An underwater target recognition method based on improved YOLOv4 in complex marine environment

Jili Zhou\textsuperscript{a}, Qing Yang\textsuperscript{a}, Huijuan Meng\textsuperscript{a} and Dexin Gao\textsuperscript{b}

\textsuperscript{a}College of Information Science and Technology, Qingdao University of Science & Technology, Qingdao, People’s Republic of China; \textsuperscript{b}College of Automation and Electronic Engineering, Qingdao University of Science & Technology, Qingdao, People’s Republic of China

ABSTRACT

In the marine environment, there are problems such as complex background and low illumination, resulting in poor picture quality, and the aggregation of small targets and multiple targets brings difficulties to target recognition. In order to improve the accuracy of marine target detection, the image enhancement and improved YOLOv4 algorithm are used to identify marine organisms. Firstly, aiming at the problems of image blur, some of the images are enhanced with the multi-scale retinex with colour restoration (MSRCR) enhancement algorithm so that the image is clearer and it is easier to extract features. Secondly, Mosaic augmentation is added to YOLOv4 to enrich the data set and increase network robustness. Then the Spatial Pyramid Pooling (SPP) module of YOLOv4 is improved by changing the size of the pooling core, which increases the range of feature extraction and improves the detection capabilities, and its mAP value reaches 97.06\%. The experimental results show that the detection accuracy of the image enhancement and improved YOLOv4 algorithm is 7.16\% higher than that of the original algorithm. And the improved YOLOv4 algorithm is improved in average precision and recall rate compared with other algorithms, which verifies the effectiveness of the algorithm.

ARTICLE HISTORY

Received 16 March 2022
Accepted 23 May 2022

KEYWORDS

Complex marine environment; underwater target recognition; improved YOLOv4; image enhancement

1. Introduction

More and more attention has been paid to the construction of marine pastures in recent years. One aspect of the construction of marine pastures is building monitoring capabilities, which include ecological environmental quality monitoring and ecological resource monitoring (Du & Cao, 2021). Ecological resources monitoring needs to know the situation of underwater organisms in time. If problems are not found in time, a large number of marine organisms will die. Generally, the monitoring mainly depends on divers, but they are affected by seabed pressure during work, and the time for underwater activities in a day is limited with low efficiency and poor safety. Therefore, using underwater robots to replace operators for various underwater tasks can reduce the loss and danger caused by operators working underwater, and can well adapt to the underwater environment and carry out underwater tasks. Currently, there are two kinds of underwater target recognition: acoustic vision and optical vision (Liu, 2020). Acoustic vision has advantages in identifying long-range targets, but it is limited in identifying short-range targets due to the influence of the acoustic blind area and multi-channel effect. The image has defects such as less information and high noise. The underwater robots use artificial intelligence and deep learning technology to detect marine targets through optical vision, and it has gradually become the main research direction of short-range underwater target recognition.

The mainstream target detection algorithms include Faster R-CNN (Ren et al., 2016), Mask R-CNN (He et al., 2017), SSD (Liu et al., 2016, October), YOLOv3 (Redmon & Farhadi, 2018), and YOLOv4 (Bochkovskiy et al., 2020). Faster R-CNN and Mask R-CNN are two-stage algorithm that need to extract candidate regions from the image, which is time-consuming. SSD and YOLO algorithms are one-stage target detection algorithm that do not need to generate region suggestions. They can directly obtain the location and category information of the target, which greatly improves the detection speed. Currently, deep learning algorithms are used in many fields, rail surface defect detection (Feng et al., 2020), traffic (Zheng et al., 2022), ship target recognition (Zhang et al., 2021), agriculture (Zhang et al., 2021), et al. With the continuous development of target detection technology, many researchers have applied deep learning technology...
to marine target detection. Shi et al. (2021) proposed an underwater organism detection algorithm based on improved Faster-RCNN, which improved the accuracy of marine organism detection, but its detection speed was relatively slow. Song et al. (2020) used the MSRCR algorithm to enhance the images, and combined Mask R-CNN to achieve a mAP value higher than 90% on a small underwater organism dataset. Its detection accuracy was improved, but the training time was too long. Hu et al. (2020) proposed using the improved SDD algorithm to identify echinus, which uses Resnet50 as the basic architecture of network feature extraction to replace the original VGG16 infrastructure in the SSD algorithm to increase the recognition accuracy of small targets. However, it is only the detection of one kind of marine organism, and the recognition speed is slow. Han et al. (2020) used the improved YOLOv3 algorithm to detect the enhanced marine biological images and applied it to the underwater robot to satisfy the real-time detection. But the detection has the problem of missed detection, and the marine biological detection with blurred edge is not detected. Hu et al. (2021) proposed an improved YOLOv4 network to detect uneaten feed pellets in underwater images. The YOLOv4 network was improved by modification of the feature maps using dense connections and a de-redundancy operation. Mao et al. (2021) proposed Model for marine organism detection in shallow sea using the improved YOLO-v4 network. Embedded Connection EC (Embedded Connection) components were built and embedded into the end of the YOLO-V4 network to improve detection accuracy. The detection accuracy was improved and the amount of calculation was reduced. Therefore, the YOLOv4 algorithm with fast detection accuracy and speed is selected for underwater marine biometrics.

