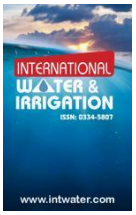

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

Underwater Object Detection Based on Spatial Pyramid and Channel Attention

Yuxin Long[1], Huili Xia[2*], Xuexiang Li[1], Weixing Zhang[1]
[1]School of Cyber Science and Engineering, Zhengzhou University, China
[2]Zhengzhou University of Economics and Business, China

Abstract: In order to solve the problems of low accuracy and high detection delay of conventional object detector in underwater environment, underwater object detection network model based on single-stage target detection is introduced. The specific work is as follows: In order to solve the problem that conventional detectors have low detection accuracy in detecting blurry small targets in underwater scenes due to water quality, a channel spatial attention mechanism is designed, which enables the model to focus more on the feature learning of target objects. This improves the extraction of information within the channel and enhancing the extraction of salient features in cases where the distinction between the foreground and background is not obvious and improves the accuracy on small target in underwater detection. On the basis of the conventional object detector, the spatial pyramid pooling module is designed, which reduces the number of computational parameters required for the extraction of features of the model while maintaining the same receptive field. This improves the inference efficiency of the network, and effectively alleviates the detection delay. The result shows that the improved model can identify underwater targets more accurately, and the detection speed of the model is also improved. The detection accuracy of the model achieved to 80.61% and the FPS achieved to 64.23.

Keywords- YOLO, underwater object detection, channel attention, spatial pooling channel.

1. Introduction

Computer vision is widely used in underwater detection to detect targets in underwater images[1]. It has significant applications in fields such as ocean science, exploration, security, and environmental protection[2]. Underwater object detection is of great importance in improving aquaculture efficiency, protecting endangered species, and monitoring the ecological environment[3]. In recent years, deep learning technology has achieved significant results in these fields[4]. Deep learning achieves automatic processing of large amounts of data by constructing and training neural network models, making underwater image detection more accurate, robust, and faster than traditional methods[5]. However, underwater object detection also faces many challenges, such as interference factors such as water currents and light, as well as lacking of underwater dataset. These problems lead to low detection accuracy and low efficiency[6]. Therefore, searching new object detection technologies is crucial for improving the accuracy and efficiency of underwater detection.

In order to solve the problems of low accuracy and high detection delay of object detection in underwater environment, space pyramid pool module and channel space attention module are designed respectively in the detector structure. The channel attention module can improve the feature extraction and integration of the model in space and channel dimension. In this way the model can focus more on the small and fused objects in the detection process, and thus improve the accuracy of underwater target detection. By reducing the repeated extraction of target features, the space pyramid pool module reduces the redundant calculation and speeds up the extraction of candidate boxes, thereby improving the detection efficiency and effectively alleviating the problem of high detection delay.

Compared with the conventional object detection tasks, there are few researches on underwater target detection. Underwater detection has a lot challenges such as color distortion, noise, blurred texture and low visibility[7]. With the establishment of underwater dataset, the universal target detection model is applied to underwater scene. Knausgård[8] et al used the Squeeze and Excitation(SE) module to improve the backbone network of the YOLO model and apply it to underwater target detection. Ye[9] et al. improved the SPP backbone network of YOLO to

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

improve the performance of small target detection. Zhang[11] et al. used channel attention to improve the ability of backbone network to extract high-frequency information from underwater images, which greatly improved the model performance in underwater dataset.

Spatial pyramid pooling

He. et al firstly proposed spatial pyramid pooling (SPP) in 2015, which has the ability to extract multi-scale feature information while receiving input of any size[10]. In this way the convolutional neural network can be better adapted to images of different scales and improves its object recognition ability. Additionally, pyramid pooling can also reduce the possibility of overfitting and improve the generalization ability of the model. Compared with the merely using maximum pooling, this method expands the inclusion range of backbone features and separate contextual features in images more effectively. The outputs feature of the structure is in a specified size through multi-receptive field feature extraction, which increases the receptive field of the network and relives the repeated extraction of target-related features by the convolutional neural network model. To a great extent, it speeds up the efficiency of generating candidate boxes, reduces the redundant calculation, and the number of parameters[13]. Researchers afterwards optimized and improved the structure, and proposed the optimized version SPPF, etc., which has made great progress in speed and feature extraction ability[12].

Attention mechanism

The attention mechanism is firstly proposed in 2017[14]. It mainly processes inputs through three attributes, Query, Key, and Value(Q, K, V). In the attention mechanism, inputs are transformed into three vectors: query, key, and value. In image processing tasks, the attention mechanism can help the model focus on important areas in the input image, so as to better extract features[15].

