machines

MDPI

Article

A Deep Learning Model Applied to Optical Image Target Detection and Recognition for the Identification of Underwater Biostructures

Huilin Ge, Yuewei Dai *, Zhiyu Zhu and Runbang Liu

School of Electronic Information, Jiangsu University of Science and Technology, Zhenjiang 212003, China
* Correspondence: dyw@nuist.edu.cn

Abstract: Objective: We propose a deep-learning-based underwater target detection system that can effectively solve the problem of underwater optical image target detection and recognition. Methods: In this paper, based on the depth of the underwater optical image target detection and recognition and using a learning model, we put forward corresponding solutions using the concept of style migration solutions, such as training samples. A lack of variability and poor generalization of practical applications presents a challenge for underwater object identification. The UW_YOLOv3 lightweight model was proposed to solve the problems of calculating energy consumption and storage resource limitations in underwater application scenarios. The detection and recognition module, based on deep learning, can deal with the degradation process of underwater imaging by embedding an image enhancement module into the detection and recognition module for the joint tuning and transferring of knowledge. Results: The detection accuracy of the UW_YOLOv3 model designed in this paper outperformed the lightweight algorithm YOLOV3-TINY by 7.9% at the same image scale input. Compared with other large algorithms, the detection accuracy was lower, but the detection speed was much higher. Compared with the SSD algorithm, the detection accuracy was only 4.7 lower; the speed was 40.9 FPS higher; and the rate was nearly 16 times higher than Faster R-CNN. When the input scale was 224, although part of the accuracy was lost, the detection speed doubled, reaching 156.9 FPS. Conclusion: Based on our framework, the problem of underwater optical image target detection and recognition can be effectively solved. Relevant studies have not only enriched the theory of target detection and glory, but have also provided optical glasses with a clear vision for appropriate underwater application systems.

Keywords: underwater imaging; deep learning; object detection; image enhancement; UW_YOLOv3

Citation: Ge, H.; Dai, Y.; Zhu, Z.; Liu, R. A Deep Learning Model Applied to Optical Image Target Detection and Recognition for the Identification of Underwater Biostructures. *Machines* **2022**, *10*, 809. https://doi.org/10.3390/machines10090809

Academic Editors: Kelvin K.L. Wong, Dhanjoo N. Ghista, Andrew W.H. Ip and W.J. (Chris) Zhang

Received: 25 July 2022
Accepted: 29 August 2022
Published: 15 September 2022

1. Introduction

Underwater target detection tasks can be divided into two categories according to the different signals of the target to be detected [1]. The first category uses acoustic images collected by sonar to detect underwater targets, which is only suitable for the long-distance detection and tracking of large targets [2]. The second type is underwater target detection based on the optical image of a machine-vision system.

Visual images have advantages in short-range underwater target detection, with a high resolution and rich information. Therefore, target detection based on light vision has gradually become the leading research direction of underwater short-range target recognition and detection [3]. To accurately identify a target, the key is to determine the category and location of the underwater target. The most direct method is to collect images through underwater cameras and implement detection through a deployed underwater target-detection algorithm. However, shallow aquatic environments are complex and often lead to problems such as color shifts, uneven illumination, blurring, and distortion in the imaging process. These scenarios are very unfavorable for the results of the detection

network. In addition, due to the existence of image domain offset, it is difficult for general object-detection algorithms to maintain high robustness in underwater environments.

An autonomous underwater vehicle (AUV) is the most advanced aquatic monitoring and operation equipment, which can replace humans to complete specific tasks in complex underwater environments [4]. To conduct underwater monitoring and operation tasks, an AUV must quickly perceive the complex aquatic environment through the visual system and accurately identify the target of interest. However, due to the attenuation of light at different degrees in the underwater propagation process, the collected underwater images generally have problems such as unclear low illumination and color distortion, making the research of underwater target detection technology face many difficulties.

The underwater target detection process can be divided into three stages: underwater image acquisition, image feature extraction, and target recognition [5]. Traditional target detection algorithms generally use manual methods to extract image features, which is cumbersome; the extracted adequate image feature information is not rich enough, resulting in an extremely low detection accuracy. In recent years, deep learning has been developed [6–10], the excellent feature extraction ability of which can achieve recognition and detection accuracy that traditional methods cannot. Deep learning also has a strong migration ability for obtaining features [11]. It has a universality and a strong generalization ability for the feature extraction of targets in various fields. Therefore, applying deep learning to underwater target detection technology to reduce the impact of uncertain aquatic environments, and to improve underwater image recognition and detection performance, is a hot topic in underwater target technology.

Based on the above analysis, it can be seen that effectively solving the problem of underwater image degradation and introducing deep learning methods to improve the detection of underwater targets are vital points for breaking through the current development status of underwater target detection technology. Meanwhile, this area of research is also of great significance to AUV development.

2. Related Work

2.1. Underwater Image Enhancement

In underwater propagation, the light attenuates to different degrees due to water quality, leading to problems such as unclear low illumination and color distortion in collected underwater images [12]. These degraded underwater images will significantly affect the feature extraction process of the deep neural network, thus making it difficult to detect underwater targets.

In 2017, Perez et al. [13] proposed a deep-learning method for underwater image defogging. This method first uses an image-recovery algorithm to process underwater photos, and then trains the deep neural network on the original and processed images. Finally, the trained network is used to process other fuzzy underwater images. Scholars have used a convolutional neural network model to learn the process of aquatic image degradation. This method can enhance unclear and biased images, but the dataset used for training was artificially stimulated, and thus different from the actual underactuated image degradation process. In the same year, Fabbri et al. [14] proposed a generative network for image enhancement that could improve the sharpness of underwater images and restore actual color. Still, it had strict requirements for the quality of the training dataset [15].

2.1.1. Underwater Target Detection

In recent years, with the rapid development of AUVs, underwater short-range target detection technology based on vision systems has become more critical, and scholars have conducted many studies. In 2007, Yamashita et al. [16] established a model for the influence of the underwater environment on color to judge typical underwater artificial targets through color information. Still, this method did not consider the similarity between continuous sample data. In 2011, Mukherjee et al. [17] proposed an underwater target recognition method based on a boundary. However, it achieved better results only when

the color of the target of interest differed significantly from that of its surroundings. In the following year, Elberink et al. [18] proposed a method for detecting the reconstructed target shape of lines. The experimental results showed that the accuracy of non-noise image detection was 96% and that of noise image detection was 88%, but the algorithm was only able to detect artificial targets with relatively simple shapes, and was not able to be applied to biological targets with complex conditions. In 2014, Hsiao et al. [19] proposed an underwater fish-recognition framework composed of unsupervised feature learning and fault-tolerant detection, which had a high detection accuracy for imbalanced underwater fish images.

