


Article

# G-Net: An Efficient Convolutional Network for Underwater Object Detection

Xiaoyang Zhao <sup>1</sup>, Zhuo Wang <sup>2</sup>, Zhongchao Deng <sup>1,\*</sup> and Hongde Qin <sup>2</sup>

<sup>1</sup> Science and Technology on Underwater Vehicle Technology Laboratory, Harbin Engineering University, Harbin 150001, China; zhaoxiaoyang@hrbeu.edu.cn

<sup>2</sup> Qingdao Innovation and Development Center, Harbin Engineering University, Qindao 266000, China; wangzhuo@hrbeu.edu.cn (Z.W.); qinhongde@hrbeu.edu.cn (H.Q.)

\* Correspondence: dengzhongchao@hrbeu.edu.cn

**Abstract:** Visual perception technology is of great significance for underwater robots to carry out seabed investigation and mariculture activities. Due to the complex underwater environment, it is often necessary to enhance the underwater image when detecting underwater targets by optical sensors. Most of the traditional methods involve image enhancement and then target detection. However, this method greatly increases the timeliness in practical application. To solve this problem, we propose a feature-enhanced target detection network, Global-Net (G-Net), which combines underwater image enhancement with target detection. Different from the traditional method of reconstructing enhanced images for target detection, G-Net realizes the integration of image enhancement and target detection. In addition, our feature map learning module (FML) can effectively extract defogging features. The test results in a real underwater environment show that G-Net improves the detection accuracy of underwater targets by about 5%, but also has high detection efficiency, which ensures the reliability of underwater robots in seabed investigation and aquaculture activities.

**Keywords:** image enhancement; image reconstruction; underwater object detection; feature enhancement



**Citation:** Zhao, X.; Wang, Z.; Deng, Z.; Qin, H. G-Net: An Efficient Convolutional Network for Underwater Object Detection. *J. Mar. Sci. Eng.* **2024**, *12*, 116. <https://doi.org/10.3390/jmse12010116>

Academic Editor: Sergei Chernyi

Received: 3 December 2023

Revised: 3 January 2024

Accepted: 5 January 2024

Published: 7 January 2024



**Copyright:** © 2024 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) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In recent years, research on underwater intelligent sensing technology has become increasingly popular due to the rapid development of modern technologies, such as the internet and artificial intelligence. Underwater unmanned vehicles perform underwater information detection, underwater aquaculture, and underwater rescue tasks with the assistance of intelligent perception technology. These tasks depend heavily on the quality of captured images [1]. However, due to the complex underwater environment, underwater images obtained from optical imaging devices cannot be directly used for target identification. These underwater images are blurred and color distorted due to scattering and absorption by water, which seriously affects the performance of underwater target detection. Some research works have been presented in the literature for underwater image enhancement [1–8]. More specifically, underwater image enhancement can be seen as an image defogging problem, and a defogging method with minimum information loss and histogram distribution prior is proposed in [2]. To better represent underwater images, an improved underwater image model was proposed in [6]. The model uses ocean monitoring data to derive the physical effective space of backscattering, which will be conducive to restoring image enhancement of complex scenes. Both [7,8] employed depth images to enhance underwater image. A lightweight adaptive feature fusion network is proposed in [5] for underwater image enhancement. The simplified U-Net was used to enhance the paired underwater images in [3]. An algorithm based on color feature fusion is proposed in [4], which adopts multi-channel feature extraction strategy to achieve underwater image enhancement.

With the widespread use of convolutional neural networks, the research on target detection techniques has been greatly improved, such as a series of faster one-stage target detection algorithms Single Shot MultiBox Detector (SSD) [9], You Only Look Once (YOLO) [10], RetinaNet [11], Fully Convolutional One-Stage Object Detector (FCOS) [12], and more accurate two-stage target detection algorithms Regions with CNN features (RCNN) [13], Fast RCNN [14], Faster RCNN [15], and Cascade RCNN [16]. These detection methods extract features through the backbone network and then classify extracted features and perform bounding box regression through the head network. However, underwater target detection remains a challenging task. These methods are only applicable to clear datasets and have difficulty detecting underwater degraded images. Moreover, underwater image enhancement methods restore degraded images to clear images and perform target detection on clear images. The traditional method firstly enhance the underwater degradation image and then detect the objects. However, the difficulty of underwater target detection is the degradation of image features extracted by backbone network, which makes it difficult for the detector to classify the target and perform boundary regression. Ref. [17] improved Fast RCNN and adversarial occlusion network for underwater target detection. Ref. [18] improved RCNN for underwater target detection. Ref. [19] improved one-stage detector yolov5 for underwater target detection. Ref. [20] improved yolox and tested it on underwater data sets. Also, image enhancement networks aim to enhance image sharpness and contrast and color correction, which may not necessarily be applicable to target detection networks. Furthermore, the efficiency of reconstructing clear images and performing target detection is inefficient, and it is difficult to reconstruct clear images and perform target detection in real time during underwater operations. To solve these two issues, the underwater target detection task is divided into two subtasks in G-Net, i.e., underwater image feature enhancement and enhanced target detection. Previous studies typically separate the two without considering the impact of enhanced images on subsequent higher-order vision tasks, but underwater enhanced image reconstruction seriously affects the operational efficiency of underwater target detection. Therefore, we construct an end-to-end underwater target detection method that does not require the reconstruction of enhanced images.

The main contributions of this work are summarized as follows:

- We propose a data-driven approach based on an end-to-end target detection model. It fuses image enhancement with target detection in the feature extraction part and enables accurate detection of target information of underwater blurred images based on enhanced features, which can be generalized to real-world underwater scenarios.
- In order to simplify the process of image enhancement, feature extraction, object classification and boundary regression in traditional methods, image enhancement and feature extraction for object detection are combined to realize the integration of image enhancement and object detection. By inputting hazy images, our G-Net network can enhance the detection features of degraded images and directly output the target detection results. The perception efficiency of the underwater unmanned boat is greatly improved.
- In order to better capture the degradation features, we propose a feature map learning (FML) module, using clear image features to guide the network to enhance the detection features of degraded images.
- Our proposed lightweight neural network not only exhibits good performance on the dataset but also performs well on real underwater environment images. It has significant performance improvement compared with other underwater object detection algorithms.

