

Research Article

Underwater Image Processing and Object Detection Based on Deep CNN Method

Fenglei Han, Jingzheng Yao , Haitao Zhu, and Chunhui Wang

College of Shipbuilding Engineering, Harbin Engineering University, No. 145 Nantong Street, NanGang District, Harbin, Heilongjiang Province 150001, China

Correspondence should be addressed to Jingzheng Yao; yaojingzheng_heu@163.com

Received 18 January 2020; Revised 17 February 2020; Accepted 6 May 2020; Published 22 May 2020

Academic Editor: Xavier Vilanova

Copyright © 2020 Fenglei Han et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Due to the importance of underwater exploration in the development and utilization of deep-sea resources, underwater autonomous operation is more and more important to avoid the dangerous high-pressure deep-sea environment. For underwater autonomous operation, the intelligent computer vision is the most important technology. In an underwater environment, weak illumination and low-quality image enhancement, as a preprocessing procedure, is necessary for underwater vision. In this paper, a combination of max-RGB method and shades of gray method is applied to achieve the enhancement of underwater vision, and then a CNN (Convolutional Neural Network) method for solving the weakly illuminated problem for underwater images is proposed to train the mapping relationship to obtain the illumination map. After the image processing, a deep CNN method is proposed to perform the underwater detection and classification, according to the characteristics of underwater vision, two improved schemes are applied to modify the deep CNN structure. In the first scheme, a 1×1 convolution kernel is used on the 26×26 feature map, and then a downsampling layer is added to resize the output to equal 13×13 . In the second scheme, a downsampling layer is added firstly, and then the convolution layer is inserted in the network, the result is combined with the last output to achieve the detection. Through comparison with the Fast RCNN, Faster RCNN, and the original YOLO V3, scheme 2 is verified to be better in detecting underwater objects. The detection speed is about 50 FPS (Frames per Second), and mAP (mean Average Precision) is about 90%. The program is applied in an underwater robot; the real-time detection results show that the detection and classification are accurate and fast enough to assist the robot to achieve underwater working operation.

1. Introduction

With the development of computer vision and image processing technology, the application of image processing methods to improve the underwater image quality to satisfy the requirements of the human vision system and machine recognition has gradually become a hot issue. At present, the methods of underwater image enhancement and restoration can be divided into nonphysical model image enhancement and physical model-based image restoration.

For underwater image enhancement, traditional image processing methods include color correction algorithms and contrast enhancement algorithms, the white balance method [1], gray world hypothesis [2], and gray edge hypothesis [3] are the typical color correction methods, and the

contrast enhancement algorithms include the histogram equalization [4] and restricted contrast histogram equalization [5], which are commonly used to enhance underwater images. Compared with the good results obtained by common image processing, the results obtained by these methods are unsatisfactory for underwater vision. The main reason is that the ocean environment is complex, and many unfavorable factors, such as the scattering and absorption of light by water, and the underwater suspended particles have serious interference on image quality.

More complex and comprehensive underwater image enhancement methods are proposed for solving the degradation of color fading, contrast reduction, and detail blurring problems. For example, Ghani et al. [6] proposed a method to solve the low contrast problem of underwater images; the

Rayleigh stretch limited contrast adaptive histogram was used to normalize the global contrast-enhanced image and the local contrast-enhanced image, so as to realize the enhancement for the low quality of underwater images. Li et al. [7] considered the multiple degradation factors of the underwater image, adopted image dehazing algorithm, color compensation histogram, equalization saturation, illumination intensity stretching, and bilateral filtering algorithm to solve the problems of blurring, color fading, low contrast, and noise problems. Braik et al. [8] used particle swarm optimization (PSO) to enhance underwater images by reducing the influence of light absorption and scattering. In addition, the Retinex theory is often applied in the underwater image enhancement process [9]; Fu et al. [10] proposed an underwater image enhancement method based on the Retinex model. This method applied different strategies to enhance the reflection and illumination components of the underwater image on the basis of color correction, and then the final enhancement results are synthesized. Perez et al. [11] proposed an underwater image enhancement method based on deep learning, which constructed a training data set consisting of groups of degraded underwater images and restored underwater images. The model between degraded underwater images and restored underwater images was obtained from a large number of training sets by deep learning method, which is used to enhance the underwater image quality.

Underwater detection mainly depends on the digital cameras, and the image processing is commonly used to enhance the quality and reduce the noise; contour segmentation methods are commonly used to locate the objects. A lot of such methods are proposed to realize the target detection. For instance, Chen Chang et al. [12] proposed a new image-denoising filter based on a standard median filter, which is used to detect noise and change the original pixel value to a newer median. Prabhakar et al. [13] proposed a novel denoising method to remove additive noise present in the underwater images, homomorphic filtering for correcting nonuniform illumination is used, and anisotropic filtering is applied for smoothing. A new approach for denoising combining wavelet decomposition with high-pass filter is applied to enhance the underwater images (Sun et al., 2011); both the low-frequency components of the back-scattering noise and the uncorrelated high-frequency noise can be effectively depressed simultaneously. However, the unsharpness in the processed image is serious based on the wavelet method. Kocak et al. [14] used a median filter to remove the noise, the quality of the images are enhanced by RGB color level stretching, the atmospheric light is obtained through the dark channel prior, and this method is helpful in the case of images with minor noise. For noisy images, a bilateral filtering method is utilized by Zhang et al. [15], the results are good, but the time processing is very high. An exact unbiased inverse of the generalized Anscombe transformation is introduced by Markku et al. [16]; the comparison shows that the method plays an integral part in ensuring accurate denoising results.

A Laser Underwater Camera Image Enhancer system is designed and built by Forand et al. [17] to enhance the laser

underwater image quality, and it is testified that the system has a range of 3 to 5 times than that of a conventional camera with floodlights. Yang et al. [18] proposed a method of detecting underwater laser weak target based on Gabor transform, which is processed on laser underwater complicated nonstationary signal to turn it to become an approximate stationary signal, and then the triple correlation is computed with Gabor transform coefficient and it can eliminate random interference and extrude target signal's correlation. Ouyang et al. [19] investigated the application of light field rendering (LFR) to images taken from a distributed bistatic nonsynchronous laser line scan imager using both line-of-sight and non-line-of-sight imaging geometries to create a multiperspective rendering of an unknown underwater scene.

Chang et al. [20] introduced a significant amount of polarization into light at scattering angles near 90 degrees: This light can then be distinguished from light scattered by an object that remains almost completely unpolarized. Results were obtained from a Monte Carlo simulation and from a small-scale experiment, in which an object was immersed in a cell filled with polystyrene latex spheres suspended in water. Gruev et al. [21] described two approaches for creating focal plane polarization imaging sensors. The first approach combines polymer polarization filters with a CMOS active pixel sensor and computes polarization information at the focal plane. The second approach outlines the initial work on polarization filters using aluminum nanowires. Measurements from the first polarization image sensor prototype are discussed in detail, and applications for material detection using polarization techniques are described. Underwater Polarization Imaging Technology is introduced in detail by Li et al. [22].

The above methods are based on wavelet decomposition, statistical methods or by means of laser technology, or color polarization theories, the results show that the methods are reasonable and effective, but the common weakness is that the processing is very time consumable, and it is difficult to achieve real-time detection.

The Convolution Neural Network (CNN) is recognized as the fastest detection method by many ways in different research fields; Krizhevsky et al. [23] applied CNN method to deal with classification problem winning the champion of ILSVRC (ImageNet Large Scale Visual Recognition Challenge), which reduce the top 5 error rate to 15.3%, from then on deep CNN has been widely applied. Girshick [24] proposed Region Convolutional Neural Network (RCNN) through combining the RPN (Region Proposal Network) and CNN methods, which are testified on Pascal VOC 2007, mAP reaches 66%. Based on RCNN, SPP-Net (Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition) is presented by He K. et al. [25] to improve the detection efficiency. RESNET is proposed by [26]; the success of RESNET is to solve the problem of network migration with the help of the introduction of residual module, so as to improve the depth of the network, which can obtain the features with stronger expression ability and higher accuracy. Multilayer Perceptron (MLP) is applied to replace SVM (Support Vector Machine); the training and classification

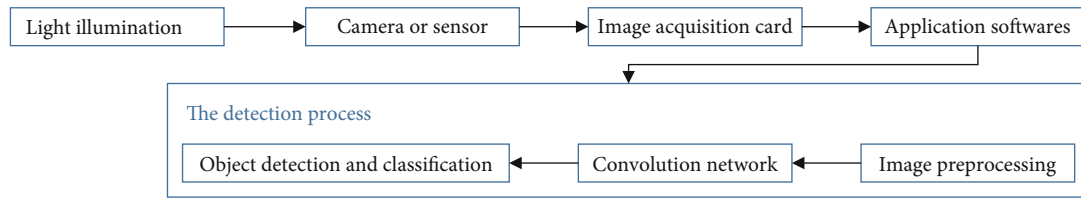


FIGURE 1: The image processing and object detection structure.

are optimized significantly, which is named Fast RCNN [6]. In Fast RCNN, Ren S, He K, and Girshick [27] added RPN to select and modify the region proposals instead of selective search, which is aimed at solving the end-to-end detection problem; this is the Faster RCNN method. Liu Wei proposed a SSD (Single Shot MultiBox) method in ECCV2016 (European Conference on Computer Vision). Compared with Faster RCNN, it has a distinct speed advantage, which is able to directly predict the coordinates and categories of bounding box without processing of generating a proposal.

In 2016CVPR(IEEE Conference on Computer Vision and Pattern Recognition), Redmon proposed YOLO (You Only Look Once) [28] Regression object detection algorithm; by this method, the detection speed is improved significantly, and the real-time detection is possible to be realized. When the YOLO algorithm was put forward, the accuracy and speed of computation were not as good as that of the SSD algorithm. Then, Redmon proposed YOLO V2 [29] version to optimize the original YOLO multitarget detection framework through a series of methods, and the accuracy is greatly improved under the advantage of maintaining the original speed. Earlier of 2018, Redmon put forward the YOLO v3 [30], which is generally recognized as the fastest detection method, and the accuracy and the detection speed are greatly improved compared with the other methods.

In this paper, we applied a combination of max-RGB method and shades of gray method to enhance the underwater images, and a CNN method is used for weakly illuminated image. For the underwater object detection, a new CNN method is proposed to solve the underwater object detection problem; considering the particularity of underwater vision, two improved schemes are proposed to improve the detection accuracy, and the results are compared with Fast RCNN[6], Faster RCNN [27], and original YOLO V3[30]. It is testified through comparison that the modification is effective, and the program is installed on an underwater robot to test the real-time detection.

