Journal of *Marine Science and Engineering*

MDPI

Article

A Marine Organism Detection Framework Based on Dataset Augmentation and CNN-ViT Fusion

Xiao Jiang [1], Yaxin Zhang [1], Mian Pan [1], Shuaishuai Lv [1], Gang Yang [2], Zhu Li [1], Jingbiao Liu [3] and Haibin Yu [1,4,5,*]

[1] College of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China
[2] First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China
[3] Ocean Technology and Equipment Research Center, Hangzhou Dianzi University, Hangzhou 310018, China
[4] Zhejiang Provincial Key Lab of Equipment Electronics, Hangzhou 310018, China
[5] Ningbo Institute of Oceanography, Ningbo 315832, China
* Correspondence: shoreyhb@hdu.edu.cn; Tel.: +86-18969137528

Abstract: Underwater vision-based detection plays an important role in marine resources exploration, marine ecological protection and other fields. Due to the restricted carrier movement and the clustering effect of some marine organisms, the size of some marine organisms in the underwater image is very small, and the samples in the dataset are very unbalanced, which aggravate the difficulty of vision detection of marine organisms. To solve these problems, this study proposes a marine organism detection framework with a dataset augmentation strategy and Convolutional Neural Networks (CNN)-Vision Transformer (ViT) fusion model. The proposed framework adopts two data augmentation methods, namely, random expansion of small objects and non-overlapping filling of scarce samples, to significantly improve the data quality of the dataset. At the same time, the framework takes YOLOv5 as the baseline model, introduces ViT, deformable convolution and trident block in the feature extraction network, and extracts richer features of marine organisms through multi-scale receptive fields with the help of the fusion of CNN and ViT. The experimental results show that, compared with various one-stage detection models, the mean average precision (mAP) of the proposed framework can be improved by 27%. At the same time, it gives consideration to both performance and real-time, so as to achieve high-precision real-time detection of the marine organisms on the underwater mobile platform.

Keywords: vision-based underwater object detection; marine organism detection; YOLOv5; dataset augmentation; CNN-ViT fusion

check for updates

Citation: Jiang, X.; Zhang, Y.; Pan, M.; Lv, S.; Yang, G.; Li, Z.; Liu, J.; Yu, H. A Marine Organism Detection Framework Based on Dataset Augmentation and CNN-ViT Fusion. *J. Mar. Sci. Eng.* **2023**, *11*, 705. https://doi.org/10.3390/jmse11040705

Academic Editor: Hiroshi Kuroda

Received: 5 February 2023
Revised: 13 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, with the increasingly large scale of development and utilization of marine resources, underwater optical visual target detection with extensive application scenarios has played an increasingly important role in the field of underwater security [1], marine exploration [2], fish farming [3], and marine ecology [4]. Therefore, achieving underwater autonomous operation through visual target detection by using underwater optical images has become a research hotspot in the field of computer vision.

With the rapid development of deep learning framework, the visual target detection algorithm has shifted from a traditional algorithm based on manual feature extraction to detection technology based on a deep neural network. From the RCNN detection algorithm proposed in 2013, Fast RCNN [5], and Faster RCNN [6], two-stage target detection algorithms obtained by improving detection precision to one-stage target algorithms, including SSD [7] and YOLO [8–12] have considered both detection precision and detection speed. At present, many scholars have applied target detection methods based on deep neural networks to underwater target detection.

Water has absorption and scattering effects on light. Compared with ordinary images, underwater images have low color contrast, blue–green tone, fog, and other phenomena [13,14]. Thus, the imaging quality is poor. In addition, different from most target detection tasks on land, which can be implemented on general computer platforms, underwater target detection tasks are generally conducted on ROV, AUV [15], and other platforms with a limited volume or power supply [16]. Thus, they can only run in embedded systems. Moreover, underwater target detection usually faces two major difficulties: serious image degradation and limited computing resources.

When the visual target detection method based on a deep learning framework is applied to marine organism detection, they will overlap and block each other because of the clustering effect. In addition, the underwater movement of the carrier carrying the camera is usually limited to a certain extent. Consequently, the marine organisms in underwater images are usually not uniform in scale and size. Some marine organisms are small in size and distributed unevenly, which aggravates the difficulty of organism detection. In addressing the problem on small target detection, Kisantal et al. [17] proposed replication and enhancement for small targets, which are used to increase the number of training samples of small targets by copying and pasting small targets in images many times to improve the detection properties of small targets. Moreover, Chen et al. [18] balanced the quantity and quality of medium and small targets by scaling and splicing [19]. The existing coping strategies aim to increase the weight of small target samples in the loss function through increasing the number of small targets. However, it cannot fundamentally solve the problem that small targets are difficult to detect because of the small proportion of pixels in the image and the small amount of information.

In addition, visual target detection based on the deep learning framework is a typical data-driven method, which requires a large number of high-quality image data to train the network. However, obtaining underwater images is difficult. Furthermore, images containing a specific underwater organism are limited. Therefore, the sample imbalance is significant. In the field of land visual target detection, the proportion of different quantities of targets in the dataset are usually balanced by under sampling/oversampling the samples or in accordance with the loss function of the samples. For example, online hard example mining (OHEM) algorithm proposed by Abhinav Shrivastava et al. can screen out a hard example based on the loss of input samples for specific training [20]. Tsung-Yi Lin et al. proposed the loss function Focal Loss [21] to allow the model to focus on difficult and error-detected samples. The existing sample balance methods have alleviated the negative impact of unbalanced sample number to a certain extent. However, when these methods are applied to underwater target recognition, the effects are generally poor because the sample imbalance of underwater images is more serious than that on land.

Aimed at the problems on existing methods in underwater organism detection, this study proposes a one-stage marine organism detection framework based on dataset augmentation and the fusion of Convolutional Neural Networks (CNN) and Vision Transformer (ViT). By introducing novel data augmentation methods, this framework can effectively cope with the problems such as the small size of underwater organisms and the scarcity of some species of organism samples. In addition, the one-stage improved YOLO-v5 model based on CNN-ViT fusion can complete real-time and high-precision underwater organism detection under limited computing resources. The main contributions of this paper are as follows:

(1) For the problems such as low recall and precision of target detection caused by different scales and sizes of marine organisms in underwater images and uneven location distribution, this study proposes a random expansion method for small targets used in the training dataset. Considering that no additional dataset is introduced and the sample image is not increased, the sizes of the targets in the sample image are randomly expanded, and the quantity of equal-sized images in the training set is greatly expanded, thereby decreasing the difficulty of model detection of small targets and effectively improving the recall and precision of model detection of small targets.

(2) Aimed at the problem that the specific target samples are limited because of the difficulty in obtaining underwater images, this study proposes a non-overlapping filling method for scarce samples combined with the random expansion of small targets. An efficient one-dimensional search method for the filling space is adopted to directly fill the scarce targets into the existing sample images, while ensuring that they do not overlap with the existing targets. Without the help of additional datasets, the scarce targets are effectively supplemented in the training stage, thereby alleviating the imbalance of the number of samples in the dataset.

(3) In view of the limited energy of autonomous underwater vehicles and the serious limitation of computing resources, this study proposes a light detection model taking YOLOv5 as the baseline. It introduces the ViT module, deformable convolution and trident block, and utilizes Focal Loss as the objective function of classification loss and target confidence loss to achieve high-precision and real-time detection of marine organisms on the resource-constrained underwater mobile platform.

2. Related Works

2.1. Target Detection Algorithm

Since Krizhevsky Alex et al. proposed AlexNet based on convolution neural network architecture and won the ILSVRC 2012 challenge championship, CNN has become the mainstream technology in computer image processing.

In 2014, R. Girshick proposed the use of features extracted by Region-Convolutional Neural Networks (RCNN) in target detection. Then, target detection began to develop at an unprecedented speed. The RCNN series algorithm, which extracts the target recommendation box obtained by selective search, is a classic two-stage deep learning target detection algorithm.

In 2017, Ashish Vaswani innovatively proposed the Transformer architecture based on encoder–decoder and abandoned RNN or CNN architecture. Transformer can effectively solve the problems that RNN cannot be processed in parallel and CNN cannot efficiently obtain long-distance dependence [22].

In utilizing the inherent information in the feature to make the interaction of attention and reduce the dependence on external information, the self-attention mechanism can be introduced. In screening out a small amount of important information from a large amount of information and focusing on this important information, ignoring most unimportant information, an attention mechanism combined with a spatial attention mechanism and channel attention mechanism is adopted.

Transformer has a significant impact in the field of Natural Language Processing (NLP). Its applications, such as Bidirectional Encoder Representation from Transformers [23], Generative Pre-trained Transformer [24], and Text-To-Text Transfer Transformer [25], promote this field to obtain the best performance. Transformer's success in NLP has aroused its interest in the field of computer vision because it can integrate global semantic information.

Transformer architecture has been directly applied to non-overlapping medium image blocks for image classification by ViT's groundbreaking work [26]. Compared with the work based on CNN, the impressive accuracy in image classification is achieved. CNN can capture local-range information, whereas ViT can capture long-range information. In extracting image features, ViT is a useful supplement to CNN. Their combination can effectively extract the features of different targets with evident scale differences, such as Deformable-Detr [27].

Although the highest detection accuracy of target detection algorithms is constantly refreshing, these detection algorithms perform poorly in underwater AUV or ROV. These high-precision target detection algorithms are restricted by limited computational power and limited power supply; therefore, a higher requirement for underwater target detection algorithms is proposed in the actual application.

2.2. Data Augmentation

The data augmentation strategy can generate more additional data based on existing data to improve the performance of the machine learning system, deep learning classification model, target detection model, and so on. The correct use of a data augmentation strategy can effectively improve the accuracy and robustness of the machine learning model. Different from the fine-tuning of the neural network structure, the data augmentation strategy only increases the training cost and time, but does not increase the inferencing cost.

The imbalance of the sample number is a major difficulty in the target detection task, which can be simply described as follows: Some target classes to be detected are less, whereas some target classes to be detected are more. We can increase the number of samples with a small number of target classes, i.e., oversampling, and decrease the number of samples with a large number of target classes, i.e., undersampling, to solve the problem by balancing the proportion of different numbers of targets in the dataset. OHEM algorithm screens out hard examples based on the loss of input samples, which are samples that have a great impact on classification or detection [20]. Then, these samples obtained from screening are applied to training in random gradient descent. Tsung-Yi Lin et al. [21] proposed a new loss function Focal Loss [21] to increase the weight of difficult samples and allow the model to focus on difficult and error-detected samples. This loss function alleviates the negative impact of the imbalance of the sample number to a certain extent, but it does not solve the difficulties in essence. Chia-Hung Yeh et al. [28] skillfully solved the imbalance of a small sample number by pasting the target in the dataset into the underwater background image, which improved the performance of the model. However, this strategy requires enough images with underwater scenes, but obtaining the suitable and available images with underwater scenes similar to those in the dataset is difficult.

