Journal of
Marine Science and Engineering

MDPI

Article

An Improved YOLOV5 Based on Triplet Attention and Prediction Head Optimization for Marine Organism Detection on Underwater Mobile Platforms

Yan Li [1,2,*], Xinying Bai [1,2,3] and Chunlei Xia [4,*]

[1] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[2] Institutes of Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China
[3] College of Information, Liaoning University, Shenyang 110136, China
[4] Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, China
[*] Correspondence: liyan1@sia.cn (Y.L.); clxia@yic.ac.cn (C.X.)

Abstract: Machine vision-based automatic detection of marine organisms is a fundamental task for the effective analysis of production and habitat changes in marine ranches. However, challenges of underwater imaging, such as blurring, image degradation, scale variation of marine organisms, and background complexity, have limited the performance of image recognition. To overcome these issues, underwater object detection is implemented by an improved YOLOV5 with an attention mechanism and multiple-scale detection strategies for detecting four types of common marine organisms in the natural scene. An image enhancement module is employed to improve the image quality and extend the observation range. Subsequently, a triplet attention mechanism is introduced to the YOLOV5 model to improve the feature extraction ability. Moreover, the structure of the prediction head of YOLOV5 is optimized to capture small-sized objects. Ablation studies are conducted to analyze and validate the effective performance of each module. Moreover, performance evaluation results demonstrate that our proposed marine organism detection model is superior to the state-of-the-art models in both accuracy and speed. Furthermore, the proposed model is deployed on an embedded device and its processing time is less than 1 s. These results show that the proposed model has the potential for real-time observation by mobile platforms or undersea equipment.

Keywords: marine organism; target identification; deep learning; attention mechanism; model optimization

check for updates

Citation: Li, Y.; Bai, X.; Xia, C. An Improved YOLOV5 Based on Triplet Attention and Prediction Head Optimization for Marine Organism Detection on Underwater Mobile Platforms. *J. Mar. Sci. Eng.* **2022**, *10*, 1230. https://doi.org/10.3390/jmse10091230

Academic Editors: Simone Marini, Jacopo Aguzzi, Giacomo Picardi, Damianos Chatzievangelou, Sascha Flögel, Sergio Stefanni, Peter Weiss and Daniel Mihai Toma

Received: 27 July 2022
Accepted: 30 August 2022
Published: 2 September 2022

1. Introduction

Underwater organism observation is an important topic in the field of underwater object detection, which can offer an effective means to evaluate the abundance of marine organisms and sensitively predict environmental changes. For instance, it can autonomously and intelligently identify and analyze the number of sea cucumbers, scallops, and other seafood, as well as invasive organisms in marine ranches, which were previously done mainly manually. Therefore, the autonomous monitoring and accurate identification of the seafood in marine ranches not only helps farmers to control the growth status of seafood and the habitat changes in real time but also releases manpower from dangerous and heavy workloads. Numerous advanced acoustic or optical-based detection tools have been applied to the identification of marine organisms [1], and meanwhile underwater robots are also further considered to provide a larger-scale observation based on their ability to be incorporated into autonomous mobile devices [2].

In contrast to the acoustic-based approach, the optic-based approach has the advantages of high resolution, low cost, and ease of operation. Thus, optic-based marine organism identification methods have attracted increasing attention and are becoming a research trend. However, there are still many significant technical challenges in the optic-based marine organism identification scenario. (i) Physical phenomena such as light scattering and absorption in the water environment result in unclear underwater imaging, coupled with the intensification of marine pollution in recent years, further reducing the quality of underwater imaging [3]. These factors not only limit the observation range and efficiency but also reduce the target identification accuracy when implementing some automatic methods with low-quality images. (ii) Marine organisms are frequently present in small-sized images and the benthic organisms are often intertwined with the complex background, which makes the accurate extraction of the marine organism difficult from the complex background. The conventional target identification approaches are not adaptive and robust to these challenges.

Following the rapid development of image processing technologies, the deep learning theory has been convincingly and successfully applied in many fields due to its superior performance in feature extraction [4,5], such as face recognition [6], maritime ship target recognition [7], plankton image analysis [8], and so on.

Underwater enhancement approaches proposed in the recent literature can be divided into traditional methods and deep learning-inspired methods. Since underwater images are considered to be similar to hazy images in some ways, dehazing algorithms based on dark channel prior were used to improve degraded underwater images [9]. This approach and its derivatizations were almost entirely based on prior knowledge or statistical assumptions about a scene to enhance the images globally [10]. To solve the problem of inaccurate depth estimation of underwater scenes, IBLA and ULAP were proposed based on image blurriness and light absorption, as well as based on underwater light attention prior [11,12]. These approaches will be unsatisfactory because of prior invalidation in complex scenes. Apart from image formation model (IFM)-based approaches, IFM-free methods are simple in principle and can also effectively improve the visual effect of images. As one typical IFM-free method, relative global histogram stretching (RGHS) [13] achieved better image quality and less noise by contrast correction and color correction in shallow water image enhancement. According to the diverse network models, deep learning-inspired approaches are separated into two categories: CNN-based methods and GAN-based methods [14,15]. The representatives of CNN-based methods include Water-net [16], LANet [17], and so on. Water-net performed a weighted fusion of the input image by the network self-learning method to obtain the enhanced underwater image. LANet solved the problems of color cast and low illumination on underwater images. To expand the training data, GANs were utilized to generate realistic underwater images. WaterGAN trained synthetic underwater images using a two-stage color correction network [18]. The underwater generative adversarial network (UGAN) converted blurred underwater images into high-resolution images [19]. However, these data-driven enhancement approaches demand a large number of training samples and a long training time. In addition, changes in the environment often require the model to be retrained. These result in difficulties in practical applications.

