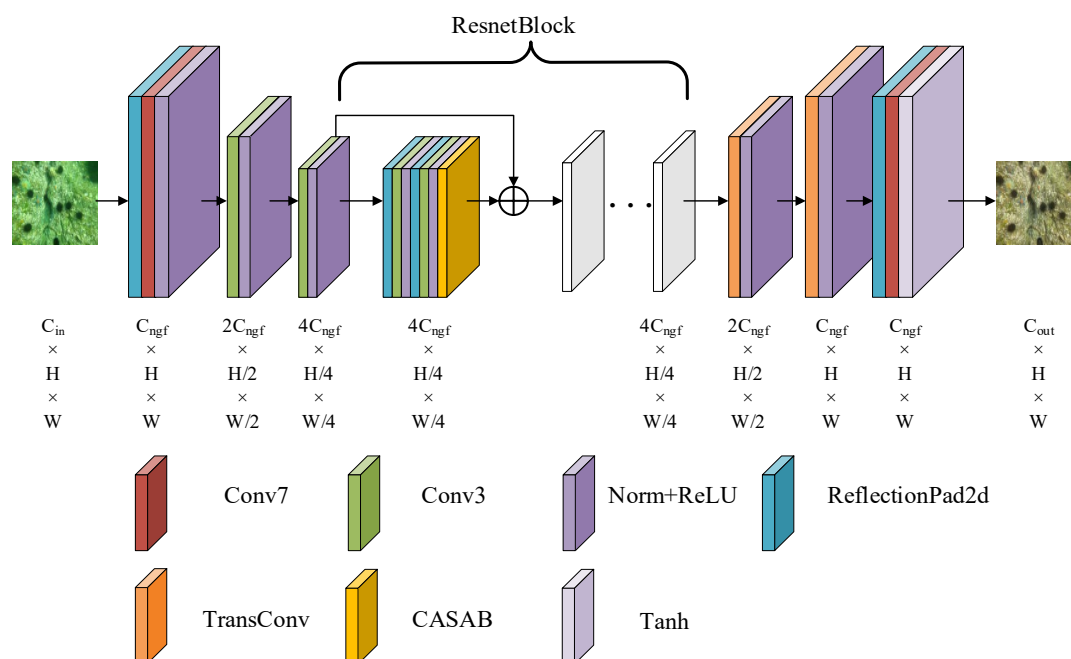


Figure 5. Schematic diagram of the RBE-CycleGAN generator network structure.

The input image is first subjected to edge padding to avoid the loss of edge information caused by convolution operations. Subsequently, a large-kernel convolution is applied to extract large-scale background and texture features of underwater images. Two down-sampling operations with a convolution kernel size of 3 and a stride of 2 are performed to reduce the spatial size of the feature map and further extract middle-level features such as local edges, corner points and texture details. Nine ResNet residual blocks are stacked to capture high-level deep features including complete target contours and semantic structural information. Two deconvolution operations are used to restore the original image resolution and map the extracted high-level semantic features back to pixel space. Finally,

a large-kernel convolution is adopted to compress the channel dimension of the feature map, and the Tanh activation function is employed to constrain the range of pixel values, yielding the final enhanced generated image.

As shown in Figure 6, the discriminator follows the PatchGAN [24] paradigm to expand the receptive field and maintain global consistency in patch-level discrimination. Instead of directly judging the authenticity of the entire image, PatchGAN performs multi-layer convolution and downsampling to extract deep features from the input image. The output feature map enables real-fake discrimination for each local patch within the image. The overall discrimination result is obtained by averaging the prediction scores of all patches.

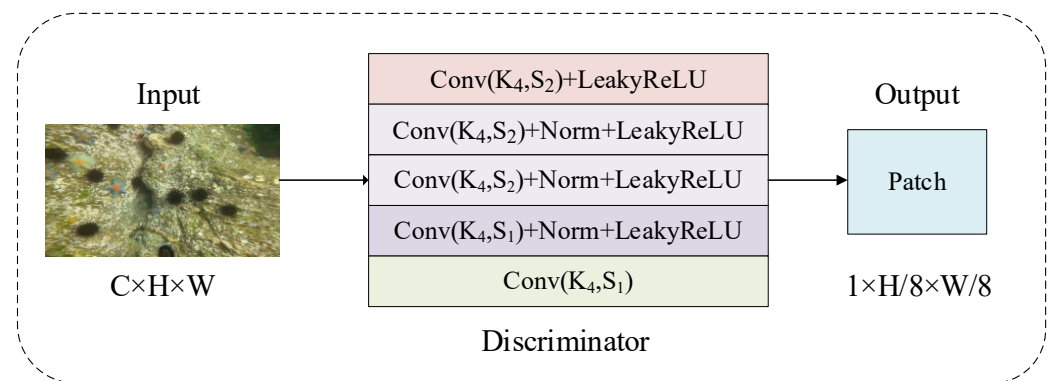


Figure 6. RBE-CycleGAN discriminator network structure diagram.

3.1.4. RBE-CycleGAN Loss Function

The loss function consists of adversarial cycle loss, identity consistency loss and cycle consistency loss. The adversarial cycle loss ensures the adversarial training effect between the generator and discriminator by measuring the difference between original images and generated images, so as to optimize the two networks. The identity consistency loss constrains the difference between original images and the outputs generated by the intra-domain generator, which guarantees that the generator can learn the correct image mapping and avoid excessive generation for images within the same domain. The cycle consistency loss compares original images with the restored images reconstructed by the generator, ensuring that the generator does not excessively alter image content and preserve the original image structure. The total loss function is obtained by weighted summation of the above three loss functions, as shown in Equation (1):

$$\text{Loss}(G, F, D_X, D_Y) = L_{GAN}(G, D_Y) + L_{GAN}(F, D_X) + \lambda_1 L_{cyc}(G, F) + \lambda_2 L_{idt}(G, F) \quad (1)$$

where G and F denote the generators of the high-quality domain and low-quality domain, respectively; D_X represents the discriminator of the high-quality domain; and D_Y denotes the discriminator of the low-quality domain. L_{GAN} is the adversarial cycle loss, L_{cyc} represents the cycle consistency loss, and L_{idt} stands for the identity consistency loss. λ_1 and λ_2 are the weight coefficients corresponding to the above loss functions, which are set to $\lambda_1 = 10$ and $\lambda_2 = 5$ in the equation.

