Small Target Underwater Sonar Image Target Detection Based on Adaptive Global Feature Enhancement Network

Kun Zheng[1], Zhe Chen[2], Jianxun Tang[3], and Jun Kit Chaw[1](✉)

[1] Institute of IR4.0, National University of Malaysia (UKM), 43600 Bangi, Selangor, Malaysia
chawjk@ukm.edu.my

[2] School of Information and Communication, Guilin University of Electronic Technology, Guilin, China

[3] School of Ocean Engineering, Guilin University of Electronic Technology, Guilin, China

Abstract. The utilization of sonar imaging for detecting underwater targets is crucial for both maritime trade and military protection at sea. Current target detection methods employing underwater sonar images primarily rely on traditional machine learning techniques. These approaches necessitate extensive a priori knowledge and face difficulties in adapting to the low resolution of sonar images, blurry target edges, and severe distortion. Consequently, target recognition accuracy is suboptimal, particularly for detecting smaller targets. To address these challenges, this study introduces an adaptive global feature enhancement network for small target detection in underwater sonar images, building upon the original YOLOv7 model. To achieve richer gradient information, the focus of the network is directed to the target rather than the background by parallelizing additional gradient flow branches, we integrate the C2f feature extraction module of YOLOv8 into Multi-Headed Self-Attention (MHSA) to form a new model C2fMHSA.Subsequently, SPD-Conv is incorporated to preserve fine-grained information and minimize the impact of less-effective features during the convolution process. This adaptation allows the model to accommodate the low-resolution nature of sonar images and the high prevalence of small targets. Lastly, by replacing the original Intersection over Union (IoU) metric with Normalized Wasserstein Distance (NWD), the model's sensitivity to small targets is decreased, effectively enhancing the network's detection results demonstrate that the proposed model significantly outperforms three baseline models in terms of recognition accuracy, highlighting the value of this innovative approach to underwater target detection.

Keywords: Underwater Target Detection · YOLOv7 · Sonar Image · Deep Learning

1 Introduction

With the ongoing advancement of human exploration into the ocean's depths, underwater target detection has emerged as a prominent research area, boasting an array of diverse applications. These span various domains, including industry and everyday life,

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 35–47, 2025.
https://doi.org/10.1007/978-3-031-60347-1_4

where fish and underwater reef positioning are vital; scientific research, where underwater environment monitoring and bathymetry play a significant role; and defense and military sectors, where the identification of underwater torpedoes and submarine targets is of paramount importance. Nonetheless, underwater sonar imaging is challenged by the inherent complexity of hydroacoustic channels and the sound wave propagation's attenuation and scattering. As a result, sonar images often exhibit prominent noise, considerable distortion, blurry target boundaries, and low resolution. Conventional identification and detection approaches primarily rely on manual interpretation and pattern recognition. Unfortunately, these methods necessitate extensive a priori knowledge, suffer from low efficiency and subjectivity, and struggle to adapt to the intricacies of underwater targets. Consequently, they fall short of addressing the present-day practical application requirements effectively.

In recent years, with the booming development in the field of computer vision, deep learning-based target recognition methods have gradually become mainstream. McKay J [1] et al. have proposed a method to incorporate convolutional neural networks into sonar image detection and recognition by using migration learning for multi-instance target detection and recognition in sonar data sets. Abu Avi [2] et al. proposed a Constant False Alarm Rate(CFAR) detection algorithm to detect 270 real sonar images in different environments, and the results showed the superiority of the algorithm over existing algorithms in terms of receiver operating characteristic curves.Santos MMD et al. [3] proposed an algorithm combining image processing techniques with convolutional neural networks to integrate satellite images and underwater sonar images in the detection process to aid underwater navigation in partially structured environments.Gu Jet al. [4] proposed an algorithm that can be applied to ROV target detection, which uses high-frequency forward-looking imaging sonar for automatic detection and recognition, and segmentation of underwater target echo shapes or acoustic shadow shapes; Tang et al. [5] used Faster R-CNN networks to achieve automatic detection of wreck targets on the seabed with side-scan sonar. However, to address the problems of complex structure and low training and detection efficiency of this model, KONG et al. [6] proposed a YOLOv3-dpfin network architecture based on YOLOv3 for sonar datasets with small effective samples and low signal-to-noise ratio, which extracted effective features of sonar images through a dual-path network module and a fusion transition module. CAO et al. [7] proposed a network based on YOLOv3 to provide reliable information for Autonomous Underwater Vehicle (AUV) obstacle avoidance by detecting obstacles in sonar images. Tang et al. [8] proposed an improved YOLOv3 model based on migration learning for side-scan sonar wreck target detection, but the performance of target detection in sonar images remains poor due to the non-adaptive nature of the YOLOv3 model. Finally, Tang [9] significantly improved the detection of side-scan sonar wreck targets by enhancing the loss function and detection frame of the YOLOv5 model. Although the training and detection efficiency were improved to a certain extent, there were still problems such as high missed alarm rate for small targets and the detection speed could not meet the real-time requirements.