The remainder of this paper is organized as follows. Section 2 briefly describes the current research status on underwater target detection. Section 3 presents the network structure and loss function of the proposed novel lightweight neural network and the FML learning module combining feature enhancement and target detection. Section 4 compares

and analyzes the experiments of G-Net and existing methods. The experimental results are summarized in Section 5.

## 2. Related Works

In this section, we will briefly introduce the development of deep learning-based target detection and the related content of underwater target detection.

### 2.1. Object Detection Based on Deep Learning

Target detection, as one of the most fundamental and important tasks in computer vision, has received widespread attention in recent years [21]. Traditional manual feature-based detection frameworks, including the Viola–Jones detector [22], histogram of gradients (HoG) [23], and deformable models [24], have been gradually replaced by neural networks.

Deep learning-based methods have achieved success in perception tasks [25–28]. With the development of deep learning, deep learning-based target detection can be divided into two categories: one-stage target detection represented by the YOLO series [29–32] and two-stage target detection represented by Faster RCNN. The two-stage detection network first forms candidate regions and then detects targets in the candidate regions. The advantage of two-stage target detection is its higher accuracy, and the advantage of one-stage target detection is its faster speed. However, the accuracy of the one-stage detector gradually becomes higher than that of the two-stage detector with development, which indicates its wider industrial application.

### 2.2. Underwater Object Detection Based on Deep Learning

Deep learning techniques have also developed rapidly in underwater target detection. For example, the R-CNN-based fish detection method proposed in [33], the classification of underwater plankton using deep residual networks [34], and the lightweight neural network for underwater fish detection [35]. Li et al. [36] improved the YOLOv4 target detection network by gradually using MobileNetv2 [37] and proposed a multiscale attentional feature fusion mechanism to improve the detection accuracy of underwater small targets. Yeh [38] et al. used a color conversion module and detection module jointly trained to enhance underwater target detection through the joint training of color conversion and detection modules. The underwater target single aggregation network was proposed in [39] by using multiscale features and complementary context information [40] proposed a fast underwater target detection network [41] fine-tuned YOLOv2 and tested it on the underwater dataset.

## 3. Proposed Methods

The traditional underwater image enhancement and target detection process is inefficient, and the detection performance is insufficient, which can not meet the needs of underwater target detection. Therefore, we propose a framework that integrates image enhancement and target detection. The enhanced network no longer needs to rebuild a clear image before detection, as shown in Figure 1. This will greatly improve the efficiency of underwater target detection. In this section, we will focus on G-Net.

### 3.1. Model Workflow

The flow chart of the proposed G-Net network structure is shown in Figure 1. G-Net consists of an Object Detection Feature Map Enhancement (ODFME) module for image feature enhancement and a Detect Target and Output Result (DTR) module for classification and regression, as shown in Figure 2. In the proposed FML learning module, the ODFME module learns feature maps of clear images without reconstructing the enhanced images, which enables the integration of underwater image enhancement and detection. The ODFME module outputs the enhanced detection features, and the DTR module inputs enhanced features for the regression of bounding boxes and target classifications. Compared

with the traditional independent tasks of underwater image enhancement and target detection, it can be better applied to underwater target detection.

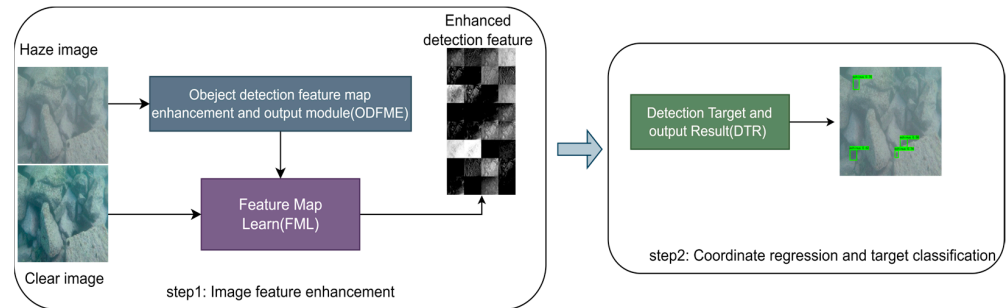


Figure 1. G-Net framework flow.

