Article

Self-Supervised Pre-Training Joint Framework: Assisting Lightweight Detection Network for Underwater Object Detection

Zhuo Wang, Haojie Chen, Hongde Qin *, Qin Chen

College of Shipbuilding Engineering, Harbin Engineering University, Harbin 150001, China
* Correspondence: qinhd@hrbeu.edu.cn

Abstract: In the computer vision field, underwater object detection has been a challenging task. Due to the attenuation of light in a medium and the scattering of light by suspended particles in water, underwater optical images often face the problems of color distortion and target feature blurring, which greatly affect the detection accuracy of underwater object detection. Although deep learning-based algorithms have achieved state-of-the-art results in the field of object detection, most of them cannot be applied to practice because of the limited computing capacity of a low-power processor embedded in unmanned underwater vehicles. This paper proposes a lightweight underwater object detection network based on the YOLOX model called LUO-YOLOX. A novel weighted ghost-CSPDarknet and simplified PANet were used in LUO-YOLOX to reduce the parameters of the whole model. Moreover, aiming to solve the problems of color distortion and unclear features of targets in underwater images, this paper proposes an efficient self-supervised pre-training joint framework based on underwater auto-encoder transformation (UAET). After the end-to-end pre-training process with the self-supervised pre-training joint framework, the backbone of the object detection network can extract more essential and robust features from degradation images when retrained on underwater datasets. Numerous experiments on the URPC2021 and detecting underwater objects (DUO) datasets verify the performance of our proposed method. Our work can assist unmanned underwater vehicles to perform underwater object detection tasks more accurately.

Keywords: underwater object detection; self-supervised learning; lightweight detection network; deep learning

1. Introduction

Due to the shortage of land resources, the exploitation of the marine field has been paid more attention to recently. As an essential part of many underwater tasks, underwater object detection technology plays an important role in various fields, including marine resource exploitation, marine biological fishing and breeding, and underwater human–robot interaction [1,2]. A lightweight and accurate underwater object detection algorithm is also of great significance to the intelligence of unmanned underwater vehicles (UUVs). It is the key for UUVs to explore the ocean autonomously. Object detection networks (ODNs) have made rapid development with the application of deep learning (DL) technology to the computer vision field in recent years. However, underwater object detection has been a challenging task. First, the underwater environment is more complex than the terrestrial environment. Compared with blue and green lights with a shorter wavelength, red light has a longer wavelength and decays faster in the water, which causes the color of underwater images to tend to be blue-green. Due to the scattering of incident light by suspended particles and microorganisms in water, as well as the interference of artificial light sources, the contrast in underwater images is low, and targets in the images are not clear. All mentioned greatly affect the detection accuracy of underwater object detection models. Second, in practical industrial applications, underwater object detection networks

often need to be deployed on UUVs with limited power for their intelligence and flexibility in the exploration of the ocean environment [3]. This means that most ODNs based on DL with a large number of parameters and high computational complexity are not applicable to these applications. Therefore, how to design a high-precision lightweight ODN that can be deployed on low-power processors has been one of the main problems in underwater object detection. Finally, there have been no large universal underwater datasets to support the research of underwater object detection, such as COCO [4] and PASCAL VOC [5] datasets.

Most of the existing methods use image enhancement technology to correct the color of underwater images and enhance the clarity of these images before underwater object detection [6–10]. Unfortunately, some of the image enhancement algorithms cannot be applied to underwater object detection tasks based on UUVs. Image enhancement algorithms aim to generate underwater images suitable for viewing with human eyes, and the evaluation criteria typically include the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). However, the common metric in underwater object detection is average precision (AP), indicating a gap between the two tasks. In addition, the application of image enhancement algorithms can increase the complexity of the entire underwater object detection task, which is also not beneficial to their deployment on low-power processors. Aiming at the requirement that an underwater ODN needs to be deployed on a low-power processor of UUVs, this paper proposes a lightweight ODN based on the YOLOX model [11] called LUO-YOLOX. The object detection network achieves an accuracy of 80.7% on the URPC2021 dataset and 80.9% on the DUO dataset [12], having 5.31 million (5.31 M) parameters and 3.81 giga floating-point operations (3.81GFLOPs). This indicates that a good trade-off between the lightweight network structure and the detection accuracy is achieved.

Different from the methods that use image enhancement algorithms before object detection, this paper proposes a self-supervised pre-training joint framework (SPJF) based on the underwater auto-encoder transformation (UAET) to solve the problems of poor image quality and scarce datasets. In the proposed framework, a simple and efficient underwater image degradation method was designed. Given random degradation parameters, an image from the land dataset PASCAL VOC2012 was randomly degraded into an underwater image using the underwater degradation pipeline. Then, the LUO-YOLOX's backbone was used to extract features from the image and the corresponding degraded image. Further, the UAET was used to decode the features to predict random degradation parameters. With the help of the UAET, the backbone of the proposed ODN can learn the intrinsic characteristics of targets from degraded underwater images. It should be noted that the UAET module can be applied to different ODNs. In this paper, this module was used to assist the proposed lightweight LUO-YOLOX. Many experiments have been conducted to prove that the proposed object detection network can achieve a good trade-off between the lightweight design and the detection accuracy and to demonstrate the high effectiveness of the proposed joint framework for self-supervised pre-training.

In summary, the main contributions to this study are as follows:

(1) A lightweight ODN based on the YOLOX model called LUO-YOLOX is proposed. The LUO-YOLOX network meets the requirement of lightweight design and can be deployed on an underwater robot, maintaining a high detection accuracy;

(2) An improved SPJF is introduced in order to handle several issues, i.e., underwater image color distortion, inconspicuous target features, and scarce underwater datasets. This framework uses the UAET module to guide the backbone of ODNs to learn more essential and robust features of a target from underwater degradation images using self-supervised learning. The UAET module is not contained in the forward inference process of the ODN, unlike in the image enhancement algorithms, so the parameter number and computation complexity of the entire ODN are not high.

The rest of this article is organized as follows. Section 2 reviews the related work in the underwater object detection field. Section 3 introduces the proposed lightweight ODN

and SPJF. Section 4 presents and analyzes the experimental results. Finally, Section 5 draws the main conclusions and presents future work.

2. Related Work

This section briefly reviews the current research on ODN based on DL and introduces some of the related studies on underwater object detection.

2.1. Object Detection Based on DL

As a basic task in computer vision, object detection has a wide range of industrial applications. With the help of DL, object detection has been developed rapidly in recent years. Due to the powerful feature extraction capability of convolutional neural networks (CNNs), the object detection algorithms have been greatly improved—by employing CNNs—in terms of both detection speed and detection accuracy compared to the traditional object detection methods, such as histograms of oriented gradients (HOG) [13] and a deformable part-based model (DPM) [14].

ODN based on DL can be roughly divided into two categories. The first category includes two-stage ODNs represented by the Fast-RCNN and Faster-RCNN [15] series. They select the candidate region of a target first and then detect the target in the candidate region. The second category includes the one-stage ODNs represented by the YOLO series [16–19]. The one-stage networks do not require the generation of candidate regions but can directly detect the target. Each of the two types of detection networks has certain advantages. For instance, the two-stage ODNs have high detection accuracy, while the one-stage ODNs have a fast detection speed. However, with the advent of one-stage detection networks, such as RetinaNet [20] and YOLOX [11], in recent years, they have surpassed the two-stage networks in terms of detection accuracy in some of their application areas. Namely, the one-stage ODNs have been more favored in industrial applications than the two-stage networks.