2.1.2. Target Detection Algorithm Based on Deep Learning

After the emergence of a deep-learning network, a multi-layer neural network could be used to fully extract target image features and learn abstract deep representation information to obtain more accurate recognition and detection results. Region-based recommendations and regression-based methods have become the two most commonly used target detection methods.

(1) The R-CNN-series algorithm is a classical algorithm based on a recommended region. It first estimates the last frame containing the target through the region recommendation method, then uses CNN to perform a feature extraction operation, and finally inputs the data to the detector for classification and positioning. In 2014, Girshick et al. [20] proposed the R-CNN detection algorithm [21], which firstly uses target detection and the segmentation of a deep neural network CNN and then applies transfer learning to improve network performance. Girshick et al. [22] further improved the R-CNN network. They proposed a FASTR-CNN network, which combined feature extraction and detection with ROI and a multi-task loss function; they also tested its speed and found that the training speed was significantly improved. In the same year, Ren Shaoqing et al. [23] proposed the well-known Faster R-CNN, which offered a region proposal network (RPN) instead of the traditional region selection method. The feature graph after convolution was shared and integrated into a network. The end-to-end training of the detection algorithm was realized for the first time.

(2) Regression-based detection algorithms mainly include the SSD series and YOLO series, which directly set the last frame on the input image and perform regression operations on the target in stand-through feature extraction. In 2016, Redmon et al. proposed YOLO, a single pipeline network viewed only once, to directly perform regression operations on BBO in the grid and predict the target's coordinate information and category probability. This network dramatically improved the speed of target detection, reaching a real-time detection rate of 45 FPS. In the following year, Redmon et al. made improvements based on the YOLO algorithm and proposed the YOLOv2 algorithm, which introduced batch normalization (BN) [24]. The anchor frame mechanism and the pass-through operation improved the detection performance of the network. In 2018, Redmon et al. [25] proposed the YOLOv3 algorithm based on YOLOv2. YOLOv3 used DarkNet-53 as the master of the thousand networks, which is composed of multiple ResNet stacks, making the depth of the entire network up to 152 layers. Moreover, multi-scale fusion was adopted in the prediction network, which further improved the feature extraction capability of the network.

2.2. Our Contributions

The research content of this paper mainly includes three parts:

The first part proposes an underwater image enhancement method based on the generative adversarial network for aquatic image degradation.

In the second part, which aims to solve the problem of the energy consumption calculation and storage resource limitations in underwater application scenarios, a lightweight model based on YOLOv3 is proposed for lightweight underwater detection.

In the third part, which aims to solve the problems of insufficient training samples, a lack of variability, and the flawed generalization of practical applications, a data network parameter transfer based on transfer learning is proposed.

3. Methods and Materials

3.1. Underwater Image Enhancement Based on a Generative Adversarial Network

Establishment of Underwater-Style Transfer Dataset

The lack of an underwater paired image dataset restricts the application of image enhancement algorithms based on deep learning in underwater scenes. This dataset needs to collect images of underwater targets in two states (with and without water), so it is difficult to organize data that meet the requirements in actual scenes. Therefore, it is possible to use a theory based on multi-scale retinex (MSR), multi-scale retinex with color restoration (MSRCR), and automated multi-scale retinex with color restoration (AMSRCR). The underwater image is enhanced by combining the algorithm with a dehaze net algorithm based on deep learning to establish an underwater-style transfer dataset, as shown in Figure 1. Due to the traditional manual method, the generalization ability is terrible, so after using the conventional method of image enhancement processing, part of the image experiences color distortion. The phenomenon of supersaturation cannot be directly used to train convolutional neural networks, as they require further screening of the enhanced image. The enhancement of the underwater photos in pairs with a better effect was chosen as the final training for the underwater image enhancement of the network dataset.

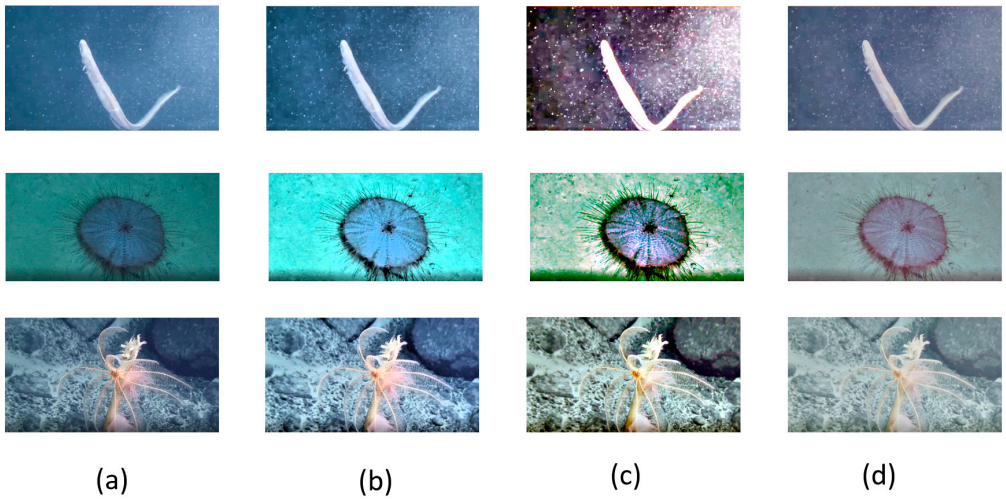

Figure 1. Image enhancement result map that relies on retinex theory based on (**a**) raw image, (**b**) MSR, (**c**) AMSRCR, and (**d**) MSRCR.

Generative adversarial networks (GAN) make up one of the research hotspots in computer vision. Confrontation refers to the conflict between a generator and a discriminant. The generator generates realistic samples as much as possible, judging the device as far as possible to identify the authenticity of the sample through the constant iterative training of the generator and discriminant criteria. In this way, the thinking ability is continuously strengthened and the samples from the generator become more accurate, ultimately achieving improvements in the performance. GAN has been applied in many fields in recent years, such as image style transfer, image super-resolution reconstruction, image repair, image segmentation, etc. The traditional image enhancement algorithm is based on GAN. Image enhancement does not require prior knowledge, and the network can automatically learn the data distribution of samples.

By GAN's successful application in the field of migration image style, Section 3.1 of this paper examined underwater images of the dataset in pairs, divided into two styles of analysis: a technique for green underwater blurred images and a method for color

averaging of underwater clear images. Then, by using the concept of image transfer, Pix2Pix underwater image enhancement was realized. This represented an improvement in the network generation conditions. The input of network pairs of images is mainly used to complete a transformation between vision and the idea of image translation. For the image style, the migration effect has proven to be an excellent application. Therefore, this paper used the Pix2Pix network model as the basis for underwater image enhancement network architecture. The underwater image enhancement network designed in this paper also included two parts: a generator and a discriminator. The network structure is shown in Figure 2. The design of this network's generator discriminator and loss function are introduced.

