Improvements Towards the Sonar Image Dataset for Yolov7

Guohao Xie[1], Jianxun Tang[1], Zhe Chen[2,3(✉)], and Mingsong Chen[1,2]

[1] School of Ocean Engineering, Guilin University of Electronic Technology, Beihai 536000, China

[2] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China
chenzhe@mail.nwpu.edu.cn

[3] Cognitive Radio and Information Processing Key Laboratory Authorized By China's Ministry of Education Foundation, Guilin University of Electronic Technology, Guilin, China

Abstract. Sonar imaging technology has been continuously improving, leading to its widespread use in recognizing underwater targets. However, sonar images often suffer from low contrast, blurred edges, and high noise, which can make it difficult to extract target information during deep learing image feature extraction. This can result in the loss of target features and ultimately affect recognition accuracy. To address the issue at hand, we propose the addition of dynamic ODConv to the original yolov7 model. This will help tackle the problems of error and omission detection wuring complex background extraction and target feature loss in the feature recognition process of sonar images.By incorporating a channel attention mechanism and activation function that are more attuned to spatial features, the feature extraction and target recognition process can avoid the issue of target feature loss, ultimately leading to improved recognition accuracy.

Keywords: SEnet attention · FReLU · ODConv · sonar image target recognition

1 Introduction

In order to build a strong marine state, marine science and technology innovation is a fundamental driving force. It plays a decisive role and runs through the entire situation, accelerating the process of marine development and revitalizing the marine economy. Science and technology are the key factors in achieving this goal.The 20th Party Congress report proposed the implementation of a science and education strategy to improve the innovation system of science and technology. It also suggested accelerating the implementation of an innovation-driven development strategy to make special arrangements. This would allow for the seizing of a historical opportunity through the cross-combination of deep-sea exploration technology and artificial intelligence. This approach would be used to compete for the high ground of the future international strategic pattern. As the process of ocean power advances, several challenges have emerged, with underwater target identification in sonar images being a significant technical bottleneck that limits the development of underwater information countermeasures [1].

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 57–67, 2025.
https://doi.org/10.1007/978-3-031-60347-1_6

The field of underwater target recognition from sonar images has a broad development space with the rapid development of deep learning. The emerging deep learning method [2] can effectively improve the generalization ability of recognition models and accomplish end-to-end information transfer and autonomous recognition.This study highlights the relevance of underwater target detection, identification, and tracking in abstract learning and heterogeneous features. However, these tasks face several challenges, such as enhancing the accuracy and speed of detection and recognition, minimizing the cost of computation and communication, and reducing the model's complexity. Addressing these issues is crucial for improving the overall performance of underwater target detection systems [3].

In the field of deep learning, the Yolo [4–6] algorithm has been recognized as a significant milestone. Leilei Jin and colleagues [7] have developed a model for recognizing underwater targets using a CNN neural network, which incorporates image salient region segmentation and pyramid pooling based on sonar image characteristics. This model has significantly improved the speed and accuracy of target image prediction. Wang et al. [8] utilized the bilinear interpolation preprocessing technique in combination with the YOLO v3 network to achieve successful sonar image target recognition. Similarly, Sheng et al. [9] addressed the issue of insufficient sonar image data by utilizing a combination of simulation and real samples. They ultimately employed the YOLO v3 network for detection and achieved favorable results. Additionally, the effectiveness of simulation data was also demonstrated.In their study, Jin et al. [7] were able to improve the accuracy of sonar image recognition by enhancing the significant region segmentation and pyramid pooling in YOLO v3, thereby reducing the impact of background noise on feature extraction. Meanwhile, Zhang et al. [10] utilized the transfer learning method to introduce IoU as a distance function in the YOLO v5 model. They also improved the traditional K-means algorithm in obtaining a priori anchor frames, leading to an enhancement in target detection performance. Wenguan Zeng [11] For the sonar image target recognition task, five different CNN network models were designed, and the effects of different network layers, convolution kernel size and activation functions on the sonar image recognition results were studied and compared in the experiments; for the sonar image target detection task, an improved FW-YOLOv3 algorithm model was proposed, and a single-scale target prediction network was designed in the new algorithm, and a A new weighted feature fusion algorithm is proposed. It is shown that the single-scale prediction network can improve the detection speed of the model, and the weighted feature fusion algorithm can effectively improve the detection accuracy of the model.

