Improved Targeted Recognition Model in Underwater Sonar Images Based on YOLOv8

Yu Huang[1], Zhe Chen[1,2], Jianxun Tang[1], and Mingsong Chen[1(✉)]

[1] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China
cms@guet.edu.cn

[2] Cognitive Radio and Information Processing Key Laboratory Authorized by China's Ministry of Education Foundation, Guilin University of Electronic Technology, Guilin 541004, China

Abstract. Deep learning based underwater sonar image target recognition has been a popular research direction. However, target recognition in sonar images continues to pose significant challenges when compared to target recognition research involving optical images. To address this issue, this study aims to propose an underwater sonar image target recognition model based on YOLOv8, which is specifically crafted to enhance the accuracy of target recognition in complex underwater scenarios. This paper addresses the issue of low resolution in sonar images by introducing the SPD-Conv CNN module, designed specifically for low-resolution images. The integration of this module leads to a significant improvement of mean average precision (mAP) by 1% when compared to the original model. Further, the Coordinate attention module leads to an additional 1.2% improvement in the recognition performance compared to the original model. Additionally, we replace the original activation function with Gaussian Error Linear Unit, resulting in a further performance enhancement of 1.9% when compared to the original model. Overall, all these model improvements collectively improved the mAP by 3.1% to reach 98.4%. The experimental results show that our model has excellent performance in target recognition in underwater sonar images.

Keywords: Underwater Sonar Image · Deep Learning · YOLOv8 Model · Target Recognition

1 Introduction

At present, the target recognition algorithm based on optical images has been more mature, and the application on underwater sonar images has been developed to a certain extent, but there is still more room for development. China has 2.997 million square kilometers of marine land area, the development of marine resources and the strategic deployment of the ocean is particularly important, with the ocean power strategy, the seaward economy, marine defense and marine ecological protection have become the focus of China's attention [1, 2]. With target recognition being a widely-used application in

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 68–75, 2025.
https://doi.org/10.1007/978-3-031-60347-1_7

the maritime domain, underwater sonar images' low resolution, background noise interference, and distorted target information place significant limitations on the application of existing recognition algorithms. Numerous target recognition algorithms are based on sonar images. For instance, Yulin et al. [3] improved the Faster R-CNN model for target detection in underwater sonar images. Although this improved the detection accuracy by 4.3 on a sonar image dataset, the method lost vital location information. Joseph Redmon, and others proposed the YOLO network [4], which uses a single-stage training method, resulting in a substantial enhancement in real-time detection speed compared to the two-stage model. YOLOv2 [5], which soon followed, gave up the use of dropout layers and utilized batch normalization for boundary prediction, leading to more accurate boundary regression predictions and improved recognition accuracy. The YOLOv3 network [6], again proposed by Joseph Redmon, mainly addresses the shortfalls of small target detection and considerably improves the model's ability to recognize small targets. Yu et al. [7] further improved the YOLOv5 model by augmenting the model with an attention mechanism to better extract features from small targets. The experimental results confirm that Yu et al.'s method produces better accuracy results and has lower time consumption.

This paper presents an improvement to the YOLOv8 target detection model making it suitable for the target detection task with underwater sonar images. The newly improved YOLOv8 model has demonstrated a remarkable performance improvement concerning the STCD [8] underwater sonar image dataset. The introduced model introduces the Coordinate attention module [9], effectively utilizing the channel features map and direction-related location information while considering the problem of long-range dependence, leading to an excellent outcome in terms of the number of parameters and computation amount. The original model used under sampling pooling, which resulted in a loss of some essential information. This issue is more pronounced for underwater sonar images with lower resolution. Consequently, this paper introduces the SPD-Conv CNN module [10] specifically designed for low-resolution images and small objects, replacing the original under sampling process, resulting in the model retaining more complete image feature information. The GELU activation function [11] is also introduced, which combines the probability of taking a value of activation parameter 1 or 0 with the activation value of the neural network, allowing the neural network to have a definite decision outcome, which preserves not only the probability but also the dependence on input. The overall performance improvement of the model is credited to these improvements.

2 Model Structure

In this section, the original structure of the YOLOv8 model and specific methods for improving the model are described. The overall structure of the improved model is shown in Fig. 1.