2. Image Preprocessing

For underwater computer vision, the image preprocessing is the most important procedure for object detection. Because of the effects of light scattering and absorption in the water, the images obtained by the underwater vision system show the characteristics of uneven illumination, low contrast, and serious noise. By analyzing the current image processing algorithms, enhancement algorithms for underwater images are proposed in this paper.

2.1. The Underwater Vision Detection Architecture. The typical underwater visual system is composed of light illumination, camera or sensor, image acquisition card, and application software. The software process of the underwater visual recognition system generally includes several parts, such as image acquisition, image preprocessing, convolution neural network, and target recognition, as shown in Figure 1.

Image preprocessing is at the low level, the fundamental purpose is to improve image contrast, to weaken or suppress the influence of various kinds of noise as far as possible, and it is important to retain useful details in the image enhancement and image filtering process. Convolutional Neural Network is used to divide images into multiple nonoverlapping regions; the basis of object detection and classification is based on feature extraction, which is aimed at extracting the most effective essential features that reflect the target. Every aspect is closely related, so every effort should be made to achieve satisfactory results. The research of this paper mainly focuses on image preprocessing and recognition of typical targets from the underwater vision.

2.2. Combination of Max-RGB Method and Shades of Gray Method. The absorption of water to light leads to the decline of the color of underwater images. As the red and orange light are completely absorbed at 10 meters deep in the water, the underwater images generally get blue-green color. In order to eliminate the color deviation of underwater images, color correction of underwater images must be carried out.

The color correction of the normal image has been very mature. Many white balance methods, such as Gray Word method, max-RGB method, Shades of Gray method, and Gray Edge method, are used to correct the color deviation of the image according to the color temperature. Generally, the application scenarios of these methods are general partial color conditions, and the treatments for severe underwater vision are not satisfied. In this paper, the original max-RGB method and shades of gray method are combined to identify the illuminant color.

$$I(x) = \int_w e(\lambda)s(\lambda, x)c(\lambda)d\lambda, \quad (1)$$

where $I(x)$ is the input underwater image, $e(\lambda)$ is the radiance given by the light source, λ is the wavelength, $s(\lambda, x)$ represents the surface reflectance, $c(\lambda)$ denotes the sensitivity of the sensors, and w is the visible spectrum.

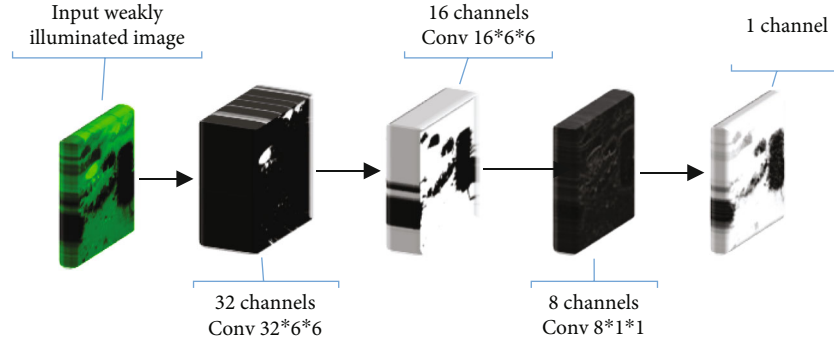


FIGURE 2: The mapping relationship prediction between the input image and illumination map CNN structure.

The illuminant e is defined as

$$e = \int_w e(\lambda) c(\lambda) d\lambda. \quad (2)$$

The average reflectance of the scene is gray according to the Grey-World assumption [31]

$$k = \frac{\int s(\lambda, x) dx}{\int dx}. \quad (3)$$

Assume k is a constant value, the physical meaning of equation (1) can be simply described as that the observed image $I(x)$ can be decomposed into the product of the reflectance of image $S(x)$ and the illumination map $e(\lambda)$. Thus, weak illumination image enhancement means removing weak illumination from the input image; equation (3) is substituted in equation (1)

$$\frac{\int s(\lambda, x) dx}{\int dx} = \frac{1}{\int dx} \iint_w e(\lambda) s(\lambda, x) c(\lambda) d\lambda dx. \quad (4)$$

The illumination by explaining that the average color of the entire image raised to a power n

$$ke = \left[\frac{\int I^n dx}{\int dx} \right]^{1/n} \quad (5)$$

According to the max-RGB method, the above equation can be modified as

$$ke = \max I(x) * \left[\frac{\int I^n dx}{\int dx} \right]^{1/n}, \quad (6)$$

where n can take any number between 1 and ∞ , the default value of $n = 6$, which is defined in shades of gray method proposed by Finlayson [31].

2.3. CNN Method for Weakly Illuminated Image Enhancement. Retinex model can be used to enhance the image based on the estimated illumination map; for underwater vision, the images are always weakly illuminated, so a trainable CNN method is applied to predict the mapping

relations between weakly illuminated image and the corresponding illumination map. A four-layer convolutional network is used, the first and the third layers focus on the high light regions, and the second layer focuses on low-light regions while the last layer is to reconstruct the illumination map. The Convolutional Neural Network directly learns from an end-to-end mapping between dark and bright images. Low-light image enhancement in this paper is regarded as a machine learning problem. A weakly illuminated image is input, and a $32 * 6 * 6$ convolution layer is applied to change the image into 32 channels; the 3-D view figure means multilayers feature map, and then $16 * 6 * 6$ and $8 * 1 * 1$ convolution layers are added in the network; the output is a one channel feature map. In this model, most of the parameters are optimized by back-propagation, while the parameters of traditional models depend on the neutral network. The four-layer convolutional network structure is shown in Figure 2.

The input image is the weakly illuminated image, and the output is the corresponding illumination map. Similar with Chongyi Li et al. [32] and Dong et al. [33], the network contains four convolutional layers with specific tasks. Observing the feature maps in Figure 2, different convolutional layers have different effects on the final illumination map. For example, the first two layers focus on the high-light regions and the third layer focuses on low-light regions, while the last layer is to reconstruct the illumination map. The specific operation form of the four convolutional layers is described as shown in Figure 2.

The enhancement effects are shown in Figure 3, the underwater background color is improved significantly, and the weakly illuminated images are enhanced using the trainable CNN method.

3. The Object Detection Theories

The images are resized into $448 * 448$, the input images are resized, the image will stretch, and the label will be recalculated too. In this case, in fact, a scale factor is calculated to record the scale of width and height, respectively, and x_{\min} , x_{\max} , y_{\min} , and y_{\max} are calculated, respectively, but the output images are resized to be same as the original images. A CNN method is used to predict the bounding boxes and classification probabilities. For the underwater detection, the

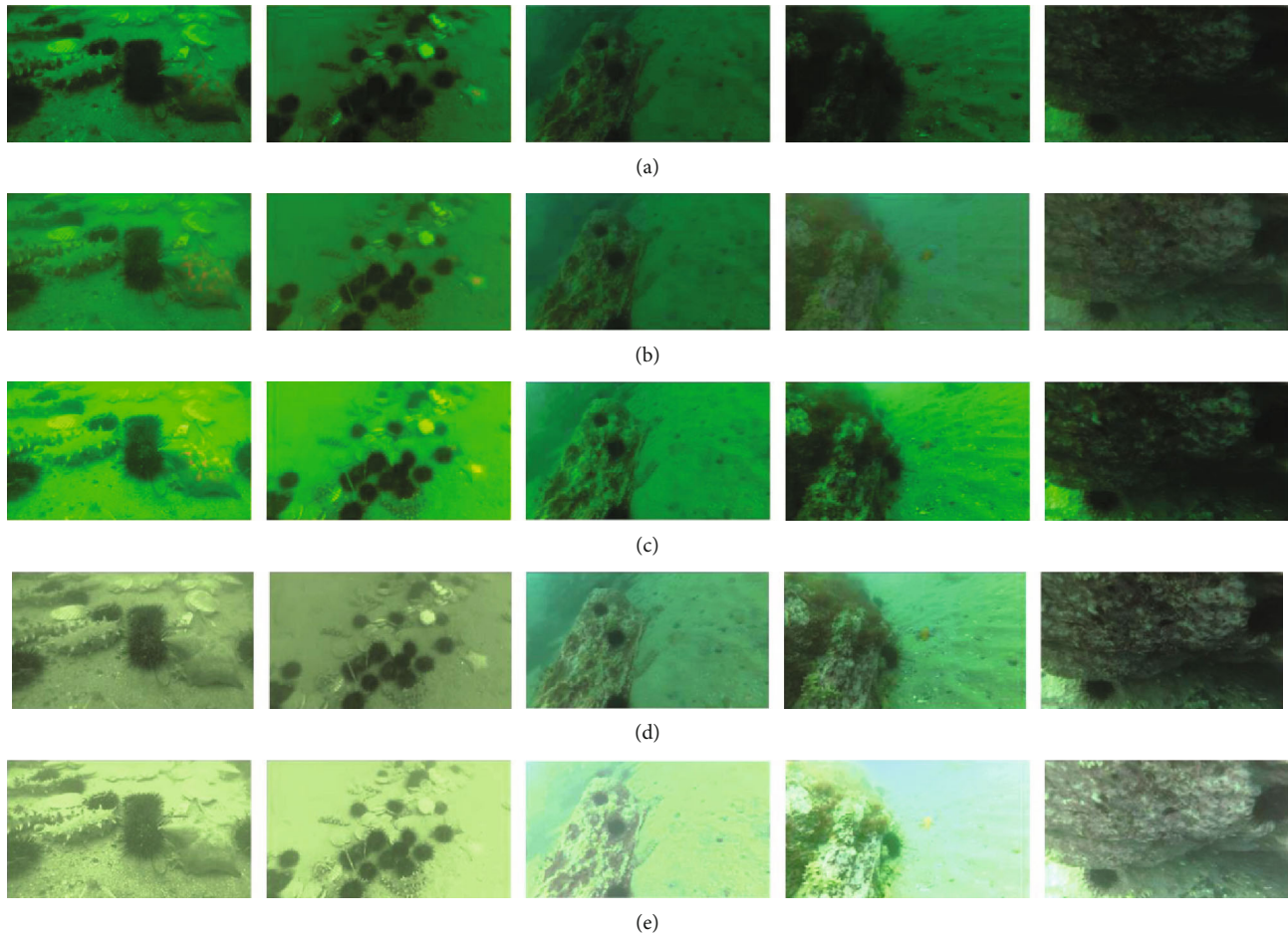


FIGURE 3: The enhancement effect by different methods: (a) original image; (b) the method proposed by D.J. [34]; (c) the method proposed by [35]; (d) max-RGB and shape of gray method; (e) weakly illuminated image enhancement.

targets are difficult to be identified from the background. In order to improve the detection accuracy, the whole image information is used to predict the bounding boxes of the targets and classify the objects at the same time; through this proposal, the end-to-end real time targets detection can be realized.

