frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 12 March 2024
DOI 10.3389/fmars.2024.1348883

Check for updates

OPEN ACCESS

EDITED BY
Xinyu Zhang,
Dalian Maritime University, China

REVIEWED BY
Lanyong Zhang,
Harbin Engineering University, China
Tingkai Chen,
Dalian Maritime University, China
Boguslaw Cyganek,
AGH University of Science and Technology, Poland

*CORRESPONDENCE
Feihu Zhang
feihu.zhang@nwpu.edu.cn

RECEIVED 03 December 2023
ACCEPTED 20 February 2024
PUBLISHED 12 March 2024

CITATION
Cheng C, Wang C, Yang D, Wen X, Liu W and Zhang F (2024) Underwater small target detection based on dynamic convolution and attention mechanism.
Front. Mar. Sci. 11:1348883.
doi: 10.3389/fmars.2024.1348883

COPYRIGHT
© 2024 Cheng, Wang, Yang, Wen, Liu and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Underwater small target detection based on dynamic convolution and attention mechanism

Chensheng Cheng, Can Wang, Dianyu Yang, Xin Wen, Weidong Liu and Feihu Zhang*

School of Marine Science and Technology, Northwestern Polytenical University, Xi'an, China

In ocean observation missions, unmanned autonomous ocean observation platforms play a crucial role, with precise target detection technology serving as a key support for the autonomous operation of unmanned platforms. Among various underwater sensing devices, side-scan sonar (SSS) has become a primary tool for wide-area underwater detection due to its extensive detection range. However, current research on target detection with SSS primarily focuses on large targets such as sunken ships and aircraft, lacking investigations into small targets. In this study, we collected data on underwater small targets using an unmanned boat equipped with SSS and proposed an enhancement method based on the YOLOv7 model for detecting small targets in SSS images. First, to obtain more accurate initial anchor boxes, we replaced the original k-means algorithm with the k-means++ algorithm. Next, we replaced ordinary convolution blocks in the backbone network with Omni-dimensional Dynamic Convolution (ODConv) to enhance the feature extraction capability for small targets. Subsequently, we inserted a Global Attention Mechanism (GAM) into the neck network to focus on global information and extract target features, effectively addressing the issue of sparse target features in SSS images. Finally, we mitigated the harmful gradients produced by low-quality annotated data by adopting Wise-IoU (WIoU) to improve the detection accuracy of small targets in SSS images. Through validation on the test set, the proposed method showed a significant improvement compared to the original YOLOv7, with increases of 5.05% and 2.51% in *mAP*@0.5 and *mAP*@0.5: 0.95 indicators, respectively. The proposed method demonstrated excellent performance in detecting small targets in SSS images and can be applied to the detection of underwater mines and small equipment, providing effective support for underwater small target detection tasks.

KEYWORDS

side-scan sonar, underwater target detection, YOLOv7, K-Means++, ODConv, GAM, WIoU

1 Introduction

Due to the distinctive attributes of the underwater environment, optical imaging techniques face substantial limitations when deployed underwater. Conversely, sound waves experience minimal attenuation in water, rendering side-scan sonar (SSS) a prevalent tool for underwater target detection.

Sonar target detection methods can be categorized into traditional techniques and Convolutional Neural Network (CNN)-based approaches. Conventional sonar image detection methods predominantly employ pixel-based (Chen et al., 2014), feature-based (Mukherjee et al., 2011), and echo-based (Raghuvanshi et al., 2014) strategies. These methods utilize manually crafted filters founded on pixel value characteristics, grayscale thresholds, or *a priori* information about the targets for detection. However, underwater settings are intricate, and sonar echoes contend with self-noise, reverberation noise, and environmental noise. Consequently, sonar images exhibit low resolution, blurred edge details, and significant speckle noise, complicating the identification of dependable pixel traits and grayscale thresholds. Furthermore, owing to the diminutive illuminated regions and ambiguous target features in acoustic images, even for the same target, discrepancies in the sonar's position, depth, and angle can lead to variations in the morphological attributes of the target within sonar images. Hence, existing conventional algorithms encounter notable constraints in terms of technical feasibility, time requirements, and applicability when confronted with intricate sonar target detection scenarios. A pressing necessity exists for a detection algorithm that remains robust against fluctuations in target morphology in sonar images, mitigates erroneous detections and omissions induced by background noise interference, and exhibits commendable generalization capabilities.

In comparison to traditional methodologies, deep learning approaches rooted in CNN offer substantial advantages due to their capacity to autonomously acquire and extract deep-level features from images. The learned feature parameters often outperform manually devised counterparts, resulting in significantly heightened detection accuracy when applied to large datasets, as compared to traditional methods. Presently, CNN-based object detection methodologies within the domain of optical image processing have attained a mature stage of development. Researchers have progressively extended the application of these technologies to various inspection tasks, such as steel defect detection (Yang et al., 2021; Zhao et al., 2021), medical image analysis (Bhattacharya et al., 2021; Jia et al., 2022), marine life detection (Chen et al., 2021; Wang et al., 2023c), radar image interpretation (Hou et al., 2021; Zhang et al., 2021a), agricultural product inspection (Soeb et al., 2023; Yang et al., 2023), and more. Significant achievements have been made in each of these fields. Moreover, CNN-based methods can also be employed for image enhancement to improve the quality of blurry images and enhance the recognition of regions of interest (Chen et al., 2023; Wang et al., 2023b), thereby enhancing the effectiveness of target detection. Therefore, investigating how to apply CNN-based object detection methods more efficiently to the field of underwater acoustic image target detection is a highly worthwhile research endeavor. Furthermore, this research can contribute to addressing the challenges associated with underwater acoustic image target detection difficulties.

As of now, employing deep learning techniques for target detection in SSS images still faces several challenges (Le et al., 2020; Neupane and Seok, 2020; Hożyń, 2021). Firstly, current target detection networks typically rely on anchor box initializations derived from extensive optical datasets, which may not necessarily be suitable for our unique SSS dataset. Consequently, there is a need to re-cluster and generate anchor box initializations customized to specific dataset. Secondly, factors such as sound wave propagation loss, refraction, and scattering often result in acquired sonar images exhibiting characteristics such as low contrast, strong speckle noise, and blurry target edges. In comparison to conventional camera images, sonar images significantly differ in terms of texture diversity, color saturation, and feature resolution. Hence, it is imperative to enhance the feature extraction capability of the backbone network and apply appropriate attention mechanisms to target features in sonar images, aiming to improve detection accuracy. Lastly, due to the formidable challenges associated with collecting SSS image data, obtaining a sufficient quantity of thoroughly comprehensive and high-quality image data for network training is challenging. This necessitates making the most of all available data, including some lower-quality data, to maximize the average detection accuracy.

In response to these challenges, this paper takes full consideration of the unique characteristics of the SSS dataset. Four improvements are made to the YOLOv7 network to enhance its detection performance for small targets in SSS images. The effectiveness of the proposed improvements is validated through multiple experiments. The main contributions of this paper are as follows:

1) We replaced the k-means algorithm with k-means++ to recluster the annotated bounding boxes in the SSS dataset, thereby obtaining initial anchor boxes that are more suitable for the sizes of small targets in the dataset.
2) We replaced the static convolutional blocks in the backbone network with Omni-dimensional Dynamic Convolution (ODConv), considering the multi-dimensional information of convolutional kernels. This substitution enhances the feature extraction capability of the network without significantly increasing the number of parameters.
3) In the neck network, five global attention mechanism (GAM) modules are introduced, taking into account global information and enhancing the capability to extract target features. This addresses the challenge of feature sparsity commonly found in SSS images.
4) In the loss function section, we introduced Wise-IoU (WIoU) to address the issue of poor quality in SSS data. Such an improvement can alleviate the adverse impact of low-quality data on gradients, leading to higher data utilization and, consequently, an improvement in the detection accuracy of the trained model.

The remaining sections of this paper are structured as follows. Section 2 elaborates on related research concerning underwater acoustic target detection. In Section 3, we detail the methodology

adopted in this study. The experimental procedure and outcome presentation are outlined in Section 4. Finally, Section 5 provides a summary of this paper and offers prospects for future work.

2 Related work

Extensive research has been undertaken in the domain of underwater acoustic image target detection (Lee et al., 2018; Zhang et al., 2021b; Kim et al., 2022; Tang et al., 2023). These endeavors encompass the design of specialized functional modules tailored to data characteristics or the adaptation and enhancement of networks originally well-suited for optical data to underwater acoustic data.

