Detection Method of Marine Biological Objects Based on Image Enhancement and Improved YOLOv5S

Peng Li 1, Yibing Fan 1, Zhengyang Cai 1, Zhiyu Lyu 2 and Weijie Ren 1,*

1 College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
2 School of Automation Engineering, Northeast Electric Power University, Jilin 132012, China
* Correspondence: renweijie@hrbeu.edu.cn; Tel.: +86-0451-8258658

Abstract: Marine biological object detection is of great significance for the exploration and protection of underwater resources. There have been some achievements in visual inspection for specific objects based on machine learning. However, owing to the complex imaging environment, some problems, such as low accuracy and poor real-time performance, have appeared in these object detection methods. To solve these problems, this paper proposes a detection method of marine biological objects based on image enhancement and YOLOv5S. Contrast-limited adaptive histogram equalization is taken to solve the problems of underwater image distortion and blur, and we put forward an improved YOLOv5S to improve accuracy and real-time performance of object detection. Compared with YOLOv5S, coordinate attention and adaptive spatial feature fusion are added in the improved YOLOv5S, which can accurately locate the target of interest and fully fuse the features of different scales. In addition, soft non-maximum suppression is adopted to replace non-maximum suppression for the improvement of the detection ability for overlapping objects. The experimental results show that the contrast-limited adaptive histogram equalization algorithm can effectively improve the underwater image quality and the detection accuracy. Compared with the original model (YOLOv5S), the proposed algorithm has a higher detection accuracy. The detection accuracy AP50 reaches 94.9% and the detection speed is 82 frames per second; therefore, the real-time performance can be said to reach a high level.

Keywords: marine biological object; object detection; image enhancement; deep learning; improved YOLOv5S

1. Introduction

Oceans account for 70 percent of the earth’s surface; they are rich in natural resources and play an important part in the existence and development of mankind [1]. The exploration and development of marine resources have recently become the focus of many countries along with the continuous development of marine information technology and the increasing demand for human life [2]. Marine biological detection technology is regarded as a critical link for the development of marine resources. It is of great significance for rescue missions, artificial structure inspection, ecological monitoring and marine life tracking [3]. In practical applications, the complexity of the underwater environment increases the difficulty of marine biological detection, so the research of marine object detection methods is particularly important.

The challenge of the underwater environment poses two difficulties: (i) wavelength-selective attenuation reduces and distorts the contrast between objects and backgrounds [4], (ii) polarization and atomization effects may cause false positives [5]. Most underwater object detection has resulted in a significantly poor performance due to these difficulties above [6]. Usually, the key technology of underwater object detection based on optical images includes image preprocessing and object detection. Underwater image preprocessing methods are divided into underwater image enhancement methods and underwater
image restoration methods [7]. The underwater image enhancement method only uses a computer graphics method to improve image clarity, and there is no need for an implementation process of an algorithm to consider the specific physical imaging process. The underwater image restoration methods aim to inversely solve the underwater imaging model to obtain the restored underwater image [8]. The algorithm process depends on the underwater image degradation model. Marine biometric detection algorithms are generally divided into two categories: traditional object detection algorithms and convolution neural network-based object detection algorithms. The process of traditional object detection consists of three parts: image preprocessing, image feature extraction and classification. In practical application, the complexity of the underwater environment increases the difficulty of marine biological detection [9].

Abstract deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [10]. With the rapid development of deep learning, people have introduced more powerful tools to solve the problems in traditional architecture. Therefore, it is widely used in the field of vision, such as image and video semantic segmentation [11], object detection [12], disease prediction [13], etc. Multiple studies have gradually introduced deep learning to underwater object classification and detection along with the mass progress of computer vision technology, which provides a new direction for marine biological detection. Object detection algorithms based on deep learning are the mainstream detection algorithms at present [14].

Regardless of traditional methods or algorithms based on deep learning, the detection effect is usually poor for objects with poor image quality, blurred backgrounds, and complex backgrounds. The detection of underwater biological object requires both high accuracy and high real-time. At present, the common detection algorithms struggle to balance the relationship between accuracy and speed. In this paper, relevant object detection models are selected by reading papers, and a large number of simulation experiments are carried out on the selected models. YOLOv5S is discovered to have certain advantages in accuracy and speed. Thus, YOLOv5S is chosen as the main method for marine biological object detection.