The traditional method of object detection is to train the detector utilizing prior artificially defined features, such as color, texture, and geometric features. Qiao et al. [20] proposed an approach to recognize the sea cucumber using principal component analysis extraction and support vector machine classification. Hasija et al. [21] classified fish species using an enhanced discriminant analysis method of display graph embedding based on display image processing. These artificially defined features of objects have high pertinency and interpretability but poor scalability, which results in the loss of high-dimensional features, especially in the complex and unstructured environment [22]. Therefore, most of the research published in recent years mainly employs deep learning models to extract more enrichment features of marine organisms. Peng et al. [23] designed a detection strategy for sea cucumbers based on a modified feature pyramid network with a shortcut connection. Cao et al. [24] proposed a real-time and robust detector to detect underwater live crabs by using a lightweight MobileNetV2 as the backbone of a single-shot multi-box detector (SSD) and replacing the standard convolution with depthwise separable convolution in the prediction layers. Li et al. [25,26] designed an in situ zooplankton detection neural network with a densely connected YOLOV3 model to reduce the loss of features in the network transmission and generated samples with a CycleGAN to address the problem of species distribution imbalances. In addition, there are many deep learning-inspired approaches that have been successfully applied to fish identification [27,28].

Human vision has the potential to quickly focus on the targets or interesting regions when facing a complex sense. Inspired by the human visual system, the attention mechanism has been used as an effective module in deep learning tasks in recent years [29–31]. Google first combined the attention mechanism with recurrent neural networks (RNN) and successfully applied it in image classification tasks [32]. Hu et al. [33] proposed the squeeze and excitation network (SE-Net), to obtain global information in the channel dimension by global average pooling operation, and then selectively emphasized important features and suppressed less important features with this information. However, this only considers the attention in the channel dimension and ignores the information in the spatial dimension. Another representative attention module is the convolutional block attention module (CBAM), which fuses attention information in both channel and spatial dimensions by adding global average pooling and global max pooling [34]. Subsequently, Park et al. [35] proposed a bottleneck attention module (BAM) based on channel and spatial dimensions to weight significant features; the difference with CBAM is its attention computation in a parallel manner. Although the above methods exhibit a significant improvement over previous achievements in target recognition integrated with CNNs, the semantic interaction between features of different dimensions is not taken into account. To avoid missing spatial detail information, a structure with three parallel branches is proposed to capture dependencies between the channel and spatial dimensions of the input tensor, respectively [36].

Therefore, this paper presents a novel marine organisms detection model based on a deep learning approach for automatic marine organism surveys or harvesting in underwater scenes. The proposed model is developed based on a lightweight deep neural network that ensures the potential for mobile observation of marine organisms by underwater robots. In this work, an adaptive underwater image enhancement method named relative global histogram stretching (RGHS) is employed primarily to optimize the imaging quality and to improve the accuracy of object detection. Subsequently, convolutional triplet attention mechanism modules are introduced into a lightweight deep neural network, YOLOV5, to enhance the identification performance of organism detection from complex scenes by encoding inter-channel and spatial information. Moreover, the structure of the prediction head is modified to improve the capability of detecting small targets. Figure 1 visualizes the improvement of our proposed detection model by comparing it with the original YOLOV5 model.

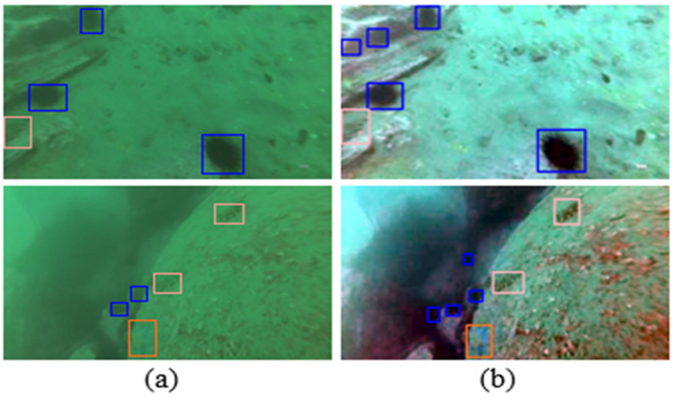

Figure 1. Marine object detection: (**a**) detection by YOLOV5; (**b**) detection by the proposed model.

The main contributions of this work are summarized as follows:

(1) The performance of three state-of-the-art attention modules for underwater object detection with YOLOV5 was evaluated. Triplet attention-based YOLOV5 achieved

the best accuracy since triplet attention can capture the cross-dimension interaction of the channel dimension and spatial dimension.

(2) An optimization strategy for underwater small target detection was presented by developing a four-scale prediction head of YOLOV5, which ensured the performance of detection to the targets with significant variation in size.

(3) The improved YOLOV5 was further tested on an embedded device (Nvidia Jetson Nano), and the inference time reached real-time performance (0.25 s). The overall processing time for one frame was 0.98 s, including 0.25 s for detection and 0.73 s for image enhancement.

(4) An underwater image enhancement algorithm, relative global histogram stretching, was utilized to improve underwater image quality. Experimental results proved that image enhancement was efficient in improving the performance of underwater target detection.

The remainder of this article is organized as follows. The proposed detection model and evaluation metric are described in Section 2. The experiment design and the discussions of the obtained results are reported in Section 3. Finally, the conclusions are drawn in Section 4.

2. Materials and Methods

2.1. Experimental Data

The optical image data of marine organisms were captured by the Underwater Robot Picking Contest hosted at Zhangzidao island in China. Four different species of marine organisms are selected in this work, including urchin, sea cucumber, starfish, and scallop. In a total of 3100 images, 2480 images are randomly selected as the training data, and the remaining 620 images are used as the testing data. The marine organisms were labeled as ground truth based on our existing knowledge with a graphical annotation tool named LableImg and converted to PASCAL VOC format to validate the accuracy of the model identification results.

2.2. Model Backbone

The backbone is one of the core parts of the deep neural network to extract the features of the input images. Considering the requirement both for the accuracy of the object detection and time consumption, a typical one-stage detection model, YOLOV5, is employed in this work due to its lightweight parameters and excellent detection performance compared to the two-stage detection model, e.g., Faster RCNN. YOLOV5 is the latest version of the YOLO series, released in 2020, surpassing the other previous versions in accuracy and speed.