(Jin et al., 2019) devised EchoNet, a deep neural network architecture that leverages transfer learning to detect sizable objects like airplanes and submerged vessels in forward-looking sonar images (Fan et al., 2021). introduced a 32-layer residual network to replace ResNet50/101 in MASK-RCNN, streamlining the network's parameter count while upholding object detection accuracy. They also adopted the Adagard optimizer in place of SGD and evaluated the detection accuracy of the network model through cross-training with a collection of 2500 sonar images (Singh and Valdenegro-Toro, 2021). conducted a comparison of diverse target segmentation networks, including LinkNet, DeepLabV3, PSPNet, and UNet, based on an extensive dataset of over 1800 forward-looking sonar images. Their investigation revealed that a UNet network employing ResNet34 as the backbone, tailored for their sonar dataset, achieved the most favorable outcomes. This network was subsequently applied to the detection and segmentation of marine debris (Xiao et al., 2021). addressed shadow information in acoustic images by introducing a shadow capture module capable of capturing and utilizing shadow data within the feature map. This module, compatible with CNN models, incurred a modest parameter increase and displayed portability. The incorporation of shadow features improved detection accuracy (Wang et al., 2021). proposed AGFE-Net, a novel sonar image target detection algorithm. This algorithm extended the receptive field of convolutional kernels through multi-scale receptive field feature extraction blocks and self-attention mechanisms, thus acquiring multi-scale feature information from sonar images and enhancing feature correlations. Employing a bidirectional feature pyramid network and an adaptive feature fusion block enabled the acquisition of deep semantic features, suppression of background noise interference, and precise prediction box selection through an adaptive non-maximum suppression algorithm, ultimately enhancing target localization accuracy. To address the issue of suboptimal transfer learning results due to significant domain gaps between optical and sonar images (Li et al., 2023a), introduced a transfer learning method for sonar image classification and object detection known as the Texture Feature Removal Network. They considered texture features in images as domain-specific features and mitigated domain gaps by discarding these domain-specific features, facilitating a more seamless knowledge transfer process. This innovative approach aims to bridge the gap between optical and sonar image analysis, enhancing the effectiveness of transfer learning techniques.

Due to the YOLO series of networks' excellent detection performance and ease of deployment, they have found wide application in the field of underwater target detection. Additionally, researchers have made numerous enhancements to the YOLO series networks, making them even more suitable for object detection in underwater acoustic images. In order to address the limitations in detection performance and low detection accuracy resulting from multi-scale image inputs (Li et al., 2023b), proposed an underwater target detection neural network based on the YOLOv3 algorithm, enhanced with spatial pyramid pooling. The improved neural network demonstrated promising results in the detection of underwater targets, including shipwrecks, schools of fish, and seafloor topography (Li et al., 2021). introduced an enhanced RBF-SE-YOLOv5 network that reallocates channel information weights to enhance effective information extraction. This enhancement entailed refining the backbone network of the original model and integrating it with RBFNet, thus improving the network's receptive field, feature representation, and capacity to learn vital information. The study demonstrated that amplifying perception information in high receptive fields and integrating multi-scale information augments the efficacy of vital feature extraction. The proposed algorithm notably enhances effective feature extraction, comprehensively captures global information, and mitigates prediction errors and issues of low credibility. Addressing the deficiency in detecting small targets in underwater sonar images (Wang et al., 2022), harnessed the YOLOv5 framework for marine debris detection. They introduced a multi-branch shuttle network into YOLOv5s and replaced YOLOv5s' neck network with BiFPN to augment detection performance. The study also analyzed the impact of uneven target data distribution and network scale on model performance, thereby furnishing reference solutions for ensuring accuracy and speed in target detection (Zhang et al., 2022a), grounded in the YOLOv5 framework, employed the IOU value between initial anchor boxes and target boxes instead of YOLOv5's Euclidean distance as the clustering criterion. This refinement brought the initial anchor boxes closer to true values, enhancing network convergence speed. Additionally, they introduced coordinate information by appending pixel coordinates of the image as extra channels to the feature map and performing convolution operations, consequently amplifying the accuracy of the detection module's localization regression (Li et al., 2023c). proposed MA-YOLOv7, a YOLOv7-based network that incorporates multi-scale information fusion and attention mechanisms for target detection and filtering in images. They also introduced a target localization method to determine target positions (latitude and longitude).

However, current research primarily revolves around employing SSS to detect large targets such as airplanes and sunken ships, or using forward-looking sonar to detect small targets at close range. There remains a significant dearth of research focused on utilizing SSS for wide-ranging detection of small underwater targets. This paper constructs a small target SSS dataset based on data collected during experiments and conducts a comprehensive study on small target detection methods in SSS. The

primary objective is to facilitate the advancement of the field of small target detection in SSS.

3 Improved methods

YOLOv7, introduced in 2022, stands as a one-stage object detection network (Wang et al., 2023a). It demonstrates outstanding proficiency in both detection speed and accuracy compared to other detection algorithms. In this study, we improved the YOLOv7 model and, through multiple experimental validations, identified four effective improvement points, as illustrated in Figure 1. We applied these enhancements to small target detection in SSS images, achieving notable improvements in detection performance compared to the original YOLOv7, as evidenced by significant enhancements in detection metrics.

3.1 K-means++

To enhance both efficiency and accuracy in detection, this study employs the k-means++ (Arthur and Vassilvitskii, 2007) technique to supplant the k-means approach, initially employed in YOLOv7, for clustering anchor boxes within the dataset. In the conventional k-means method, the first phase involves the random generation of n cluster centers from the data samples. Subsequently, the Euclidean distance between each sample and the cluster centers is computed, and the sample is assigned to the cluster center exhibiting the smallest Euclidean distance. In the subsequent phase, the cluster centers are reevaluated, and samples are reclassified. This iterative process is repeated until the cluster centers reach stability.

The k-means++ method represents an enhancement over the conventional k-means approach. Unlike generating all cluster centers randomly in a single instance, k-means++ generates one cluster center at a time. It calculates the Euclidean distance $D(x)$ between all samples and the cluster center, subsequently deriving the probability of each sample being chosen as the next cluster center through the Equation 1.

$$P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \quad (1)$$

Subsequently, the next cluster center is chosen via the roulette wheel selection method. This sequence of steps is reiterated until n cluster centers are generated. After this stage, the ensuing process resembles that of the conventional k-means algorithm: the cluster centers are updated, samples are reclassified, and these steps are iterated until the cluster centers achieve stability. While the k-means++ algorithm invests more time in selecting initial cluster centers, once these initial centers are established, the convergence speed accelerates, yielding cluster centers that hold greater representativeness. This approach mitigates the challenge of becoming trapped in local optima.

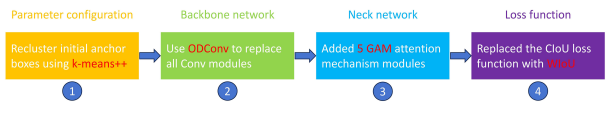

FIGURE 1
Improvements made to YOLOv7.

3.2 Omni-dimensional dynamic convolution

In current neural networks, the majority typically employ static convolutional kernels. However, recent research on dynamic convolutions suggests calculating relevant weights based on the input and linearly combining n convolutional kernels according to these weights. This makes the convolution operation dependent on the input, leading to a significant improvement in neural network accuracy. The experimental results from (Li et al., 2022) demonstrate that the use of ODConv enhances the detection performance for small targets. Therefore, in this study, all convolutional operations in the YOLOv7 backbone network are replaced with ODConv to enhance the detection performance of the network.

The core innovation of ODConv lies in its multi-dimensional dynamic attention mechanism. Traditional dynamic convolution typically achieves dynamism only in the dimension of the number of convolutional kernels, by weighting and combining multiple kernels to adapt to different input features. ODConv extends this concept further by dynamically adjusting not only the number of convolutional kernels but also three other dimensions: spatial size, input channel number, and output channel number. This means that ODConv can adapt more finely to the features of input data, thereby improving the effectiveness of feature extraction.

Additionally, ODConv employs a parallel strategy to simultaneously learn attention across different dimensions. This strategy allows the network to efficiently process features on each dimension while ensuring complementarity and synergy among the dimensions. This is particularly beneficial for handling complex features in SSS images. The network structure is illustrated in Figure 2.

The output after ODConv can be expressed using the Equation 2.

$$y = (d_1 \odot c_1 \odot b_1 \odot a_1 \odot W_1 + \ldots + d_n \odot c_n \odot b_n \odot a_n \odot W_n) * x \quad (2)$$

where a represents the attention parameter for the spatial dimensions of the convolutional kernel, b represents the attention parameter for the input channel dimensions, c represents the attention parameter for the output channel dimensions, d represents the attention parameter for the convolutional kernel W.