The brief introduction of some main-stream attention modules is as follows. SENet[16] is a channel-dimension attention mechanism, in which SE attention mechanism is mainly divided into two steps. The first step is to reduce the input feature graph to 1×1 size by global average pooling operation. And the second step is to fully join the results obtained after global pooling to obtain C/r-dimension vectors, and then perform ReLu activation. Then the output makes a full connection, and changes the vector of C/r dimension back to C dimension vector, and then activates sigmoid (making the output value between 0 and 1) to get the weight matrix. But this attention module ignores the weight of the target space position[17]. ECA is also a channel attention mechanism. It obtains the enhanced weight of each feature layer by obtaining cross-channel information. Although it has better cross-channel information, it ignores the spatial information of the target. The CA attention mechanism is different from other channel attention mechanisms. It transforms the channel attention into a one-dimensional coding process that aggregates features along two different coordinates. By this method, the remote dependence is captured along one coordinate direction and the precise location information of the detected target is retained along the other spatial direction. CBAM combines channel and spatial attention mechanisms to process channel weights and spatial weights respectively, focusing on both channel information and spatial information. Different attention mechanisms pay attention to information in different ways, and different models have different compatibility with different attention mechanisms[13].

The attention mechanism is firstly proposed in 2017[14]. It mainly processes inputs through three attributes, Query, Key, and Value(Q, K, V). In the attention mechanism, inputs are transformed into three vectors: query, key, and value. In image processing tasks, the attention mechanism can help the model focus on important areas in the input image, so as to better extract features[15].

The brief introduction of some main-stream attention modules is as follows. SENet[16] is a channel-dimension attention mechanism, in which SE attention mechanism is mainly divided into two steps. The first step is to reduce the input feature graph to 1×1 size by global average pooling operation. And the second step is to fully join the results obtained after global pooling to obtain C/r-dimension vectors, and then perform ReLu activation. Then the output makes a full connection, and changes the vector of C/r dimension back to C dimension vector, and then activates sigmoid (making the output value between 0 and 1) to get the weight matrix. But this attention module

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

ignores the weight of the target space position[17]. ECA is also a channel attention mechanism. It obtains the enhanced weight of each feature layer by obtaining cross-channel information. Although it has better cross-channel information, it ignores the spatial information of the target. The CA attention mechanism is different from other channel attention mechanisms. It transforms the channel attention into a one-dimensional coding process that aggregates features along two different coordinates. By this method, the remote dependence is captured along one coordinate direction and the precise location information of the detected target is retained along the other spatial direction. CBAM combines channel and spatial attention mechanisms to process channel weights and spatial weights respectively, focusing on both channel information and spatial information. Different attention mechanisms pay attention to information in different ways, and different models have different compatibility with different attention mechanisms[13].

2. Objectives

Experimental Environment and Setup

The performance comparison experiment proved whether the improved algorithm can improve the detection rate effectively. We chose mAP and FPS as the performance indicator to analyze quantitatively. The data of the experiment is shown in Table.1.The computer hardware used in the algorithm experiments in this article is an HP computer, the processor parameters used are Intel(R) Core(TM) i9-10850K CPU@3.60 GHz, the memory is 32GB, the graphics card is NVIDIA GeForce RTX 3080; the system environment is Windows 11 Professional Edition; the Python version is 3.8.13, the Pytorch version is 1.13.0, and the CUDA version is 11.6.

All the experimental data in this paper were measured in the same environment setup. During the training process, the learning rate is updated by the learning rate decline; the maximum learning rate was 1e-3; the frozen training batch size was 16; the unfrozen training batch size was 4; the momentum was 0.937; and the training epoch was 2000. The frozen training model only trained the parts other than the Backbone network, and the unfrozen training trained the entire network model.

Datasets and Evaluation Indicators

The dataset used in this experiment is mainly based on the URPU(2019) dataset[19] . In order to enable the model to learn more features, more images of the same category are added to make the dataset richer. In order to make the distribution of training and test set categories more reasonable, we expand the total amount of the dataset. The dataset contains 7000 images, including 5400 training images, 800 verification images and 800 test images. There are five detection categories: scallops, holothurians, starfish, waterweeds and echinus. The number of each category in the dataset ranges from several thousand to thirty thousand, simulating the actual situation of uneven distribution of underwater organisms. The images in the dataset are captured in real ocean environments, with issues such as color distortion, low contrast, and blurry feature information. Additionally, there are occlusions, target density, and significant differences in the distribution of each category, leading to a highly imbalanced dataset. This poses a significant challenge for underwater image detection tasks. Therefore, before feeding the dataset images into the model for training and discrimination, image enhancement preprocessing is applied.

There are five evaluation indicators that access the model performance. Precision represents the proportion of true positive predictions among all predicted positive instances. A higher value indicates a stronger ability of the model to discriminate between positive and negative classes. Recall represents the proportion of true positive instances detected by the model among all actual positive instances. A higher value indicates a stronger ability of the model to identify all positive instances. Average Precision(AP) represents the area under the precision-recall curve for each class that is calculated appropriately. A higher value indicates better recognition accuracy for the current category. Mean average precision(mAP) is the average value of all classes' AP. A larger value indicates that the model has better accuracy in recognizing the target. Frame Per Second(FPS) represents the number of frames processed by the model per second, reflecting the reasoning speed of the model. The larger the value, the faster the inference speed of the model. Some of the formulas are shown as below.