3.2. Underwater Target Detection Based on MSFDC-Net

Aiming at the problem of low proportion of pixels of underwater small targets and frequent missed detection in single-stage detectors, this paper adopts a two-stage detection framework for underwater small target detection. Based on Faster R-CNN, an improved network named MSFDC-Net is proposed in this work. It redesigns the feature extraction mode of the backbone network and adopts the FPN strategy. The middle-level features

containing target edges and texture structures are selected from inverted residual blocks, and then fused with high-level features that carry global semantic information of targets. The fused multi-scale features are fed into the RPN Head to generate candidate detection boxes. Furthermore, the structure of the RPN Head is optimized to alleviate the limited receptive field caused by single-scale convolution. A parallel dilated convolution module embedded with coordinate attention is designed. The attention mechanism is used to screen potential target regions, while dilated convolutions with different dilation rates extract target features in a parallel manner. Feature fusion is subsequently conducted to further expand the receptive field of small target features.

3.2.1. Object Detection Basic Theory

As a classic two-stage object detection network, Faster R-CNN [35] consists of four core components: backbone network, Region Proposal Network (RPN), ROI Pooling, and detection classification head. The backbone network encodes the input image to obtain deep feature maps with rich semantic information. The RPN generates a series of candidate regions containing potential targets on the feature maps, preliminarily distinguishes foreground from background, and completes coarse target localization. The ROI Pooling layer maps candidate regions of different scales to the feature maps and achieves accurate feature alignment, so as to retain effective feature information of small targets. Finally, the detection head performs category classification and precise bounding box regression on the features of candidate regions to output the final detection results.

3.2.2. Multi-Scale Feature Fusion Module

Shallow residual blocks extract corner points, edges and other detailed information of underwater targets; middle-level residual blocks capture local contours and shape fragments; and deep residual blocks learn complete target contours and global semantic features. To visualize the features extracted by convolutional layers at different levels, this paper adopts heat maps to mark the attention regions of multi-scale convolution layers, as illustrated in Figure 7.

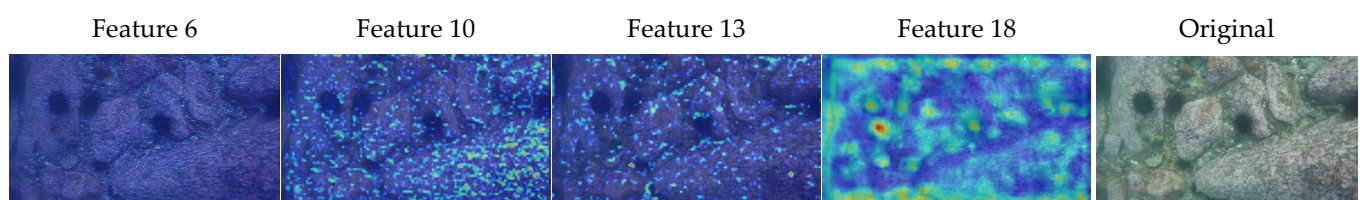


Figure 7. Thermal schematic diagram of features at different levels.

As for the original MobileNetV2 backbone, it implements five downsampling operations. After feature extraction via the backbone, a small object of 32×32 pixels is compressed into a single pixel. At this extreme scale, almost all discriminative cues such as object edges, textures and geometric shapes vanish. If deep features are utilized alone, the network can barely confirm the presence of targets but delivers inferior localization accuracy, resulting in severe missing detections of small objects.

To address the problem of small target feature loss caused by repeated downsampling, this paper fuses selected shallow and middle-level features with deep features to supplement missing information of small targets. The Feature6 layer corresponds to small targets mapped to a 4×4 -pixel resolution, retaining shallow-level feature information such as target corners and edges; the Feature10 layer matches small targets at 2×2 pixels, carrying local object contours and abundant semantic information; and the Feature13 layer also corresponds to 2×2 -pixel small targets with high-level semantic features that enable superior category classification capability. For instance, element-wise addition between

Feature10 and Feature6 via the MSF module injects semantic information into Feature6. This empowers Feature6 to distinguish target edges from background textures within each 4×4 -pixel region, thereby improving the network's overall target discrimination performance. The specific implementation process is shown in Figure 8.

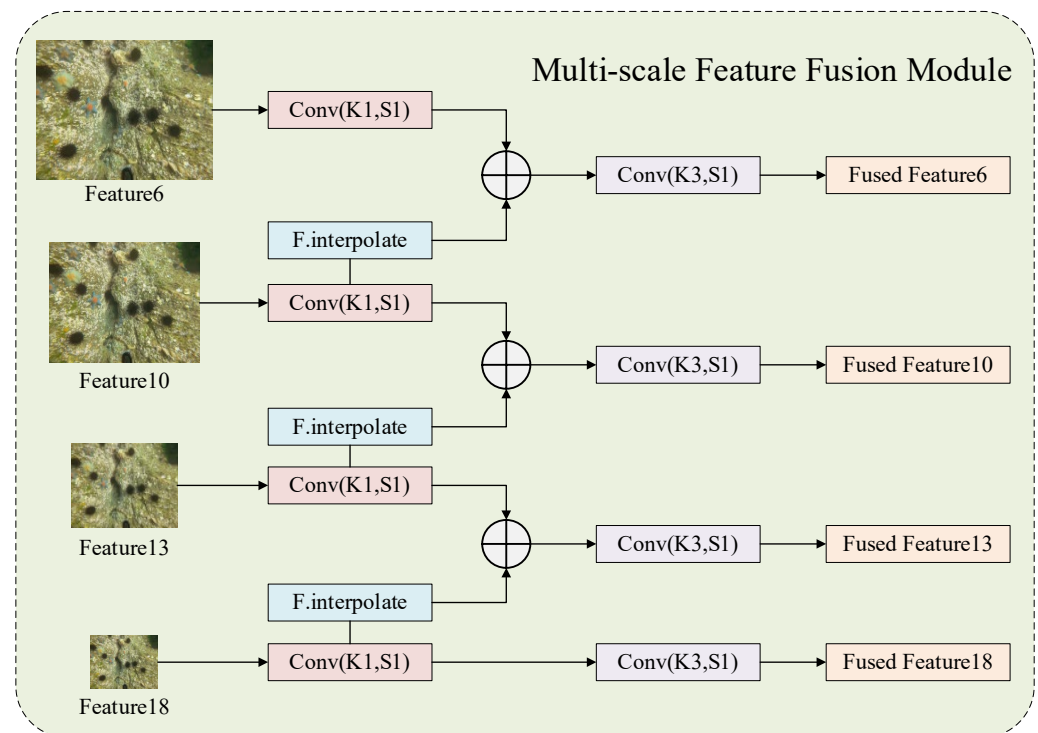


Figure 8. Multi-scale feature fusion module structure diagram.

First, a 1×1 -pixel convolution mask is applied to the deepest feature map Feature 18 to adjust its channel dimension to 256. The processed feature map is then divided into two branches. One branch adopts a 3×3 -pixel convolution mask for feature smoothing to obtain the output feature. The other branch adopts nearest-neighbor upsampling to align its spatial size with the upper feature map Feature 13. The two feature maps are added element-wise, followed by a 3×3 -pixel convolution mask smoothing operation to generate the Fused Feature 13.