3.3 Global attention mechanism

The incorporation of attention mechanisms within neural networks draws inspiration from human visual attention, enhancing feature extraction by assigning distinct weights to various channels within neural network feature layers. This strategy enables the model to concentrate on pertinent

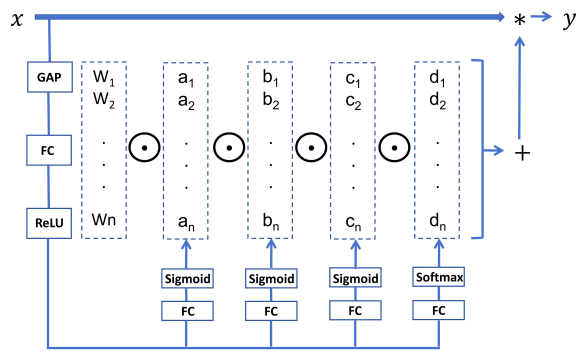

FIGURE 2
The architecture of the ODConv module.

information while disregarding irrelevant data, leading to resource conservation and augmented model performance. Several mainstream attention mechanisms, such as SE-Net (Hu et al., 2018), ECA-Net (Wang et al., 2020), BAM (Park et al., 2018), CBAM (Woo et al., 2018) and GAM (Liu et al., 2021), have been demonstrated to enhance the detection performance of models.

The GAM represents a form of global attention mechanism that curtails information loss and amplifies interactions across global dimensions. Consequently, the neural network's aptitude for extracting target features is bolstered. The schematic depiction of the GAM module structure is presented in Figure 3.

GAM employs a sequential channel-spatial attention mechanism with the aim of amplifying global inter-feature interactions while reducing information dispersion. In the channel attention submodule of GAM, a three-dimensional configuration is employed to preserve tridimensional information. The input feature map undergoes dimensional transformation and subsequently undergoes an MLP operation. The result is then reverted to the original dimension, and a sigmoid function is applied to produce the final output.

In the spatial attention submodule, aimed at intensifying focus on spatial information, two convolutional layers facilitate spatial data fusion. Initially, a convolution employing a kernel size of 7 is executed to diminish channel count and computational complexity. Subsequently, another convolution with a kernel size of 7 enhances the number of channels while maintaining uniform channel consistency. The resulting output is then processed through a sigmoid function.

In order to enhance the detection performance of the detection network, we introduced GAM modules at five distinct locations in the neck network. The architecture of the YOLOv7 network with added GAM modules, as well as the specific structures of individual sub-modules within the network, are illustrated in Figure 4.

3.4 Wise-IoU

The bounding box regression function holds a pivotal role in object detection by enhancing object localization accuracy, accommodating objects of varying scales, rectifying object orientations and shapes, and bolstering algorithmic robustness. This collective functionality contributes significantly to the advancement of object detection algorithms.

However, the majority of current research on Intersection over Union (IoU) (Yu et al., 2016) assumes that the training data consists of high-quality samples, with their primary focus being on enhancing the fitting capability of bounding box regression loss functions, such as Generalized-IoU (GIoU) (Rezatofighi et al., 2019), Distance-IoU (DIoU) (Zheng et al., 2020), Complete-IoU (CIoU) (Zheng et al., 2020), and Efficient-IoU (EIoU) (Zhang et al., 2022b), as shown in Table 1, where their advantages and disadvantages are compared. Yet, when dealing with datasets that

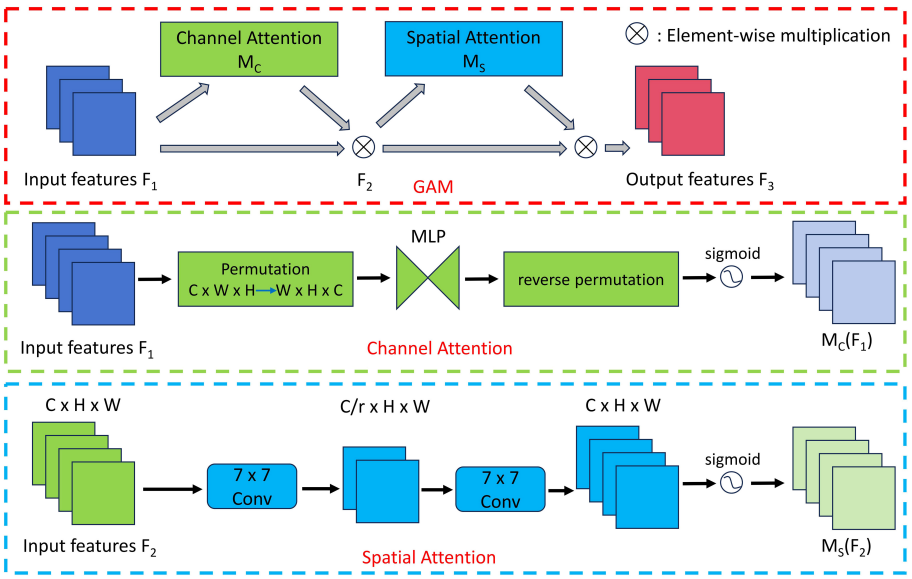

FIGURE 3
The structure diagram of the GAM module.

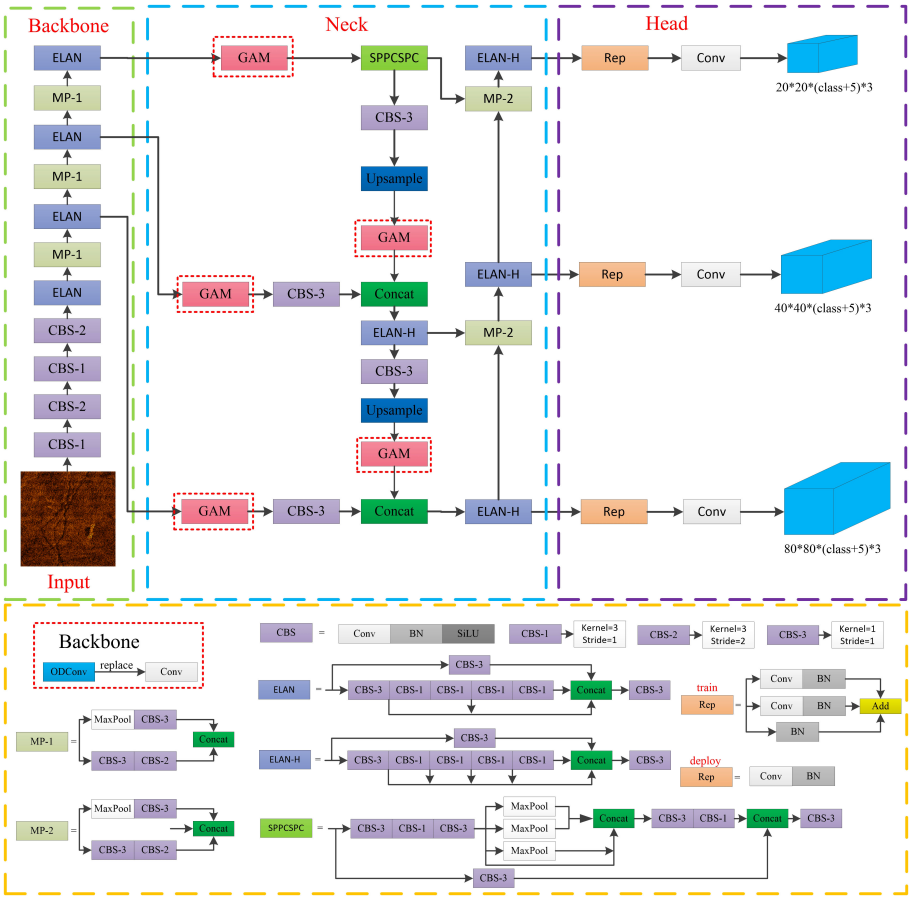

FIGURE 4
The network architecture diagram of the improved YOLOv7.

contain a significant amount of inaccurately annotated low-quality data, blindly intensifying the fitting ability of the bounding box regression loss function can have detrimental effects on the model's learning process.

In SSS imagery, targets are highly susceptible to noise interference in the generated images, which presents significant challenges for annotation. In the process of manual annotation, inaccuracies inevitably arise, as illustrated in Figure 5. If the annotation boxes are initially flawed, when an excellent detection model generates high-quality anchor boxes for low-quality sample data, the loss function L_{IoU} will have a relatively large value, leading to a substantial gradient gain. In such cases, the model will learn in an unfavorable direction. This phenomenon is particularly relevant in the context of scientific research and analysis for SSS imagery.

To address the issue of poor quality in underwater SSS data, we introduce the WIoU (Tong et al., 2023) as the bounding box loss

TABLE 1 Comparison of the advantages and shortcoming of different IoU methods.

	Overlapping	Center Point	Aspect Ratio	Advantage	Shortcoming
IoU	✓	×	×	Taking into account scale invariance and non-negativity.	If two boxes do not intersect, it cannot reflect the distance and cannot accurately reflect the degree of overlap between the two boxes.
GIoU	✓	×	×	Addressing the issue where the loss equals zero when there is no overlap between the detection box and the ground truth box.	When there is containment between the detection box and the ground truth box, GIOU degenerates into IOU, and when the two boxes intersect, convergence is slow in both the horizontal and vertical directions.
DIoU	✓	✓	×	Directly regressing the Euclidean distance between the centers of the two boxes to accelerate convergence.	Considering the aspect ratio of bounding boxes during the regression process, there is still room for further improvement in accuracy.