Due to the complexity of the marine environment, the collected images have some problems, such as noise, blue–green bias, low contrast, blur, and so on. The low image quality causes some interference to the accuracy of marine organisms. Therefore, the image enhancement method is introduced to clarify the image of marine organisms (Guo et al., 2021). Xie et al. (2018) proposed a new underwater image enhancement method based on the dark channel prior model and underwater backscatter model, which avoids the influence of noise and retains more underwater image information. However, the detection speed of the method is slow and the effect on colour deviation is not very good. Fu and Cao (2020) combined the merits of deep learning and conventional image enhancement technology to improve the underwater image quality, designed a compressed-histogram equalization to complement the data-driven deep learning. The quality of the image has improved. Lu et al. (2019) proposed an underwater image restoration method based on transferring an underwater style image into a recovered style using the Multi-Scale Cycle Generative Adversarial Network (MCycle GAN) System, which used a dark channel prior (DCP) algorithm to get the transmission map of an image and designed an adaptive SSIM loss to improve underwater image quality. Chen et al. (2021) proposed an image super-resolution reconstruction method using attention mechanism with feature map to facilitate reconstruction from original low-resolution images to multi-scale super resolution images. The method can effectively improve the visual effect of images. In this paper, a multi-scale retinal with colour restoration (MSRCR) (Petro et al., 2014) enhancement algorithm is used for image enhancement to improve the clarity of the images, and the accuracy of marine biological detection is improved.

Based on the above situation, an underwater target recognition method based on image enhancement and improved YOLOv4 in the complex marine environment is proposed. In this method, MSRCR enhancement and Mosaic augmentation are used to preprocess the marine biological images to obtain a data set with a better visual effect. The enhanced data set is input into the improved YOLOv4 model for marine biological image detection. The experimental results show that the detection accuracy is improved by adding the enhanced images. Moreover, the improved YOLOv4 model has a good effect on a small target, multi-target, and occluded target detection. The main contributions of this paper can be summarized as follows:

1. Aiming at the problems of blurred and bluish-green bias of the collected images in complex marine environment, MSRCR enhancement algorithm is introduced to improve the quality of the images and better extract the features of marine organisms, so as to improve the accuracy of marine organism recognition.
2. Since the underwater image collection is restricted by conditions, the number of collected images is small, and the training effect is not good. Mosaic enhancement algorithm is introduced to enrich the dataset of marine biological images and improve the robustness of the network.
3. For the problem of unbalanced distribution of marine biological samples, aggregation of small targets and multiple targets, which is prone to missed detection. The SPP structure of YOLOv4 algorithm is improved by changing the size of SPP pooling kernel, which can increase the range of feature extraction, improve the accuracy of small target recognition and improve the leakage detection situation.
The rest of this article is described as follows. The second part introduces the algorithms used in image preprocessing, including the principles of the MSRCR algorithm and Mosaic algorithm. The third part analyses the principle and improvement of the YOLOv4 algorithm. The fourth part shows the experimental results and verifies the effectiveness of the algorithm. The fifth part is the experimental conclusion.

2. MSRCR enhancement algorithm

When detecting images in a complex marine environment, the collected images will have problems such as uneven illumination, low illuminance, serious noise, and complex backgrounds. It will have a certain impact on underwater target recognition. In order to achieve better results in detection, images need to be enhanced to improve the clarity. The image preprocessing process is shown in Figure 1.

The image preprocessing part uses the MSRCR algorithm and the Mosaic enhancement algorithm. The MSRCR algorithm enhances the collected marine biological images, and obtains clearer images by improving the contrast and brightness. It is convenient for feature extraction. And then using the Mosaic algorithm randomly cuts and splices the enhanced images to enrich the detection data set.

2.1. MSRCR enhancement algorithm

In the process of preprocessing the collected marine biological images, firstly, the MSRCR algorithm is used to enhance original images. The enhanced image features are clearer and easy to extract. The formula for the MSRCR algorithm with colour recovery is shown in Equation (1):

\[ R_{\text{msrcr}}(x, y) = D(x, y) \cdot R_{\text{msr}}(x, y) \]  

where \( R_{\text{msr}}(x, y) \) is the form before the improvement of the MSRCR algorithm, which has no colour restoration, and the formula is shown in Equation (2):

\[ R_{\text{msr}}(x, y) = \sum_{i=1}^{n} W_i \cdot \lg[I(x, y) - \frac{\lg[F(x, y) \times l(x, y)]}{\max - \min}][255] \]

where \( W_i \) is the weight of the \( i \)-th scale; \( n \) is the number of scales; \( F(x, y) \) is the Gaussian function. The Gaussian function is convolved with the input marine biological images to obtain the illumination component.

\[ D(x, y) = g \cdot [\lg(\alpha \cdot I(x, y) + 1) - l'(x, y)] \]

where \( g \) and \( \alpha \) are hyperparameters, which are the gain constant and the controlled nonlinear intensity respectively. After many experiments obtain the best image enhancement results, and they are set to 1 and 128.