Figure 2 illuminates the architecture of the proposed marine organism identification model, which includes four parts: enhancement, backbone, neck, and prediction. The input of the backbone of the detection model is an enhanced optical underwater image with the RGHS method at the enhancement part after comparing the performance and analyzing the suitability to the backend structures of the detection model.

There are four versions depending on the length and width of the backbone networks, including YOLOV5s, YOLOV5m, YOLOV5l, and YOLOV5x. In this work, the YOLOV5s, with the smallest parameters and fastest speed, is embedded into the proposed marine organism detection model. The backbone is composed of Focus, Conv-BN-SiLU (CBS), Cross-Stage Partial (CSP), and Spatial Pyramid Pooling (SPP) modules. The detailed architecture of these modules is also exhibited in Figure 2. The Focus block mainly consists of four parallel slice layers to extract pixels from input high-resolution images periodically and reconstruct them into low-resolution images. The CSP module, inspired by CSPNet, halves the number of channels by performing separate convolution operations to allow the model to learn more distinguishing features. Aiming to further broaden the receptive field and aid in segregating contextual characteristics, the SPP module is plugged at the end of the backbone network.

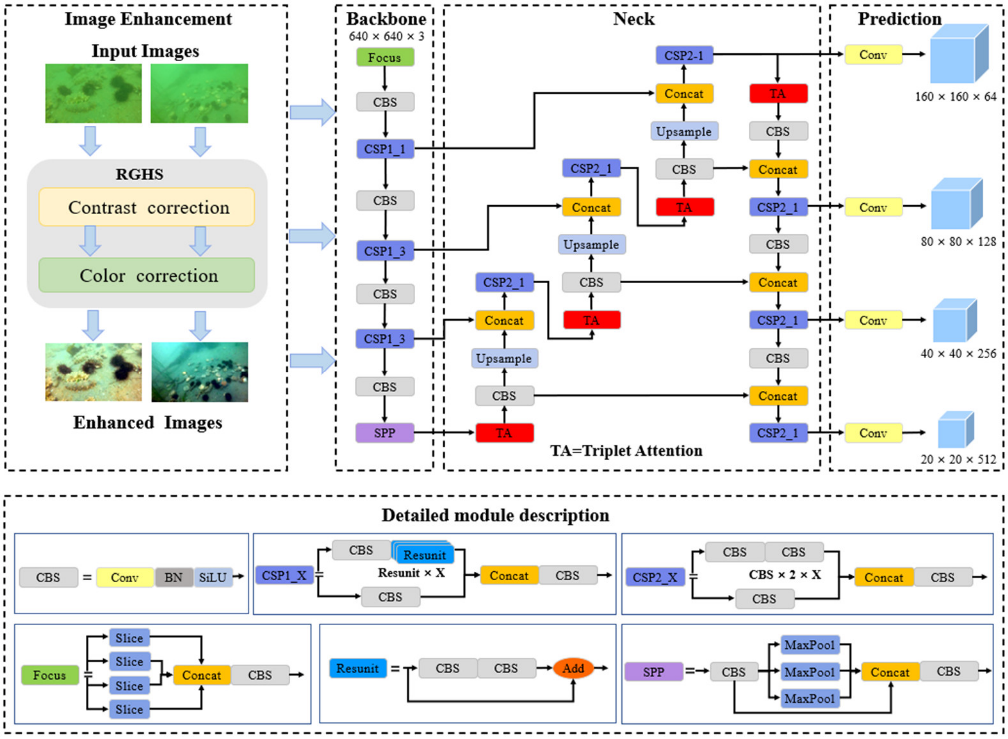

Figure 2. The architecture of the proposed marine organism identification model.

2.3. Triplet Attention Module

The neck part aims to further improve the feature extraction ability of the proposed detection model. It is succeeded by PANet, using an FPN structure and PAN module to convey strong semantic features and positioning features from top to bottom and bottom to top, respectively. Many triplet attention modules are embedded into the neck part of the identification model to improve the feature representations. Figure 3 shows the structure and framework of the triplet attention module.

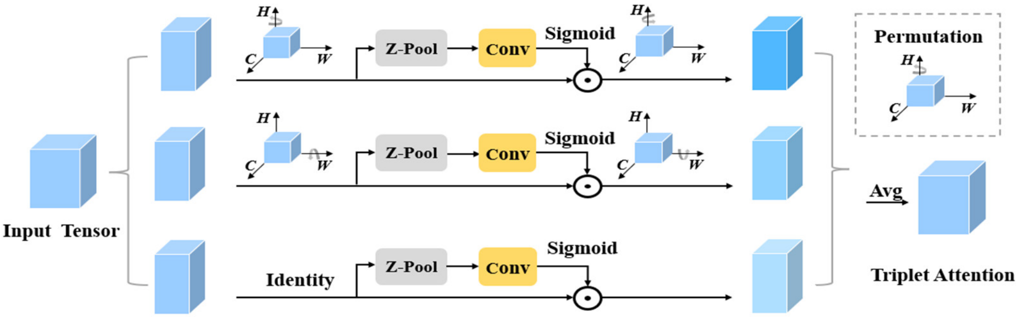

Figure 3. Structure and framework of triplet attention module.

Different from the other most prominent attention mechanism modules, such as SE-Net, CBAM, and BAM, the triplet attention modules compute attention weights by capturing cross-dimension interactions between (Channel, Height), (Channel, Width), and (Height, Width) dimensions of the input tensor, respectively, using three parallel branches. By concatenating the average pooled and max pooled features across each dimension, the Z-pool layer reduces the zeroth dimension of the tensor to two. The advantage of the Z-pool layer is to obtain a detailed representation of the actual tensor while also reducing its depth, to make the following computations more efficient. The Z-Pool is formulated as:

$$\mathrm{Z}-\mathrm{pool}(\chi) = \left[\mathrm{MaxPool}_{0\mathrm{d}}(\chi), \mathrm{AvgPool}_{0\mathrm{d}}(\chi)\right] \quad (1)$$

where 0d is the zeroth dimension across which the module reduces the dimension of the tensor χ to two dimensions via max and average pooling.