Deep features contain high-level semantic information of the image, while middle-level features retain texture, edge and contour details. This fusion strategy endows middle-level features with high-level semantic cues and facilitates the network in target category judgment.

The above operation is repeated for Feature 13. After channel adjustment via 1×1 -pixel convolution mask and nearest-neighbor upsampling, Feature 13 is added element-wise with the upper feature map Feature 10 to obtain Fused Feature 10. This process enriches Feature 10 with multi-scale edge and texture information and improves the ability of the network to identify underwater targets.

Similarly, Feature 10 is upsampled and fused with Feature 6. This fusion introduces structured semantic information into shallow features and expands their receptive field, enabling the network to effectively distinguish background noise from real targets.

3.2.3. Parallel Dilated Convolution Fused with CA Module

CA is a hybrid attention mechanism that models attention in both channel and spatial dimensions. It explicitly embeds spatial coordinate information into channel attention, and

simultaneously captures channel dependencies and long-range spatial dependencies of feature information.

Standard channel attention methods such as SENet perform global average pooling on input feature maps of size $C \times H \times W$ to produce $C \times 1 \times 1$ feature vectors. This operation completely discards spatial position information of features and only learns the importance of individual channels, failing to inform the network of the spatial locations of target objects. In contrast, the coordinate attention assigns weights for each pixel by aggregating global statistical information from its entire corresponding row and column. Such a globally statistics-driven weighting mechanism behaves like a low-pass filter, which effectively suppresses local and discontinuous high-frequency noise.

The network structure of the CA attention mechanism is illustrated in Figure 9a. Global average pooling is performed along the vertical and horizontal directions to encode directional feature information, which captures long-range dependencies in a single direction while retaining precise positional information in the other direction.

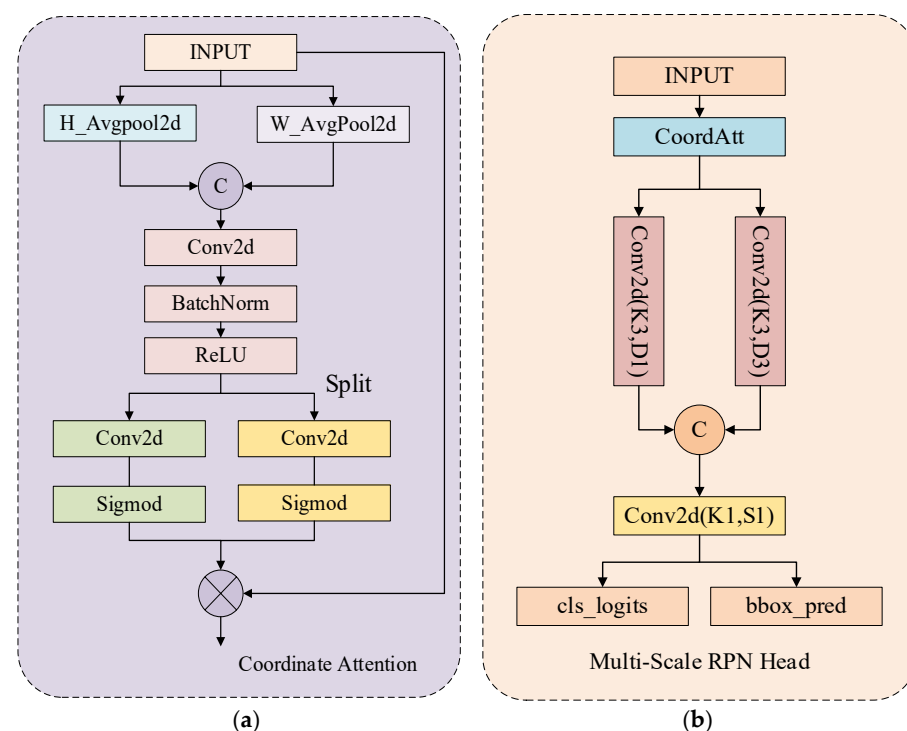


Figure 9. (a) Coordinate attention structure diagram, (b) Multi-scale RPN head structure diagram.

Feature maps from the two directions are concatenated in the spatial dimension, followed by convolution dimensionality reduction, feature splitting, convolution dimensionality elevation, and Sigmoid activation to generate coordinate attention weights. Finally, the generated attention weights are multiplied element-wise with the original input features to obtain the enhanced feature representation.

The Region Proposal Network (RPN) is a core component of Faster R-CNN, which is responsible for generating preliminary candidate regions on the shared feature map. The RPN Head undertakes feature extraction, as well as classification and bounding box regression tasks. The conventional RPN Head adopts a single 3×3 -pixel convolution mask to extract image features. However, underwater small targets occupy only a small proportion of pixels. Limited by the fixed receptive field of single-scale convolution, it is difficult to capture contextual information around small targets. Consequently, small targets are easily interfered by complex background surroundings and eventually lead to missed detection.

To solve this issue, this paper designs a parallel dilated convolution module integrated with coordinate attention, whose structure is shown in Figure 9b. The features output from the backbone are first enhanced by the coordinate attention module, which assigns pixel-level weights to suppress background noise interference. Relying on the long-range modeling capability in the spatial dimension, the module achieves preliminary perception of potential target regions.

Subsequently, two groups of convolutions with different dilation rates are adopted to extract features with diverse receptive fields at the same feature scale for small targets. The 3×3 -pixel dilated convolution mask with a dilation rate of 1 captures features of small targets and their adjacent backgrounds to preliminarily distinguish foreground targets from background clutter. The 3×3 -pixel dilated convolution mask with a dilation rate of 3 further expands the receptive field and captures long-range contextual information, which helps refine the separation between small targets and backgrounds and strengthens the perception ability of surrounding spatial information.

3.2.4. MSFDC-Net Network Structure

MSFDC-Net adopts the improved MobileNetV2 as the backbone feature extraction network. Through multi-scale feature fusion of middle-level and high-level features, the network achieves effective feature extraction for underwater small targets. The overall network structure is illustrated in Figure 10, and its workflow is described as follows.