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{mAP} = \frac{\sum_{i=1}^{k} \text{AP}_i}{k} \quad (3)$$

In these formulas, TP is the number of predicted positive samples in positive samples; FP is the number of predicted positive samples in negative samples; FN is the number of predicted negative samples in positive samples; TN is the number of predicted negative samples in negative samples; k is the number of sample types.

SPPFCSPC Performance Contrast Experiment

In order to further improve the detection speed of the one-stage network YOLOv7 and enhance its real-time performance, we introduced SPPFCSPC module into the network. This enhances the model's efficiency in integrating multi-scale feature information to accelerate the detection process. To verify the effectiveness of modified network in underwater object detection, we conducted contrast experiment and quantitatively analyze the experimental results by comparing mAP and FPS. The experiment was designed to contrast the modified model with three models. Among them, Exp.1 represents the original YOLOv7 model. Exp. 2 introduced SPPFCSPC on the architecture of Exp. 1. The performance comparison experiment proved whether the improved algorithm can improve the detection rate effectively. We chose mAP and FPS as the performance indicator to analyze quantitatively. The data of the experiment is shown in Table 1.

Experiment	SPPFCSPC module	Image size	FPS(f/s)
Exp.1	×	640*640	65.91
Exp.2	√	640*640	66.5

Table 1 Performance comparison of the improved model

It can be seen from the result that after SPPFCSPC is introduced, the model has an improvement on FPS compared with original model, which indicates the improvement enhanced the detection speed.

Attention Mechanism and Pooling Module Performance Contrast in Model

The ultimate goal of the underwater image target detection task is to detect the target to be detected accurately and quickly. Therefore, after improving the detection rate of the detection algorithm, we also conducted research on improving the accuracy of the detection algorithm. We introduced the channel space attention mechanism into the detection algorithm and enhanced the feature extraction capability of the model to improve the detection accuracy of the detection algorithm.

Because different attention mechanisms focus on information in different ways, different models are compatible with different attention mechanisms in different ways. We chose CBAM attention mechanism and added it with different sizes of feature maps that processed by feature process net(Neck). To validate the compatibility of CBAM with the model, it is compared with network models incorporating None attention mechanism, Fusion SENet attention mechanism, Fusion ECA attention mechanism, and Fusion CA attention mechanism through comparative experiments. The result is shown in Fig. 1 and **Error! Reference source not found.**.

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

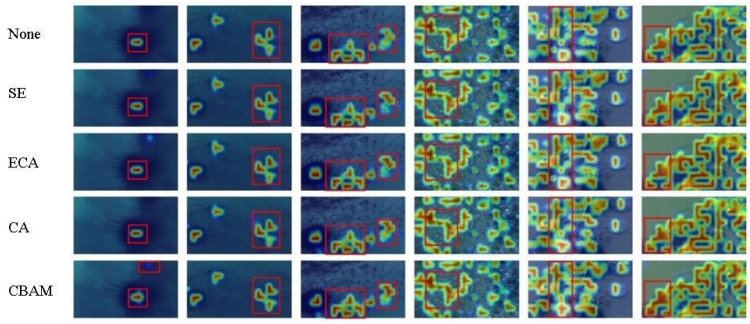

Fig. 1 Attention result comparison

Experiment	backbone	Attention mechanism	Accuracy mAP(%)
Exp.1	YOLOv7	-	78.93
Exp.2	YOLOv7	SETNet	79.28
Exp.3	YOLOv7	ECA	79.03
Exp.4	YOLOv7	CA	79.36
Exp.5	YOLOv7	CBAM	80.64

Table 2 Comparison on different attention mechanism

The experiment employs visualization methods in deep learning for qualitative analysis, which involves gradient-weighted class activation mapping called Grad-CAM. When Grad-CAM predicting images, the weight parameters obtained after pre-training are passed to the network layer that needs to be visualized to obtain a gradient matrix with the same size as the feature information map of this layer. The gradient matrix is then compressed to 1×1 through global average pooling to obtain a vector whose length is the number of channels. This vector performs weighted processing on all channels of the feature information to obtain a heat map. It is used to show the difference in the attention area after different attention mechanisms are introduced into the network model, reflecting the influence of different areas on the results. The gradient from blue light to red light indicates that the importance of the feature is increasing.

To verify the compatibility between module SPPFCSPC and the imported attention mechanism, we designed three experiments to compare with the improve model. Exp.1 represents the original YOLOv7, and Exp.2 introduces CBAM attention on the architecture of Exp. 1. Exp. 3 replaces SPPCSPC module to SPPFCSPC module on the basis of Exp. 1. Exp. 4 combined SPPFCSPC and spatial channel attention mechanism. Ablation experiments conducted on spatial channel attention mechanism and SPPFCSPC using these four experiments, test and evaluate their compatibility with each other, and determine if they can achieve the intended goal of complementary model accuracy and inference speed. The experiment result is shown in Table 3.

