water

MDPI

Article

An Effective Method for Underwater Biological Multi-Target Detection Using Mask Region-Based Convolutional Neural Network

Zhaoxin Yue [1,2,3], Bing Yan [4], Huaizhi Liu [5] and Zhe Chen [4,*]

1 School of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China; yzx10000@163.com
2 Key Laboratory of River Basin Digital Twinning of Ministry of Water Resources, China Institute of Water Resources and Hydropower Research, Beijing 100038, China
3 Industrial Perception and Intelligent Manufacturing Equipment Engineering Research Center of Jiangsu Province, Nanjing Vocational University of Industry Technology, Nanjing 210023, China
4 College of Computer and Information, Hohai University, Nanjing 211100, China; hhucomputer@163.com
5 CSIC PRIDe (Nanjing) Atmospheric & Oceanic Information System Co., Ltd., Nanjing 211106, China; liuhz1985@126.com
* Correspondence: chenzhe@hhu.edu.cn

Abstract: Underwater creatures play a vital role in maintaining the delicate balance of the ocean ecosystem. In recent years, machine learning methods have been developed to identify underwater biologicals in the complex underwater environment. However, the scarcity and poor quality of underwater biological images present significant challenges to the recognition of underwater biological targets, especially multi-target recognition. To solve these problems, this paper proposed an ensemble method for underwater biological multi-target recognition. First, the CutMix method was improved for underwater biological image augmentation. Second, the white balance, multiscale retinal, and dark channel prior algorithms were combined to enhance the underwater biological image quality, which could largely improve the performance of underwater biological target recognition. Finally, an improved model was proposed for underwater biological multi-target recognition by using a mask region-based convolutional neural network (Mask-RCNN), which was optimized by the soft non-maximum suppression and attention-guided context feature pyramid network algorithms. We achieved 4.97 FPS, the mAP was 0.828, and the proposed methods could adapt well to underwater biological multi-target recognition. The recognition effectiveness of the proposed method was verified on the URPC2018 dataset by comparing it with current state-of-the-art recognition methods including you-only-look-once version 5 (YOLOv5) and the original Mask-RCNN model, where the mAP of the YOLOv5 model was lower. Compared with the original Mask-RCNN model, the mAP of the improved model increased by 3.2% to 82.8% when the FPS was reduced by only 0.38.

Keywords: underwater biological multi-target recognition; CutMix; image fusion; deep learning; Mask-RCNN

Citation: Yue, Z.; Yan, B.; Liu, H.; Chen, Z. An Effective Method for Underwater Biological Multi-Target Detection Using Mask Region-Based Convolutional Neural Network. *Water* **2023**, *15*, 3507. https://doi.org/10.3390/w15193507

Received: 30 August 2023
Revised: 6 October 2023
Accepted: 7 October 2023
Published: 8 October 2023

1. Introduction

Underwater creatures have a strong relationship with the hydrological, life, and ecology environment and affect the delicate balance of the ocean ecosystem. However, the scarcity and poor quality of underwater biological images present significant challenges to the recognition of underwater biological targets [1,2]. For instance, image quality relies on the underwater vision system, which can execute image processing, feature extraction, and recognition tasks [3]. However, due to the unique features of underwater environments, the underwater data acquisition equipment has demanding specifications, and the collected images may present problems such as distortion, small data scale, and target category imbalance [4]. Furthermore, the capture and dispersal of light by water molecules and

many suspended objects lead to blurred images, color distortion, image contrast reduction, and other issues that make the recognition of underwater targets difficult. Apart from the aforementioned challenges, the collected single images often contain multiple recognition targets, which can be problematic due to occlusion, overlap, and other issues. Moreover, these targets are typically small in size and occupy a relatively small portion of the image.

In recent years, many machine learning methods have been developed to identify underwater biologicals in complex underwater environments. Traditional methods first preprocess the image to make the target information more prominent, then classify or recognize the target to obtain good results [5,6]. For example, Wang et al. [5] investigated a novel regional saliency model for underwater object detection and obtained good results under both uniform and uneven illumination conditions. Despite the previous methods having achieved some advancements in specific scenarios, extracting feature information is complicated and time-consuming; thus, they are unable to fulfill the demands for real-time performance. In addition, most conventional approaches are employed for single-target recognition. Unlike traditional methods, deep learning methods possess greater capacity for expression and achieve higher accuracy in target recognition but require large-scale datasets [7–9]. With the rapid development of information technology, the utilization of deep learning-based target recognition technology has gradually been applied more to underwater environments. Song et al. [1] believe that the intricate and constantly changing underwater environment, along with the lack of adequate datasets, contribute to subpar optical imaging quality, which in turn might result in overfitting issues in target recognition models. Zhou et al. [2] posited that certain challenges encountered in the marine environment such as complex backgrounds and low illumination could lead to subpar picture quality. Additionally, the presence of small targets and multiple targets pose difficulties for target recognition.

In summary, the challenges of underwater biological multi-target detection arise from limited datasets, the poor quality of images, and machine learning model selection. Existing research has struggled to achieve satisfactory recognition accuracy, particularly for multi-target and multi-class recognition tasks. Consequently, it is vital to enhance, reconstruct, and augment underwater images, and then select a robust deep learning model for multi-target detection.

In order to address these issues, this paper proposed an image augmentation, image enhancement method, and improved Mask-RCNN combined for underwater biological multi-target detection and used three types of underwater creatures as examples of recognition targets: starfish, sea urchin, and sea cucumber. First, image augmentation and enhancement algorithms were applied to an original sample set to expand the dataset and improve image quality. Then, an improved Mask-RCNN model was proposed to effectively realize the underwater biological multi-target recognition. The flow of the proposed method is presented in Figure 1.

The primary contributions of this research include:

(1) A commonly used image augmentation method and improved CutMix algorithm were applied to expand dataset samples and solve the overfitting problem in deep learning training due to the class imbalance that occurs in multi-target recognition.
(2) A novel method for underwater image enhancement based on simple weighted fusion was proposed to enhance the image quality in complex underwater environments.
(3) The Mask-RCNN model was improved to prevent problems such as missed and false detections, thereby improving the identification accuracy of the model. The results revealed that the proposed model exhibited superior performance in comparison to the other models.

The subsequent sections of this paper are structured as follows. Section 2 briefly reviews the related works, and Section 3 proposes the processes and fundamental structure of the proposed method. The data used in the study are described in Section 4, and the results of the different methods are presented in Section 4. Section 5 discusses the findings of the study and offers our conclusions.

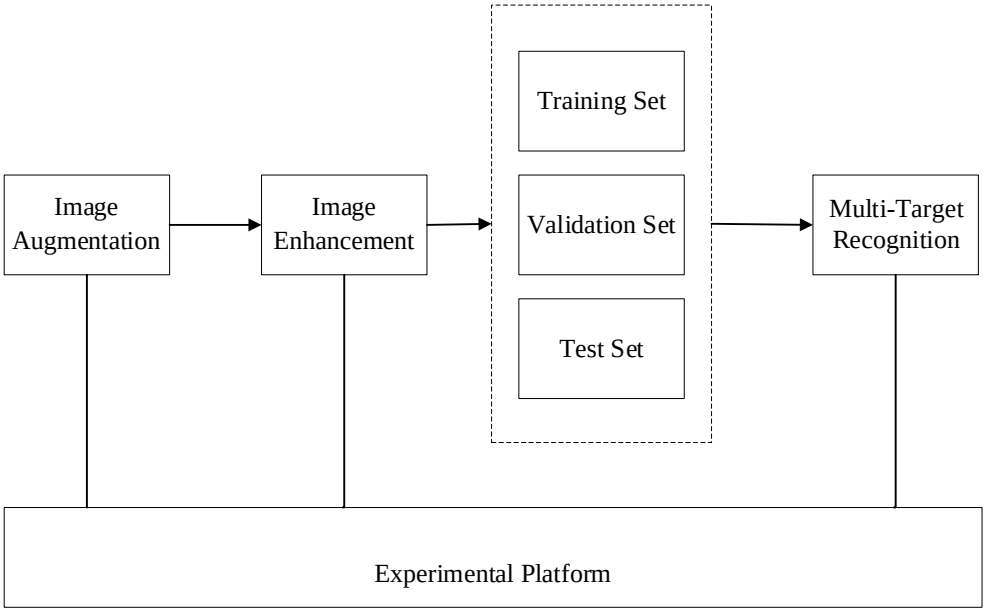


Figure 1. Framework of our proposed method.

2. Related Works

Current research in underwater target recognition primarily focuses on three aspects: image augmentation, image enhancement, and target recognition. Related works are briefly reviewed in this section.

2.1. Underwater Image Augmentation

Image augmentation is a strategy used for increasing the quantity and diversity of images in a limited dataset and to extract more information from images [10]. Common image augmentation methods include (1) geometric transformation [11], which changes the perspective of the original dataset and improves the robustness and recognition accuracy of the model, (2) optics-based transformation augmentation [12], which involves two main processes of illumination transformation and color space transformation, and (3) the generation of new samples without changing the position of the original data to promote the robustness of the model by adding noise, changing the image illumination, and applying sharpness transformation.