The value after colour restoration needs to be mapped to the interval 0-255. The mapping function is shown in Equation (4):

\[ \text{image} = \text{Clip}\left(\frac{R_{\text{msrcr}} - \min}{\max - \min} \times 255\right) \]

where \( \text{Clip} \) represents a clipping function, which cuts values other than 0 ~ 255 to the boundary of the marine biological image. The formulas of min and max are shown in Equation (5):

\[ \min = \text{mean} - \text{dynamic} \cdot \text{std} \]
\[ \max = \text{mean} + \text{dynamic} \cdot \text{std} \]

where \( \text{mean} \) is the average value of all pixels in the \( R_{\text{msrcr}} \); std is the standard deviation; dynamic is the hyperparameter. The smaller the dynamic value, the contrast of the image is higher. In order to have a better display effect on the enhanced images of marine organisms, set it to 2.5. The comparison before and after MSRCR enhancement is shown in Figure 2.

The enhanced image eliminates the blue–green bias, improves the contrast and makes the images clearer. It has better visual effect and better image quality. It is conducive to feature extraction.

![Figure 1. Image preprocessing.](image-url)
2.2. Mosaic data augmentation

After the original marine biological images is enhanced by MARCR, clearer images are obtained, and then Mosaic (Bochkovskiy et al., 2020) augmentation is performed on images. Mosaic data augmentation is to randomly select four enhanced marine biological images. The Mosaic data augmentation algorithm first randomly selects four enhanced marine biological images for random scaling and clipping, then the corresponding parts are crossly spliced randomly, after the spliced images are transmitted to the YOLOv4 network for training. The target box corresponding to each original image will be limited by the cross clipping, and will not exceed the clipping range of the original image. The specific process is shown in Figure 3.

Mosaic data augmentation algorithm enriches the detection data set, adds many small targets, and makes the network more robust. In addition, when calculating batch normalization, Mosaic calculates the data of four images at a time, so that the size of the mini-batch does not need to be large, and one GPU can achieve better results.

3. YOLOv4 algorithm analysis and improvement

In the early stage, the MSRCR algorithm and Mosaic data augmentation algorithm are used to preprocess the marine biological image, and then YOLOv4 is used for target detection. The pre-processed marine biological images are stretched and scaled to 608 \times 608, and then input to the YOLOv4 algorithm module. In the backbone module of YOLOv4, CSPDarknet53 (Wang et al., 2020, June) is used for network training to extract the
features of marine biological images. Neck structure introduces technologies such as Spatial Pyramid Pooling (SPP) (He et al., 2015) and Path Aggregation Network (PANet) (Liu et al., 2018, June) to fuse the previously extracted marine biological image features at different scales and output three feature maps at different scales. Head module firstly merges the features of the three feature layers output by PANet after one convolution operation, then predicts the detection results after one convolution operation, and obtains the prediction box by adjusting the priority box. Finally, the detection results of the marine biological targets are output, including the bounding box, category and confidence parameters. Figure 4 shows the improved YOLOv4 network structure.

Due to the living habits of marine organisms, it is easy to missed detection and error detection. In order to detect marine organisms of different sizes, the SPP structure is improved. The improved SPP module can increase the global receptive field of the feature extraction network and enhance the feature extraction ability. It can not only reduce the impact of environmental background and obstacle occlusion on the detection of marine organisms effectively, but also improve the target detection ability to avoid error detection and missed detection.

3.1. Feature extraction

YOLOv4 uses a new backbone network CSPDarknet53 (Wang et al., 2020, June) to extract features from pre-processed marine biological images. CSPDarknet53 is an improvement of Darknet53. The Darknet53 network has five large residual blocks, and the number of small residual units contained in each large residual block is 1, 2, 8, 8 and 4. CSPDarknet53 adds Cross Stage Partial Network (CSP) (Wang et al., 2020, June) module to each large residual block of the Darknet53 network, which solves the problem of gradient disappearance caused by the continuous deepening of the network, greatly reduces the network parameters, and makes it easier to train a deeper convolutional neural network. The network parameters of CSPDarknet53 are shown in Table 1.

The structure of the CSP module is shown in Figure 5. CSP is composed of the convolutional block (CBM) and residual layer. CBM includes the convolution layer, batch normalization layer and (Misra, 2019) activation function.

