Journal of *Marine Science and Engineering*

MDPI

Article

Detection of Small Objects in Side-Scan Sonar Images Using an Enhanced YOLOv7-Based Approach

Feihu Zhang *, Wei Zhang, Chensheng Cheng, Xujia Hou and Chun Cao

School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; 2021200685@mail.nwpu.edu.cn (W.Z.); chensheng.cheng@mail.nwpu.edu.cn (C.C.); hxj1363947894@mail.nwpu.edu.cn (X.H.); caochun@mail.nwpu.edu.cn (C.C.)
* Correspondence: feihu.zhang@nwpu.edu.cn

Abstract: Deep learning-based object detection methods have demonstrated remarkable effectiveness across various domains. Recently, there has been growing interest in applying these techniques to underwater environments. Conventional optical imaging methods face severe limitations when operating in underwater conditions, restricting their ability to identify objects with good visibility and at close distances. Consequently, side-scan sonar (SSS) has emerged as a common equipment choice for underwater detection due to its compatibility with the characteristics of sound waves in water. This paper introduces a novel method, termed the Enhanced YOLOv7-Based Approach, for detecting small objects in SSS images. Building upon the widely-adopted YOLOv7 method, the proposed approach incorporates several enhancements aimed at improving detection accuracy. First, a dedicated detection layer tailored for small objects is added to the original network architecture. Additionally, two attention mechanisms are integrated within the backbone and neck components of the network, respectively, to strengthen the network's focus on object features. Finally, the network features are recombined based on the BiFPN structure. Experimental results demonstrate that the proposed method outperforms mainstream object detection algorithms. In comparison to the original YOLOv7 network, it achieves a Precision of 95.5%, indicating a significant improvement of 4.8%. Moreover, its Recall reaches 87.0%, representing an enhancement of 5.1%, while the mean Average Precision (mAP) at an IoU threshold of 0.5 (mAP@.5) reaches 86.9%, reflecting a 6.7% improvement. Furthermore, the mAP@.5:.95 reaches 55.1%, a 4.8% enhancement. Therefore, the method presented in this paper enhances the performance of YOLOv7 for object detection in SSS images, providing a fresh perspective on small object detection based on SSS images and contributing to the advancement of underwater detection techniques.

Keywords: side-scan sonar; object detection; YOLOv7; small objects; attention mechanism; BiFPN; feature fusion

Citation: Zhang, F.; Zhang, W.; Cheng, C.; Hou, X.; Cao, C. Detection of Small Objects in Side-Scan Sonar Images Using an Enhanced YOLOv7-Based Approach. *J. Mar. Sci. Eng.* **2023**, *11*, 2155. https://doi.org/10.3390/jmse11112155

Academic Editors: Adriano Mancini, Anna Nora Tassetti and Pierluigi Penna

Received: 7 October 2023
Revised: 9 November 2023
Accepted: 10 November 2023
Published: 12 November 2023

1. Introduction

SSS is commonly used for underwater detection and image acquisition, with extensive applications in underwater sediment classification [1], underwater objects detection [2], underwater image segmentation [3], and other fields. The detection of small underwater objects has become a prominent research focus both domestically and internationally due to technical challenges such as limited availability of acoustic image data, difficulties in feature extraction, variable scales, and challenging detection scenarios. Traditional sonar image detection methods are mainly based on pixel [4], feature [5], and echo [6] methods, which manually design filters for object detection according to pixel value features, the gray threshold, or prior information about the object. However, because the underwater environment is very complicated, sonar echoes are affected by self-noise, reverberation noise, and ambient noise, resulting in low resolution, blurred edge details, and serious speckle noise. Therefore, it is difficult to find good pixel characteristics and gray thresholds.

In recent years, with the rapid development of deep learning and convolutional neural networks, researchers have designed many detectors; these are mainly divided into single-stage detectors and two-stage detectors. A two-stage detector tries to find any number of objects in the image in the first stage, then classifies and locates them in the second stage. Two-stage detectors mainly rely on R-CNN [7], SPP-Net [8], Fast R-CNN [9], Faster R-CNN [10], FPN [11], R-FCN [12], MaskR-CNN [13], DetectoRS [14], etc. Single-stage detectors use intensive sampling to classify and locate semantic objects in a single scan. They use predefined boxes/key points of varying proportions and aspect ratios to locate objects. Single-stage object detection algorithms treat this task as a regression problem, using a single end-to-end network from the input raw image to the object location and category output. Compared with two stage-detection, single-stage detection algorithms typically have better real-time performance. The single-stage detectors mainly include YOLO [15], SSD [16], RetinaNet [17], EfficientDet [18], YOLOv7 [19], etc.