Unlike the commonly used image augmentation methods, underwater image augmentation methods are generally effective in various scenarios, but may lose efficacy for underwater biological datasets due to the complex underwater environment, and labeling a large number of underwater images is both expensive and time-consuming. For instance, Huang et al. [13] proposed three data augmentation methods designed for underwater imaging that validated the effectiveness of marine organism detection and recognition. Noh et al. [14] developed a data augmentation method that used the unique properties of light to improve the accuracy of object detection in underwater environments and thus reduced the training effort. Despite these methods performing well in underwater image augmentation, it should be noted that labeling numerous underwater images is costly and time-consuming. Currently, the CutMix augmentation strategy has been applied for image augmentation in other fields and has obtained good effects. For example, Yun et al. [15] put forward the CutMix augmentation strategy, where patches were cut and pasted into training images. In their strategy, ground truth labels were mixed in proportion to the patched areas. Through the efficient utilization of training pixels and the retention of the regularization effect from regional dropout, the CutMix augmentation strategy consistently demonstrated superior performance compared to other state-of-the-art augmentation strategies on tasks. With respect to the unique underwater environment, this study introduced the CutMix augmentation strategy for marine underwater biological image augmentation.

2.2. Underwater Image Enhancement

Existing methods used in underwater image enhancement can be divided into the spatial domain- [16,17], transformation domain- [18,19], color constancy- [20], and deep learning-based [21]. Spatial domain-based techniques include logarithmic transformation, contrast stretching, histogram equalization, and sharpening. Transformation domain-based methods transform the original image to the frequency domain through wavelet, Fourier, and color space transforms. Color constancy-based methods employ the idea that when the color of light on the surface of the object changes within a certain range, the perception of the surface color of the object basically remains unchanged. Finally, deep learning techniques employ deep neural networks to learn features for image enhancement.

For instance, Ghani et al. [16] integrated a color model with Rayleigh distribution to improve underwater image quality, and the results revealed that the proposed model exhibited superior performance in comparison to other models in terms of contrast and noise reduction. Vasamsetti et al. [17] proposed a wavelet-based variational enhancement method for underwater image enhancement, which might be useful in boosting the development of underwater detection. Iqbal et al. [18] introduced the Laplace decomposition method for underwater image enhancement and obtained good enhancement results. Jobson et al. [19] introduced a solution called the multiscale retinex method to address the disparity between color images and the human perception of scenes. Their approach successfully achieved simultaneous enhancements in dynamic range compression, color consistency, and lightness rendition. Li et al. [20] proposed the WaterGAN method for real-time color correction of monocular underwater images, which had been successfully applied to underwater image enhancement. Li et al. [21] introduced a CNN model trained using the UIEB to enhance underwater images, and the results showcased the versatility of the constructed UIEB.

2.3. Underwater Target Recognition

The aforementioned traditional recognition methods are complicated and time-consuming, thereby hardly meeting the requirements of underwater target recognition. Thus, deep learning models such as convolutional neural networks have been developed for target recognition. Existing deep learning-based target detection methods can be divided into two main categories: candidate window-based for higher detection accuracy and end-to-end target detection for better real-time performance. Because convolutional neural networks and deep learning techniques have exhibited impressive performance on datasets such as Pascal VOC [3] and ImageNet [22], these methods and their improved versions are increasingly being applied to underwater target recognition. Mittal et al. [23] presented a survey of deep learning techniques for performing underwater image classification to identify their similarities and differences. Chen et al. [24] proposed a modified you-only-look-once version 4 (YOLOv4) neural network for underwater target recognition that improved the target accuracy and recognition speed. Yeh et al. [25] introduced a deep model that combined the learning of color conversion and object detection for underwater images, where the image color conversion module transformed color images to corresponding grayscale images to solve the problem of underwater color absorption. This method improved the object detection performance while reducing computational complexity.

In conclusion, deep learning has strong generalization capabilities [26,27] and is effective in fish target recognition, benefitting from clear and rich underwater datasets [28–31]. For instance, Shi et al. [28] introduced an improved Faster-RCNN algorithm for underwater biological detection, and the proposed model performed better than the YOLOv4 and Faster-RCNN models. Li et al. [29] designed an underwater biological detection algorithm that integrated the channel attention mechanism, and the proposed model was superior to the original YOLOv4 algorithm. However, for multiple targets, realizing target recognition under complex underwater environments is difficult because of the small datasets, uneven sample categories, and poor underwater image quality. Few existing methods can achieve high recognition accuracy, particularly for multiple targets. For example, Li et al. [32] proposed an improved CME-YOLOv5 network to detect fish in dense groups and small

targets, and the proposed algorithm exhibited good detection performance when applied to densely spaced fish and small targets.

3. Proposed Method

In order to address the issues above-mentioned, this study proposed a novel approach for underwater biological multi-target recognition. First, an improved CutMix-based image augmentation method was proposed to expand the dataset. Second, a fusion-based underwater image enhancement algorithm was presented to enhance the image quality. Finally, a Mask-RCNN model was developed to realize underwater biological multi-target recognition.

3.1. Improved CutMix Based Underwater Image Augmentation

We first introduce the idea of CutMix, which enables the background to be converted into an underwater background color before filling the sample in the cropped area to achieve a better fusion effect. The flow of the improved CutMix algorithm is presented in Figure 2. This study used starfish image augmentation as an example, and the specific steps are as follows:

(1) Select the images of starfish with good clarity from the dataset and use the image segmentation techniques to segment the starfishes, resulting in dataset X where the background is black and only the main body of the starfish is retained.
(2) Select images with fewer organisms in the dataset to obtain dataset Y, in order to avoid or reduce overlap with other types of organisms when expanding the number of starfish in the dataset.
(3) Randomly select images from dataset Y without putting them back, calculating the color channel components of the image background R, G, and B. Based on the requirements, add a certain number of starfish, randomly select multiple starfish images from dataset X, and convert their black background into the calculated color components to achieve a better fusion effect.
(4) Resize the selected starfish images and adopt the image fusion algorithm to fuse the starfish images into the images selected from dataset Y in step 3, thus completing the image segmentation of the starfish samples.
(5) Repeat steps 3 and 4 until there are no more images in dataset Y.

3.2. Image Fusion-Based Underwater Image Enhancement

Underwater images often present color decay, low contrast, and blurred details. In order to address these issues and improve the accuracy of multi-target recognition, this study combined the white balance algorithm (WBA), multiscale retinal with color restoration (MSRCR), and dark channel prior (DCP) algorithms to enhance the underwater image quality. The flow of image fusion-based underwater image enhancement is illustrated in Figure 3.

3.2.1. White Balance Algorithm

The white balance algorithm (WBA) can effectively solve the color bias problem. Common WBAs include the gray world (GW) and perfect reflector (PR) algorithms [33]. The GW algorithm performs very fast calculations, but often fails when the image color transformation is not obvious. The PR algorithm also performs fast calculations, but the selection of ratio parameters significantly affects the effectiveness of image processing. The white balance processing results using the two algorithms are shown in Figure 4. This study used the GW algorithm for underwater image enhancement, where the mean values of the R_{avg}, G_{avg}, and B_{avg} channels were calculated as follows [24]:

$$\begin{cases} R_{avg} = \frac{1}{m_{number}} \sum_{0}^{m_{number}-1} R \\ G_{avg} = \frac{1}{m_{number}} \sum_{0}^{m_{number}-1} G \\ B_{avg} = \frac{1}{m_{number}} \sum_{0}^{m_{number}-1} B \end{cases} \quad (1)$$

where the R, G, and B values denote the red, green, and blue components of each pixel, respectively, R_{avg}, G_{avg}, and B_{avg} are the respective averages of all pixels, and m_{number} is the quantity of pixels in the image. Based on Equation (1), the mean value of RGB can be obtained as