A triplet attention module is inserted before each prediction head to enrich the target features, so as to improve the recognition accuracy at different scales, as shown in Figure 2.

2.4. Prediction Head Optimization

In underwater observation, marine organisms could be present on various scales due to their body size and their locations relative to the optical imaging device. Various scales and body poses could reduce the accuracy of detection and identification to a certain extent. Detection of small targets is a common issue occurring in underwater object detection. In this work, the distribution of the target size in the marine organism dataset was analyzed by examining the object/image ratio. The object/image ratio was calculated by the object area divided by the total image size. In other words, the object/image ratio calculated the proportion of the object in the entire image. The target size distribution is presented in Figure 4. The target size was presented in significant variations. The maximal object size was approximately 20 times larger than the minimum size. Moreover, a large proportion of targets were of a small size, and their object/image ratio was less than 0.05.

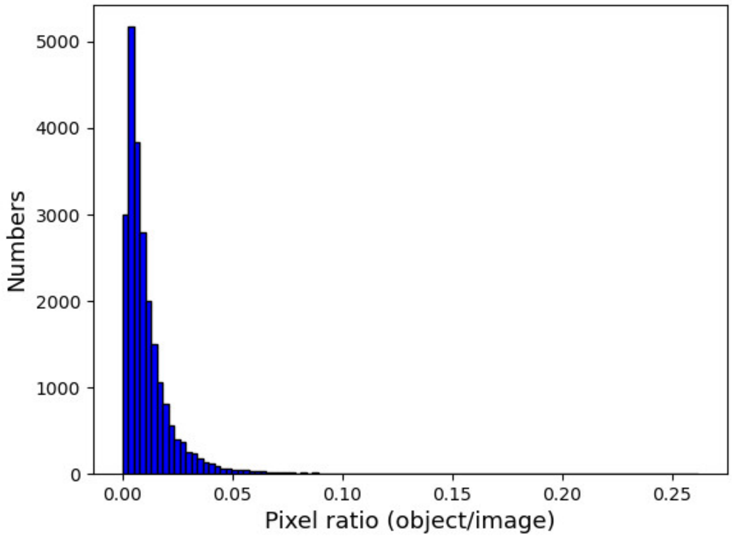

Figure 4. Target size distribution in marine organism dataset.

However, the prediction head of the original YOLOV5 was not designed to deal with such a significant variation in target size. To suppress the effect of drastic scale changes on detection accuracy, the prediction head of YOLOV5 is modified to capture objects with large scale variation. An additional prediction scale is added to the prediction head of the proposed detection model. In the proposed model, the output feature maps of four different scales are 160 × 160, 80 × 80, 40 × 40, and 20 × 20, respectively, when the input image is resized as 640 × 640.

2.5. Evaluation Metrics

The performance is evaluated from two aspects covering detection performance and time consumption performance. In the evaluation of the marine organism detection performance, AP is the integral over the precision–recall curve and represents the average precision. mAP is the average precision of all species of marine organisms. AP and mAP are formulated as follows:

$$\mathrm{AP} = \int_0^1 \mathrm{Precision} - \mathrm{Recall}(\mathrm{Recall})\mathrm{d}(\mathrm{Recall}) \tag{2}$$

$$\mathrm{mAP} = (1/\mathrm{N}) \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP}_{\mathrm{i}} \tag{3}$$

where N is the number of species of marine organisms. The precision represents the proportion of the correct number of positive samples predicted to all predicted samples. The recall is the ratio of positive samples correctly predicted to all samples. Formulas for precision and recall are as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

Among them, true positives, false positives, and false negatives, respectively, represent correctly classified positive samples, incorrectly classified negative samples, and incorrectly classified positive samples.

Otherwise, the average time consumption of marine organism identification is also considered to evaluate the performance of the proposed model, which is an important factor to evaluate the availability of real-time observations.

3. Experiments and Discussion

In order to evaluate the detection performance of marine organisms, ablation studies were implemented to verify the validity of each optimization module. In these experiments, the YOLOV5s model is selected as the original architecture to provide a performance baseline. Thereafter, an overall performance evaluation of the proposed model is discussed.

Subsequently, the detection model of the marine organisms is trained on a server under a Linux environment, which is equipped with Intel XEON Gold 5217 CPU and NVIDIA RTX TITAN GPU cards. Finally, the trained model is deployed on an embedded system platform, the Jetson Nano development kit, with an ARM A57 processor (Quad-core @ 1.43 GHz), 4 GB RAM, and Maxwell GPU (128 cores), to evaluate both the detection and real-time performance.

3.1. Performance of Image Enhancement

A multiple color model-based image enhancement method, RGHS, is selected as the image enhancement module in the proposed detection model of marine organisms by comparing the enhanced performance with other methods published in the last decade. From the enhancement results of these methods, illustrated in Figure 5, the performance of RGHS is superior under human visual judgment. After image enhancement, the invisible targets in the original images become visible. Thus, the enhanced images are re-labeled to evaluate the performance of the image enhancement module.

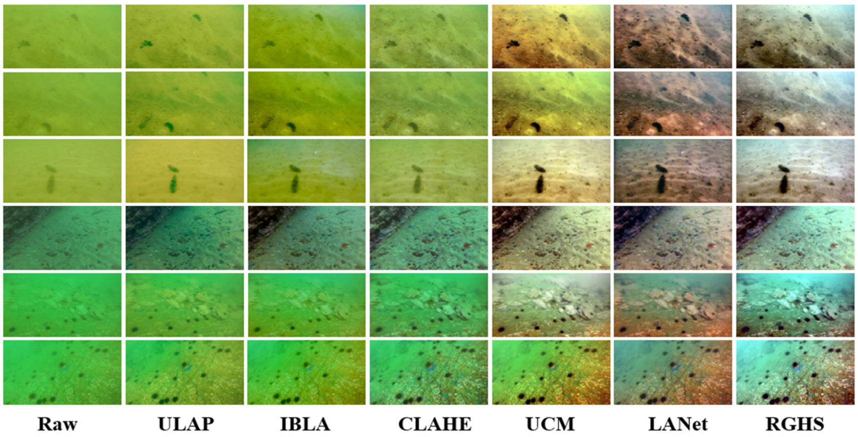

Figure 5. Experimental comparison of enhancement performance on underwater images.