There are several difficulties in sonar image small object detection based on deep learning. First, it is difficult to learn the correct representation from the limited and distorted information of small objects. Second, there is a scarcity of large-scale datasets for small object detection. Third, there is very low tolerance for positioning error; because the annotation box of small objects is generally relatively small, even a small positioning error can cause a large visual deviation. Yang et al. [20] proposed a high quality sonar image generation method based on a diffusion model, which is used to generate a large number of high quality sonar images with obvious features and can be used for training to ensure that the objects in the image are more obvious. This approach can be further applied to engineering applications such as target detection and image classification. Fu et al. [21] proposed an improved YOLOv5 method based on an attention mechanism and an improved anchor frame for real-time detection of underwater small objects in SSS images, aiming to address the shortcomings of high miss rate and high false detection rate in the detection of underwater small objects based on YOLOv5. Wang et al. [22] proposed a new sonar image object detection algorithm called AGFE-Net, which uses multi-scale sensing domain feature extraction blocks and a self-attention mechanism to expand the convolution kernel sensing domain in order to obtain multi-scale feature information of sonar images and enhance the correlation between different features. Based on the YOLOv5 framework, Zhang et al. [23] used the IOU values of the initial anchor frame and the object frame instead of the Euclidean distance typically used in YOLOv5 as the basis for clustering. This approach provides an initial anchor frame that is closer to the real value, increasing the convergence speed of the network. In addition, the pixel coordinates of the image were added to the feature graph as the information of the two channels. The accuracy of detection module positioning regression is improved. Li et al. [24] proposed a real-time SSS image object detection algorithm based on YOLOv7. First, a method based on threshold segmentation and pixel importance value was used to quickly identify any suspicious objects in the SSS images, then scale information fusion and an attention mechanism were introduced to the network. The proposed algorithm achieved advanced performance and can be applied to real underwater tasks.

In the current mainstream object detection models, single-stage objects detection networks are more suitable for object detection in SSS images because of their high accuracy and good real-time performance. Among them, YOLOv7 is a leading algorithm known for its exceptional detection accuracy, fast speed, and scalability. Therefore, in this paper we propose an improved YOLOv7 method on the basis of YOLOv7 model, including an increased detection scale, re-fusion of features, and introduction of an attention mechanism.

In summary, the main contributions of this paper are as follows:

- To enhance the network's capability in detecting small objects, a 160 × 160 detection layer is incorporated, allowing the network to capture additional feature representations.
- The feature extraction capability of the network is improved by incorporating the CoT module [25] to extend the ELAN module, introducing contextual self-attention to the backbone part.

- The network incorporates a CA module [26] both before and after feature fusion in the neck part, with the aim of enhancing the network's focus on object features and improving the accuracy of detection.
- Building upon the BiFPN structure [18], the network employs learnable weights to fuse features of varying scales in order to accommodate the requirements of different detection scales.

The dataset used in this paper was collected by the laboratory using SSS in a sea test, and contains four kinds of objects. To assess the effectiveness of the method proposed in this study, an ablation study was conducted. The experimental results demonstrate that the incorporation of each key innovation described in this paper significantly improves the detection accuracy of the original network. The proposed enhanced YOLOv7-based approach presents a viable solution for improving the detection of small underwater objects in SSS images. This approach offers a new option for addressing the challenges of small object detection in SSS applications.

The rest of this paper is organized as follows: Section 2 introduces related work, including the YOLOv7 model, two attention mechanisms, and BiFPN structure; Section 3 introduces the enhanced YOLOv7-based approach; Section 4 describes the datasets, experiments, and resulting analysis; finally, Section 5 summarizes the work.

2. Methods

In this section, an overview of the YOLOv7 network is provided, along with an introduction to the two attention mechanisms and the BiFPN structure.