To address the problems of existing models, this paper proposes a small target underwater sonar image target detection model based on an adaptive global feature enhancement network. Firstly, the C2fMHSA module is designed to obtain richer gradient information by parallelizing more gradient stream branches while making the network focus more on the target rather than other unimportant information such as the background. Then, by introducing SPD-Conv [10], the fine-grained information and less effective features are retained during the convolution process, adapting to the low-resolution nature of sonar images and the large proportion of small targets. Finally, by introducing Normalized Wasserstein Distance (NWD) [11] to replace the original metric of Intersection over Union (IoU), the sensitivity of the model to small targets is reduced, thus effectively improving the detection performance of the network.

2 Related Work (YOLOv7)

The model structure of YOLOv7 [12] consists of three parts: Input, Backbone and Head, which are shown in Fig. 1.

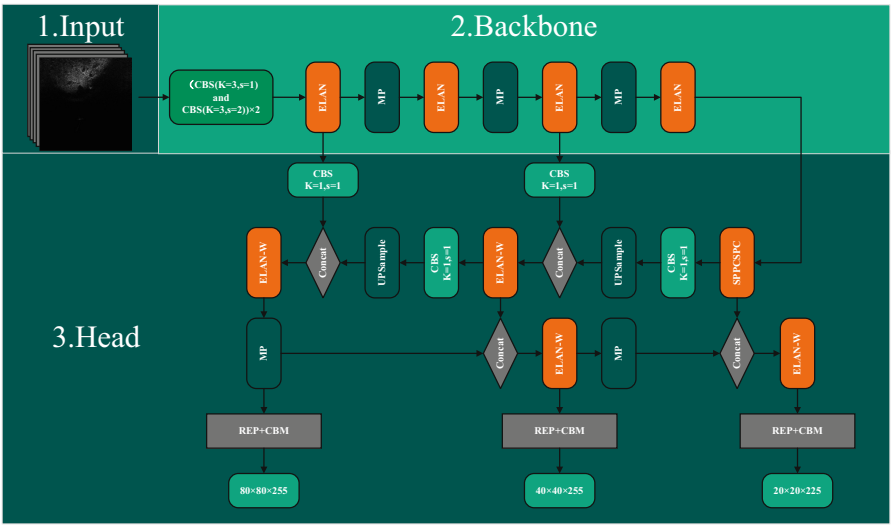

Fig. 1. Model structure of YOLOv7.

The input module primarily employs the Mosaic data augmentation technique, wherein four images from the original dataset are initially flipped, scaled, and subject to color gamut adjustments. These images subsequently undergo random cropping and are stitched together, yielding a new image that serves as the model's input sample. This strategy not only enhances the detection dataset, particularly by increasing the number of small targets, but also bolsters the model's robustness. During Batch Normalization (BN), the data from the four images are processed simultaneously, significantly improving the efficiency of model training.