(Continued)

TABLE 1 Continued

	Overlapping	Center Point	Aspect Ratio	Advantage	Shortcoming
CIoU	✓	✓	✓	Introducing loss terms for the scale of the detection box, as well as for its length and width, which makes the predicted box better match the ground truth.	The aspect ratio describes relative values, introducing some degree of ambiguity and not considering the balance of difficulty levels among samples.
EIoU	✓	✓	✓	Calculating differences in width and height instead of aspect ratio, while also incorporating Focal Loss to tackle the problem of imbalanced difficulty levels among samples.	More attention is given to high-quality anchor boxes, with insufficient focus on low-quality anchor boxes.

function. This aims to alleviate the impact of low-quality anchor boxes generated during annotation. The WIoU function employs a dynamic non-monotonic focus mechanism that evaluates anchor box quality through outliers, instead of IoU. This approach furnishes a judicious gradient allocation strategy, curbing the competitiveness of high-quality anchor boxes while attenuating detrimental gradients arising from low-quality instances. Consequently, WIoU prioritizes anchor boxes of moderate quality, ameliorating detector performance overall.

The symbols defined in WIoU are illustrated as shown in Figure 6. In this figure, the blue box represents the smallest bounding box, and the red line represents the line connecting the centers of the true box and the predicted box, where the union area is denoted as $S_u = wh + w_{gt}h_{gt} - W_iHi$.

The WIoU methodology, founded on distance metrics, incorporates a two-tier attention mechanism known as WIoU v1. WIoU v1 can be represented by Equations 3, 4.

$$\mathcal{L}_{WIoUv1} = \mathcal{R}_{WIoU}\mathcal{L}_{IoU} \tag{3}$$

$$\mathcal{R}_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{4}$$

where $\mathcal{L}_{IoU} \in [0, 1]$, W_g and H_g are the dimensions of the minimum bounding box, $(x, y)$and (x_{gt}, y_{gt}) represent the center coordinates of the predicted box and the ground truth box.

Subsequently, building upon WIoU v1, the incorporation of outliers is achieved through the Equation 5.

$$\beta = \frac{\mathcal{L}^*_{IoU}}{\overline{\mathcal{L}_{IoU}}} \in [0, +\infty) \tag{5}$$

Finally, a non-monotonic focus coefficient β is formulated and integrated into WIoU v1. As a result, we obtain Equation 6.

$$\mathcal{L}_{WIoU} = r\mathcal{L}_{WIoUv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{6}$$

A reduced outlier score implies a higher quality anchor box, yielding a diminished gradient gain assigned to it. Consequently, the bounding box regression concentrates on anchor boxes of intermediate quality. In contrast, anchor boxes exhibiting larger outlier scores are allocated lesser gradient gains, effectively curtailing the generation of significant harmful gradients from low-quality instances. Notably, as $\overline{\mathcal{L}_{IoU}}$ remains dynamic, the categorization threshold for anchor boxes' quality also remains adaptive. This adaptability empowers WIoU to judiciously allocate

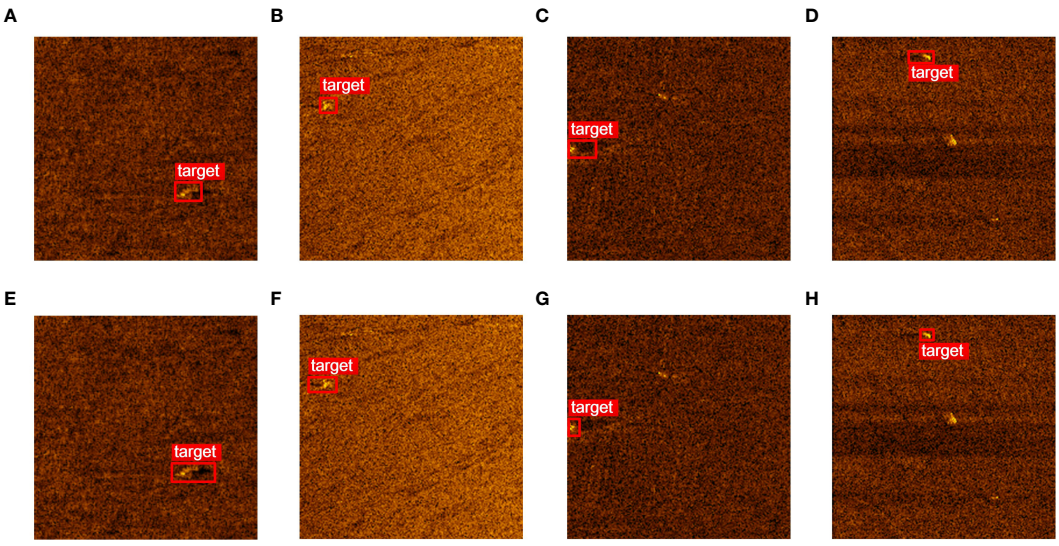

FIGURE 5
Low quality annotated samples. **(A–D)** are low-quality samples with inaccurate annotations, while **(E–H)** have accurate annotations.

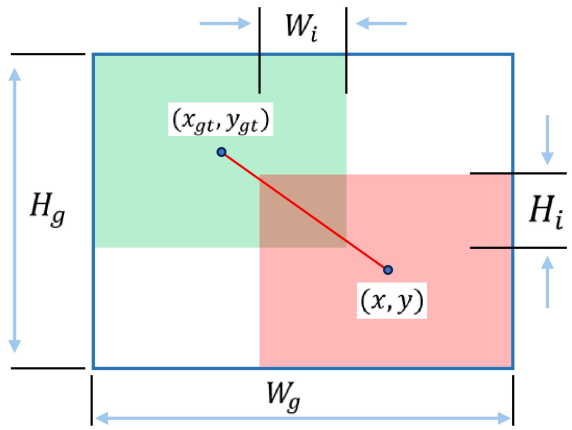

FIGURE 6
The symbol definitions in WIoU.

gradient gains that are suitable for real-time scenarios, enhancing its effectiveness in each instance.

4 Experiments

4.1 Experiment platform

The experiments presented in this study were carried out on an Ubuntu 20.04 system, serving to corroborate the efficacy of the proposed enhanced detection algorithm. Detailed configuration parameters of the system are provided in Table 2.

4.2 Model evaluation metrics

When evaluating the detection performance of the improved YOLOv7, we employed evaluation metrics including Recall (R), Precision (P), Average Precision (AP), and mean Average Precision (mAP). The calculation methods of these four indicators can be expressed by Equations 7–10 respectively.

$$R = TP/(TP + FN) \tag{7}$$

$$P = TP/(TP + FP) \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

$$mAP = \sum_{i=1}^{N} AP_i/N \tag{10}$$

Within the array of evaluation metrics mentioned, True Positive (TP) signifies the tally of correctly identified positive samples, False Positive (FP) corresponds to the count of erroneously identified negative samples, and False Negative (FN) stands for the tally of positively labeled samples that remain undetected. The variable N represents the overall number of detected categories.

TABLE 2 Experimental environment settings.

Component	Specification
Operating system	Ubuntu 20.04(64-bit)
Deep learning framwork	Pytorch 1.11
Programming language	Python 3.9
GPU accelerated environment	CUDA 11.3
Graphics Card (GPU)	Nvidia GeForce RTX 3090
Processor (CPU)	Platinum 8255C CPU @ 2.50GHz

4.3 Dataset preparation

In the data acquisition phase, we deployed objects of two types, namely cylindrical and conical structures, as detection targets in the experimental marine area. We utilized the SS3060 dual-frequency SSS as the detector for data collection. The SSS and detection targets are illustrated in Figure 7. The size of the SSS is 100mm in diameter and 1250mm in length, with a weight of 25kg in air and 12kg in water. And the performance parameters of SSS are presented in Table 3. For the experiment's execution, the SSS was affixed beneath an unmanned boat. The utilization of GPS signals emanating from the unmanned boat enabled the verification of congruence between features visible in the SSS images and the physically predetermined targets. This methodology thereby facilitated the creation of a dataset characterized by high quality.

After deploying the targets, to ensure the diversity of the collected dataset, we employed two different survey paths in the target water area to perform a comprehensive scan of underwater targets. The placement of the targets and the scanning paths are illustrated in Figure 8. In the figure, the lateral distance between the targets is approximately 50 meters, and the longitudinal distance is approximately 100 meters. Due to the influence of underwater currents, some degree of deviation in this distance is inevitably present.

Due to the complex and variable underwater environment, as well as the susceptibility of images to noise interference, the images acquired using SSS also exhibit significant variations, as shown in Figure 9.

Discerning distinct target features within SSS images presents a formidable challenge. Manual annotation subsequent to data collection is arduous, making on-site, real-time labeling the optimal strategy. To attain the real-time processing of SSS images, we adopt a tactic wherein image segments are extracted from the sonar waterfall plot at intervals of 30 seconds, illustrated in Figure 10. This approach facilitates the annotation of targets on SSS images in real-time, while accounting for the field environment and GPS coordinates.