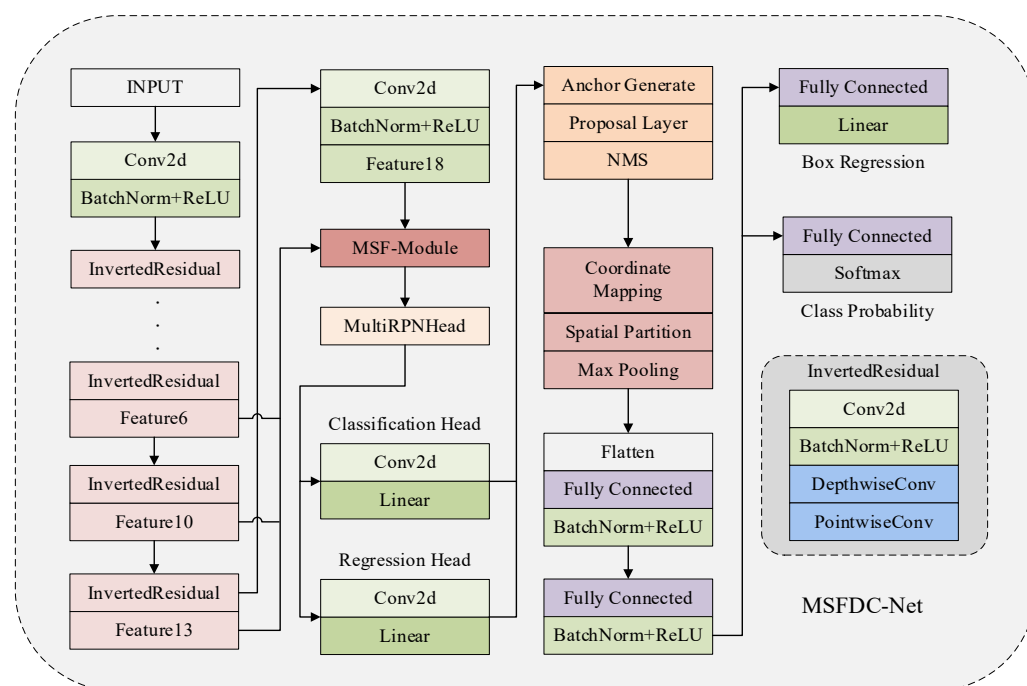


Figure 10. MSFDC-Net structure.

The input image first passes through a 3×3 -pixel convolution mask to expand channel dimensions and extract shallow features such as corner points and texture details. Then the features are fed into inverted residual blocks to extract multi-scale information including edges, contours and global semantics. The fused shallow, middle and deep features are subsequently delivered to the detection head for target detection.

In the detection head, the coordinate attention mechanism assigns pixel-wise weights to the input feature maps. It focuses on key target information, suppresses background noise, highlights potential target regions, and completes the rough localization of small targets. Meanwhile, parallel dilated convolution is employed to extract image features and

expand the receptive field, which fully captures spatial contextual information and assists the RPN in distinguishing foreground targets from complex backgrounds.

For the screened anchor boxes, non-maximum suppression (NMS) is adopted to eliminate redundant proposals. The remaining candidate boxes are mapped to the shared backbone feature maps, and ROI Pooling is used to unify the spatial size of different candidate regions. Finally, the detection branch and classification branch jointly complete the category recognition and precise location regression of underwater targets.

3.2.5. Object Detection Loss Function

The loss function of Faster R-CNN consists of two components: RPN loss and detection head loss. The overall loss is the weighted sum of these two parts, as shown in Equation (2):

$$Loss = L_{cls}^{RPN} + L_{reg}^{RPN} + L_{cls}^{R-CNN} + L_{reg}^{R-CNN} \quad (2)$$

where L_{cls}^{RPN} and L_{reg}^{RPN} are the classification loss and regression loss of the RPN network, which are used to distinguish foreground from background and generate preliminary candidate proposals. L_{cls}^{R-CNN} and L_{reg}^{R-CNN} denote the losses of the ROI Head, which are responsible for refining the position of candidate boxes and completing the final category classification.

4. Experimental Results and Analysis

This paper verifies and analyzes the performance of the improved RBE-CycleGAN and MSFDC-Net from two dimensions: subjective visual evaluation and objective quantitative indicators.

4.1. Experimental Setup

The network models are trained on the Windows operating system based on the PyTorch 12.8 framework. The CPU is Intel Core i7-14700KF with 32 GB of RAM, and the graphics card is NVIDIA GeForce RTX 5070Ti with 16 GB video memory.

4.2. Underwater Image Enhancement Results

The unpaired subset of the Enhancement of Underwater Visual Perception (EUVP) [36] dataset is selected as the training set, containing 3205 low-quality underwater images and 3140 high-quality underwater images. The test set adopts the dataset for downstream underwater object detection tasks, and the enhancement performance is validated from both subjective and objective perspectives. From the objective evaluation dimension, several unsupervised image enhancement metrics are employed, including the Underwater Image Quality Measure (UIQM), Underwater Color Image Quality Evaluation (UCIQE), and Natural Image Quality Evaluator (NIQE).

The Adam optimizer is adopted for model training, and the learning rate is dynamically adjusted with the training epochs to achieve stable model convergence. The parameter settings for the generator and discriminator are listed in Table 1. Specifically, the learning rate is set to 0.0002 during the first 100 training epochs, and then linearly decays to 0 in the subsequent 100 epochs. All other parameters use the default configuration.

The Underwater Robot Professional Contest (URPC) [37] and Detecting Underwater Objects (DUO) [38] datasets are adopted as the test set. Partial visual results of underwater image enhancement are shown in Figure 11.

The left column shows original underwater images covering typical degradation scenarios such as blue tone deviation, green tone deviation and blurriness, while the right column presents the corresponding enhanced results. As can be seen from Figure 11, the

overall color distortion of enhanced underwater images is effectively corrected, which helps alleviate the background occlusion problem caused by color deviation.

Table 1. Image enhancement network parameter.

Parameters	Values
BatchSize	3
Generator Learning Rate	0.0002
Discriminator Learning Rate	0.0002
First-order Decay Rate	0.5
Second-order Decay Rate	0.999

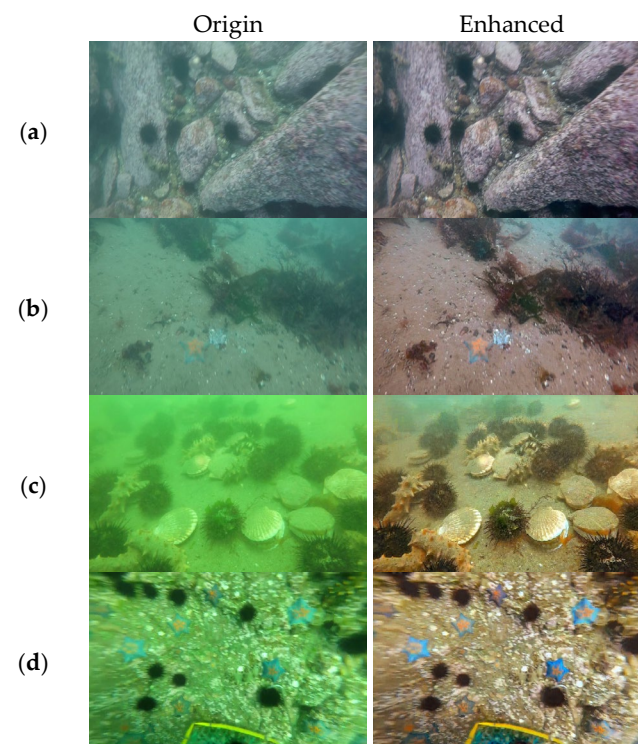


Figure 11. Illustration of image enhancement effects for several typical underwater degradation scenarios. (a) slight color deviation; (b) moderate color deviation and blurriness; (c) severe color deviation and blurriness; (d) severe color deviation.