Experiment	CBAM	SPPFCSPC	Image size	Accuracy mAP(%)	FPS(f/s)
Exp.1	×	×	640*640	78.93	65.91
Exp.2	√	×	640*640	80.64.	62.26
Exp.3	×	√	640*640	77.60	66.5
Exp.4	√	√	640*640	80.61	64.23

Table 3 Result of ablation experiments

It can be learned from the experiment above that the introduction of the channel spatial attention module between the Neck network and the prediction network in analysis of data from Exp. 1 and 2 has led to a 1.71% improvement in the mean Average Precision (mAP) compared to the original model. The FPS of the network model has slightly decreased, indicating that the introduction of the channel spatial attention mechanism has increased the computational parameters of the network model, thereby increasing its complexity and decreasing the inference speed. Nevertheless, this attention mechanism utilizes channel attention to establish the correlation between different channels within the convolutional layers, enhancing the model's capability to extract important feature

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

information and suppress background features in images. Furthermore, the spatial attention mechanism effectively extracts overall feature information from the images and captures the spatial position information of the targets. The parallel action of both mechanisms allows the network model to focus more on the feature information of the detection targets, thus improving the quality of the feature map significantly and enhancing the overall accuracy of the model. Comparing the data from experiments 1 and 3, the FPS of the network model with the addition of the SPPFCSPC module has improved by 0.85% compared to the original model. This suggests that replacing the SPPCSPC module with the SPPFCSPC module has slightly reduced the computational parameters of the network model while maintaining the receptive field unchanged, leading to an increase in model inference speed. However, there is a significant decrease in mAP after introducing the sigmoid module, despite the improvement in inference speed.

By comparing Exp. 2 and 4, the average precision is just slightly deduced, and the detection speed had a greatly improved. So we considered simply introducing SPPFCSPC module to change the individual pooling to joint pooling deduced the extraction ability of target feature, causing that the detection results of the network model decreased. If channel spatial attention and SPPFCSPC are introduced simultaneously, the problem is alleviated to a great extent. Comparing the result of Exp. 1 and 4, the mAP increases by 1.68%. Meanwhile, the FPS is sightly reduced but higher than that of network models with channel spatial attention solely. This indicates integrating the channel spatial attention mechanism and SPPFCSPC in YOLOv7 model, obtains considerable accuracy improvement in a slightly price of model detecting speed because it increases the amount of parameters.

Introducing both of them balances the unilateral performance degradation caused by the single use of one of the modules, which makes the network model greatly improve the ability of extracting feature information of targets in underwater images, and is more suitable for underwater image target detection tasks. The loss function about model training after introducing channel spatial attention mechanism and SPPFCSPC pooling module is shown in **Error! Reference source not found.**.

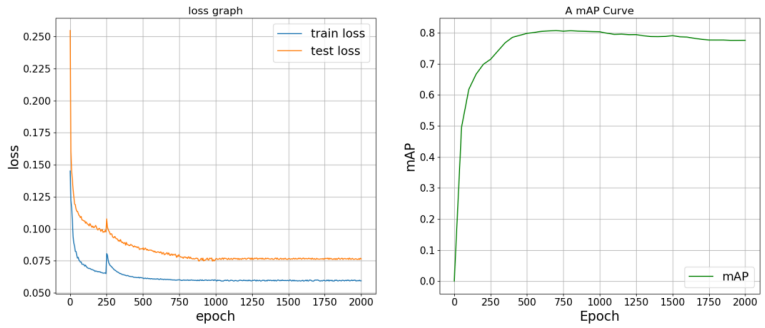

Fig. 2 Model loss curve and mAP curve

It can be seen that the model starts to convergence at around 750 epoch, and the mAP reaches to the a maximum value, then it begins to go down. This means at this moment, the model has reached the global optimal effect.

After incorporating the attention mechanism into the detection algorithm, the detection algorithm's performance on each category in the dataset is statistically analyzed. Specifically, the accuracy and Average Precision (AP) values of each category are calculated to compare the detection performance of the algorithm on different categories. The result is illustrated by Fig. 3.

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

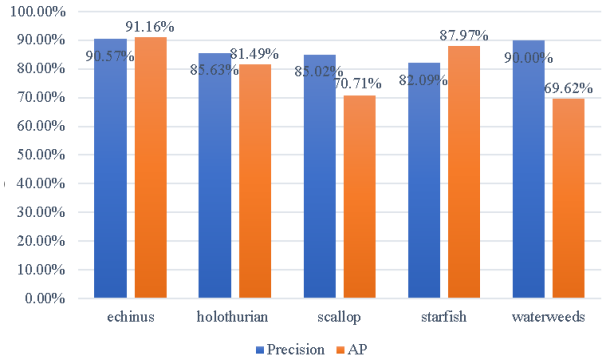

Fig. 3 Effect on every class

It can be seen that in various categories of targets, the accuracy and AP values for echinus achieved the best results. The AP values for scallops and waterweeds are relatively lower compared to other categories. After analysis, we believe that the reason for the excellent performance on echinus is because the dataset contains the largest number of echinus instances, reaching around 22000, and their color features are more distinctive compared to the underwater environment images with predominantly green and blue colors. Therefore, the detection algorithm has learned more prominent features for echinus. The higher accuracy but lower AP values for scallops and waterweeds are due to the fact that there are only around five thousand instances of these categories in the dataset, leading to insufficient learning by the model. However, this does not affect the detection algorithm's accuracy on these two categories. This also indicates that sufficient dataset and data augment is effective.