In order to make YOLOv5S have a better detection effect, this paper first improves the underwater image quality, background blurring and background complexity. By comparing with other traditional unsupervised preprocessing algorithms, the contrast-limited adaptive histogram (CLAHE) is selected to apply to ocean image enhancement to improve image brightness, correct blue-green color, and enhance image contrast skew. At the same time, in order to make YOLOv5S have better model capability, a coordinated attention (CA) mechanism and an adaptive spatial feature fusion (ASFF) were added to YOLOv5S, which fully fused the features of different scales, thus improving the performance of detection models. Soft non-maximum suppression (Soft NMS) is used instead of non-maximum suppression (NMS) to enhance the detection ability of overlapping objects.

The rest of this document is organized as follows: Section 2 briefly introduces the related work. Section 3 describes the image preprocessing methods in detail. Section 4 introduces the YOLOv5S object detection model and improvement strategies. Section 5 carries out the experimental research, and the analysis of the experimental results on the test dataset are presented. Then, Section 6 concludes the whole work. Finally, a new method based on image enhancement and improved YOLOv5S applied to marine biology detection has been obtained to achieve detection with excellent accuracy and outstanding real-time performance.

2. Related Work

Due to the special underwater imaging environment, underwater image processing is of great significance for marine biological object detection. As is well-known, most underwater images have blur and color distortion, which are caused by light scattering and absorption. When light propagates in water, it scatters when it meets suspended particles,
and forward scattering blurs the details of the image. Backscattering causes foggy blur and reduces the image contrast [15]. In addition, owing to light being selectively absorbed by water, red light with longer wavelength attenuates the fastest, and shorter-wavelength blue light propagates the farthest, so the underwater image is often blue-green. These two characteristics of underwater images have a great impact on object detection; therefore, enhancing and restoring the original underwater image is essential. The conventional unsupervised image processing methods can be generally divided into two categories: unphysical model enhancement methods and physical model-based restoration methods [16]. There is no need for underwater image enhancement methods to concern the physical imaging process and not require obtaining the optical parameters of the water body in advance [17]. These methods directly use image processing methods to improve image clarity, mainly including histogram-based enhancement algorithms, Retinex-based enhancement algorithms, filtering and information processing-based enhancement algorithms, image fusion-based enhancement algorithms and convolutional neural network (CNN) based enhancement algorithms [18]. Pizer et al. [19] proposed adaptive histogram equalization (AHE). It is AHE that redistributes the image brightness through the calculation of the image local histogram to make the image contrast changed, so AHE is an algorithm that has the ability to improve the image local contrast. For better performance of the AHE algorithm, Zuiderveld [20] proposed contrast-limited adaptive histogram equalization. Jobson et al. proposed a multi-scale Retinex enhancement algorithm [21] and a multi-scale Retinex with color restoration (MSRCR) enhancement algorithm [22]. A priori knowledge is considered in underwater image restoration and the dark channel prior algorithm proposed by H. Kaiming et al. [23]. In addition to the dark channel prior method, Biancoet et al. [24] proposed a new prior method, which can calculate the transmittance map by using the difference of attenuation of each color channel. Wang et al. [25] proposed the adaptive attenuation curve prior (AACP) method, and Shin et al. [26] proposed a general convolution structure, from which the transmission image and background light of underwater image can be learned at the same time to realize underwater image restoration. In order to further improve the restoration effect, Hu et al. [27] proposed a parallel convolution neural network to process the underwater image. The network includes two parallel branches to estimate the transmission image and background light. Finally, the background light and transmission image are input into the underwater imaging model to restore the underwater image.