2.1. YOLOv7 Network

There are three types of YOLOv7 [19] network models applicable to different GPUs: YOLOV7-tiny for edge GPUs, YOLOv7 for ordinary GPUs, and YOLOv7-W6 for cloud GPUs. The three models gradually increase in depth, complexity, and detection accuracy. This section briefly introduces the network structure of YOLOv7.

YOLOv7 is the most advanced algorithm in the YOLO series at present, and is the most typical representative of one-stage object detection algorithms, surpassing the previous YOLO series in both accuracy and speed. The YOLOv7 network is mainly composed of three parts: a backbone part, neck part, and head part, as shown in Figure 1.

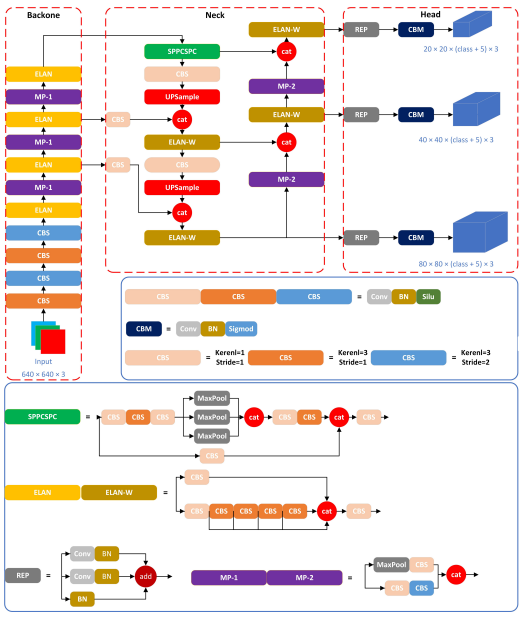


Figure 1. YOLOv7 original network. The input image size of the network is 640 × 640 × 3, which is mainly divided into three parts: backbone, neck, and head. The structural diagram of the main modules of the network are shown below the figure.

The backbone part plays a crucial role in capturing low-level features such as edges, textures, and shapes in the image, then progressively transforming them into higher-level semantic features. The neck part serves as an intermediary layer between the backbone network and the head network; its main role is to further integrate and fuse features from different levels. The head part is responsible for performing object classification and position regression on the feature map generated by the neck component. Typically, it is composed of a sequence of convolutional layers and fully connected layers that extract features and predict the object's category label and bounding box.

In this paper, the YOLOv7 network is utilized as a benchmark network for comparison with the proposed enhanced YOLOv7-based method. The size of the input side scan sonar image is $640 \times 640 \times 3$, and detection is carried out on three scales: 80×80, 40×40, and 20×20.

2.2. CA Module

Attention mechanisms are commonly employed in object detection to enable neural networks to learn the content and location that demand increased attention. The majority of existing methods primarily emphasize the development of more sophisticated attention modules to improve performance. However, in the context of SSS image object detection, it is equally important to prioritize detection efficiency alongside performance. Therefore, in this paper we make the decision to introduce a lightweight attention module to the neck component in a plug-and-play manner, aiming to strike a balance between model performance and complexity. Among the lightweight attention modules, the most popular methods are SE Attention [27] and CBAM [28]. However, the SE module primarily focuses on inter-channel information encoding, overlooking the importance of location information, while CBAM only captures local correlations and lacks the ability to capture the essential long-range dependencies required for visual tasks.

The Coordinate Attention (CA) module [26] splits channel attention into two 1D feature coding processes that aggregate features along different directions, capturing long-range dependencies along one spatial direction and retaining precise location information along the other spatial direction, thereby effectively integrating spatial coordinate information into the generated attention map, as shown in Figure 2.