$$P = (R_{avg} + G_{avg} + B_{avg})/3 \quad (2)$$

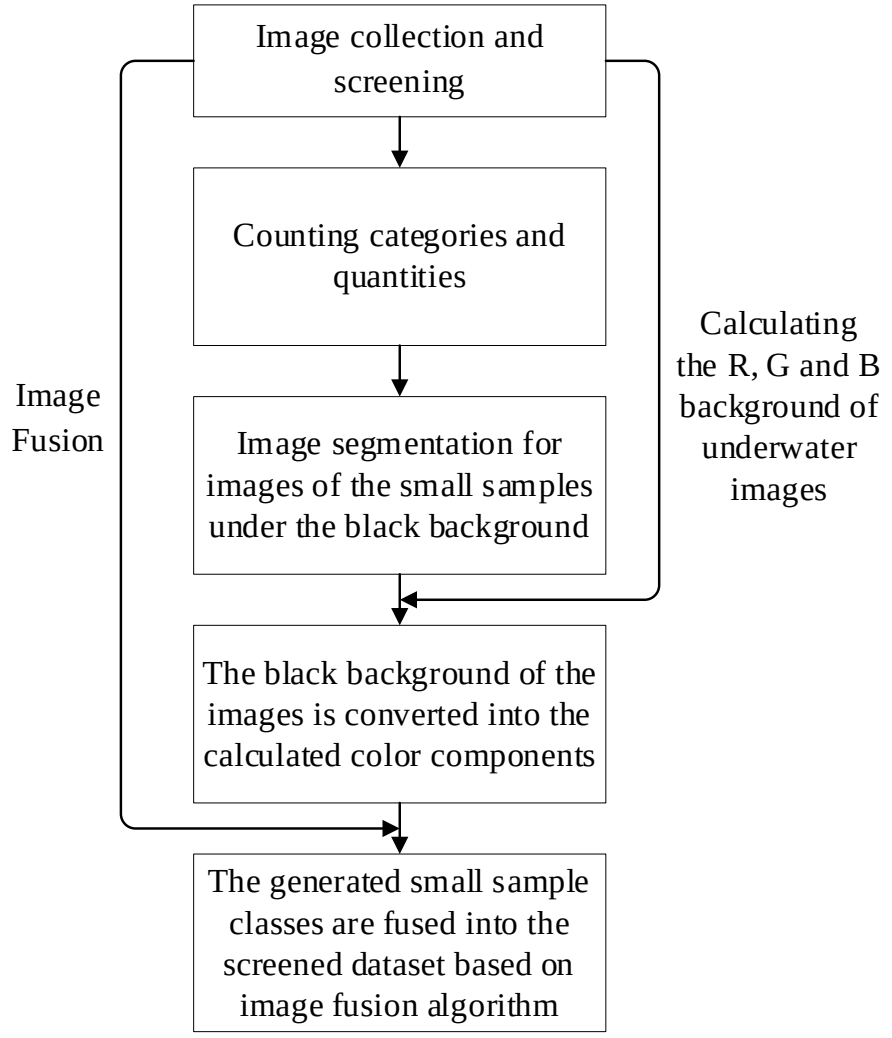


Figure 2. Flow of the improved CutMix algorithm.

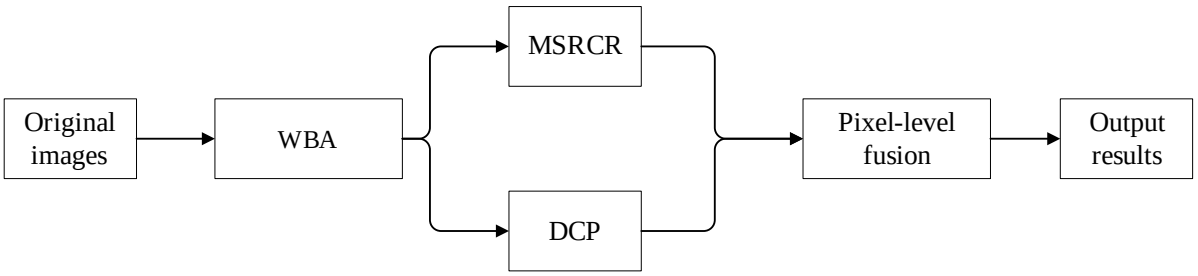


Figure 3. Flow of image fusion-based underwater image enhancement.

(**a**) (**b**) (c)

Figure 4. White balance algorithm processing for underwater images using the GW and PR algorithms: (**a**) original image; (**b**) result of using the GW algorithm; (**c**) result of using the PR algorithm.

Next, the relative gain of each channel with respect to P can be determined in the following manner:

$$\begin{cases} P_R = P/R_{avg} \\ P_G = P/G_{avg} \\ P_B = P/B_{avg} \end{cases} \tag{3}$$

The pixel values are subsequently modified individually based on the calculated gains:

$$\begin{cases} R_{update} = P_R * R \\ G_{update} = P_G * G \\ B_{update} = P_B * B \end{cases} \tag{4}$$

where R_{update}, G_{update}, and B_{update} represent the new values for each pixel.

The PR algorithm: Assume that there exists a pure white pixel in the image, and then use this as a reference to perform automatic white balance on the image. As a result, the pure white pixel is defined as the maximum values of R, G, and B. The algorithm for this process is as follows:

(1) Calculate the sum of the R, G, and B values for each pixel and save the coordinates of the brightest point in the image.
(2) Calculate the threshold T from the top 10% of the sum of R, G, and B or other ratio reference points.
(3) Traverse through each point in the image and calculate the cumulative sum and average of the R, G, and B components for all points where the sum of R, G, and B is greater than the threshold T.
(4) Calculate the gain coefficients of each channel in the image according to the brightest point value and the average calculated results in the previous step.
(5) Quantize each pixel to [0, 255] according to the gain coefficients.

3.2.2. Multiscale Retinal with Color Restoration

Jobson et al. [19] presented the retinex theory for image enhancement. The MSRCR algorithm is a development of the retinex algorithm, making it suitable for a wide range of applications. This approach utilizes the principle of color constancy and its mathematical model, which can be solved through calculations [34]. The underlying assumption is that an optimal image can be represented as

$$I(a,b) = R(a,b) \cdot L(a,b) \tag{5}$$

where L represents the luminance component, which is independent of the scene and determines the dynamic range of the image, while R denotes the reflection component, which remains independent of scene lighting. In their study, Jobson et al. [19] put forward the single-scale retinex (SSR) algorithm, which relied on the principles of homomorphic filtering. Nonetheless, this approach failed to address both the dynamic range and tonal

contrast. In order to address these challenges, Rahman et al. [35] introduced the multi-scale retinex (MSR) and MSRCR algorithms.

To address the limitations of the SSR algorithm, the MSR algorithm was developed by incorporating various weighted scales of the SSR. The MSR algorithm strikes a balance between image dynamic range and color fidelity and can be expressed as

$$R_{MSR_i}(a,b) = \sum_{n=1}^{N} \omega_n \cdot (\ln I_i(a,b) - \ln(G(a,b) \cdot I_i(a,b))) \tag{6}$$

$$G(a,b) = \frac{1}{2\pi\sigma^2} e^{-\frac{a^2+b^2}{2\sigma^2}} \tag{7}$$

where $RMSR_i$ represents the resulting transformed image of the ith component image I (a, b) by employing the MSR algorithm. The wrap supporting function G (a, b) is utilized in N different scales, where N denotes the number of dimensions or scales. For practical purposes and improved calculation efficiency, it is common to employ three dimensions (N = 3) in real-world applications. Typically, small, medium, and large scales are $\sigma < 50$, $50 \leq \sigma < 100$, and $\sigma \geq 100$, respectively. Finally, ω_n denotes the weights, where $\sum_{n=1}^{N} \omega_n = 1$ in practical applications.

Note that the resulting images generated by the MSR algorithm may suffer from significant color distortions. In order to address this issue, MSRCR is proposed, which can be formulated as follows:

$$R_{MSRCR_i}(a,b) = C_i(a,b) \cdot R_{MSR_i}(a,b) \tag{8}$$

where $C_i(a, b)$ represents the color restoration function employed to adjust the proportions of the three color channels. Since we employed single-color channel imaging in MSRCR, we redefined $C_i(a, b)$ as

$$C_i(a,b) = \ln\left(\frac{I_i(a,b)}{\frac{1}{M}\sum_{(a,b)\in\Omega} I_i(a,b)}\right) \tag{9}$$

where M represents the overall number of pixels within the input image. We considered the dynamic range of the hazy image in the transmittance estimation because it can provide insights into the haze concentration and aids in accurate transmittance estimation. In addition, this prevents oversaturation in the recovered image when the transmittance and hazy image are too closely related. The results of different retinex algorithms are shown in Figure 5.

3.2.3. Dark Channel Prior Algorithm

He et al. [36] proposed the DCP algorithm. Given an arbitrary image J, its dark channel J^{dark} is given by

$$J^{dark}(a) = \min_{y\in\Omega(a)} \left(\min_{c\in\{r_{colar}, g_{colar}, b_{colar}\}} J^c(b) \right) \tag{10}$$

where J^c is a color channel of J, and $\Omega(a)$ is a local patch centered at a.

In computer vision, a main model for foggy images can be expressed as

$$I(a) = J(a) \cdot t(a) + A \cdot (1 - t(a)) \tag{11}$$

where $I(a)$ is the brightness of the observed image, $J(a)$ is the clear image, $t(a)$ is the transmittance, and A is the ambient light.