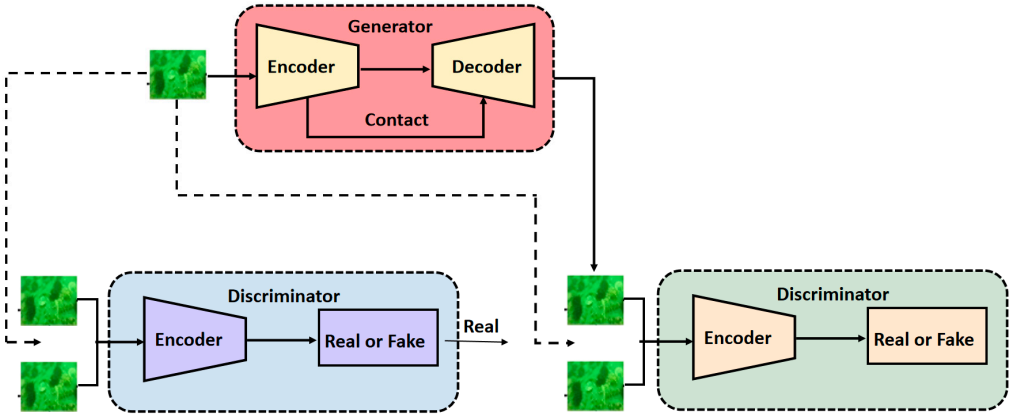

Figure 2. Schematic diagram of underwater image enhancement network.

Firstly, the generator generated underwater high-quality images from an input of underwater degraded images. The concrete structure is shown in Figure 3. The generator used a system containing an encoder and a decoder, a reference for the design of the network model for the U-Net network, and a network for the convolutional neural network; the input and output were the three RGB channels. The overall structure of the image and the network was composed of two symmetric parts of down-sampling and up-sampling. The basic unit structure of the down-sampling stage was the convolution batch regularization ReLU. The sampling phase structure and the similar deconvolution replaced the convolution operation sampling, and the sampling under corresponding parts used the same number of convolution kernels. Through the jump layer connection manner, which lowered the amount of information fusion and high-level information, the details of the pixel were kept under different resolutions to a certain extent, thus improving the generation of the image details.

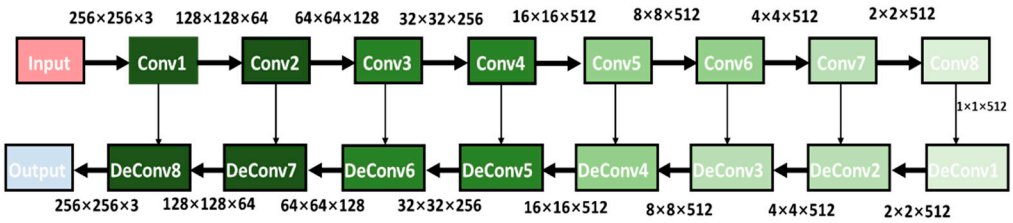

Figure 3. Underwater image enhancement network generator. The generator used a system with an encoder and decoder, a reference for the design of the network model for the U-Net network, and a network for the convolutional neural network; the input and output were the three RGB channels.

The second part was the discriminator part, the function of which was to judge whether the image input by the network was the accurate data or the data generated by the generator. The specific structure of the discriminator part is shown in Figure 4. The basic unit structure of the discriminator part was the same as that of the generator part,

and the input of the discriminant network was two 3-channel RGBs. The positive sample of the network was the image pair, composed of the input image and its corresponding truth value, and the negative sample was the image pair consisting of the input image and the image generated by the generator.

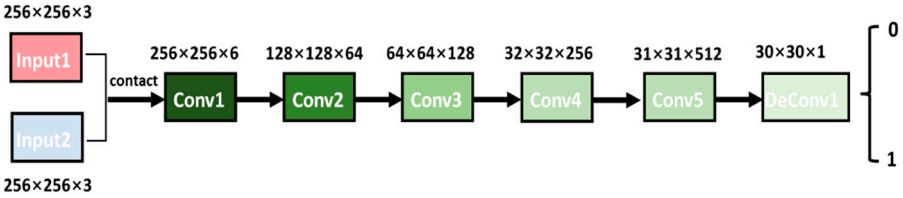

Figure 4. Underwater image enhancement network discriminator. The basic unit structure of the discriminator part was the same as that of the generator part, and the input of the discriminant network was two 3-channel RGBs.

The loss function of the network consisted of three parts: the conditional admixture loss, the reconstruction loss, and the perceived loss in a generative admixture network; the goal of the loss function was to make the distribution of the generated samples and actual samples as close as possible. The Pix2Pix network is a variant of a conditional GAN. The loss function is shown in Equation (1):

$$L_{cGAN}(G, D) = E_{x,y}[logD(x, y)] + E_{x,z}[log(1 - D(x, G(x, z)))] \quad (1)$$

where x is an underwater degradation image; y is a clear underwater image; and Z is random noise.

The goal of generator G was to learn the mapping relationship between the sample space X and the sample space Y. The purpose of the discriminator D was to identify whether the image was an actual image as accurately as possible. If the input of the discriminator was a virtual image, the output of the discriminator was 1; otherwise, the result was 0. The goal was to minimize the loss function and maximize the loss function of D, thus achieving learning. Besides the conditions against loss, the loss function of the network design included refactoring losses to constrain the similarity between the input image and the generated picture; the most commonly used refactoring losses include the L1 and L2 equidistance measure functions. Studies have shown that images with the L1 distance constraint, based on Laplacian priors, are more transparent; therefore, the L1 distance, as shown in Equation (2), was used as the reconstruction loss:

$$L_{L_1}(G) = E_{x,y,z}[[\| y - G(x, z) \|]_1] \quad (2)$$

To preserve the details of the generated image and prevent the loss of the textural information of the image, perceptual loss was introduced into the loss function. Perceptual loss is not a constraint on the pixel level of the image, but rather a constraint on the high-level semantic information of the image. The perceptual loss was calculated by the generated image and the truth value in the pre-trained VGG16. The distance between feature maps generated at specific layers of the network and the perceived loss function are defined in Equation (3):

$$L_{Perceptual}(G) = \| \varphi(x) - \varphi(G(x)) \|_2^2 \quad (3)$$

Therefore, the optimization objective function of the final underwater image enhancement network is shown in Equation (4), where λ and μ are the weights of the L_1 loss and perception loss, respectively:

$$G^* = arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L_1}(G) + \mu L_{Perceptual}(G) \quad (4)$$

where φ is the pretrained feature mapping for the RELU4_2 layer of the VGG16 networks on ImageNet; λ is the L_1 loss weight; and μ is the weight of the perceived loss.