Image enhancement operation as a preprocessing module is plugged into the original YOLOV5 model to evaluate its performance. The quantified detection results of the marine organisms between the original YOLOV5 model and the YOLOV5 model with the image enhancement module are listed in Table 1. Here, to represent the results more clearly, the detection models are named with the suffixes 'Raw' and 'En' to classify the model using the raw images or the enhanced images as the dataset. The ground truth is raised to 6530 from 6207, which means that there are 323 additional organisms visible after the image enhancement operation. Notably, the metric AP of all four species of the marine organisms is improved obviously after image enhancement, while the mAP for the Intersection of Union (IoU) threshold with 0.5 rises to 78.0% from 71.5%. This indicates that the image enhancement module is necessary and helpful to increase the capability for observing and detecting targets in the marine environment.

Table 1. Performance comparison of image enhancement module.

Model	AP(@0.5)				mAP(@0.5)	Ground Truth
	Urchin	Sea Cucumber	Starfish	Scallop		
YOLOV5 (Raw)	82.2%	62.6%	80.4%	60.9%	71.5%	6207
YOLOV5 (En)	91.3%	62.9%	86.2%	71.5%	78.0%	6530

3.2. Evolution of Attention Modules

Recently, attention mechanisms have achieved great success in computer vision tasks, such as object detection/recognition and segmentation [37,38]. SENet and CBAM are two state-of-the-art attention modules and have been widely applied to improve the performance of object detection [38,39]. However, these attention mechanism modules overlook the interaction between the channel dimension and spatial dimension. In this work, triplet attention (TA) is proposed to improve the feature representational ability of YOLOV5. Triple attention is a lightweight module and is able to capture the cross-dimension interaction between channel dimension and spatial dimension without increasing the computational burden.

Performance analysis of YOLOV5, YOLOV5-SE, YOLOV5-CBAM, and YOLOV5-TA is listed in Table 2. There is a 1.3% rise in the mAP, increasing from 71.5% to 72.8% after the triplet attention modules are integrated. The mAP values of YOLOV5-SE and YOLOV5-CBAM are 71.4% and 72.1%, respectively, which are both lower than the mAP of YOLOV5-TA. Especially for the species starfish and scallop, which are often indistinguishable from the background, the triplet attention outperforms the SE and CBAM attention.

Table 2. Performance comparison of attention mechanism modules.

Model	AP(@0.5)				mAP(@0.5)	Ground Truth
	Urchin	Sea Cucumber	Starfish	Scallop		
YOLOV5 (Raw)	82.2%	62.6%	80.4%	60.9%	71.5%	6207
YOLOV5-SE (Raw)	83.8%	63.7%	81.2%	56.7%	71.4%	6207
YOLOV5-CBAM (Raw)	85.4%	61.6%	81.2%	60.4%	72.1%	6207
YOLOV5-TA (Raw)	83.5%	61.9%	82.7%	63.2%	72.8%	6207

Subsequently, the attention heatmaps of these modules were visualized by implementing Grad-CAM to explain the effectiveness of triplet attention. A couple of samples for each species of the organism are provided in Figure 6. Tighter and more relevant bounds on images are captured, which indicates that the triplet attention modules provide more meaningful internal representations of the image through cross-dimensional interaction. In terms of the above results, the performance and efficiency of the triplet attention mechanism are demonstrated.

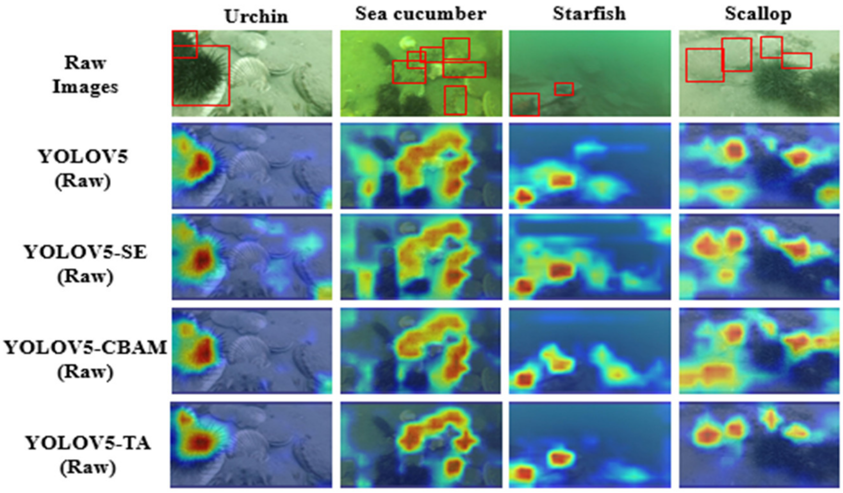

Figure 6. Visualization of Grad-CAM results for each species of marine organisms; the ground truth in the raw images is marked with red boxes.

3.3. *Performance of Prediction Head Optimization*

In this subsection, one additional prediction head is integrated into the YOLOV5 model for tiny organism detection to reduce the missed targets. Similar to YOLOV5-TA, this model with an additional prediction head is abbreviated as YOLOV5-PH. In Table 3, the AP has a minimum 1% increase in three quarters of the target species of marine organisms, but the AP of the scallop species has a slight decrease (0.8%). Nevertheless, there is an average 1.4% improvement in the metric AP after optimizing the prediction head to the original detection model YOLOV5.

Table 3. Performance comparison of prediction head optimization.