For instance, in Group a and Group d, obvious improvements in color deviation can be observed. The images no longer exhibit a blue-green tint. In Group c, the enhanced image achieves improved clarity and contrast. The edge texture and detailed features of sea urchin targets become more distinguishable, providing strong support for accurate recognition of underwater small targets. In Group b, the original degraded image suffers from blurred starfish edges and similar color distribution between targets and background, which easily causes occlusion and missed detection. By contrast, the enhanced image exhibits clearer starfish contours and higher color contrast. It enables obvious separation between targets and background, and effectively improves the recall rate of underwater object detection.

The unsupervised image evaluation metrics of the URPC and DUO datasets are listed in Table 2.

UIQM is adopted to comprehensively reflect the color richness and clarity of underwater images, with a higher value corresponding to better overall visual quality. On the public URPC and DUO datasets, the images enhanced by RBE-CycleGAN obtain superior UIQM scores compared with the original images. This verifies that the proposed underwater

image enhancement method can effectively enhance the expressive ability of target edge and texture features.

Table 2. URPC and DUO dataset no-reference image evaluation metric values.

		UIQM	UCIQE	NIQE
URPC	Original	1.69	0.53	11.07
	Enhanced	1.86	0.58	9.71
	Gain	10.06%	9.43%	12.29%
DUO	Original	1.93	0.52	11.73
	Enhanced	2.17	0.57	9.89
	Gain	12.44%	9.62%	15.69%

UCIQE mainly evaluates color saturation and tonal deviation, and acts as a core metric to quantify the correction performance of underwater color distortion. The results on two test datasets demonstrate that RBE-CycleGAN can effectively eliminate the inherent color deviation of raw underwater images and achieve balanced tonal restoration.

NIQE is a natural image quality evaluation indicator; a lower NIQE value indicates less image distortion. Experimental results show that the NIQE values of images processed by RBE-CycleGAN are significantly reduced. It is validated that the proposed method can greatly improve underwater image quality and alleviate visual distortion caused by complex underwater optical environment.

4.3. Underwater Target Detection Effect

The URPC dataset is selected for model training, which contains images of four categories of underwater targets: sea cucumber, sea urchin, starfish and scallop under diverse underwater environment. The official bounding box annotations are provided in XML format, which can be directly applied to the training of object detection networks.

The training set of URPC includes 5543 images and the validation set contains 1200 images. For model testing, the URPC and DUO datasets are adopted. The URPC test set consists of 800 images, while the DUO test set contains 1111 images.

To evaluate the detection performance of the proposed model, comparative experiments are conducted with mainstream algorithms including Mask R-CNN [39], SSD [40], RetinaNet [41], and the YOLO series.

Transfer learning is adopted in this work. MobileNetV2 pre-trained on the ImageNet dataset is used as the backbone initialization weight, and a freeze-then-unfreeze training strategy is applied. The backbone is frozen for the first 5 epochs, then trained for the next 200 epochs. The relevant trainable parameters are presented in Table 3.

Table 3. Object detection parameter.

Parameters	Values
Epochs	205
BatchSize	4
Workers	4
LR Scheduler	Cosine Annealing LR
Optimizer	SGD
Momentum	0.9
Lr_unfreze	0.005
Lr_min	1×10^{-6}

Mean Average Precision (mAP) is employed to evaluate the detection accuracy of the model, which reflects the accuracy of target localization and classification. Mean Recall

(mRecall) is used to measure the model's ability to detect targets and characterize its missed detection level. The detection results are shown in Table 4.

Table 4. URPC and DUO dataset object detection results.

Method	URPC		DUO	
	mAP	mRecall	mAP	mRecall
Faster R-CNN	0.64	0.70	0.65	0.72
Mask R-CNN	0.63	0.68	0.62	0.70
SSD	0.46	0.48	0.43	0.45
RetinaNet	0.58	0.61	0.55	0.61
YOLOv8	0.65	0.69	0.64	0.68
YOLOv10	0.62	0.66	0.60	0.64
YOLOv11	0.63	0.66	0.64	0.67
MSFDC-Net	0.69	0.76	0.67	0.76
Gain	7.81%	8.57%	3.08%	5.56%

Values in bold black font represent the optimal performance among all compared methods.

All comparative experiments share identical dataset partitioning for training and validation sets, with consistent input resolution, training epochs, optimizer and learning rate decay strategy. All models are trained via transfer learning.

As shown in Table 4, MSFDC-Net achieves the optimal performance in terms of both mAP and mRecall on the URPC and DUO datasets. This result fully verifies the effectiveness of MSFDC-Net in enhancing the feature representation capability of underwater targets. The proposed improvements can significantly strengthen the model's ability to extract and perceive underwater object features, and further boost the detection performance.

The detection visualization results of several object detection algorithms on the URPC dataset are presented in Figure 12.