This literature is based on recent studies of sonar image detection methods and their results. While the models discussed in the literature have shown high detection speed and accuracy, further research is necessary to address the challenge of merging noise and target information during the feature extraction process of sonar images. This challenge currently impedes the improvement of recognition accuracy.

This paper proposes using ODConv to replace conv for better background extraction and target feature loss during sonar image feature recognition. The attention mechanism is utilized to model the importance between individual features, and for different tasks, features can be assigned according to the input for feature assignment, which is a simple

and effective solution to address error and omission detection problems. The current models are capable of identifying dynamic features assigned to the convolutional kernel through one dimension of the convolutional kernel space while disregarding the other three dimensions. To address this limitation, the ODConv utilizes multidimensional attention mechanism and parallel strategy to consider features of multiple dimensions, which effectively enhances the model's performance.

To address the issue of target feature loss in sonar image recognition due to low contrast, blurred edges, and high noise, an attention mechanism can be added. This paper proposes the use of the SE attention mechanism in different modules based on the original yolov7. The SE module is a simple and easy-to-implement solution that can be easily integrated into existing network model frameworks. SENet [12] filters out the correlation between the main learning channels, which leads to a slight increase in computational effort. However, this process improves the feature extraction capability by creating a global sensory field for the features close to the data input.

Further the original activation function is insensitive to the spatial information of the context, which is aggravated by the blurred target and background boundaries of the sonar images, so the FReLU activation function is introduced for the image recognition task, which extends ReLU and PReLU to 2D activation functions by adding negligible spatial conditional overhead. The ability to implement pixel-level spatial information modeling in the activation function stage can be used simply and efficiently for target recognition tasks such as target detection and semantic segmentation.

2 Correlation Principle Design

2.1 YOLOv7 Algorithm Overview

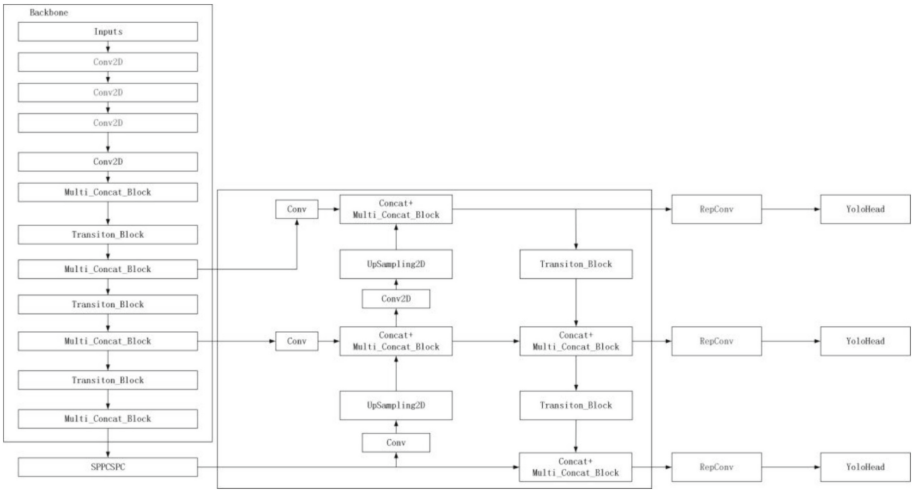

Fig. 1. YOLOv7 structure diagram

Figure 1 displays the detailed architecture of the improved model based on YOLOv7. The input image is initially extracted within the back-end network, producing a set of features known as feature layers. Three effective feature layers are then obtained for the subsequent step of network construction.The FPN module merges the three effective feature layers obtained in the back-end to combine feature information at different scales. These layers are then used for additional feature extraction in YoloV7. The combination of Backbone and FPN allows for the creation of three improved, effective feature layers in YoloV7. Each feature layer is defined by its width, height, and number of channels, as well as three prior boxes for each feature point and a specific number of feature channels for each prior box. The decoupling head utilized in YoloV7 is similar to previous versions of Yolo, in which classification and regression are accomplished through 1×1 convolution.

The REP module is employed to modify the number of image channels for the output features of varying scale sizes. These features are then transformed into bounding box, category, and confidence information. The convolution layer is utilized as the detection head for downsampling to achieve multi-scale detection of targets of different sizes, including large, medium, and small ones.