Model Optimizer Comparison

Apart from the structure of the neural network model, the choice of optimizer also has a significant impact on the effectiveness of the model in deep learning. The core of the optimizer lies in two points. The first one is optimizing the trend. The effectiveness of the model training direction greatly affects the accuracy of the model, which is represented as a gradient in the optimizer[18]. There are many optimizers, such as gradient descent, stochastic gradient descent(SGD), and Adam.

Considering the significant changes in model structure due to the introduction of channel spatial function and SPPFCSPC module, the adaptability of optimizer types also has a significant impact on the final accuracy of the network model. Therefore, comparative experiments were conducted on the following optimizers: SGD, momentum SGD, momentum Nesterov, RMSProp and Adam optimizer. The experiment result is shown in Table 4.

Experiment	optimizer	Accuracy mAP(%)
Exp.1	SGD	73.57
Exp.2	Momentum SGD	76.36
Exp.3	Nesterov Momentum	79.63
Exp.4	RMSProp	80.36
Exp.5	Adam	80.61

Table 4 Comparison on different optimizer

It can be seen from the above table that the Adam optimizer has a better adaptive effect on the improved and optimized network model, followed by the RMSProp, while the SGD has a poor effect. This can be proof from starting training. The other optimizers improved the model very quickly and had good initial accuracy for all five targets in the dataset, while SGD and Momentum SGD has low accuracy on performance. Adam optimizer combines the strengths of both adaptive gradient descent algorithms and momentum gradient descent algorithms. It not only adapts to discrete gradients but also alleviates the problem of gradient fluctuation. The main formulas are as follows:

$$m_t := \text{beta}_1 * m_{t-1} + (1 - \text{beta}_1) * g \quad (4)$$

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

$$v_t := beta_2 * v_{t-1} + (1 - beta_2) * g * g \quad (5)$$

$$variable := variable - lr_t * m_t / \left(\sqrt{v_t} + \epsilon\right) \quad (6)$$

Formula 4 is used to calculate the exponentially smoothed value of historical gradients, obtaining gradient values with momentum. Formula 5 is used to calculate the exponentially smoothed value of its square, obtaining the learning rate for each weight parameter in the network layer. Formula 6 is used to calculate the variable update value, which is proportional to the smoothed historical gradient value and inversely proportional to the smoothed historical gradient square value. Adam optimizer can update variables based on the oscillation of historical gradients and the filtered true historical gradients after oscillation. So experimentally, Adam optimizer is the most suitable for the improved and optimized network.

3. Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Orci a scelerisque purus semper eget duis at tellus at. Quisque egestas diam in arcu cursus. Pulvinar mattis nunc sed blandit. Tempus iaculis urna id volutpat lacus laoreet non curabitur. Morbi tincidunt ornare massa eget egestas purus viverra accumsan in. Vehicula ipsum a arcu cursus. Sapien et ligula ullamcorper malesuada proin. Ut diam quam nulla porttitor. Tincidunt dui ut ornare lectus sit. Neque ornare aenean euismod elementum nisi quis eleifend. Mus mauris vitae ultricies leo integer. In nulla posuere sollicitudin aliquam ultrices. Eget duis at tellus at urna condimentum mattis. Tellus molestie nunc non blandit. Quam quisque id diam vel quam elementum pulvinar. Integer quis auctor elit sed vulputate mi. Pellentesque elit eget gravida cum sociis natoque penatibus et. Aliquet risus feugiat in ante. Commodo ullamcorper a lacus vestibulum sed.In this model, the preprocessed images are first affine transformed to 640×640 size images. Then we put the fined images into the network for training. The architecture of the network we designed is shown in Fig. 4. The original YOLOv7 architecture modules are replaced with the modules in red rectangles as the improvements. First, in order to increase the detection speed in advance, we replaced the FPN(Feature Pyramid Networks) pooling module into SPPFCSPC module. Then we introduced channel spatial attention mechanism for increasing the accuracy of the model after Neck. It improved the feature extraction and integration ability of the model in space and channel dimension of the model. The experiment showed that the accuracy was improved efficiently.

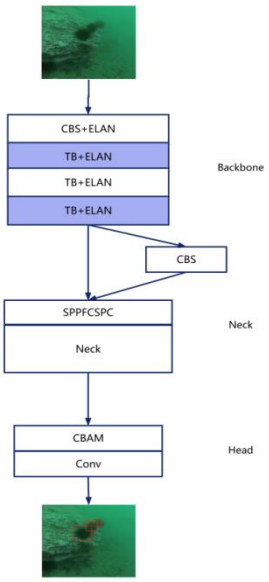

Fig. 4 The architecture of model

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

Space Pyramid Pooling Method

We used SPPFCSPC in replace of the pooling module in the origin network. The architecture is shown in Fig. 5.