3.2. YOLOv3-Based Lightweight Detection Model

Research has shown that the YOLOv3 algorithm can maintain accuracy while ensuring a fast detection speed. However, underwater robots are generally equipped with low-performance embedded computing equipment, so it is difficult for this algorithm to meet the real-time detection requirements. Therefore, this chapter reconstructs the backbone network and prediction network of the YOLOv3 algorithm and designs a lightweight model UW_YOLOV3 for the real-time detection of underwater targets. The experiments showed that the model had a good detection performance and met practical engineering applications.

3.2.1. UW_YOLOv3 Trunk Network Establishment

To carry out real-time aquatic product detection on the embedded device Jeston T 2, this section designs a lightweight detection model based on YOLOv3. Table 1 shows the number of parameters and the calculation amount of the YOLOv3 trunk network, DarkNet-53. The table only considers the parameters of the convolution operation. It can be seen that the total number of parameters was over 41 million, and the calculation amount was 48.9 bf. However, the total number of parameters in the YOLOv3 model was about 62 million, and the number of calculations was 65.426 bf. After the analysis, it can be seen that the number of parameters and the amount of research in the trunk network accounted for 66% and 75% of the overall model, respectively. It can be seen that most of the calculations of YOLOv3 were concentrated in the trunk network. Based on this, the leading network and prediction network of the YOLOv3 model were improved in this section to reduce the parameters and computations of the model and maintain the detection performance with a guaranteed accuracy. The improved lightweight model was named UW_YOLOv3.

Table 1. Number and computations of DarkNet-53 parameters.

Type	Filter	Size	Output	Calculated Quantities/10^6	Number of Arguments
Conv2d	32	3 × 3	416 × 416	299.04	864
Conv2d	64	3 × 3/2	208 × 208	1594.88	18,432
Conv2d	32	1 × 1			
Conv2d	64	3 × 3		×1 = 1772.09	×1 = 20,480
Residual			208 × 208		
Conv2d	128	3 × 3/2	104 × 104	1594.88	73,728
Conv2d	64	1 × 1			
Conv2d	128	3 × 3		×2 = 3544.18	×2 = 163,840
Residual			104 × 104		
Conv2d	256	3 × 3/2	52 × 52	1594.88	294,912
Conv2d	128	1 × 1			
Conv2d	256	3 × 3		×8 = 14,176.00	×8 = 2,621,440
Residual			52 × 52		
Conv2d	512	3 × 3/2	26 × 26	1594.88	1,179,648
Conv2d	256	1 × 1			
Conv2d	512	3 × 3		×8 = 14,176.00	×8 = 10,485,760
Residual			26 × 26		
Conv2d	1024	3 × 3/2	13 × 13	1594.88	4,817,592
Conv2d	512	1 × 1			
Conv2d	1024	3 × 3		×4 = 7088.04	×4 = 20,971,520
Residual			13 × 13		
Sum	DarkNet-53 structure			48,960.11	40,549,216

The YOLOv3 backbone network had a total of 52 convolution layers, was made up of 23 pairs in 1 × 1 and 3 × 3 convolutions of a residual block, and used five convolution

layers for the sampling operation. The 1 × 1 channel convolution was mainly used for compression characteristics. To reduce the model size, a 3 × 3 traditional convolution was used to extract the feature, while at the same time extending the output channel number. The number of convolutional layers in the UW_YOLOv3 trunk network designed in this paper was reduced, as shown in Figure 5. The trunk network had only 20 layers in total. Secondly, the convolution mode was improved. The original standard 3 × 3 convolution was replaced by the improved depth-separable network module RI-DSC. In addition, the Conv module with a step size of 2 was used for the down-sampling operation, which further strengthened the information communication ability between the channels of the feature graph. The above process reduced the network parameters and calculation amount and maintained specific feature extraction abilities.

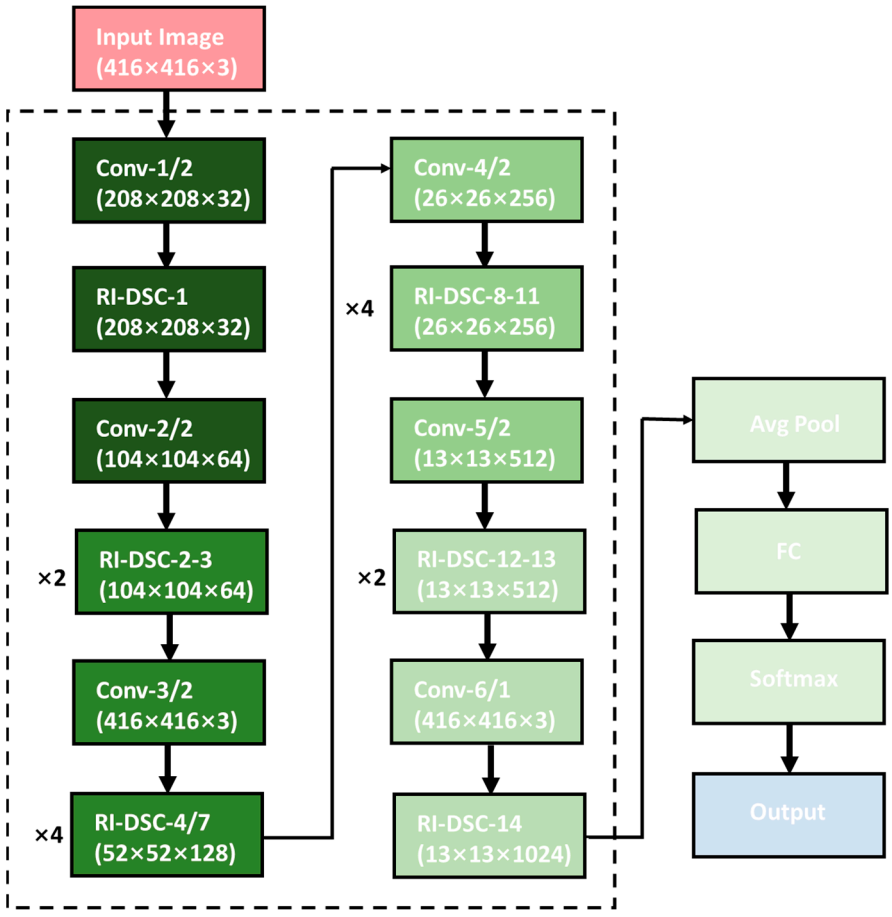

Figure 5. UW_YOLOv3 network structure diagram. The shape of the output feature graph of the layer is in parentheses for each module.