2.2. Underwater Object Detection

In recent years, various underwater ODNs have been proposed. Different solutions have been introduced to address the difficulty of underwater object detection. For instance, to solve the problem of underwater image color distortion, Yeh et al. [21] placed a color conversion network before the ODN to transform underwater images into gray images and then jointly trained the color conversion network and the detection network. Further, to alleviate the adverse effects of a complex underwater environment in underwater images, Zhao et al. [22] proposed using an auxiliary backbone network to learn the source-domain data and considered the connection between the backbone of ODN and the auxiliary backbone to eliminate the interference caused by complex source-domain information. Wang et al. [23] proposed adding an adaptive image enhancement module to the detection model before the ODN to solve the problems of unclear images and color distortion in underwater images. They used the object detection loss function to guide the added module to optimize the object detection parameters. Chen et al. [24] jointly trained a GAN network for underwater image restoration and an ODN and used the gradient backpropagation method to train the GAN network to generate underwater images conducive to object detection tasks. Zhang et al. [25] improved the YOLOv4 object detection model by using the MobileNetv2 [26] and proposed a multi-scale attention feature fusion mechanism to improve the detection accuracy of small targets under the condition of a large number of small targets in the underwater environment.

Although the above-mentioned studies have solved the problems in underwater object detection to a certain extent, they increased the complexity of the detection task. Namely, they require much effort in underwater image preprocessing and use complex modules to improve the accuracy of underwater ODNs. To overcome these limitations, this paper proposes a self-supervised pre-training method that does not require adding any module to the forward inference of the ODN while improving the detection accuracy and ensuring

the light weight of a detection model. The proposed method is described in detail in the following section.

3. Proposed Method

This section introduces the LUO-YOLOX model. In the proposed model, the backbone of YOLOX is lightened by using the weighted ghost-CSPDarknet to extract features from underwater images. In addition, the path aggregation network (PANet) [27] is improved, and a simplified PANet is used as a feature synthesis part in the proposed object detection network. Finally, the decoupled anchor-free head is used to detect the target. After introducing the LUO-YOLOX model, the proposed self-supervised pre-training joint framework is described in detail. The overall architecture of the proposed method is shown in Figure 1.
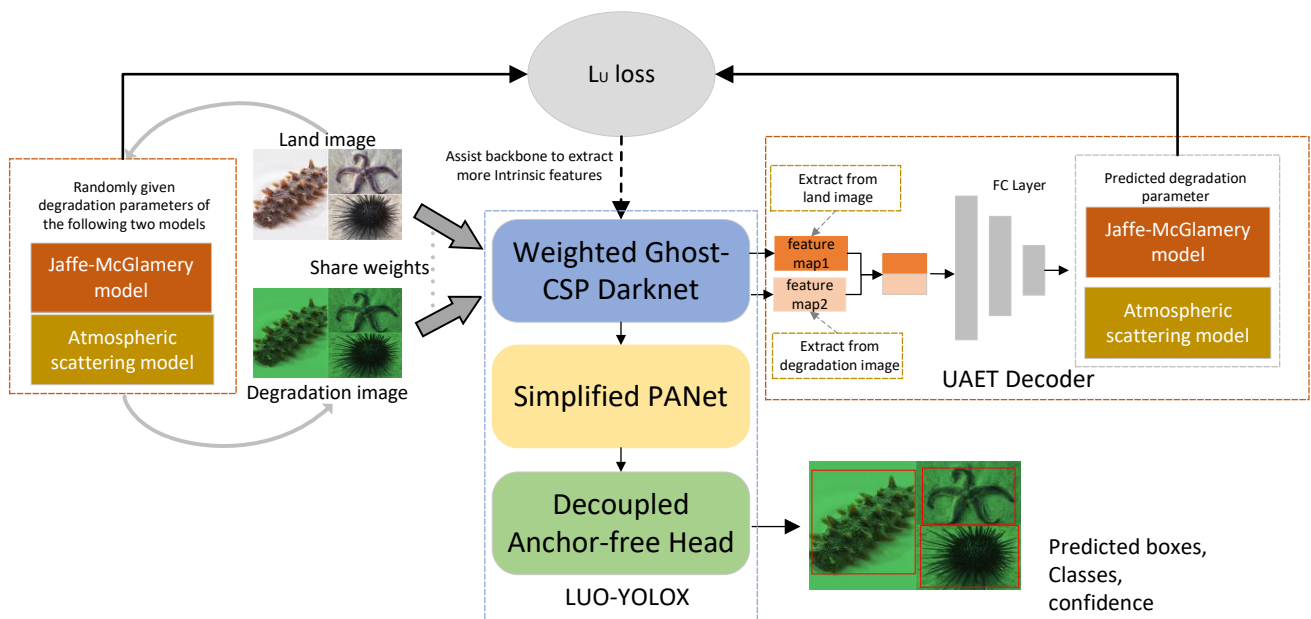

Figure 1. The overall architecture of the proposed self-supervised pre-training joint framework; L_U loss is the loss of the UAET decoder.

3.1. Weighted Ghost-CSPDarknet

This study designs a lightweight backbone based on the CSPDarknet model [11] called weighted ghost-CSPDarknet (WG-CSP). The WG-CSP structure is presented in Figure 2. In the figure, Base conv consists of Conv2d, BatchNorm, and SiLU, where Conv2d denotes the convolution operation, BatchNorm is the batch normalization, and SiLU represents the activation function; K3 denotes the convolution kernel size of 3×3; S2 indicates that the stride of convolution is two; P1 shows that the padding of convolution is one. The default values of K, S, and P are one, one, and zero, respectively. Further, Maxpool2d denotes the max-pooling downsampling operation; concat shows that the feature maps are concatenated in the channel dimension; $\oplus$ stands for the element-wise addition; num_cls in the head denotes the number of classes in datasets. The CSPDarknet is the backbone of the YOLOX model. It should be noted that a large number of parameters are taken over by the bottlenecks in the CSPDarknet. Therefore, more attention is paid to the lightweight design of the bottleneck.

The main innovation of this work relates to the GhostNet [28], which assumes that there is redundant information in the feature maps extracted by the common convolutional layer. Therefore, the Ghost Block, which uses convolution operations to extract rich, informative features while the redundant features are formed by cheap operations, is proposed. The Ghost Block was used to replace the common convolutional layers in bottlenecks, and depthwise separable convolutions were used as cheap operations. However, the Ghost Block simply concatenates rich, informative feature maps and redundant feature maps in

the channel dimension. It is obvious that the importance of the two feature maps is different, so the squeeze-and-excitation (SE) block [29] was added after the Ghost Block to weigh the feature maps extracted by the Ghost Block in the channel dimension. The SE block assigns larger weights to rich, informative feature maps and smaller weights to redundant feature maps. The structures of the Ghost Block and SE block are shown in Figure 3.
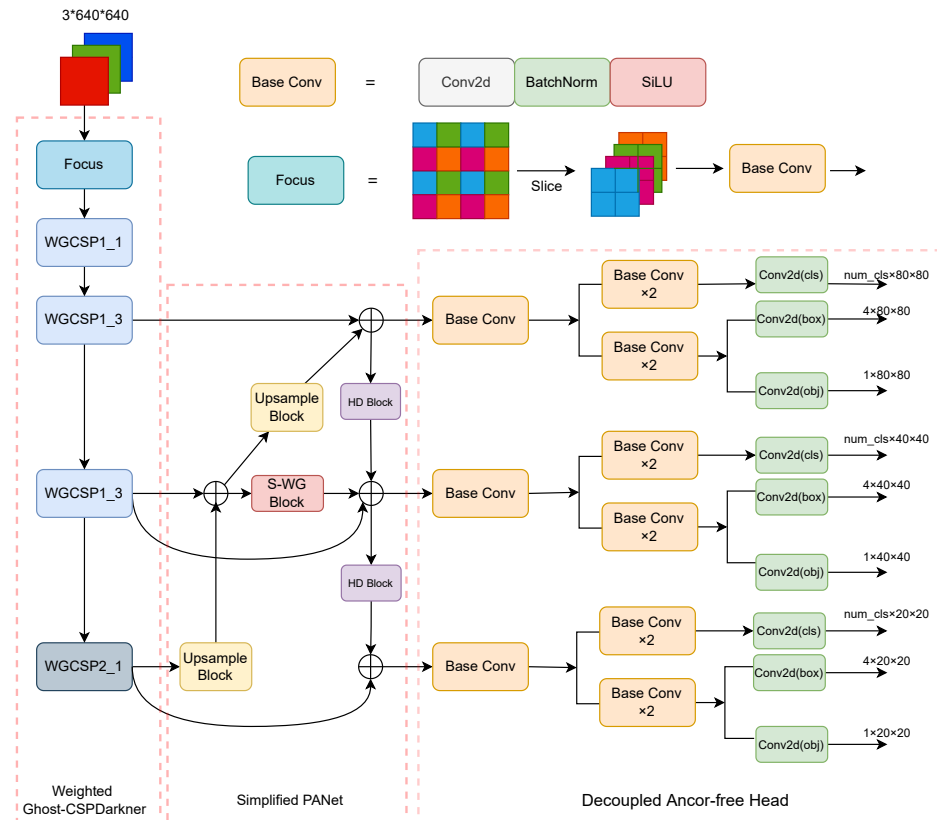