The transmittance is given by

$$t = 1 - \min_{\Omega}(\min_{c}\frac{I^c}{A^c}) \tag{12}$$

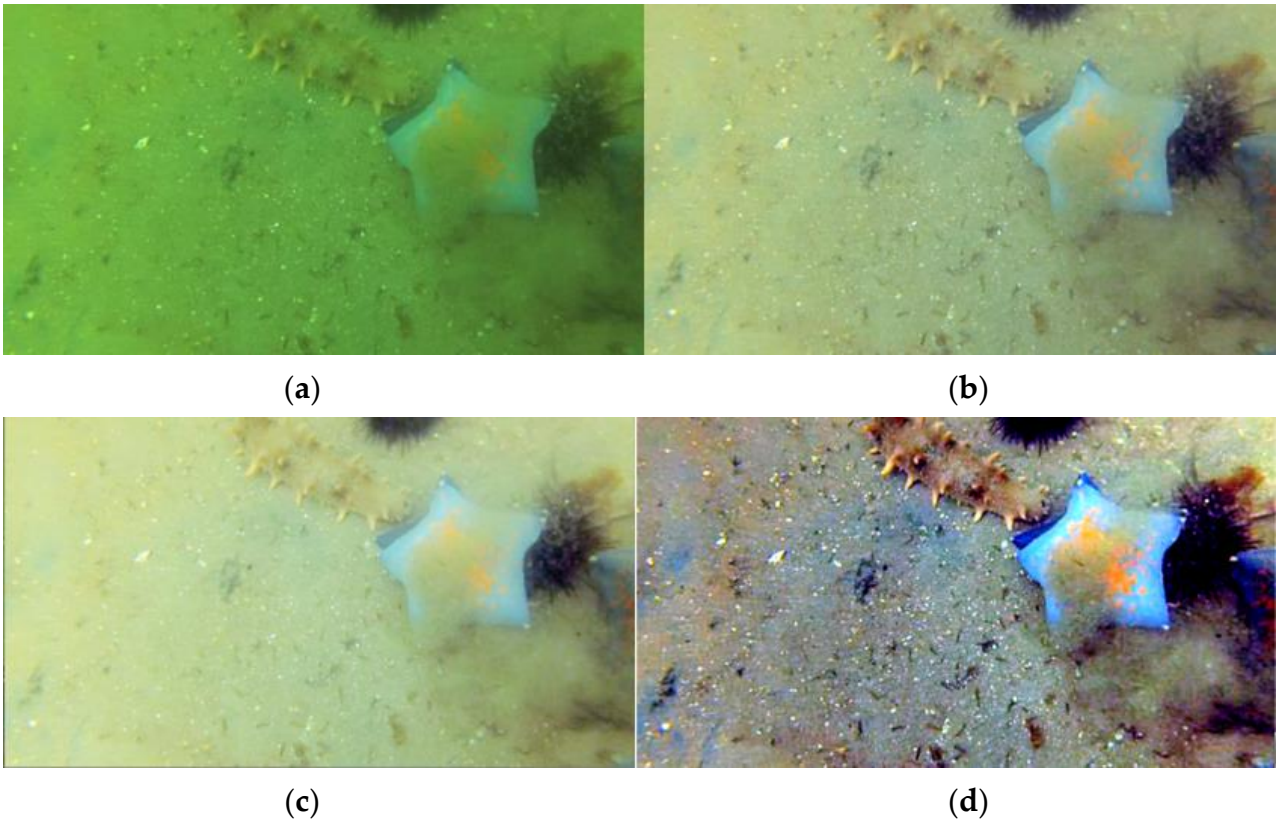


Figure 5. Results of the different retinex algorithms: (**a**) original image; (**b**) result of using SSR; (**c**) result of using MSR; (**d**) result of using MSRCR.

According to the transmittance and estimated atmospheric ambient light, a clear image can be derived from

$$J(a) = \frac{I(a) - A}{t(a)} + A \tag{13}$$

The result of using the DCP algorithm is shown in Figure 6.

Figure 6. Processing result of a defogged image: (**a**) original image; (**b**) result of using the DCP algorithm.

3.2.4. Image Fusion-Based Underwater Image Enhancement

Due to complex and variable underwater environments, the underwater image enhancement algorithm cannot easily adapt to different scenes. To solve this problem, a novel method for underwater image enhancement was designed to enhance the robustness of the algorithm. First, the GW algorithm was applied to the original image to improve the color deviation of the underwater image, and then the MSRCR and DCP algorithms were applied to enhance the image. Finally, the resulting image was generated by weighted fusion and contrast enhancement.

The image-weighted fusion formula is expressed as

$$img = img1 \cdot p + img2 \cdot (1 - p) \tag{14}$$

where $img1$ and $img2$ denote the first and second images to be fused, respectively, and p is the weight coefficient. In weight fusion, the weight coefficient is mainly set by human experience, which diminishes the effect of image fusion. Therefore, this study designed a method for image fusion based on the sum of modulus of gray difference (SMD) to adjust the weight coefficients to achieve better fusion results. SMD [37,38] is expressed as

$$SMD(k) = \sum_{a}\sum_{b} |f_k(a,b) - f_k(a+1,b)| \cdot |f_k(a,b) - f_k(a,b+1)|;\ k = 1,2,3,\cdots \tag{15}$$

where $f_k(a, b)$ represents the gray value of the kth pixel at point (a, b). $SMD(k)$ represents the result of the pixel traversal calculation for the entire image.

The MSRCR algorithm is first applied to an image to generate Image A, followed by the GW and DCP algorithms to generate Image B. Then, the sharpness values q_A and q_B of the two images are obtained by calculating them with the SMD. Finally, the fusion weight coefficients are calculated according to the sharpness of the image fusion.

The weight coefficient p is expressed as

$$p = \frac{q_A}{q_A + q_B} \tag{16}$$

A sample of the results of our underwater image enhancement is shown in Figure 7.

(**a**) (**b**)

Figure 7. Underwater image enhancement algorithm based on image fusion: (**a**) original image; (**b**) result of using the proposed algorithm.

3.3. Underwater Biological Multi-Target Recognition Based on the Improved Mask-RCNN

To address the problems of target occlusion and overlap, this study used the soft non-maximum suppression (soft-NMS) algorithm to enable the model to more effectively detect and recognize occluded objects, and then used the attention-guided context feature pyramid network (AC-FPN) to enable its application to small targets.

3.3.1. Non-Maximum Suppression Algorithm

(1) NMS Algorithm

The NMS algorithm is a post-processing technique widely used in computer vision applications [39–41]. The flow of the NMS algorithm can be described as follows:

Step 1: Arrange all of the bounding boxes in set B in descending order according to their confidence scores.

Step 2: Calculate the intersection-over-union (*iou*) of the first bounding box M, which has the highest confidence score, and the sequenced bounding boxes b_i. The *iou* is generally set manually. If *iou* (M, b_i) exceeds the rigid threshold N_t, the confidence score of b_i will be set to zero.

Step 3: Move the proposal m, with bounding box M, into the set F, which is initialized with an empty set.

Step 4: Repeat the above three steps for the remaining bounding boxes in B until complete traversal.

Set F represents the final prediction results. The NMS algorithm is given by

$$s_i = \begin{cases} s_i, iou(M,b_i) < N_t \\ s_i, iou(M,b_i) \geq N_t \end{cases} \quad (17)$$

where s_i and b_i represent the confidence score and bounding box of the ith proposal, and N_t is a constant rigid threshold that ranges between 0 and 1.

However, NMS has two problems. First, all frames are sorted by confidence, but the detection frame with the highest classification confidence is not necessarily the most accurate position. Then, when the two objects are close to each other, the intersection of the two objects and *iou* are higher than the threshold, and the phenomenon of false filtering occurs.

(2) Soft-NMS

To solve the aforementioned problems, Bodla et al. [42] proposed an improved NMS algorithm (soft-NMS) that reduces the confidence by replacing the deleted box with an *iou* greater than the threshold. It is evident that the scores for detection boxes with a higher overlap with M should decrease more significantly as they are more likely to be false positives. In this study, the pruning step could be modified according to the following guideline [42]:

$$s_i = \begin{cases} s_i, iou(M,b_i) < N_t \\ s_i(1 - iou(M,b_i)), iou(M,b_i) \geq N_t \end{cases} \quad (18)$$

This function reduces the scores of detections that surpass a threshold N_t based on the extent of their overlap with M, using a linear equation. As a result, detection boxes that are distant from M remain unaffected, while those in close proximity receive a more significant penalty.

However, this function exhibits a lack of continuity in terms of overlap, resulting in an abrupt imposition of penalties when the non-maximum suppression threshold of N_t is reached. Ideally, the penalty function should be continuous; otherwise, abrupt changes to the ranked list of detections could occur. A continuous penalty function should have no penalty when no overlap occurs, and a very high penalty when there is a high overlap. In addition, when the overlap is low, the penalty should be increased as M should not affect the scores of boxes that have a very low overlap with it. However, when the overlap of a box b_i with M becomes close to 1, b_i should be severely penalized. Accordingly, this study adopted updating the pruning step with a Gaussian penalty function. The steps of Algorithm 1 are as follows [42]:

Algorithm 1: soft-NMS

Input: $B = \{b_1, \ldots, b_n\}$, $S = \{s_1, \ldots, s_n\}$, N_t
B is the list of initial detection boxes
S contains corresponding detection scores
N_t is the NMS threshold
begin
$D \leftarrow \{\}$
While B ≠ empty **do**
$m \leftarrow$ argmax S
$M \leftarrow b_m$
$D \leftarrow D \cup M$; $B \leftarrow B - M$
for b_i in B **do**
if $iou\ (M, b_i) \geq N_t$ **then**
$B \leftarrow B - b_i$; $S \leftarrow S - s_i$
end **NMS**
$s_i \leftarrow s_i\, f(iou\ (M, b_i))$
soft-NMS
end
end
return D, S
end

3.3.2. Improved Feature Pyramid Network

(1) FPN

FPN was introduced to harness the inherent multiscale feature representation of deep convolutional networks. In particular, by incorporating a top–down pathway, FPN integrates low-resolution large-receptive-field features with high-resolution small-receptive-field features to effectively detect objects across various scales. Thus, the FPN alleviates the conflicting requirements of the feature map resolution and receptive fields. However, the following problems remain in current FPN-based approaches:

(i). The nearest neighbor interpolation method is adopted for the process of upsampling, but the high-level semantic information may not be transmitted effectively.
(ii). A lack of effective communication exists among multi-size receptive fields.
(iii). The FPN network applies four stages of backbone network output, which may not be sufficient for output multi-scale information.