Although the YOLOv7 algorithm has excellent detection accuracy and detection speed, there are still many problems when it is directly applied to sonar image detection. Therefore, this paper is based on The YOLOv7 algorithm is used as the basis for improvement.

2.2 ODConv

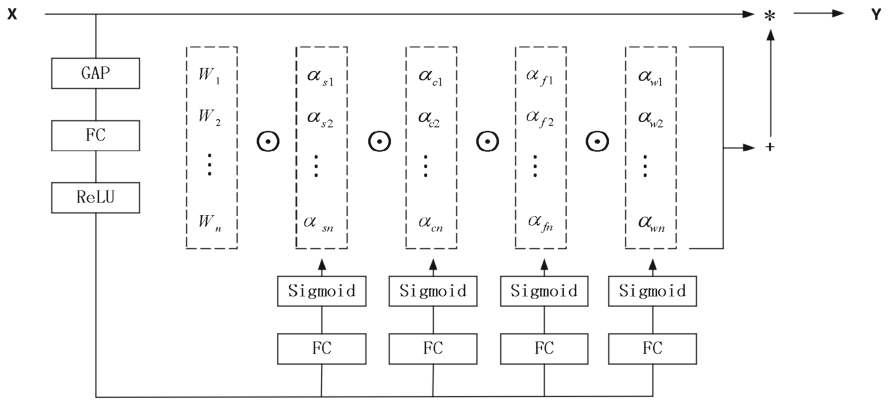

Fig. 2. ODConv

The original yolov7 model demonstrated errors and missed detections in processing sonar images, likely due to complex background issues (Fig. 2). To address this, the use of ODConv (as shown in Fig. 5) was explored to learn a complementary attention kernel space using a multidimensional attention mechanism through parallel strategy. This approach aimed to improve the model's feature extraction capability.

ODConv introduces a multidimensional attention mechanism through a parallel strategy to learn more flexible attention on the four dimensions of the convolutional kernel space, ODConv introduces three new attentions, and it uses a multidimensional attention mechanism to learn complementary attentions along the four dimensions of the kernel space through a parallel strategy, which can improve detection performance, especially for small-sized targets.

2.3 SE Attention Module

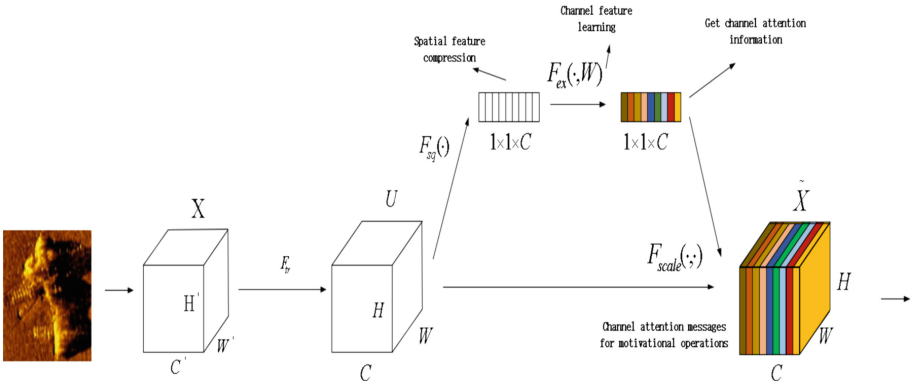

Fig. 3. Structure of the SE Attention Module

In this paper, we propose the use of SE, a pre-built attention module, which can be easily added to any part of the YOLOv7 network. Specifically, we incorporate the SE module into the C2 module, resulting in three feature layers with enhanced attention mechanisms. Unlike the original C2 module, this approach prevents the loss of target features during feature extraction and recognition, while also enabling a global perceptual field for features near the data input. Overall, our proposed method improves the network's feature extraction capability.

The action process of SE attention mechanism is shown in Fig. 3:

①. Starting from the input single image, extract the current image features, and set the feature map dimension of the current feature layer U as [C, H, W]

②. The average pooling or maximum pooling operation is performed on the two dimensions of [H, W] of the feature map, and the size of the pooled feature map is from[C, H, W] -> [C, 1, 1]. [C, 1, 1] is for each channel C, there is a number corresponding to it one by one.

③. For [C, 1, 1] the features can be understood as the weights extracted from each channel itself. The weights represent the influence of each channel on the feature extraction, and the vector after global pooling is passed through the MLP network to obtain the weights of each channel.