An excessively small target size is another difficulty in the target detection task. A small target has a low proportion of pixels, which contains less information, causing great difficulty in positioning. In recent years, solutions to small target detection have emerged. In view of data augmentation, Yu et al. [29] proposed a scale matching strategy. Based on the differences of targets in size, the size of the whole image was adjusted. Considering limited GPU resources, the image was cut again to narrow the gap between targets with different sizes and to prevent the information of small and medium targets from becoming lost in the conventional zoom operation. This strategy is effective and feasible, but it is very cumbersome. Aimed at the small area covered by small targets, the lack of diversity of occurrence positions, and the intersection ratio between the detection frame and the truth value box is far less than the expected threshold, Kisantal et al. [17] proposed a replication and enhancement method, which increases the number of training samples of small targets by copying and pasting small targets in the image multiple times, thus improving the detection performance of small targets. This strategy is very ingenious, without losing additional GPU resources, but it will cause an imbalance of targets in number. Based on the small proportion of the quantity of small targets and less information contained in the image, Chen et al. [19] zoomed and spliced images during training and transferred large targets in the dataset into medium targets and medium targets into small targets. In addition, they improved the quantity and quality of medium/small targets considering the consumption from GPU computation. In particular, Kisantal et al.'s strategy that copies small targets and Chen et al.'s strategy that zooms and splices small targets only increase the weight of small target samples in loss function by increasing the quantity of small targets, which does not completely solve the problem of small target detection, but makes small targets smaller and more difficult to detect.

To sum up, underwater target detection is usually hindered by small size, uneven distribution, serious imbalance of samples, and other situations. Most of the existing methods directly follow the sample augmentation and equalization methods on land, without considering the current situation that underwater images are difficult to obtain and sample sets are extremely scarce. Aimed at the problems appearing in the application of existing methods to underwater target detection, this study introduces novel data augmentation methods in the model training to effectively address the problems such

as the small size of underwater organisms and the scarcity of some species of organism samples. Moreover, a lightweight detection model taking YOLOv5 as the baseline is adopted to achieve real-time and high-precision underwater organism target detection under limited computing resources.

3. Method

This study proposes a one-stage marine organism detection framework based on dataset augmentation and CNN-ViT fusion. The overall structure of the framework is shown in Figure 1, and its neural network structure is shown in Figure 2. This framework primarily consists of three parts; namely, data augmentation strategy, backbone feature extraction module, and detection head module.

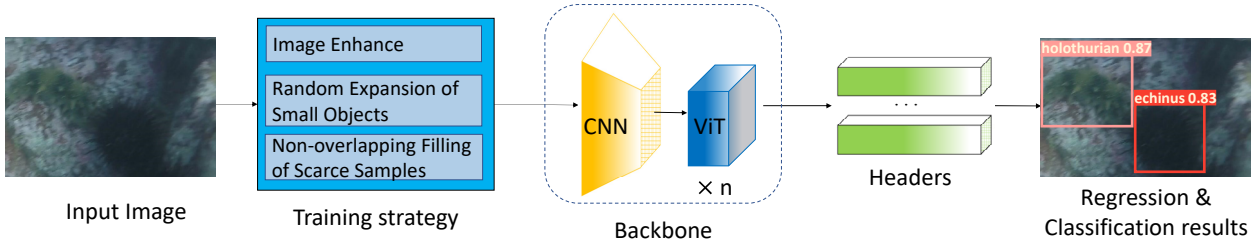


Figure 1. The overall structure of the proposed framework.

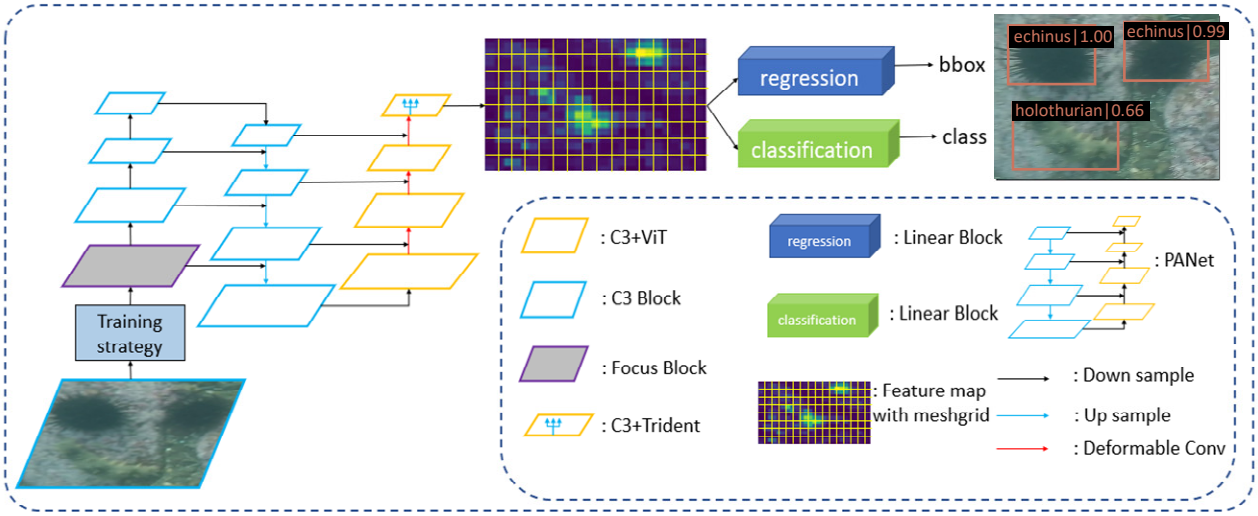


Figure 2. The neural network structure of the proposed framework.

The data augmentation strategy is applied in the training stage of the framework. At this stage, three methods; namely, image enhancement, random expansion of small objects, and non-overlapping filling of scarce samples; are adopted to address the deterioration of underwater imaging, the underscaling of marine organism targets, and the extreme lack of some organism samples. In the feature extraction module, YOLOv5 is taken as a baseline, and ViT is introduced in backbone to achieve the fusion of CNN and ViT, which has both the local-range information capture ability of CNN and the long-range information capture ability of ViT. In addition, deformable convolution [30] and trident block [31] are introduced to extract more abundant features through a multi-scale receptive field to deal with the feature extraction of marine organism targets with different shapes and scales. In the detection head module, based on the multi-scale features output by the path aggregation network (PANet) [32], the category of targets in the detection box is output by classification branch, and the target in the detection box output by regression branch is mapped back to the position in the original image.

3.1. Data Augmentation

3.1.1. Random Expansion of Small Objects

In improving the precision of small target detection, the pixel size of small targets must be increased as much as possible. The common practice is to directly expand the size of the input image. Although this practice can improve the precision of small target detection, it will bring a higher computational load during training. In order to reduce the amount of computation during training, we can use the label of the original dataset to randomly expand the size of small targets in the image, so as to obtain a new image. Consequently, by expanding the size of the small target, the number of images in the training set can be expanded without introducing additional datasets, thereby completing data augmentation. It is helpful to train the detector by reducing the recognition difficulty of hard cases, which is inspired by Dynamic-RCNN [33].

As shown in Figure 3, when the target to be detected is small in size, the model can hardly find the target, and it cannot generate enough and effective prediction boxes (Figure 3a). After the size of the small target is slightly enlarged, the target may be detected in the blue prediction box. However, the target is incorrectly recognized as a starfish, resulting in a classification error (Figure 3b). When the target is expanded to be large enough, a white prediction box is observed in which the target is detected and correctly determined as a scallop (Figure 3c).

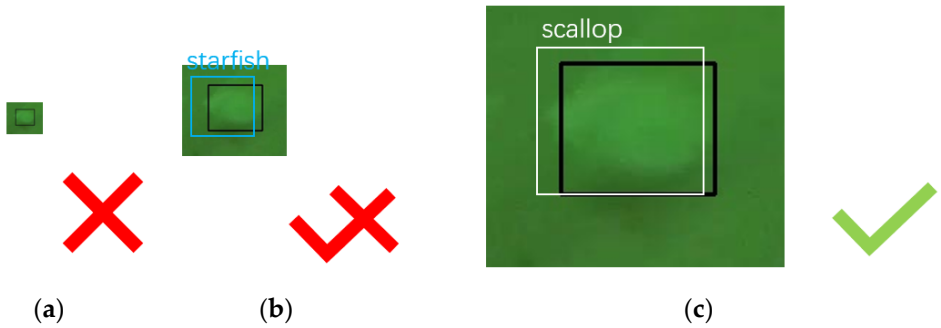


Figure 3. An example of a small target (scallop) being correctly detected after size expansion: (**a**) the target is in small size; (**b**) the small target is slightly enlarged; and (**c**) the target is expanded to be large enough.

Based on the abovementioned ideas, this study proposes a data augmentation method based on random expansion of small targets. The specific practices of this method are as follows: In the training stage, the positions of the center points of the rectangular boxes where all the targets in the image are located as (x, y) and the widths and heights of the rectangular boxes (w, h) are also obtained. The rectangular boxes where small targets are located are selected. The positions (x, y) of the chosen rectangular boxes in the image remain unchanged, and their widths and heights are changed to $(w + 2 \times \Delta w, h + 2 \times \Delta h)$. Bilinear interpolation is adopted to expand the image in the rectangular box and cover the small target at the original position, thereby acquiring a new image. To limit the expansion range of the label image, the super parameter $expand$ is set, representing the maximum number of randomly expanded pixels, that is, $\Delta w, \Delta h \in [0, expand]$. In order to avoid the overlapping of the expanded target label images as much as possible, which will affect the effectiveness of the target, the expansion selected in this study is 5. A randomly expanded image example of small targets is shown in Figure 4. The two echini on the rightmost part of Figure 4 overlap after size expansion. Therefore, the upper limit expansion size $expand$ should be carefully set.

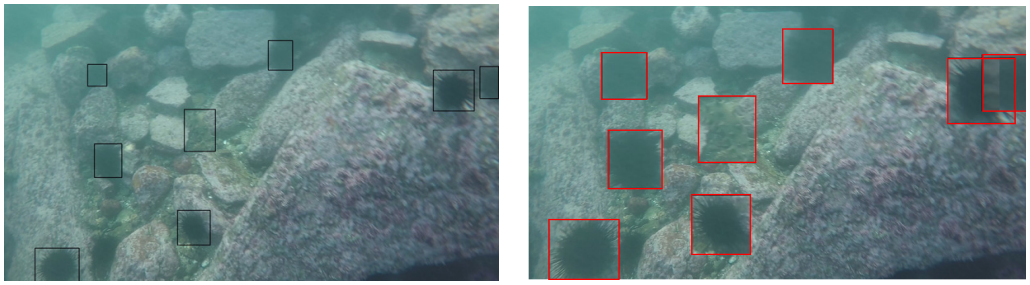


Figure 4. A randomly expanded image example of small targets. The black boxes represent the targets before expansion, and the red boxes represent the targets after expansion.

Notably, random expansion of small objects can achieve the expansion of the training set without introducing additional datasets. In addition, it only changes the training strategy rather than modifying the structure of the target detection model, manually setting the shape and size of the anchor frame and increasing the weight for difficult samples. Therefore, it only increases the time cost of the training model, but it does not affect the speed of model inferencing and will not increase the complexity of the algorithm.

3.1.2. Non-Overlapping Filling of Scarce Samples

Compared with land images, underwater images have limited acquisition channels. Therefore, there are scarce underwater images of specific targets. Furthermore, images containing certain underwater organisms are scarcer. Sample imbalance is very significant.