The backbone, which serves as the feature extraction module of the entire target detection network, comprises four Convolutional Batch-normalization Sigmoid linear Unit (CBS) layers, four Efficient Layer Aggregation Network (ELAN) layers, and three Max-Pooling (MP) layers. The four CBS layers consist of two convolutional layers with a kernel size of 3×3 and stride size of 1 (primarily for feature extraction) and two convolutional layers with a kernel size of 3×3 and stride size of 2 (mainly for downsampling). These layers enable the rapid acquisition of the input image's original shallow features. ELAN has two branches: the first branch involves a 1×1 convolution to alter the number of channels, while the second branch is more complex, also incorporating a 1×1 convolution module for channel count adjustment. The MP module is instrumental in downsampling and optimizing the features obtained from the first branch, which undergoes a max-pooling operation for this purpose, followed by a 1×1 convolution to change the number of channels. The second branch also goes through a 1×1 convolution for channel count alteration, as well as a 3×3 convolution kernel with a stride of 2 convolution block for downsampling. Ultimately, the outcomes of both branches are combined to produce the final downsampled result.

Head, the feature fusion and target regression detection module of the entire target detection network, consists of the SPPCSPC layer, the ELAN-W layer, the REP layer and the PaFPN. The Spatial Pyramid Pooling Layer (SPP) is able to increase the perceptual field, allowing the algorithm to adapt to different resolutions of the image, by means of maximum pooling to obtain different perceptual fields. It consists of two branches. In the first branch, there are four branches that go through the max-pool. These four different scales of maximum pooling have four perceptual fields, which are used to distinguish between large and small targets. The REP module is divided into two, train and deploy. The train module has three branches. The top branch is a 3×3 convolution for feature extraction. The middle branch is a 1×1 convolution for smoothing features. The last branch is an identity, which is summed together. Deploy module, contains a 3×3 convolution. It is converted from the training module reparameterisation. The FPN is top-down, fusing the deep features with the underlying features by up-sampling to obtain the predicted feature map, but this operation only passes down the strong semantic features from the top to the bottom, which enhances the model's This operation only passes down the strong semantic features at the top level, which enhances the model's ability to learn the image features, but some localization features may be lost. For this reason, PAN is added after FPN to complement FPN by conveying strong localization features from the bottom up. By combining these two modules, parameters are aggregated from different backbone layers to different detection layers, thus improving the robustness and learning performance of the model in general. Finally, YOLOv7 inherits the three-layer 1×1 convolution of YOLOv5 to predict objectness, class and bbox.

Although YOLOv7 has improved on the overall model compared to YOLOv5, the adaptation to the sonar image target recognition task still has the following shortcomings. Firstly, due to the large depth of the Backbone network, although it has an advantage over other feature extraction models in that it can focus more on more shallow and deep features of the target during feature extraction, small targets can easily be overlooked by the model during the large number of convolutions. The second point is that, as the original feature extraction process mainly focuses on the target features, there is not

enough attention to the target-related feature features, thus causing some targets to be undetectable. The third point is that the original IoU metric is highly sensitive to small targets, so it will fail to identify small targets in the detection regression process, thus reducing the recognition rate of the model.

3 Model Improvement Strategy

Addressing the challenges posed by the YOLOv7 model in sonar image target detection tasks, this study introduces a small target underwater sonar image detection model based on an adaptive global feature enhancement network derived from YOLOv7. The model aims to alleviate issues related to low accuracy and the loss of numerous small targets during detection, which arise from the low resolution and shape distortion of underwater sonar images. To obtain richer gradient information, a C2fMHSA module is first designed, which parallelizes multiple gradient stream branches and shifts the network's focus from unimportant information, such as background elements, onto the target. Subsequently, a SPD-Conv approach is incorporated to preserve fine-grained information and less effective features while convoluting, thereby adapting to the low resolution of sonar images and the large proportion of small targets. Finally, the model's sensitivity to small targets is reduced by replacing the original IoU metric with the Normalized Wasserstein Distance (NWD), which in turn effectively improves the detection performance of the network.

3.1 C2fMHSA

Multi-Headed Self-Attention (MHSA) [13] is one of the basic layers of the Transformer, a method used to compute the self-attentiveness of pixels before detection, allowing the network to focus on the target rather than the background or other unimportant things.

Based on the principle of multi-headed attention mechanism, this paper introduces this method into the C2f feature extraction module of YOLOv8. Since C2f obtains rich gradient flow information through a lightweight method, but the differentiation between target and background is insufficient in the feature extraction process, this paper introduces MHSA to replace the 3×3 convolution in Bottleneck to form BottleNeckMHSA, thus proposing C2fMHSA on the basis of C2f.The details of C2fMHSA are shown in Fig. 2.