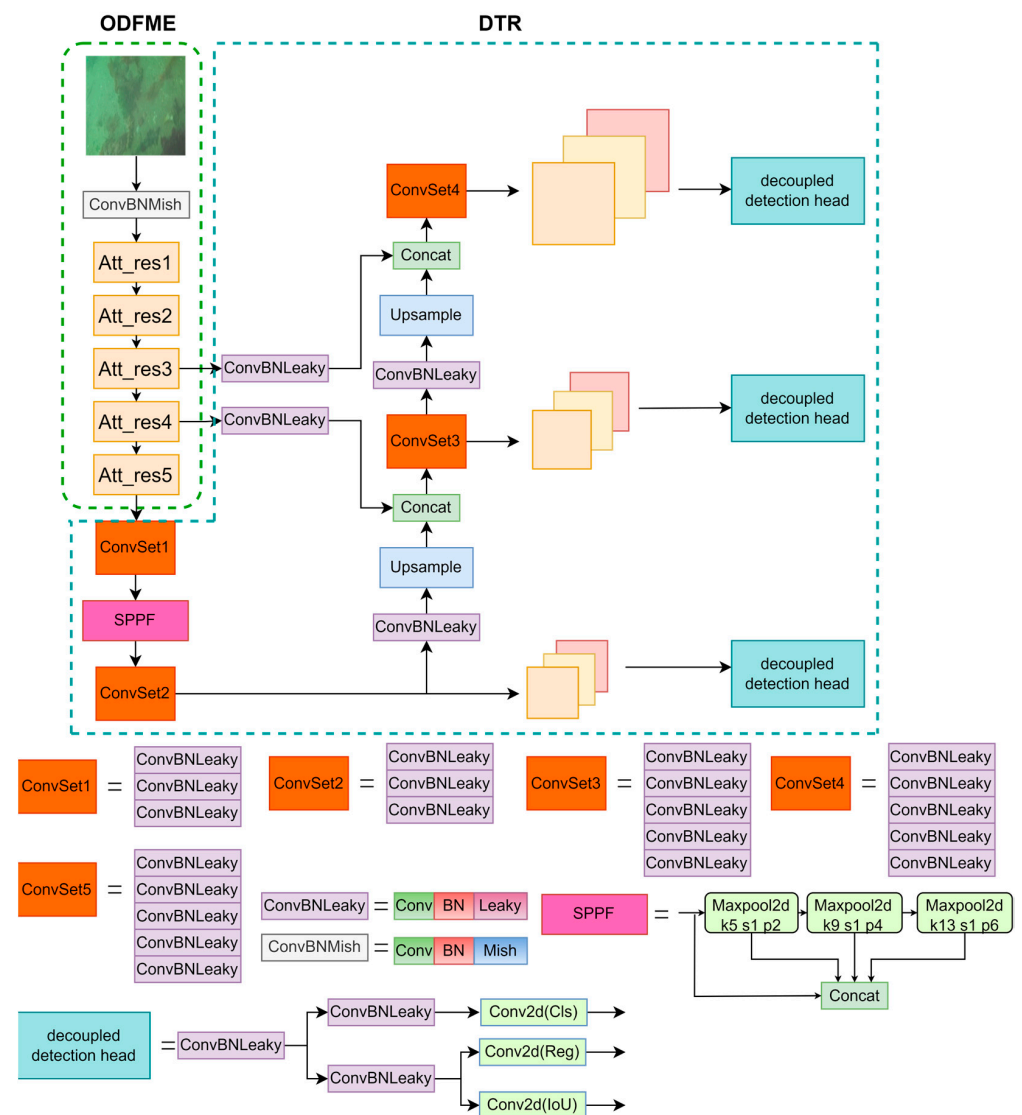
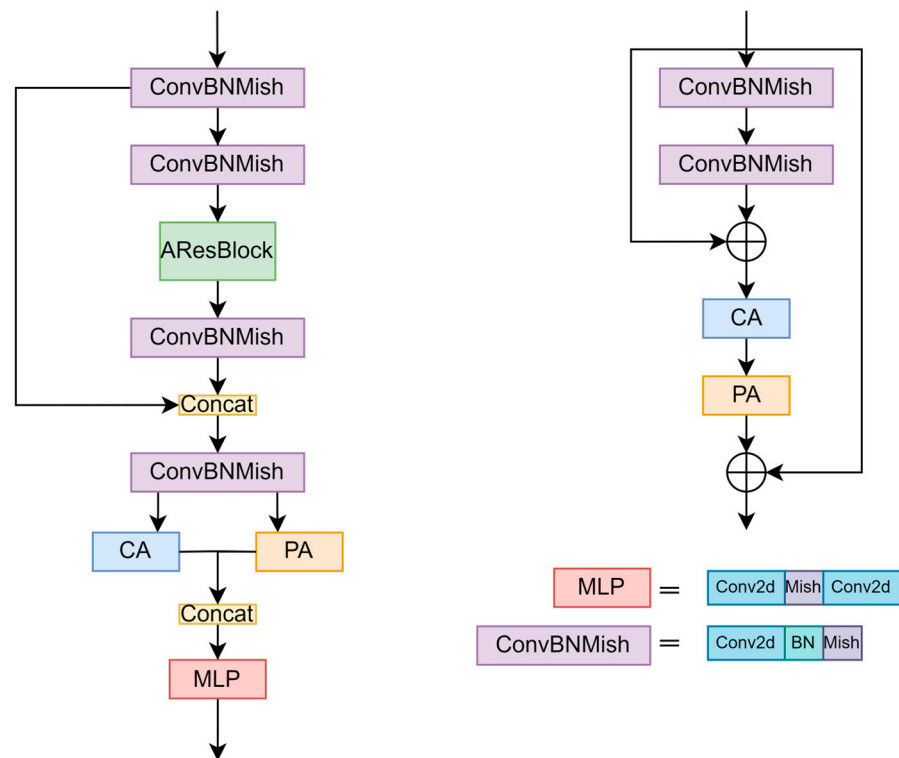


Figure 2. The image on the left shows the ODFME module and enhanced features. The ODFME module consists of five Att\_res blocks. On the right is the DTR structure for coordinate regression and target classification. Att\_res indicates attention residual block, Conv and Conv2d indicate convolution, BN indicates batch normalization, and Leaky and Mish indicate activation function. SPPF means spatial pyramid pooling-faster.

### 3.2. ODFME Module

The G-Net network structure consists of an Object detection feature map enhancement and output (ODFME) module and a detect target and output (DTR) module, as shown in Figure 2. ConvBNLeaky consists of Conv2d, Batch normalization, Leaky. ConvBNMish consists of Conv2d, BN, Mish. ConvSet consists of ConvBNLeaky. SPPF consists of parallel Maxpool2d. Where Conv2d denotes the convolution, BN denotes Batch Normalization, Mish denotes Mish activation function. Leaky denotes Leaky activation function. Since our feature enhancement module does not need to reconstruct a clear underwater image but uses the enhanced feature map directly for coordinate regression and classification, our network design focuses on the ODFME part.

Our ODFME module is designed based on the CSP-DarkNet model. The main structure of the network is shown in Figure 2. Our ODFME module is mainly composed of five Attention Resual Blocks (Att\_res Block). The structure of Att\_res block is shown in Figure 3 (five Att\_res blocks only differ in the number of attention resistance block (Aresblock), and the rest are exactly the same). Att\_res module adds an attention resistance block (AresBlock) among the three ConvBNMish. It adopts serial channel attention and spatial attention, which can extract and enhance degradation features. AResBlock Concat operation can fuse shallow information with deep information. Then, parallel channel attention (CA) and spatial attention (PA) modules encode the transmission value and global atmospheric light value of the fused image, respectively. The multi-layer perceptron (MLP) module is used to re-fit the enhanced dehazing features.



**Figure 3.** The network of the Att\_res block and Aresblock. On the left is the Att\_res block network structure. The feature map is enhanced by series-parallel channel attention and spatial attention. The image on the right shows the structure of AResBlock, which is enhanced with local residual and tandem channel attention and spatial attention. CA means channel attention, PA means spatial attention, MLP means multi-layer perceptron. AresBlock means attention resistance block. ⊕ means point-wise addition.