The Ultralytics' latest YOLO model in the series, YOLOv8 network model, brings some significant changes compared to its predecessors. The model replaces the original C3 module with C2f, reducing model weight, while adding cross-layer connections, augmenting model gradient levels. YOLOv8 model implements Decoupled-Head, discarding Anchor-Base and incorporating Anchor-Free approach, which bolsters the detector's

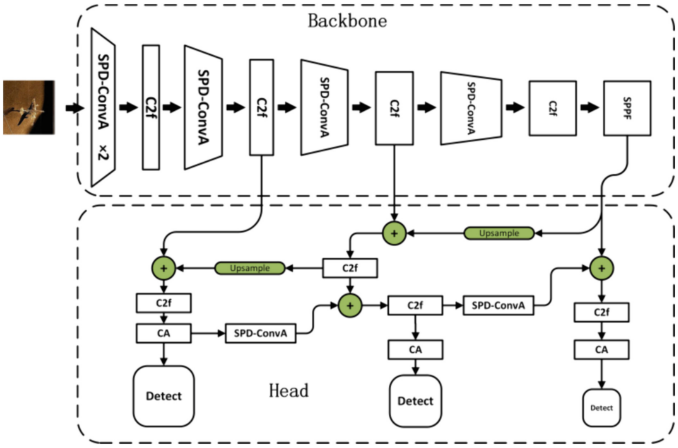

Fig. 1. Diagram of the general structure of the model.

performance, reducing the need for presetting anchors. The model only requires regressing the target's centroid, width, and height feature maps at different scales, significantly decreasing time consumption while requiring minimal computational power.

The improved model in this paper adds the CA module (Coordinate attention) to the original model. The attention module is added at each of the three detection heads, and the CA module is shown in Fig. 2.

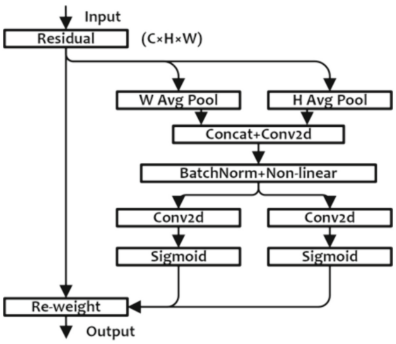

Fig. 2. Coordinate attention.

The module automatically gets the input W and H and then averages the pooling of W and H.

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < H} x_c(h, i) \tag{1}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{2}$$

After pooling by H AvgPool (formula (1)) and W AvgPool (formula (2)) to obtain C $\times$ H $\times$ 1, and C $\times$ W $\times$ 1, the pooled results are stitched on spatial latitude (dim = 2), and then the results are convolved and the channels are compressed. After the normalization and activation function in the fourth layer, the complete feature vector is then re-divided into two directional vectors, and the number of channels of the two directional feature vectors is adjusted.

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{3}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{4}$$

The final attention weight value is obtained after the Sigmoid function (formula (3–4)), and then the original input information is weighted in two directions, and the output is a weighted feature map containing both channel information and location information.

The SPD-ConvA module is used to replace the down-sampling convolution process of the original model. The main principles in the SPD-ConvA module are shown in Fig. 3.

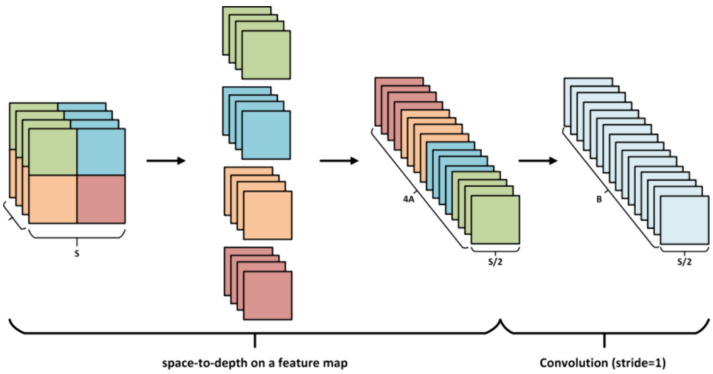

Fig. 3. SPD-Conv module schematic.