The development of marine biological detection is synchronized with general object detection. Traditional underwater object detection methods are mostly based on SIFT [28], HOG [29] and shape features [30]. The accuracy of object detection is greatly improved by object detection methods based on CNNs along with the development of deep learning. Algorithms based on CNN are mainly divided into two major classes, two-stage target detection and one-stage target detection[31]. Two-stage target detection algorithms divide the detection problem into two phases: in the first phase, the algorithm generates candidate regions, while in the second phase, the algorithm classifies and refines candidate regions. Such algorithms include R-CNN [32], Fast R-CNN [33], Faster R-CNN [34] and Cascade R-CNN [35]. Two-stage detection algorithms can achieve the most accurate detection effects. Although they are accurate, the prediction speed is very slow, and it has high hardware requirements, which is difficult to meet in terms of real-time requirements. One-stage object detection algorithms detect and classify simultaneously and directly outputs the classification probability and location coordinate values of the targets. Typical algorithm models, such as the YOLO series [36–39], SSD [40], RetinaNet [41], FreeAnchor [42], FSAF [43] and FCOS [44], may be slightly less accurate than the two-stage model, but have higher real-time predictive power than the two-stage model. Therefore, the greatest advantage of this kind of network model is its fast speed, although the accuracy is slightly lower than that of the two-stage detector. Yang et al. [45] used two representative target detection algorithms (YOLOv3 and Fast R-CNN) for underwater object detection, which can represent two major kinds of object detection. Song et al. [46] proposed an automatic real-time
underwater object detection method based on an improved convolutional neural network. Salman et al. [47] proposed an automatic method to detect and locate fish instances in unconstrained underwater video by using a depth R-CNN network. Cao et al. [48] proposed a real-time and lightweight multi-scale object detector to detect underwater live crabs. Ammar Mahmood et al. [49] developed a novel method for capturing feature information by adding a convolutional block attention module to the YOLOv5 backbone network with a self-adaptive global histogram stretching algorithm designed to eliminate degradation problems. Ref. [50] proposed an improved YOLOv3 real-time detection model called YOLOv3-DPFIN to achieve accurate detection of multi-level sonar objects. Sun et al. [51] proposed a method to detect and remove crosstalk noise in forward scanning sonar images using CNN. Hu et al. [52] proposed a shrimp net based on CNN to realize shrimp recognition, but there are few CNN layers in shrimp net, and the ability to extract deep features is insufficient, so it is difficult for it to detect shrimp of different sizes. Because underwater biological detection is mostly small object detection, and the biological size is very small, it is necessary to deeply extract and enhance the fusion of features. Ref. [53] proposed an improved YOLO-SC (YOLO-Submarine Cable) detection method based on the YOLO-V3 algorithm, built a testing environment for submarine cables, and created a submarine cable image dataset.

3. Image Enhancement Algorithm

CLAHE is an image enhancement algorithm derived from adaptive histogram equalization (AHE). The image contrast is changed by AHE algorithm through the calculation of the image local histogram and the redistribution of the brightness. The algorithm can not only obtain more image details but also improve the local contrast of the image. The CLAHE mechanism is as follows.

3.1. Contrast Ratio

Contrast refers to the measurement of different brightness levels between the brightest white and the darkest black in the light and dark areas of an image. The image contrast is measured by the grayscale range, which can be obtained by observing the grayscale histogram. The larger the grayscale range, the higher the contrast. The most commonly used quantitative measurement method is Michelson contrast. Michelson contrast is defined by Formula (1).

\[
C = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}
\]  

In Formula (1), \(I_{\text{max}}\) represents the lightest brightness, and \(I_{\text{min}}\) represents the darkest brightness.

3.2. Histogram Equalization

Histogram equalization is often used to increase the global contrast of many images, especially when the contrast of useful data in an image is fairly close. In this way, the brightness can be better distributed on the histogram. This can be used to enhance local contrast without affecting the overall contrast, which is achieved by effectively extending the commonly used brightness.

Suppose the input image is \(I\), the height is \(H\), the width is \(W\), the \(\text{Hist}_I\) represents the gray histogram of \(I\), and the \(\text{Hist}_I(k)\) represents the number of pixels with a gray value equal to \(k\), where \(k \in [0, 255]\). The global histogram equalization operation maps image \(I\) to make the gray histogram \(\text{Hist}_O\) of output image \(O\) equalized, so that the number of pixels in each gray level is approximately equal, as per Formula (2).

\[
\text{Hist}_O(k) \approx \frac{H \times W}{256}k \in [0, 255]
\]
Then, for any gray level $p (0 \leq p \leq 255)$, $q (0 \leq q \leq 255)$ can always be found, making
\[
p \sum_{k=0}^{p} \text{Hist}_I (k) = q \sum_{k=0}^{q} \text{Hist}_O (k).
\]
\[
p \sum_{k=0}^{p} \text{Hist}_I (k) \text{ and } q \sum_{k=0}^{q} \text{Hist}_O (k)
\]
called the cumulative histograms of $I$ and $O$, respectively. Additionally, because of $\text{Hist}_O (k) \approx \frac{H \times W}{256}$, Formula (3) is obtained.

\[
p \sum_{k=0}^{p} \text{Hist}_I (k) \approx (q + 1) \times \frac{H \times W}{256} \tag{3}
\]