Furthermore, the targets occupy a minuscule proportion within the complete SSS image. Training the network directly with large-scale SSS images would generate an excessive number of negative samples, potentially impeding the training process and squandering computational resources. Moreover, considering practical applications, the network needs to be deployed on resource-

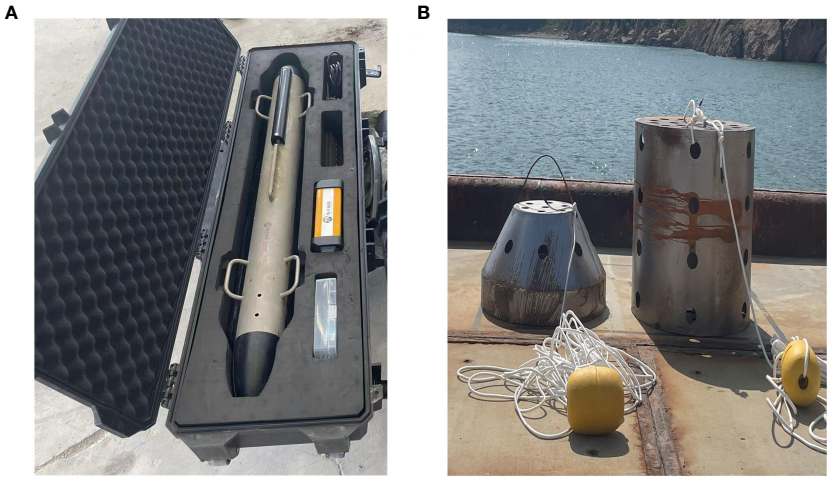

FIGURE 7
The SSS and preset targets. **(A)** SSS. **(B)** preset targets.

TABLE 3 Performance parameters of the SSS.

Frequency	300kHz	600kHz
Maximum range	150m	100m
Maximum slope distance	230m	200m
Horizontal beam width	0.5°	0.26°
Vertical beam width	50°	50°
Horizontal resolution	1.3m	0.45m
Vertical resolution	2.5m	1.25m

constrained underwater autonomous vehicles, making it imperative to restrict the image size fed into the detection network. To address these challenges, we partitioned the images into diminutive patches with dimensions of 200 × 200. Each patch features a 50-pixel overlap to prevent the loss of target characteristics. From these patches, we selectively identified those containing targets for training, significantly reducing the generation of irrelevant negative samples stemming from extraneous background information. Similarly, during the detection phase, we performed the same cropping operation before inputting the complete image into the detection network.

Finally, we filtered out unusable data and conducted data augmentation using high-quality data, yielding a total of 975 sample images. These images include 293 Cones, 318 Cylinders, and 364 Non-target instances. ("Non-target" refers to miscellaneous items on the seafloor, such as rocks or accidentally dropped artificial objects, which were not intentionally deployed by us. Despite not being the primary focus of the experiment, these Non-target items share certain similarities with the intentionally deployed targets. Including them in the dataset is essential, as their presence could potentially impact our ability to detect the deployed targets.) These samples were then randomly divided into training, validation, and test sets in a 7:1:2 ratio, with the specific number of samples for each set as shown in Table 4.

4.4 Experiment results

To validate the effectiveness of the algorithm proposed in this study for detecting small targets in SSS imagery, we tested the

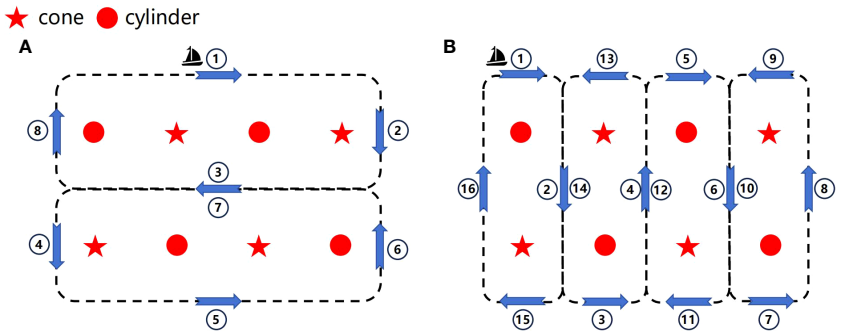

FIGURE 8
The target deployment locations and scanning paths. **(A)** Scanning path 1. **(B)** Scanning path 2.

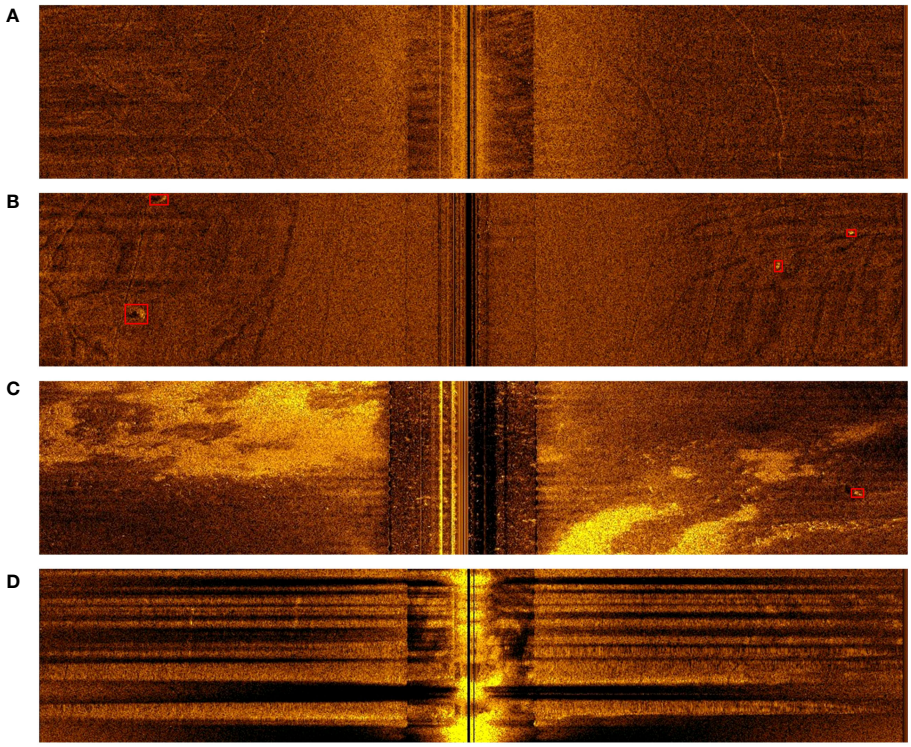

FIGURE 9
Acquired Sonar Images. **(A)** Background Images. **(B)** Images with Targets. **(C)** Target Images in Complex Environments. **(D)** Interfered Images.

algorithm on a real dataset collected during our sea trials. The variations in various loss functions and accuracy metrics during the training process are illustrated in Figure 11.

To ensure that all the introduced modifications exerted a positive influence on the network, a sequence of ablation experiments was carried out. The results of these experiments are presented in Table 5. In the table, *mAP*@0.5 represents the average precision at an IoU threshold of 0.5, while *mAP*@0.5: 0.95 represents the average of mAP values at IoU thresholds ranging from 0.5 to 0.95. It is apparent that the integration of k-means++, ODConv, GAM, and WIoU enhancements has resulted in an improved detection performance of the original YOLOv7 model on our assembled SSS dataset. The comparison of Precision-Recall (PR) curves on the test set between the improved YOLOv7 network and the original YOLOv7 network is shown in Figure 12, while the comparison of confusion matrices is shown in Figure 13. From Figure 12, it can be observed that the improved YOLOv7 network achieved an average precision improvement of 5.05% on the test set.

From Figures 12, 13, it can be observed that the improved YOLOv7 network demonstrates a noticeable enhancement in the detection performance of Non-target objects. In Figure 12, the PR curve of the enhanced YOLOv7 network for the Non-target

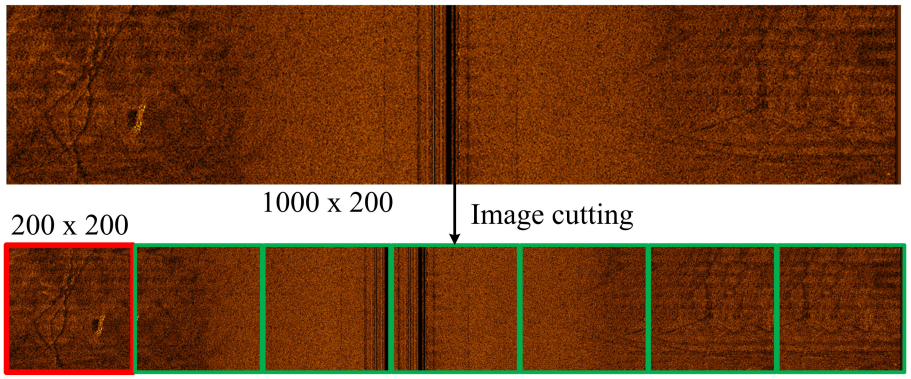

FIGURE 10
Preprocessing of SSS images. We partitioned the images into diminutive patches with dimensions of 200×200. Each patch features a 50-pixel overlap to prevent the loss of target characteristics.