The attention logic in MHSA is $Q^{K^t} + Q^{R^t}$ where Q^{K^t} denotes content, Q^{R^t} denotes content location, q denotes query, k denotes key, and R denotes encoding. The formula for MHSA is derived as follows:

$$Attention(Q_h, K_h, V_h) = softmax(\frac{Q_h K_h^T + Q_h R_h^T}{\sqrt{d_k}})V_h \tag{1}$$

where d_k is the variance of the dot product of Q_h and K_h to mitigate the gradient vanishing problem of softmax.

The MHSA connects the multiheaded self attention and then undergoes the following linear transformation:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_h)L \tag{2}$$

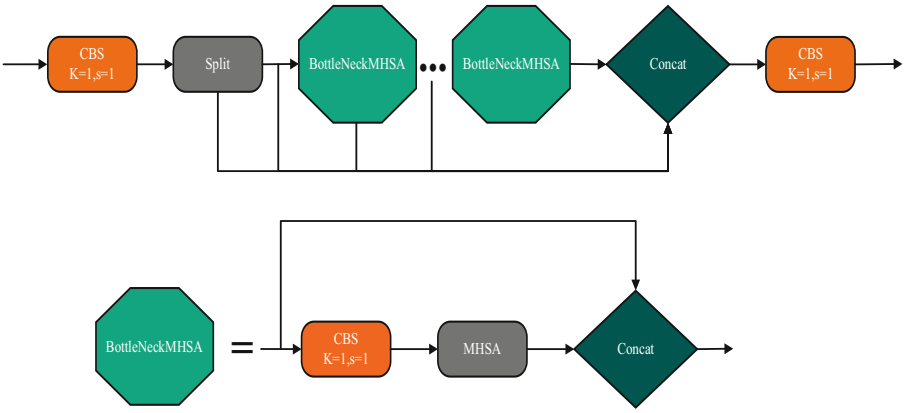

Fig. 2. Structure of the C2fMHSA model.

L is the matrix used for linear transformations. Multiheaded self attention linearly projects queries, keys and values to different dimensions through different learned linear projections, thus enabling the model to jointly process information from different representation subspaces at different locations and facilitating parallel computation.

3.2 SPD-Conv

The SPD-Conv method primarily comprises an SPD feature conversion layer and a Non-Strided Convolution layer. The fundamental principle of SPD-Conv is twofold: initially, it downsamples the internal feature mapping of Convolutional Neural Networks (CNNs), and subsequently, it preserves valid feature information to the maximum extent possible via the non-strided convolution layer. In contrast to traditional strided convolution operations, this method emphasizes the importance of target features while simultaneously preventing non-discriminatory feature loss, leading to more effective and accurate results.

SPD (Space-to-Depth). Assuming an input feature map X of size $S \times S \times C_1$, its sub-feature mapping sequence when subjected to SPD is

$$\begin{pmatrix} f_{0,0} = X[0:\mathrm{S}:\mathrm{scale}, 0:\mathrm{S}:\mathrm{scale}] & \cdots & f_{scale-1,0} = X[\mathrm{scale}-1:\mathrm{S}:\mathrm{scale}, 0:\mathrm{S}:\mathrm{scale}] \\ \vdots & \ddots & \vdots \\ f_{0,scale-1} = X[0:\mathrm{S}:\mathrm{scale}, \mathrm{scale}-1:\mathrm{S}:\mathrm{scale}] & \cdots & f_{scale-1,scale-1} = X[scale-1:\mathrm{S}:\mathrm{scale}, \mathrm{scale}-1:\mathrm{S}:\mathrm{scale}] \end{pmatrix} \tag{3}$$

From the above sequence of feature maps, the workflow of the SPD is as follows: firstly, for any feature map X, the sub-map $f_{x,y}$ consists of all feature maps to form the feature maps X (i, j), $i+x$ and $j+y$ which are divisible by scale, secondly, each sub-map is down-sampled by a scale factor x, and finally, the feature maps are joined to form a new feature map. Figure 3 shows the detailed workflow of SPD when $scale = 2$ is used.