This study proposes a sample number balancing method that directly fills the scarce targets into the existing sample images. In avoiding the overlap between filling the target and existing target during filling, we need a strategy to search a region to fill the target rapidly. In addition, before the training stage, we need to create a scarce target set. The filling process is shown in Figure 5: firstly, as shown in the red box, we should select an image from the scarce target set. Secondly, the image will go through the label augmentation pipeline, the selector will randomly choose a scarce target and the target will flip and resize. Thirdly, as shown in the blue box, the target will be pasted in the input image.

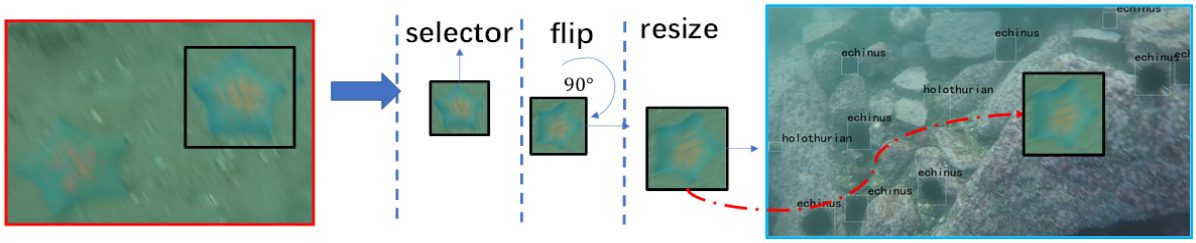


Figure 5. The process of the non-overlapping filling of scarce samples: (**a**) Image with scarce target; (**b**) label augmentation pipeline; and (**c**) input image with a scarce target. The red box represents the image selected from the scarce target set. The black boxes represent the scarce target selected from the area surrounded by the red box.

Considering that the image to be filled also contains other targets, the existing targets in the image cannot be covered during target filling. The label image can be easily filled in the input image that contains only one target. However, in filling the label image without overlapping the original target present in the input image that contains more than one target, there is a regional search problem in the two-dimensional space. When the number of images to be filled is large, the time consumption of the two-dimensional space search algorithm becomes unacceptable. In lowering the complexity of target filling space search, the idea of dimensionality reduction is adopted to transform the problem of

finding the appropriate two-dimensional area into the problem of finding the appropriate one-dimensional area.

The specific methId is shIwn in Figure 6: when we are looking for areas to horizontally fill targets, as shown in Figure 6a, the rectangular area demarcated along the horizontal direction of the ordinate interval (the interval marked by the red arrow) is the unfillable area, and the remaining green rectangular areas are fillable. Filling the targets in these fillable areas must not overlap with the original targets; similarly, when we are looking for areas to vertically fill targets, as shown in Figure 6b, the rectangular area demarcated along the vertical direction of the abscissa area (the area marked by the red arrow) is the unfillable area, and the remaining green rectangular areas are fillable. Consequently, the region that can fill scarce sample targets in the image with uncertain target distribution can be obtained at a very rapid speed, thereby greatly reducing the complexity of filling area search. Its time complexity is only $O(nlogn)$, and the spatial complexity is $O(n)$. n indicates the number of existed targets in the image.

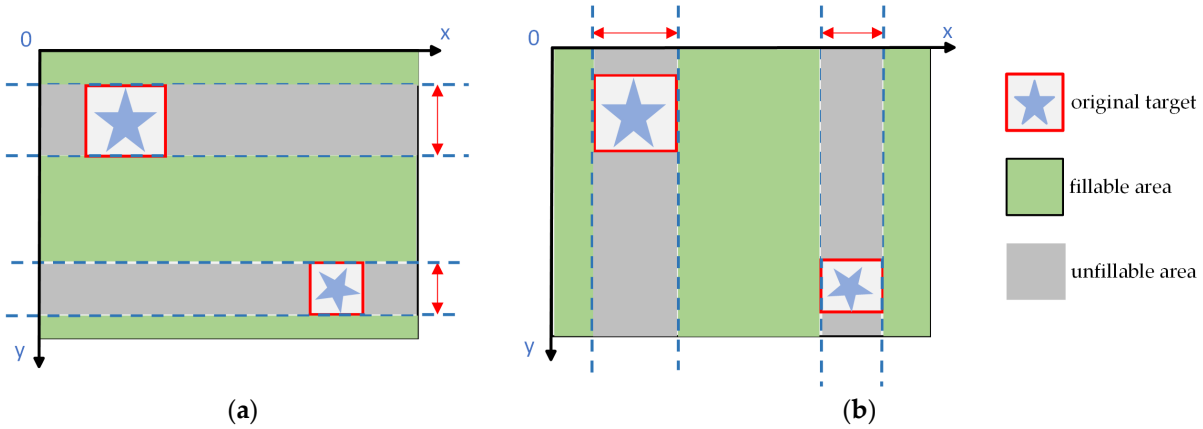


Figure 6. The one-dimensional non-overlapping filling area search: (**a**) Filling targets along the horizontal direction; and (**b**) filling targets along the vertical direction.

If a starfish is taken as a scarce sample, the effect of filling in the vertical direction (y direction) in the image with multiple echini is shown in Figure 7. As shown in the left sub-figure of Figure 7, black boxes identify the targets. The red box is the minimum closure of all black boxes, and it is easy to fill targets out of the red box. Thus, we mainly discuss the area in the red box. With the help of the label augmentation pipeline and the one-dimensional non-overlapping filling area search method described in this section, we filled several starfish of different sizes along the vertical direction into the left subfigure of Figure 7, and the results are shown in the right subfigure of Figure 7.

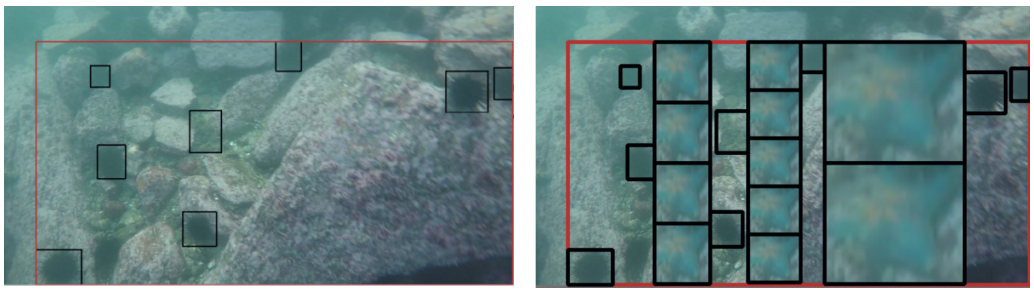


Figure 7. An example of the non-overlapping filling in the vertical direction. The red box represents the minimum closure of all black boxes. The black boxes show the position of the targets.

3.2. Target Detection

The marine organisms to be detected in the proposed framework are mainly echini, holothurians, starfish, and scallops, which are all included in the images of URPC2019 dataset [34]. These marine organisms have different colors, shapes, and textures. The underwater environment has serious color deviation and dark light. The distances and angles of taking pictures are different. In addition, a certain number of starfish and scallops are covered by the deposited dust. All the factors cause difficulties in target detection. In addition, considering that the detection framework proposed in this paper will run on the embedded platforms with limited computing resources such as AUV [15], strict requirements are provided for the size and complexity of the framework. In view of the above difficulties and requirements, we improve YOLOv5 by introducing a ViT module, deformable convolution, and trident block, as shown in Figure 8, and adopting Focal Loss [21] as the loss function of classification loss and object confidence loss.

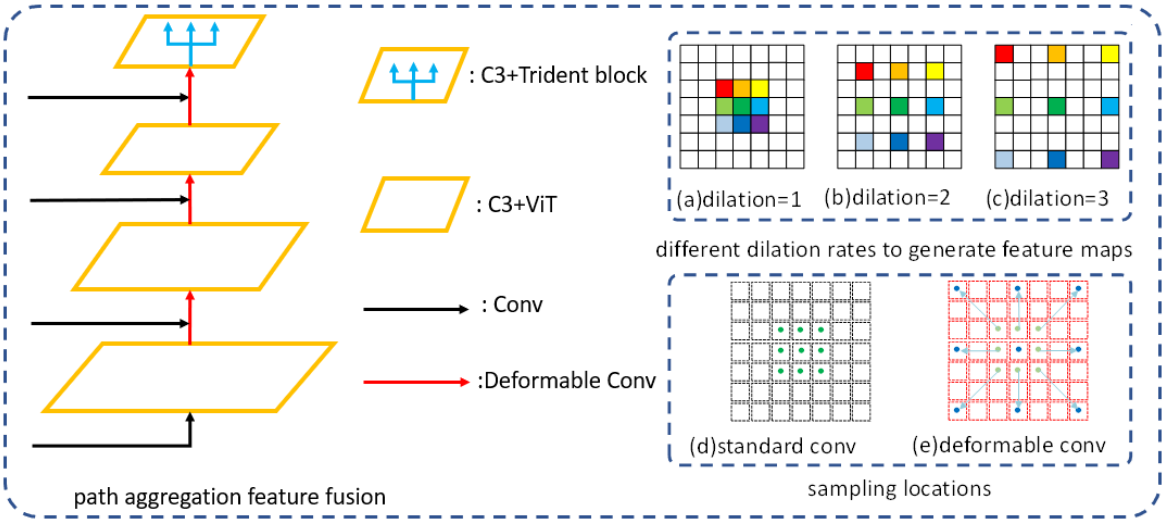


Figure 8. Overall structure of the proposed target detection model: (**a**–**c**) Feature maps generated by trident block with different dilation rates. (**d**,**e**) Sampling locations of standard convolution and deformable convolution.

3.2.1. Backbone of Feature Extraction Network

YOLOv5 successfully absorbed the idea of MobileNet [35] to reduce the amount of computation by separating convolution, and adopted a large number of combinations of conv and C3 modules, as shown in Figure 9. The conv module undertakes the task of deep-level convolution, which is composed of multiple single-layer convolutions; the C3 module undertakes the task of pixel-level convolution, which primarily contains multiple convolutions with a core size of 1, and it is composed of a residual structure.

Given the successful application of Transformer in the NLP field, ViT attempts to apply the standard Transformer structure directly to the image and make the least modifications to the entire image classification process. In particular, in the ViT algorithm, the whole image will be divided into small image blocks, and then these small image blocks will be sent to the network as the input of Transformer in the form of sequence. Afterward, image classification training is performed through supervised learning.

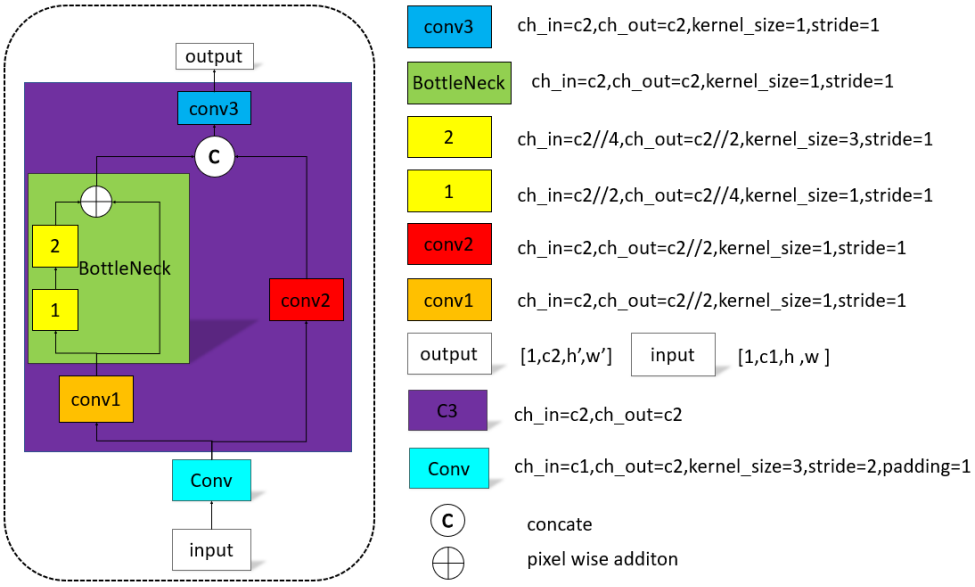


Figure 9. Network structure of YOLOv5.