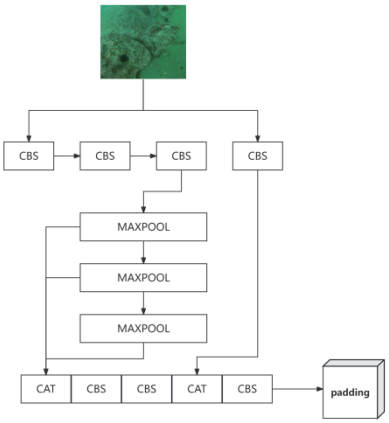

Fig. 5 SPPFCSPC module

The architecture connects the separate pooling operations in the original core module and performs calculations on the basic output results of the smaller pooling kernel. The input of the pooling layer below will be the results of pooling and no pooling. After doing some, and then outputting, the speed is improved while keeping the receptive field unchanged. The specific pooling calculation formulas are as follows:

$$R_1(F) = \text{MaxPool}_{k=5}^{p=2}(F) \quad (7)$$

$$R_2(R_1) = \text{MaxPool}_{k=5}^{p=2}(R_1) \quad (8)$$

$$R_3(R_2) = \text{MaxPool}_{k=5}^{p=2}(R_2) \quad (9)$$

$$R_4 = R_1 \odot R_2 \odot R_3 \quad (10)$$

In these formulas, R is pooling result, P is padding number, K is the size of kernel size and F is the input of convolutional layer. In this way, the pooling operations that process separately originally would be linked together and information of the whole feature map will be completed, making a better connection of contextual information.

Channel Spatial Attention Mechanism

We adopted convolutional block attention module(CBAM) in our neural network. Before detection, we introduced channel attention to extract and analyze important information about feature maps in output net, generalizing high quality bounding boxes and increasing prediction accuracy. The specific experimental details will be introduced in the experimental section. Channel spatial attention mechanism combines channel feature information and spatial feature information, The structure is shown in Fig. 6.

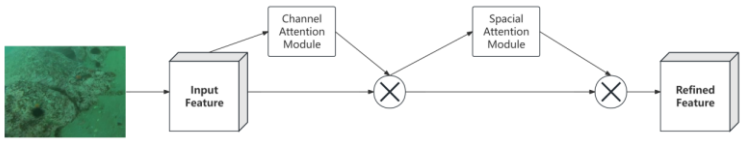

Fig. 6 Channel spatial attention module

The processed feature information of the input image obtained through the backbone feature extraction network and neck network is subjected to channel attention mechanism, which enhances important features and suppresses background features in the channel dimension of the output, obtaining the feature information by spatial attention. Subsequently, the feature information undergoes spatial attention mechanism to calculate the attention weights on

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

both channel and spatial dimensions, resulting in the final output that enhances and suppresses features in both channel content and spatial positions. This output feature map is obtained by applying weights to each element in the feature map, producing a weighted feature map. The weighted feature map is then processed through convolutional layers and pooling layers for further processing, ultimately yielding the model's output result. The formulas are shown as follows.

$$F' = M_c(F) \otimes F \quad (11)$$

$$F'' = M_s(F') \otimes F' \quad (12)$$

Channel spatial attention is basically made with two sub modules. Data is input to channel attention module to compute the importance of feature map in each channel. During the process of training iteration, the weights are obtained and they are used to weight each channel accordingly for improving object feature and decreasing irrelevant feature to this task. Channel attention module contains two parts of full connection layers. The first one aims to lower the dimensions, and the second one aims to higher the dimensions. Channle attention module is illustrated in Fig. 7, which is composed with Maxpool, Avgpool and Shared MLP.

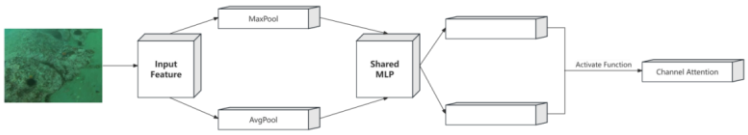

Fig. 7 Channel spatial attention module

For a single input feature layer F(H×W×C), it undergoes Maxpool and Avgpool with the dimension of H×W. And the feature layer is compressed to 1×1×C. After compressed by Maxpool, the one-dimensional vector retains the feature of the original feature layer, and it contains the essential information that distinguishes object feature. And the one-dimensional vector that produced by Avgpool contains the global vision information before compressed, which means it has large receptive field. The result of Maxpool and Avgpool is proposed by shared MLP, and the afterwards output sums up to obtain the channel weight of feature maps. The formula is shown in formula 13. σ represents the sigmoid function. W_0 and W_1 represent the shared full connection layers that constructed to MLP module.

$$\begin{aligned} M_c(F) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F_a vg^c\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \quad (13) \end{aligned}$$

After the channel attention module, images are processed on spatial attention module for spatial feature progress. The module is used to compute the importance of different parts of images, in order to weight different areas. The module is consist of Maxpool, Avgpool and convolutional layer. It is used to extract spatial information from images. The architecture is illustrated in Fig. 8.