Simplify Formula (3) to Formula (4).

\[
q \approx \frac{\sum_{k=0}^{p} \text{Hist}_I (k)}{H \times W} \times 256 - 1 \tag{4}
\]

The above formula gives a mapping from an input pixel with a gray level of $p$ to an output pixel with a gray level of $q$, and thus, Formula (5) can be obtained.

\[
O(r, c) = \frac{I(r, c)}{H \times W} \times 256 - 1 \tag{5}
\]

where $I(r, c)$ is the gray value of row $r$ and column $c$ of input image $I$, and $O(r, c)$ is the gray value of the output at the corresponding position of output image $O$, where $0 \leq r < H$, $0 \leq c < W$.

3.3. Local Processing

From the above, histogram equalization is the global processing of images. For a particular image data, there are many obvious noise points using the histogram equalization algorithm, even weakening the details of the image, making the image appear paradoxically worse.

The traditional histogram equalization is for the whole image, but in fact, it is unreasonable to directly equalize the whole image. The global histogram equalization is not effective when there are some differences among the parts of the image. For a particular image data, there are many obvious noise points using the histogram equalization algorithm, even weakening the details of the image, making the image appear paradoxically worse. Therefore, an adaptive histogram equalization algorithm (AHE) is generated. It follows from this that for the grayscale values $I(r, c)$ of the pixels in column $r$ of row $r$, equalizing them with histograms takes the mapping of Formula (5). For the adaptive histogram equalization algorithm, on the basis of the conventional histogram equalization algorithm, the images are divided into blocks and processed separately, and no small piece of the images is statistically a distribution function that is exclusive to each other.

3.4. Contrast Limit

AHE does not make a transitional treatment of the block to the block’s edge and suffers from overmagnifying the image. Without noise, the gray histogram of each small block is limited to a small grayscale. However, if there is noise, the noise will be amplified after histogram equalization of each segmented block. To avoid noise, a new method called “contrast-limited adaptive histogram equalization” is proposed based on adaptive histogram equalization. If the histogram exceeds the preset “Limit Contrast”, it will be clipped, and the clipped parts will be evenly distributed to other parts; thus, the histogram will be reconstructed.

4. The Improved YOLOv5S Object Detection Model

YOLOv5 is a one-stage object detection model based on convolution neural networks. There are four different models, YOLOv5S, YOLOv5M, YOLOv5L and YOLOv5X. From the
perspective of model size and real-time, YOLOv5S is used as the detection network in this paper. At the same time, to make YOLOv5S have better model capability, a coordinated attention (CA) mechanism and an adaptive spatial feature fusion (ASFF) were added to YOLOv5S, which fully fused the features of different scales, thus improving the performance of detection models. Soft non-maximum suppression (soft NMS) is used instead of non-maximum suppression (NMS) to enhance the detection ability of overlapping objects.

4.1. YOLOv5S

YOLOv5S is the model with the smallest depth and the smallest width of characteristic graphs in the YOLOv5 series. The model selected in this paper is YOLOv5S, which has the smallest size. It can meet our requirements with precision and is much faster than other models. The YOLOv5S model structure is shown in Figure 1.

From Figure 1, as can be seen, the YOLOv5 model mainly consists of three parts: BACKBONE, PAFPN and HEAD. A network of classifiers with outstanding performance is generally used as the BACKBONE to extract some common feature representations. YOLOv5 retains the CSPDarknet53 structure and adds Focus and SPP structures to make up the skeleton in comparison with the previous YOLO series. The neck network of YOLOv5 still uses the FPN+PAN structure, named PAFPN, and uses the CSP-net structure for reference to enhance the ability of network feature fusion. It is the head that complete the output of the object detection result. Different detection algorithms have different numbers of output dimensions of the model, but usually include classification and regression parts. Furthermore, YOLOv5 takes GIOU-Loss to replace the Smooth L1 Loss function for the further improvement of detection accuracy.

4.2. Improvement Strategy

In this paper, three improvement measures are proposed for YOLOv5S:

1. ASFF is added before the detection layer to achieve the cross-layer fusion of the feature layer, make full use of the features of different scales, and improve the model detection effect.

Many models use FPN to capture multilevel features to take full advantage of the semantic information of high-level features and fine-grained features of low-level features, but either YOLOv3 or RetinaNet uses direct connections or additions, which do not take full advantage of features of different scales. ASFF can adaptively learn the spatial weights
of feature map fusion at different scales and has a better feature fusion effect. In this paper, ASFF is added after the PAFPN layer for further feature fusion. The ASFF structure can be represented in Figure 2.