Figure 2. The block diagram of the LUO-YOLOX model.
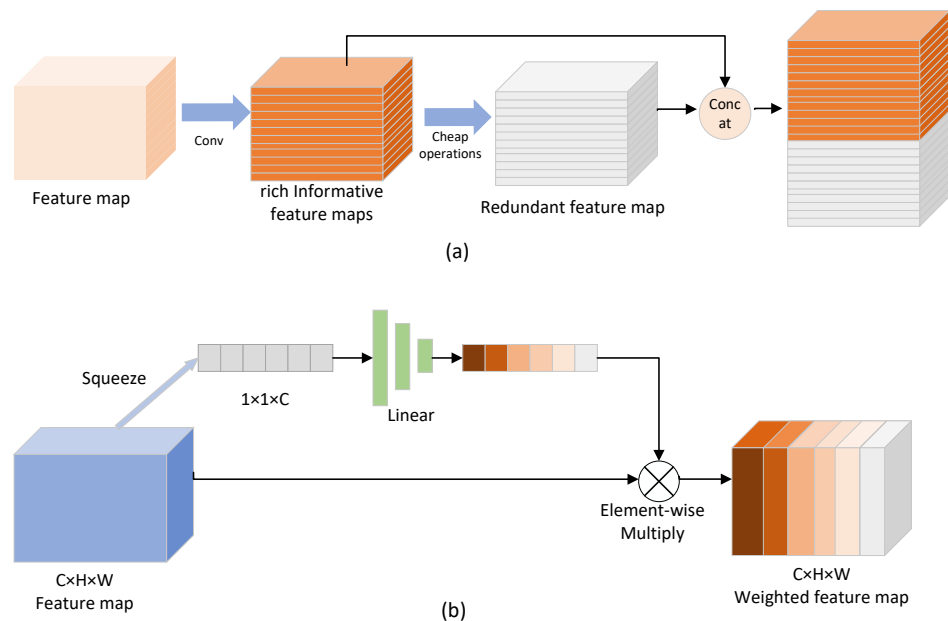

Figure 3. The structures of the Ghost Block and SE block. (**a**) The Ghost Block structure, where Conv denotes the convolution operations, and cheap operations denote linear transformation; (**b**) the SE block structure.

By combining the Ghost Block and the SE block, a weighted ghost (WG) block was designed and used to reconstruct the bottleneck. The structures of the WG Block, the new bottleneck, and the WG-CSP block are shown in Figure 4. Moreover, different from the YOLOX model, which uses spatial pyramid pooling (SPP) [30] to fuse features of different scales, the proposed WG-CSP uses a spatial pyramid pooling-faster (SPPF) block. The SPPF block can achieve the same results as the SPP block while reducing the calculation cost of the ODN and increasing the detection speed of the network. After applying the WG-CSP, the number of parameters of ODN reduces from 8.94 M to 7.93 M, while the floating-point operation number reduces from 5.63 GFlOPs to 4.86 GFLOPs compared to the YOLOX model.
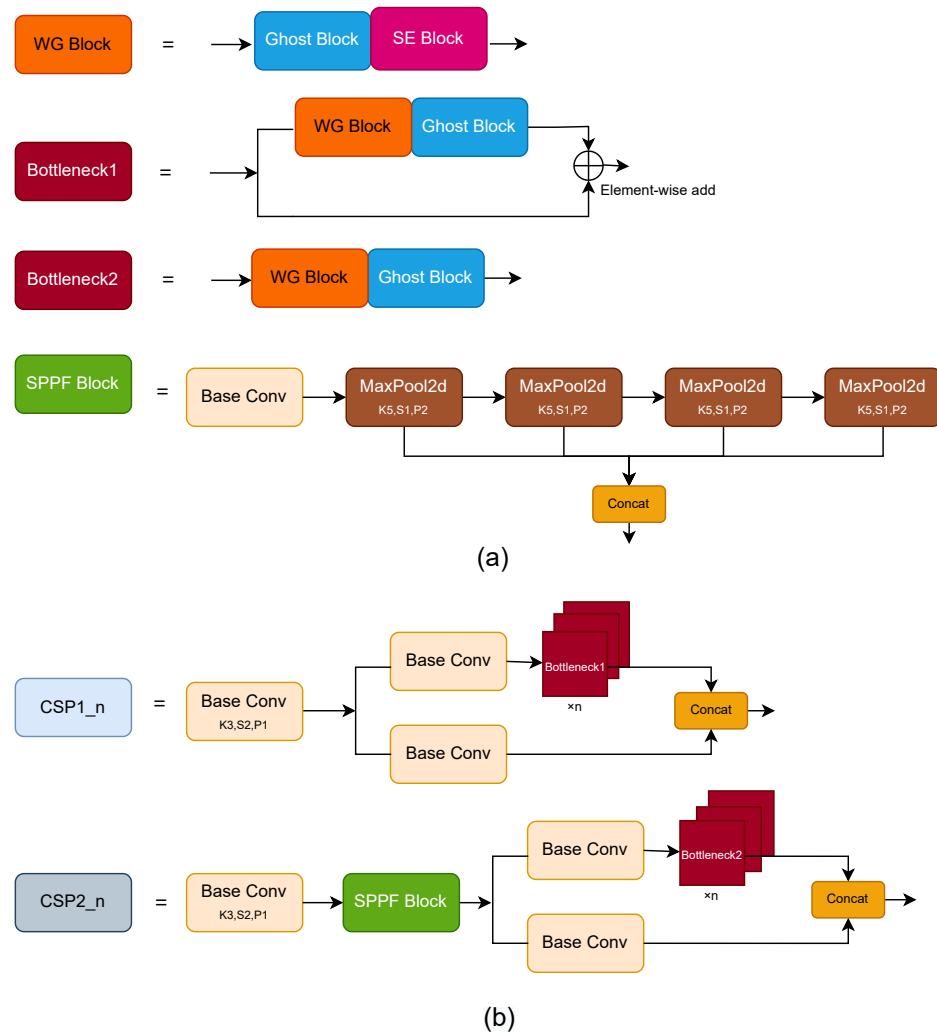

Figure 4. The details of the WG-CSP. (**a**) The structure of the WG Block, Bottleneck, and SPPF; (**b**) basic WGCSP Blocks in the WG-CSP consisting of the modules shown in (**a**).