Model	AP(@0.5)				mAP(@0.5)	Ground Truth
	Urchin	Sea Cucumber	Starfish	Scallop		
YOLOV5 (Raw)	82.2%	62.6%	80.4%	60.9%	71.5%	6207
YOLOV5-PH (Raw)	85.5%	63.6%	82.6%	60.1%	72.9%	6207

Figure 7 shows the performance comparison of prediction head optimization on the detection of tiny marine organisms. More organisms with small scale in images are detected after adding one additional prediction head with a scale of 160 × 160 compared to the common prediction structure with three heads. This indicates that the prediction head optimization makes the detection model more sensitive to organisms with different scales.

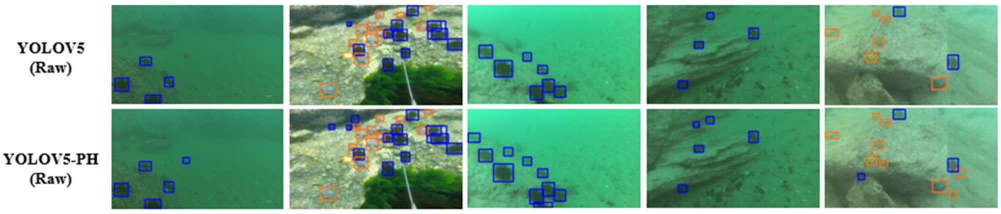

Figure 7. Performance comparison of prediction head optimization on detection of tiny marine organisms.

3.4. *Overall Identification Performance Evaluation*

To verify and validate the overall performance of the marine organism detection model, we implement other ablation experiments based on incorporating the YOLOV5 model for image enhancement by plugging the triplet attention module and optimizing the prediction head, respectively. These model variants are marked as YOLOV5 (En), YOLOV5-TA (En), and YOLOV5-PH (En), respectively. In addition, a well-known two-stage approach, a Faster RCNN model with the structures resnet50 and resnet101 as the backbone, respectively, is also selected to compare with the proposed model.

Figure 8 shows that the loss curves of all the YOLOV5 architecture-based models experience a steady decline while training these models, and eventually converge to a low constant. After 30 epochs, the loss of the proposed model nearly remains stable at a constant 0.07, similar to YOLOV5-PH (En), while the training loss of the other models maintains a slightly decreasing trend until the 80th epoch. This means that the weights of the proposed model could be trained with lower time consumption.

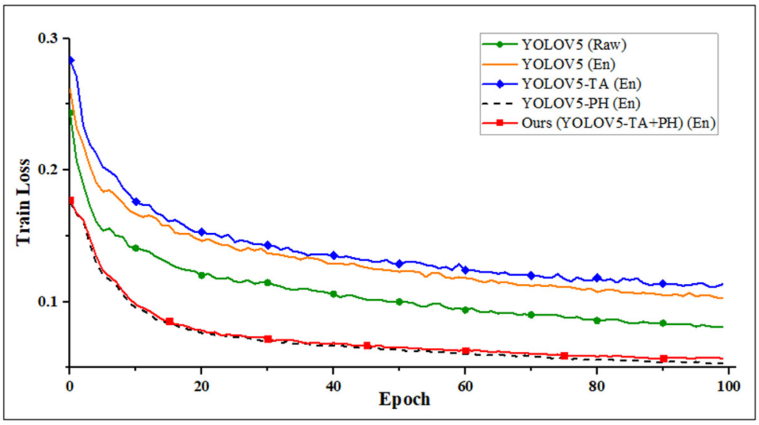

Figure 8. Loss curves of the proposed model and the other comparable models.

The indices of performance evaluation for the proposed model in this paper and the other comparison models are listed in Table 4. Values in bold denote that the related model has the best performance for the corresponding evaluation indexes. mAP(@0.5:0.95) represents the average mAP for increasing IoU threshold values from 0.5 to 0.95 by 0.05. As reflected in these results, both the mAP(@0.5) and mAP(@0.5:0.95) of the proposed model are higher than those of the other comparison models, achieving 83.1% and 42.2%, respectively. Even in each species, the detection performance of the proposed model is better than the other models. Compared to urchin and starfish, the AP values for sea cucumber and scallop are reduced to some extent, reaching only 69.8% and 79.7%, respectively. The reason is that these two species of marine organisms are similar in color to the background and often lay flat on the surface of the seafloor or even buried in the sand, which poses a huge challenge for detection.

A couple of samples of detection results are shown in Figure 9. What can be clearly seen is that each approach accurately detects more organisms after image enhancement, which proves the necessity of image enhancement in the marine organism observation and detection scenario. As the depth of the backbone network increases, Faster RCNN detects many more organisms. However, there are more misdetections, which results in a lower AP of 61.6% for the Faster RCNN with resnet101 as the backbone. The above results confirm that the proposed model is helpful to improve the detection performance of marine organisms by integrating the triplet attention module and optimizing the prediction head.

Table 4. Overall identification performance evaluation.

Model	AP(@0.5)				mAP (@0.5)	mAP (@0.5:0.95)	Detection Time	Ground Truth
	Urchin	Sea Cucumber	Starfish	Scallop				
YOLOV5 (Raw)	82.2%	62.6%	80.4%	60.9%	71.5%	35.1%	**0.248**	6207
YOLOV5 (En)	91.3%	62.9%	86.2%	71.5%	78.0%	40.7%	0.917	6530
YOLOV5-TA (En)	92.0%	65.6%	87.1%	73.1%	79.4%	41.1%	0.924	6530
YOLOV5-PH (En)	93.1%	63.9%	88.0%	74.5%	79.9%	40.6%	0.999	6530
Faster RCNN (resnet50) (En)	73.0%	31.0%	57.1%	44.1%	51.3%	20.7%	6.817	6530
Faster RCNN (resnet101) (En)	85.9%	46.0%	78.0%	36.7%	61.6%	29.8%	11.370	6530
Ours (YOLOV5-TA+PH) (En)	**93.4%**	**69.8%**	**89.3%**	**79.7%**	**83.1%**	**42.2%**	0.982	6530