3.1. Convolutional Neural Network. The image is divided into 4×4 grid cells, which is used to locate the center of the detection object. For each grid cell, the bounding boxes (bbox) are predicted, which includes 5 parameters, (x, y) is the center location of the bounding box, (w, h) is the width and height of the box, confidence is the Intersection of Union (IoU), which equals the intersection divided by the union between the bbox and the ground truth, the process is shown in Figure 4.

The bounding box is predicted through a fully-connected layer; if the width and height are only related to the scales and ratios of the input images, the location of the different objects in different shapes cannot be very accurate. Therefore, Region Proposal Network is applied to predict the bounding box and confidence [27], in which the predicted boxes with different scales and ratios are used, and the offsets of the boxes are calculated in RPN, as shown in Figure 5. The fully

connected layer is removed, and the convolution layer with anchor boxes is added to predict the bounding box. In order to keep the high quality of the original image, a pooling layer is removed, and the input image is 448×448 , the scale of the final feature map is 14×14 with only one center.

Through a series of convolutions, a common feature map is obtained, then, RPN is applied. Firstly through a convolution, a new feature map is given, which can also be seen as high dimensional feature vectors, then through two 1×1 convolutions, a $18 \times 16 \times 16$ feature map and a $36 \times 16 \times 16$ characteristic map are obtained. That is $16 \times 16 \times 9$ results, each result contains 2 scores and 4 coordinates, and then combined with the predefined anchors; after preprocessing, the bounding box is calculated.

In deep learning process, grid cell data is input in the deep learning results, the center of some pixels is within the certain range of a specific grid cell, and then, all the pixels satisfied the feature of the object are clustered in a certain range. After many times of trial training with penalty, it can find the exact range through sliding window. However, the center position cannot exceed the range of the grid cell. This greatly limits the model's computation when it is sliding around in the picture. In this way, position detection and category recognition are combined into a CNN network to predict, you

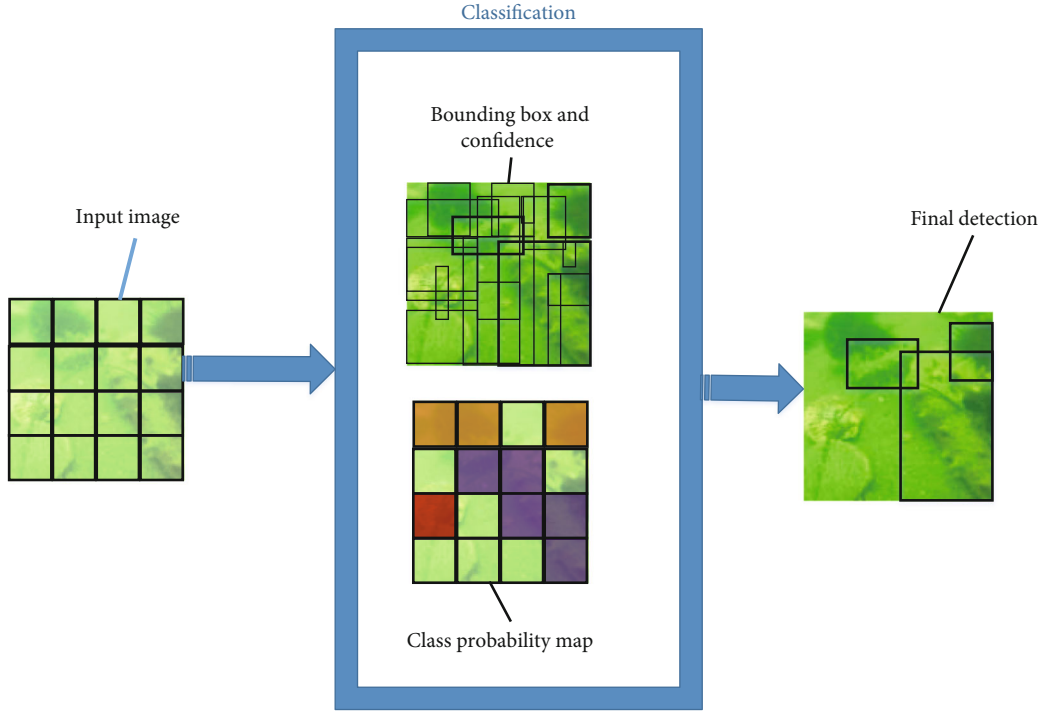


FIGURE 4: Detection process.

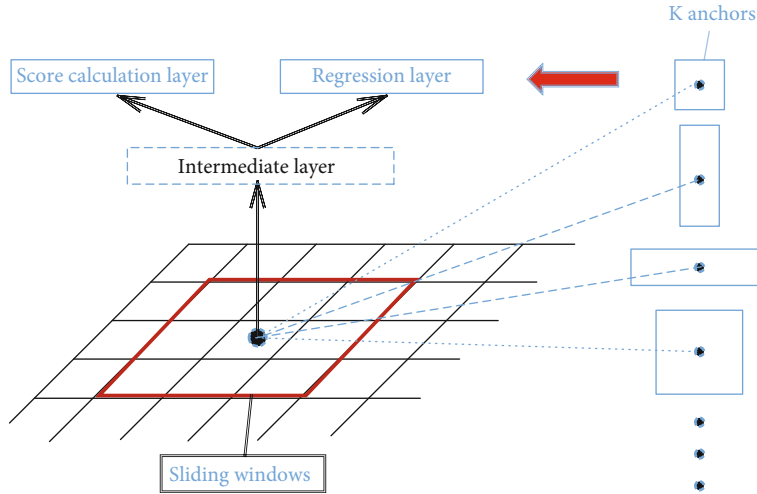


FIGURE 5: The convolutional feature map.

only need to scan the picture once to infer the position information and category of all objects in the picture.

3.2. Cluster Analysis. The k -means cluster method is used to train the bounding boxes, the target is to obtain a better IoU between the bbox and the grounding truth, so the distance from the center of bbox to the cluster center is calculated as a parameter:

$$d(\text{box}, \text{centroid}) = 1 - \text{IoU}(\text{box}, \text{centroid}). \quad (7)$$

The Euclidean distance is applied in the traditional k -means cluster method, which means that the bigger boxes

with more errors compared with the smaller boxes, the result may be deviated from the true value. So the IoU score is proposed to substitute the traditional method.

The convolutional kernel is 3×3 , the max-pooling size is 2×2 , and the dimension of the feature map is reduced 2 times. The global average pooling is applied to complete the prediction; the 1×1 convolution is used to compress the channels of the feature maps, so as to reduce the parameters and the amount of calculation. A batch normalization layer is added to accelerate the convergence speed and avoid the overfitting.

Data preprocessing (unified format, equalization, noise reduction, etc.) can greatly improve the speed of training

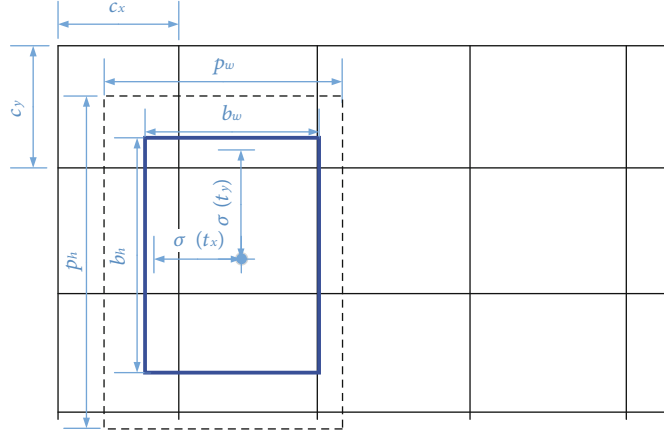


FIGURE 6: Bounding box prediction.

and enhance the training effect. Batch Normalization (BN) is proposed by Google, which is commonly used in the CNN network. After the convolution or pooling and before the activation function, all of the input data is normalized as follows:

$$\hat{x}^{(k)} = \frac{x^{(k)} - E(x^{(k)})}{\sqrt{\text{Var}(x^{(k)})}}, \quad (8)$$

$$\hat{y}^{(k)} = r^{(k)}\hat{x}^{(k)} + \beta^{(k)}$$

where E is Batch mean value and Var is variance; γ and β are the scale and shift coefficients, which are obtained from training.

3.3. Location Prediction. In order to solve the unstable problem for using the anchor boxes, especially in the process of early iteration, the following procedures are applied to predict the location of boxes:

$$x = (t_x * w_a) - x_a, \quad (9)$$

$$y = (t_y * h_a) - y_a,$$

where (x, y) is the predicted value, (x_a, y_a) is the coordinates of anchor, (x^*, y^*) is the real coordinates value, (t_x, t_y) is the offset value, and (w_a, h_a) is the width and height of the box.

When $t_x = 1$, the box is offset a distance equalled the width of the box to the right; if $t_x = -1$ the offset is to the left, so every predicted box can be located at any position on the image, which is the reason why the model is unstable, and the prediction is very time consumable. The prediction box is limited in the grid cell, and sigmoid function is used to calculate the offset value, which is defined between 0-1; the t_x , t_y , t_w , and t_h can be computed from the following equations:

$$b_x = \sigma(t_x) + c_x,$$

$$b_y = \sigma(t_y) + c_y$$

$$b_w = p_w e^{t_w}$$

$$b_h = p_h e^{t_h} \quad (10)$$

In the above equations, (c_x, c_y) is the upper left corner coordinates of the grid cell, as shown in Figure 6; when the scale of the grid cell is 1, the center is limited in the internal of the cell by the sigmoid function. The p_w and p_h are the priori width and height.

3.4. Loss Function. In the process of training, the loss function form is a key technique; for the method proposed in this paper, a sum squared error loss is used to balance the errors. For the boxes in different size prediction, the width and height of the bounding box are substituted by the square root value; thus, the smaller box has a relatively large value offset to make the prediction more effective. The loss function can be divided into 2 parts:

$$L_1 = \sum_{i=0}^{s^2} \sum_{j=0}^B l_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2 \right]. \quad (11)$$

L_1 is aimed at determining the j -th box in i -th grid cell is in charge for the object or not, which is a coordinate prediction for the loss.

$$L_2 = \sum_{i=0}^{s^2} \sum_{j=0}^B l_{ij}^{\text{obj}} (c_i - \hat{c}_i)^2 + \sum_{i=0}^{s^2} l_i^{\text{obj}} \sum_{\text{classes}} [p_i(c) - \hat{p}_i(c)]^2. \quad (12)$$

L_2 is the confidence prediction loss of the box with the object. The total loss is the sum of the L_1 and L_2 , which can give a better balance between the coordinates, confidence, and classification.