In ViT, the dependence of the window-based self-attention mechanism on external information is reduced, and the inherent information in the local feature is used as much as possible to make interactions of attention and obtain the correlation in the spatial dimension. However, the attention information stored between channels is also noteworthy. Channel attention aims to show the correlation among different channels modeled and automatically acquire the importance of each feature channel through network learning. Finally, different weight coefficients are assigned to each channel to strengthen important features and inhibit non-important features.

Drawing lessons from the advantages in the accuracy of ViT in solving computer vision tasks, we combine the architecture of ViT with YOLOv5 (Figure 10). The ViT module aims to extract self-attention information of each layer in the feature map; the role of the convolutional module is to allow downsampling of the input feature map, which narrows the size of the feature map and improves the abstraction of the feature.

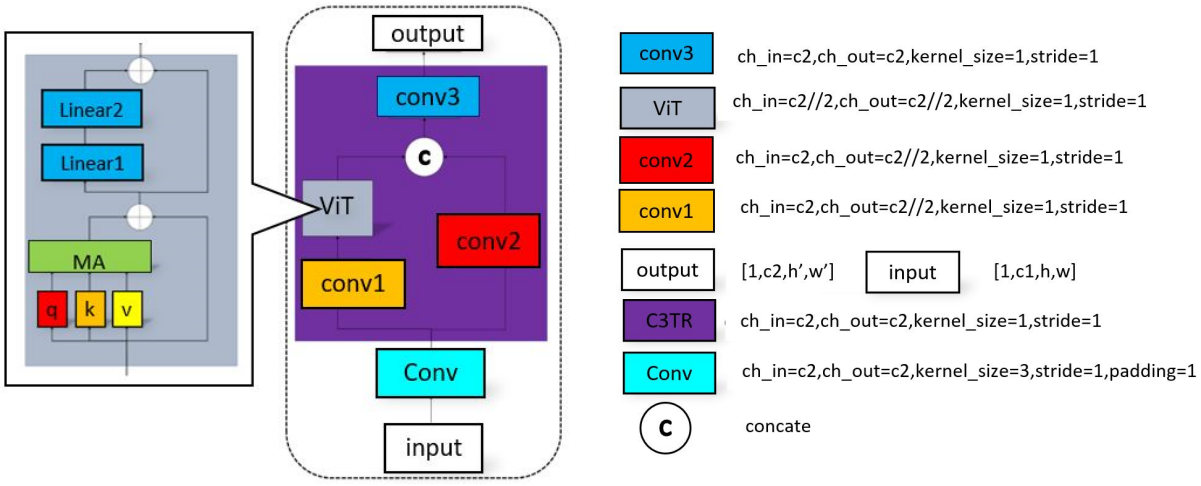
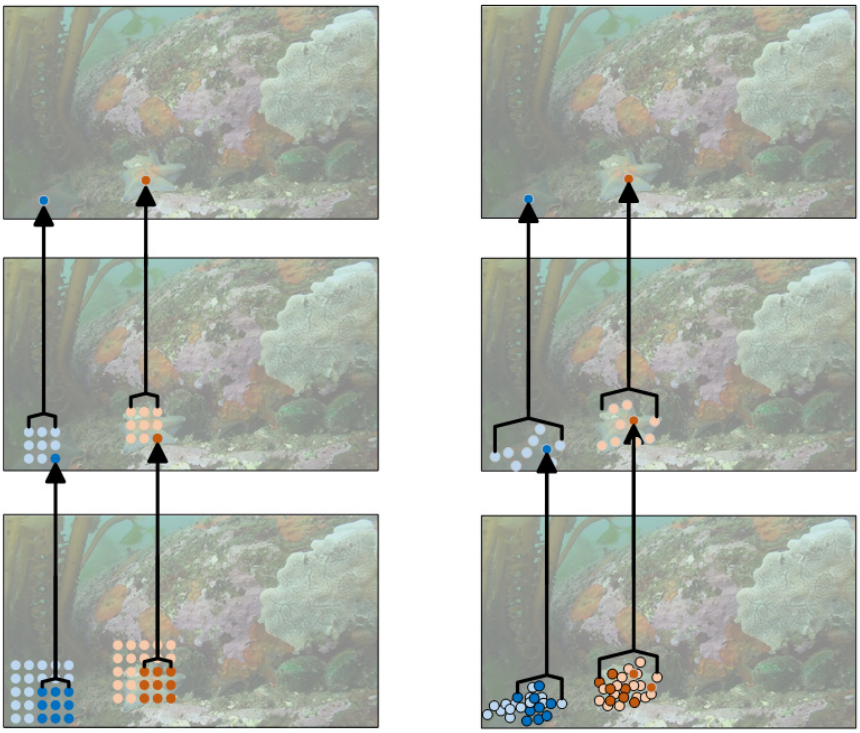


Figure 10. Proposed feature extraction network based on CNN-ViT fusion.

3.2.2. Deformable Convolution

In order to make the features output by the feature extraction network better adapt to underwater targets with different scales and shapes, adaptive offset deformable convolution is adopted in this study, as shown in Figure 11. The geometric structure of the

convolution kernel in the general convolution neural network is fixed; thus, its ability in geometric transformation modeling is limited. Without additional supervision, deformable convolution can add a spatial sampling position through additional offset and learn the offset from the target task. Therefore, compared with the fixed-sized convolution kernel, deformable convolution can improve the detection performance by slightly increasing the computational load.



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

Figure 11. The deformable convolution adopted in this study: (**a**) Convolution is to calculate the pixels within a regular rectangular range and extract features. (**b**) Deformable convolution has one more position mapping matrix, which represents the convolution kernel and the pixel after position mapping for calculation.

3.2.3. Trident Block

According to the previous discussion on the underwater dataset to be used in this study, the underwater organism targets in the dataset have large scale differences, which may be the main reason for the low recall rate of the detection model. To improve the recall rate of the detection model, the number and diversity of recommendation boxes in this study must be increased. Thus, different dilation rates are adopted in the C3 module to ensure multi-scale target detection with a multi-scale receptive field. In addition, inspired by Trident Net, parameter sharing is adopted for multiple bifurcations of convolution to obtain the C3 trident block, as shown in Figure 12. It has various dilation rates, can extract rich features, and reduce the number of parameters of the model. Although the parameters of multiple bifurcated convolution kernels are the same, different dilation rates can ensure the diversity of extracted features. Therefore, they can detect the targets in different receptive fields.

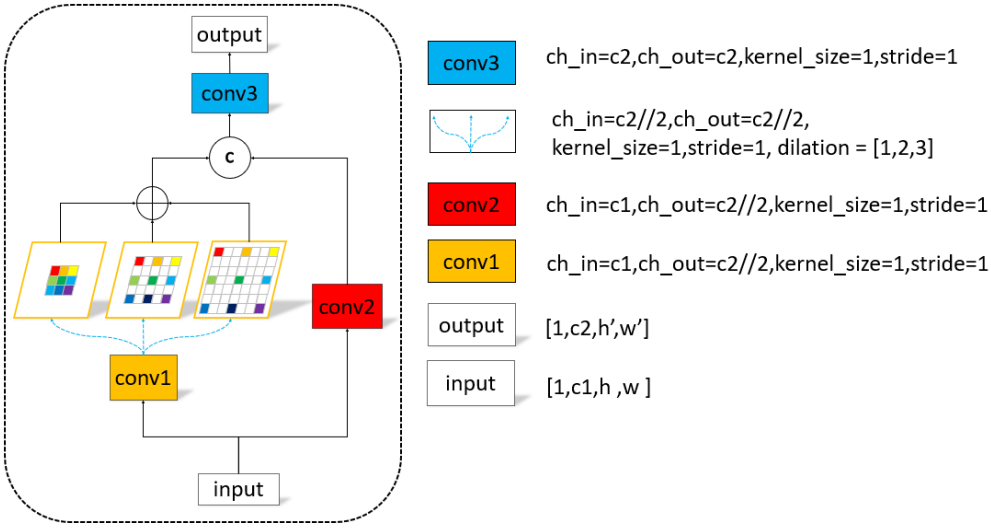


Figure 12. The trident block adopted in this study.

3.2.4. Loss Function

As previously mentioned, in order to achieve the balance of the number of samples, Focal Loss [21] is used as the classification loss and object confidence loss to weigh the loss of samples that are difficult to classify and the loss of a small number of positive samples. In addition, CIoU Loss is used as the regression loss to produce an accurate detection effect during dense target detection.

The objective function of Focal Loss is shown in Equation (1).

$$Loss_{cls} = FL(p, y) = \begin{cases} -\alpha(1-p)^{\gamma}\log(p), & if\ y = 1 \\ -(1-\alpha)p^{\gamma}\log(1-p), & otherwise \end{cases} \quad (1)$$

where p refers to the predicted object score, $p \in [0, 1]$; and $y \in \{\pm 1\}$ specifies the ground-truth class.

$$Loss_{obj} = FL(p, y) = \begin{cases} -\alpha(1-p)^{\gamma}\log(p), & if\ y = 1 \\ -(1-\alpha)p^{\gamma}\log(1-p), & otherwise \end{cases} \quad (2)$$

For each foreground pixel, q of the real class is set as the IoU value of the generated predicted boundary box and real boundary box, whereas q of the non-real class is set as 0. For each background pixel, its target score q is set to 0. In this study, α, γ are hyperparameters, which are valued as 0.25 and 1.5, respectively.

Focal Loss retains the penalty of the model for negative sample error judgment. Considering that the positive samples are scarcer than the negative samples, it is unbalanced in the number of positive and negative samples. The model is under the supervision of Focal Loss, which allows the model not only to correctly determine positive samples, but also to punish the wrong judgment of positive samples. Because high-quality positive samples are more conducive to achieving high average precision value (AP) than low-quality positive samples, the Focal Loss function focuses more on high-quality positive samples during model training. Hyperparameter α is used to balance the loss of positive and negative samples; and γ is used to lower the influence of a large number of negative samples on the loss function.

In this study, CIoU Loss is used as the boundary box regression loss. IoU represents the area ratio of intersection and union of rectangles A and B, which is used to measure the overlapping degree of two intersecting rectangles, as shown in Equation (3).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

The disadvantage of directly using *IoU* as the loss function is that $IoU = 0$ when the prediction box does not overlap with the real box, particularly when both of them are far away. Adjusting the position of the prediction box still cannot change *IoU*; thus, the gradient is always 0. Furthermore, finding the correct regression direction through gradient descent is difficult. In order to compensate for *IoU*'s inability to measure the distance between two disjoint rectangles, Zhaohui Zheng et al. proposed CIoU [36]. The main improvement of CIoU on *IoU* is to consider the similarity between the prediction box (green) and the real box (black), as shown in Figure 13.