As can be seen from the figure above, four sub-mappings $f_{0,0}$, $f_{1,0}$, $f_{0,1}$, $f_{1,1}$ are generated as a result of $scale = 2$, and the feature map consisting of the four feature

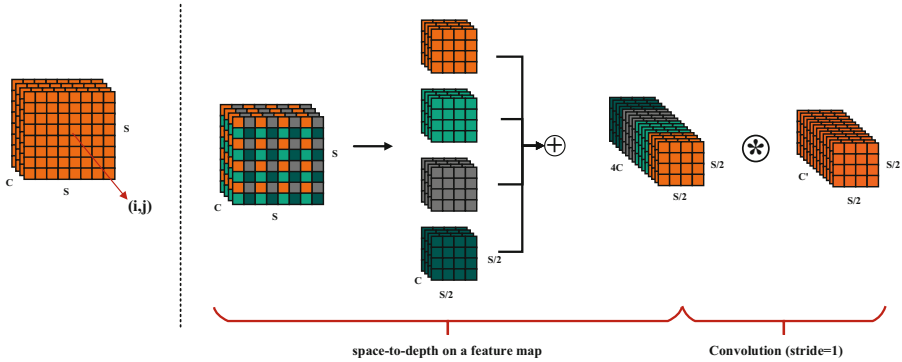

Fig. 3. SPD-Conv operational flow.

maps is shaped as (S, S, C_1), and each sub-mapping is then downsampled by a scale factor of 2. Finally, the sub-maps are concatenated along the channel dimension to give the new feature map X', which has a reduced spatial dimension of $1/2$ and an increased channel dimension of four times the original.

Non-strided Convolution. This is because when the stride size of the convolution is greater than 1 will result in a non-discriminatory loss of information, e.g. when using a convolution operation with a convolution kernel size of 3×3 and stride $= 3$, the feature map will shrink (each pixel is sampled only once) and if stride $= 2$, asymmetric sampling will occur (even rows or columns are sampled differently to odd rows or columns). Therefore in order to retain as much information as possible about all discriminative features, a non-stride convolution layer with a C_2 filter (i.e. stride $= 1$) is added after the SPD feature conversion layer, where $C_2 < scale^2 C_1$, and then the features are further converted after the SPD conversion, i.e. $X'(S/sacle, S/sacle, scale^2 C_1) \rightarrow X''(S/sacle, S/sacle, C_2)$. Thus the whole SPD-Conv module functions as $X(S, S, C_1) \rightarrow X''(S/sacle, S/sacle, C_2)$.

3.3 IoU Loss Based on NWD

In conventional methods, bounding boxes (bboxes) are represented by rectangles, utilizing Intersection over Union (IoU) to measure the alignment between bboxes. However, this approach is not well-suited for small targets, as their detection relies more on localization, particularly the position of the target center. Hence, we introduce the Normalized Wasserstein Distance (NWD) method to represent bboxes with a 2D Gaussian distribution. This distribution results in an elliptical 2D projection, where the center holds the highest weight, and the weight gradually decreases from the center to the edges. By incorporating both the original IoU and the NWD method, we effectively decrease sensitivity and enhance the detection capabilities for small targets. The steps of this approach are outlined below:

1. From matrix to ellipse: Since the planar projection of a 2D Gaussian distribution is an ellipse, the inner tangent circle of the bbox rectangle can be constructed. Assume

that bbox is (cx,cy,w,h), where (cx,cy) is the centre of the matrix and (w,h) is the length and width of the matrix, and its inner tangent circle is formulated as follows:

$$\frac{(x-cx)^2}{(w/2)^2}+\frac{(y-cy)^2}{(h/2)^2}=1 \tag{4}$$

Assuming that X is the location variable and $(\mu, \sum)$ is the mean vector and covariance matrix, the probability density function of this two-dimensional Gaussian distribution is

$$f(X|\mu,\sum)=\frac{\exp(-\frac{1}{2}(X-\mu)^T\sum^{-1}(X-\mu))}{2\pi\left|\sum\right|^{\frac{1}{2}}} \tag{5}$$