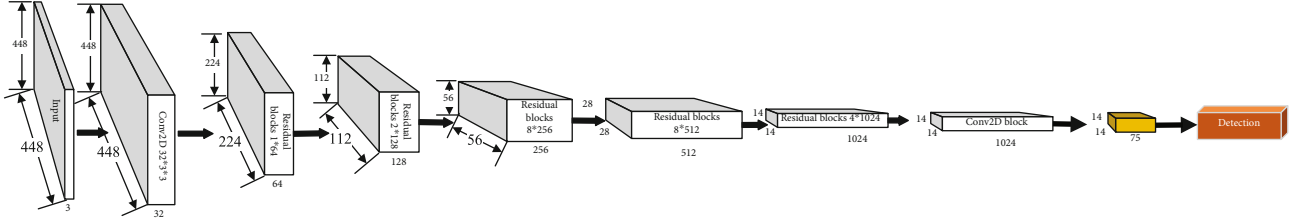


FIGURE 7: Original object detection network structure.

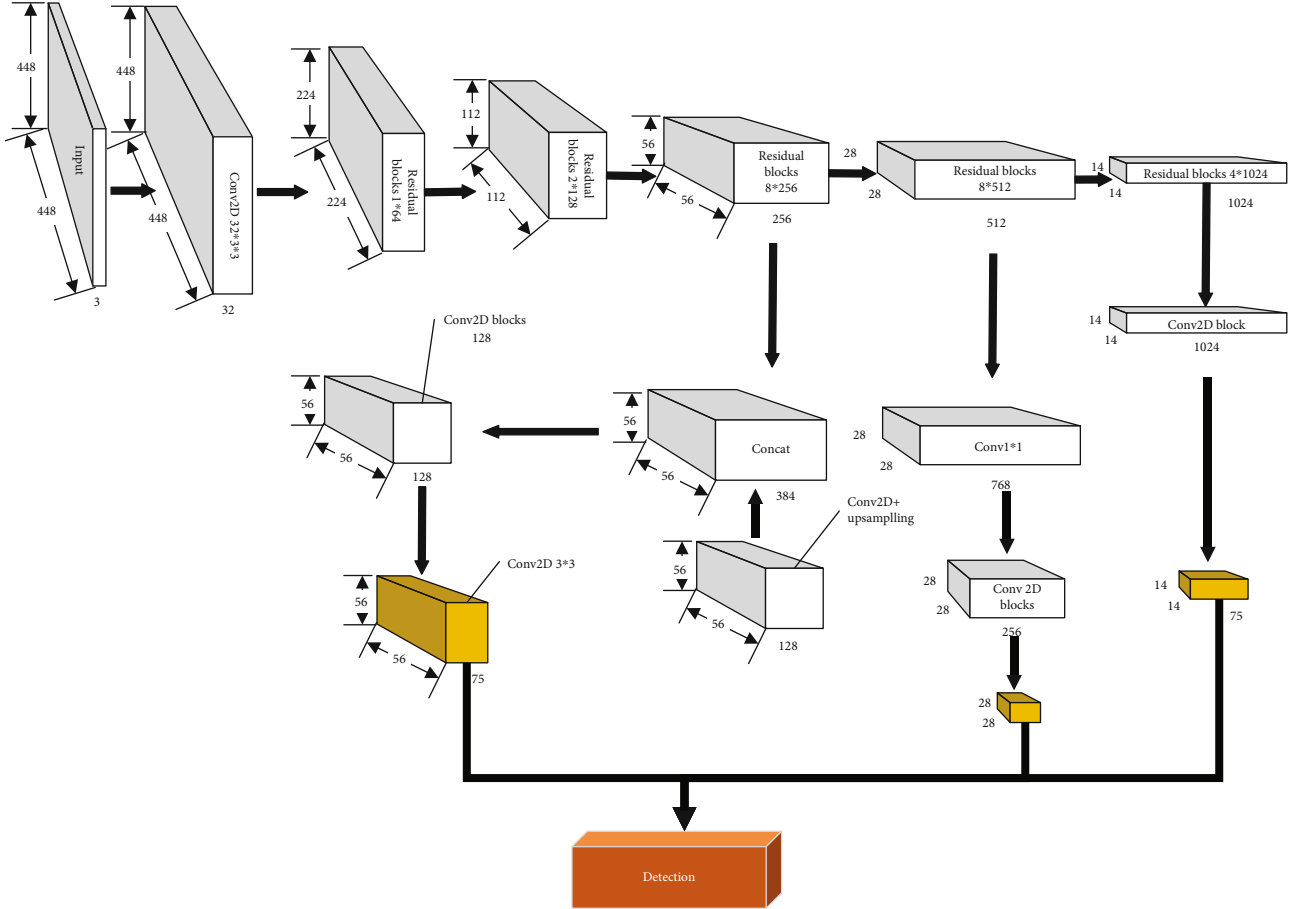


FIGURE 8: The network structure modification scheme 1.

4. Underwater Detection CNN Network

For underwater detection, the commonly used methods are not applicable because of the low-quality vision and the small objects for detection. Our original neutral network is shown in Figure 7, the input image is resized into $448 \times 448 \times 3$, the resized images should be batch normalized (BN), the convolution kernels is 3×3 and 1×1 , the stride is 1, and the output feature map is $14 \times 14 \times 75$. In order to solve the phenomenon of gradient dispersion or explosion of the network, the better proposal is to change the layer-by-layer training of deep neural network to step-by-step training. The deep neural network is divided into several subsegments,

each subsegment contains shallow network layers, then, short cut is used to make each subsegment train residual, and each subsegment has a total learning error. At the same time, the proposed method can control the propagation of gradient well and avoid the situation of vanishing gradient or exploding gradient, which is not conducive to training.

Firstly, 3×3 convolution is used to reduce the number of channels and training parameters; then, convolution kernel of different sizes is used to perform convolution operation; finally, each feature image is combined according to the channel. In order to get more advanced features, the previous way is to increase the depth of the network, and we proposed this network to achieve this goal by

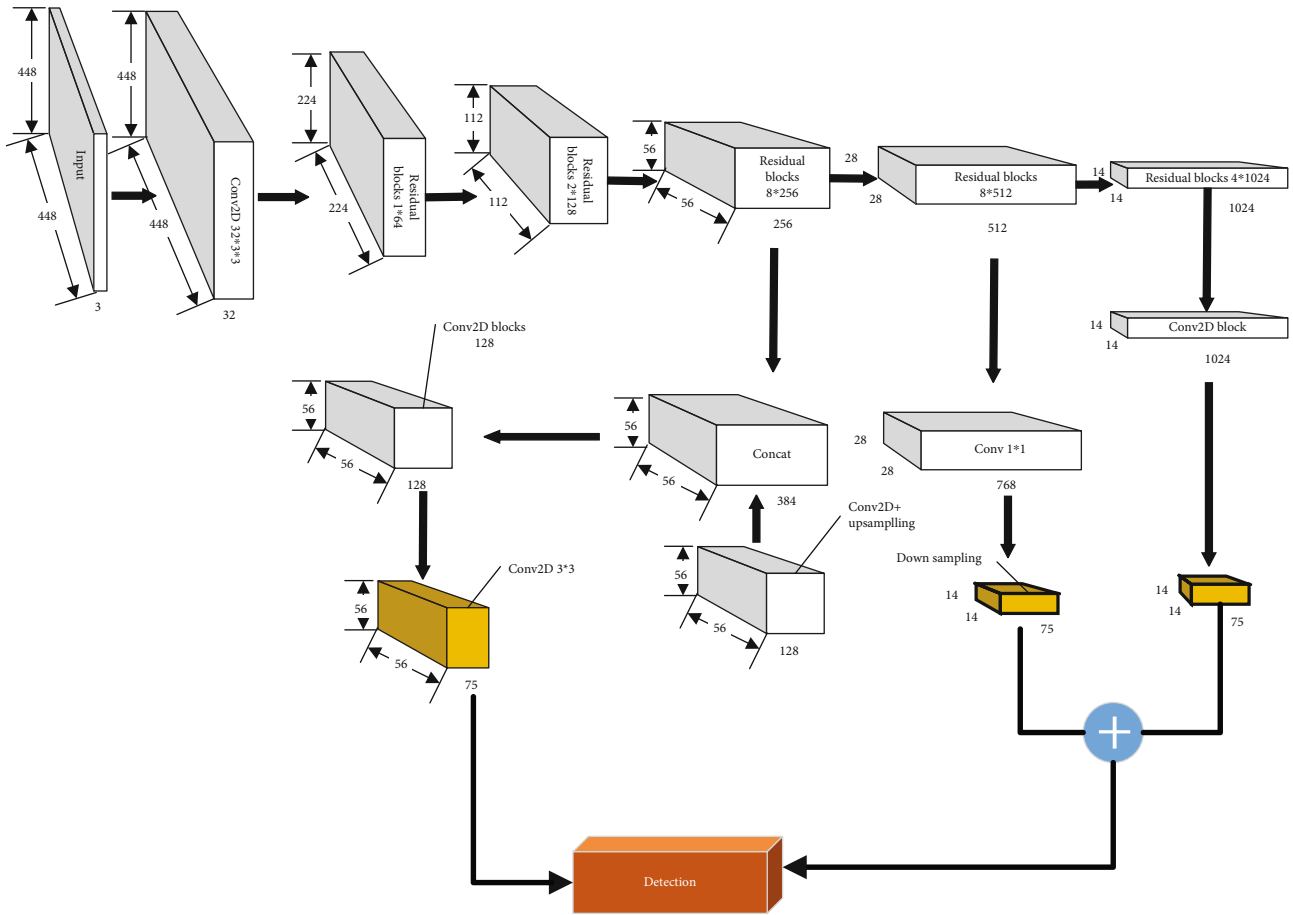


FIGURE 9: The network structure modification scheme 2.

increasing the width of the network. The concept module comprehensively considers the results of multiple convolution kernels, different information of the input image and better image representation are obtained. In order to prevent the middle part of the vanishing gradient process of the network structure, we introduced two auxiliary classifiers. Softmax operations are used on the output of two of the perception modules, and then, the auxiliary loss is calculated. Auxiliary loss is only used for training, not for the prediction process.