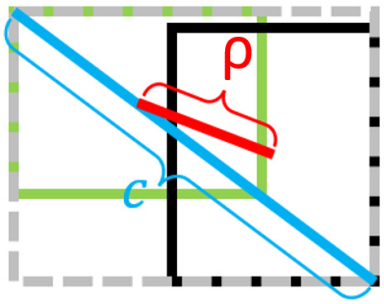


Figure 13. The similarity between the prediction box (green) and the real box (black) adopted by CIoU.

The expression of CIoU is as follows:

$$\begin{cases} Loss_{CIoU} = 1 - IoU + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha\nu \\ \nu = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \\ \alpha = \frac{\upsilon}{\upsilon - IoU + 1} \end{cases} \tag{4}$$

where b, b^{gt} refer to the central points of the prediction box and real box, respectively; ρ refers to the Euclidean distance between two central points. c refers to the diagonal distance in the minimum closure area that contains the prediction box and real box. ν is used to measure the similarity of the length–width ratio of the prediction box and real box; α refers to a weight function.

The complete loss function is calculated as follows:

$$Loss = Loss_{reg} + Loss_{cls} + Loss_{obj} = \lambda_1 \sum_i Loss_{CIoU}(bbox_i, bbox_{gt}) + \lambda_2 \sum_i \sum_c Loss_{cls}(p_{c,i}, q_{c,i}) + \lambda_3 \sum_i Loss_{obj}(conf_i, 1) \tag{5}$$

where $p_{c,i}, q_{c,i}$ refer to the predicted and targeted confidence scores of category c at position I in each level of the feature map output by the network; $bbox_i$ and $bbox_{gt}$, respectively, represent the prediction box and corresponding real box at position i on the feature map of each level output by the PANet; hyperparameters λ_1, λ_2, and λ_3 refer to the weights of regression loss, classification loss, and IoU loss of the boundary box, which are 0.05, 0.5, and 1.0, respectively, in this study.

4. Experiment

4.1. Experimental Details

The underwater organism dataset URPC2019 used for the overall framework performance test is provided by the "Underwater Robot Picking Contest" organized by the National Natural Science Foundation of China. Some of the images in this dataset were obtained by underwater robots with cameras. The robot is a remote-controlled robot designed and produced for the fishing of underwater marine organisms.

The target detection model is pre-trained on the COCO2017 dataset [37]. Then, the detection model proposed in this study is evaluated on the URPC2019 dataset. URPC2019 contains 5543 marked images, and the test set contains 1400 unmarked images. The dataset

contains four targets; namely, echinus, holothurian, scallop, and starfish. The detection model proposed in this study takes the YOLOv5 model released by ultralytics as a backbone. The pre-training model is obtained on the basis of the pre-training of the COCO2017 dataset. In addition, after a number of methods described in Section 3.2 are introduced, the detection model proposed in this study is obtained after finetuning. In this experiment, the training and detection processes of the model are run on two NVIDIA GeForce RTX 2080Ti.

In the training process, HSV image enhancement, image normalization, image random flip, image mosaic, proposed random expansion of small objects, proposed non-overlapping filling of scarce samples, and other training strategies are adopted, as shown in Table 1. batch_size is set to 4. SGD optimizer is adopted. $lr_0 = 0.01$, $lr_f = 0.2$, $momentum = 0.937$, $weight_decay = 0.0005$ (optimizer weight decay), $warmup_epochs = 3$, $warmup_momentum = 0.8$, $warmup_bias_lr = 0.1$. The training model is defaulted as 300 epochs. In the detection process, the input image size is fixed to (640, 640).

Table 1. Summary of the training strategies adopted.

Schedule	Value	Description
hsv_h	0.015	image HSV-Hue augmentation (fraction)
hsv_s	0.7	image HSV-Saturation augmentation (fraction)
hsv_v	0.4	image HSV-Value augmentation (fraction)
degrees	0.0	image rotation degrees (+/− deg)
translate	0.1	image translation (+/− fraction)
scale	0.5	image scale (+/− gain)
shear	0.0	image shear (+/− deg)
perspective	0.0	image perspective (+/− fraction), range 0–0.001
flipud	0.0	image flip up-down (probability)
fliplr	0.5	image flip left-right (probability)
mosaic	1.0	image mosaic (probability)
mixup	0.0	image mixup (probability)
expand	1.0	label bbox expand (pixel)
labelbalance	true	balance the number of labels (true/false)

4.2. Ablation Experiments

In this section, YOLOv5 is taken as the baseline model to carry out the experiments on the baseline model by random expansion of small objects and non-overlapping filling of scarce samples as well as ablation studies of various modules in the improved detection model.

4.2.1. Random Expansion of Small Objects

YOLOv5 is taken as the baseline model to verify the effectiveness of random expansion of small objects in model training. Figure 14a,b show the performance of the model in the test set when trained with and without random expansion of small objects, respectively. In the training process, the precision of the model without random expansion gradually increases, and the recall rate gradually decreases; mAP@0.5 and mAP@0.5:0.95 increase initially and then decrease and converge. By contrast, the precision, recall rate, mAP@0.5, and mAP@0.5:0.95 of the model with random expansion increase, and the convergence is higher. Consequently, the detection performance of the model trained with random expansion for small objects has been significantly improved on the URPC2019 dataset.

Figure 15a,b show the curves of the F1-score of the models trained with and without random expansion under different confidence levels. For models trained without and with expansion, when the confidence level is 0.461 and 0.560, respectively, the F1-score of all categories reaches the best 0.62 and 0.91. In addition, as shown in Figure 15a, the baseline model encountered great challenges when detecting scallops with very small size. However, after using random expansion strategy in training, the precision of scallops is greatly improved. Therefore, proposed random expansion strategy is not only beneficial to detect small size targets, but also helpful to regular size targets.

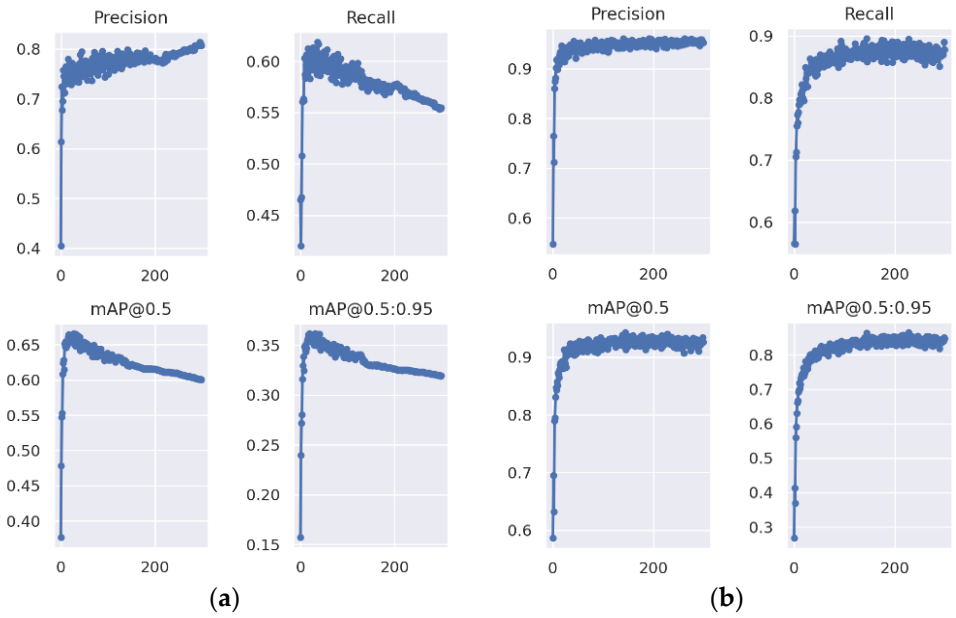


Figure 14. The performance of the model in the test set when trained with and without random expansion of small objects: (**a**) the performance of the model trained without random expansion; and (**b**) the performance of the model trained with random expansion.

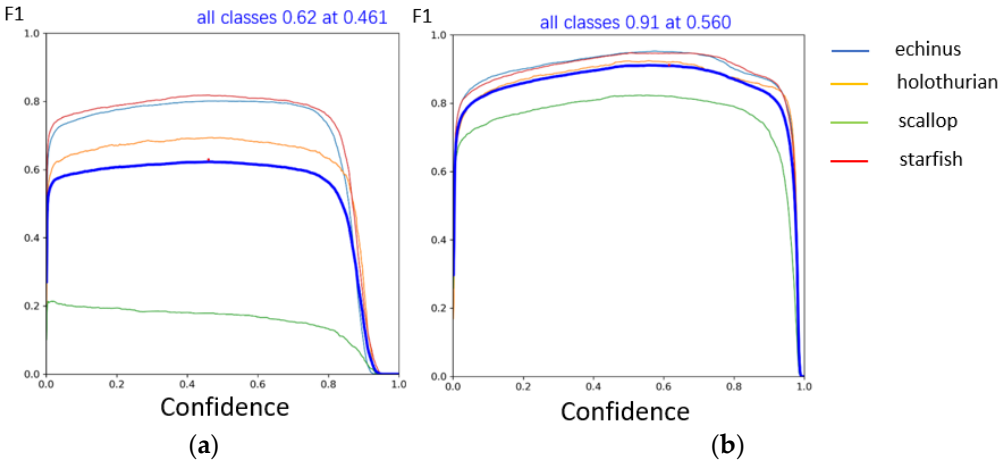


Figure 15. F1-score curve of the models trained with and without random expansion under different confidence levels: (**a**) F1-score curve of the model trained without random expansion; and (**b**) F1-score curve of the model trained with random expansion.

Figure 16 shows the comparison of detection results of some images in the test set, including the ground truth, the results of the baseline model trained without random expansion strategy and the results of the baseline model trained with random expansion strategy. It can be seen from Figure 16b,c that the model trained without random expansion can only detect a small number of underwater targets, whereas the ability of the baseline model trained with random expansion to recognize underwater small targets is significantly improved.

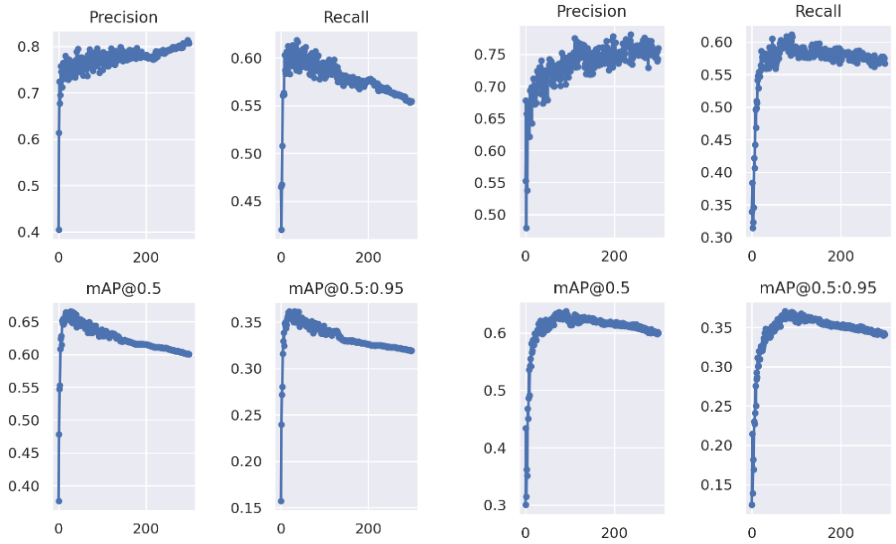


(a) (b) (c)

Figure 16. The comparison of detection results of some images in test set: (**a**) ground truth; (**b**) detection results of the baseline model trained without random expansion; and (**c**) detection results of the baseline model trained with random expansion. In (**b**,**c**), orange boxes represent starfish and red boxes represent echini.