④. The weights of each channel C [C, 1, 1] are obtained, and the weights are applied to the feature map U [C, H, W], i.e., each channel is multiplied by its respective

weight. When the weight is large, the value of that channel feature map increases accordingly and the impact on the final output becomes larger; when the weight is small, the value of that channel feature map becomes smaller and the impact on the final output becomes smaller.

2.4 FReLU

On the other hand, due to the original features such as low resolution of underwater sonar images and small detection objects, the native head layer of YOLO v7 loses fine information and retains invalid background features after the original convolution or clustering process, and the native SiLU activation function is insensitive to contextual and spatial information, which seriously hinders the image recognition accuracy problem. Replacing the SiLU activation function with FReLU activation function converts the activation function from one-dimensional to two-dimensional by introducing contextual relationships, which is more beneficial to extract spatial information in addition to its own image information.

The activation function solves the spatial insensitivity problem in the activation function by converting the activation function from 1D to 2D mainly through the introduction of contextual relations, so that the regular (ordinary) convolution also has the ability to capture complex visual layouts and make the model capable of pixel-level modeling. The spatial dependency condition T(X) is created by adding a parameter pooling window to effectively represent the spatial contextual feature extractor. The flow is shown in Fig. 4, and the detailed implementation is shown in Eqs. 1 and 2

$$f(x_{c,i,j}) = max(x_{c,i,j}, T(x_{c,i,j})) \tag{1}$$

Equation 1, the body of the function is chosen from the same simple nonlinear function max(·), the function max(·) gives each pixel a choice of looking at the spatial background or not, **and** $x_{c,i,j}$ represents the input pixel of the nonlinear activation $f(\cdot)$ on the cth channel.

$$T(x_{c,i,j}) = x^{\omega}_{c,i,j} \cdot p^{\omega}_{c} \tag{2}$$

Equation 2, the function $T(\cdot)$ is the Funnel condition, $x^{w}_{c,i,j}$ denotes the parameterized pool window centered on the input pixel of the nonlinear activation function $f(\cdot)$ on the cth channel at the 2D spatial location (i, j), p^{w}_{c} denotes the coefficients shared in the same channel on this window, and $(\cdot)$ denotes the dot product operation.

2.5 Improved YOLOv7 Model

This paper presents the modified YOLOv7 model network structure, which is illustrated in Fig. 5. The ODConv is used to replace the convolutional layer in the neck of the network, which helps to capture richer context and focus more on the target information. Furthermore, the SE attention mechanism is integrated into the three effective feature layers for feature extraction and fusion, allowing the neural network to automatically distinguish important and unimportant channels. To enhance the feature extraction capability, appropriate weights are assigned to amplify important features and suppress unimportant ones. Finally, the F-ReLU function is adopted as the activation function of the