### Table 1. CSPDarknet53 network parameters.

| CSPDarknet53 | CBM | CSP1 | CSP2 | CSP8 | CSP8 | CSP4 |
|--------------|-----|------|------|------|------|------|
| parameter    | [3 × 3.32] | 1 × 1.32 | 1 × 1.64 | 1 × 1.128 | 1 × 1.256 | 1 × 1.512 | 3 × 3.312 | 4 |
| output size  | 608 × 608,32 | 304 × 304,64 | 152 × 152,128 | 76 × 76,256 | 38 × 38,512 | 19 × 19,1024 |
The size of the convolution core in front of each CSP module is $3 \times 3$ and the stripe value is 2, which can play the role of down sampling. The structure of Backbone extraction and residual edge separation in the CSP module can better extract image features, which can maintain high accuracy and reduce the amount of calculation while keeping lightweight.

The pre-processed marine biological images are stretched and scaled to $608 \times 608$, and input to the convolution neural network. After five CSP modules, the image features are fully extracted by the Backbone feature extraction network CSPDarkNet53, and the later three scales of $76 \times 76$, $38 \times 38$ and $19 \times 19$ image features are transmitted to the SPP network and PANet network for feature fusion.

### 3.2. Feature fusion network

After extracting marine organism features using feature extraction network CSPDarknet53 in Backbone module, it is input to Neck part for feature fusion using feature fusion network to improve the detection effect. As shown in Figure 6, the Neck part includes SPP and PANet. The SPP module is located at the connection between the Backbone network and Neck, whose network structure uses four different scales maximum pooling cores for multi-scale feature fusion, so that the input of the convolutional neural network is not limited by the fixed scale. On the premise of not reducing the network running speed, the SPP module increases the receptive field, obtains the feature layer, and separates significant context features. SPP stacks and convolutes the new feature map and the feature map before entering the network, and then outputs it to the feature fusion network PANet structure. PANet merges up sampling, down sampling and features, the results of up sampling and down sampling are concatenated with the results of the convolution of the corresponding effective feature layers. After merging multi-layer information, start to predict the information, effectively use the underlying information, and finally obtain three YOLO Head effective feature layers of $19 \times 19$, $38 \times 38$ and $76 \times 76$.

PANet structure is a circular pyramid structure composed of convolution operation, up sampling operation, feature layer fusion and down sampling operation. It replaces the feature pyramid network (FPN) (Lin et al., 2017, July) in YOLOv3 as a parameter aggregation method. YOLOv4 adds a top-down path expansion.
structure behind the FPN layer to repeatedly extract the incoming image features, and merge the features extracted from different feature layers of the backbone network from top to bottom to form strong positioning features, so as to further improve the feature positioning ability. FPN layer conveys strong semantic features from bottom to top and aggregates parameters of different detection layers from different backbone layers with PANet to further improve the feature extraction ability of the network.

And the PAN structure is improved in YOLOv4, as shown in Figure 7. In PANet paper (Liu et al., 2018, June), the addition method is used to merge the feature information of feature maps of different sizes and the number of output feature mapping channels remains unchanged. In YOLOv4, the concat method is used to connect two input feature maps and merge the channel numbers of the two feature maps.

3.3. Network prediction

The feature output extracted from PANet is three different scale YOLO Head prediction layers. The prediction layer information includes target score, prediction anchor box size coordinates, classification confidence, etc. The position and size of the anchor box are adjusted appropriately by using the information of the YOLO Head prediction layer, and some anchor boxes are removed by using the target scoring screen. Finally, the results of the three prediction layers are combined through Non-Maximum Suppression and output the marine target detection results. YOLOv4 continues the basic idea of YOLOv3 bounding box prediction and adopts a prediction scheme based on a priori box. YOLOv4 bounding box prediction is shown in Figure 8.

\[(c_x, c_y)\] is the coordinates of the upper left corner of the grid unit where the target centre point is located, \((p_w, p_h)\) is the width and height of the prior box, \((b_w, b_h)\) is the width and height of the actual prediction box for marine organisms, and \((\sigma(t_x), \sigma(t_y))\) is the offset value predicted by the convolutional neural network. The position information of the marine organisms bounding box is calculated by Equations (6)–(10), where \(t_w, t_h\) is predicted by the convolutional network, and \((b_x, b_y)\) is the coordinates of the centre point of the actual prediction box of marine organisms. In the obtained characteristic graph, the length and width of each grid cell are 1. Therefore, in Figure 8 \((c_x, c_y) = (1, 1)\), the sigmoid function is used to limit the prediction offset between 0 and 1.

\[
b_x = \sigma(t_x) + c_x \tag{6}
\]
\[
b_y = \sigma(t_y) + c_y \tag{7}
\]
\[
b_w = p_we^{t_w} \tag{8}
\]
\[
b_h = p_he^{t_h} \tag{9}
\]
\[
\sigma(x) = \frac{1}{1 + e^{-x}} \tag{10}
\]

The loss function can be divided into three parts: bounding box prediction loss, confidence loss and classification loss. In the bounding box loss prediction, YOLOv4 uses CIOU (Zheng et al., 2020) (complete intersection over union) loss instead of mean square error (MSE) loss on the basis of YOLOv3, which makes the boundary regression faster and more accurate. By minimizing the loss function between the marine organism prediction box and the actual box, the network is trained, and the weight confidence and classification loss are continuously updated.