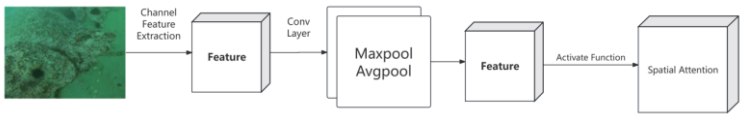

Fig. 8 Spatial attention module

For a single input feature layer F(H×W×C), it undergoes Maxpool and Avgpool. The feature layer are compressed to H×W×1, and the compressed feature layer focuses on the effective information of the spatial region, and is used to extract the efficient information region along the channel. Then, the results of the two are joined on the channel, and then the convolution dimension reduction is carried out to obtain the spatial weight of the feature map, so as to capture the local correlation of the feature information. The formula of spatial attention module is shown as formula 14. σ represents the sigmoid function. And $f^{7\times7}$ represents the convolutional layer of a kernel size with 7×7.

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

$$M_s(F) = \sigma(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])) = \sigma\left(f^{7\times7}\left(\left[F^S_{avg}; F^S_{max}\right]\right)\right) \quad (14)$$

The channel-spatial attention mechanism can adaptively focus on different parts of the input image by calculating channel and spatial attention, thereby improving the accuracy of the model. Compared with ordinal CNN architecture, channel spatial attention dig better on image information, concentrating on essential feature while network extraction. The mechanism weights different areas of images, reducing the influence of noise and reluctant information, in order to adapt different input data in the complex environment underwater with the trait of much noise. It can reduce the excessive dependence on training data, thereby improving the generalization ability of the model.

4. Results

To validate the superiority of the optimization-improved detection network algorithm proposed in this paper on underwater environment images, the detection algorithm in this paper is compared with mainstream object detection algorithms such as Efficientdet, Faster-RCNN, SSD, YOLOv3, YOLOv4, YOLOv5 and YOLOv7 on the same dataset. The image data are adjusted to the required input size without distortion according to the model requirements. A uniform set of training parameters is used to train the models, and each model is trained to convergence. A comparison analysis is conducted based on two metrics: mean Average Precision (mAP) and model detection speed (Frames Per Second, FPS). The result is shown in Table 5.

Model	Image size	Accuracy mAP(%)
Efficientdet-D1	640*640	51.32
Faster-RCNN-VGG	600*600	59.41
Faster-RCNN-Resnet	600*600	62.62
SSD	300*300	52.46
YOLOv3	416*416	65.4
YOLOv4-tiny	416*416	52.19
YOLOv4	416*416	61.69
YOLOv5-s	640*640	72.47
YOLOv5-l	640*640	77.17
YOLOv7-tiny	640*640	72.73
YOLOv7	640*640	78.93
Ours	640*640	80.61

Table 5 Comparison on mainstream detection model

From the result it can be seen that, from the point of view of the accuracy of detection, the improved YOLOv7 meets 80.61\% on mAP. The experiment indicates the improved algorithm has more advantage on underwater object detection tasks. From the point of view of detection speed, compared with the model of the same scale, the detection speed of the improved model is maintained at the medium level, and it has good real-time performance. Compared with the main-stream object detection model Faster-RCNN-Resnet, the improved YOLOv7 is 17.99% higher on accuracy. Compared with the most widely-used one-stage detector YOLOv5, its mAP is 3.44% higher. Compared with YOLOv7, mAP rises by 1.68%. It is shown that the improved algorithm is a high detection precision algorithm, and the discriminant accuracy of the algorithm is improved. Improved training experiments were conducted for 2000 epochs. From the loss function and average precision obtained during the algorithm training, it was found that the network model reached a converged state at around the 700th epoch. Additionally, the model achieved optimal network model structure weight parameters near the 700th epoch, resulting in a mAP of 80.61%. The performance of the model was significantly enhanced compared to the previous one, reaching higher levels of average precision and accuracy. The improved model demonstrates better performance in underwater target detection, effectively addressing challenges such as target overlap, complex and blurry backgrounds in underwater scenes. The details are illustrated by Fig. 9.

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

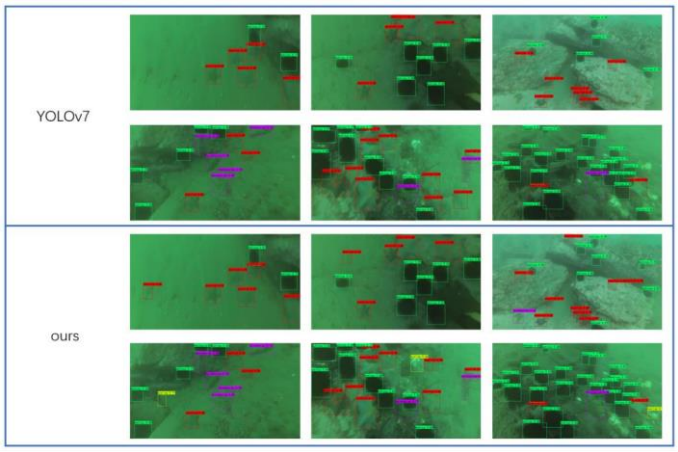

Fig. 9 Actual effect of the detection

In summary, compared with various mainstream algorithms, our improved algorithm shows the highest detection accuracy and has a moderate detection speed, which has significant advantages in underwater tasks with high detection accuracy.