TABLE 4 The actual dimensions of underwater targets and the final dataset sample size.

Category	Target			Dataset			
	Diameter	Height	Number	Train	Val	Test	Total
Cone	0.30m/0.50m	0.60m	4	205	29	59	293
Cylinder	0.50m	1.00m	4	223	31	64	318
Non-target	/	/	/	255	36	73	364

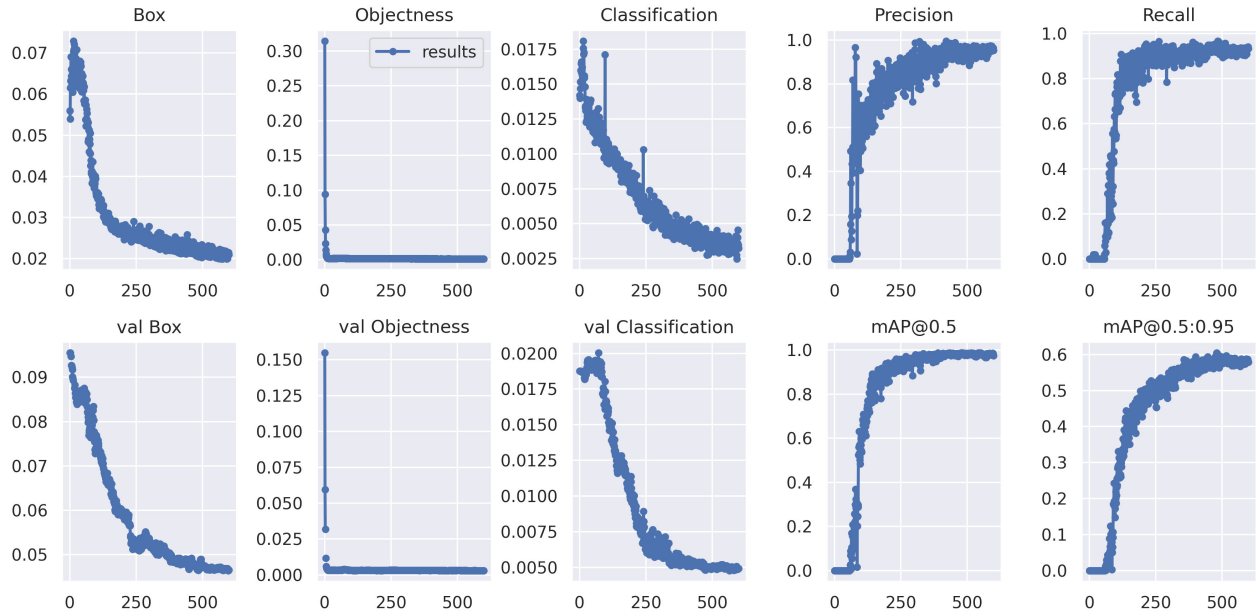

FIGURE 11
The loss function and relevant metrics during the training process of the improved YOLOv7. The horizontal axis in the figure represents the number of training epochs.

category shows a value of 0.909, which represents an improvement of 0.095 compared to the original network's 0.814. In Figure 13, within the improved YOLOv7's confusion matrix, the Non-target category registers a value of 0.94, as opposed to the original network's 0.92, marking a 0.02 improvement.

In addition, our experimental results provide evidence of the enhanced network's superior performance in detecting Non-target objects, as depicted in Figure 14. The original YOLOv7 network misclassified Non-target objects as Cylinder and Cone, whereas the improved YOLOv7 network can accurately identify Non-target categories. This advancement has reduced the false detection rate for Non-target, which holds significant practical significance in engineering applications. During the search process, it prevents wasting time on Non-target objects.

Furthermore, a comparative analysis was conducted between our enhanced detection algorithm and prominent detection networks to validate the efficacy of the proposed methodology. The comparative visualization of detection outcomes is illustrated in Figure 15. Detailed detection metrics are presented in Table 6. These findings collectively furnish compelling evidence for the superior performance of the approach proposed in this paper within the domain of small target detection using SSS.

TABLE 5 Ablation experiment.

Model	K-means++	ODConv	GAM	WIoU	*mAP*@0.5(%)	*mAP*@0.5: 0.95(%)
	×	×	×	×	90.73	49.78
	✓	×	×	×	91.77(1.04↑)	50.39(0.61↑)
YOLOv7	✓	✓	×	×	93.28(2.55↑)	51.17(1.39↑)
	✓	✓	✓	×	94.49(3.76↑)	51.79(2.01↑)
	✓	✓	✓	✓	95.78(5.05↑)	52.29(2.51↑)

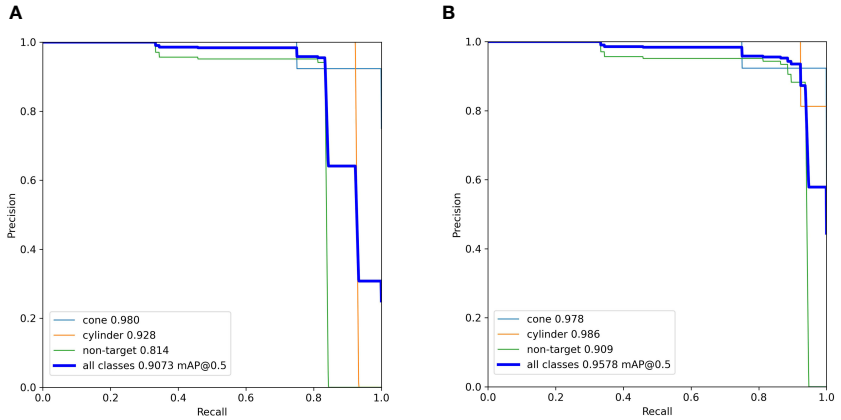

FIGURE 12
The PR curve on the test set. **(A)** initial YOLOv7 network. **(B)** improved YOLOv7 network.

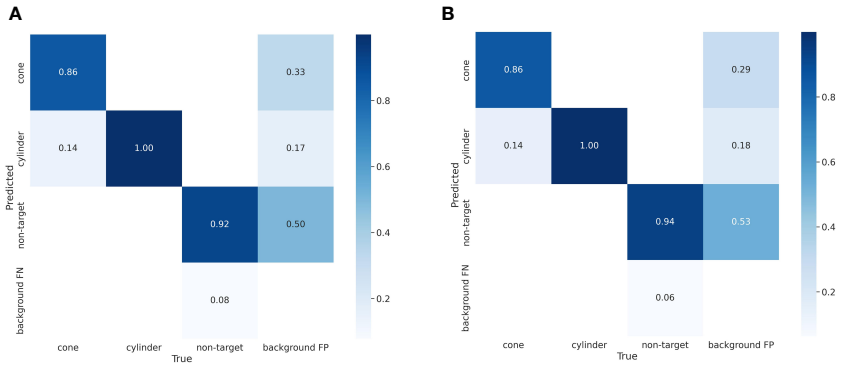

FIGURE 13
The Confusion Matrix on the test set. **(A)** initial YOLOv7 network. **(B)** improved YOLOv7 network.

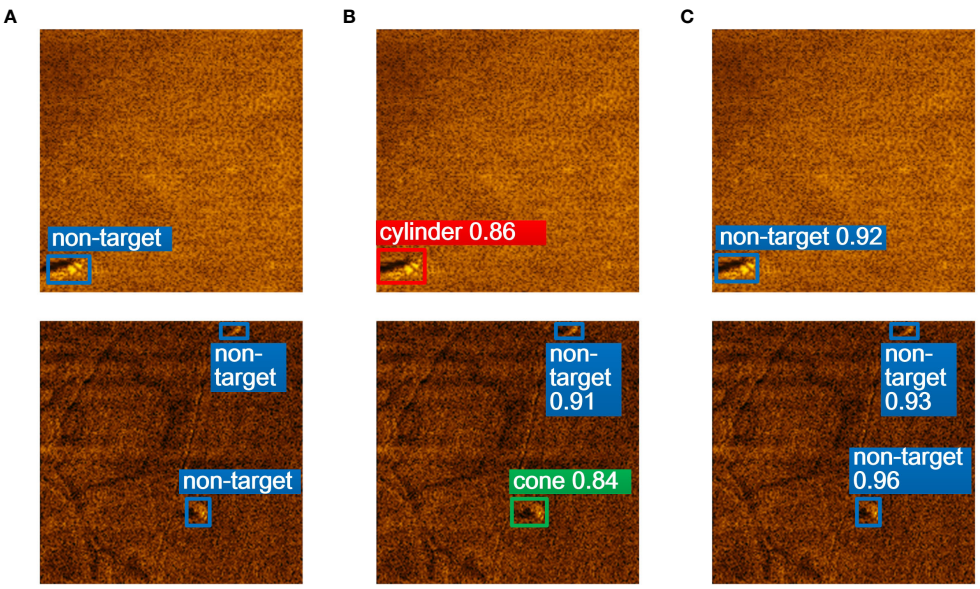

FIGURE 14
Comparison of Non-target category detection results between the improved YOLOv7 and the original YOLOv7 networks. **(A)** Labels. **(B)** Initial YOLOv7 network. **(C)** Improved YOLOv7 network.