2. From 2D Gaussian distribution to ellipse: When the Eq. 1 $(X-\mu)^T\sum^{-1}(x-\mu)=1$ is satisfied, the inner tangent circle of the bbox becomes the contour of the 2D Gaussian distribution, so the derivation of the transformation of the bbox into a 2D Gaussian distribution is

$$bbox\sim N(\mu,\sum)|\mu=\begin{pmatrix}cx\\cy\end{pmatrix},\sum=\begin{pmatrix}\frac{w^2}{4}&0\\0&\frac{h^2}{4}\end{pmatrix} \tag{6}$$

3. From the similarity of the predicted and true frame distributions to the IoU between the two frames: assuming that the predicted and true distributions of the predicted and true frames (A, B) are $A\sim N(\mu_a,\sum_a), B\sim N(\mu_b,\sum_b)$, the Wasserstein distance in optimal transmission theory is chosen to calculate the distance between the two distributions and simplified by trace operations to obtain Eq. 1, and finally the exponent is normalized to (0, 1) to obtain the final WD, as shown in Eq. 1. Where C is determined by the average size of the targets in the dataset:

$$W_2^2(A,B)=\left\|\left[cx_a,cy_a,\frac{w_a}{2},\frac{h_a}{2}\right]^T,\left[cx_b,cy_b,\frac{w_b}{2},\frac{h_b}{2}\right]^T\right\|_2^2 \tag{7}$$

$$NWD(\mathcal{N}_a,\mathcal{N}_b)=exp(-\frac{\sqrt{W_2^2(\mathcal{N}_a,\mathcal{N}_b)}}{C}) \tag{8}$$

4. From NWD to NWD-based loss: IoU loss can reduce gap during training and testing, but IoU-based loss cannot provide gradients to optimize the network when there is no overlap and when ground-truth (gt) fully contains the prediction frame or vice versa. Both situations are highly likely to occur if there are small targets in the prediction objective. Although DIoU and CIoU can handle both cases, they are still too sensitive to location information for them, so NWD-based loses are introduced, with the following equation

$$\mathcal{L}_{NWD}=1-NWD(\mathcal{N}_a,\mathcal{N}_b) \tag{9}$$

Although the NWD-based loss can solve the problem of small target miss detection, due to the uneven distribution of targets in sonar images, this will also lead to a large number of additional targets being detected, thus making the model's Precision and computational effort increase.

$$IoU = 0.5\mathcal{L}_{NWD} + 0.5\alpha_{IoU} \tag{10}$$

Of which α_{IoU} Reference [14].

4 Experimentation and Analysis

The experiments in this study were tested on a public dataset of a real underwater environment containing a wide variety of target objects, The performance of the proposed approach was evaluated by comparing the proposed model qualitatively and quantitatively with the original YOLOv7, YOLOv5, YOLOv8 network models.

4.1 Data Sets

The public dataset UATD [15] used in this experiment consists of over 9000 MFLS images captured by Peng Cheng Laboratory in lakes and shallow waters using the Tritech Gemini 1200ik sonar. It contains cube, ball, cylinder, human body, type, circle cage, square cage, metal bucket, plan, and ROV 10 classes of objects. In the experiment, we used 7600 images were divided into training set (about 70%), validation set (about 10%) and test set (20%), where some of the dataset images are shown in Fig. 4.

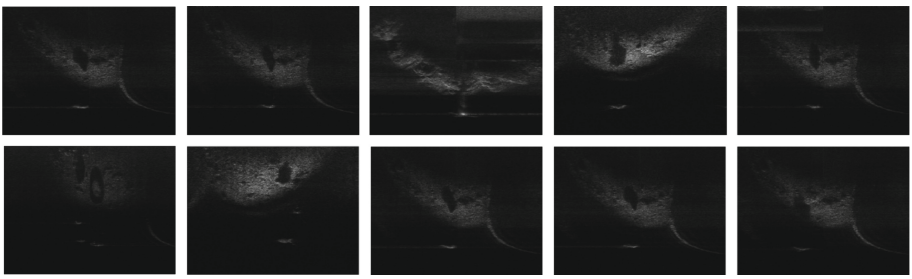

Fig. 4. Selected images of typical targets in 10 categories from the UATD dataset.