(2) AC-FPN

To tackle these problems, Cao et al. [43] proposed the adaptive context feature pyramid network (AC-FPN) architecture, which takes advantage of distinctive data from different extensive receptive fields by combining features guided by attention across multiple paths. The model comprises two modules: a context extraction module (CEM) that investigates extensive contextual information from multiple receptive fields and an attention-guided module (AM) that intelligently captures the significant dependencies among objects through the implementation of the attention mechanism. The AM is introduced to address the issue of misleading localization and recognition caused by redundant contextual relations. It is composed of two submodules: the context attention (CxAM) and the content attention (CnAM) modules. These submodules are responsible for capturing discriminative semantics and locating precise positions, respectively. Notably, the AC-FPN can be easily integrated into existing FPN-based models, offering enhanced performance and flexibility. The architecture of the AC-FPN model is shown in Figure 8.

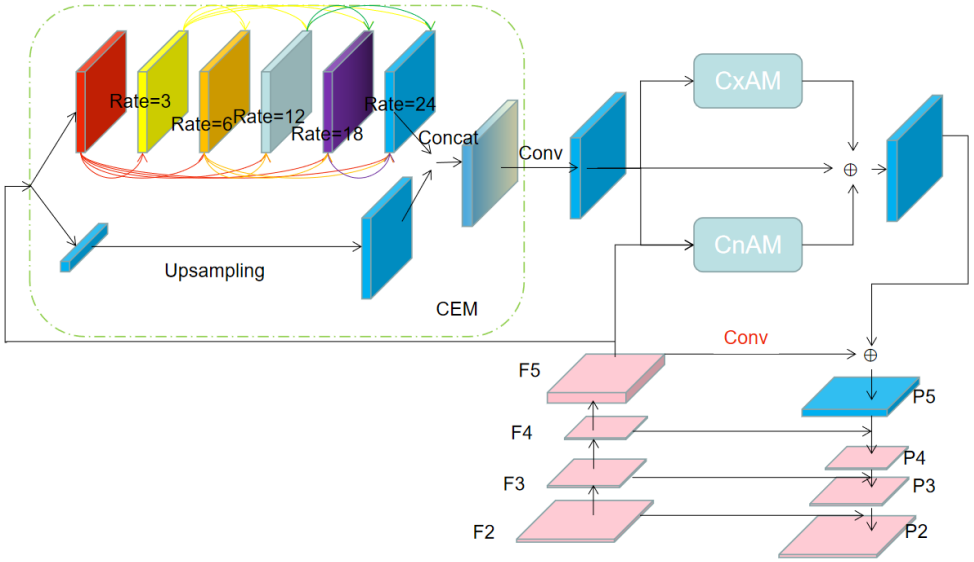


Figure 8. Architecture of the AC-FPN model.

As shown in Figure 8, the CEM can obtain a considerable amount of context information from multiple experience fields and capture large receptive field information with different expansion rates of the multipath convolution layer by increasing the amount of calculation. The CEM takes the F5 of FPN as the input and retains the high-resolution information of input features. However, the redundant context of the CEM module may have adverse effects on localization and recognition.

Because the context extraction module has excess receptive field information, the AM adopts the attention mechanism to adapt to the saliency dependence of the captured object. The AM consists of the CxAM, as shown in Figure 9, and CnAM, as shown in Figure 10, which play the roles of capturing discriminative semantics and locating precise positions, respectively [43].

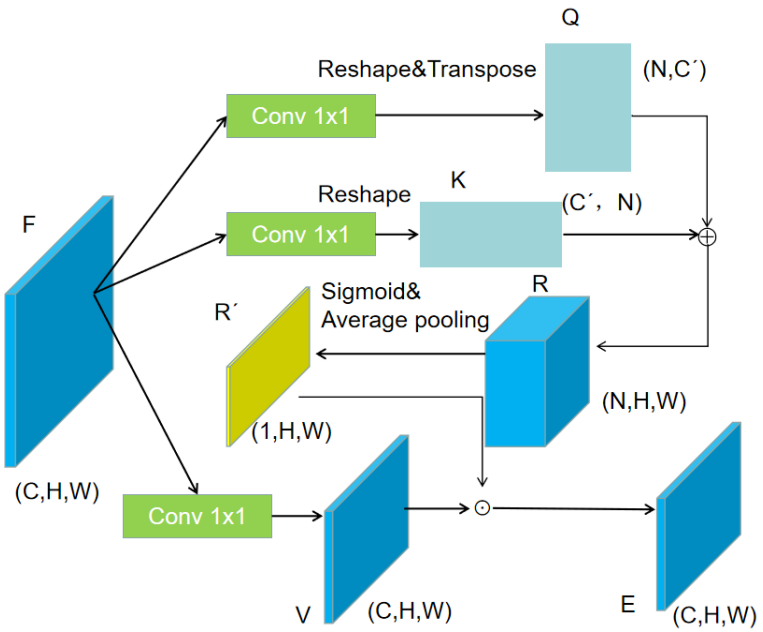


Figure 9. Architecture of CxAM.

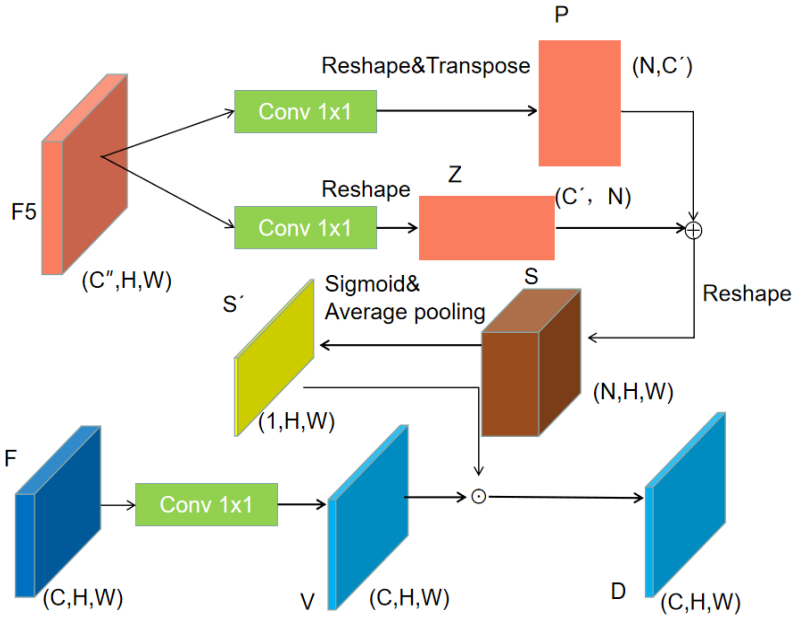


Figure 10. Architecture of CnAM.

4. Experimental Results and Analysis

The proposed approach for underwater biological multi-target recognition was applied and included developing an underwater dataset, configuring parameters, and establishing evaluation criteria. Underwater biological multi-target recognition results were then obtained. The performance of the proposed method was demonstrated by experimental comparison with other methods of underwater biological multi-target recognition to demonstrate the superiority of using the common criteria.

4.1. Development of the Underwater Dataset

The dataset used in this study was obtained from the 2018 Underwater Robot Picking Contest (URPC2018) provided by the organizing committee of the National Underwater Robot Competition (http://www.urpc.org.cn/index.html# (accessed on 15 January 2022)). The dataset was first screened to remove blurred and bio-dense underwater images, and a new dataset was then constructed using image augmentation and annotation. The construction process is presented in Figure 11.

4.2. Selection of Underwater Dataset

The dataset used in this study for real underwater environments (which included underwater reef and sediment environments) was derived from the Dalian Zhangzidao Marine Ranch. The marine ranch is one of the first national marine ranch demonstration areas in China and is rich in sea cucumbers, urchins, starfish, and other marine animals. The shooting data are real and representative. Some partial representative images are shown in Figure 12. The sizes of the images were different (Figure 12a), and the environments complex (Figure 12a,b). As the red parts of Figure 12a,b shows, the color of sea cucumbers approximates that of sediment or rocks at the sea bottom, which is more difficult to identify in multi-aquatic environments.

To ensure the quality of the model training images, blurred and dense images were cleaned. The results are shown in Figure 13. The initial dataset had 400 underwater images including 644 sea urchins, 375 sea stars, 543 sea cucumbers, and 1562 marine organisms. The dataset had an imbalanced number of sea stars and sea urchins represented in the images.