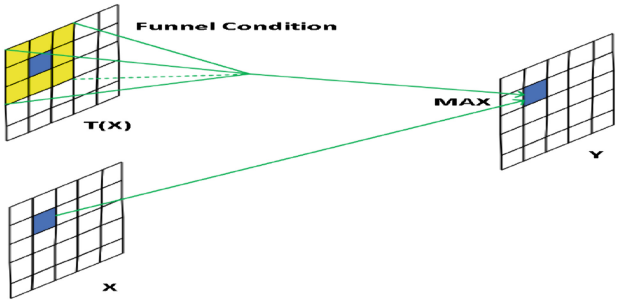

Fig. 4. FReLU calculation flow

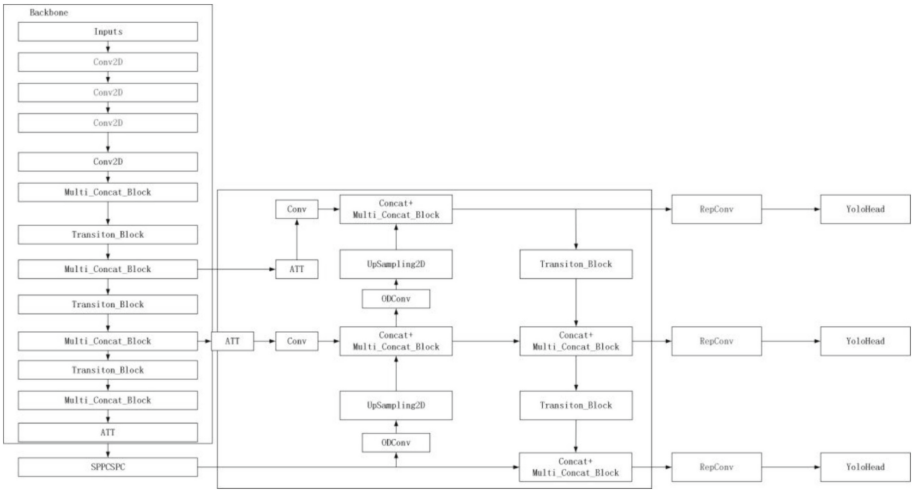

Fig. 5. Improved YOLOv7 model

model convolution module, which expands the sensory field range of the convolutional network and improves the accuracy of target recognition.

3 Experimental Results and Analysis

In order to verify the effectiveness of the model, comparative experiments were conducted on the SCTD-A dataset to validate the detection of the model. The operating system is Windows 11, the deep learning framework is Pytouch 1.4.0, the CPU is Intel Core i7 12700H, the RAM is 32 GB, and the GPU is NVIDIA GeForce GTX 3070TI.

3.1 Experimental Data

To conduct image processing and identification research in this paper, we collected the STD-A dataset showcasing sonar imaging results of different targets underwater in

various sea trials. These sonar images were filtered and sorted into categories, resulting in sonar images of three real target categories: human, ship, and aircraft. We selected three images of each type for display, as shown in Fig. 6.

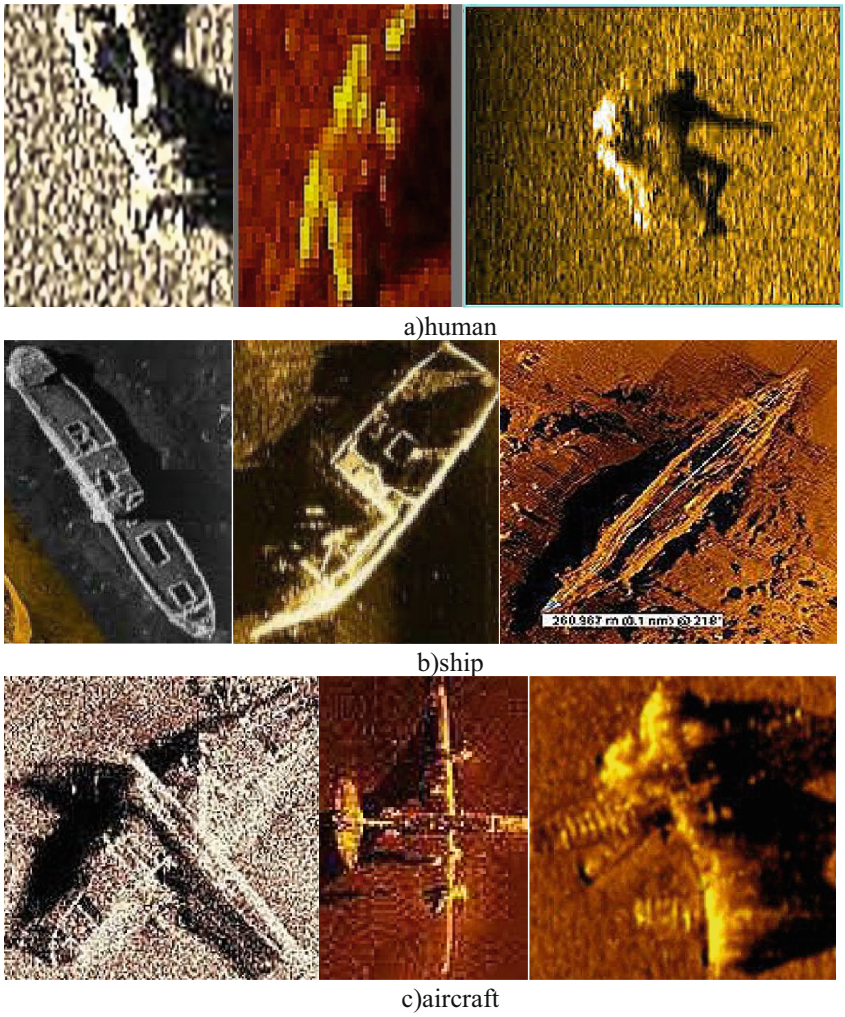

Fig. 6. Sonar image dataset

The dataset consists of 1620 sonar images, each representing one of three types of underwater targets. These images will be used as input to the deep learning model. The target images are labeled separately into three classes through manual annotation. The training and validation sets are split in a 9:1 ratio for training purposes.