According to the image degradation model widely used in [42–45], the underwater image degradation formula can be expressed as

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

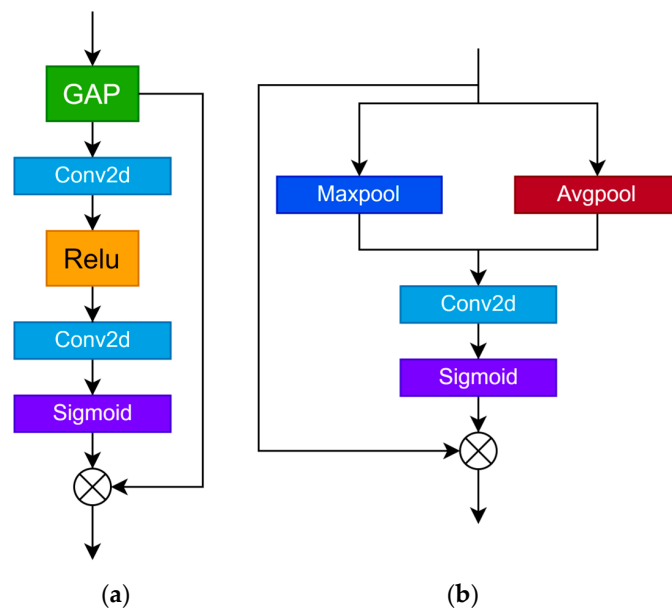
where  $I(x)$  represents underwater degradation image,  $J(x)$  white represents clear image,  $t(x)$  represents transmission value and  $A$  represents global atmospheric light value.

We believe that transmission value  $t(x)$  in the underwater image degradation Formula (1) is a location-related local variable, and the global atmospheric light value  $A$  is a global shared variable. Meanwhile, most image enhancement works treat channel and pixel features equally and cannot process images with uneven distributions of underwater degradation and weighted channels. Inspired by FFA-Net, AResBlock adopts a local residual structure and a combination of a serial channel attention module and spatial attention module, which can address different features and pixel regions unequally and provide more flexibility when processing different types of information. We believe that the channel attention mechanism can better obtain the global shared variable  $A$ , and the spatial attention mechanism can better obtain the local variable  $t$  related to position. Therefore, in addition to the serial attention module used in AResBlock, we use parallel channel attention and spatial attention at the end of the Att\_res structure to enhance different channel and pixel features, respectively, and the MLP module to adjust the number of channels and combine different enhancement features. We believe that the parallel use of the channel attention module and spatial attention module can prevent features from being overenhanced and can better extract the global shared variable  $A$  and the location-related local variable  $t(x)$  from the original features.

The channel attention structure is shown in Figure 4. Let  $X$  be the feature map. According to the gating mechanism in [46], our channel attention can be expressed as

$$\begin{aligned} X_w &= \text{Sigmoid}(\text{Conv2d}(\text{Relu}(\text{Conv2d}(\text{GAP}(X)))))) \\ X_{out} &= X_w \otimes X \end{aligned} \tag{2}$$

where the GAP means Global average pooling,  $\text{Relu}$  means activation function, and  $\otimes$  refers to the Hadamard product.



**Figure 4.** Channel attention and spatial attention structure. (a) represents the channel attention module, and (b) represents the spatial attention module. GAP means global average pooling, Sigmoid means activation function.  $\otimes$  means Hadamard product. Maxpool means max pooling; Avgpool means average pooling.



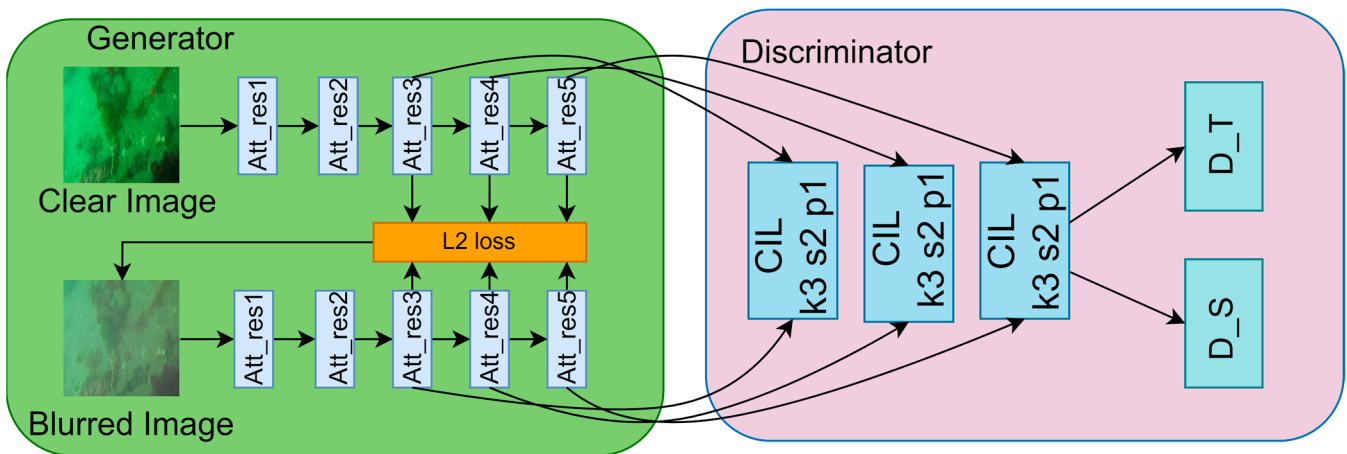
We use global averaging pooling, Conv2d, ReLU functions, and sigmoid functions to obtain its weight on the channel and then multiply the corresponding elements with  $x$ .

The structure of spatial attention is shown in Figure 5. The gating mechanism is used as the main content of the spatial attention mechanism, the left branch is used as the feature branch, and the right branch is used as the spatial pixel selection communication number. According to [47], if the input feature graph is  $x$ , the spatial attention mechanism can be expressed as

$$W = \text{Sigmoid}(\text{Conv2d}(\text{Avgpool}(X) \oplus \text{maxpool}(X)))$$

$$X_{out} = X \otimes W \tag{3}$$

where the  $\oplus$  means point-wise addition, *Avgpool* means average pooling, and *maxpool* means max pooling.