The development of the loss function has been improved many times, as shown in Figure 9. The IOU Loss (Yu et al., 2016, October) is used firstly, but when there is no overlapping area between the prediction box and the target box, IOU Loss will not work (Rezatofighi et al., 2019). In 2019, GIOU Loss (Rezatofighi et al., 2019) (Generalized IOU) was proposed, which added an intersecting scale measurement method on the basis of IOU Loss, and solved the problem of gradient disappearance caused by the non-overlapping of the two boxes of IOU Loss. However, when the prediction box is inside the target box...
and the size is the same, the relative position relationship cannot be distinguished, and the convergence speed of loss is affected. In 2020, DIOU Loss (Rezatofighi et al., 2019) (Distance IOU) was proposed. DIOU Loss considers the overlapping area and the centre points distance and adds a simple penalty term to minimize the standardized distance between the two detection centre points and accelerate the convergence process of loss. However, when the centre point distance of the marine organism prediction box and the target box are the same, the value of DIOU Loss is the same, and can’t make distinctions. Therefore, a more comprehensive CIOU Loss (Zheng et al., 2020) (Complete IOU) is proposed, which adds an impact factor based on the DIOU Loss and takes into account both the prediction box and the target box. The CIOU Loss function equation is shown in Equation (11).

\[
CIOU\_Loss = 1 - CIOU = 1 - \left(\text{IOU} - \frac{\text{Dis} \tan \text{ce}_{-}^2}{\text{Dis} \tan \text{ce}_{-}^2} - \frac{\nu^2}{(1 - \text{IOU}) + \nu}\right)
\]

where \(\nu\) is a parameter that measures the consistency of the aspect ratio. The calculation formula is Equation (12).

\[
\nu = \frac{4}{\pi^2} \left(\text{arctan} \frac{w^R}{h^R} - \text{arctan} \frac{w^P}{h^P}\right)^2
\]

where \(w^R\) and \(h^R\) are the width and height of the marine organism real box, \(w^P\) and \(h^P\) are the width and height of the marine organism prediction box.

YOLOv4 uses the CIOU regression method to make the prediction box regression faster and more accurate.

**3.4. SPP structure improvement**

The size of the input images cannot meet the size required during input, so it needs to be cropped and warped. In this way, the aspect ratio of the images and the size of the input images will be changed, which will distort the original images. In order to transform any size feature map into a fixed-size feature vector, spatial pyramid pooling is introduced.

After the feature map enters the SPP structure, the maximum pooling operation is carried out by four cores with different scales respectively, and then the pooled results are output by channel splicing. Compared with YOLOv3 using a maximum pooling of \(n \times n\) scale, improved YOLOv4 can more effectively enrich the expression ability of the feature map, and finally obtain the local feature map and global receptive field information.

The SPP network uses pooling cores with sizes of \(1 \times 1\), \(5 \times 5\), \(9 \times 9\), and \(13 \times 13\) for maximum pooling, and then merge feature maps of different scales. For feature map of any size is first divided into 16, 4, 1 blocks, then maximum pooling is done on each block, and the pooled features are spliced to get a fixed dimensional output. Since the input image size is \(608 \times 608\), and the size change rule of the feature map is \(608, 304, 152, 76, 38,\) and 19. After 5 CSP modules, the output dimension is \(19 \times 19\) size feature map. According to the particularity of the experimental detection object, in order to improve the accuracy of target recognition, the SPP structure is adjusted and the multi-scale maximum pool is improved.

As shown in Figure 10, the pooling core size is set to \(1 \times 1\), \(5 \times 5\), \(9 \times 9\), and \(13 \times 13\) for maximum pooling, and then merge feature maps for concat operation. The maximum pooling adopts a padding operation, the number of moving steps is 1, and the size of the feature map remains unchanged \(19 \times 19 \times 512\) after pooling. The input of SPP is the output of the last down sampling of the backbone feature extraction network. It compresses the last down sampled feature layer in three different degrees and performs comprehensive extraction.

The improved SPP module can increase the global receptive field of the feature extraction network and enhance the feature extraction ability, so that can effectively reduce the impact of environmental background and obstacle occlusion on marine biological detection, and can improve the detection ability of marine organisms and avoid false detection and missed detection.
4. Experiment and result analysis

4.1. Data set and experimental environment

The data set used in this paper is some marine organism images collected in offshore marine pastures. The marine organism data set has four species: echinus, starfish, scallop, and holothurian. After preliminary preprocessing of the data to remove some invalid data, a total of 3408 images were selected, among which the ratio of the training set to the validation set is 8:2, and 512 pictures were enhanced. Due to their different living habits, single species aggregation and uneven distribution of species samples are easy to appear in images, and the number of echinus is higher than other species, which adds some difficulty to model training.