4.1. Network Structure Improvement. For underwater object detection, the vision sensors are installed on the underwater robot. For the real operation, the common method performs not well in small objects detection, because the regular dataset used in the experiment are normal images, which are high-quality and well-lighted images. For underwater detection, the objects are always overlapped by other things, such as rocks and corals, and the underwater vision is always vague, the clarity is low. Under these conditions, the network structure should retain more original features. In deep CNN, the more layers always extract features that are more abstract, and the deep semantic information can be extracted more clearly. On the other hand, the fewer layers can retain more representation information. The

deep semantic information and the representation information can be combined to give a more accurate detection. In this paper, the structure is proposed by two schemes, the first one is that a 1×1 convolution kernel is used on the 28×28 feature map, and then, a downsampling layer is added to resize the output to equal 14×14 , which is combined with the last output to complete the detection; the improvement is shown in Figure 8.

Because of the original information loss in convolution operation, in the second scheme, the downsampling is added firstly, and then the convolution layer is inserted in the network, the result is combined with the last output to achieve the detection; the modification is shown in Figure 9.

There are three full convolution feature extractors, respectively, corresponding to the convolutional set, which is the internal convolution kernel structure of the feature extractor, 1×1 convolution kernel is used for dimensionality reduction, 3×3 convolution kernel is used for feature extraction, and multiple convolution kernels are interleaved to achieve the purpose. Each full convolution feature layer is connected. The input of the current feature layer has a part of the output from the previous layer. Each feature layer has an output prediction results. Finally, the results are regressed according to the confidence level to get the final prediction results.

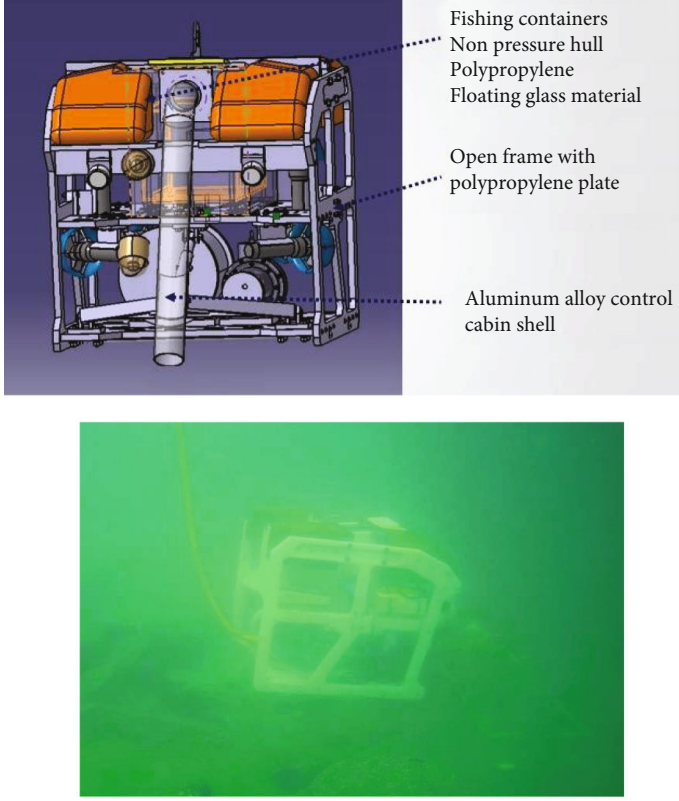


FIGURE 10: Underwater ROV for marine organisms fishing.

4.2. Dataset Augmentation. Underwater dataset is difficult to prepare, the underwater images and video are not easy to obtain on the internet, and for underwater images, the background is almost the same in the same area, so the images in the dataset are similar, because of these factors the training output model is always not effective to be used in other sea areas. Therefore, the dataset should be modified and augmented, so as to make the deep learning model more generally used. The dataset augmentation is mainly based on rotation, flipping, zoom, shift, etc.

The dataset used in this paper is obtained from the video recorded by an underwater robot. The total number of images is about 18000, and the images are similar with each other, so the rotation and color transformation is applied to transform the original patterns.

The three channels of images are dimensionality reduced; the R (Red), G (Green), and B (Blue) direction vectors are obtained, respectively.

$$I_{xy} = [R_{xy}, G_{xy}, B_{xy}] \quad (13)$$

The eigenvalues and eigenvectors of R , G , and B are defined as

$$\begin{aligned} R_{xy} &= p_r \lambda_r, \\ G_{xy} &= p_g \lambda_g \\ B_{xy} &= p_b \lambda_b \end{aligned} \quad (14)$$

	True (relevant)	False (not relevant)
Positive (retrieved)	TP	FP
Negative (not retrieved)	TN	FN

FIGURE 11: The parameters definition.

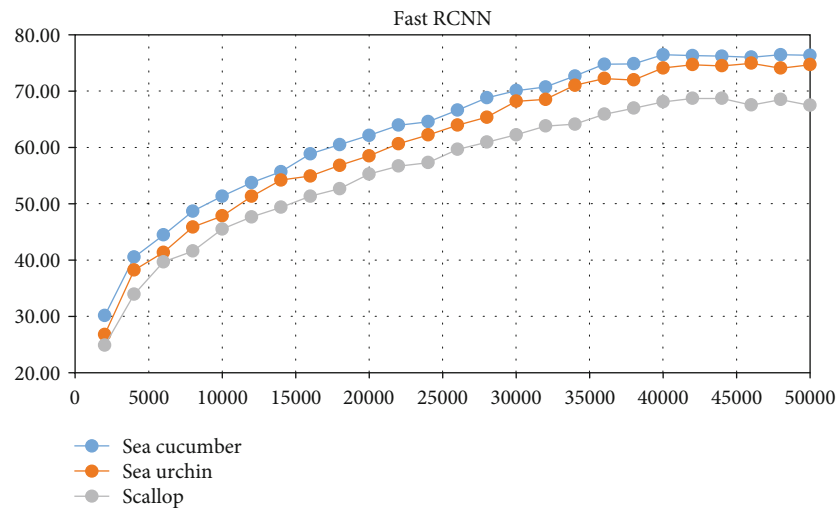
α is a random variable with a mean value of 0 and a variance of 0.1, and it is added in the transformation function as follows:

$$I_{xy} = [p_r, p_g, p_b] [\alpha_r \lambda_r, \alpha_g \lambda_g, \alpha_b \lambda_b]^T. \quad (15)$$

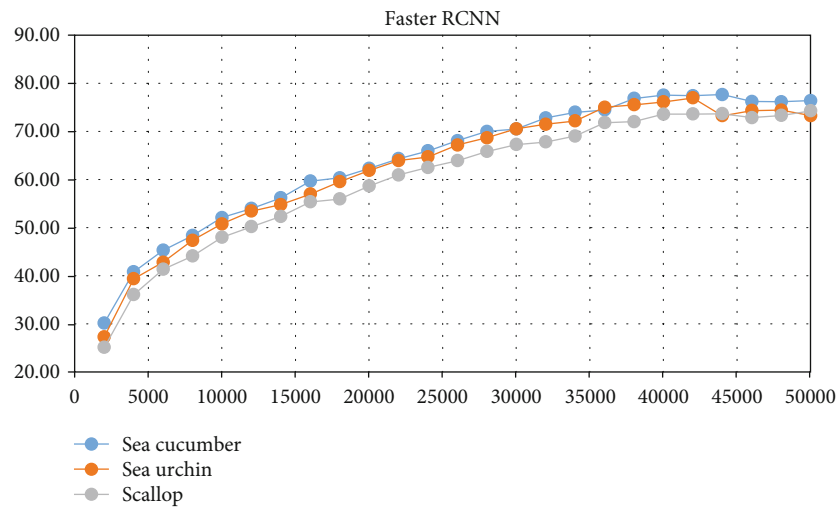
The rotation transformation is presented as

$$\begin{cases} x' = x_i \cos \theta_1 - y_i \sin \theta_1, \\ y' = x_i \sin \theta_1 + y_i \cos \theta_1. \end{cases} \quad (16)$$

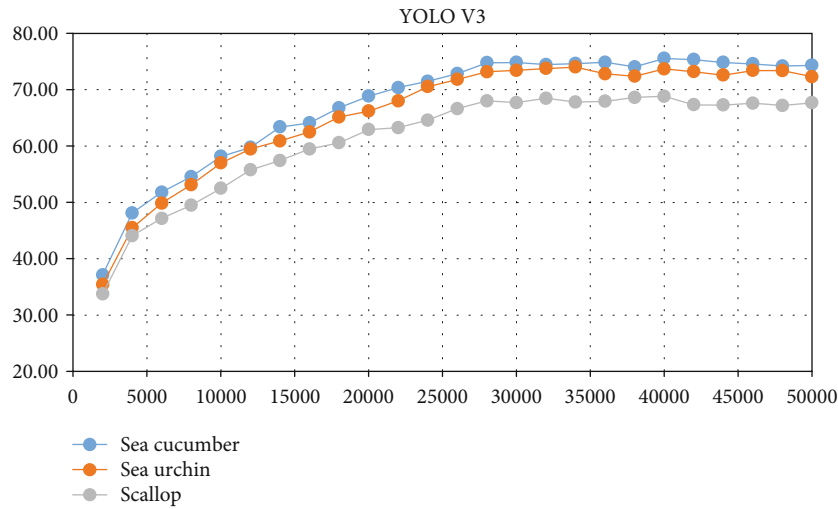
where (x', y') is the transformed location coordinates, and θ_1 is the rotation angle.



(a)



(b)



(c)

FIGURE 12: Continued.