**Figure 5.** The FML module contains a generator and discriminator.

### 3.3. DTR Detection Module

The DTR structure is improved on the basis of YOLOv4. Multiple convolution blocks are used in DTR to detect feature information on features of different scales, and the up-sampled low-resolution features are combined with high-resolution features in the channel dimension to enrich the semantic and spatial information in high-resolution features. In the DTR structure, we use the SPPF module to replace the original SPP module, which can pool different feature maps without changing the size of the feature map, to capture the details of the target at different scales. It can also reduce the calculation amount of the model and improve the reasoning speed. The model structure of SPPF is shown in Figure 2. For the detection header of YOLOv4, the decoupling header can extract target location information and category information, which can effectively reduce the number of parameters and computational complexity and enhance the generalization ability and robustness of the model. The structure of the decoupling header is shown in Figure 2. In the G-Net proposed in this paper, since the ODFME module has completed the enhancement of degraded image detection features, DTR can directly perform accurate target classification and coordinate regression according to ODFME output features, which greatly improves the detection efficiency. Therefore, we did not make much improvement on the DTR part, and our improvement mainly lies in the enhanced feature extraction part of ODFME.

### 3.4. FML Module

The FML module proposed by us can be regarded as a new loss, and the FML learning module enables the ODFME module to better obtain the feature mapping of clear images. The concrete structure of the FML net is shown in Figure 5. As a whole, our learning module network is similar to gan. In the generator, we freeze the DTR part of G-Netnd and then let ODFME module learn the features of the corresponding clear image (the clear image is used as the input of G-Net and then trained to obtain the model and frozen). The

generator outputs three kinds of characteristic maps with different scales, and the enhanced network is trained by gradient back propagation of  $L_2$  loss. As for the discriminator, our discriminator adopts the design of patch gan, and the input of different scales is converted into the feature map with the size of  $4 \times 4$  by full convolution method, so that the discriminator can more accurately identify whether the feature map comes from a clear underwater image or a fuzzy underwater image. In the discriminator, we use the clear image output under ODFME module as the true value. Through iterative training, the feature mapping of underwater blurred image generated by ODFME enhancement network is closer to clear images. FML learning module only participates in the training phase of the model, but not in the reasoning phase of the model. Through the FML module, our ODFME module can obtain clearer detection features, which plays an important role in the following target classification and boundary regression.

### 3.5. Loss Function

In the training stage, we first train our G-Net network, then freeze the DTR part of the G-Net network and use the FML learning module to further train the ODFME module.

In the first phase, G-Net uses exactly the same loss functions as yolov4 (bounding box loss, confidence loss, and classification loss).

During FML module training, in the generator, we use an ODFME network with clear image as input to instruct the ODFME network with the haze image as input to learn image degradation features. According to [48], we use  $L_{2loss}$  as the generator's loss function.  $L_{2loss}$  can be formulated as:

$$L_{2loss}[X, Y] = \sum_{i=1, j=1}^{M, N} [X(i, j) - Y(i, j)]^2 \quad (4)$$

where  $X$  and  $Y$  correspond to the feature maps of paired clear and blurred underwater images, respectively;  $M$  and  $N$  correspond to the width and height of the feature maps, respectively; and  $i$  and  $j$  denote pixel points.

The loss function of the discriminator can be expressed intuitively as

$$L_{dis}[D_s, D_t] = ([D_s - 0] + [D_t - 1])/2 \quad (5)$$

where  $D_s$  and  $D_t$  indicate whether the discriminator identifies the input feature map as coming from a clear or degraded image.

## 4. Experiment

### 4.1. Datasets

It is difficult to obtain clear and degraded underwater images simultaneously. Meanwhile, to simulate the real environment as much as possible, we added haze to the underwater dataset based on the underwater imaging model based on Equation (6). To generate different concentrations of haze, the global atmospheric light value is set to (0.5, 0.95), and the transmission map is set to (0.3, 0.95). Our clear underwater dataset (UPRC) applies the data from real offshore environments and includes a total of 5542 images of four objects: holothurian, echinus, scallop, and starfish, with a total of 5542 images, as shown in Figure 6. Among them, 4987 images are used for training, and 555 images are used for testing. To reduce the memory consumption during training, our image size is set to (256, 256). Although G-Net functions include two parts, underwater image feature enhancement and object detection, underwater image enhancement (detection feature enhancement) is designed to improve the accuracy of object detection. Therefore, the experiment mainly verifies the target detection performance of G-Net.





**Figure 6.** Samples of underwater objects (holothurian, echinus, and scallop, and starfish) mainly considered to be detected in this article.

#### 4.2. Evaluation Metric

To evaluate the target detection performance, the average accuracy mAP is applied to measure the detection performance of the network, and the IOU is set to 0.5. The mAP equation is expressed as

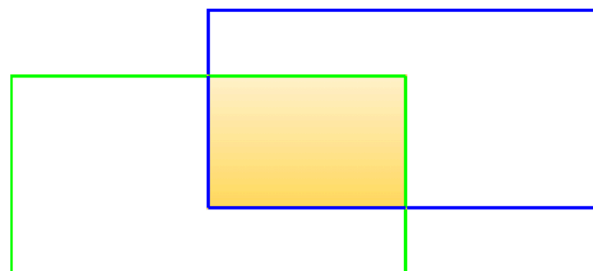
$$mAP = \frac{AP}{num\_classes} \tag{6}$$

where AP is the detection accuracy of each class of targets, and num\_classes is the number of classes.

The IOU can be denoted by:

$$IOU = \frac{B_d \cap B_{gt}}{B_d \cup B_{gt}} \tag{7}$$

As shown in Figure 7, the green box indicates  $B_d$ , and the blue box indicates  $B_{gt}$ .  $B_d$  represents the detected bounding box, and  $B_{gt}$  represents the real bounding box. If the IOU of  $B_d$  and  $B_{gt}$  is greater than 0.5 and belongs to the same category, the test result is a true positive sample. Otherwise, it is a false positive sample. The intersection over union (IOU) is shown in Figure 7.



**Figure 7.** Intersection over union (IOU). The yellow part indicates the intersection.

The *Precision* can be expressed as

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

where *TP* indicates a correct prediction as a true value and *FP* indicates an incorrect prediction as a true value. The details are shown in Table 1.

**Table 1.** *TP* denotes true positive samples, *FP* denotes false-positive samples, *FN* denotes false-negative samples, and *TN* denotes true negative samples.