The experiment is carried out on a server with NVIDIA GeForce GTX 3090 Ti graphics card and 32G memory. The system uses Ubuntu 18.04LTS, CUDA version is 11.0, Cudnn version is 8.04, and the deep learning environment is Anaconda3, python3.7, and PyTorch1.8.0.

4.2. Experimental process

During the experiments, the training parameters are continuously adjusted to find the parameters with fast training speed and good detection results. If the learning rate is too low, the gradient will drop slowly, and the training time will be long. If the learning rate is too high, the training speed will be accelerated but it is prone to occur gradient explosion. After many experiments, it is found that the step size is set to 16 and the learning rate is set to $10^{-3}$ for the best results. A higher step size or a lower learning rate will make the final loss too high and thus result in a poor weight effect. Lower step size will make the training speed slower. In the experiment, the ‘warm-up’ strategy is used to set the learning rate. The learning rate is set to $10^{-3}$ for the first 60 cycles, $10^{-4}$ for 60–100 cycles, and $10^{-5}$ for 140 cycles. And the step size is set to 16, 8, and 4, respectively. The input pixel is $608 \times 608$, the confidence is 0.5, and the IOU is 0.5. The loss function curve image is shown in Figure 11.

Due to the use of the pre-training model, the loss value decreases rapidly in the first few iterations, so it soon decreases to a lower value. It can be seen from the loss function curve that the loss value tends to decline until it finally converges. When each time the learning rate and step size change, the loss value increases firstly, then starts to decrease at a relatively fast speed, and finally gradually tends to level off. When the number of iterations is 100, the loss value decreases to a relatively small value and reaches a relatively stable level. When the number of iterations reaches 140, the loss value finally decreases to 2.00, and the overall training effect is ideal.

4.3. Evaluation parameters

The mean average precision (mAP) is used as the model evaluation index, whose calculation equation is as
predicted targets to the total number of positive predictions, whose calculation equation is as Equation (14), and R represents the ratio of the number of accurate targets predicted to the actual number of targets, whose calculation equation is as Equation (15). The threshold value of the cross-merge ratio is greater than 0.5 as the true target.

\[
mAP = \frac{1}{N} \sum_{i=1}^{N} \int_{0}^{1} PdR \tag{13}
\]

\[
P = \frac{TP}{TP + FP} \tag{14}
\]

\[
R = \frac{TP}{TP + FN} \tag{15}
\]

where TP is the number of true targets predicted by the network; FP is the number of false targets predicted; FN is the number of true targets that cannot be successfully predicted; TP + FP is the number of all predicted targets; TP + FN is the number of all marked targets.

As shown in Figure 12, it is the FP and TP values of the detection set,

**4.4. Detection result**

Figure 13 shows the detection results of four marine organisms based on improved YOLOv4, and their AP values, accuracy rate, and recall rate are shown in Table 2.

AP values are used as indicators to measure various species. It can be seen from Table 2 that echinus has the highest AP values, followed by starfish, scallop, and holothurian. Their AP values are 98.53%, 97.94%, 96.02%, and 95.78%, respectively. The main reason is that the number of echinus is the largest in the training set accounting for a large proportion, while the proportion of holothurian is the smallest.

**4.5. Comparison of algorithms**

Compare the detection results of different algorithms, such as Faster-RCNN, SSD, YOLOv3, YOLOv4, ie-YOLOv4 enhanced by MSRCR, and the improved M-YOLOv4 enhanced by MSRCR, and the improved M-YOLOv4.

**Table 3. Comparison of target detection results of different algorithms.**

| Network       | echinus | starfish | scallop | holothurian | mAP/% |
|---------------|---------|----------|---------|-------------|-------|
| Faster RCNN   | 90.43   | 85.78    | 74.31   | 81.75       | 83.07 |
| SSD           | 91.61   | 85.83    | 75.07   | 80.06       | 83.29 |
| YOLOv3        | 89.36   | 88.90    | 85.19   | 83.68       | 86.78 |
| YOLOv4        | 95.75   | 91.45    | 87.17   | 85.24       | 89.90 |
| ie-YOLOv4     | 98.13   | 96.59    | 94.62   | 93.78       | 95.78 |
| M-YOLOv4      | 98.53   | 97.94    | 96.02   | 95.78       | 97.06 |
algorithm. Resnet50 is the model used by the Faster-RCNN algorithm, and VGG is the model used by the SSD algorithm. The model used by YOLOv3 is Darknet53, the model used by YOLOv4 is CSPDarknet53, and the enhancement algorithm used by ie-YOLOv4 is the MSRCR algorithm. Their detection results are evaluated objectively using AP and mAP. Table 3 shows the comparison of target detection results of different algorithms.