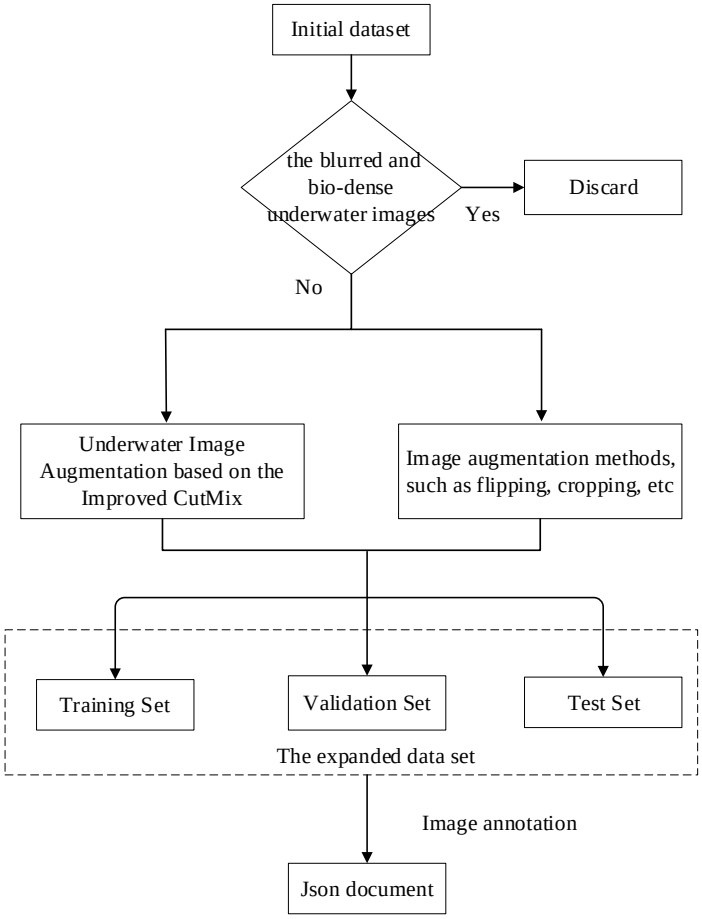
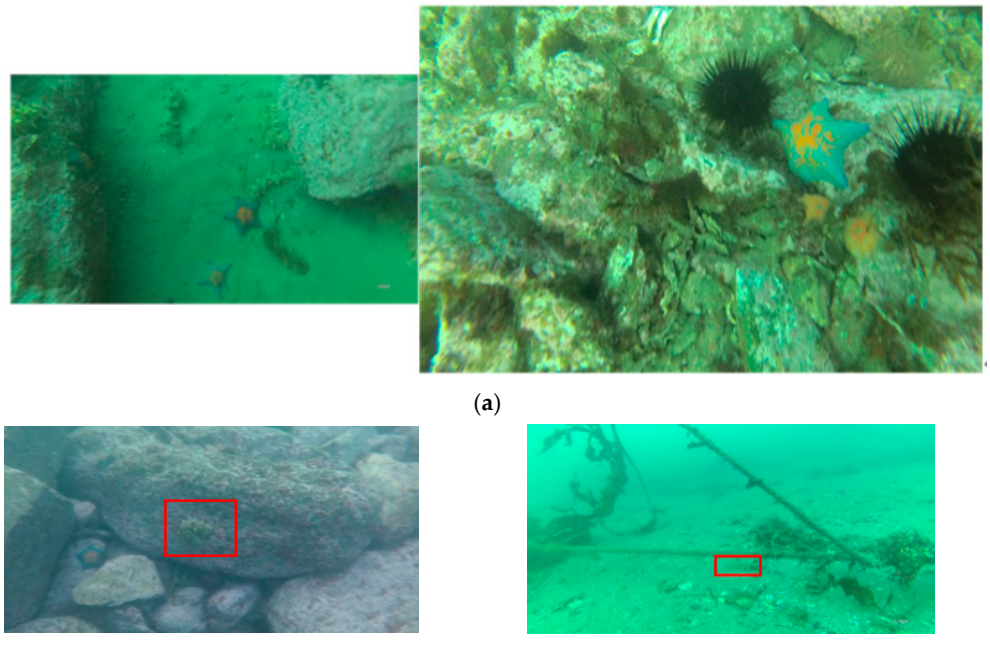


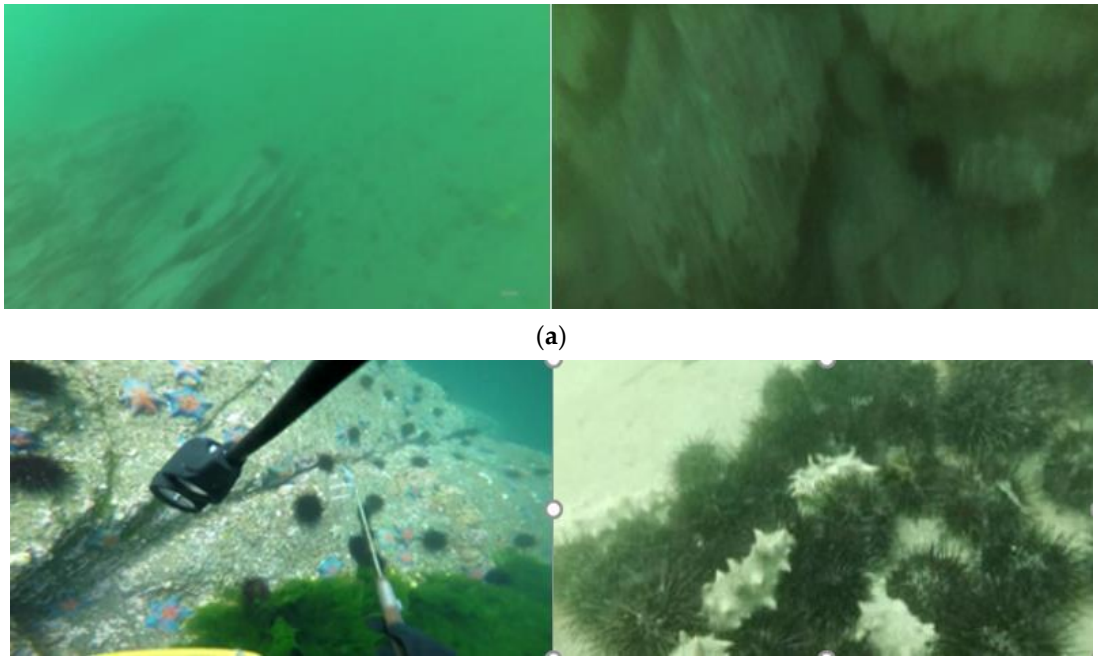
Figure 11. Process of dataset construction.



(**a**)

(**b**) (**c**)

Figure 12. Representative images in the dataset: (**a**) images of different sizes; (**b**) sea cucumbers among rocks; (**c**) sea cucumbers in a multi-aquatic environment.



(**a**)

(**b**)

Figure 13. Examples of images that had to be cleaned: (**a**) blurred images; (**b**) dense images.

4.3. Parameter Configuration and Evaluation Criteria

This study used the Ubuntu operating system as the working platform and the TensorFlow GPU 1.13.1 as the deep learning framework for dataset training. In addition, the Mask-RCNN model was designed for underwater biological multi-target recognition, and the parameter settings were as follows. ResNet101 was selected as the backbone network, and the initial learning rate was 0.001. Because only one GPU was required for accelerated training in the experiments, images_per_gpu = 1, and each epoch included 100 steps (Steps_per_epoch = 100), with a total number of epochs = 120.

To estimate the performance of the convolutional neural network model, the evaluation criteria included the precision, recall, average precision (AP), and mean AP (mAP) [31].

4.4. Underwater Image Augmentation Results

The initial dataset contained 400 images. The image augmentation method based on CutMix expanded the dataset to 1000 images, where the numbers of sea urchins, starfish, and sea cucumbers shown in these 1000 images were 1367, 1298, and 1326, respectively, totaling 3991 sea creatures. The augmented dataset was then divided based on a 6:2:2 ratio; that is, the training, validation, and test sets contained 600, 200, and 200 images, respectively. The augmented images based on the improved CutMix method are shown in Figure 14. The number of organisms to be identified between the initial and augmented datasets is listed in Table 1.

Table 1. Number of organisms to be identified.