3.2. Simplified PANet

Currently, the majority of one-stage ODNs use a PANet as their neck to fuse features from different layers, which is beneficial for detecting targets of different scales. The PANet is designed based on the feature pyramid network (FPN) [31], and its main function is to add the bottom-to-top feature fusion path on the basis of the top-down feature fusion path of an FPN. This significantly alleviates the deficiency of low semantic information in the top fusion layer of the FPN. However, due to its complex structure, the PANet is not suitable for lightweight ODNs. In view of this, this study proposed a simplified PANet, whose structure is shown in Figure 2.

There are a large number of 1×1 convolution kernels in a PANet. Therefore, to make the network lightweight, this study proposes a new module called the shuffle-weighted ghost (S-WG) block to replace the 1×1 convolution kernels. First, the input feature maps were split in the channel dimension. Half of the input feature maps were calculated by the WG block, while the other half of the feature maps were obtained by identity mapping. Finally, the channels [32] were combined to enhance the feature maps' interactions in the channel dimension. With the help of the S-WG block, the number of network parameters was decreased significantly. In addition, a downsampling block called hybrid downsampling (HD) was designed based on the S-WG block. In the HD block, maximum pooling downsampling and S-WG downsampling were performed on the input feature maps in turn. Then, the two output feature maps were concatenated in the channel dimension. In addition, inspired by the bidirectional feature pyramid network (BiFPN) [33], the cross-level links were used at the feature level to achieve better feature fusion. The structures of the S-WG, upsample, and HD blocks in the S-PANet are shown in Figure 5.
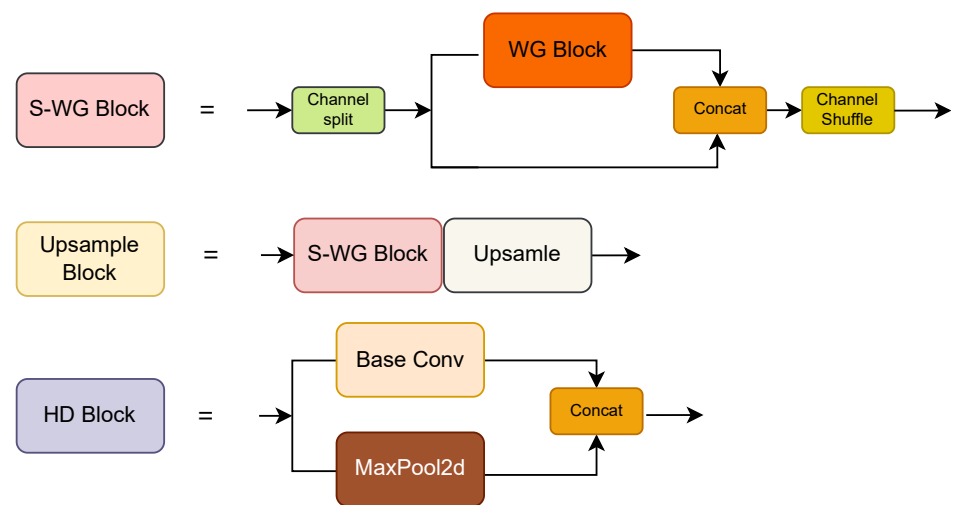

Figure 5. The structures of the S-WG, upsample, and HD blocks in the S-PANet.

After using the S-PANet, the parameters of our ODN were reduced from 7.93 M to 5.31 M, and the number of floating-point operations was reduced from 4.86 GFLOPs to 3.81 GFLOPs while ensuring the feature fusion ability of the proposed object detection network.

3.3. Decoupled Anchor-Free Head

The decoupled anchor-free head (DAH) was used for underwater object detection and its structure is shown in Figure 2. Decoupled anchor-free detection heads are at the end of LUO-YOLOX. It has three branches for detecting large, medium, and small targets, respectively. Compared to the anchor-based detector, an anchor-free detector omits the complex anchor generation process. In addition, due to a large number of small targets in the underwater object detection task and difficulty in matching the anchor generated by the anchor-based detector to the ground truth for small targets, a high missing rate of small targets can be caused. The proposed anchor-free detector can solve this problem based on keypoint detection. Moreover, the decoupled technology was used, and two branches were used to perform the regression of the detection box and the classification of the target.

3.4. SPJF

The attenuation and scattering of the light medium have a great influence on the detection accuracy of an underwater ODN. To overcome this limitation, most of the existing methods aim to enhance an underwater image using image enhancement algorithms. However, in some cases, this approach does not yield satisfactory results. Unlike this approach, this study proposes an innovative SPJF to solve the mentioned problems in underwater

object detection. In addition, inspired by the multi-task auto encoding transformation (MAET) [34], this study proposes the UAET. Applying self-supervised learning, the proposed UAET handles the adverse effects of color bias and blurring in underwater images when extracting the intrinsic feature of objects.

First, an underwater degradation pipeline was designed, as shown in Figure 6. Considering the time cost of the ODN training, the pipeline was designed in a way to omit the design complexity. Further, the Jaffe–McGlamery model [35] was used to degrade land optical images to underwater images, which can be expressed as follows:

$$I_\lambda(x) = t_\lambda(x) \cdot J_c(x) + [1 - t_\lambda(x)] \cdot A_c \tag{1}$$

$$t_\lambda(x) = \exp[-\beta_\lambda \cdot d(x)], \tag{2}$$

where λ represents r, g, and b channels; $I_\lambda(x)$ is a degraded underwater image in a channel λ; $J_\lambda(x)$ is the clear land optical image in channel λ; A_c is the background light; $t_\lambda(x)$ is the underwater attenuation rate of channel λ, which is calculated by Equation (2); β_λ represents the attenuation coefficients of different wavelengths of the light in water; $d(x)$ represents the depth information on image pixels.
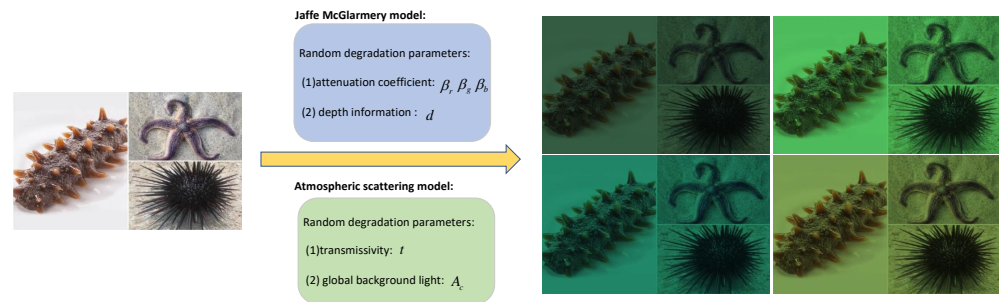

Figure 6. The general view of the proposed underwater degradation pipeline.

The atmospheric scattering model is used to simulate the foggy effect of underwater optical images, which can be expressed as follows:

$$I(x) = J(x) \cdot t(x) + A \cdot [1 - t(x)], \tag{3}$$

where $I(x)$ represents a foggy image; J(x) is a clear image; $t(x)$ is the transmittance of light; A is the intensity of background light.

Given a clear land image, this study aimed to degrade it to an overall greenish-blue, fuzzy optical image of the water using the underwater degradation pipeline. The degradation coefficients of the pipeline were randomly set within a certain range. According to the UWCNN research [36], four types of underwater attenuation coefficients were adopted in this study to generate different characteristics of underwater images. In addition, the depth of an underwater optical image was randomly assigned between 1.5 m and 8 m, and the depth value of the whole underwater image obeys the normal distribution. Further, the transmittance and background light intensity of an underwater scene were set randomly and used by the atmospheric scattering model to generate foggy underwater images with different degrees of ambiguity. The ranges of the degradation coefficients are shown in Table 1, where $U(a, b)$ denotes a number randomly set between a and b, and $N(a, b)$ indicates the normal distribution with a mean value of a and a standard deviation of b.