5. Discussion

The optimization network algorithm proposed is aiming at improving the speed and accuracy of underwater images. The structure optimization is made to improve the detection efficiency and feature extraction ability of the network. Firstly, the principle and advantages of the SPPFCSPC module and the channel space attention mechanism are introduced into the network model . Secondly, the SPPFCSPC module is introduced and the performance comparison experiment is carried out to link the original three independent pooling layers together. While keeping the perception field of the algorithm unchanged, the calculation amount of parameters inside the algorithm is reduced. Then, the detection accuracy of the detection algorithm is improved, and the channel space attention mechanism is introduced to enhance the extracted features in the channel and spatial dimensions, which avoids the information loss caused by the feature extraction and improves the overall feature extraction capability of the network. The YOLOv7 model, which integrates the attention mechanism, has higher detection accuracy in the underwater target scene. The compatibility test of these two parts is carried out, and the performance test is compared with the mainstream algorithm, which fully proves the advantages of the improved underwater detection algorithm, and effectively improves the detection accuracy of the detection algorithm.

Refrences

[1] Yu, Y.; Guo, B.; Chu, S.; Li, H.; Tang, P. Suvery of underwater biological object detection methods based on deep learning. 423 Shangdong Science 2023, 36, 1–7.

[2] Sankar A, Jacob J. Exploration of Beamforming Approach to Enhance the Detection Rate of Underwater Targets in Distributed Multiple Sensor Systems[J]. Smart Science, 2020, 8(4): 227-241.

[3] LU CY，JIA F G，YU L M, A Review of Underwater Image Target Detection Research[J]. Digital Ocean and Underwater Warfare, 2023,6(01):34-40.

[4] Moniruzzaman M, Islam S M S, Bennamoun M, et al. Deep learning on underwater marine object detection: A survey[C]//Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18. Springer International Publishing, 2017: 150-160.

[5] Qian X Q, Liu W F, Zhang J, Cao Y. Underwater-relevant image object detection based feature-degraded enhancement method[J], Journal of Image and Graphics, 2022, 27(11): 3185-3198.

[6] Velit laoreet id donec ultrices tincidunt arcu non sodales neque. Non curabitur gravida arcu ac tortor dignissim convallis aenean et.

International Water and Irrigation
Volume 44 Issue 1 (2025)
ISSN: 0334-5807
Pages 265 - 277

[7] Luo Y H, Liu Q P, Zhang Y et al. Review of Underwater Image Object Detection Based on Deep Learning[J]. Journal of Electronics & Information Technology, 2023, 45(10): 3468-3482
[8] KNAUSGÅRD K M, WIKLUND A, SØRDALEN T K, et al. Temperate fish detection and classification: A deep learning based approach[J]. Applied Intelligence, 2022, 52(6): 6988–7001. doi: 10.1007/s10489-020-02154-9.
[9] Ye Z, Duan X, Zhao C. Research on Underwater Target Detection by Improved YOLOv3-SPP[J]. Computer Engineering and Applications, 2023, 59(6): 231–240.
[10] Zaidi S S A, Ansari M S, Aslam A, et al. A survey of modern deep learning based object detection models[J]. Digital Signal Processing, 2022, 126: 103514.
[11] Zhang Y, Li X X, Sun Y M et al. Underwater object detection algorithm based on channel attention and feature fusion[J]. Journal of Northwest Polytechnical University, 2022,40(2):433–441.doi:10.3969/j.issn. 1000-2758.2022.02.025.
[12] Zhang L, Zhang L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(2): 270-294.
[13] Etienne A, Ahmad A, Aggarwal V, et al. Deep learning-based object detection system for identifying weeds using uas imagery[J]. Remote Sensing, 2021, 13(24): 5182.
[14] Aberman K, He J, Gandelsman Y, et al. Deep saliency prior for reducing visual distraction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 19851-19860.
[15] Hafiz A M, Parah S A, Bhat R U A. Attention mechanisms and deep learning for machine vision: A survey of the state of the art[EB/OL]. (2021-6-3) [2024-7-1] https://doi.org/10.48550/arXiv.2106.07550.
[16] Hu J, Shen L, Sun G. Squeeze-and- excitation networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[17] Wang L, Zhang X, Su H, et al. A comprehensive survey of continual learning: theory, method and application[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
[18] Bharati P, Pramanik A. Deep learning techniques—R-CNN to mask R-CNN: a survey[J]. Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019, 2020: 657-668.
[19] Lin W H, Zhong J X, Liu S, et al. Roimix: proposal-fusion among multiple images for underwater object detection[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 2588-2592.