Category	Sea Urchins	Sea Cucumbers	Starfish
Initial number of creatures	644	543	375
Number of creatures after augmentation	1367	1326	1298

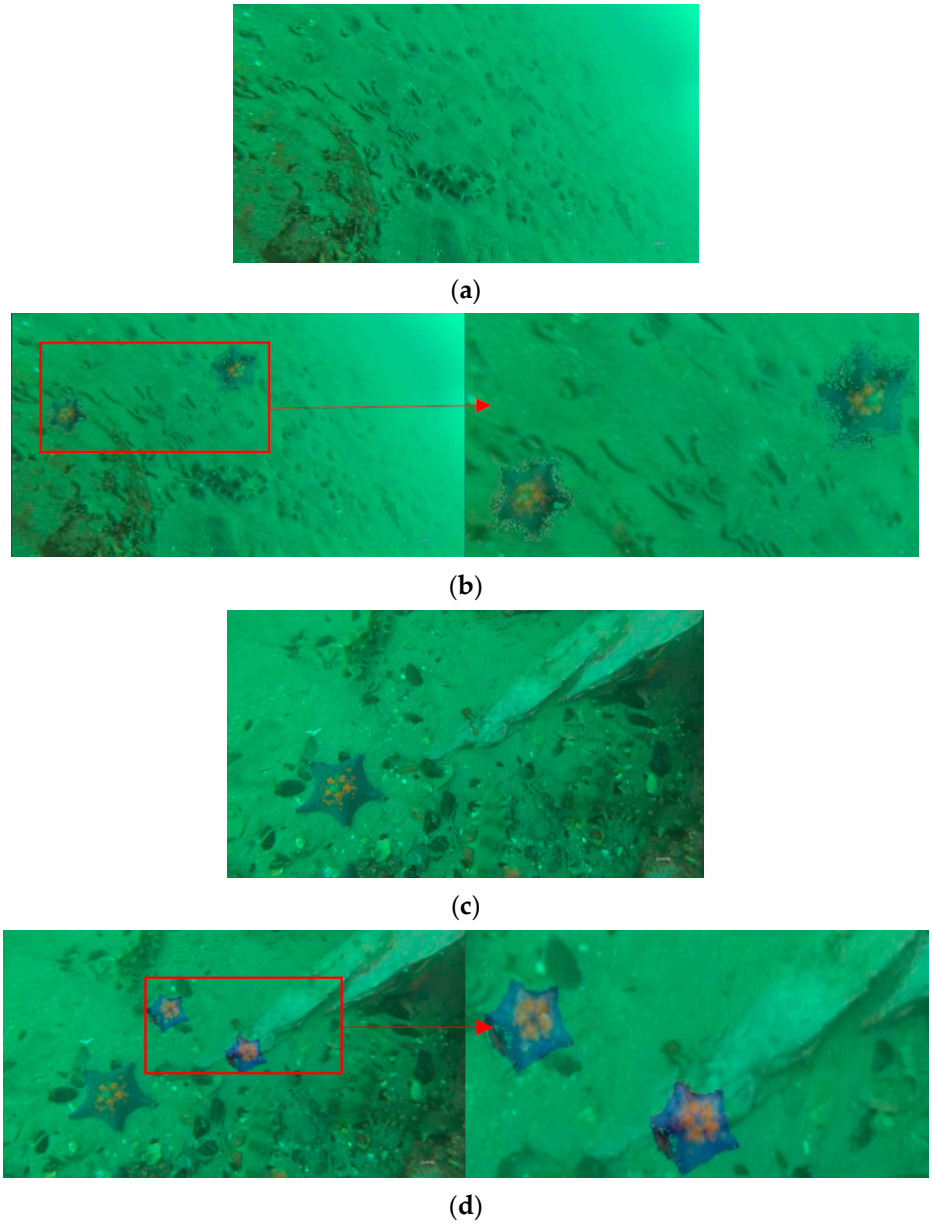


Figure 14. Image augmentation results based on the improved CutMix method: (**a**) original image A; (**b**) augmentation result for Image A; (**c**) original Image B; (**d**) augmentation result for Image B.

To verify the generalization ability of the improved CutMix method, this study used the original and augmented datasets for training in conjunction with the Mask-RCNN model, and the comparison results are listed in Table 2.

Table 2. Comparison of the image augmentation results.

Image Augmentation	Sea Urchins (AP)	Sea Cucumbers (AP)	Starfish (AP)	mAP
No	0.759	0.733	0.691	0.728
Yes	0.804	0.794	0.791	0.796

As shown in Table 2, the identification accuracy for starfish with a small number was relatively low due to the imbalanced number of creatures in the original dataset, which was 0.037 less than the mAP. Under the image augmentation approach, the identification accuracy of sea cucumbers, starfish, and sea urchins increased by 0.061, 0.10, and 0.045, respectively. In addition, the mAP was 0.796, which was 0.068 greater than that of the initial

dataset. The number of starfish was approximately balanced with the other categories of creatures after image augmentation, and the recognition accuracy was the most significantly improved compared with other methods. Compared with the other two types of organisms, the AP value for sea cucumbers was the lowest because sea cucumbers have a similar color to sediment or rock, making them more difficult to distinguish.

In summary, the image augmentation method improved the recognition accuracy, and the augmented dataset could be used for training in subsequent comparison experiments.

4.5. Underwater Image Enhancement Results

To verify the effectiveness of the underwater image enhancement method based on image fusion, the unenhanced and enhanced datasets were trained using the Mask-RCNN model based on the augmented image dataset. The comparison results are listed in Table 3.

Table 3. Comparison of the image enhancement results.

Image Enhancement	Recall	Precision	mAP
No	0.809	0.791	0.796
Yes	0.817	0.823	0.812

The recall, accuracy, and mAP were improved after the image enhancement method was applied, revealing that the appropriate enhancement algorithm could improve the identification accuracy of the model. In addition, the mAP improvement was relatively minimal because more feature information is learned after the dataset is augmented, and further improving the accuracy is difficult.

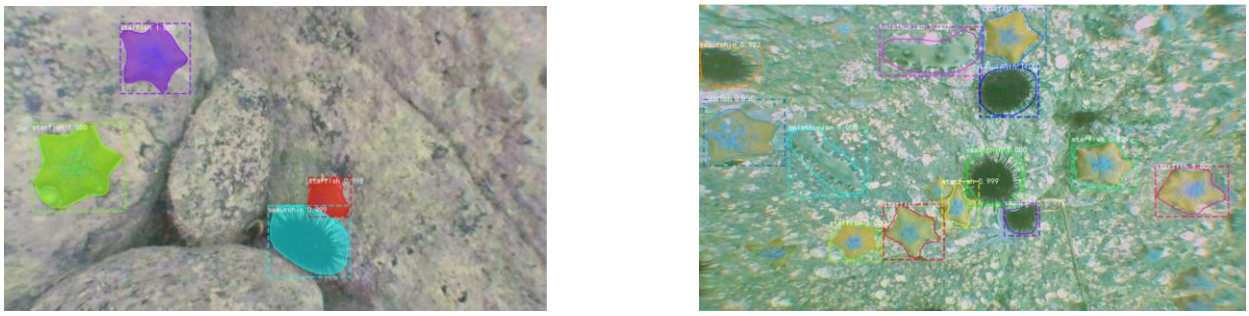
4.6. Underwater Biological Multi-Target Recognition Results

To address the problems of occlusion and overlap, this study adopted the soft-NMS algorithm to improve the model's ability to detect and recognize occluded objects. An improved FPN (AC-FPN) was then designed to ensure that the model was suitable for small targets. To sufficiently demonstrate the superiority of the proposed improved Mask-RCNN model for underwater biological multi-target recognition, current state-of-the-art deep learning models including YOLOv5 and the original Mask-RCNN were compared. YOLOv5 (Ultralytics) is an end-to-end single-stage algorithm, but no studies have reviewed YOLOv5. The comparison results are listed in Table 4.

Table 4. Comparison of the recognition results.

Model	mAP	FPS
YOLOv5	0.661	26.72
Mask-RCNN	0.796	5.35
Proposed	0.828	4.97

It can be seen that the mAP of YOLOv5 was lower, but the detection speed was the fastest. Compared with the original Mask-RCNN model, the mAP of the proposed model increased by 3.2% to 82.8%, while the speed was reduced by only 0.38, proving the effectiveness of the improved method. The recognition and instance segmentation results finished by the proposed method are shown in Figure 15. As can be seen from Figure 15, the detection and instance segmentation results finished by the proposed improved Mask-RCNN model were satisfactory, and different underwater creatures were labeled by different color boxes.



(**a**) (**b**)

Figure 15. Recognition and instance segmentation results finished by the proposed method: (**a**) targets detection among rocks; (**b**) detection of multi-scale targets.

5. Conclusions

The scarcity and poor quality of underwater images present significant challenges to underwater target recognition. To solve these problems, this paper proposed an image augmentation, image enhancement method, and improved Mask-RCNN combined for underwater biological multi-target detection. Image augmentation and enhancement algorithms were applied to the original sample set to expand it and improve the image quality. An improved Mask-RCNN model was further proposed to realize underwater biological multi-target recognition. The novelty of our proposed method is in its use of image augmentation based on the improved CutMix, image enhancement based on image fusion, and multi-target recognition based on the improved Mask-RCNN. Instance segmentations of three underwater creatures (sea cucumbers, urchins, and starfish) were used to evaluate the proposed methods, and the results showed that the proposed methods could perform underwater biological multi-target recognition effectively, obtaining highly accurate and reliable recognition results.

This article provides a viable solution to underwater biological multi-target recognition. We acknowledge that the practical application of the proposed method remains difficult because of the poor quality of underwater images, the limited datasets, and scarcity of computing resources. Furthermore, like most underwater biological multi-target recognition systems, this method is not suitable for long-distance underwater biological multi-target recognition. In future research, we will improve the quality of underwater images and expand our database. In addition, future studies will include investigating a greater number of deep learning models and their parameter-optimization algorithms.