In Figure 5, the shape of the output feature graph for the layer is in parentheses for each module. The number on the left of the module represents how many times the module was repeated. Crow-1 represents the first lower sampling layer of the image input. The scale transformation was realized by extracting features with a step size of 2 through standard convolution. After each Conv module, the depth of the feature graph doubled, but it did not increase in the RI-DSC module. The specific structural parameters of the UW_YOLOv3 trunk network are shown in Table 2, where the filter shape represents the width, height, and depth of the convolution kernel, which were 1, 1, and 3 × 3, respectively. The parentheses after the convolution represent the number of convolution kernels. Since multi-scale convolution kernels were used in this paper, the output channel of the features of the layer was equal to the sum of the number of the two convolution kernels. By comparing the UW_YOLOv3 and YOLOv3 trunk networks, it can be seen that the parameter number and

computation amount of the main thousand networks after lightweight were significantly reduced by orders of magnitude. The parameter number and computation amount were only 20.3% and 9.93% of the YOLOv3 trunk network, respectively, while the effective 3×3 convolution layer of feature extraction was only reduced to layer 20 and the 1×1 convolution layer of the compression channel was canceled.

Table 2. Comparison of quantitative results of underwater image enhancement.

Image	Original Image	MSRCR	MSRCR + DehazeNet	Methods
Entropy	6.35	6.48	7.18	7.26
Standard deviation	20.4	23.6	36.0	38.4

3.2.2. UW_YOLOv3 Prediction Network Establishment

A high-resolution net (HRNet) is a network that can maintain high-resolution features throughout the whole process. The network can support a large resolution feature map through the same operation and use it as the leading network. Then, low-resolution sub-networks are gradually added in parallel to the top grid, and the parallel networks are connected to form the feature of multi-scale fusion. In this way, semantic information from different low-resolution parts can be received in the main primary work to improve the representation ability of the large-scale aspects of the network. As a result, the predicted key features are more spatially accurate [17].

Figure 6 is a schematic diagram of the HDNet network model structure. The abscissa represents the depth of the network and the ordinate represents the width of the network. The sub-networks of the second and third lines were low-resolution networks built in parallel in the leading network process, and the information exchange between networks of different scales was realized through up-sampling and down-sampling operations. The features of different scales were fused many times.

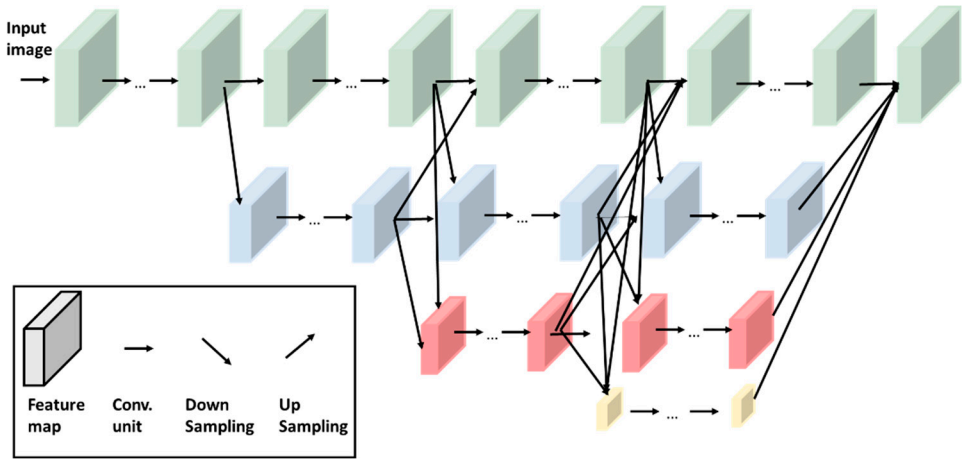

Figure 6. Schematic diagrams of HRNet's network structure.

The network had two advantages:

(1) HRNet can always maintain high-resolution features and will not lose feature details due to down-sampling operations, nor will it be unable to fully represent all upper-level information due to up-sampling recovery features.

(2) The HRNet feature fusion method can make predictions more accurate. In this network, different scale fusions are used many times, and low-resolution high-level semantic information is used to improve the capability of the high-resolution feature representation.

To improve the detection accuracy of the network, this paper improved the original feature fusion part of the network, introduced the HRNet network to increase the capability of high-resolution feature characterization, and realized multi-scale prediction from the output of low-resolution features. As shown in Figure 7 for the improved feature fusion part, components with scales between 13, 26, and 52 were connected in parallel. DSC was used at the same-scale layer to amplify the network depth. When multiple features were combined, they were stacked along the direction of the depth of the feature map, and then the 1×1 convolution was adopted to carry out the channel fusion. Finally, the multi-scale prediction was retained with the YOLOV3 sample.

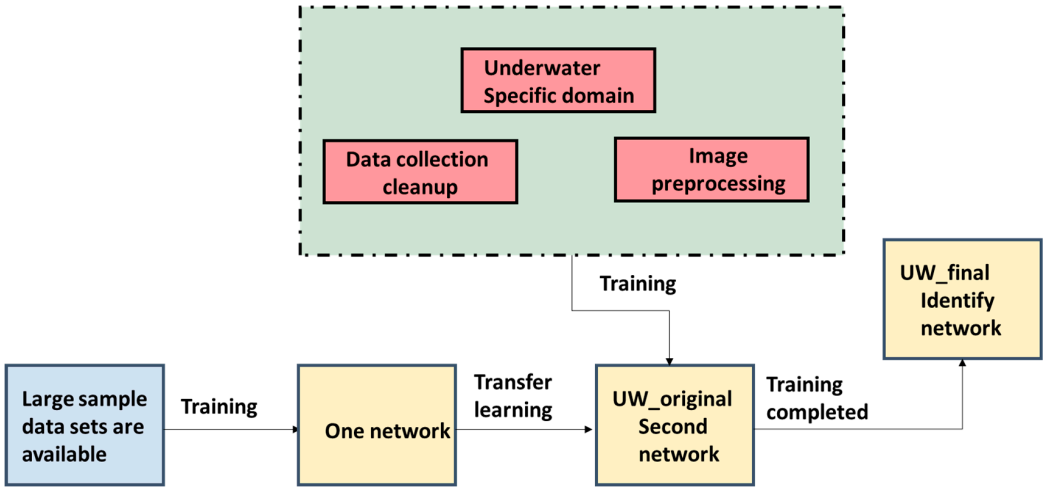

Figure 7. Underwater target detection network structure. The Faster R-CNN network used in this paper adopted the VGG16 classification network as its backbone.

3.2.3. UW_YOLOv3 Network Construction

In this paper, the UW_YOLOv3 network model was built by combining the designed backbone network and the prediction network. The prediction part first used convolution for scale adjustment. The characteristic output depth of the last layer was required to conform to the detection principles of the YOLOv3 algorithm, with a depth = $3 \times (5 + \text{Len (class)})$. The overall network structure was improved as follows compared with the YOLOv3 network:

(1) The standard 3×3 convolution was replaced with a profoundly separable network, significantly compressing the network model;
(2) Inception was introduced for multi-scale feature extraction to improve the feature-extraction capability of the convolutional layer;
(3) A reset structure was adopted so that the network only learned residuals and speed-up training;
(4) The parallel connection mode of HRNet was used to improve the expression ability of the high-resolution features of the network.