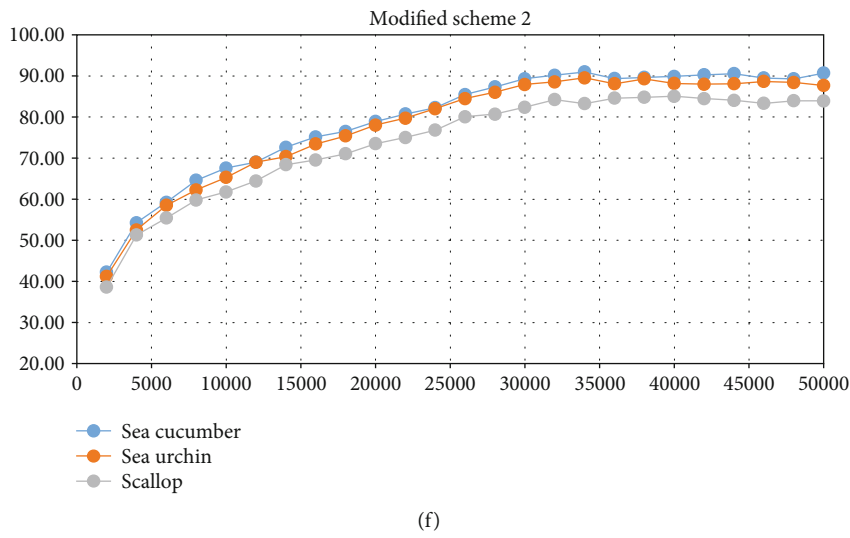
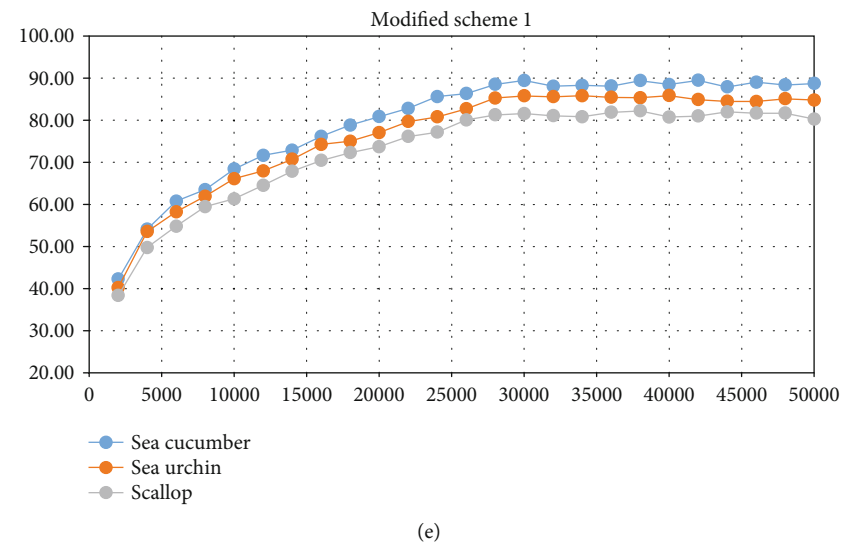
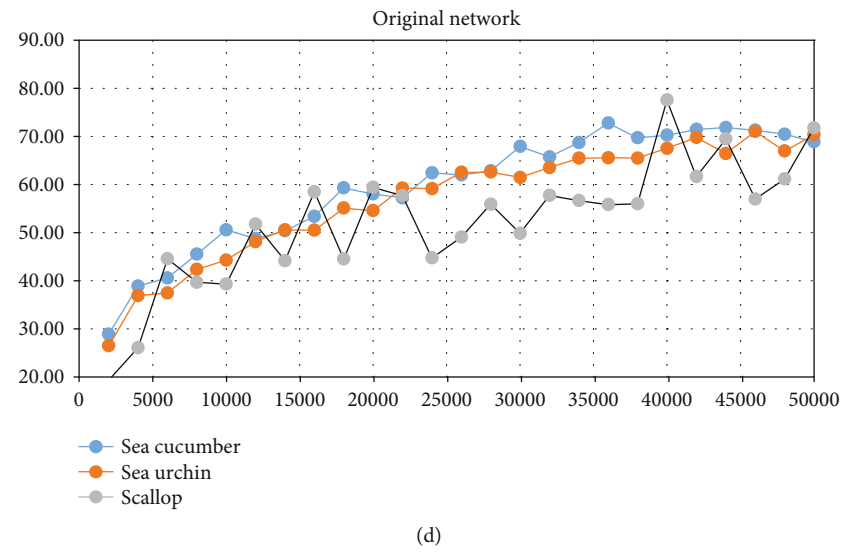


FIGURE 12: mAP results obtained by different methods.

TABLE 1: mAP and precision of different iteration times by Fast RCNN, Faster RCNN, and YOLO V3. (IoU = 0.7).

Iteration	mAP (%)	Fast RCNN			Faster RCNN			YOLO V3				
		Sea cucumber	Sea urchin	Scallop	mAP	Sea cucumber	Sea urchin	Scallop	mAP (%)	Sea cucumber	Sea urchin	Scallop
2000	27.26	30.13	26.79	24.87	27.53	30.18	27.29	25.13	35.43	37.14	35.42	33.74
4000	37.56	40.51	38.23	33.93	38.74	40.80	39.35	36.06	45.90	48.12	45.50	44.08
6000	41.83	44.45	41.36	39.67	43.15	45.30	42.80	41.35	49.61	51.81	49.87	47.16
8000	45.37	48.67	45.85	41.59	46.59	48.35	47.35	44.08	52.40	54.56	53.14	49.51
10000	48.22	51.33	47.84	45.50	50.28	52.09	50.76	48.00	55.89	58.17	56.99	52.50
12000	50.90	53.75	51.31	47.65	52.53	53.96	53.44	50.20	58.34	59.77	59.48	55.77
14000	53.09	55.69	54.20	49.38	54.43	56.18	54.78	52.31	60.58	63.39	60.91	57.44
16000	55.04	58.85	54.92	51.35	57.32	59.66	56.98	55.34	62.02	64.09	62.50	59.47
18000	56.66	60.49	56.81	52.67	58.62	60.35	59.55	55.95	64.18	66.79	65.15	60.58
20000	58.63	62.12	58.49	55.27	60.93	62.30	61.86	58.63	66.00	68.87	66.20	62.93
22000	60.42	63.95	60.63	56.67	63.07	64.33	63.93	60.95	67.22	70.37	68.02	63.26
24000	61.35	64.57	62.19	57.29	64.37	65.94	64.67	62.51	68.88	71.50	70.54	64.60
26000	63.40	66.60	63.94	59.65	66.38	68.07	67.17	63.90	70.44	72.85	71.83	66.64
28000	65.03	68.81	65.36	60.92	68.15	69.98	68.67	65.82	72.00	74.79	73.17	68.03
30000	66.84	70.09	68.19	62.24	69.42	70.49	70.53	67.24	71.99	74.84	73.44	67.70
32000	67.68	70.73	68.53	63.78	70.68	72.78	71.47	67.79	72.24	74.47	73.78	68.47
34000	69.26	72.65	71.03	64.11	71.72	73.96	72.17	69.01	72.15	74.62	74.01	67.81
36000	70.96	74.75	72.23	65.90	73.75	74.44	75.02	71.79	71.88	74.87	72.82	67.95
38000	71.25	74.83	71.98	66.95	74.80	76.83	75.53	72.03	71.69	74.05	72.37	68.63
40000	72.88	76.46	74.09	68.08	75.75	77.53	76.13	73.58	72.70	75.55	73.70	68.83
42000	73.23	76.30	74.68	68.71	75.99	77.41	76.96	73.59	71.96	75.34	73.21	67.33
44000	73.13	76.19	74.49	68.70	74.86	77.65	73.26	73.67	71.57	74.83	72.59	67.28
46000	72.82	76.00	74.93	67.53	74.46	76.19	74.33	72.85	71.87	74.58	73.41	67.61
48000	73.01	76.46	74.07	68.49	74.62	76.16	74.41	73.30	71.59	74.19	73.39	67.20
50000	72.84	76.35	74.69	67.48	74.64	76.40	73.26	74.25	71.44	74.32	72.32	67.69

The detection results obtained by the methods proposed in this paper are shown in Table 1.

TABLE 2: mAP and precision of different iteration times by YOLO v3 and modified methods. (IoU = 0.7).

Iteration	mAP (%)	Original network				Scheme 1				Scheme 2			
		Sea cucumber	Precision (%)	Scallop	mAP	Sea cucumber	Precision (%)	Scallop	mAP	Sea cucumber	Precision (%)	Scallop	mAP
2000	24.90	28.88	26.48	19.35	40.29	42.25	40.25	38.37	40.72	42.25	41.25	38.65	
4000	33.95	38.90	36.90	26.07	52.47	54.10	53.55	49.76	52.71	54.28	52.54	51.31	
6000	40.85	40.57	37.45	44.54	57.94	60.76	58.25	54.81	57.73	59.24	58.51	55.43	
8000	42.51	45.51	42.34	39.67	61.63	63.49	61.93	59.46	62.22	64.60	62.26	59.79	
10000	44.72	50.58	44.29	39.29	65.27	68.39	66.11	61.31	64.91	67.60	65.33	61.79	
12000	49.53	48.75	48.11	51.74	68.04	71.64	67.95	64.53	67.49	69.04	68.98	64.45	
14000	48.37	50.35	50.59	44.16	70.50	72.86	70.74	67.89	70.47	72.64	70.35	68.41	
16000	54.12	53.40	50.48	58.49	73.61	76.17	74.23	70.44	72.72	75.16	73.45	69.54	
18000	52.95	59.29	55.08	44.48	75.38	78.84	74.99	72.31	74.32	76.51	75.40	71.07	
20000	57.33	58.04	54.54	59.42	77.20	80.87	77.05	73.69	76.82	78.91	78.04	73.51	
22000	58.04	57.21	59.24	57.67	79.54	82.80	79.70	76.13	78.48	80.72	79.70	75.02	
24000	55.44	62.44	59.14	44.75	81.18	85.61	80.80	77.14	80.37	82.28	82.03	76.80	
26000	57.85	61.98	62.52	49.05	83.02	86.31	82.70	80.06	83.34	85.46	84.48	80.07	
28000	60.42	62.81	62.56	55.89	85.01	88.52	85.25	81.25	84.69	87.32	86.04	80.70	
30000	59.74	67.91	61.47	49.85	85.60	89.44	85.78	81.58	86.54	89.33	87.92	82.39	
32000	62.32	65.73	63.52	57.72	84.90	88.06	85.56	81.06	87.65	90.15	88.55	84.25	
34000	63.60	68.72	65.44	56.64	84.96	88.28	85.79	80.82	87.92	90.96	89.55	83.26	
36000	64.70	72.76	65.55	55.80	85.13	88.11	85.42	81.87	87.34	89.36	88.08	84.57	
38000	63.73	69.71	65.50	55.98	85.65	89.40	85.33	82.21	87.90	89.62	89.28	84.79	
40000	71.77	70.26	67.52	77.54	85.02	88.46	85.83	80.76	87.69	89.88	88.17	85.04	
42000	67.62	71.46	69.78	61.63	85.12	89.47	84.89	80.99	87.58	90.27	87.99	84.48	
44000	69.25	71.82	66.45	69.47	84.79	87.91	84.47	81.99	87.57	90.56	88.08	84.08	
46000	66.42	71.26	71.02	56.99	85.04	89.01	84.45	81.66	87.15	89.48	88.64	83.33	
48000	66.19	70.46	66.96	61.15	85.05	88.37	85.10	81.66	87.21	89.27	88.42	83.95	
50000	70.33	68.86	70.35	71.78	84.59	88.73	84.78	80.27	87.42	90.69	87.68	83.91	

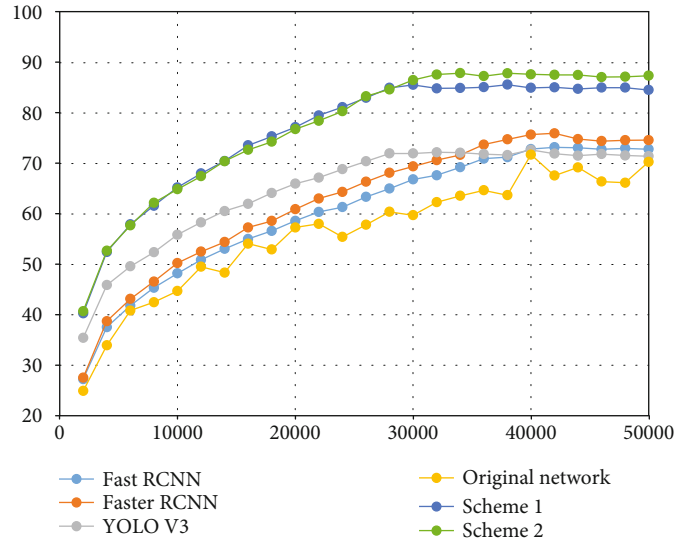


FIGURE 13: mAP results and comparison with other methods (%).