	True Detect	False Detect
True ground	<i>TP</i>	<i>FN</i>
False ground	<i>FP</i>	<i>TN</i>

#### 4.3. Implementation Details

All experiments are built on the PyTorch framework using a single NVIDIA GTX 4090 GPU and an Intel Core i9-13900 CPU. The resolution of the experimental image is  $256 \times 256$ . To better fit the training data and make full use of the GPU, we use a batch size of bit 32, a learning rate of 0.0001, and an Adam optimizer with a total of 300 epochs.

#### 4.4. Contrast Experiment

The underwater degraded image feature enhancement and target detection included in G-Net are aimed at improving the accuracy of underwater target detection. Therefore, our comparison experiments mainly verify the target detection performance of G-Net. Since our G-Net was improved based on the YOLOv4 network, we adopted YOLOv4 as the target detection network for comparative testing. In this section, we first employ YOLOv4 to test different underwater enhancement methods to explore the effectiveness of our proposed method. Next, ablation experiments are conducted to investigate the effects of different components of our method using qualitative and quantitative approaches.

Underwater degraded images severely affect visibility. To verify the effectiveness of G-Net in real underwater environments, we compare the detection results of different advanced underwater degradation image enhancement methods through YOLOv4. UDCP, Ugan, CycleGan and PUIE were used as image enhancement networks, and YOLOv4 was used as a detection network to perform comparative experiments with G-Net. Table 2 demonstrates that compared to using ordinary detectors to detect underwater degraded images with the lowest detection accuracy, our G-Net has the highest detection accuracy of approximately 75.46% for the overall mAP. It can visually demonstrate the importance of G-Net in underwater degraded images without reconstructing the enhanced images and directly detecting the enhanced features. Although the parameters of G-Net are slightly higher than those of CycleGan and PUIE, but the time only needs 7.1 ms, which is lower than others, and overall, our performance is the best.

To better display the detection results of G-Net, the detection results of different enhancement algorithms are shown in Figure 8. Figure 8a is the detection result of YOLOv4 on clear images, and Figure 8b–d denote the detection results after image enhancement using the U-Gan, CycleGan, and PUIE methods, respectively. Figure 8e is the detection result of G-Net. In the second column, three different image enhancement algorithms of Figure 8b–d have missed detection behavior in holothurian detection. When detecting scallops in the third column, Figure 8b,c were misdetections; when detecting starfish in the fourth column, Figure 8b–d were also missed. Only Figure 8e can successfully detect all results.

**Table 2.** The detection results of traditional methods on underwater degraded images and G-Net detection results. “Params” is the total number of parameters of underwater image enhancement methods and detection methods. “FLOPs” is the number of floating-point operations of different algorithms. “Time” is the time needed for enhancement and detection by different methods. The second rows “C1” to “C4” are holothurian, echinus, scallop, and starfish.

Image Enhance Method	Detection Method	mAP(%)	AP(%)				Params (M)	FLOPs (G)	Time (ms)
			C1	C2	C3	C4			
-	YOLOv4	70.92	86.67	52.86	67.19	76.96	63.98	22.72	4.5
UDCP	YOLOv4	71.62	86.74	53.54	64.30	81.89	63.98	22.72	12.0
UGan	YOLOv4	70.64	83.86	53.96	70.48	74.27	105.81	58.85	8.2
CycleGan	YOLOv4	72.49	87.29	54.43	70.29	77.49	78.12	113.75	7.7
PUIE	YOLOv4	74.02	87.94	58.38	71.48	78.27	65.38	82.69	7.3
G-Net		75.46	87.08	59.63	74.02	81.11	78.75	33.41	7.1

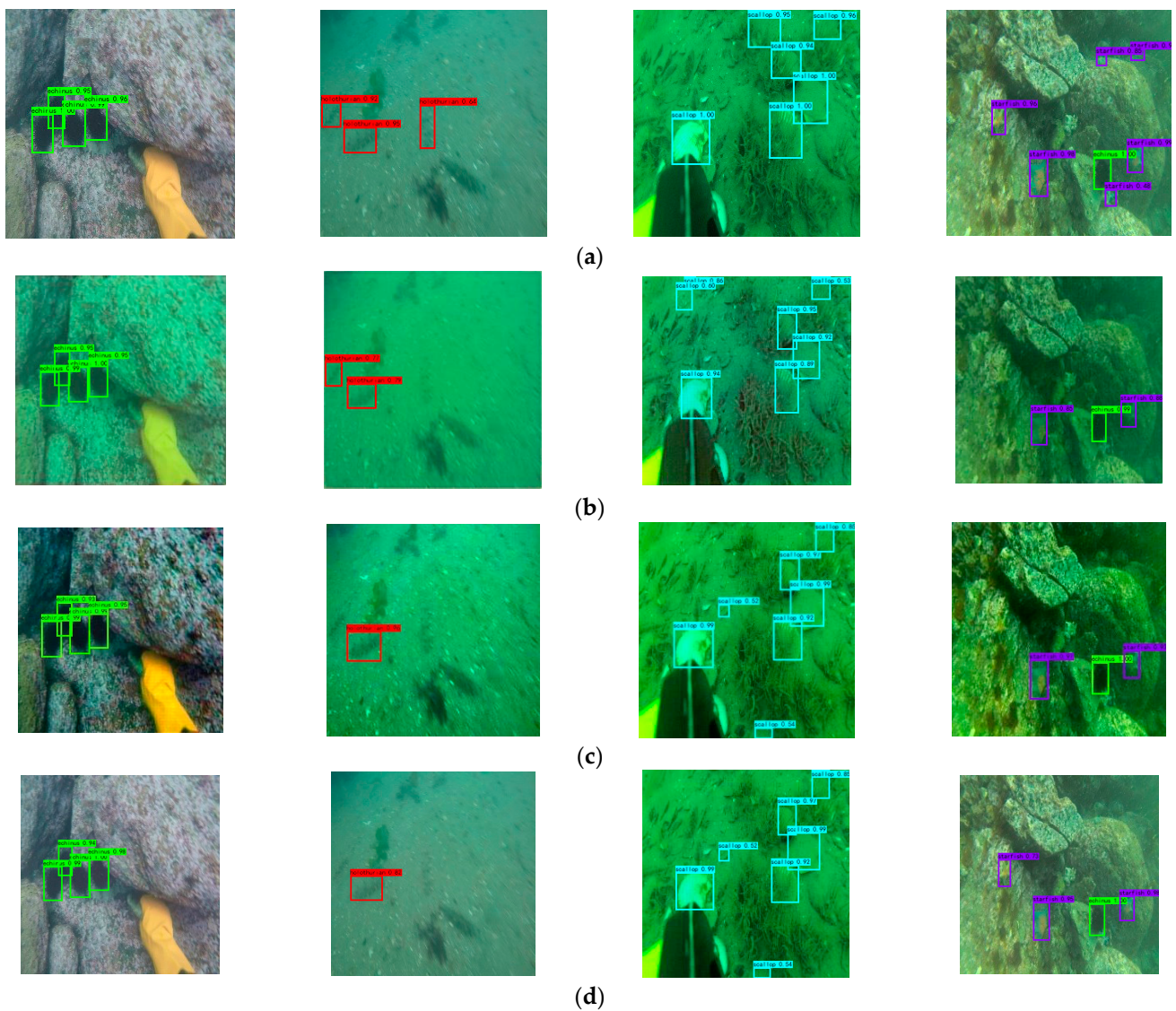
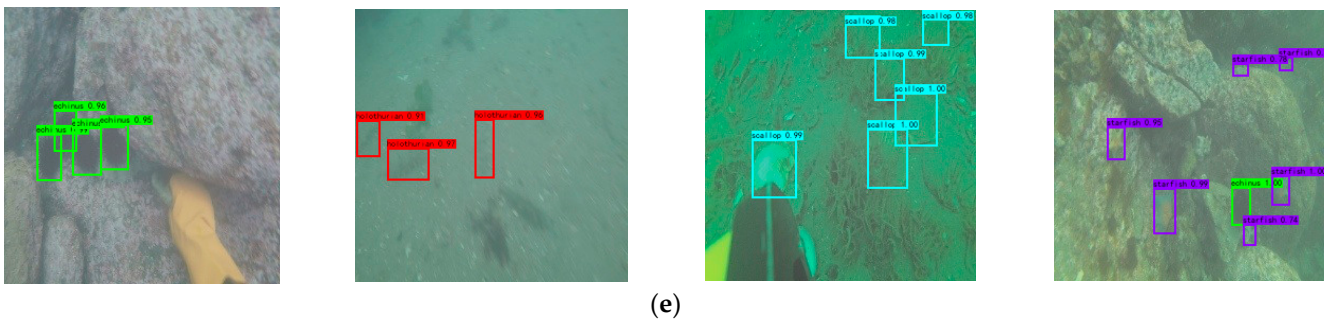