4.2.2. Non-Overlapping Filling of Scarce Samples

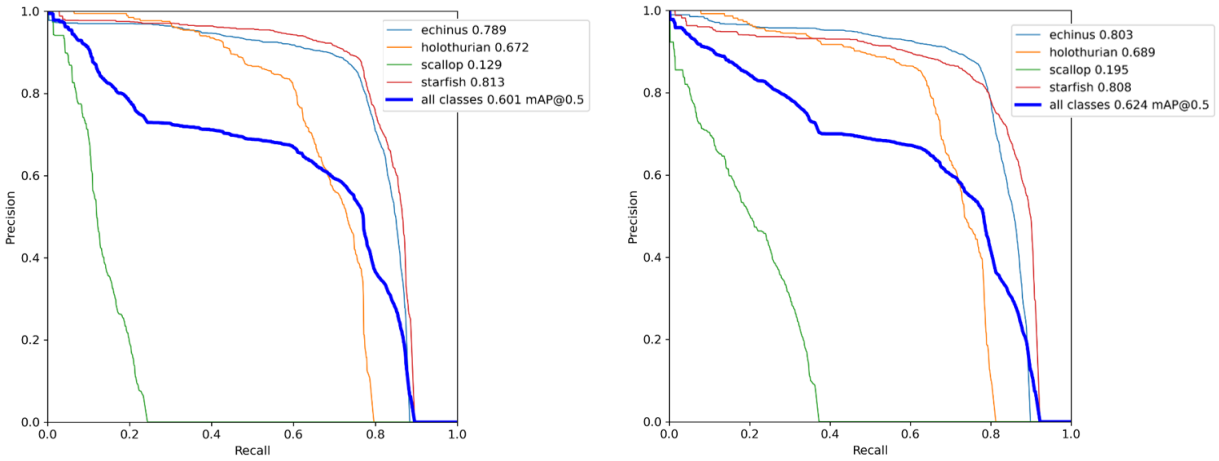
YOLOv5 is also taken as the baseline model to verify the effectiveness of non-overlapping filling of scarce samples in model training. Figure 17a,b show the performance of the model in the test set when trained with and without non-overlapping filling for scarce samples. In the training process, the precision of the model without non-overlapping filling gradually increases, and the recall rate gradually decreases; mAP@0.5 and mAP@0.5:0.95 initially increase and then decrease. By contrast, the precision, recall rate, mAP@0.5, and mAP@0.5:0.95 of the model with non-overlapping filling have a relatively stable change trend, with evident convergence and slow attenuation.



(a) (b)

Figure 17. The performance of the model in the test set when trained with and without non-overlapping filling of scarce samples: (**a**) the performance of the model trained without non-overlapping filling; and (**b**) the performance of the model trained with non-overlapping filling.

Figure 18a,b show the PR curves of the model trained without and with non-overlapping filling of scarce samples, respectively. In the URPC2019 dataset, there are more echini, fewer scallops and starfish, and the least holothurians. As shown in the figure, the area enclosed by the PR curve and the coordinate axis of the model with non-overlapping filling for scarce samples are large. Therefore, this strategy improves the performance of the model.



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

Figure 18. PR curve of the models trained without and with non-overlapping filling of scarce samples: (**a**) PR curve of the model trained without non-overlapping filling; and (**b**) PR curve of the model trained with non-overlapping filling.

Figure 19 shows the comparison of the detection results of some images in the test set, including the ground truth, the results of the baseline model trained without non-overlapping filling strategy and the results of the baseline model trained with non-overlapping filling strategy. By contrast, the baseline model with non-overlapping filling detects more targets. Combining with the results of Figures 17 and 18, we can conclude that non-overlapping filling of scarce samples can help the model obtain better detection results.



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

Figure 19. The comparison of detection results of some images in test set: (**a**) ground truth; (**b**) detection results of the baseline model trained without non-overlapping filling; and (**c**) detection results of the baseline model trained with non-overlapping filling.

4.2.3. Ablation Experiment of Each Module in The Proposed Model

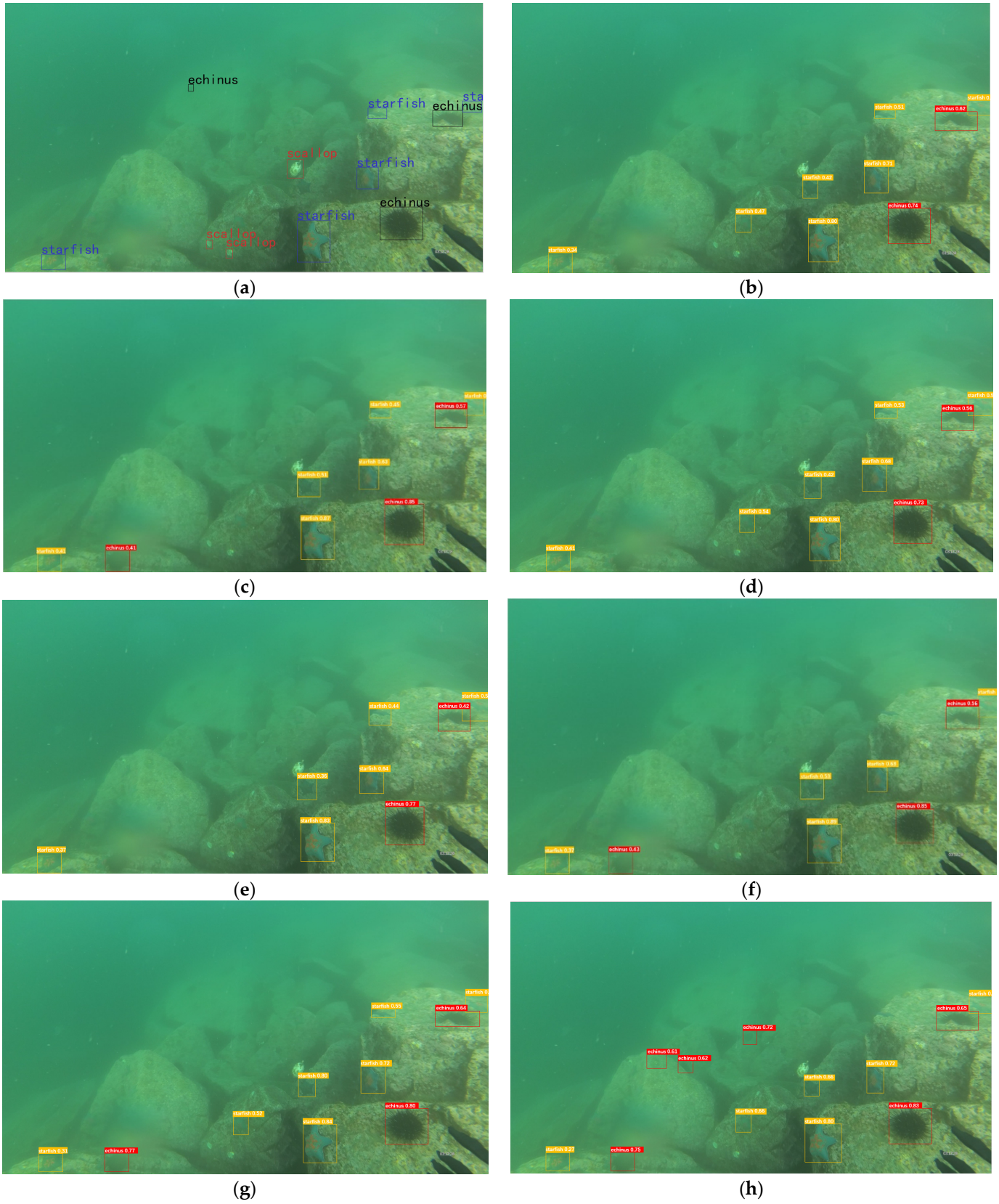
Based on the discussion in Section 3.2, the proposed detection model introduces ViT, deformable convolution (DfConv) and trident block (Trident) on the baseline of YOLOv5. In order to verify the role of these modules, we conducted ablation experiments on these modules on the URPC2019 dataset in this section. In the ablation experiments, the parameters for performance comparison include precision, recall, map@0.5, mAP@0.5:0.95, and F1-score, as shown in Table 2.

Table 2. Performance comparison of the ablation experiment of each module.

Method	Pretrained Model	Recall	Precision	mAP@0.5	mAP@0.5:0.95	F1-Score
YOLOv5	yolov5 × 6	0.7344	0.5975	0.6542	0.3536	0.6589
YOLOv5 + Trident [1, 2, 3]	yolov5 × 6	0.7492	0.6134	0.6829	0.3623	0.6745
YOLOv5 + Trident [1, 1, 1]	yolov5 × 6	0.7371	0.6375	0.6734	0.3571	0.6837
YOLOv5 + Trident [2, 2, 2]	yolov5 × 6	0.7149	0.6081	0.6647	0.3469	0.6572
YOLOv5 + Trident [3, 3, 3]	yolov5 × 6	0.7067	0.5747	0.6314	0.3294	0.6339
YOLOv5 + ViT	yolov5 × 6	0.7476	0.6036	0.6633	0.3614	0.6679
YOLOv5 + ViT + Trident [1, 2, 3]	yolov5 × 6	0.738	0.6307	0.6762	0.3651	0.6801
YOLOv5 + ViT+DfConv	yolov5 × 6	0.7596	0.6095	0.6783	0.3686	0.6763
Proposed model (YOLOv5 + ViT + Trident [1, 2, 3] + DfConv)	yolov5 × 6	0.7621	0.6422	0.6848	0.3717	0.6970

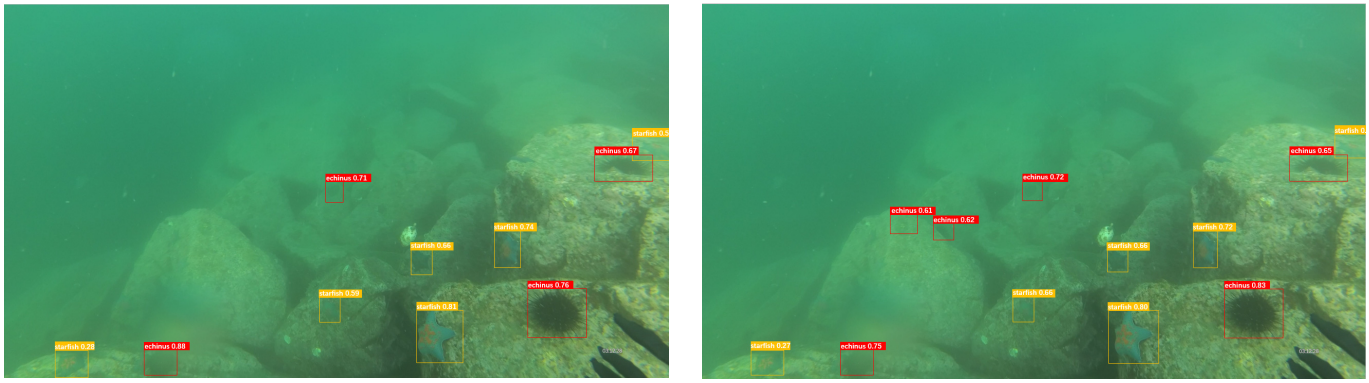
In addition to the performance comparison, Figure 20 also shows the typical detection results on the test set of the ground truth, the baseline model and the model after the introduction of each module.