The features of each layer can be fused according to Formula (6):

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{l1} + \beta_{ij}^l \cdot x_{ij}^{l2} + \gamma_{ij}^l \cdot x_{ij}^{l3}$$

(6)

where $x_{ij}$ stands for input characteristics. For example, $x_{ij}^{l1}$ represents the features of the location $(i,j)$ of the layer $l$. $\alpha_{ij}^l$, $\beta_{ij}^l$, $\gamma_{ij}^l$ are the weights for the corresponding positions of the signature map, and $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$. These three parameters can be adaptively learned by the network. $y_{ij}$ represents the characteristics of the output, and $y_{ij}^l$ represents the characteristics of the location of the output to the $l_{th}$ ASFF layer.

Figure 2. Structure diagram of ASFF.

(2) Coordinate attention is introduced after the PAFPN layer. The content that needs more attention can be picked out by using an attention mechanism, which has extensive use in deep neural networks. Compared with previous attention methods on lightweight networks, coordinate attention has the following advantages. First, it captures both direction-aware and location-aware information while extracting cross-channel information, which makes the model locate and identify objects of interest more accurately. Second, coordinate attention is flexible and lightweight and can be easily plugged into classical modules, bringing huge benefits to downstream tasks. The precise location information is used to encode the relationship between channel and long-term correlations by CA. Similar to the SE module, there are two steps in the process of CA, with coordinate information embedding as the first step and coordinating attention generation as the second step. Its specific structure is described in detail in Figure 3.

It is the CA module that decomposes the solution of channel attention into two parallel 1D feature coding processes; thus, the generated attention map by CA effectively integrates the spatial coordinate information. Specifically, each channel of the input $X$ is encoded respectively along horizontal and vertical coordinates through pooled kernels of dimensions $(H, 1)$ and $(1, W)$. Therefore, the output for channel $c_h$ with a height of $h$ is described below.

$$z^h_c(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i)$$

(7)

Similarly, the output for the c-th channel with a width of $W$ is described below;

$$z^w_c(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w)$$

(8)

These two transformations fuse features along two spatial directions, returning two direction-aware attention maps. Then, specific direction information is embedded into two feature maps, and the feature map is encoded into two attention maps, each capturing long-range dependence of the input feature map along one spatial direction. Therefore, the
representation of the feature map is enhanced by the multiplication between input feature map and attention maps which save the location information.

Figure 3. CA module architecture.

(3) Soft-NMS is used to improve the detection capability of overlapping objects. The specific steps of the algorithm can be shown in Algorithm 1. Non-maximum suppression (NMS) plays a significant role in the process of the object detection. When object detection is in progress, the object detection box with high confidence is extracted, while the false detection box with low confidence is suppressed. Generally speaking, when the parsing model is output to the object frame, there will be many object frames, depending on the number of anchors. Many of the duplicate boxes are located at the same object, and NMS is used to remove these duplicate boxes and obtain the real object boxes. The NMS algorithm first sorts the recommendation boxes by score from high to low; then the detection box $M$ with the highest score would be selected, and the other boxes that overlap significantly with the highest score recommendation box $M$ would be suppressed. The remaining check boxes will be recursively manipulated by this procedure. According to the principle of the algorithm, if a real existence object is within the preset overlap threshold, it may not be detected. That is, when the two object boxes are close together, the boxes with lower scores are deleted because they overlap too much.

Forcing the scores of neighboring low-confidence detection boxes to zero is the biggest problem of the NMS algorithm. If the same kind of object is dense and there is occlusion, it is easy to miss the detection. Soft-NMS learned the lessons of NMS. In specific algorithm implementation, instead of directly deleting the bounding box where the IOU is higher than the setting threshold, it reduces the confidence score of the box. It can improve the detection ability of the model for overlapping occluded objects without increasing the amount of calculations. The improved YOLOv5S structure is shown in Figure 4:
Algorithm 1: Soft non-maximum suppression.

Input: \( B = b_1, b_2, b_3, ..., b_n \);
\( S = s_1, s_2, s_3, ..., s_n ; N_t \);
\( B \) is the list of initial detection boxes;
\( S \) contains corresponding detection boxes;
\( N_t \) is the NMS threshold.