4.2 A Subsection Sample

In the experiments in this paper, the mean average precision (mAP) is used to evaluate different models. The mAP refers to the proportion of all positive samples that are predicted to be positive. The mAP measures the average precision of each category of data and is important in assessing model localization performance, target detection

model performance and segmentation model performance. In the formula, Q represents the number of target categories, and q represents the specific target of the qth category.

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{11}$$

4.3 Experimental Environment

The experiments in this paper were conducted on a Pytorch deep learning platform with two Intel Xeon Sliver 4310 CPUs, one NVIDIA Tesla A100 80G GPU, and 256GB RAM environment. To ensure the accuracy and authority of the experimental results, all models were not loaded with pre-training weights during the experiments, and different models were trained and tested, and then the performance of the algorithms were compared and analyzed.

4.4 Comparative Analysis of Model Performance Results

The change in loss and change in mAP of the improved model proposed in this paper during the training of 400 epochs in the UATD dataset are shown in Fig. 5(a), (b), respectively.

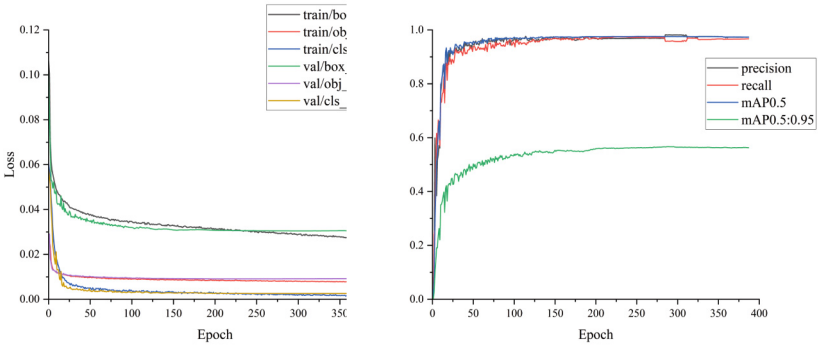

Fig. 5. Parameter changes during model training: a. Loss change plot. b. mAP change plot.

From the loss curve, it can be seen that the network gradually stabilized after the 150th epoch. The stability of the network can be judged by observing the change curves of the precision, recall and mAP of the validation set.

Tables 1 and 2 show the results of YOLOv5, YOLOv7, YOLOv8 and the improved model proposed in this paper based on YOLOv7 for different intersection ratios of detection frames and different classes of targets for a given dataset. As can be seen from Table 1, the detection results of YOLOv5, YOLOv7 and YOLOv8 are very similar when the intersection ratio threshold is greater than 0.5, while the proposed method

improves the mAP by 3.5%, 2.8% and 3.2%, respectively, compared with the other three models. mAP was 5.8%, 3.9% and 1.6%, respectively. Therefore, as shown in Table 1, the proposed target detection framework inherits the good performance of the original YOLOv7 target detection and achieves optimal detection results compared with the other three models when different intersection ratios are set for the detection frames.

Table 1. Table of mAPs corresponding to different IoUs of the detection.

Network Model	mAP0.5	mAP0.5:0.95
YOLOv5	0.942	0.518
YOLOv7	0.949	0.537
YOLOv8	0.945	0.56
Our Model	0.977	0.581

Table 2. Table of mAP for different categories of targets.

Network Model	Human Body	Ball	Circle Cage	Square Cage	Tyre	Metal Bucket	Cube	Cyclinder	Plane	Rov
YOLOv5	0.956	0.952	0.948	0.957	0.933	0.891	0.955	0.923	0.97	0.939
YOLOv7	0.954	0.954	0.937	0.963	0.941	0.911	0.961	0.949	0.971	0.949
YOLOv8	0.957	0.952	0.943	0.96	0.942	0.904	0.958	0.922	0.963	0.953
Our Model	0.995	0.982	0.995	0.995	0.972	0.992	0.955	0.931	0.995	0.959

To further validate the performance of this model, the mAP of the four network models for detecting different classes of targets were also compared, as shown in Table 2. Due to the larger width of the target detection model resulting in some targets, there is a partial loss in the feature extraction process compared to the original model, for example, the mAPs of Cube and Cyclinder differ from the original baseline model by 0.6% and 1.8% respectively. However, the detection results of the other eight categories of objects were better than the other three models. Compared to the original YOLOv7, the improved version proposed in this paper has an overall performance improvement of 2.8%, which validates the improvement of the model.