As shown in Figure 20b, echinus and scallop with a small size are missed when the baseline model is applied. As shown in Figure 20c, a trident block with different dilation rates is introduced to YOLOv5 and different dilation rates help the model extract various features. As shown in Figure 20d, the performance of the trident block whose dilation rates are (1,1,1) is similar with YOLOv5. However, with the increase of the dilation rates, the performance of the trident block drops, as shown in Table 2 and Figure 20e,f. This is mainly because the bigger dilation rates lead to a bigger receptive filled and only help the model to detect bigger objects. As shown in Figure 20g, after the introduction of ViT in the baseline, echinus that is not labeled and located at the lower left is detected, but the overall detection effect is not satisfactory. As shown in Figure 20h, trident block is introduced to YOLOv5 + ViT, which supplements the multi-scale receptive field of convolution kernel and detects more underwater targets. As shown in Table 2, after the trident block is introduced, the recall rate is slightly decreased, but precision and map@0.5:0.95 are slightly increased. Considering that the boundary boxes of underwater organisms are not all approximate square rectangles, it is difficult to detect flat or slender holothurians with an approximate square rectangular box, thus deformable convolution is introduced on the basis of YOLOv5+ViT. As shown in Table 2, the recall rate and mAP@0.5:0.95 are improved after the introduction of deformable convolution, but precision and F1-score decreased. Since the trident block improves precision at the expense of recall, and deformable convolution improves recall at the expense of precision, we combine the two to further improve the performance of the final model. As shown in Figure 20j, the proposed synthesis model detects a scallop that other models could not detect, as well as an unlabeled starfish in the lower left corner of the image, and two unlabeled blurry echini in the center of the image. Meanwhile, Table 2 also shows that the recall rate, precision, map@0.5, and map@0.5:0.95 reach the highest.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 20. *Cont.*



(i) (j)

Figure 20. The typical detection results on the test set of the ablation experiment of each module: (**a**) ground truth; (**b**) detection results of the baseline YOLOv5; (**c**) detection results of the model YOLOv5 + Trident [1, 2, 3]; (**d**) detection results of the model YOLOv5 + Trident [1, 1, 1]; (**e**) detection results of the model YOLOv5 + Trident [2, 2, 2]; (**f**) detection results of the model YOLOv5 + Trident [3, 3, 3]; (**g**) detection results of the model YOLOv5 + ViT; (**h**) detection results of the model YOLOv5 + ViT + Trident [1, 2, 3]; (**i**) detection results of the model YOLOv5+ViT+DfConv; and (**j**) detection results of the proposed model. In (**b**–**j**), the red, blue and orange boxes represent the echinus, scallop, and starfish, respectively.

4.3. Comparative Experiments

The detection model proposed in this paper is compared with the model proposed by Zhang et al. in [38], the EDR-D0 modal proposed by Jia et al. in [39], and the classical one-stage models, including SSD, YOLO-v3, YOLO-v4 and YOLO-v5, on the URPC2019 dataset to evaluate the performance of the proposed detection model. Meanwhile, in reflecting the role of the proposed data augmentation method, the training set using the proposed data augmentation method is integrated with the proposed detection model to form the proposed framework. The performance of the proposed framework is compared with abovementioned models including the proposed detection model, and the comparative experiment includes the detection performance comparison and the detection real-time comparison.

4.3.1. Detection Performance Comparison

In the model detection performance comparison, the size of the input image of each model is uniformly set to (640, 640), and the pre-training model with the same level and specification is selected. The performance indicators used for comparison include recall rate, precision, AP at IoU = 0.5 and mAP at IoU = 0.5:0.95. The comparison results are shown in Table 3. Typical test results in the test set are shown in Figure 21.

Table 3. Detection performance comparison of the proposed framework with the two-stage model proposed by Zhang et al. in [38], the EDR-D0 modal proposed by Jia et al. in [39], the classical one-stage models, and the proposed model.

Method	Pretrained Model	Recall	Precision	mAP@0.5	mAP@0.5:0.95
Zhang et al. [38]	ResNet-50 + Cascade	0.6629	0.5612	0.6791	0.4142
EDR-D0 [39]	EfficientNet-B0	0.6120	0.6267	0.6443	0.3374
SSD	VGG-16	0.5420	0.5257	0.5676	0.2927
YOLO-v3	DarkNet-53	0.6434	0.5596	0.5713	0.3248
YOLO-v4	CSPDarkNet-53	0.6767	0.6024	0.6374	0.3427

Table 3. *Cont.*

Method	Pretrained Model	Recall	Precision	mAP@0.5	mAP@0.5:0.95
YOLO-v5	yolov5-l	0.7344	0.5975	0.6542	0.3536
Proposed model	yolov5-l	0.7621	0.6422	0.6848	0.3717
Proposed framework (Data augmentation + Proposed model)	yolov5-l	0.8853	0.8917	0.9249	0.8091

As shown in Table 3 and Figure 21, with the continuous improvement of the model, the performance is becoming better and better. Compared with SSD, YOLOv3, and YOLOv4, YOLOv5 introduces the Focus structure, which lowers the computation load of the model by reducing the width and height of the image and then stacking along the channel. In addition, it further uses the training skills adopted by YOLOv4 to achieve the best performance among these models.

Taking YOLOv5 as the baseline model, the detection model proposed in this study has made several targeted improvements based on the characteristics of marine organisms. Therefore, its performance on the URPC2019 dataset exceeds other models for comparison. As shown in Figure 21h, it detected an incomplete echinus at the bottom of the image and a starfish at the top of the image that other models did not detect.

In addition to improving the detection model, we also introduce two data augmentation strategies aiming at the problems of marine organism datasets. After integrating the training set using the data augmentation strategy with the proposed detection model, compared with other models for comparison (including the proposed detection model in this study), the proposed framework detected more small echini and a holothurian (scarce sample in URPC2019) at the top of the image, as shown in Figure 21i. As shown in Table 3, its detection precision and recall rate are greatly improved compared to the models without data augmentation. In addition, the proposed data augmentation strategy only consumes the time required by the training, but it does not bring a negative impact on the model size and inferencing time. In terms of detection precision, compared with the baseline model YOLOv5, mAP@0.5 and mAP@0.5:0.95 of the proposed framework have been improved by 27% and 45.6%, respectively.

Based on the inferencing test results of the model, the confusion matrix of YOLOv5, the proposed detection model, and the proposed framework are drawn in Figure 22. In Figure 22a, it shows that YOLOv5 only detected 12% of the scallops correctly and recognized 88% of the scallops as the background. It indicates that most of the scallops were not detected. In Figure 22b, after the model finetuning, the proposed model detected 16% of the scallops correctly and recognized 7% of the scallops incorrectly as starfish, and recognized 77% of the scallops as background. In Figure 22c, it shows that after data augmentation, the proposed framework detected 79% of the scallops, only missed 16% of the scallops, and recognized 4% of the scallops incorrectly. Similarly, Figure 22a shows that YOLOv5 detected 61% of the holothurians, but it did not detect 33% of the holothurians. After the model finetuning, the proposed model increased the precision to 65%, as shown in Figure 22b. In Figure 22c, it shows that after data augmentation, the precision of the holothurians was increased to 91%, and only 5% of the holothurians were missed.

Based on the comparison among the confusion matrices, before using the data augmentation strategy, compared with the YOLOv5 baseline model, the proposed model has slightly improved the precision of detecting scallops (small target) and holothurians (scarce target). After using the training set with data augmentation, the precision of all kinds of targets has been improved, especially the precision of scallops (small target) and holothurians (scarce target) has been greatly improved, which fully demonstrates the effectiveness of the proposed data augmentation strategy.