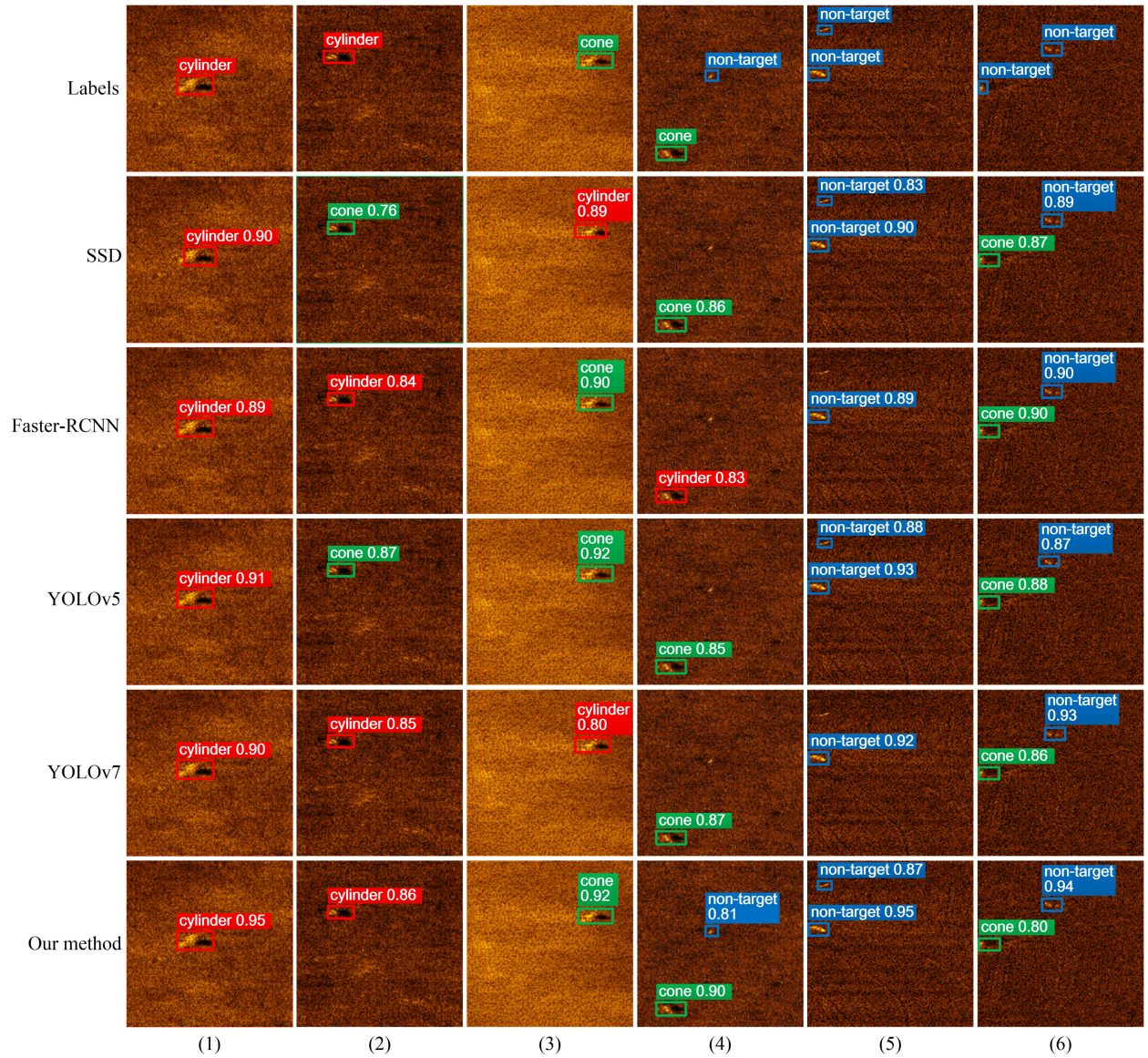

FIGURE 15
Comparison of detection results between our method and other detection networks. The first row in the figure represents the ground truth labels for different target categories, while the second to fifth rows depict the detection results of various algorithms.

The results illustrated in Figure 15 provide empirical validation of the efficacy of the approach introduced in this research. As demonstrated in columns (2), (3), and (4) of Figure 15, some mainstream detection networks often exhibit mis-detections when accurately distinguishing between the categories of cylindrical and conical objects. In contrast, the proposed method in this paper demonstrates accurate detection for objects that are challenging to differentiate, with higher

TABLE 6 Comparison of detection metrics between our method and other detection networks.

Method	Precision(%)	Recall(%)	*mAP*@0.5(%)	*mAP*@0.5: 0.95(%)
SSD	88.31	89.76	89.28	48.24
Faster-RCNN	85.33	83.91	87.19	46.73
YOLOv5	88.72	90.46	89.98	49.80
YOLOv7	93.56	89.12	90.73	49.78
Our method	92.99	89.10	95.78	52.29

probability values assigned. This highlights the superiority of the algorithm presented in this paper.

Nonetheless, it is important to note that the enhanced network in this study does exhibit certain limitations. For example, as depicted in column (6) of Figure 15, all networks misclassify a Non-target as a Cone. This misclassification arises due to the distinct shadow surrounding the Non-target and the similarity in the size of the bright spot to the Cone category, resulting in a false positive detection. At present, there is a lack of definitive solutions for scenarios in which acoustic image features exhibit extremely high similarity, yet the actual objects belong to different categories. Using a higher-precision device to acquire images with increased resolution may be beneficial for addressing this issue.

5 Conclusions

This study collected a dataset of small target SSS images during sea trials and proposed an enhancement method based on the YOLOv7 model for detecting small targets in SSS images. The method utilizes the k-means++ algorithm to obtain more accurate initial anchor box sizes. Subsequently, it employs ODConv to replace static convolution modules in the YOLOv7 backbone network and integrates a GAM attention mechanism into the YOLOv7 neck network, thereby enhancing the feature extraction capabilities of the detection network. In the loss function section, a WIoU loss function is introduced to balance the impact of high-quality and low-quality anchor boxes on gradients, enhancing the network's focus on average-quality anchor boxes. Experimental results demonstrate the effectiveness of the proposed YOLOv7-based enhancement algorithm, with *mAP*@0.5 and *mAP*@0.5: 0.95 metrics reaching 95.78% and 52.29%, respectively, representing improvements of 5.05% and 2.51% over the original YOLOv7 network. Furthermore, comparisons with mainstream underwater detection networks confirm the superiority of the proposed method in small target detection in SSS images.

The proposed method can be applied to autonomous target detection in Unmanned Underwater Vehicles (UUVs) and Unmanned Surface Vehicles (USVs), enhancing the autonomous operational capabilities of unmanned autonomous ocean observation platforms. In the future, we plan to collect more diverse small target data and continue researching SSS-based small target detection methods to further contribute to underwater exploration.

Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

Author contributions

CC: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. CW: Methodology, Writing – review & editing. DY: Software, Writing – review & editing. XW: Software, Writing – review & editing. WL: Project administration, Writing – review & editing. FZ: Conceptualization, Project administration, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Key Research and Development Program (2023YFC2808400).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arthur, D., and Vassilvitskii, S. (2007). "K-means++ the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, SODA 2007*. (New Orleans, Louisiana, USA: ACM), 1027–1035. doi: 10.1145/1283383.1283494

Bhattacharya, S., Maddikunta, P. K. R., Pham, Q.-V., Gadekallu, T. R., Chowdhary, C. L., Alazab, M., et al. (2021). Deep learning and medical image processing for coronavirus (covid-19) pandemic: A survey. *Sustain. cities Soc.* 65, 102589. doi: 10.1016/j.scs.2020.102589

Chen, T., Wang, N., Chen, Y., Kong, X., Lin, Y., Zhao, H., et al. (2023). Semantic attention and relative scene depth-guided network for underwater image enhancement. *Eng. Appl. Artif. Intell.* 123, 106532. doi: 10.1016/j.engappai.2023.106532

Chen, Z., Wang, H., Shen, J., and Dong, X. (2014). "Underwater object detection by combining the spectral residual and three-frame algorithm," in *Lecture Notes in Electrical Engineering* (Berlin, Germany: Springer). 279, 1109–1114. doi: 10.1007/978-3-642-41674-3_154