Table 1. The underwater degradation pipeline parameters.

Transformation	Parameters	Value
Jaffe-McGlamery model	Attenuation coefficients β_r, β_g, and β_b (four groups of parameters)	$\begin{bmatrix} 0.750 & 0.885 & 0.830 \\ 0.750 & 0.885 & 0.725 \\ 0.710 & 0.820 & 0.650 \\ 0.670 & 0.730 & 0.630 \end{bmatrix}$
	Depth information $d(x)$	$d(x) = \alpha \times N(0,1) + U(1,10)$
Atmospheric scattering model	Background light A_c	$U(0.6, 0.95)$
	Transmissivity t	$U(0.3, 0.95)$

After degenerating clear land images into underwater images, both the land images and the simulated degraded underwater images were input into the weight-sharing WG-CSP backbone. The features of the two types of images were extracted by the WG-CSP. In the channel dimension, the two extracted feature maps were concatenated and input into the UAET decoder. The UAET decoder consists of two fully connected layers that decode the concatenated feature maps to predict the degradation parameters that are randomly set in the underwater degradation pipeline. Thus, the UAET adopts self-supervised learning to predict the degradation parameters to assist the WG-CSP in extracting more intrinsic features. The more accurate the prediction result of UAET decoder of randomly set degradation parameters is, the more essential the features extracted by the WG-CSP are.

The UAET self-supervised module can assist the backbone of LUO-YOLOX through joint training to extract more intrinsic features from degraded underwater images. The operational principle of the SPJF is presented in Figure 1. It should be noted that the SPJF framework represents a pre-training method based on land images and randomly degraded underwater images. After pre-training, the pre-trained model was retrained on real underwater datasets to achieve better results in underwater object detection.

3.5. Loss Function

The total loss of the proposed SPJF consists of two parts, which can be expressed by:

$$L_{total} = \omega_1 \cdot L_{obj} + \omega_2 \cdot L_U, \tag{4}$$

where L_{obj} is the loss of the ODN; L_U is the loss of the UAET decoder; ω_1 and ω_2 are two hyper-parameters used to balance the loss function, and they are set to 1.0 and 0.01, respectively.

Moreover, L_{obj} includes the classification loss, box regression loss, and confidence loss, which can be expressed by:

$$L_{obj} = L_{cls} + L_{conf} + l_{bbox}$$
$$= \sum_{i=0}^{s^2} I_{ij}^p \sum_{c \in cls} BCE(P_i(c), \widehat{p}_i(c)) + \sum_{i=0}^{s^2} BCE(c_{ij}, \widehat{c}_{ij})$$
$$+ \sum_{i=0}^{s^2} I_{ij}^p (1 - Iou(box_{ij}, pbox_{ij})), \tag{5}$$

where L_{cls} denotes the classification loss, L_{obj} is the confidence loss, and L_{bbox} represents the box regression loss; $IouLoss$ is the box regression loss; Iou stands for the intersection over union, which can be expressed by $Iou_{AB} = \frac{A \cap B}{A \cup B}$; the s^2 is the number of anchor points of the output feature maps; I_{ij}^p indicates whether the anchor point is a positive sample; BCE stands for the binary cross entropy loss, which can be expressed by $BCE(c_{ij}, \widehat{c}_{ij}) = -c_{ij} \log(\widehat{c}_{ij}) - (1 - c_{ij}) \log(1 - \widehat{c}_{ij})$; $P_i(c)$ and $\widehat{p}_i(c)$ are the probabilities of the cth predicted class network and the corresponding ground truth, respectively; c_{ij} and $\widehat{c}_{ij}$ are the confidence degrees of the prediction box (i, j) and its ground truth, respectively; box_{ij} and $pbox_{ij}$ denote the prediction box and its ground truth, respectively.

Further, L_U includes the deviation loss and cosine similarity loss. The cosine similarity loss is used for gradient shunting for a multi-task framework to reduce the impact of the loss function, which is designed for the UAET in the object detection task. The L_U can be expressed as follows:

$$L_U = \lambda_1 \cdot L_d + \lambda_2 \cdot L_{cs}$$
$$L_d = \|d_p - d_{gt}\|_1$$
$$L_{cs} = \frac{\frac{\partial E}{\partial D_{obj}} \cdot \frac{\partial E}{\partial D_U}}{\left\|\frac{\partial E}{\partial D_{obj}}\right\| \cdot \left\|\frac{\partial E}{\partial D_U}\right\|}, \tag{6}$$

where L_d denotes the loss deviation between the predicted degradation parameter (d_p) and the corresponding ground truth (d_{gt}); L_{cs} is the cosine similarity loss; λ_1 and λ_2 are hyperparameters; E represents the encoder of the joint framework; D_{obj} is the object detection decoder; D_U denotes the UAET decoder.

4. Experiments and Results

4.1. Training Parameters and Evaluation Metrics

The experiments were based on the general object detection framework MMDetection [37]. The operating system was Ubuntu20.04, and Nvidia GeForce 2080ti GPU was used to train the proposed network. The proposed SPJF was trained on the Pascal VOC2012 dataset, which is a land image dataset. This dataset was randomly degraded into underwater images using the proposed underwater degradation pipeline. Then, the land image and the corresponding degraded image were input into the proposed joint framework together and used for the end-to-end training. The size of the training image was 512×512, the batch size was 16, and the initial learning rate was 0.01. After the joint training for 100 epochs, the trained model is retrained on the URPC2021 and detecting underwater object (DUO) dataset for further model optimization. The URPC2021 contained four types of targets: holothurian, echinus, starfish, and scallop, including a total of 7600 underwater images divided into the training and test sets according to the ratio of 9:1. The DUO also contained four types of targets: holothurian, echinus, starfish, and scallop, including 7782 images, of which 6671 were used for training and 1111 for testing. Some of the samples of the two datasets are shown in Figure 7.
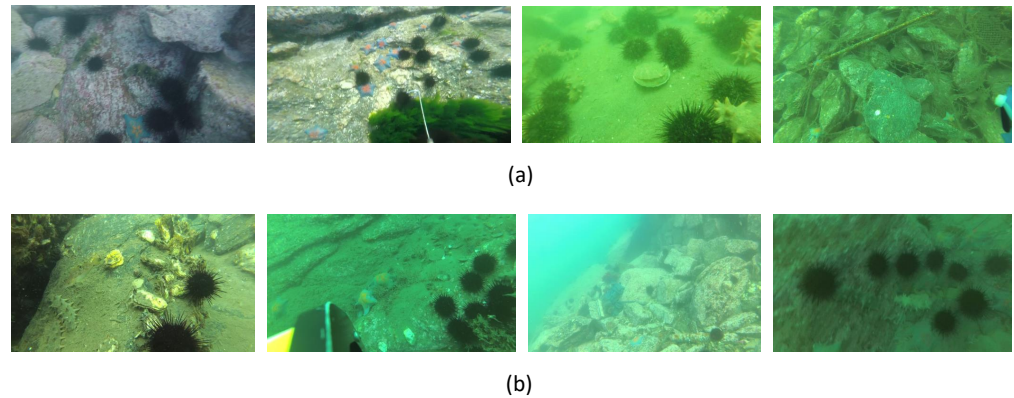
(a)

(b)

Figure 7. Samples of the URPC2021 and DUO datasets. (**a**) URPC2021 dataset; (**b**) DUO dataset.

In the training process on the underwater datasets, the image size was set to 640×640, the batch size was 12, and the stochastic gradient descent (SGD) optimizer was used to optimize network parameters. The initial learning rate was set to 0.01, and the cosine annealing strategy [38] was used to decrease the learning rate automatically as follows:

$$l_r = \{[1 + \cos(\frac{\pi \cdot e}{e})] \times \frac{1 - l_{rf}}{2} + l_{rf}\} \times init\_l_r, \tag{7}$$

where e is the number of epochs; $init\_l_r$ is the initial learning rate; l_{rf} is a hyper-parameter.