The module comprises two steps - first, undersampling the input feature map. In the illustration shown in the figure, the original feature map undergoes division into four copies. The longitudinal specification of the initial feature map is transformed from (S, S, A) to (S/2, S/2, 4A). The non-stepwise convolution changes the channel number to B. This transformation of the intermediate feature map preserves the discriminative feature information of the feature map as much as possible. It enables the model to enhance its ability to utilize detailed information, thereby improving the model's learning efficiency of the feature representation. The model demonstrates a sharp performance improvement on underwater sonar images with lower resolution.

In this paper, we replaced the original activation function with the GELU activation function. This function has shown to perform well in neural networks due to its stochastic

regular transformation with a nonlinear variation, as described by the following equation:

$$xP(X \leq x) = x\Phi(x) = x \int_{-\infty}^{x} \frac{e^{-\frac{(X-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dX \tag{5}$$

The probability of $xP(X \leq x)$ (x being the activation value input of the current neuron), i.e. the cumulative distribution $\Phi(x)$ of the Gaussian-normal distribution $\Phi(X)$ of X varies with x. As x increases, $\Phi(x)$ increases, and as x decreases, $\Phi(x)$ decreases. This means that when x is smaller, the activation result is more likely to be 0 with the current activation function, and when x is larger, it's more likely to be retained.

The loss calculation in our model consists of two branches: classification and regression. The classification loss is determined by the VFL (Vector Focal Loss) function, while the regression loss is calculated using a combination of CIOU (Complete Intersection Over Union) Loss and DFL (Distribution Focal Loss).

$$VFL(p, q) = \begin{cases} -q(q\log(p) + (1-q)\log(1-p)) & q > 0 \\ -\alpha p^{\gamma}\log(1-p) & q = 0 \end{cases} \tag{6}$$

As depicted in the formula above, p represents the label, q denotes the value obtained from norm_align_metric for positive samples, and p equals 0 for negative ones. We utilize BCE (Binary Cross-Entropy) for positive samples and apply adaptive norm_align_metric weighting to emphasize the most important samples. For negative samples, the standard FL (Focal Loss) is used.

$$DEL(S_i, S_{i+1}) = -((y_{i+1} - y)\log\log(S_i) + (y - y_i)\log(S_{i+1})) \tag{7}$$

S_i is the sigmod output of the network, y_i and y_{i+1} are the interval order, and y is the label value. The aim of the Distribution Focal Loss function is to enable the network to swiftly focus on values close to the label, achieving maximum probability density at the label. The approach involves leveraging the cross-entropy function to improve the probability of the two adjacent positions left and right of the label y. This helps to ensure that the network focuses on the values near the label.

3 Experimental Results

3.1 Data Set

The experimentation leverages the STCD-A dataset, which is an extension of the STCD dataset. This extension mitigates the imbalanced class samples in the STCD dataset. The STCD dataset comprises 357 underwater sonar images of crashed aircraft (57), human remains (34), and shipwrecks (266). The augmentation techniques include random cropping, rotation, zooming in and out, among others. The augmented STCD-A dataset comprises a total of 1620 images, consisting of crashed aircraft (588), human remains (568), and shipwrecks (494). Figure 4 displays the targets' three types.

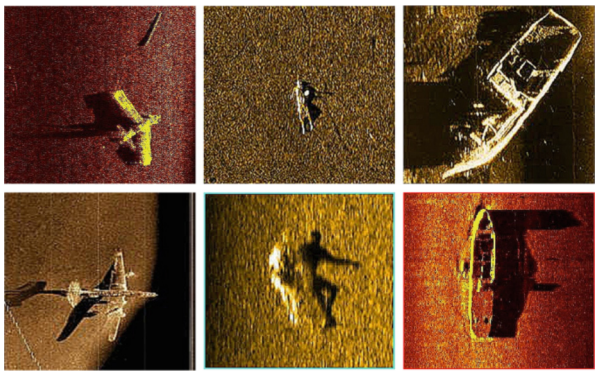

Fig. 4. STCD dataset with three types of targets.