Chen, T., Wang, N., Wang, R., Zhao, H., and Zhang, G. (2021). One-stage cnn detector-based benthonic organisms detection with limited training dataset. *Neural Networks* 144, 247–259. doi: 10.1016/j.neunet.2021.08.014

Fan, Z., Xia, W., Liu, X., and Li, H. (2021). Detection and segmentation of underwater objects from forward-looking sonar based on a modified mask rcnn. *Signal Image Video Process.* 15, 1135–1143. doi: 10.1007/s11760-020-01841-x

Hożyń, S. (2021). A review of underwater mine detection and classification in sonar imagery. *Electronics* 10, 2943. doi: 10.3390/electronics10232943

Hou, F., Lei, W., Li, S., and Xi, J. (2021). Deep learning-based subsurface target detection from gpr scans. *IEEE sensors J.* 21, 8161–8171. doi: 10.1109/JSEN.2021.3050262

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-Excitation Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Salt Lake City, UT, USA: IEEE). pp. 7132–7141. doi: 10.1109/CVPR.2018.00745

Jia, Y., Wang, H., Chen, W., Wang, Y., and Yang, B. (2022). An attention-based cascade r-cnn model for sternum fracture detection in x-ray images. *CAAI Trans. Intell. Technol.* 7, 658–670. doi: 10.1049/cit2.12072

Jin, L., Liang, H., and Yang, C. (2019). Accurate underwater atr in forward-looking sonar imagery using deep convolutional neural networks. *IEEE Access* 7, 125522–125531. doi: 10.1109/ACCESS.2019.2939005

Kim, W.-K., Bae, H. S., Son, S.-U., and Park, J.-S. (2022). Neural network-based underwater object detection off the coast of the korean peninsula. *J. Mar. Sci. Eng.* 10, 1436. doi: 10.3390/jmse10101436

Le, H. T., Phung, S. L., Chapple, P. B., Bouzerdoum, A., Ritz, C. H., and Tran, L. C. (2020). Deep gabor neural network for automatic detection of mine-like objects in sonar imagery. *IEEE Access* 8, 94126–94139. doi: 10.1109/ACCESS.2020.2995390

Lee, S., Park, B., and Kim, A. (2018). Deep learning from shallow dives: Sonar image generation and training for underwater object detection. *arXiv preprint arXiv:1810.07990*. doi: 10.48550/arXiv.1810.07990

Li, J., Chen, L., Shen, J., Xiao, X., Liu, X., Sun, X., et al. (2023b). Improved neural network with spatial pyramid pooling and online datasets preprocessing for underwater target detection based on side scan sonar imagery. *Remote Sens.* 15, 440. doi: 10.3390/rs15020440

Li, L., Li, Y., Yue, C., Xu, G., Wang, H., and Feng, X. (2023c). Real-time underwater target detection for auv using side scan sonar images based on deep learning. *Appl. Ocean Res.* 138, 103630. doi: 10.1016/j.apor.2023.103630

Li, W., Wang, J., Zhao, X. F., Wang, Z., and Zhang, Q. J. (2021). "Target detection in color sonar image based on yolov5 network," in *2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. (Xi'an, China: IEEE), 1–5.

Li, C., Ye, X., Xi, J., and Jia, Y. (2023a). A texture feature removal network for sonar image classification and detection. *Remote Sens.* 15, 616. doi: 10.3390/rs15030616

Li, C., Zhou, A., and Yao, A. (2022). Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947*. doi: 10.48550/arXiv.2209.07947

Liu, Y., Shao, Z., and Hoffmann, N. (2021). Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv preprint arXiv:2112.05561*. doi: 10.48550/arXiv.2112.05561

Mukherjee, K., Gupta, S., Ray, A., and Phoha, S. (2011). Symbolic analysis of sonar data for underwater target detection. *IEEE J. Oceanic Eng.* 36, 219–230. doi: 10.1109/JOE.2011.2122590

Neupane, D., and Seok, J. (2020). A review on deep learning-based approaches for automatic sonar target recognition. *Electronics* 9, 1972. doi: 10.3390/electronics9111972

Park, J., Woo, S., Lee, J.-Y., and Kweon, I. S. (2018). Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*. doi: 10.48550/arXiv.1807.06514

Raghuvanshi, D. S., Dutta, I., and Vaidya, R. (2014). "Design and analysis of a novel sonar-based obstacleavoidance system for the visually impaired and unmanned systems," in *2014 International Conference on Embedded Systems (ICES)*. (Coimbatore, India: IEEE), 238–243.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (Long Beach, CA, USA: IEEE), 658–666.

Singh, D., and Valdenegro-Toro, M. (2021). "The marine debris dataset for forward-looking sonar semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (Montreal, BC, Canada: IEEE), 3741–3749.

Soeb, M. J. A., Jubayer, M. F., Tarin, T. A., Al Mamun, M. R., Ruhad, F. M., Parven, A., et al. (2023). Tea leaf disease detection and identification based on yolov7 (yolo-t). *Sci. Rep.* 13, 6078. doi: 10.1038/s41598-023-33270-4

Tang, Y., Wang, L., Jin, S., Zhao, J., Huang, C., and Yu, Y. (2023). Auv-based side-scan sonar real-time method for underwater-target detection. *J. Mar. Sci. Eng.* 11, 690. doi: 10.3390/jmse11040690

Tong, Z., Chen, Y., Xu, Z., and Yu, R. (2023). Wise-iou: Bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051*. doi: 10.48550/arXiv.2301.10051

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023a). "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Vancouver, BC, Canada: IEEE), 7464–7475.

Wang, N., Chen, T., Kong, X., Chen, Y., Wang, R., Gong, Y., et al. (2023b). Underwater attentional generative adversarial networks for image enhancement. *IEEE Trans. Human-Machine Syst.* 53 (3), 490-500. doi: 10.1109/THMS.2023.3261341

Wang, N., Chen, T., Liu, S., Wang, R., Karimi, H. R., and Lin, Y. (2023c). Deep learning-based visual detection of marine organisms: A survey. *Neurocomputing* 532, 1–32. doi: 10.1016/j.neucom.2023.02.018

Wang, J., Feng, C., Wang, L., Li, G., and He, B. (2022). Detection of weak and small targets in forward-looking sonar image using multi-branch shuttle neural network. *IEEE Sensors J.* 22, 6772–6783. doi: 10.1109/JSEN.2022.3147234

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (Seattle, WA, USA: IEEE), 11534–11542.

Wang, Z., Zhang, S., Huang, W., Guo, J., and Zeng, L. (2021). Sonar image target detection based on adaptive global feature enhancement network. *IEEE Sensors J.* 22, 1509–1530. doi: 10.1109/JSEN.2021.3131645

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision. (ECCV)* (Munich Germany: Springer), 3–19.

Xiao, T., Cai, Z., Lin, C., and Chen, Q. (2021). A shadow capture deep neural network for underwater forward-looking sonar image detection. *Mobile Inf. Syst.* 2021, 1–10. doi: 10.1155/2021/3168464

Yang, D., Cui, Y., Yu, Z., and Yuan, H. (2021). Deep learning based steel pipe weld defect detection. *Appl. Artif. Intell.* 35, 1237–1249. doi: 10.1080/08839514.2021.1975391

Yang, H., Liu, Y., Wang, S., Qu, H., Li, N., Wu, J., et al. (2023). Improved apple fruit target recognition method based on yolov7 model. *Agriculture* 13, 1278. doi: 10.3390/agriculture13071278

Yu, J., Jiang, Y., Wang, Z., Cao, Z., and Huang, T. (2016). "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*. (Amsterdam, The Netherlands: ACM), 516–520.

Zhang, Y.-F., Ren, W., Zhang, Z., Jia, Z., Wang, L., and Tan, T. (2022b). Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing* 506, 146–157. doi: 10.1016/j.neucom.2022.07.042

Zhang, P., Tang, J., Zhong, H., Ning, M., Liu, D., and Wu, K. (2021a). Self-trained target detection of radar and sonar images using automatic deep learning. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi: 10.1109/TGRS.2021.3096011

Zhang, P., Tang, J., Zhong, H., Ning, M., Liu, D., and Wu, K. (2021b). Self-trained target detection of radar and sonar images using automatic deep learning. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi: 10.1109/TGRS.2021.3096011

Zhang, H., Tian, M., Shao, G., Cheng, J., and Liu, J. (2022a). Target detection of forward-looking sonar image based on improved yolov5. *IEEE Access* 10, 18023–18034. doi: 10.1109/ACCESS.2022.3150339

Zhao, W., Chen, F., Huang, H., Li, D., and Cheng, W. (2021). A new steel defect detection algorithm based on deep learning. *Comput. Intell. Neurosci.* 2021, 1–13. doi: 10.1155/2021/5592878

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2020). "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*. (New York, USA: AAAI), Vol. 3, 12993–13000.