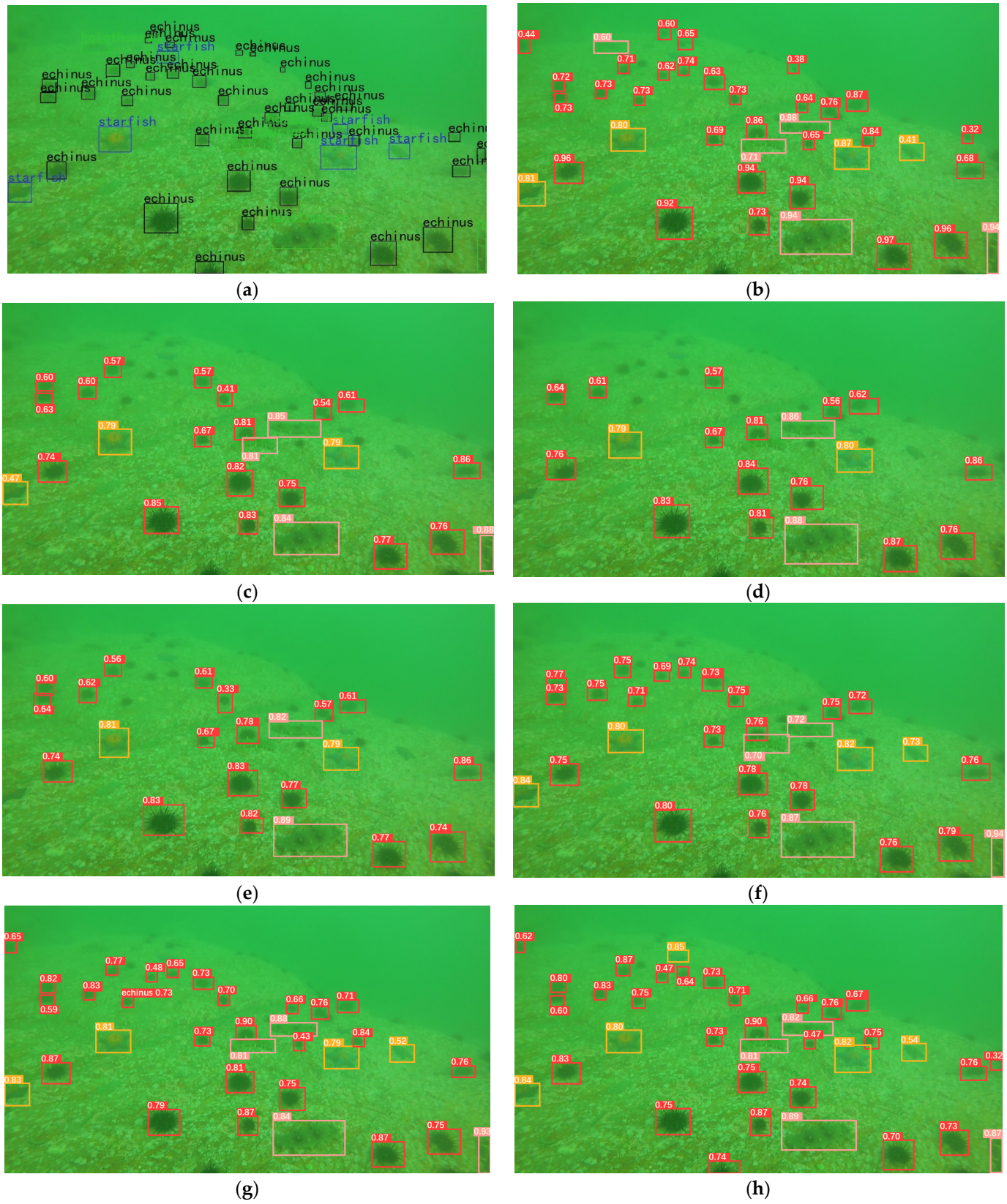
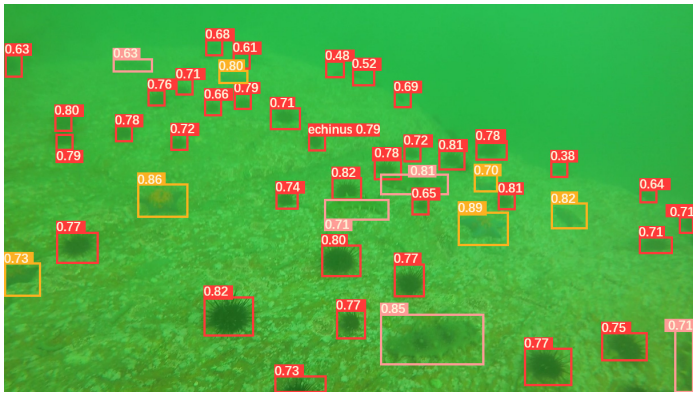
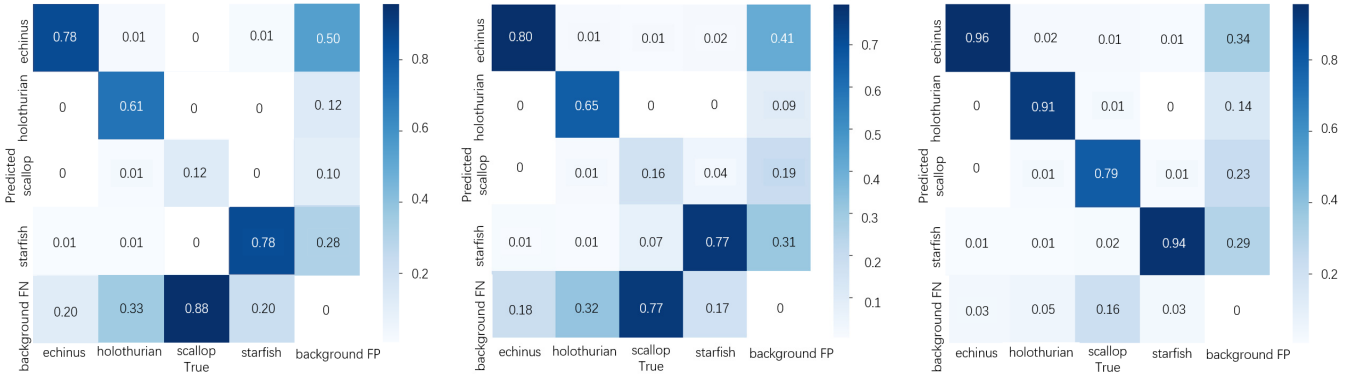


Figure 21. *Cont.*



(**i**)

Figure 21. The typical detection results on the test set of the detection performance comparison: (**a**) ground truth; (**b**) detection results of the model proposed by Zhang et al. in [38]; (**c**) detection results of EDR-D0 modal proposed by Jia et al. in [39]; (**d**) detection results of SSD; (**e**) detection results of YOLOv3; (**f**) detection results of YOLOv4; (**g**) detection results of YOLOv5; (**h**) detection results of the proposed model; and (**i**) detection results of the proposed framework. In (**b**–**i**), the red, orange and pink boxes represent the echinus, starfish and holothurian, respectively. In (**b**–**i**), red boxes represent echini, pink boxes represent holothurians and orange boxes represent starfish).



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

Figure 22. The confusion matrices based on the inferencing test results: (**a**) YOLOv5; (**b**) proposed model; and (**c**) proposed framework.

4.3.2. Detection Real-Time Comparison

According to the foregoing, the proposed framework will eventually be applied to embedded platforms, thus it should have a small size and practical application feasibility. To verify the feasibility of the practical application, the complete proposed framework is compared with SSD, YOLOv3, YOLOv4, and YOLOv5 with regard to the framework size and inferencing frame rate. In real-time comparative experiment, the test set with 1400 underwater images is first inferred on the PC with GTX1650. Second, the proposed framework is deployed on a Jetson Xavier NX board to complete the inferencing of the same 1400 underwater images, and the fps is also measured. During the experiments, the size of the input image is set to $(640,\ 640)$, and the batch size is set to 1. The experimental results are shown in Table 4.

Table 4. Detection real-time comparison of proposed framework with the classical one-stage model.

Method	Parameters	Fps	
		On PC	On Jetson Xavier NX
SSD	33.0 M	15.82	20.32
YOLO-v3	61.5 M	14.83	17.43
YOLO-v4	52.5 M	15.01	18.95
YOLO-v5	47.0 M	15.34	17.08
Proposed Framework	64.2 M	14.66	17.46

It can be seen from Table 4 that although the proposed framework has the largest size, it is similar to other models and can be normally deployed on an embedded platform. Whether on PC or an embedded platform, the inferencing fps of the proposed framework is slightly different from that of other models except for SSD, and all of them can achieve quasi-real-time/real-time detection. Therefore, the proposed framework achieves the balance of speed and precision under the condition of optimal precision, which can meet the requirements of high-precision, real-time detection on underwater mobile platforms.

5. Conclusions

This study proposes a marine organism detection framework that includes a dataset augmentation strategy and improved one-stage detection model. Using random expansion of small objects and non-overlapping filling of scarce samples, the proposed framework solves the problems that some targets in the dataset are too small to detect, and the uneven number of samples between different organisms leads to the low precision of scarce organisms. The proposed framework takes YOLOv5 as the baseline in the target detection. Aimed at the characteristics of marine organism images, ViT is introduced into the feature extraction network to achieve the fusion of CNN and ViT, which has both the local-range information capture ability of CNN and long-range information capture ability of ViT. In addition, deformable convolution and trident block are also introduced to extract abundant features through the multi-scale receptive field and deal with the feature extraction of marine organisms with different shapes and scales. Focal Loss is used as the classification loss and the target confidence loss to weigh the loss of samples that are difficult to classify and the loss of a small number of positive samples. Moreover, CIoU loss is used as the regression loss to produce more accurate detection results in the dense target detection. The experimental results show that, compared with the baseline model YOLOv5, the mAP of the proposed framework can be increased by 27%, while the detection speed remains unchanged. The frame rate of the proposed framework on the Jetson Xavier NX embedded platform can reach 17.46, and it gives consideration to both performance and real-time, so as to achieve high-precision real-time detection of the marine organisms on the underwater mobile platform.

Author Contributions: Conceptualization, H.Y. and Y.Z.; methodology, X.J. and M.P.; software, X.J. and Y.Z.; validation, S.L. and Z.L.; writing—original draft preparation, X.J. and Y.Z.; writing—review and editing, H.Y., M.P. and G.Y.; supervision, J.L.; funding acquisition, H.Y. and G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Project of China (grant No. 2022YFC2803600); the Key Research and Development Program of Zhejiang Province (grant No. 2022C03027, 2022C01144); the Public Welfare Technology Research Project of Zhejiang Province (grant No. LGF21E090004, LGF22E090006); Professional Development Program for Domestic Visiting Teachers in Zhejiang Universities (grant No. FX2021006); and Zhejiang Provincial Key Lab of Equipment Electronics.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Teng, B.; Zhao, H. Underwater target recognition methods based on the framework of deep learning: A survey. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420976307. [CrossRef]
2. Qi, J.; Gong, Z.; Xue, W.; Liu, X.; Yao, A.; Zhong, P. An Unmixing-Based Network for Underwater Target Detection From Hyperspectral Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5470–5487. [CrossRef]
3. Rova, A.; Mori, G.; Dill, L.M. One fish, two fish, butterfish, trumpeter: Recognizing fish in underwater video. *DBLP* **2007**, 404–407.
4. Yuan, F.; Huang, Y.; Chen, X.; Cheng, E. A Biological Sensor System Using Computer Vision for Water Quality Monitoring. *IEEE Access* **2018**, *6*, 61535–61546. [CrossRef]
5. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
7. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
9. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
10. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
11. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
12. Github. YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 28 May 2021).
13. Li, H.; Zhuang, P.; Wei, W.; Li, J. Underwater Image Enhancement Based on Dehazing and Color Correction. In Proceedings of the 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Xiamen, China, 16–18 December 2019; pp. 1365–1370.
14. Luo, W.; Duan, S.; Zheng, J. Underwater Image Restoration and Enhancement Based on a Fusion Algorithm With Color Balance, Contrast Optimization, and Histogram Stretching. *IEEE Access* **2021**, *9*, 31792–31804. [CrossRef]
15. Inzartsev, A.V.; Pavin, A.M. AUV Cable Tracking System Based on Electromagnetic and Video Data. In Proceedings of the OCEANS 2008—MTS/IEEE Kobe Techno-Ocean, Kobe, Japan, 8–11 April 2008; pp. 1–6.
16. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. [CrossRef]
17. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
18. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Qi, L.; Sun, J.; Jia, J. Dynamic Scale Training for Object Detection. *arXiv* **2020**.
19. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Meng, G.; Xiang, S.; Sun, J.; Jia, J. Stitcher: Feedback-driven Data Provider for Object Detection. *arXiv* **2020**, arXiv:2004.12432.
20. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
21. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 318–327. [CrossRef] [PubMed]
22. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:1706.03762.
23. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
24. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
25. Raffel, C.; Shazeer, N.M.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2019**, arXiv:1910.10683.
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.
27. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
28. Yeh, C.H.; Lin, C.H.; Kang, L.W.; Huang, C.H.; 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. [CrossRef]

29. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale Match for Tiny Person Detection. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020.
30. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
31. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z.X. Scale-Aware Trident Networks for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6053–6062.
32. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
33. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. In Proceedings of the Computer Vision–ECCV 2020 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 260–275.
34. Han, F.; Yao, J.; Zhu, H.; Wang, C. Marine Organism Detection and Classification from Underwater Vision Based on the Deep CNN Method. *Math. Probl. Eng.* **2020**, *2020*, 3937580. [CrossRef]
35. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
36. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *Neurocomputing* **2021**, *506*, 146–157. [CrossRef]
37. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Zitnick, C.L. *Microsoft COCO: Common Objects in Context*; Springer International Publishing: Berlin/Heidelberg, Germany, 2014.
38. Zhang, X.; Fang, X.; Pan, M.; Yuan, L.; Zhang, Y.; Yuan, M.; Lv, S.; Yu, H. A Marine Organism Detection Framework Based on the Joint Optimization of Image Enhancement and Object Detection. *Sensors* **2021**, *21*, 7205. [CrossRef] [PubMed]
39. Jia, J.; Fu, M.; Liu, X.; Zheng, B. Underwater Object Detection Based on Improved EfficientDet. *Remote Sens.* **2022**, *14*, 4487. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.