5 Conclusion

This study examines and explores recognition techniques for sonar images, focusing on enhancing the performance of a small target underwater sonar image target detection model. Leveraging YOLOv7, we proposed an adaptive global feature enhancement network to address the challenges of low accuracy and the significant loss of small targets

during the detection process, which stem from the low resolution and distorted shapes of underwater sonar images. Our primary goal is to improve the network's recognition performance. The method's generalizability is substantiated through experiments conducted on open datasets with measured sonar images. Nevertheless, the method exhibits potential limitations in detecting small targets within densely packed areas. As a result, our next research endeavor will center on enhancing the detection rate for targets with specific characteristics in sonar images.

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

References

1. McKay, J., Gerg, I., Monga, V., Raj, R.G.: What's mine is yours: Pretrained CNNs for limited training sonar ATR. In: OCEANS 2017-Anchorage, pp. 1–7. IEEE (2017)
2. Abu, A., Diamant, R.: CFAR detection algorithm for objects in sonar images. IET Radar Sonar Navig. **14**(11), 1757–1766 (2020)
3. Dos Santos, M.M., De Giacomo, G.G., Drews, P.L.J., Botelho, S.S.C.: Satellite and underwater sonar image matching using deep learning. In: 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), pp. 109–114. IEEE (2019)
4. Gu, J., Pyo, J., Joe, H., Kim, B., Kim, J., Cho, H., Yu, S.-C.: A method for automatic detection of underwater objects using forward-looking imaging sonar. In: OCEANS 2015-MTS/IEEE Washington, pp. 1–5. IEEE (2015)
5. Yulin, T., Shaohua, J., Gang, B., Yonzhou, Z., Fan, L.: Wreckage target recognition in side-scan sonar images based on an improved faster r-cnn model. In: 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), pp. 348–354. IEEE (2020)
6. Kong, W., et al.: YOLOv3-DPFIN: a dual-path feature fusion neural network for robust real-time sonar target detection. IEEE Sens. J. **20**(7), 3745–3756 (2019)
7. Cao, X., Ren, L., Sun, C.: Research on obstacle detection and avoidance of autonomous underwater vehicle based on forward-looking sonar. IEEE Trans. Neural Netw. Learn. Syst. **34**, 9198 (2023). https://doi.org/10.1109/TNNLS.2022.3156907
8. Yulin, T., Jin, S., Bian, G., Zhang, Y.: Shipwreck target recognition in side-scan sonar images by improved YOLOv3 model based on transfer learning. IEEE Access **8**, 173450–173460 (2020). https://doi.org/10.1109/ACCESS.2020.3024813
9. Tang, Y., Bian, S., Zhai, G., Liu, M., Zhang, W.: Improved YOLOv5 method for detecting shipwreck target with side-scan sonar. Geom. Inf. Sci. Wuhan Univ. (2021). https://doi.org/10.13203/j.whugis20210353
10. Sunkara, R., Luo, T.: No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Amini, M.R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., Tsoumakas, G. (eds.) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2022. Lecture Notes in Computer Science, vol. 13715. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-26409-2_27
11. Wang, J., Xu, C., Yang, W., Yu, L.: A Normalized Gaussian Wasserstein Distance for Tiny Object Detection (2021). arXiv preprint arXiv:2110.13389

12. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors (2022). arXiv preprint arXiv:2207.02696
13. Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 16514–16524. https://doi.org/10.1109/CVPR46437.2021.01625
14. He, J., Erfani, S., Ma, X., et al.: α-IoU: a family of power intersection over union losses for bounding box regression. Adv. Neural. Inf. Process. Syst. **34**, 20230–20242 (2021)
15. Xie, K., Yang, J., Qiu, K.: A dataset with multibeam forward-looking sonar for underwater object detection. Sci. Data **9**, 739 (2022)