Output: \( D, S \)

while \( B \neq \text{empty} \) do
  \( n_i \leftarrow \arg\max S \);
  \( M \leftarrow b_{m_i} \);
  \( D \leftarrow D \cup M \);
  \( B \leftarrow B - M \);
  foreach \( b_i \) in \( B \) do
    if \( \text{iou}(M, b_i) \) then
      \( B \leftarrow B - b_i ; \)
      \( S \leftarrow S - s_i ; \)
    end
    \( s_i \leftarrow s_i f(\text{iou}(M, b_i)) \)
  end
end

Figure 4. The Improved YOLOv5S structure diagram. Compared with the original model, it adds CA in the feature extraction section and ASFF in the detection header section.

5. Experimental Research and Result Analysis

5.1. Dataset Establishment

(A) Dataset. This paper uses the underwater biological image dataset, which was officially supplied by the China Underwater Vehicle (Zhanjiang) Competition in 2020. The dataset contains 5534 images of holothurian, echinus, starfish and scallops, as shown in Figure 5.
Figure 5. Dataset images of marine life (holothurian, echinus, scallops and starfish). It can be seen that, unlike normal images, underwater images have chromatic aberration and low contrast, resulting in poor image quality.

(B) Dataset image enhancement. Because many images in the original dataset have low contrast and poor color distortion quality, some images with poor quality are selected for enhancement. The CLAHE underwater image enhancement algorithm is used to enhance the dataset. The enhanced image is compared with the original image as shown in Figure 6. It can be seen that the RGB three-channel histogram of the enhanced image is stretched, and the image contrast and image quality is improved. The enhanced image and the original image constitute the dataset of this paper, totaling 8938 images.

Figure 6. Comparison between original (top) and enhanced (bottom). It can be seen that the grayscale histogram of the processed image has a larger grayscale range.

5.2. Model Training

Transfer learning was used as the training method. The model was pretrained on the standard coco dataset, and then 100 epochs were trained using the sample dataset enhanced in the paper. In addition to the image enhancement technology mentioned above, cutout, random cropping, mosaic, flipping, rotating were also used in the preprocessing of training. The optimizer was SGD, the initial model learning rate was set to 0.001 and the input image shape was set to 640 × 640. The batchsize was set to 16, and the learning rate decreased to 10% at 75 and 90 epochs. The dataset was divided into train val:test = 9:1, train:val = 9:1, with 7241 images in the training set, 804 images in the cross-validation set, and 893 images in the test set. The test set was independent from the validation set, and test images were not used during training.

5.3. Experimental Results
5.3.1. Validation of Image Enhancement

The image enhancement algorithms of CLAHE, Retinex, ACE and DCP image restoration methods based on physical models were used in this paper. The image processing effect is shown in Figure 7.
From Figure 7, it can be easily seen that the image contrast enhancement after ACE and CLAHE processing is better than other methods, but some of the image color deviation problems after ACE processing still exist; the image contrast after DCP processing is reduced and the bluish-green deviation is more serious. After RGHS treatment, the brightness of the image is improved, but some of the images are distorted and the color is red.

Different image processing algorithms that enhance the image for model training and detection are used to verify the impact of different image enhancement algorithms on model detection results. The enhanced images of each algorithm are trained on the yolov5s model and tested on the test set. The experimental results are shown in Table 1. From the experimental results, it can be known that, compared with the original image, the detection accuracy of the model is improved after CLAHE, RGHS, ACE and DCP image preprocessing, especially for holothurian. The CLAHE algorithm $AP^{50}$ was improved by 3.6%, the ACE algorithm $AP^{50}$ was improved by 2.5%, the RGHS algorithm $AP^{50}$ was improved by 2.3%, the DCP algorithm $AP^{50}$ was improved by 1.9%, and the CLAHE algorithm achieved the best detection results. Compared with the original image, $AP^{50}$ was improved by 2.6% and $AP$ was increased by 2.5%, which fully demonstrates the effectiveness of CLAHE algorithm for underwater image enhancement.
Table 1. Comparison of image processing algorithm experimental results.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP50 (%)</th>
<th>AP (%)</th>
<th>AP50 (%)</th>
<th>AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMAGE</td>
<td>83.8</td>
<td>93.0</td>
<td>87.5</td>
<td>92.3</td>
</tr>
<tr>
<td>CLAHE</td>
<td>87.4</td>
<td>93.0</td>
<td>90.1</td>
<td>93.2</td>
</tr>
<tr>
<td>DCP</td>
<td>85.7</td>
<td>93.2</td>
<td>89.2</td>
<td>92.0</td>
</tr>
<tr>
<td>ACE</td>
<td>86.3</td>
<td>92.7</td>
<td>89.7</td>
<td>93.7</td>
</tr>
<tr>
<td>RGHS</td>
<td>86.1</td>
<td>92.9</td>
<td>88.9</td>
<td>93.0</td>
</tr>
</tbody>
</table>