Figure 14. Target recognition results: (a) images without enhancement and YOLOv4 without improvement; (b) images with enhancement and YOLOv4 without improvement; (c) images with enhancement and YOLOv4 with improvement.
(1) In Table 3, the detection mAP values of the Faster RCNN, SSD, YOLOv3 and YOLOv4 algorithms are 83.07%, 83.29%, 86.78%, and 89.90%, respectively. According to their map value, it can be seen that the value detected by the YOLOv4 algorithm is higher and the detection effect is better. It shows that YOLOv4 is feasible in marine target recognition.

(2) As can be seen from the comparison of the detection results of YOLOv4 without image enhancement and ie-YOLOv4 enhanced by MSRCR and Mosaic algorithms, the mAP value of the unenhanced YOLOv4 image detection is 89.90%, and the mAP value of the enhanced ie-YOLOv4 image detection is 95.78%, an increase of 5.88%. It shows that using the enhancement algorithm for preprocessing is conducive to target detection.

(3) The detection results of unimproved YOLOv4 and the improved M-YOLOv4 algorithm are analysed, and the mAP value of the improved M-YOLOv4 algorithm reaches 97.06%, which is an increase of 1.28% over the unimproved YOLOv4 algorithm. The result verifies the effectiveness of the improved M-YOLOv4 algorithm.

From the visual detection effect, the unenhanced YOLOv4, enhanced ie-YOLOv4 and improved M-YOLOv4 are evaluated subjectively. Figure 14(a–c) shows the marine biometric recognition results of the images without enhancement, enhanced images without YOLOv4 improvement, and improved YOLOv4 respectively. Echinus are marked with green boxes, starfish with purple boxes, scallops with blue boxes, and holothurians with red boxes.

Observation of the three sets of images in Figure 14, it can be seen that the enhanced images in Figure 14(b,c) have higher recognition accuracy than the images without enhancement in Figure 14(a). There is an echinus in the lower right corner of the first column of images, which can be seen that the images in Figure 14(a) are not recognized and there is a small target loss phenomenon. But the images in Figure 14(b and c) are recognized, which means that the enhanced images have better results in terms of recognition accuracy and precision.

For the unimproved YOLOv4 in Figure 14(a and b) and the improved YOLOv4 in Figure 14(c), it shows that the difference in accuracy between unimproved YOLOv4 and M-YOLOv4 is not significant. But there is a small holothurian at the bottom of the middle of the second column of images, which is not detected by the unimproved YOLOv4 algorithm, however, M-YOLOv4 can detect it. There is a missed detection in ie-YOLOv4, and improved YOLOv4 extracts more detailed features. The improved YOLOv4 has a good detection effect on small targets and underwater creatures with occlusion.

5. Conclusions
This article is aimed at the recognition of organisms in the marine environment. Due to the environment, the obtained underwater image has a problem such as unclear images and colour bias. Therefore, this article uses the MSRCR enhancement algorithm and improved YOLOv4 for target recognition. The experimental results show that the effect of enhanced image detection is better than the original image, and the value of mAP increases from 89.90% to 90.89%. Then Mosaic augmentation is added, which enriches the data sets and increases the complexity of training. The value of the map is 95.78%, which is 4.89% higher than the original algorithm. On this basis, the SPP module of YOLOv4 is improved, and the size of the pooling core is changed. The improved algorithm map reaches 97.06%, which is higher than the original accuracy. The improved YOLOv4 algorithm has a good detection effect on small targets and multi targets. The experimental results verify the feasibility and effectiveness of the improved algorithm.

For the improved YOLOv4 network in this paper, its application has strong expansibility. It can be applied to other underwater target detection and recognition fields, which are of high scientific use and research value. Future research will focus on improving the recognition speed, so that real-time detection of marine organisms can be achieved.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Funding
This work is partially supported by the National Natural Science Foundation of China (grant number U1806201).

Data availability
The data that support the findings of this study are openly available in GitHub at https://github.com/Eric9906/JPEGImages and https://github.com/Eric9906/yolov4-code

References
Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv.
Chen, Y., Liu, L., Phonevilay, V., Gu, K., Xia, R., Xie, J., Zhang, Q., & Yang, K. (2021). Image super-resolution reconstruction based on feature Map attention mechanism. Applied Intelligence, 51(7), 4367–4380. https://doi.org/10.1007/s10489-020-02116-1
Du, Y. W., & Cao, W. M. (2021). Theoretical framework for the supervision of the ecological security of China’s marine ranching. China Population, Resources and Environment, 31(1), 182–191.