Figure 8. Cont.



**Figure 8.** Detection result. (b–d) denote the detection results after image enhancement using the U-Gan, CycleGan, and PUIE methods, respectively. (a) is the detection result of YOLOv4 on clear images. (e) is the detection result of G-Net.

The combination of traditional image enhancement and detection has resulted in missed and false detection. The detection results of the G-Net network are very close to those of the clear image, and there is no need to rebuild clear images, which further proves the effectiveness of the proposed G-Net network.

#### 4.5. Ablation Study

Qualitative and quantitative methods are used to validate the effectiveness of the FML learning module. We use the same dataset to compare G-Net with and without FML learning modules separately. The results in Table 3 demonstrate that G-Net with the FML learning module has the highest average detection accuracy. G-Net has a 75.46 mAP. Except for the detection accuracy of holothurian and scallop, which is slightly lower than that of G-Net without the FML module, the detection accuracy of echinus, scallop, and starfish is higher than that of G-Net without the FML module. Furthermore, the FML learning module does not participate in the forward inference of the model, so there is no increase in the number of parameters and the model time. When the backbone network does not use our attention structure, the number of parameters and the complexity of the model are reduced, the reasoning speed is relatively increased, but the detection accuracy is decreased. As shown in Table 3, the number of parameters, FLOPs, and runtime of our model are the same regardless of whether the FML module is used. The results show that the FML module can improve the detection accuracy of the G-Net network without reducing the speed. The attention structure of the backbone network also shows that we can improve our detection accuracy with a small increase in inference time.

**Table 3.** Experimental results of G-Net with or without the FML module. “Params” represents the number of parameters of the model. “FLOPs” indicates the complexity of the model. “C1” to “C4” and “Time” are the same as those in Table 2.

FML Module	Att_res Module	mAP(%)	AP(%)				Params (M)	FLOPs (G)	Time (ms)
			C1	C2	C3	C4			
✓		74.86	88.06	58.90	73.40	79.08	65.031	27.351	6.3
	✓	74.52	87.43	58.62	75.30	79.28	78.75	33.41	7.1
✓	✓	75.46	87.08	59.63	74.02	81.11	78.75	33.41	7.1

## 5. Conclusions

To reduce the interference of underwater degraded images on target detection performance and to improve the underwater detection efficiency, we propose an image feature enhancement detection network, G-Net. It simultaneously completes feature enhancement and detection tasks of underwater images in a single network, achieving the unification of underwater image enhancement and detection, and greatly improving underwater detec-



tion efficiency. Moreover, the FML unsupervised learning module uses an unsupervised approach to extract clear underwater image features, which guide the ODFME module for learning. It enables the ODFME module to learn clear underwater image features even in degraded underwater conditions. The experimental results demonstrate that our G-Net effectively reduces the interference of underwater degraded images and achieves high-accuracy target detection, which is crucial for tasks such as underwater grasping and surrounding obstacle detection.

In the follow-up study, due to the complexity and variety of underwater scenes, the visual interference caused by fog only being a part of it, as well as the effects of motion blur and light intensity, we should improve the FML module, so that it can learn different kinds of degradation features. Therefore, in the future, we will study the ODFME module and FML learning module with more general and better performance, so that they can better serve underwater target detection. At the same time, we will also try to combine some newer detectors with our G-Net to investigate more feasibility.

**Author Contributions:** Conceptualization, X.Z. and Z.D.; methodology, X.Z.; software, X.Z.; validation, X.Z. and Z.D.; formal analysis, X.Z.; investigation, X.Z.; resources, Z.D. and Z.W.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z.; visualization, X.Z.; supervision, Z.D., Z.W. and H.Q.; project administration, Z.D., Z.W. and H.Q.; funding acquisition, H.Q. 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 under Grant 52025111, China.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All the data that support the findings in this study can be obtained by contacting the corresponding author.