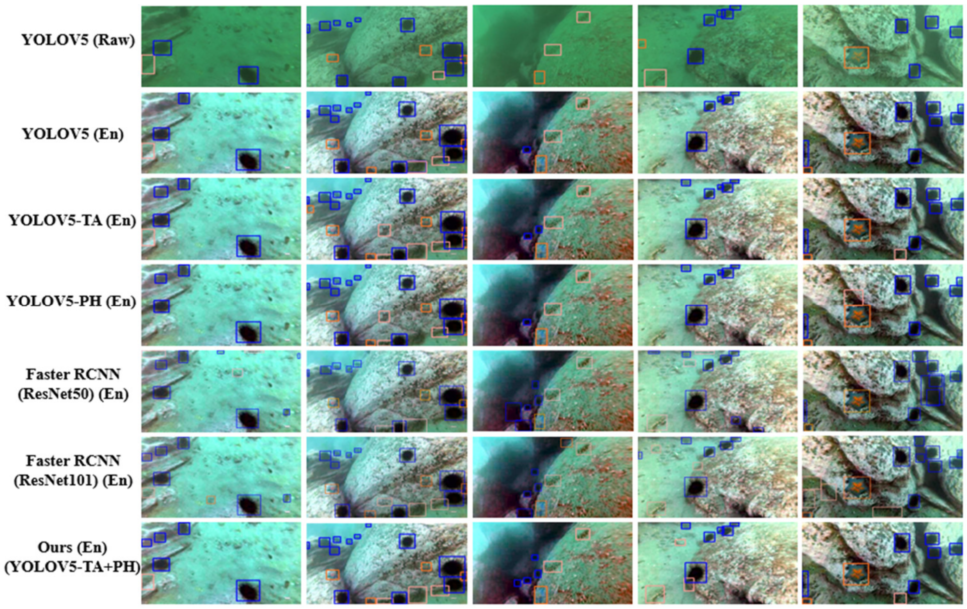

Figure 9. Marine organism identification results of the proposed model and the other comparable models.

3.5. Adaptation Performance Evaluation

To reduce the computational burden of the embedded system, the raw images are resized to 640 × 640, and the resized images as input are fed to the detection model. The time consumption for each frame is shown in Table 4. The trained detection model has a weight of 360.15 MB and requires 11.370 s to process one frame by implementing the Faster RCNN model with resnet101 as the backbone. Compared to the Faster RCNN, the weight of our proposed detection model is approximately 14.1 MB, with the time consumption of 0.982 s for processing one frame, including 0.25 s for detection and 0.73 s for image enhancement, which ensures its implementation in the embedded system to carry out the marine organism observation in real time. Notably, since we replace the complex CSP module with a lightweight triplet attention module, our proposed model consumes less time compared to the YOLOV5-PH model, with 0.999 s for one frame.

4. Conclusions

This study aims to improve the ability of automatic marine organism detection in the real marine environment, which could release workers from heavy workloads and is considered an effective tool in the management of the marine ranch. To achieve this goal, an adaptive deep neural network model is proposed based on the YOLOV5 architecture by integrating it with the image enhancement module and triplet attention mechanism modules, as well as optimizing the number of prediction heads. In these optimizations, the image enhancement module aims to improve the visual quality of images, which brings out more targets buried in noise and also extends the observation range. The purpose of the other optimizations, covering triplet attention mechanism modules and prediction head optimization, is to improve the detection performance, which makes the proposed approach more sensitive to the complex environment. The experimental results demonstrate the effectiveness of each optimization in the proposed approach to marine organism detection, and the mAP reaches 83.1%, experiencing an 11.6% arise. Therefore, the proposed approach is suitable for deployment on an embedded system due to its small volume and low time consumption.

Currently, the proposed approach is deployed on the deep learning development board Jetson Nano, the core process unit of our autonomous underwater vehicle. In further work, many experiments will be carried out with the mobile platform to evaluate its performance in the real marine ranch after intermodulation of the perception system and control system.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L. and X.B.; Formal Analysis, Y.L.; Writing—Review and Editing, Y.L. and C.X.; Funding Acquisition, Y.L.; Data Curation, X.B.; Software, X.B.; Validation, X.B.; Resources, C.X.; Investigation, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China, grant number 61821005; in part by the Liaoning Provincial Natural Science Foundation of China, grant number 2020-MS-031; in part by the State Key Laboratory of Robotics at Shenyang Institute of Automation, grant number 2021-Z10L01; and in part by the Liaoning Revitalization Talents Program, grant numbers XLYC2007035 and XLYC1902032.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. 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 Networks Learn. Syst.* **2021**, *99*, 1–15. [CrossRef]
2. Han, M.; Lyu, Z.; Qiu, T.; Xu, M. A review on intelligence dehazing and color restoration for underwater images. *IEEE Trans. Syst. Man, Cybern. Syst.* **2020**, *50*, 1820–1832. [CrossRef]
3. 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. [CrossRef]
4. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
5. Jiao, L.C.; Zhang, F.; Liu, F.; Yang, S.Y.; Li, L.L.; Feng, Z.X.; Qu, R. A Survey of Deep Learning-Based Object Detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
6. Guo, G.; Zhang, N. A survey on deep learning based face recognition. *Comput. Vis. Image Underst.* **2019**, *189*, 102805. [CrossRef]
7. Leclerc, M.; Tharmarasa, R.; Florea, M.; Boury-Brisset, A.; Kirubarajan, T.; Duclos-Hindié, N. Ship classification using deep learning techniques for maritime target tracking. In Proceedings of the 2018 21st International Conference on Information Fusion, Cambridge, UK, 10–13 July 2018; pp. 737–744.
8. Py, O.; Hong, H.; Zhongzhi, S. Plankton classification with deep convolutional neural networks. In Proceedings of the 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, Chongqing, China, 20–22 May 2016. [CrossRef]
9. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2341–2353.
10. Drews, P., Jr.; Nascimento, E.; Moraes, F.; Botelho, S.; Campos, M. Transmission estimation in underwater single images. In Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, Sydney, NSW, 2–8 December 2013; pp. 825–830.
11. Peng, Y.-T.; Cosman, P.C. Underwater Image Restoration Based on Image Blurriness and Light Absorption. *IEEE Trans. Image Process.* **2017**, *26*, 1579–1594. [CrossRef]
12. Song, W.; Wang, Y.; Huang, D.; Tjondronegoro, D. A Rapid Scene Depth Estimation Model Based on Underwater Light Attenuation Prior for Underwater Image Restoration. In *Advances in Multimedia Information Processing—PCM 2018*; Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M., Ngo, C.-W., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11164, pp. 678–688. [CrossRef]