The shift transformation is given as

$$\begin{cases} x' = x_i + y_i \tan \theta_2, \\ y' = x_i \tan \theta_2 + y_i. \end{cases} \quad (17)$$

where θ_2 is the shift angle.

The above three methods are selected randomly to transform the original image, and the total number is augmented to 30000.

5. Experiments Results

The method proposed in this paper is going to be used on an underwater remote operated vehicle (ROV) for fishing marine products. The robot is about 1 m long, 0.8 meters wide, and weighs 90 kg. The method of collecting marine products is adsorption type; the design and real robot are shown in Figure 10. The robot is remote operated; our team is going to reconstruct the ROV to semiautonomous, so the key technology is how to detect and locate the objects.

5.1. Detection Comparison. The GPU used in these computations is NVIDIA GTX 1080ti, and the total number of images is 30000, which are labeled one by one artificially. And in deep learning, 8520 images are used for training, 8530 for validation, and 12950 for test. In object detection, Precision, Recall, and Mean Average Value are commonly used to assess the accuracy; the definition is shown in Figure 11.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN} \end{aligned} \quad (18)$$

Mean Average Precision is the mean value of precision of all the detection classes, which is widely used to evaluate the detection system. In this paper, the dataset is prepared in Pascal VOC form, the results obtained from Fast RCNN [6] and Faster RCNN [27] are shown in Figure 12, and the concrete data is shown in (Table 1 and Table 2).

In order to make clear about the convergence of different methods, the mAP values vs. iteration times are shown in Figure 13.

From the above results and comparison, it can be seen that the detection accuracy of Faster RCNN is better than the other methods, but the difference is not very large. Compared with the original YOLO V3 method [30], the proposed method can give more accurate detection, and the scheme 2 is more effective. The convergence of the methods is different; the YOLO V3 methods convergent after the 28000 iteration times, which is earlier than Fast RCNN and Faster RCNN. After 40000 times iteration, all the methods cannot improve the detection accuracy, the reason is lack of the underwater samples of the dataset, and the images of the dataset are similar, especially the background of the images is the same. This is the main reason for underwater object detection, the underwater data in deep sea is too difficult to obtain.

The original network proposed in this paper is not stable; the results fluctuated with the iteration times increasing. The modified schemes are proposed to improve the stability and accuracy, as shown in Figure 13. Compared with the other typical methods, our proposed methods can give a more accurate result.

The loss function curves are shown in Figure 14, the loss values of all of the methods are convergent, and the loss values amplitude of the YOLO V3 methods are smaller compared with Fast RCNN [6] and Faster RCNN [27]; the convergent speed of the proposed methods are slower than the original YOLO V3 method [30].

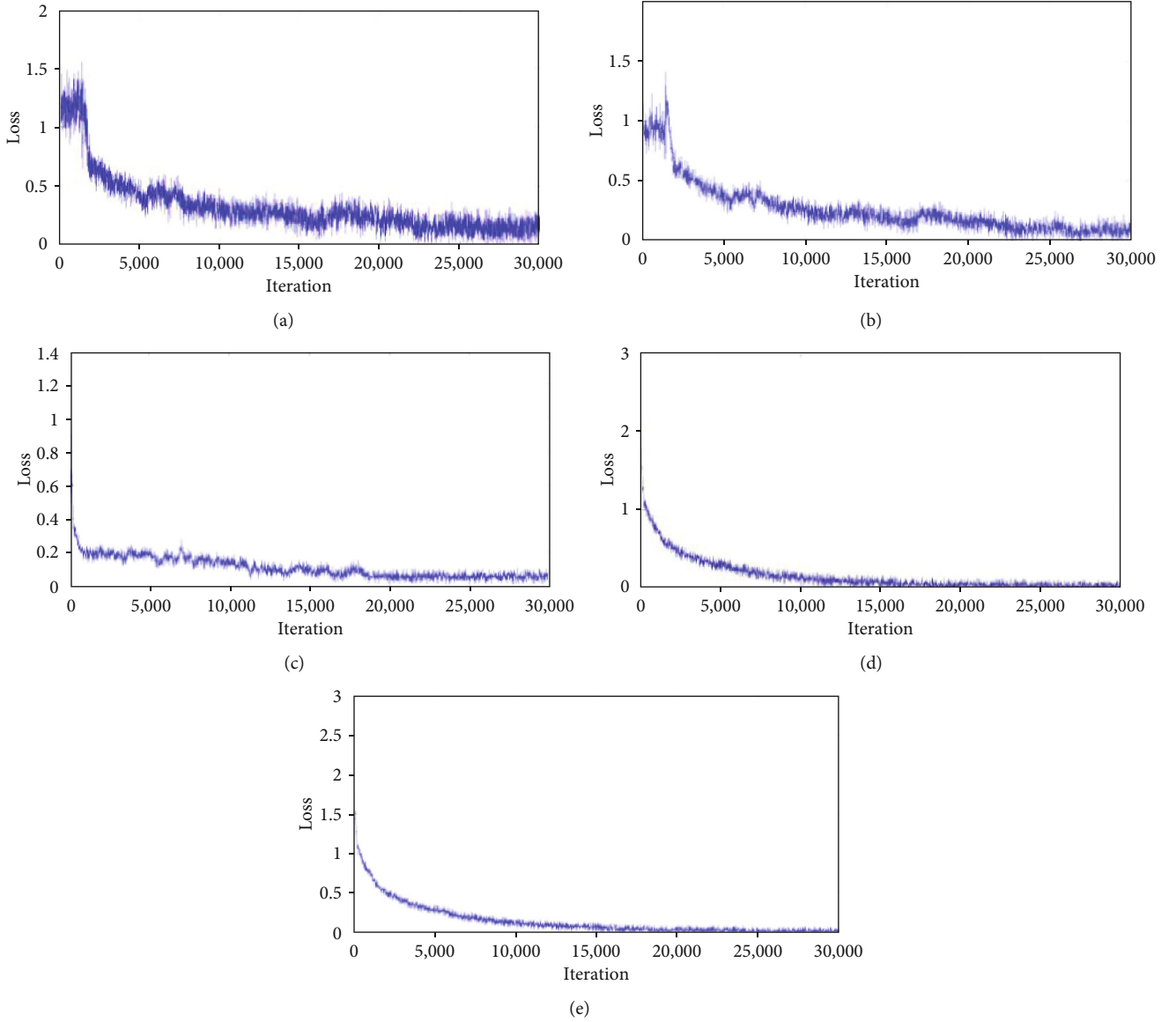


FIGURE 14: The loss curves of different methods. (a) Fast RCNN. (b) Faster RCNN. (c) YOLO v3. (d) Scheme 1. (e) Scheme 2.

TABLE 3: Detection speed of different methods (IoU = 0.7, learning rate = 0.001).

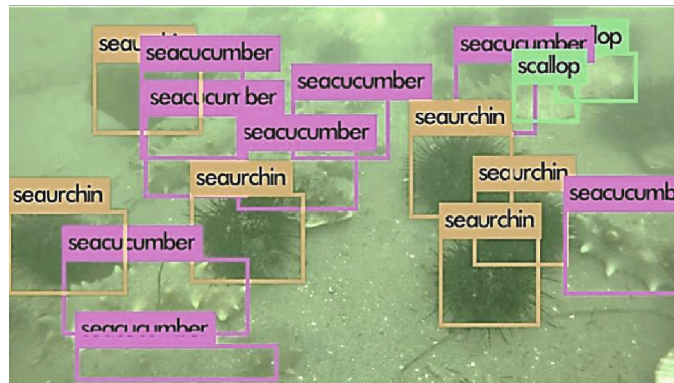
Approach	Fast RCNN	Faster RCNN	YOLO V3	Scheme 1	Scheme 2
Time cost (ms)	96	85	20	22	19

For object detection, the accuracy of all of the above methods are enough for application, the real-time detection is more important, and the detection speed is shown in Table 3.

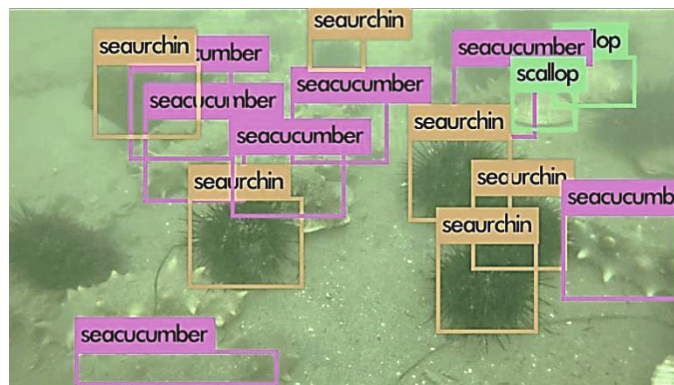
It is clear that the YOLO V3 [30] methods have a very fast detection speed, almost four times faster than Faster RCNN[27]. Based on the accuracy and detection speed analysis, the scheme 2 is better than the other methods, which has the same accuracy with the Faster RCNN, and the detection speed of this method is around 50FPS, even on a NVIDIA TX2 card, the detection speed can reach 17FPS, it is enough for real application.

5.2. Detection Results. The following typical images are used to testify the method (scheme 2) proposed in this paper, the images are provided by the “Underwater Robot Picking Contest”, and some images are filmed by the underwater ROV.