5.3.2. Validation of Improved Algorithm

In this section, the influence of the improvement strategies introduced to the model is verified with the value of detection AP. The experimental results are shown in Table 2. The detection accuracy of the model improved by the three methods is significantly higher than that of the original YOLOv5S. After adding ASFF, AP50 increased by 1.9% and AP increased by 5.6%. After adding Soft-NMS, AP50 increased by 0.8% and AP increased by 1.4%. Finally, after adding the CA model, AP50 increased by 1.3% and AP increased by 2.1%. Combining the above three improved methods, compared with YOLOv5S, the AP50 of the improved YOLOv5S increased by 4%, and the AP increased by 9.1%.

Table 2. Comparison of experimental results of improved algorithms.

<table>
<thead>
<tr>
<th>Model</th>
<th>ASFF</th>
<th>Soft-NMS</th>
<th>CA</th>
<th>AP50(%)</th>
<th>AP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv5S</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>93.6</td>
<td>60.7</td>
</tr>
</tbody>
</table>

The model detection P-R diagram of ASFF, Soft-NMS and CA added in sequence is shown in Figure 8. As can be seen from the figure, the detection accuracy of the three detection methods for holothurians, starfish and scallops has been significantly improved, especially for holothurians. The ASFF remarkably improved the AP50, and the AP50 was improved by 4.3%. The other two methods for improving also enhanced their detection ability. The three enhancement methods finally improved the detection accuracy of the improved YOLOv5S for holothurians (AP50) by 7.2%, scallops (AP50) by 5.9% and starfish (AP50) by 2.6%. Finally, the detection accuracy of holothurians, echinus, scallops and starfish is 94.6%, 93.0%, 96.0% and 95.8%, respectively.
5.3.3. Comparison of Detection Effects of Different Models

In this paper, two-stage architectures (Faster R-CNN, Cascade R-CNN) with backbone resnet50, single-stage architectures (RetinaNet, FreeAnchor, FSAF, FCOS) with backbone resnet50, YOLO series and the improved YOLOv5S are selected for comparison to evaluate the detection performance of the improved YOLOv5S. The experimental results on test datasets are shown in Table 3. The results of Table 3 show that the YOLO series has obvious
advantages in marine biological detection compared with other models. Whether it is a two-stage architecture or a one-stage architecture, its precision has no advantage over the YOLO series. However, in terms of speed, the YOLO series is much faster than other models. In particular, the improved yolov5s not only surpasses other one stage two stage models in \( AP_{50} \), but also has 3–5 times the speed. As for the YOLO series, the \( AP_{50} \) of the improved YOLOv5S is 94.9%, and the AP is 62.8%, which is 4% and 9.1% higher than that of YOLOv5S before the improvement, while the detection frames per second is only slightly lower than before the improvement. The detection accuracy of the improved YOLOv5S is equivalent to YOLOv3; the model volume is only 7% of that of YOLOv3, and the detection speed is 4 times that of YOLOv3. YOLOv4 performs poorly on the data set in this paper. Compared with YOLOv4, the improved YOLOv5S has great advantages in accuracy and speed. The model volume is only 8% of YOLOv4. Although the detection accuracy of YOLOv5M is slightly better than that of the improved YOLOv5S, the volume of the improved YOLOv5S model is only half of YOLOv5M, and the speed is faster than YOLOv5M. Overall, the improved YOLOv5S has achieved good detection results; in particular, the AP has increased by 9.1% compared with before improvement. That is because adding CA makes the model extract more important and interesting features, and adding ASFF causes the cross-layers of different scale feature maps to fuse so that the classification network can obtain more informative feature maps so as to make the detection frame of the model more accurate. Most of the datasets in this paper are dense small objects, and the overlapping occlusion is serious. The addition of Soft-NMS improves the detection ability of the model for overlapping and occluded objects, which is also an important reason for the sharp increase of model AP.