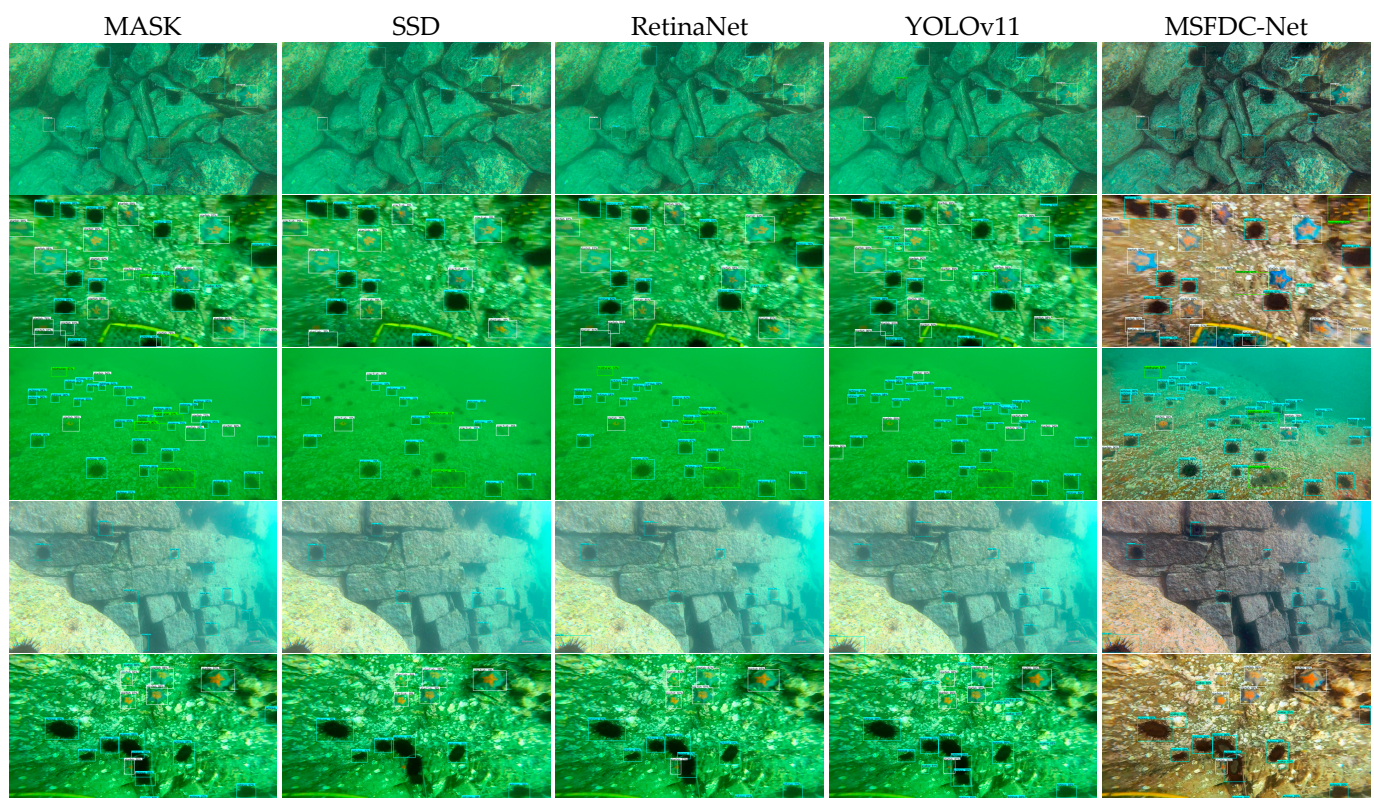


Figure 12. Result images of some object detection algorithms.

It can be observed from Figure 12 that the proposed MSFDC-Net detects and identifies more underwater targets with denser bounding boxes, which is consistent with the improved recall rate reflected in the quantitative experimental results.

To further verify the performance gain of the underwater image enhancement algorithm for the object detection task, comparative experiments are conducted in this paper. With the same detection model adopted, object detection is performed on original underwater images and enhanced images respectively.

The experimental results are presented in Table 5. Underwater image enhancement can effectively correct color deviation and enlarge the color difference between target regions and background regions, thereby improving the recall rate of object recognition. Meanwhile, the overall clarity of images is improved, which helps the detection model fully extract detailed features such as edges and textures of small objects and further boost detection accuracy. The comparative experiments demonstrate that applying underwater image enhancement as a preprocessing method can effectively facilitate the subsequent object detection task.

Table 5. Comparison of detection results on original and enhanced images.

Method	Original		Enhanced	
	mAP	mRecall	mAP	mRecall
Faster R-CNN	0.64	0.70	0.66	0.71
Mask R-CNN	0.63	0.68	0.65	0.70
SSD	0.46	0.48	0.47	0.49
RetinaNet	0.58	0.61	0.60	0.62
YOLOv8	0.65	0.69	0.67	0.70
YOLOv10	0.62	0.66	0.64	0.67
YOLOv11	0.63	0.66	0.65	0.67

4.4. Ablation Experiment

To further verify the effectiveness of the improved strategies in this paper, ablation experiments are designed on the URPC dataset to individually validate the performance gain of each module for the object detection network. Four comparative schemes are set in the experiments:

- (1) Only introducing the multi-scale feature fusion strategy into the backbone network, denoted as FPN;
- (2) Only embedding the parallel dilated convolution and coordinate attention module into the RPN Head, denoted as RPN;
- (3) Only adopting the image enhancement preprocessing strategy, denoted as Enhanced;
- (4) Integrating image preprocessing, multi-scale feature fusion and parallel dilated convolution modules simultaneously, denoted as MSFDC-Net.

The results of each ablation experiment are listed in Table 6, where Base represents the baseline method of Faster R-CNN.

Table 6. Ablation experiment results.

Method	mAP	Gain	mRecall	Gain
Base	0.64	—	0.70	—
FPN	0.67	4.62%	0.74	5.71%
RPN	0.65	1.56%	0.71	1.43%
Enhanced	0.66	3.13%	0.71	1.43%
MSFDC-Net	0.69	7.81%	0.76	8.57%

As can be seen from the experimental data in Table 6, the mAP and mRecall indicators of experimental groups (1), (2) and (3) are all superior to those of the baseline network. This indicates that all the proposed improvements can effectively enhance the model's ability to extract target features and optimize the performance of object recognition and localization.

Among them, the FPN experimental group provides multi-scale features for the detection head and effectively alleviates pixel loss caused by repeated downsampling. It equips deep features with edge information of mid-level features and enhances mid-level features with the semantic discriminative ability of deep features. It achieves the most prominent improvement in quantitative metrics, with the mAP reaching 0.67, an increase of 4.62% over the baseline model, and the mRecall reaching 0.74, a relative improvement of 5.71%.

The RPN module expands the convolutional receptive field and enriches the contextual information of feature space. Meanwhile, the coordinate attention mechanism is adopted to weight and strengthen the feature maps, highlight key regions containing object information, and improve the accuracy of initial anchor box generation. The improved quality of anchor candidates reduces invalid screening of subsequent proposals and lowers target missing rates at the anchor generation stage.

The Enhanced module corrects the color deviation of underwater images, effectively alleviating missed detection caused by the fusion of targets and complex background. It also enhances detailed features such as image edges and textures, which facilitates object recognition and further improves the detection accuracy of the model.

5. Discussion

This work integrates the CASAB attention module into RBE-CycleGAN to strengthen the network's ability to focus on critical regions, thereby achieving targeted image enhancement. The performance gain of the improved model mainly stems from the embedded CASAB attention design. This module first adopts global average pooling on feature maps along the channel dimension to capture the average activation level of each channel, and leverages global max pooling to extract salient signals within individual channels. The two pooling outputs are further combined to generate adaptive weight coefficients for all channels. As for spatial attention branch, it conducts max pooling, min pooling, mean pooling and sum pooling across all channels at each pixel position to calculate the corresponding pixel weight. The joint action of channel and spatial attention enables the network to accurately locate dominant feature channels and key object pixel regions. Consequently, the modified model delivers more focused and visually superior results in underwater image enhancement tasks.