The main evaluation metrics used in the experiments included the average precision of the ODN when the *Iou* threshold was set to 0.5 (AP50) and the mean average precision (AP). The AP was calculated by:

$$AP = \frac{AP50 + AP55 + AP60 + ... + AP95}{10}. \tag{8}$$

AP is the main metric by which to measure the performance of ODNs, which is defined as the percentage of the number of objects correctly detected in the total number of detected objects. Besides this, we also calculated the AR which denoted the average recall of the detection networks, the parameters (Params) and computational complexity (FLOPs) of the object detection networks. AR is defined as the percentage of the number of objects correctly detected of the total number of objects in the dataset.

4.2. Contrast Experiment

We compared our method with the state-of-art ODNs to verify the effectiveness of the proposed network model. For the fairness of comparison, the image size, batch size, and data augment method (random flip only) of all compared detection networks were the same during the training process. The results obtained on the URPC2021 and DUO datasets are shown in Table 2. The average precision curves of the detection networks are displayed in Figure 8. The graphs in the first row show the results on the URPC2021 dataset and those in the second row present the result on the DUO dataset. Further, the first-column figures display the AP curves, and the second-column figures show the AP50 curves. The green point denotes the result of the first epoch, while the red point represents the best result obtained after 100 epochs.

As shown in Table 2, compared with the other models, the proposed model could guarantee a small number of parameters and low computational complexity and achieve high detection accuracy, which indicated that a good trade-off between the lightweight characteristic and the detection accuracy was achieved. The proposed model had slightly lower accuracy than the baseline YOLOX model [11] but also had 41% fewer parameters. Compared with the other lightweight ODNs, such as YOLOX-tiny, SSD [39] with the MoblieNetv3 [40] backbone, and YOLOv3 [16] with the MobileNetv3 backbone, the proposed model achieved a significant advantage in detection accuracy. In addition, compared

with the RetinaNet [20] and YOLOF [41] models, the proposed network not only had a much smaller number of parameters but also achieved higher average precision on the two underwater datasets. The results in the last two rows in Table 2 show that, after using the UAET for self-supervised pre-training, the average detection accuracy in the underwater object detection task was improved.

Zhang et al. [25] proposed a lightweight ODN that also achieves a good trade-off between accuracy and speed. We compared our method with it on the URPC dataset to verify the effectiveness of our method. The compared results are shown in Table 3. We achieved the same mAP as Zhang et al. However, the parameters of our LUO-YOLOX are only half of it. Our detection speed is also faster.

Table 2. The experimental results on the URPC2021 and DUO datasets.

| Model | Backbone | URPC2021 | | DUO | | Params (M)↓ | FLOPs (G)↓ |
		AP↑ (%)	AP50↑ (%)	AP↑ (%)	AP50↑ (%)		
SSD [39]	MobileNetv3	34.5	67.5	46.3	69.9	3.07	1.19
YOLOF [41]	ResNet50	37.5	72.1	51.1	74.0	42.13	16.6
RetinaNet [20]	ResNet50	44.5	79.8	58.8	81.5	36.17	34.83
YOLOv3 [16]	MobileNetv3	40.6	79.5	53.2	79.9	3.67	2.78
YOLOv3 [16]	DarkNet53	45.5	82.7	60.2	83.9	61.54	32.77
YOLOX-tiny [11]	CSPDarkNet	42.3	78.1	57.2	78.9	5.03	3.20
YOLOX-s [11]	CSPDarkNet	47.4	82.1	60.1	81.1	8.94	5.63
Proposed LUO-YOLOX	WG-CSP	46.3	80.7	59.3	80.9	5.31	3.81
LUO-YOLOX + SPJF	WG-CSP	46.7	81.5	59.6	81.6	5.31	3.81
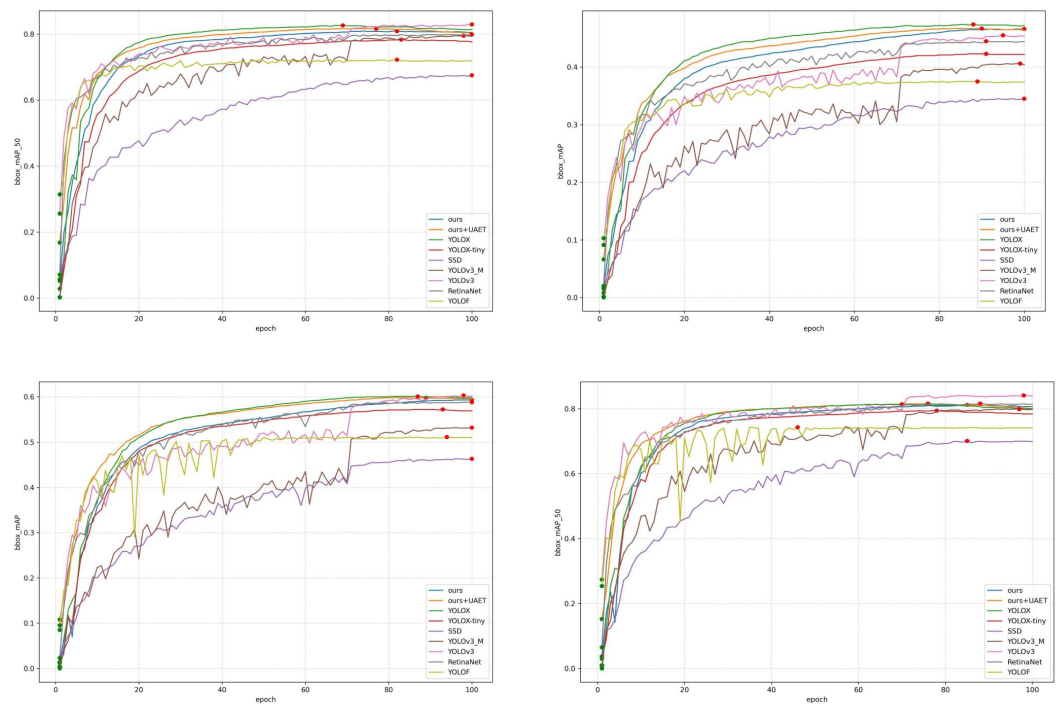
Figure 8. The average precision curves of the underwater ODNs.

Table 3. The results of the comparison with Zhang et al.

Model	Backbone	Input Size	mAP(%)	Params (M)	GPU	Speed (FPS)
Zhang et al.	MobileNet v2	416 × 416	81.67	10.73	2080ti	44.18
Ours	WG-CSP	640 × 640	81.50	5.31	2080ti	50.60

4.3. Ablation Study

Next, the ablation experiments were conducted on the URPC2021 to further verify the effectiveness of the proposed SPJF. In the experiments, we set all the training parameters, such as image size, batch size, learning rate, and optimizer type as the same, as these parameters have a great impact on the trained models. In the first experiment, the normal pre-training method was used; namely, the LUO-YOLOX model was pre-trained directly on Pascal VOC2012 and the weights of the trained model were transferred to the underwater object detection task. In the second experiment, the images in the PASCAL VOC2012 dataset were degraded into underwater images using the underwater degrade pipeline. The obtained dataset was used to pre-train the LUO-YOLOX model for 100 epochs. Then, the weights of the pre-trained model were transferred to the underwater object detection task. In the last experiment, the proposed SPJF was used. The PASCAL VOC2012 dataset was degraded into underwater images, and the UAET module was used in the self-supervised pre-training to assist the backbone of the ODN to extract inherent features in the pre-training stage. The experimental results of the three experiments on the URPC2021 dataset are shown in Table 4. The reason we did not train LUO-YOLOX with UAET without the underwater degradation pipeline is that the UAET module can be used for self-supervised learning only after using the underwater degradation pipeline to generate degraded images.