3.3. Data Network Parameter Transfer Based on Transfer Learning

3.3.1. Transfer Learning

The concept of transfer learning was born in the NIPS seminar in 1995 and has attracted extensive attention from experts and scholars at home and abroad. Its main idea is to transfer the knowledge learned from one or more source fields with a large amount of data to another target field with a small amount of data. Transfer learning has been comprehensively reviewed and is divided into three types: inductive transfer learning, direct transfer learning, and unsupervised transfer learning [22], wherein the data of the source domain and target domain of the inductive transfer learning are labeled, but the tasks of the source domain and target domain are different. The source domain of direct transfer learning has labels, while the target domain does not; however, the missions of the source

domain and target domain are the same. In machine learning, inductive transfer learning is a widely used transfer-learning method, which is also divided into instance-based, feature-based, parameter-based, and knowledge-based transfers. Parameter transfers in inductive transfer learning are the most used in deep-learning-based object detection.

3.3.2. Design and Implementation of Parameter Transfer Algorithm

In computer vision, the most influential extensive sample dataset is the ImageNet image classification dataset, which contains about 1.2 million training images, 50,000 verification images, and 100,000 test images. There are 1000 different categories. Since target detection tasks and image classification have certain commonalities, target detection networks usually use an image classification network as their backbone, as shown in Figure 7. The Faster R-CNN network used in this paper adopted the VGG16 classification network as its backbone. Because the labeling cost of object detection data is much higher than that of image classification data, the number of general object detection datasets (such as the Pascal VOC and Microsoft COCO dataset) is much smaller than the number of ImageNet image classification datasets. At present, almost all target detection algorithms based on depth studies will advance using the ImageNet dataset classification task of training. Using trained model parameters as the initial weights of target detection network avoids overfitting the problem of small sample data and can, to a certain extent, improve the convergence speed and precision of the models.

However, the ImageNet dataset categories are common categories of raw images taken on land, such as cars and cats, which are different from the target domain of underwater target detection problems. Therefore, using VGG16 for the parameter initialization of an underwater target detection network trained on ImageNet does not help improve the detection accuracy of submerged targets. To solve this problem, this section proposes specific steps of parameter migration based on the underwater target detection network as follows:

(1) An underwater biological classification dataset was constructed. The acquisition of tags in the biological classification dataset was much easier than in the underwater target detection dataset. There were many seafood image data with corresponding labels on the Internet. In this paper, many sea cucumbers and sea urchins were extracted from Google and Baidu, respectively. Crawling scallop images were screened. Finally, a dataset of three seafood categories was obtained, including 10,230 images: 3105 sea urchin images, 3715 sea cucumber images, and 3410 scallop images.
(2) The parameter migration of the ImageNet classification network was due to VGG16. The parameter space was ample, and it was easy to overfit and slow to converge when using the constructed seafood classification dataset to train the VGG16 network directly. Therefore, this paper firstly initialized the seafood classification network by using the parameters obtained by VGG16 training on the ImageNet dataset, and then carried out the first parameter migration.
(3) The network was fine-tuned. The parameters of the first 15 layers of the initialized VGG16 network were fixed and the last layer was fine-tuned.
(4) Several iterations were performed on the seafood dataset, restoring the learning rate of the first 15 layers and the trained VGG16.
(5) Parameter migration was performed for the marine classification network. The model parameters obtained in the previous step initialized the backbone network of the underwater target detection network and carried out the second parameter migration.

4. Experiments and Results

4.1. Underwater Image Enhancement Based on a Generative Adversarial Network

The dataset used in the experiment came from the offline target detection group of the Underwater Robot Target Capturing Contest (URPC2017). The dataset had 19,967 underwater images; the proposed method to deal with the dataset was to build an underwater-style migration dataset, deal with the filtered data, and choose 1500 underwater images. The

underwater-style migration dataset is established in this paper. In this chapter, through the established underwater-style migration datasets, 375 images from 18,467 photos in the URPC2017 dataset that did not participate in underwater image enhancement network training were randomly selected for testing in order to train the network on underwater image enhancement.

When training the underwater image enhancement network based on an adjunctive generative network, the image input size was set to 256 × 256; the initial learning rate was 0.0002; the number of network training samples each time was 1; the weight coefficient of L1 loss and the weight coefficient of perceived loss were set to 120 and 0.0001, respectively; and the Adam optimizer was used with a momentum of 0.5 for gradient updates.

Table 2 compares the quantitative results of different underwater image enhancement methods on the test dataset. In the table, the data with the best performance are bolded. As the original image had poor detail and low contrast, its entropy and standard deviation were small. The methods involved, to a certain extent, improving the quality of the image, the image's entropy, and the traditional deviation values; however, this paper adopted the MSRCR DehazeNet method based on further enhancing the image entropy and standard deviation values, which showed that the proposed underwater image enhancement method resulted in better image texture retention and image contrast enhancement performance.

This paper adopted the alternate training method. To further visualize the loss during the training process, loss curves of four training stages of the model were drawn, as shown in Figure 8. Note that Figure 8a–d represent YOLOv3-416, YOLOv3-tiny-416, UW_YOLOv3-416, and UW_YOLOv3-224, respectively.