This paper proposes MSFDC-Net as a novel network for underwater small object detection. The overall performance improvement of the model mainly relies on the specially designed multi-scale feature fusion module. Underwater small objects inherently occupy an extremely low proportion of image pixels. Repeated downsampling operations in conventional networks tend to severely discard critical feature details of tiny targets, which inevitably degrades detection accuracy. To address this issue, this work adopts middle-level features that retain edge contours and texture information of small targets for cross-layer feature fusion and optimization. The integrated and complete feature representations are then fed into the detection head for final inference, effectively compensating for the insufficient feature characterization of underwater small objects. Ablation experimental results fully demonstrate that the proposed multi-scale feature fusion strategy can markedly boost the detection capability of the network. Both mAP and mRecall, the two core evaluation metrics, achieve considerable improvements after adopting this optimization scheme.

From the perspective of feature utilization, this paper optimizes the feature utilization mode of both the image enhancement network and object detection network, and achieves certain improvement in overall performance. Nevertheless, the present study still has several limitations. In the training of the image enhancement network, the adopted EUVP dataset can effectively mitigate color distortion and improve the visual quality of underwater images. Even so, the scene coverage of EUVP is still limited for object detection tasks, leading to unsatisfactory generalization capability, especially under extreme underwater environmental conditions. Moreover, although the backbone of the object detection network adopts a multi-scale feature fusion strategy, it does not make targeted improvements to the convolution operation itself. The network is still constrained by the insufficient receptive field size, which restricts further performance breakthroughs.

In follow-up research, we will explore datasets covering a wider range of underwater scenarios to further enhance the model's environmental adaptability and generalization performance. Meanwhile, we will investigate feasible improvement schemes to expand the receptive field of convolution operations. This can enrich the feature information contained in each network layer and further strengthen the model's recognition capability for underwater targets.

6. Conclusions

Aiming at the common problems of color deviation and blurriness in underwater small object detection tasks, RBE-CycleGAN is adopted to effectively alleviate color distortion, improve image clarity and the quality of edge and texture features. It lays a solid foundation for downstream object detection, and reduces missed detection and background interference caused by color distortion.

To address the challenges of tiny object pixels and complex background interference, based on Faster R-CNN, this paper improves the MobileNetV2 backbone. A multi-scale feature fusion strategy is created to fuse middle-level features containing edge and texture information of small targets, supplement object feature representation, and reduce the loss of small object information caused by downsampling.

Standard convolution has a limited receptive field in the region proposal network head. To solve this problem, we adopt the coordinate attention mechanism. It can assign adaptive weights to feature maps and restrain interference from background noise. This method also builds long-range spatial dependency for object information. Apart from that, the parallel dilated convolution strategy can further expand the spatial receptive field. It enriches spatial context information of targets, and boosts the model's capability to recognize objects.

The original MobileNetV2 model has 82.37 million trainable parameters, with a weight file size of 628.64 MB and a peak GPU memory usage of 0.444 GB for single-image inference. In contrast, the improved MSFDC-Net achieves 20.21 million trainable parameters, a weight file size of 154.47 MB, and a peak GPU memory footprint of 0.213 GB during single-image inference. By optimizing the output channels of the backbone network, MSFDC-Net reduces the original 1280 output channels to 256 after feature fusion, which drastically cuts down the number of parameters in the RPN and ROI Head modules.

Experimental results show that RBE-CycleGAN is able to optimize underwater image quality. It can also well correct the existing color deviation. Ablation experimental findings further prove that image enhancement contributes positively to object detection performance. The designed MSFDC-Net reduces object information loss caused by continuous downsampling. It also enlarges the receptive field of standard convolution. The coordinate attention mechanism builds long-range spatial dependencies among points. It can supplement spatial location information for target regions, and help distinguish

foreground objects from background noise. Such design finally boosts the overall detection accuracy of the constructed model. Quantitative experimental results demonstrate that our RBE-CycleGAN improves UIQM, UCIQE and NIQE by 10.06%, 9.43% and 12.29% on the URPC dataset, with respective gains of 12.44%, 9.62% and 15.69% on the DUO dataset. Meanwhile, the proposed MSFDC-Net achieves mAP and mRecall increments of 7.81% and 8.57% on URPC, as well as 3.08% and 5.56% on DUO.

Although the above improvements effectively promote the detection performance, the proposed method still has certain limitations. The image enhancement network is trained based on unpaired datasets. In addition, complex underwater environment leads to diversified image degradation varying with water depth and water area. Since our model is only trained on the EUVP dataset under limited underwater degradation scenarios, the generalization of the enhancement module is restricted. In future work, we will construct richer training datasets tailored to underwater detection tasks to further improve the generalization capability of the overall model.

Author Contributions: Conceptualization, Z.L. and C.X.; methodology, Z.L. and C.X.; validation, Z.L. and C.X.; resources X.K.; writing Z.L. and C.X.; visualization, W.H. and R.C.; Supervision, X.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (grant number: 62571245, 62201260), the Fundamental Research Funds for the Central Universities (grant number: 30925020226, 30924010941), and the Project of Jiangsu Province Independent Scientific Research Fund (grant number:2025-JSS-LB-034-14).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the first author.

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

Abbreviations

The following abbreviations are used in this manuscript:

CASAB	Channel Attention and Spatial Attention Block
URPC	Underwater Robot Professional Contest
UIQM	Underwater Image Quality Measure
UCIQE	Underwater Colour Image Quality Evaluation
NIQE	Natural Image Quality Evaluator
mAP	Mean Average Precision
mRecall	Mean Recall
RBE-CycleGAN	ResNet Block Enhanced-CycleGAN
MSFDC-Net	Multi-Scale Feature Dilated Convolution Network
CA	Coordinate Attention
RPN	Region Proposal Network
CycleGAN	Cycle-Consistent Generative Adversarial Network
CBAM	Convolutional Block Attention Module
EUVP	Enhancement of Underwater Visual Perception

References

1. Fayaz, S.; Parah, S.A.; Qureshi, G. Underwater object detection: Architectures and algorithms—A comprehensive review. *Multimed. Tools Appl.* **2022**, *81*, 20871–20916. [[CrossRef](#)]
2. Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; Liotta, A. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* **2023**, *527*, 204–232. [[CrossRef](#)]