As shown in Table 4, the performance of the ordinary pre-training was roughly equivalent to that of the pre-training with degraded images. However, the AP and AR of the object detection network were worse than those of the direct training on the underwater dataset without a pre-training strategy. This could be caused by the domain shift phenomenon between the clear land images and the underwater images, which meant that weights obtained by the pre-training process on the land images could not match underwater target detection well. Compared with the two pre-trained cases, when the weights obtained by the proposed SPJF were transferred to the underwater object detection task, both the average precision and the average recall rate were improved. This proved that the proposed method could solve the domain shift problem to a certain extent when performing transfer learning.

Table 4. The ablation results of the proposed SPJF scheme.

	Underwater Degradation Pipeline	UAET	AP↑ (%)	AP50↑ (%)	AR↑ (%)
1			46.1	81.2	54.4
2	✓		46.0	81.1	54.6
3	✓	✓	46.7	81.5	59.6

"✓" denotes whether the proposed module model was used in the experiments.

It has been assumed that using the COCO dataset as a land dataset could ensure higher precision than using the PASCAL VOC2012 dataset. Namely, compared to the PASCAL VOC2012 dataset, the COCO dataset includes more data samples and richer objects, which is beneficial in terms of the ODN learning more features during the pre-training stage. However, training a prediction model on the COCO dataset requires using a high-performance GPU. Therefore, if the subsequent experimental equipment allows, this type of model training will be performed in future work.

4.4. Effect of Image Enhancement on Object Detection Results

In this subsection, the effect of image enhancement on the object detection result is analyzed. The MLLE [42] and FUnIE-GAN [43] methods were selected as image enhancement algorithms to enhance images in the URPC2021 dataset. The former is a traditional image enhancement algorithm and the latter is a DL-based algorithm. The results obtained by the two methods on the URPC2021 dataset are shown in Figure 9.
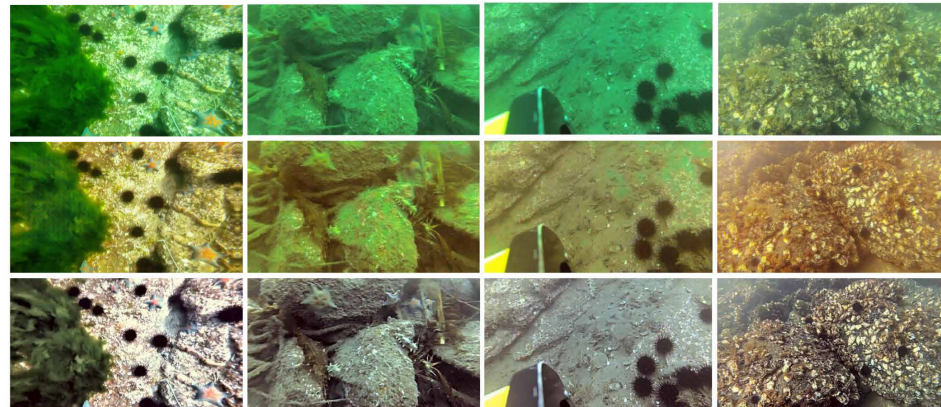

Figure 9. Enhanced samples of images in the URPC2021 dataset. The first row shows the original images; the second row displays the samples enhanced by the FUnIE-GAN method; the third row is the samples enhanced by MLLE.

The enhanced URPC2021 dataset was used to train the proposed lightweight object detection network (i.e., the LUO-YOLOX model). The results on this dataset are shown in Table 5, where it can be seen that, different from general cognition, the performance of the detector was not improved by using the enhanced images for training. The low detection accuracy after using the image enhancement method could be caused by the difference in the two tasks' evaluation metrics and optimized targets. In addition, there was a domain shift between underwater images and enhanced images, which caused low average recall efficiency [44]. After using the proposed SPJF, the recall in the underwater object detection tasks was improved significantly, as well as the AP.

Table 5. The experimental results obtained on the enhanced URPC2021 dataset.

	AP↑ (%)	AP50↑ (%)	AR↑ (%)
MLLE	46.4	81.1	55.6
FUnIE-GAN	45.3	80.6	54.6
Proposed SPJF	46.7	81.5	59.1

4.5. Visualization of Detection Results

The detection results of our methods on the URPC2021 and DUO datasets are presented in Figure 10. (a) shows the results on the DUO dataset. (b) shows the result on the URPC2021 dataset. The first rows of Figure 10a,b show the results directly obtained by the proposed LUO-YOLOX. The second rows present the results detected by LUO-YOLOX pre-trained by the proposed framework. Yellow boxes denote false detection targets and purple boxes denote missing detection targets. The method directly detected targets by the LUO-YOLOX model, facing the problem of missing and false detection. As shown in the first row in Figure 10a, a number of holothurian and starfish targets were missed; the ODN mistakenly identified water plants as holothurian, which was not beneficial for the improvement of the accuracy of the underwater ODNs. The first row in Figure 10b indicates the same problem. The underwater detection network incorrectly identified the blue gloves as starfish, and some of the holothurian targets were also missed. However, after using

the proposed SPJF to pre-train the LUO-YOLOX model, the mentioned problems were effectively alleviated. The second row in Figure 10a,b shows that the category and position of the targets were accurately identified and located without any missing or false detection.
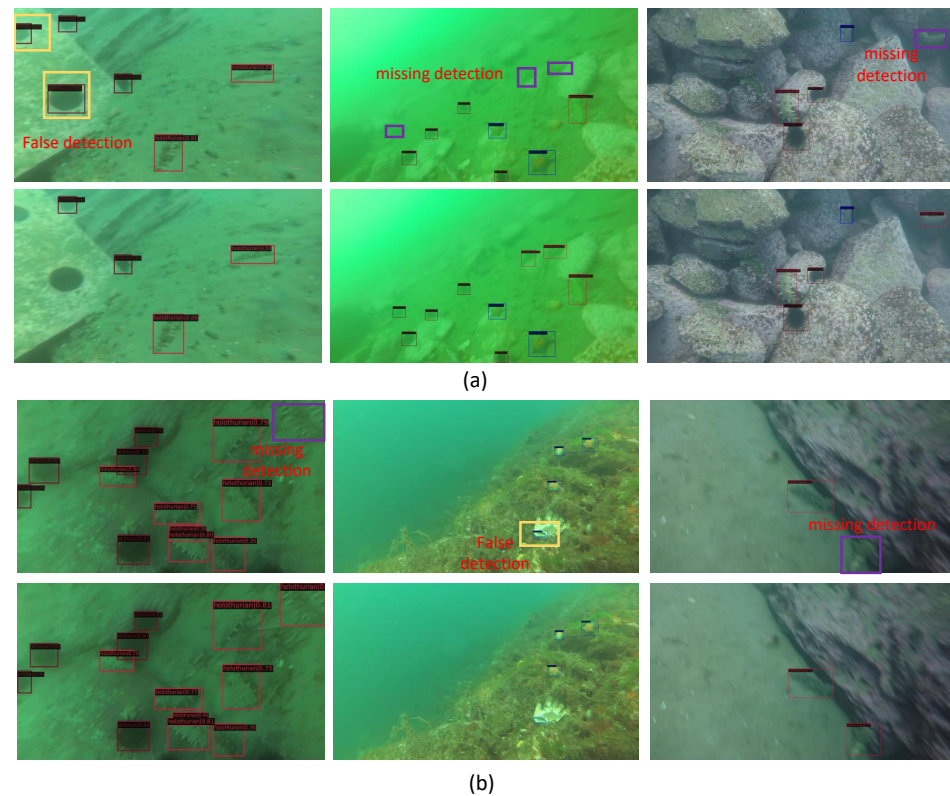

(a)

(b)

Figure 10. The detection results of the proposed method. (**a**) shows the results on the DUO dataset. (**b**) shows the result on the URPC2021 dataset

The detection result shows that the performance of the proposed object detection network, the LUO-YOLOX model, became more stable after using the proposed SPJF.