3.2 Sonar Image Recognition Results

We use map values as the experimental evaluation criteria. *mAP* is an indicator of recognition accuracy in object detection. In object detection for multiple categories, each category can draw a curve based on recall and precision. *AP* is the area under the curve, and *mAP* is the average value of multiple categories of *AP*.

To validate the performance of the method and compare it with the detection method, the evaluation index is *mAP*. The mainstream methods such as yolov3.yolov5.yolov7 and the latest methods were selected for comparison. The experimental results are shown in Table 1, and the method has the highest value in *mAP*.

Table 1. Experimental data of each model in the SCTD-A dataset

Method	AP_1	AP_2	AP_3	*mAP*
YOLO v3	0.996	1	0.926	0.974
YOLO v5	0.997	1	0.925	0.974
YOLO v7	0.989	1	0.921	0.970
YOLO v7+FReLU	0.990	1	0.923	0.971
YOLOV7+SE	0.987	1	0.947	0.978
YOLOV7+ODConv	0.996	1	0.950	0.982
YOLOV7+SE+FReLU	0.991	1	0.949	0.980
YOLOV7+ODConv+FReLU	0.993	1	0.956	0.983
YOLOV7+SE+ODConv	0.998	1	0.972	0.990
YOLOV7+ODConv+SE+FReLU	1	1	0.974	0.991

AP_1, AP_2, AP_3 denote the detection accuracy of aircraft, human, and ship targets, respectively *AP*.

As can be seen from Table 1, the improved yolov7 model has improved in experimental accuracy precision by comparison tests with the mainstream network yolov3.v5.v7.The model with dynamic convolution and attention mechanism has been improved to enhance the feature learning ability. The F-ReLU function, serving as the activation function of the convolution module, strengthens the context learning ability of the convolution layer without increasing the number of model parameters and computation. This improvement has resulted in an increase in the accuracy of the model in three ways.

The detailed variations of the values of these three categories are shown in Fig. 8 and the detailed value pairs are shown in Fig. 7. Utilizing ODConv, a multidimensional attention mechanism can be employed through a parallel strategy to learn the four dimensions of the complementary attention kernel space, resulting in improved detection performance. This attention mechanism enhances the neural network's activation function, allowing for the extraction of spatial information by introducing contextual relationships. This conversion of activation function from one-dimensional to two-dimensional