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

## References

- Zhang, S.; Zhao, S.; An, D.; Liu, J.; Wang, H.; Feng, Y.; Li, D.; Zhao, R. Visual SLAM for Underwater Vehicles: A Survey. *Comput. Sci. Rev.* **2022**, *46*, 100510. [[CrossRef](#)]
- Li, C.-Y.; Guo, J.-C.; Cong, R.-M.; Pang, Y.-W.; Wang, B. Underwater Image Enhancement by Dehazing With Minimum Information Loss and Histogram Distribution Prior. *IEEE Trans. Image Process.* **2016**, *25*, 5664–5677. [[CrossRef](#)]
- Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [[CrossRef](#)]
- Gong, T.; Zhang, M.; Zhou, Y.; Bai, H. Underwater Image Enhancement Based on Color Feature Fusion. *Electronics* **2023**, *12*, 4999. [[CrossRef](#)]
- Yang, H.-H.; Huang, K.-C.; Chen, W.-T. LAFFNet: A Lightweight Adaptive Feature Fusion Network for Underwater Image Enhancement. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May 2021; pp. 685–692.
- Akkaynak, D.; Treibitz, T. A Revised Underwater Image Formation Model. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6723–6732.
- Akkaynak, D.; Treibitz, T. Sea-Thru: A Method for Removing Water from Underwater Images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1682–1691.
- Ueda, T.; Yamada, K.; Tanaka, Y. Underwater Image Synthesis from RGB-D Images and Its Application to Deep Underwater Image Restoration. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2115–2119.
- 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 Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905, pp. 21–37.
- Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-Time Flying Object Detection with YOLOv8 2023. *arXiv* **2023**, arXiv:2305.09972.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.

12. Mieske, S.; Hilker, M.; Infante, L. Fornax Compact Object Survey FCOS: On the Nature of Ultra Compact Dwarf Galaxies. *Astron. Astrophys.* **2004**, *418*, 445–458. [[CrossRef](#)]
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
14. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. 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.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
16. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
17. 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. [[CrossRef](#)]
18. Lei, F.; Tang, F.; Li, S. Underwater Target Detection Algorithm Based on Improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [[CrossRef](#)]
19. 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. [[CrossRef](#)]
20. 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. [[CrossRef](#)]
21. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [[CrossRef](#)]
22. Viola, P.; Jones, M. Robust Real-Time Face Detection. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*; IEEE Computer Society: Vancouver, BC, Canada, 2001; Volume 2, p. 747.
23. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
24. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [[CrossRef](#)]
25. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
26. Yeh, C.-H.; Huang, C.-H.; Kang, L.-W. Multi-Scale Deep Residual Learning-Based Single Image Haze Removal via Image Decomposition. *IEEE Trans. Image Process.* **2020**, *29*, 3153–3167. [[CrossRef](#)] [[PubMed](#)]
27. Yeh, C.-H.; Lin, M.-H.; Chang, P.-C.; Kang, L.-W. Enhanced Visual Attention-Guided Deep Neural Networks for Image Classification. *IEEE Access* **2020**, *8*, 163447–163457. [[CrossRef](#)]
28. Lin, C.-Y.; Tao, Z.; Xu, A.-S.; Kang, L.-W.; Akhyar, F. Sequential Dual Attention Network for Rain Streak Removal in a Single Image. *IEEE Trans. Image Process.* **2020**, *29*, 9250–9265. [[CrossRef](#)]
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
30. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
31. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
32. 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.
33. Li, X.; Shang, M.; Qin, H.; Chen, L. Fast Accurate Fish Detection and Recognition of Underwater Images with Fast R-CNN. In Proceedings of the OCEANS 2015—MTS/IEEE Washington, Washington, DC, USA, 19–22 October 2015; pp. 1–5.
34. Li, X.; Cui, Z. Deep Residual Networks for Plankton Classification. In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 September 2016; pp. 1–4.
35. Li, X.; Tang, Y.; Gao, T. Deep But Lightweight Neural Networks for Fish Detection. In Proceedings of the OCEANS 2017—Aberdeen, Aberdeen, UK, 19–22 June 2017; pp. 1–5.
36. Jiang, Y.; Li, W.; Zhang, J.; Li, F.; Wu, Z. YOLOv4-dense: A Smaller and Faster YOLOv4 for Real-time Edge-device Based Object Detection in Traffic Scene. *IET Image Process.* **2023**, *17*, 570–580. [[CrossRef](#)]
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; pp. 346–361. Available online: [https://link.springer.com/chapter/10.1007/978-3-319-10578-9\\_23](https://link.springer.com/chapter/10.1007/978-3-319-10578-9_23) (accessed on 2 December 2023).
38. Yeh, C.-H.; 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.* **2022**, *33*, 6129–6143. [[CrossRef](#)]
39. Deng, J.; Pan, Y.; Yao, T.; Zhou, W.; Li, H.; Mei, T. Single Shot Video Object Detector. *IEEE Trans. Multimed.* **2021**, *23*, 846–858. [[CrossRef](#)]
40. Yu, K.; Cheng, Y.; Tian, Z.; Zhang, K. High Speed and Precision Underwater Biological Detection Based on the Improved YOLOv4-Tiny Algorithm. *J. Mar. Sci. Eng.* **2022**, *10*, 1821. [[CrossRef](#)]
41. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.



42. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
43. Peng, Y.-T.; Cosman, P.C. Underwater Image Restoration Based on Image Blurriness and Light Absorption. *IEEE Trans. Image Process.* **2017**, *26*, 1579–1594. [[CrossRef](#)]
44. Peng, Y.-T.; Cao, K.; Cosman, P.C. Generalization of the Dark Channel Prior for Single Image Restoration. *IEEE Trans. Image Process.* **2018**, *27*, 2856–2868. [[CrossRef](#)]
45. Chiang, J.Y.; Chen, Y.C. Underwater Image Enhancement by Wavelength Compensation and Dehazing. *IEEE Trans. Image Process.* **2012**, *21*, 1756–1769. [[CrossRef](#)]
46. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
47. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19. ISBN 978-3-030-01233-5.
48. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss Functions for Image Restoration With Neural Networks. *IEEE Trans. Comput. Imaging* **2017**, *3*, 47–57. [[CrossRef](#)]

**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.