13. Huang, D.; Wang, Y.; Song, W.; Sequeira, J.; Mavromatis, S. Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition. In Proceedings of the International Conference on Multimedia Modeling, Bangkok, Thailand, 5–7 February 2018; pp. 453–465.
14. Hou, M.; Liu, R.; Fan, X.; Luo, Z. Joint residual learning for underwater image enhancement. In Proceedings of the 2018 IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4043–4047.
15. Sun, X.; Liu, L.; Li, Q.; Dong, J.; Lima, E.; Yin, R. Deep pixel-to-pixel network for underwater image enhancement and restoration. *IET Image Process.* **2019**, *13*, 469–474. [CrossRef]
16. 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.* **2020**, *29*, 4376–4389. [CrossRef]
17. Liu, S.; Fan, H.; Lin, S.; Wang, Q.; Ding, N.; Tang, Y. Adaptive Learning Attention Network for Underwater Image Enhancement. *IEEE Robot. Autom. Lett.* **2022**, *7*, 5326–5333. [CrossRef]
18. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. Water GAN: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robot. Autom. Lett.* **2017**, *3*, 387–394.
19. Fabbri, C.; Islam, J.; Sattar, J. Enhancing Underwater Imagery Using Generative Adversarial Networks. In Proceeding of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 7159–7165.
20. Hasija, S.; Buragohain, M.J.; Indu, S. Fish Species Classification Using Graph Embedding Discriminant Analysis. In Proceedings of the 2017 International Conference on Machine Vision and Information Technology (CMVIT), Singapore, 17–19 February 2017; pp. 81–86. [CrossRef]
21. Qiao, X.; Bao, J.; Zhang, H.; Wan, F.; Li, D. fvUnderwater sea cucumber identification based on Principal Component Analysis and Support Vector Machine. *Measurement* **2019**, *133*, 444–455. [CrossRef]
22. Han, F.; Zhu, H.; Yao, J. Multi-Targets Real Time Detection from Underwater Vehicle Vision Via Deep Learning CNN Method. In Proceedings of the 29th International Ocean and Polar Engineering Conference, Honolulu, Hawaii, USA, 16 June 2019; p. 6.
23. Peng, F.; Miao, Z.; Li, F.; Li, Z. S-FPN: A shortcut feature pyramid network for sea cucumber detection in underwater images. *Expert Syst. Appl.* **2021**, *182*, 115306. [CrossRef]
24. Cao, S.; Zhao, D.; Liu, X.; Sun, Y. Real-time robust detector for underwater live crabs based on deep learning. *Comput. Electron. Agric.* **2020**, *172*, 105339. [CrossRef]
25. Li, Y.; Guo, J.; Guo, X.; Zhao, J.; Yang, Y.; Hu, Z.; Jin, W.; Tian, Y. Toward in situ zooplankton detection with a densely connected YOLOV3 model. *Appl. Ocean Res.* **2021**, *114*, 102783. [CrossRef]
26. Li, Y.; Guo, J.; Guo, X.; Hu, Z.; Tian, Y. Plankton Detection with Adversarial Learning and a Densely Connected Deep Learning Model for Class Imbalanced Distribution. *J. Mar. Sci. Eng.* **2021**, *9*, 636. [CrossRef]
27. Li, X.; Shang, M.; Qin, H.; Chen, L. Fast accurate fish detection and recognition of underwater images with fast R-CNN. In Proceedings of the OCEANS 2015—MTS/IEEE Washington, Washington, DC, USA, 19–22 October 2015; pp. 1–5. [CrossRef]
28. Li, X.; Shang, M.; Hao, J.; Yang, Z. Accelerating fish detection and recognition by sharing CNNs with objectness learning. In Proceedings of the OCEANS 2016—Shanghai, Shanghai, China, 10–13 April 2016; pp. 1–5. [CrossRef]
29. Li, X.; Tang, Y.; Gao, T. Deep but lightweight neural networks for fish detection. In Proceedings of the OCEANS 2017—Aberdeen, Aberdeen, UK, 19–22 June 2017; pp. 1–5. [CrossRef]
30. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
31. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12346, pp. 213–229. [CrossRef]
32. Mnih, V.; Heess, N.; Graves, A. Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 2204–2212.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *IEEE Trans Pattern Anal Mach Intell.* **2020**, *42*, 2011–2023. [CrossRef]
34. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19. [CrossRef]
35. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.S. BAM: Bottleneck Attention Module. *arXiv* **2018**, arXiv:1807.06514. Available online: http://arxiv.org/abs/1807.06514 (accessed on 30 May 2022).
36. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3138–3147. [CrossRef]
37. Wang, D.; He, D. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. *Comput. Electron. Agric.* **2022**, *196*, 106864. [CrossRef]
38. Qi, J.; Liu, X.; Liu, K.; Xu, F.; Guo, H.; Tian, X.; Li, M.; Bao, Z.; Li, Y. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease. *Comput. Electron. Agric.* **2022**, *194*, 106780. [CrossRef]
39. QWang, Q.; Cheng, M.; Huang, S.; Cai, Z.; Zhang, J.; Yuan, H. A deep learning approach incorporating YOLO v5 and attention mechanisms for field real-time detection of the invasive weed Solanum rostratum Dunal seedlings. *Comput. Electron. Agric.* **2022**, *199*, 107194. [CrossRef]