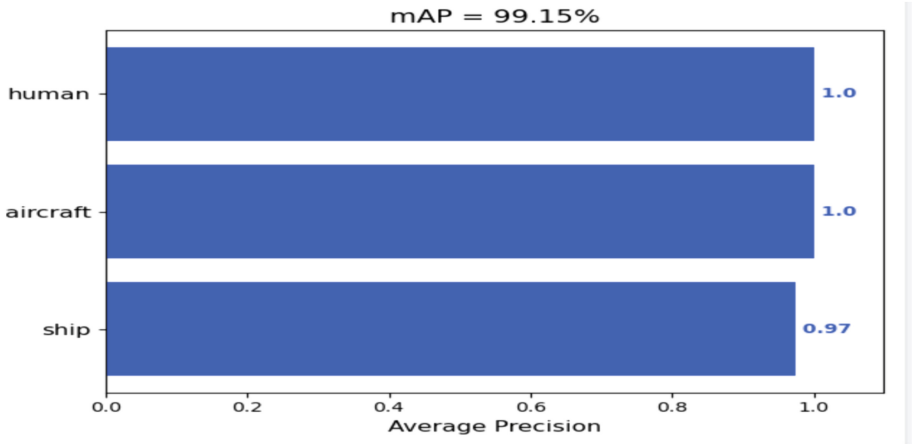

Fig. 7. Model testing on the SCTD-A test set *mAP* Results

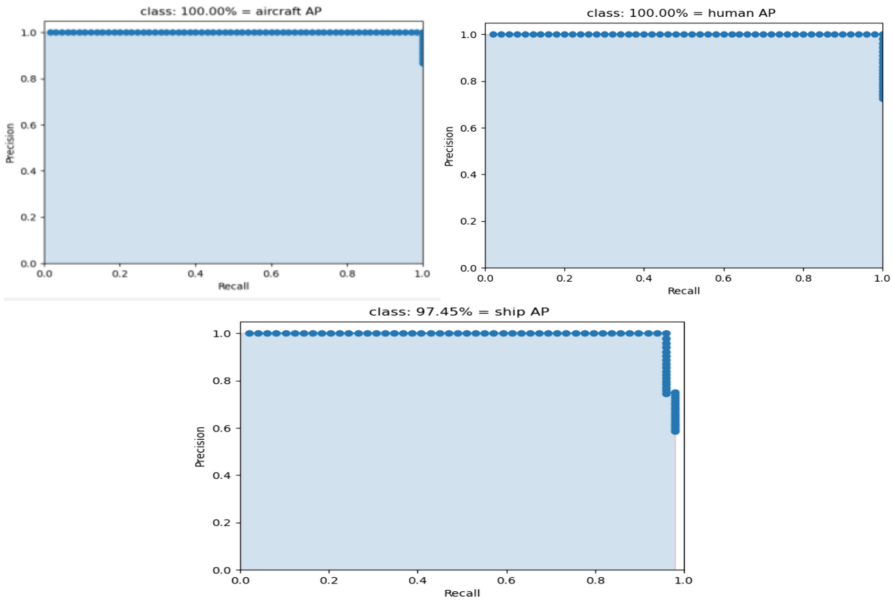

Fig. 8. Model testing on the SCTD-A test set *AP* results: (a) *AP* results for aircraft, (b) *AP* results for human, (c) *AP* results for ship

facilitates the processing of image information. It is easy to see from Figs.7 and 8 that the improved model has improved the effect on marine sonar image recognition.

4 Conclusion

To enhance the precision of sonar image recognition, this study proposes a detection algorithm that utilizes an improved version of YOLOv7. The algorithm addresses the challenges of complex background feature extraction and target feature loss in sonar image recognition by utilizing ODconv to extract multidimensional features. This approach effectively enhances the model's performance. The model is enhanced by incorporating the SE attention mechanism, which improves the global nature of the feature extraction process. Additionally, the FReLU activation function is introduced to enhance the learning capability of the convolutional layer for contextual features.

The purpose of this paper's algorithm is to enhance the accuracy of the model's detection capabilities without introducing overly complex modules. The algorithm has been successfully applied in engineering to achieve high-precision recognition of sonar images. Future work will focus on improving the model's robustness and detection speed, which will increase its practicality for real-time use.

Acknowledgement. This research was supported by the Special Program of Guangxi Science and Technology Base and Talents under Grant No. AD21220098, the Innovation Project of Guangxi Graduate Education (YCSW2022289) and the Innovation Project of Guangxi Graduate Education (YCSW2023329).

References

1. Tan, P., Wu, X., Zhang, X., et al.: A review of research on underwater target recognition based on sonar images. In: Digital Ocean and Underwater Attack and Defense, vol. 5, no. 04 (2022)
2. Xu, Huang, Z., Chen, L., et al.: Advances in deep learning for passive recognition of underwater targets. Signal Process. **35**(9), 1460–1475 (2019)
3. Review of research on underwater target detection, identification and tracking based on sonar images
4. Redmon, J., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
5. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
6. Redmon, J., Farhadi, A.: Yolov3: An Incremental Improvement (2018). arXiv preprint arXiv: 1804.02767
7. Jin, L., Liang, H., Yang, C.: Sonar image recognition of underwater target based on convolutional neural network. J. Northwest. Polytech. Univ. **39**(2), 285–291 (2021)
8. Wang, X., Guan, Z., Wang, J., Wang, Y.: Target detection of color image sonar based on convolutional neural network. J. Comput. Appl. **39**(S1), 187191 (2019)
9. Sheng, Z., Huo, G.: Detection of underwater mine target in sides-can sonar image based on sample simulation and transfer learning. CAAI Trans. Intell. Syst. **16**(2), 385–392 (2021)
10. Zhang, H., Tian, M., Shao, G., et al.: Target detection of forward-looking sonar image based on improved YOLOv5. IEEE Access **10**, 18023–18034 (2022)
11. Zeng, W.: Convolutional Neural Network Based Sonar Image Recognition and Detection. Dissertation (2022)
12. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-Excitation Networks. Journal version of the CVPR 2018 paper, accepted by TPAMI