Author Contributions: Conceptualization, Z.Y.; Data curation, B.Y. and H.L.; Formal analysis, B.Y. and H.L.; Methodology, Z.Y. and Z.C.; Software, Z.Y. and Z.C.; Validation, B.Y. and H.L.; Writing—original draft, Z.Y.; Writing—review and editing, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by 'The School Research Fund of Nanjing Vocational University of Industry Technology' (Grant No. YK21-05-05), 'the Open Research Fund of Key Laboratory of River Basin Digital Twinning of Ministry of Water Resources, (Grant No. Z0202042022)', 'the Open Foundation of Industrial Perception and Intelligent Manufacturing Equipment Engineering Re-search Center of Jiangsu Province' (Grant No. ZK22-05-13), and 'The Vocational Undergraduate Education Research Fund of Nanjing Vocational University of Industry Technology' (Grant No. ZBYB22-07).

Data Availability Statement: All data included in this study are available upon request by contacting the corresponding author.

Acknowledgments: We thank the anonymous reviewers for their comments and suggestions that greatly improved the manuscript.

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

References

1. Song, S.; Zhu, J.; Li, X.; Huang, Q. Integrate MSRCR and mask R-CNN to recognize underwater creatures on small sample datasets. *IEEE Access* **2020**, *8*, 172848–172858.
2. Zhou, J.; Yang, Q.; Meng, H.; Gao, D. An underwater target recognition method based on improved YOLOv4 in complex marine environment. *Syst. Sci. Control Eng.* **2022**, *10*, 590–602. [CrossRef]
3. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136.
4. Bao, Z.; Guo, Y.; Wang, J.; Zhu, L.; Huang, J.; Yan, S. Underwater Target Detection Based on Parallel High-Resolution Networks. *Sensors* **2023**, *23*, 7337.
5. Huibin, W.; Qian, Z.; Xin, W.; Zhe, C. Object detection based on regional saliency and underwater optical prior knowledge. *Chin. J. Sci. Instrum.* **2014**, *35*, 387–397.
6. Shi, X.U.X.; Zhang, J.L. Feature extraction of underwater targets using generalized S-transform. *J. Comput. Appl.* **2012**, *32*, 280–282.
7. Jiang, R.; Han, S.; Yu, Y.; Ding, W. An access control model for medical big data based on clustering and risk. *Inf. Sci.* **2023**, *621*, 691–707.
8. Zhou, T.; Wu, W.; Peng, L.; Zhang, M.; Li, Z.; Xiong, Y.; Bai, Y. Evaluation of urban bus service reliability on variable time horizons using a hybrid deep learning method. *Reliab. Eng. Syst. Saf.* **2022**, *217*, 108090.
9. Zhang, J.; Cui, Y.; Ren, J. Dynamic Mission Planning Algorithm for UAV Formation in Battlefield Environment. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *59*, 3750–3765. [CrossRef]
10. Zhang, W.; Dong, L.; Pan, X.; Zou, P.; Qin, L.; Xu, W. A survey of restoration and enhancement for underwater images. *IEEE Access* **2019**, *7*, 182259–182279.
11. Schettini, R.; Corchs, S. Underwater image processing: State of the art of restoration and image enhancement methods. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 746052.
12. Chang, H.H.; Cheng, C.Y.; Sung, C.C. Single underwater image restoration based on depth estimation and transmission compensation. *IEEE J. Ocean. Eng.* **2018**, *44*, 1130–1149. [CrossRef]
13. Huang, H.; Zhou, H.; Yang, X.; Zhang, L.; Qi, L.; Zang, A.Y. Faster R-CNN for marine organisms detection and recognition using data a ugmentation. *Neurocomputing* **2019**, *337*, 372–384. [CrossRef]
14. Noh, J.M.; Jang, G.R.; Ha, K.N.; Park, J.H. Data augmentation method for object detection in un-derwater environments. In Proceedings of the 2019 19th International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 15–18 October 2019; pp. 324–328.
15. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, New Orleans, LA, USA, 18–24 June 2022; pp. 6023–6032.
16. Ghani, A.S.A.; Isa, N.A.M. Underwater image quality enhancement through integrated color model with Rayleigh distribution. *Appl. Soft Comput.* **2015**, *27*, 219–230. [CrossRef]
17. Vasamsetti, S.; Mittal, N.; Neelapu, B.C.; Sardana, H.K. Wavelet based perspective on variational enhancement technique for underwater imagery. *Ocean Eng.* **2017**, *141*, 88–100.
18. Iqbal, M.; Riaz, M.M.; Ali, S.S.; Ghafoor, A.; Ahmad, A. Underwater image enhancement using laplace decomposition. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1500105. [CrossRef]
19. Jobson, D.J.; Rahman, Z.U.; Woodell, G.A. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **1997**, *6*, 965–976. [CrossRef]
20. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robot. Autom. Lett.* **2017**, *3*, 387–394. [CrossRef]
21. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.* **2019**, *29*, 4376–4389.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90.
23. Mittal, S.; Srivastava, S.; Jayanth, J.P. A survey of deep learning techniques for underwater image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 6968–6982. [CrossRef] [PubMed]
24. Chen, L.; Zheng, M.; Duan, S.; Luo, W.; Yao, L. Underwater target recognition based on improved YOLOv4 neural network. *Electronics* **2021**, *10*, 1634.
25. Yeh, C.H.; Lin, C.H.; Kang, L.W.; Huang, C.H.; Lin, M.H.; Chang, C.Y.; Wang, C.C. Lightweight deep neural network for joint learning of underwater object detection and color conversion. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6129–6143.
26. Huang, W.; Wang, Y.; Zhu, L. A Time Impulse Neural Network Framework for Solving the Minimum Path Pair Problems of the Time-Varying Network. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 7681–7692.
27. Jiang, R.; Kang, Y.; Liu, Y.; Liang, Z.; Duan, Y.; Sun, Y.; Liu, J. A trust transitivity model of small and medium-sized manufacturing enterprises under blockchain-based supply chain finance. *Int. J. Prod. Econ.* **2022**, *247*, 108469.
28. Shi, P.; Xu, X.; Ni, J.; Xin, Y.; Huang, W.; Han, S. Underwater Biological Detection Algorithm Based on Improved Faster-RCNN. *Water* **2021**, *13*, 2420.

29. Li, A.; Yu, L.; Tian, S. Underwater Biological Detection Based on YOLOv4 Combined with Channel Attention. *J. Mar. Sci. Eng.* **2022**, *10*, 469. [CrossRef]
30. Liu, Z.; Zhuang, Y.; Jia, P.; Wu, C.; Xu, H.; Liu, Z. A Novel Underwater Image Enhancement Algorithm and an Improved Underwater Biological Detection Pipeline. *J. Mar. Sci. Eng.* **2022**, *10*, 1204.
31. Yu, K.; Cheng, Y.; Tian, Z.; Zhang, K. High Speed and Precision Underwater Biological Detection Based on the Improved YOLOV4-Tiny Algorithm. *J. Mar. Sci. Eng.* **2022**, *10*, 1821. [CrossRef]
32. Li, J.; Liu, C.; Lu, X.; Wu, B. CME-YOLOv5: An Efficient Object Detection Network for Densely Spaced Fish and Small Targets. *Water* **2022**, *14*, 2412.
33. Buchsbaum, G. A spatial processor model for object colour perception. *J. Frankl. Inst.* **1980**, *310*, 1–26.
34. Wang, J.; Lu, K.; Xue, J.; He, N.; Shao, L. Single image dehazing based on the physical model and MSRCR algorithm. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2190–2199. [CrossRef]
35. Rahman, Z.U.; Jobson, D.J.; Woodell, G.A. Retinex processing for automatic image enhancement. *J. Electron. Imaging* **2004**, *13*, 100–110.
36. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353.
37. Li, Y.; Chen, N.; Zhang, J. Fast and high sensitivity focusing evaluation function. *Appl. Res. Comput.* **2010**, *27*, 1534–1536.
38. Yi, F. Research on an Auto-focusing Algorithm for Microscope. *Chin. J. Sci. Instrum.* **2005**, *26*, 1275.
39. Rothe, R.; Guillaumin, M.; Van Gool, L. Non-maximum suppression for object detection by passing messages between windows. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, 1–5 November 2014*; Revised Selected Papers, Part I 12; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 290–306.
40. Ni, J.; Shen, K.; Chen, Y.; Yang, S.X. An Improved SSD-Like Deep Network-Based Object Detection Method for Indoor Scenes. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5006915.
41. Wang, W.; Li, X.; Lyu, X.; Zeng, T.; Chen, J.; Chen, S. Multi-Attribute NMS: An Enhanced Non-Maximum Suppression Algorithm for Pedestrian Detection in Crowded Scenes. *Appl. Sci.* **2023**, *13*, 8073.
42. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
43. Cao, J.; Chen, Q.; Guo, J.; Shi, R. Attention-guided context feature pyramid network for object detection. *arXiv Preprint* **2020**, arXiv:2005.11475.