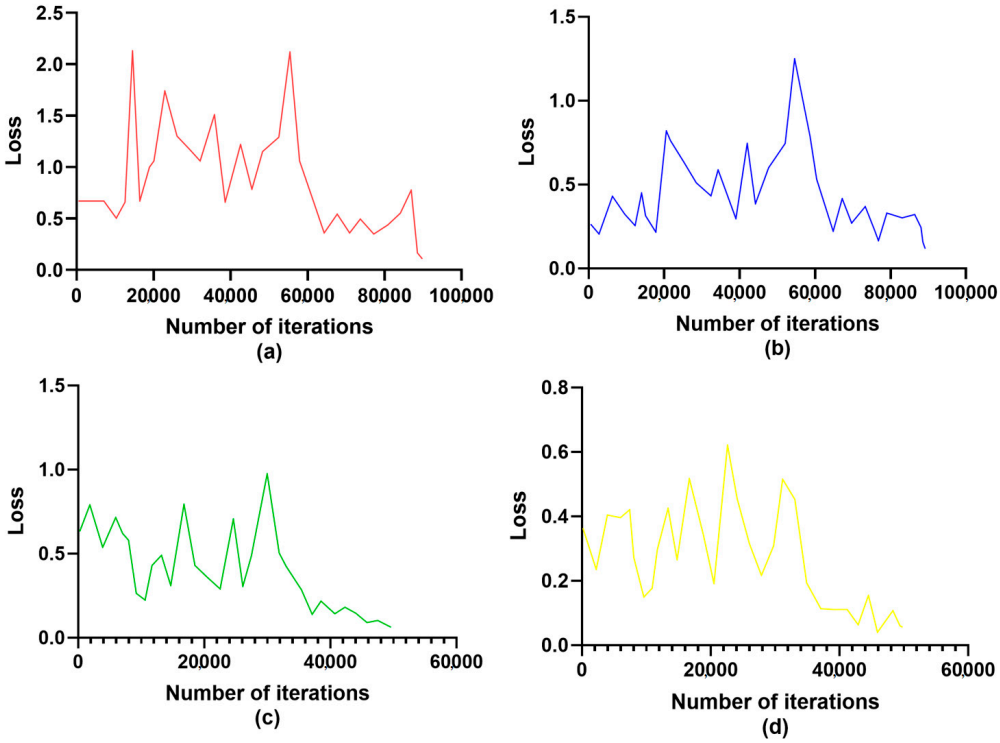

Figure 8. Loss curves of four training stages of the model (**a**–**d**), representing YOLOv3-416, YOLOv3-tiny-416, UW_YOLOv3-416, and UW_YOLOv3-224, respectively.

4.2. Yolov3-Based Lightweight Detection Model

As shown in Table 3, the detection accuracy of the UW_YOLOv3 designed in this paper improved on the lightweight algorithm YOLOV3-TINY by 7.9% at the same image scale input. Compared with other large algorithms, the detection accuracy was lower, but the detection speed was much higher. Compared with the SSD algorithm, the detection accuracy was only 4.7 times lower; the speed was 40.9 FPS higher; and the rate was nearly

16 times higher than Faster R-CNN. When the input scale was 224, although part of the accuracy was lost, the detection speed doubled, reaching 156.9 FPS. Therefore, it can be seen that the lightweight network designed in this paper was able to maintain a guaranteed detection accuracy and a high speed, basically meeting the real-time detection requirements for embedded devices.

Table 3. Comparison of the detection effects of different algorithms.

Algorithm	Trunk Network	AP50	AP_{75}	FPS
Faster RCNN-600	ResNcl-101	75.6	62.7	5.1
SSD-512	VGGNet-16	66.5	51.6	39.3
YOLOv3-416	DarkNet-53	71.9	53.5	48.7
YOLOv3-tiny-4l6	DarkNet-tiny	53.9	38.1	96.5
UW_YOLOv3-4l6	RI-DSC	61.8	49.3	80.2
UW_YOLOv3-224	RI-DSC	54.1	39.7	156.9

4.3. Data Network Parameter Transfer Based on Transfer Learning

The dataset used in the experiment came from the data of the offline object detection group of the Underwater Robot Target Picking Contest (URPC2018). The dataset included four categories: sea cucumbers, sea urchins, scallops, and starfish, with 3701 underwater images, including 2901 training data points and 800 test data points.

Figure 9 shows the comparison of the mAP curve trends for parameter migration based on the seafood classification network and parameter migration based on the ImageNet classification network for target detection networks in small-sample scenarios, with an increase in iteration times during the training process. The network parameters of migration, which were found for the marine classification, not only sped up the convergence rate to save network training time, but also improved the accuracy of the detection network, which enhanced the performance of the target detection. Table 4 shows that after 70,000 iterations, the different target detection models for different categories of detection precision were based on ImageNet only. When a parameter migration of the classification network was carried out, the average accuracies of the detection network for sea urchins, scallops, and sea cucumbers were 84.6%, 44.7%, and 64.9%, respectively, and the overall average accuracy was 69.7%. When a parameter migration based on the seafood classification network was carried out, the overall average accuracy of the detection network for sea urchins, scallops, and sea cucumbers was 84.8%. The average accuracies for sea urchins, scallops, and sea cucumbers were increased by 0.2%, 3.7% and 4.5%, respectively, compared with the former.

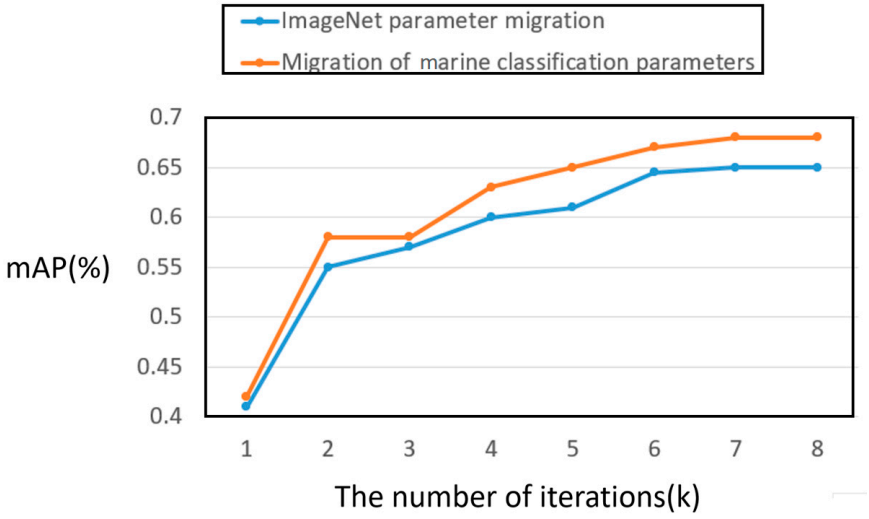

Figure 9. Comparison of MAP curves for target detection, with or without parameter migration, in a small-sample scenario.

Table 4. Accuracy comparison of detection network, with or without parameter migration, in different categories.

Method	Sea Urchin	Scallop	Sea Cucumber	MAP (%)
ImageNet Parameters of the Migration	84.6	44.7	64.9	64.7
Classification of Seafood Parameters of the Migration	84.8	48. 4	69.4	67.5

Based on the classification of marine network parameters, the use of migration to improve the detection precision of the target model was especially important for the detection of scallops and sea cucumbers, but not for sea urchins. This is mainly because sea urchins are characteristically bright, making it easy to distinguish the original network once the identification precision of sea urchins reaches a reasonable level. Hence, the parameters of the transfer operation through the marine classification network for the ascension of the detection accuracy were not significant. Still, for scallops, parameter migration can transfer the knowledge learned in the classification network to the detection network to improve the detection ability for scallops and sea cucumbers in the detection network.

5. Discussion

In this paper, the problem of underwater target detection based on deep learning was studied. The main factors restricting the improvement of underwater target detection accuracy were analyzed in depth. The research results are as follows:

(1) To solve the problem of underwater image degradation, an underwater image enhancement method based on a generative adventure network was proposed and implemented to establish an underwater-style migration dataset by combining MSRCR and DehazeNet.