3. Berman, D.; Levy, D.; Avidan, S.; Treibitz, T. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2822–2837. [[PubMed](#)]
4. Er, M.J.; Chen, J.; Zhang, Y. Marine robotics 4.0: Present and future of real-time detection techniques for underwater objects. In *Industry 4.0—Perspectives and Applications*; IntechOpen: London, UK, 2022.
5. Yang, S.D.; Chen, Z.H.; Feng, Z.P.; Ma, X.M. Underwater Image Enhancement Using Scene Depth-Based Adaptive Background Light Estimation and Dark Channel Prior Algorithms. *IEEE Access* **2019**, *7*, 165318–165327. [[CrossRef](#)]
6. Hao, Y.S.; Yuan, Y.Y.; Zhang, H.M.; Zhang, Z. Underwater optical imaging: Methods, applications and perspectives. *Remote Sens.* **2024**, *16*, 3773. [[CrossRef](#)]
7. Zhang, W.; Dong, L.; Pan, X.; Zou, P.; Qin, L.; Xu, W. A survey of restoration and enhancement for underwater images. *IEEE Access* **2019**, *7*, 182259–182279. [[CrossRef](#)]
8. Fu, X.; Cao, X. Underwater image enhancement with global-local networks and compressed-histogram equalization. *Signal Process. Image Commun.* **2020**, *86*, 115892. [[CrossRef](#)]
9. Priyadharsini, R.; Sharmila, T.S.; Rajendran, V. A wavelet transform based contrast enhancement method for underwater acoustic images. *Multidimens. Syst. Signal Process.* **2018**, *29*, 1845–1859. [[CrossRef](#)]
10. Wang, G.; Chen, C.; Xu, H.; Ru, J.; Wang, S.; Wang, Z.; Liu, Z. RUE-Net: Advancing Underwater Vision with Live Image Enhancement. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642814. [[CrossRef](#)]
11. Ji, X.; Leng, N.; Guo, H. Progress and prospect of underwater image enhancement and restoration technologies. *J. Comput.-Aided Des. Comput. Graph.* **2024**, *36*, 805–830.
12. Wang, Y.; Zhang, J.F.; Wang, L.K.; Fan, X.H. Underwater image enhancement based on attention mechanism and feature reconstruction. *Infrared Technol.* **2024**, *46*, 1006–1014.
13. Cheng, C.S.; Wang, C.; Yang, D.Y.; Wen, X.; Liu, W.; Zhang, F. Underwater small target detection based on dynamic convolution and attention mechanism. *Front. Mar. Sci.* **2024**, *11*, 1348883. [[CrossRef](#)]
14. Sun, C.; Ai, Y.; Wang, S.; Zhang, W. Mask-guided SSD for small-object detection. *Appl. Intell.* **2021**, *51*, 3311–3322.
15. Zhuang, Y.; Liu, J.; Zhao, H.; Ma, L.; Fang, Z.; Li, L.; Wu, C.; Cui, W.; Liu, Z. A deep learning framework based on structured space model for detecting small objects in complex underwater environments. *Commun. Eng.* **2025**, *4*, 24. [[CrossRef](#)] [[PubMed](#)]
16. Hou, Q.B.; Zhou, D.Q.; Feng, J.S. Coordinate attention for efficient mobile network design. *arXiv* **2021**, arXiv:2103.02907.
17. Zhang, F.; Huang, D.; Ding, L.; Wang, Y.; Song, L. Contrastive Learning-Based Underwater Image Enhancement and Detection. *Robot* **2025**, *47*, 64–75.
18. Huang, D.M.; Wang, Y.; Song, W.; Sequeira, J.; Mavromatis, S. Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition. In *Proceedings of the 24th International Conference on MultiMedia Modeling*; Springer: Heidelberg, Germany, 2018; pp. 453–465.
19. Drews, P., Jr.; Do Nascimento, E.; Moraes, F.; Botelho, S.; Campos, M. Transmission estimation in underwater single images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*; IEEE Computer Society Press: Los Alamitos, CA, USA, 2013; pp. 825–830.
20. Chiang, J.Y.; Chen, Y.C. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* **2012**, *21*, 1756–1769. [[CrossRef](#)] [[PubMed](#)]
21. Li, C.Y.; Guo, J.C.; Pang, Y.W.; Chen, S.; Wang, J. Single underwater image restoration by blue-green channels dehazing and red channel correction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE Computer Society Press: Los Alamitos, CA, USA, 2016; pp. 1731–1735.
22. Li, C.Y.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [[CrossRef](#)]
23. Wang, Y.D.; Guo, J.C.; Gao, H.; Yue, H. UIEC²-Net: CNN-based underwater image enhancement using two color space. *Signal Process. Image Commun.* **2021**, *96*, 116250. [[CrossRef](#)]
24. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv* **2017**, arXiv:1703.10593.
25. 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.* **2018**, *3*, 387–394.
26. Liu, X.D.; Gao, Z.; Chen, B.M. MLFcGAN: Multilevel feature fusion-based conditional GAN for underwater image color correction. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1488–1492. [[CrossRef](#)]
27. 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. [[CrossRef](#)]
28. Li, J.; Liu, C.; Lu, X.; Wu, B. CME-YOLOv5: An efficient object detection network for densely spaced fish and small targets. *Water* **2022**, *14*, 2412. [[CrossRef](#)]
29. Wang, J.; Qi, S.; Wang, C.; Luo, J.; Wen, X.; Cao, R. B-YOLOX-S: A lightweight method for underwater object detection based on data augmentation and multiscale feature fusion. *J. Mar. Sci. Eng.* **2022**, *10*, 1764. [[CrossRef](#)]

30. Hua, X.; Cui, X.; Xu, X.; Qiu, S.; Liang, Y.; Bao, X.; Li, Z. Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy. *Pattern Recogn.* **2023**, *139*, 109511. [[CrossRef](#)]
31. Yang, X.T.; Liu, C.Z.; Han, J.Y. Reparameterized underwater object detection network improved by cone rod cell module and WIOU loss. *Complex Intell. Syst.* **2024**, *10*, 7183–7198. [[CrossRef](#)]
32. Ouyang, W.; Wei, Y.; Liu, G. A lightweight object detector with deformable upsampling for marine organism detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5018909. [[CrossRef](#)]
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*; Springer: Munich, Germany, 2018; pp. 3–19.
34. Wazir, S.; Kim, D.Y. Rethinking Decoder Design: Improving Biomarker Segmentation Using Depth-to-Space Restoration and Residual Linear Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society Press: Nashville, TN, USA, 2025; pp. 30861–30871.
35. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
36. Islam, M.J.; Xia, Y.; Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [[CrossRef](#)]
37. Underwater Robot Professional Contest Committee. URPC Dataset: Underwater Robot Professional Contest Dataset for Object Detection. 2020. Available online: <https://www.kaggle.com/datasets/lywang777/urpc2020> (accessed on 6 May 2020).
38. Liu, C.; Li, J.; Li, X. A dataset and benchmark of underwater object detection for robot picking. In *Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
39. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2961–2969.
40. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2016; pp. 21–37.
41. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.M.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2980–2988.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.