Figure 9 shows the comparison of detection effects of different models. From the diagram, it can be seen that the YOLOv3 model has some false and missed cases. The starfish has been misdetected and the sea cucumber has been missed in Figure 9b. YOLOv5S has missed inspection, and the holothurian is missed in Figure 9b. YOLOv4, YOLOv5M and the improved YOLOv5S can achieve complete detection of objects, but the improved YOLOv5S box regression is more accurate and has higher confidence.

<table>
<thead>
<tr>
<th>Model</th>
<th>AP50(%)</th>
<th>AP(%)</th>
<th>Size</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>90.8</td>
<td>61.7</td>
<td>41.2M</td>
<td>23</td>
</tr>
<tr>
<td>Cascade R-CNN</td>
<td>90.4</td>
<td>63.8</td>
<td>68.4M</td>
<td>18</td>
</tr>
<tr>
<td>FreeAnchor</td>
<td>91.8</td>
<td>63.1</td>
<td>36.1M</td>
<td>26</td>
</tr>
<tr>
<td>RetinaNet</td>
<td>89.9</td>
<td>58.1</td>
<td>36.17M</td>
<td>26</td>
</tr>
<tr>
<td>FCOS</td>
<td>85.8</td>
<td>48.0</td>
<td>31.8M</td>
<td>17</td>
</tr>
<tr>
<td>FSAF</td>
<td>89.6</td>
<td>57.4</td>
<td>36.2M</td>
<td>26</td>
</tr>
<tr>
<td>YOLOv3</td>
<td>93.8</td>
<td>63.0</td>
<td>283.4M</td>
<td>34</td>
</tr>
<tr>
<td>YOLOv4</td>
<td>92.7</td>
<td>59.2</td>
<td>256.3M</td>
<td>47</td>
</tr>
<tr>
<td>YOLOv5S</td>
<td>90.9</td>
<td>53.7</td>
<td>14.2M</td>
<td>83</td>
</tr>
<tr>
<td>ImprovedYOLo5S</td>
<td>94.9</td>
<td>62.8</td>
<td>20.4M</td>
<td>82</td>
</tr>
<tr>
<td>YOLOv5M</td>
<td>94.5</td>
<td>65.5</td>
<td>42.5M</td>
<td>59</td>
</tr>
</tbody>
</table>

Table 3. Comparison of experimental results of improved algorithms on test datasets.
Figure 9. Comparison of test results of various models. (a–c) are three images with obvious contrast effect.

6. Conclusions

To realize the rapid detection of underwater biological images with low definition, low contrast and low quality, an improved model based on YOLOv5S is proposed for the effective real-time detection of underwater objects. In this paper, the CLAHE algorithm is used to enhance the underwater dataset. Then, in the model, the ASFF is used to make feature maps more fully fused, CA attention mechanism is introduced, and Soft-NMS is used to enhance the model’s ability to detect overlapping objects. Finally, the comparative experiment of the improved YOLOv5S and other models indicates that the detection accuracy of the improved YOLOv5S network has been greatly increased compared with the original YOLOv5S. The $AP$ reached 62.8%, representing an increase of 9.1%, and $AP^{50}$ reached 94.9%, representing an increase of 4%. The model volume is only 20.4 megabytes, the detection speed has great advantages compared with other models, and the FPS reaches 82. The improved YOLOv5S model in this paper has strong expansibility, can be applied to other marine biological detection and identification, plays a good role in promoting the identification and object detection of marine biological resources, and lays a foundation for the next deployment to mobile devices.
Author Contributions: Conceptualization, P.L. and W.R.; methodology, P.L., Y.F. and Z.C.; software, Y.F. and Z.C.; writing—original draft preparation, P.L. and Y.F.; writing—review and editing, Z.L. and W.R.; funding acquisition, Z.L. and W.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Nature Science Foundation of China (51879060), the Natural Science Foundation of Heilongjiang Province (LH2021F017), the Fundamental Research Funds for the Central Universities (3072022TS0402, 3072022(C0404) and the Research Initiation Project of Northeast Electric Power University (BSJXM-2021022).

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.

Abbreviations
AHE Adaptive histogram equalization
CLAHE Contrast-limited adaptive histogram equalization
CA CACoordinate attention
ASFF Adaptive spatial feature fusion
NMS Non-maximum suppression
Soft-NMS Soft non-maximum suppression
FPS Frames per second
AP Average precision
CNN Convolutional neural network

References