3.2 Experimental Results and Analysis

The experiments utilize the same device, dataset, training set, and test set, in conjunction with similar conditions. The improved YOLOv8 model undergoes testing on a test set subject to 1000 training cycles to compare different models, determining the effect of the model improvements on experimental results. This consists of the training set (1166), the validation set (130), and the test set (324). Finally, there are a total of 1620 sets in this dataset. The initial learning rate of the model is set to 0.01, using the Adam optimizer. Table 1 showcases the Confusion Matrix results obtained as a result of these experiments.

Table 1. Confusion Matrix.

	aircraft	human	ship	background FP
aircraft	0.91	0.02	0	0.18
human	0.09	0.98	0	0.73
ship	0	0	0.98	0.09
background FN	0	0	0.02	0

The improved model has high accuracy in recognizing the three types of targets, which reflects the high performance of the model. The background FN indicates that the model did not recognize the target, and the table reflects the low number of unrecognized phenomena in the model. The background FP indicates that the model identifies the background as the target, and the table shows that the human class has a higher probability of background FP compared to the other two classes, because the original dataset has fewer pictures of human remains and more variability between pictures.

This paper primarily focuses on enhancing three critical positions of the YOLOv8 model. A comparison experiment's results are presented in Table 2 to demonstrate the improvement achieved.

Table 2. Improved module comparison experimental results.

None	CA	SPD-Conv	GELU	P	R	mAP0.5
✔				91.3	92.0	95.3
	✔			94.2	96.4	96.5
		✔		95.2	92.2	96.2
			✔	96.6	94.4	97.1
		✔	✔	96.1	94.8	97.8
	✔	✔	✔	96.6	96.8	98.4

In Table 2, It is evident that the application of CA, SPD-Conv, and GELU modules individually yields improved accuracy, recall, and mAP compared to the original model. Notably, the optimal results are attained when all three modules are utilized in tandem. Comparison results of our model with other three network models in the YOLO family is shown in Table 3.

Table 3. Comparison results with YOLO series models.

	P	R	mAP0.5
YOLOv3	94.6%	89.4%	93.8%
YOLOv5	96.5%	92.3%	94.3%
YOLOv8	91.3%	92.0%	95.3%
Ours	96.6%	96.8%	98.4%

The experiments demonstrate that the improved YOLOv8 model delivers the best performance under similar conditions. Due to underwater sonar image quality and other related factors, humans sometimes find it difficult to identify the target and distinguish between target categories easily. Hence, the correct identification of underwater targets is a crucial aspect for successful target detection tasks. The experimental results imply that the improvement introduced in this paper is much better than the original model, as well as other models, for underwater sonar image detection tasks.

4 Conclusion

In this paper, we propose an improved model based on YOLOv8 for the target recognition task of underwater sonar images. The model is capable of accurate target detection and classification of sonar images in complex underwater environments. In this paper, the CA attention mechanism module is added to the YOLOv8 model to make full use of the channel and direction-related location information of the feature map. Introduction the SPD-Convolution module to enhance the model's information acquisition for

underwater sonar images with low resolution. The GELU function assigns the proper weight to the input according to its size, making the output of the neuron more probabilistic. Experiments show that these improvements allow the model to perform better in underwater sonar image detection tasks and outperform other models.

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

References

1. LeHardy, P.K., Moore, C.: Deep ocean search for Malaysia airlines flight 370. In: 2014 Oceans-St. John's, pp. 1–4 (2014)
2. Zhao, J., Li, J., Li, M.: Progress and future trend of Hydrographic surveying and charting. J. Geomat. **34**(04), 25–27 (2009)
3. Yulin, T., Shaohua, J., Gang, B., et al.: 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 (2020)
4. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788 (2016)
5. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271. (2017)
6. Redmon, J., Farhadi, A.: Yolov3: an Incremental Improvement (2018). arXiv preprint arXiv: 1804.02767
7. Yu, Y., Zhao, J., Gong, Q., et al.: Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. Remote Sens. **13**(18), 3555 (2021)
8. 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)
9. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)
10. Sunkara, R., Luo, T.: No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III. Springer Nature Switzerland, Cham, pp. 443–459 (2023)
11. Hendrycks, D., Gimpel, K.: Gaussian Error Linear Units (gelus) (2016). arXiv preprint arXiv: 1606.08415