5. Discussion

Compared with other ODNs, the LUO-YOLOX we proposed ensures high AP with a small number of parameters and low computational complexity in underwater object detection tasks. The SPJF we proposed guides the backbone of underwater ODNs to extract more essential features from the underwater images in the manner of self-supervised learning. This improves the accuracy and robustness of the underwater ODNs without increasing any parameters and computational complexity. In this paper, we successfully used it to assist the proposed LUO-YOLOX.

However, our methods have some limitations. Due to the large number of multi-branches in LUO-YOLOX, the detection speed of our LUO-YOLOX is slower compared with YOLOX even though its computation complexity is lower. We tested the detection speed on the Nvidia GeForce 2080ti GPU. The detection rate of the LUO-YOLOX is 50 images per second while the detection rate of the YOLOX is 70 images per second. However, compared to most other underwater ODNs, LUO-YOLOX still has an advantage in terms of detection speed.

In the following study, we will try to simplify the branch of LUO-YOLOX and design a more efficient structure to increase the detection speed of the network. Moreover, we will use a larger land dataset to train the proposed SPJF. We speculate that SPJF will achieve more satisfactory results with a larger amount of data. Moreover, our UUVs are being developed in Qindao, China. Once the project progress meets the experimental conditions, we will conduct experiments to capture real underwater images and test the performance of our method on them.

6. Conclusions

In this paper, a lightweight underwater object detection network called the LUO-YOLOX, which can be implemented in the low-power processors of UUVs, was proposed. The proposed network can achieve high detection accuracy under a small number of parameters and at low computational complexity in underwater object detection tasks. It achieves an accuracy of 80.7% on the URPC2021 dataset and 80.9% on the DUO dataset. To solve the problems of color distortion and unclear target features of underwater images, the SPJF framework, which can be used to guide the backbone of an underwater object detection network to extract more intrinsic and robust target features from degraded underwater images, was developed. A large number of comparison and ablation experiments have been conducted to verify the proposed framework and detection model. The results demonstrate the high effectiveness of the proposed lightweight object detection network and SPJF framework.

Author Contributions: Conceptualization, Z.W. and H.C.; methodology, H.C.; software, H.C.; validation, Z.W., H.C. and Q.C.; formal analysis, Z.W.; investigation, H.C.; resources, Z.W.; data curation, H.C.; writing—original draft preparation, H.C.; writing—review and editing, H.C.; visualization, H.C.; supervision, Z.W.; project administration, H.Q.; funding acquisition, H.Q. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data that support the findings in this study are given in the supporting information or by contacting the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average precision
AR	Average recall
CNN	convolutional neural network
DAH	Decoupled anchor-free head
DUO	Detecting underwater objects
DL	Deep learning
DPM	Deformable part-based model
FPN	Feature pyramid network
HOG	Histograms of oriented gradients
LUO-YOLOX	lightweight underwater object detection network based on the YOLOX
ODN	Object detection network
PSNR	Peak signal-to-noise ratio
PANet	Path aggregation network
SSIM	Structural similarity
SPJF	Self-supervised pre-training joint framework
SE	Squeeze-and-excitation
S-WG	Shuffle-weighted ghost
SPP	Spatial pyramid pooling
SPPF	spatial pyramid pooling-faster
UAET	Underwater auto-encoder transformation
UUV	Unmanned underwater vehicle
WG	Weighted ghost
WG-CSP	Weighted ghost-CSPDarknet

References

1. Yu, J.; Xianglong, P.; Mingzhu, X.; Chong, W.; Hong, Q. An Underwater Human–Robot Interaction Using Hand Gestures for Fuzzy Control. *Int. J. Fuzzy Syst.* **2021**, *23*, 1879–1889. [CrossRef]
2. Jiang, Y.; Zhao, M.; Wang, C.; Wei, F.; Qi, H. A Method for Underwater Human–Robot Interaction Based on Gestures Tracking with Fuzzy Control. *Int. J. Fuzzy Syst.* **2021**, *23*, 2170–2181. [CrossRef]
3. Wang, G.; Wei, F.; Jiang, Y.; Zhao, M.; Wang, K.; Qi, H. A Multi-AUV Maritime Target Search Method for Moving and Invisible Objects Based on Multi-Agent Deep Reinforcement Learning. *Sensors* **2022**, *22*, 8562. [CrossRef]
4. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Volume 42, pp. 740–755.
5. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–308.
6. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. . [CrossRef]
7. Yu, H.; Li, X.; Feng, Y.; Han, S. Multiple attentional path aggregation network for marine object detection. *Appl. Intell.* **2022**, *53*, 2434–2451. [CrossRef]
8. Zhang, X.; Fang, X.; Pan, M.; Yuan, L.; Zhang, Y.; Yuan, M.; Lv, S.; Yu, H. A marine organism detection framework based on the joint optimization of image enhancement and object detection. *Sensors* **2021**, *21*, 7025. [CrossRef] [PubMed]
9. Venkatesh Alla, D.N.; Jyothi, B.N. Vision-based Deep Learning algorithm for Underwater Object Detection and Tracking. In Proceedings of the International Journal of Computer Vision, Xiamen, China, 19–21 August 2022; pp. 1–6.
10. Tang, L.; Xu, H.; Wu, H.; Tan, D.; Gao, L. Research on Collaborative Object Detection and Recognition of Autonomous Underwater Vehicle Based on YOLO Algorithm. In Proceedings of the 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2022; pp. 1664–1669.
11. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
12. Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; Wang, Z. A dataset and benchmark of underwater object detection for robot picking. In Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; pp. 1–6.
13. Navneet, D.; Bill, T. 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; pp. 886–893.
14. 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]
15. Shaoqing, R.; Kaiming, H.; Ross, G.; Jian, S. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.
16. Joseph, R.; Ali, F. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
17. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
18. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
19. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
20. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
21. 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]
22. Zhao, Z.; Liu, Y.; Sun, X.; Liu, J. Composited FishNet: Fish Detection and Species Recognition From Low-Quality Underwater Videos. *IEEE Trans. Image Process.* **2021**, *30*, 4719–4734. [CrossRef] [PubMed]
23. Wang, L.; Ye, X.; Wang, S.; Li, P. ULO: An Underwater Light-Weight Object Detector for Edge Computing. *Machines* **2022**, *10*, 629. [CrossRef]
24. Chen, L.; Jiang, Z.; Tong, L.; Liu, Z.; Zhao, A. Perceptual Underwater Image Enhancement With Deep Learning and Physical Priors. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3078–3092. [CrossRef]
25. 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]
26. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
27. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
28. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.

29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

30. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]

31. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

32. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.

33. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.

34. Cui, Z.; Qi, G.J.; Gu, L.; You, S.; Zhang, Z.; Harada, T. Multitask aet with orthogonal tangent regularity for dark object detection. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 2553–2562.

35. Chiang, J.Y.; Chen, Y.C. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* **2012**, *21*, 1756–1769. [CrossRef]

36. Li, C.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [CrossRef]

37. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.

38. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 558–567.

39. 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, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.

40. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.

41. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 13039–13048.

42. Zhang, W.; Zhuang, P.; Sun, H.H.; Li, G.; Kwong, S.; Li, C. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Trans. Image Process.* **2022**, *31*, 3997–4010. [CrossRef]

43. Islam, M.J.; Xia, Y.; Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]

44. Zhang, J.; Zhu, L.; Xu, L.; Xie, Q. Research on the Correlation between Image Enhancement and Underwater Object Detection. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 5928–5933.