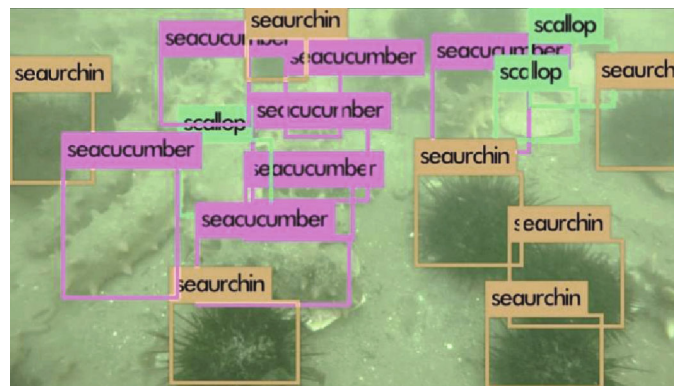
The method scheme 2 is better in underwater detection because of retaining more representation information, the comparison is shown in Figure 15, (a) and (b) are the same image, and the scheme 2 method can detect the sea cucumber and the sea urchin in the lower-left corner, but the original missed the objects. In (c) and (d), the left sea cucumber is missed by the original YOLO V3 [30] method too, so this method is more effective obviously. And from the image (a)



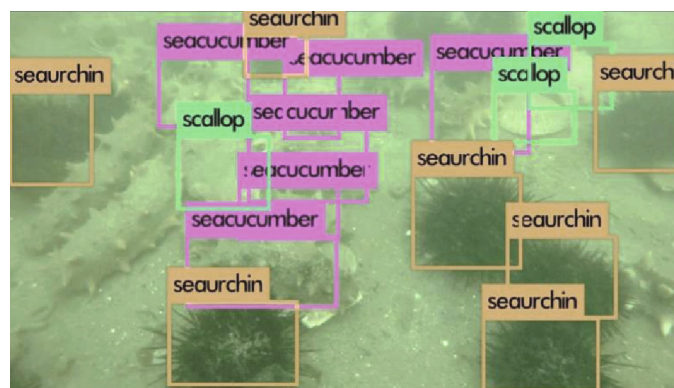
(a)



(b)



(c)



(d)

FIGURE 15: The detection comparison between YOLO V3 and the scheme 2 method. (a) Scheme 2. (b) YOLO V3. (c) Scheme 2. (d) YOLO V3.



FIGURE 16: The detection results of scheme 2 method.

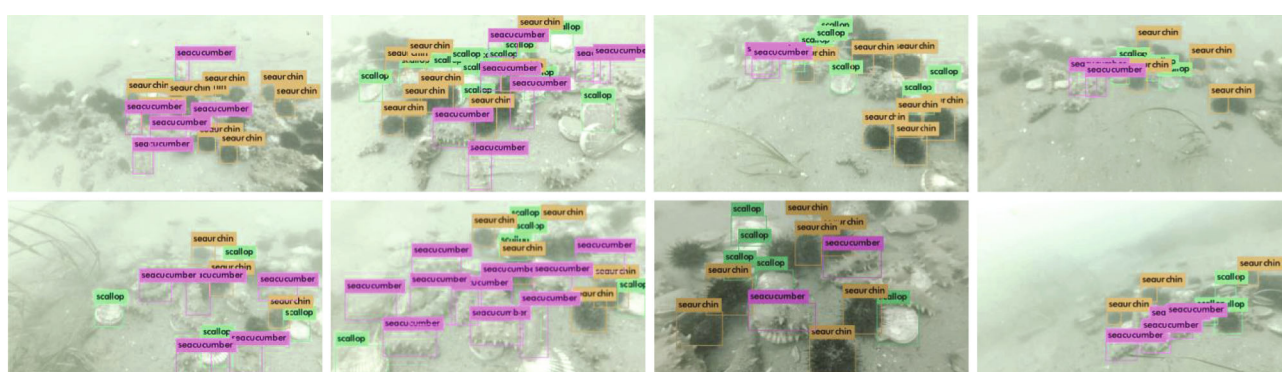


FIGURE 17: The detection applied in the ROV.

detection, we can see that the sea cucumber covered by the sands in the lower-left corner can be detected too, which is difficult to detect by human vision.

In order to verify the method, 8 images are chosen in the experiment; the detection results are shown in Figure 16.

The training model is applied in the ROV to test the detection effect, the weather is cloudy, and the sea water is very turbid; the real-time detection results are presented in Figure 17.

As seen in Figure 15, some of the objects are missed to be detected, the reason is that the dataset is not large enough, especially the images of the dataset are very similar; the light and the background are simple, so when the trained model is used to detect in the other sea area or under different environment conditions, the detection accuracy is going to reduce more or less, so our team is planning to film more underwater images in different sea area and under different conditions to make the dataset more plentiful, so as to achieve the perfect underwater detection.

6. Conclusion

Considering the underwater vision characteristics, some new image processing procedures are proposed to deal with the low contrast and the weakly illuminated problems. A deep CNN method is proposed to achieve the detection and classification of marine organisms, which is commonly recognized

as the fastest object detection method. The underwater vision is in low quality, and the objects are always overlapped and shaded, so the original YOLO V3 [30] method is not very effective for underwater detection; two methods are proposed to deal with these problems. Through detection results comparison with the other methods, the scheme 2 can give a better detection. The trained model is used to assist the ROV to detect underwater objects; although some of the objects are missed, the effectiveness and capability of the proposed method are obviously verified by the qualitative and quantitative evaluation results. The proposed method is suitable for our underwater robot to detect the objects, which is not better than the typical methods for the other dataset. And dropout layers and other technologies are not significant in this model; the reconstruction of the network by using a more complicated algorithm would be more effective.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

No potential conflict of interest was reported by the authors.

Acknowledgments

We would like to express our gratitude for support from the National Key R&D Program of China (Grant No. 2018YFC0309402) and the Fundamental Research Funds for the Central Universities (Grant No. HEUCF180105).

References

- [1] E. Y. Lam, "Combining gray world and retinex theory for automatic white balance in digital photography," in *Proceedings of the Ninth International Symposium on Consumer Electronics, 2005. (ISCE 2005)*, pp. 134–139, Macau, Macau, June 2005.
- [2] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin Institute*, vol. 310, no. 1, pp. 1–26, 1980.
- [3] J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2207–2214, 2010.
- [4] R. Hummel, "Image enhancement by histogram transformation," *Computer Graphics and Image Processing*, vol. 6, no. 2, pp. 184–195, 1977.
- [5] K. Zuiderveld, *Contrast limited adaptive histogram equalization[M]// graphics gems IV*, Academic Press Professional, Inc., 1994.
- [6] A. S. A. Ghani and N. A. M. Isa, "Enhancement of low quality underwater image through integrated global and local contrast correction," *Applied Soft Computing*, vol. 37, no. C, pp. 332–344, 2015.
- [7] C. Li and J. Guo, "Underwater image enhancement by dehazing and color correction," *Journal of Electronic Imaging*, vol. 24, no. 3, article 033023, 2015.
- [8] M. Braik, A. Sheta, and A. Ayesh, "Image enhancement using particle swarm optimization," *Journal of Intelligent Systems*, vol. 2165, no. 1, pp. 99–115, 2007.
- [9] E. H. Land, "The Retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–128, 1977.
- [10] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 4572–4576, Paris, France, Oct. 2014.
- [11] J. Perez, A. C. Attanasio, N. Nechiporenko, and P. J. Sanz, "A deep learning approach for underwater image enhancement," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pp. 183–192, Springer, Cham, 2017.
- [12] C. C. Chang, J. Y. Hsiao, and C. P. Hsieh, "An Adaptive Median Filter for Image Denoising," in *2008 Second International Symposium on Intelligent Information Technology Application*, pp. 346–350, Shanghai, China, Dec. 2008.
- [13] C. J. Prabhakar and P. U. P. Kumar, "Underwater image denoising using adaptive wavelet subband thresholding," in *2010 International Conference on Signal and Image Processing*, pp. 322–327, Chennai, India, December 2010.
- [14] D. M. Kocak and F. M. Caimi, "The current art of underwater imaging – with a glimpse of the past and vision of the future," *Marine Technology Society Journal*, vol. 39, no. 3, pp. 5–26, 2005.
- [15] M. Zhang and B. K. Gunturk, "Multiresolution bilateral filtering for image denoising," *IEEE Transactions on Image Processing*, vol. 17, no. 12, pp. 2324–2333, 2008.
- [16] M. Mäkitalo and A. Foi, "Optimal inversion of the generalized Anscombe transformation for Poisson-Gaussian noise," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 91–103, 2013.
- [17] J. L. Forand, G. R. Fournier, D. Bonnier, and P. Pace, "LUCIE: a Laser Underwater Camera Image Enhancer," in *Proceedings of OCEANS '93*, Victoria, BC, Canada, Canada, Oct. 1993.
- [18] S. Yang and F. Peng, "Laser underwater target detection based on Gabor transform," in *2009 4th International Conference on Computer Science & Education*, pp. 95–97, Nanning, China, Jul 2009.
- [19] B. Ouyang, F. Dalgleish, A. Vuorenkoski, W. Britton, B. Ramos, and B. Metzger, "Visualization and image enhancement for multistatic underwater laser line scan system using image-based rendering," *IEEE Journal of Oceanic Engineering*, vol. 38, no. 3, pp. 566–580, 2013.
- [20] P. C. Y. Chang, J. C. Flitton, K. I. Hopcraft, E. Jakeman, D. L. Jordan, and J. G. Walker, "Improving visibility depth in passive underwater imaging by use of polarization," *Applied Optics*, vol. 42, no. 15, pp. 2794–2803, 2003.
- [21] V. Gruev, J. V. D. Spiegel, and N. Engheta, "Advances in integrated polarization image sensors," in *2009 IEEE/NIH Life Science Systems and Applications Workshop*, pp. 62–65, Bethesda, MD, USA, Apr 2009.
- [22] Y. Li and S. Wang, "Underwater polarization imaging technology," in *2009 Conference on Lasers & Electro Optics & The Pacific Rim Conference on Lasers and Electro-Optics*, pp. 1–2, Shanghai, China, Aug 2009.
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [24] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, Dec 2015.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun 2016.
- [27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, Jun 2014.
- [28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, June 2016.
- [30] J. Redmon and A. Farhadi, *YOLOv3: An Incremental Improvement*, 2018.

- [31] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *The Twelfth Color Imaging Conference, Is&T-The Society for Imaging Science and Technology*, pp. 37–41, Scottsdale, AZ, USA, 2004.
- [32] C. Li, J. Guo, F. Porikli, and Y. Pang, "LightenNet: a convolutional neural network for weakly illuminated image enhancement," *Pattern Recognition Letters*, vol. 104, pp. 15–22, 2018.
- [33] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015.
- [34] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multi-scale Retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.
- [35] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2017.