Feng, J. H., Yuan, H., Hu, Y. Q., Lin, J., Liu, S. W., & Luo, X. (2020). Research on deep learning method for rail surface defect detection. IET Electrical Systems in Transportation, 10(4), 436–442. https://doi.org/10.1049/iet-est.2020.0041

Fu, X. Y., & Cao, X. Y. (2020). Underwater image enhancement with global–local networks and compressed-histogram equalization. Signal Processing: Image Communication, 86. https://doi.org/10.1016/j.image.2020.115892

Guo, Y. J., Wu, Q., Yuan, J. J., Hou, J.C. & Lü, W. H. (2021). An underwater image processing and object detection based on deep CNN method. Journal of Sensors, 2020. https://doi.org/10.1155/2020/6707328

He, K. M., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October 22–29). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2961–2969.

He, K. M., Zhang, X. Y., Ren, S. Q., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

Hu, K., Lu, F. Y., Lu, M. X., Deng, Z. L., & Liu, Y. P. (2020). A marine object detection algorithm based on SSD and feature enhancement. Complexity, 2020. https://doi.org/10.1155/2020/5476142

Hu, X. L., Liu, Y., & Zhao, Z. X. (2021). Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network. Computers and Electronics in Agriculture, 185. https://doi.org/10.1016/j.compag.2021.106135

Lin, T. Y., Dollar, P., & Girshic, R. (2017, July 21–26). Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 936–944.

Liu, J. X. (2020). Research on underwater AUV navigation and target detection based on acousto-optic image (Master’s thesis). Jilin University, Jilin, China. https://doi.org/10.27162/cdnki.jilin.2020.005943

Liu, S., Qi, L., Qin, H. F., Shi, J. P., & Jia, J. (2018, June 18–22). Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, 8759–8768.

Liu, W., Anguelov, D., & Erhan, D. (2016, October 8–16). SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Springer, Amsterdam, The Netherlands, 21–37.

Lu, J. Y., Li, N., Zhang, S. Y., Yu, Z. B., Zheng, H. Y., & Zheng, B. (2019). Multi-scale adversarial network for underwater image restoration. Optics & Laser Technology, 110, 105–113. https://doi.org/10.1016/j.optlastec.2018.05.048

Mao, G. J., Weng, W. D., Zhu, J. D., Zhang, Y., Wu, F. C., & Mao, Y. Z. (2021). Model for marine organism detection in shallow Sea using the improved. Transactions of the Chinese Society of Agricultural Engineering, 37(12), 152–158. https://doi.org/10.11999/JET190803

Misra, D. (2019). Mish: A self regularized non-monotonic neural activation function. arXiv preprint.

Petro, A. B., Sbert, C., & Morel, J. M. (2014). Multiscale retinax. Image Processing on Line, 4, 71–88. https://doi.org/10.5201.ipol.2014.107

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv, 1804.02767.

Ren, S. Q., He, K. M., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 39(6), 1137–1149. https://ieeexplore.ieee.org/document/7485869

Rezatofighi, H., Tsoi, N., & Gwak, J. Y. (2019, June 15–22). Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 658–666.

Shi, P. F., Xu, X. W., Ni, J. J., Xin, Y., Huang, W., & Han, S. (2021). Underwater biological detection algorithm Based on improved faster-RCNN. Water, 13(17), 2420. https://doi.org/10.3390/w13172420

Song, S. J., Zhu, J. X., Li, X. H., & Huang, Q. B. (2020). Integrate MSRCR and Mask R-CNN to recognize underwater creatures on small sample datasets. IEEE Access, 8, 172848–172858. https://doi.org/10.1109/ACCESS.2020.3025617

Wang, C. Y., Liao, H. Y. M., & Wu, Y. H. (2020, June 14–19). CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR), Seattle, WA, USA, 1571–1580.

Xie, K., Pan, W., & Xu, S. X. (2018). An underwater image enhancement algorithm for environment recognition and robot navigation. Robotics, 7(1), 14. Article ID 7010014. https://doi.org/10.3390/robotics7010014

Yu, J. H., Jiang, Y. N., Wang, Z. Y., Cao, Z. M., & Huang, T. (2016, October 01). Unitbox: An advanced Object Detection network. In Proceedings of the 24th ACM International Conference on Multimedia (ACM) (New York, NY, USA, 516–520.

Zhang, D., Zhan, J., Tan, L., Gao, Y., & Župan, R. (2021). Comparison of two deep learning methods for ship target recognition with optical remotely sensed data. Neural Computing and Applications, 33(10), 4639–4649. https://doi.org/10.1007/s00521-020-05307-6

Zhang, Z. G., Zhang, Z. D., Li, J. N., Wang, H. Y., Li, Y. B., & Li D. H. (2021). Potato detection in complex environment based on improved Yolov4 model. Transactions of the Chinese Society of Agricultural Engineering, 37(22), 170–178. https://doi.org/10.11975/j.issn.1002-6819

Zheng, H., Liu, J., & Ren, X. (2022). Dim target detection method based on deep learning in complex traffic environment. Journal of Grid Computing, 20(1), 1–12. https://doi.org/10.1007/s10723-021-09594-8

Zheng, Z. H., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding Box regression. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 34(07), 12993–13000. https://doi.org/10.1609/aaai.v34i07.6999