(2) Aiming to solve the problem of insufficient real-time detection of the YOLOv3 algorithm in embedded devices, the lightweight network model UW_YOLOv3 was designed by improving the structure of its backbone network and predictive network. In the trunk network, a deep convolution separable network was introduced to replace the standard convolution, a 1×1 convolution was introduced to increase the network width, and a 20-layer trunk network was built by using this unit to replace DarkNet-53. This was introduced to the predictive network to improve the traditional feature fusion method so that the network always maintained the capability of high-resolution feature representation. The experimental results showed that the number of parameters of the improved backbone network was only 20% of the original, which achieved the purpose of being lightweight, and the detection speed doubled to 98.1 FPS while maintaining a detection accuracy of 58.1%. The detection accuracy of the improved predictive network was improved by 3.7%. Compared with other algorithms, the lightweight UW_YOLOv3 network designed in this paper dramatically improved the detection speed at the expense of a small part of the accuracy, and its 61.8% accuracy and 80.2 FPS detection speed met the requirements of practical engineering applications.

(3) Given the small in-sample problem caused by the difficulty of the large-scale acquisition of underwater images, a solid supervised underwater target detection method was proposed and implemented in a small-sample scenario. The idea of transfer learning was used to realize the transfer of seafood classification knowledge to underwater target detection knowledge and improve the convergence speed and detection accuracy of the detection network. The data augmentation of an underwater image was completed using a spatial variation network, which strengthened the robustness of the detection network in the targeting of spatial transformation. This solved the problem of the network model being easy to overfit under the condition of small samples.

6. Conclusions

This paper conducted in-depth research on underwater target detection methods based on deep learning. From data enhancement, to network model design, to embedded device realization, good results were achieved. However, there are still some shortcomings, which in-depth glasses could further improve:

(1) Due to the influence of a complex underwater environment, accurate underwater image data are seriously lacking. Although data amplification can increase a part of the data in this paper, it is still much lower in order of magnitude than the above-ground database, which significantly limits the accuracy of underwater target detection. At the same time, this paper has not considered the fuzzy problem of dynamic robot images, which affects the accuracy of underwater target detection in practice. In subsequent underwater target detection technologies, more attention should be paid to the study of data quantity and quality, and the problem of non-target occlusion in underwater target detection should be further studied.

(2) The high-resolution network-holding model was introduced into the underwater lightweight model prediction network to maintain the feature expression ability of a high resolution, which improves the detection ability of small targets to a certain extent. However, there is still an issue of the partial detection of large targets, and the positioning deviation is significant, especially for some targets that account for more than 30%. Therefore, we need to investigate a detection method that accounts for a relatively large target.

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

Funding: This research was funded by National Natural Science Foundation of China, grant number 62006102.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Henke, B.; Vahl, M.; Zhou, Z. Removing color cast of underwater images through non-constant color constancy hypothesis. In Proceedings of the 2013 8th International Symposium on Image and Signal Processing and Analysis, Trieste, Italy, 4–6 September 2013; pp. 20–24.
2. Ancuti, C.; Ancuti, C.O.; Haber, T.; Bekaert, P. Enhancing underwater images and videos by fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 81–88.
3. Hu, X.; Yang, L.; Zhao, Z.; Liu, J.; Yang, X.; Sun, C.; Chen, S.; Li, B.; Zhou, C. Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network. *Comput. Electron. Agric.* **2021**, *185*, 106135. [CrossRef]
4. 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] [PubMed]
5. Land, E.H. The retinex theory of color vision. *Sci. Am.* **1977**, *237*, 108–129. [CrossRef] [PubMed]
6. Hu, G.; Wang, K.; Peng, Y.; Qiu, M.; Shi, J.; Liu, L. Deep learning methods for underwater target feature extraction and recognition. *Comput. Intell. Neurosci.* **2018**, *2018*, 1214301. [CrossRef] [PubMed]
7. Zhang, S.; Wang, T.; Dong, J.; Yu, H. Underwater image enhancement via extended multi-scale retinex. *Neurocomputing* **2017**, *245*, 1–9. [CrossRef]
8. Zhao, M.; Liu, X.; Liu, H.; Wong, K.K.L. Super-resolution of cardiac magnetic resonance images using laplacian pyramid based on generative adversarial networks. *Comput. Med. Imaging Graph.* **2020**, *80*, 101698. [CrossRef] [PubMed]
9. Lu, Y.; Fu, X.; Chen, F.; Wong, K.K.L. Prediction Christian of fetal weight at varying gestational age in the absence of ultrasound examination using ensemble learning. *Artif. Intell. Med.* **2020**, *102*, 101748. [CrossRef] [PubMed]

10. Zhao, M.; Wei, Y.; Wong, K.K.L. A generative adversarial network technique for high-quality super-resolution reconstruction of cardiac magnetic resonance images. *Magn. Reson. Imaging* **2022**, *85*, 153–160. [CrossRef] [PubMed]
11. Fu, X.; Zhuang, P.; Huang, Y.; Liao, Y.; Zhang, X.-P.; Ding, X. A retinex-based enhancing approach for single underwater image. In Proceedings of the 2014 IEEE International Conference on Image Processing, Paris, France, 27–30 October 2014; pp. 4572–4576.
12. Chiang, J.Y.; Chen, Y.C. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* **2012**, *21*, 1756–1769. [CrossRef]
13. 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] [PubMed]
14. Ali, R.; Hardie, R.C.; Narayanan, B.N.; Kebede, T.M. IMNets: Deep learning using an incremental modular network synthesis approach for medical imaging applications. *Appl. Sci.* **2022**, *12*, 5500. [CrossRef]
15. Zhao, X.; Jin, T.; Qu, S. Deriving inherent optical properties from background color and underwater image enhancement. *Ocean. Eng.* **2015**, *94*, 163–172. [CrossRef]
16. Yamashita, A.; Fujii, M.; Kaneko, T. Color registration of underwater images for underwater sensing with consideration of light attenuation. In Proceedings of the 2007 IEEE International Conference on Robotics and Automation, Rome, Italy, 10–14 April 2007; pp. 4570–4575.
17. Mukherjee, K.; Gupta, S.; Ray, A.; Phoha, S. Symbolic analysis of sonar data for underwater target detection. *IEEE J. Ocean. Eng.* **2011**, *36*, 219–230. [CrossRef]
18. Elberink, S.O. Target graph matching for building reconstruction. *Proc. Laserscanning* **2009**, *9*, 49–54.
19. Hsiao, Y.H.; Chen, C.C.; Lin, S.I.; Lin, F.-P. Real-world underwater fish recognition and identification, using sparse representation. *Ecol. Inform.* **2014**, *23*, 13–21. [CrossRef]
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
22. 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.
23. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
24. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 734–750.
25. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. Available online: https://arxiv.org/abs/1804.02767 (accessed on 8 April 2018).