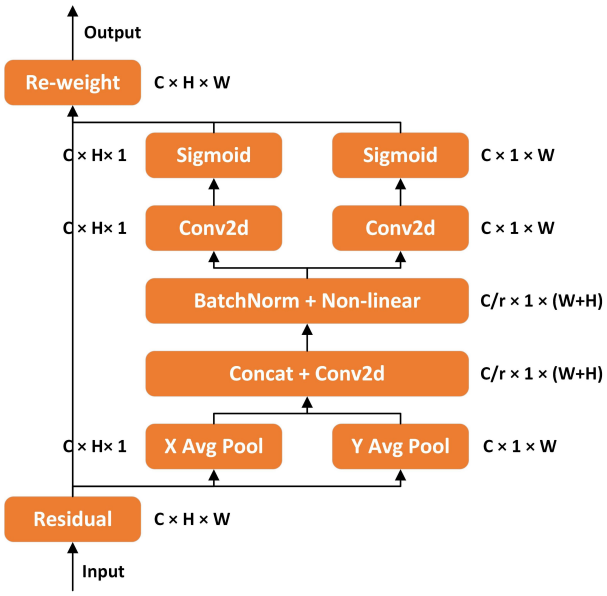


Figure 2. Network structure of the CA module; 'X Avg Pool' and 'Y Avg Pool' refer to 1D horizontal global pooling and 1D vertical global pooling, respectively.

The CA mechanism addresses the encoding of channel relationships and long-range dependencies by incorporating precise position information, which involves two distinct steps: coordinate information embedding, and the generation of coordinate attention. In the first step, coordinate information embedding uses the two spatial ranges $(H, 1)$ or $(1, W)$ of the pooled kernel to encode each channel along the horizontal and vertical coordinates, respectively. Thus, the output of the c-th channel at height h and width w can be formulated as Equation (1):

$$\begin{aligned} z_c^h(h) &= \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \\ z_c^w(w) &= \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \end{aligned} \tag{1}$$

where H and W are the height and width of the input feature, while x_c is the input directly from a convolution layer with a fixed kernel size. In the second step, the aggregate feature graphs obtained from Equation (1) are first concatenated as inputs; then, through a shared 1×1 convolution transform function F_1, the output is as follows:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{2}$$

where $[\cdot, \cdot]$ is the concatenation operation along the spatial dimension and δ is the nonlinear activation function. Then, f is divided into two separate tensors f^h and f^w along the spatial dimension. Two other 1×1 convolution transformations F_h and F_w are respectively used to transform f^h and f^w into tensors with the same number of channels as the input, yielding

$$\begin{aligned} g^h &= \sigma\left(F_h\left(f^h\right)\right), \\ g^w &= \sigma\left(F_w\left(f^w\right)\right), \end{aligned} \tag{3}$$

where σ is the sigmoid function. Finally, the input g^h and g^w are used as the attention weights. The output of CA module y can be formulated as Equation (4):

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j). \tag{4}$$

2.3. CoT Module

The transformer self-attention mechanism has gained significant traction in the field of natural language processing, demonstrating competitive results. Motivated by this success, researchers have started to investigate the applicability of self-attention mechanisms in computer vision tasks. However, most current studies primarily focus on utilizing self-attention on 2D feature graphs, which generates an attention matrix based on isolated query and key pairs at each spatial location without fully leveraging the contextual information available among adjacent keys.

The Contextual Transformer (CoT) module [25] leverages the inherent static contextual relationship among input keys to facilitate the learning of a dynamic attention matrix, and finally fuses the static and dynamic context information to enhance the representation of visual features. In computer vision tasks, CoT can serve as an alternative to standard convolution, thereby enhancing self-attention for contextual information that is lacking in backbone networks. More specifically, we can assume that the input 2D feature map is $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ and that the keys, queries, and values are defined as $K = X$, $Q = X$, and $V = XW_v$, respectively. First, $k \times k$ group convolution is performed for all adjacent keys in the $k \times k$ grid; the obtained keys $K_1 \in \mathbb{R}^{H \times W \times C}$ contain the context information of their adjacent keys. Then, the contextualized keys K_1 and queries Q are taken as inputs; after two 1×1 convolution, the attention matrix can be formulated as Equation (5):

$$A = [K_1, Q] W_\theta W_\sigma \tag{5}$$

where W_θ represents 1×1 convolution with the ReLU activation function and W_σ represents the same without the activation function. Next, according to the obtained attention matrix, as in typical self-attention, the aggregate values V are used to calculate the dynamic context representation K_2 of the inputs; this can be formulated as Equation (6):

$$K_2 = V * A \tag{6}$$

where $*$ denotes the local matrix multiplication. The final output of the CoT module **Y** is a fusion of the static context K_1 and dynamic context K_2, as shown in Figure 3.

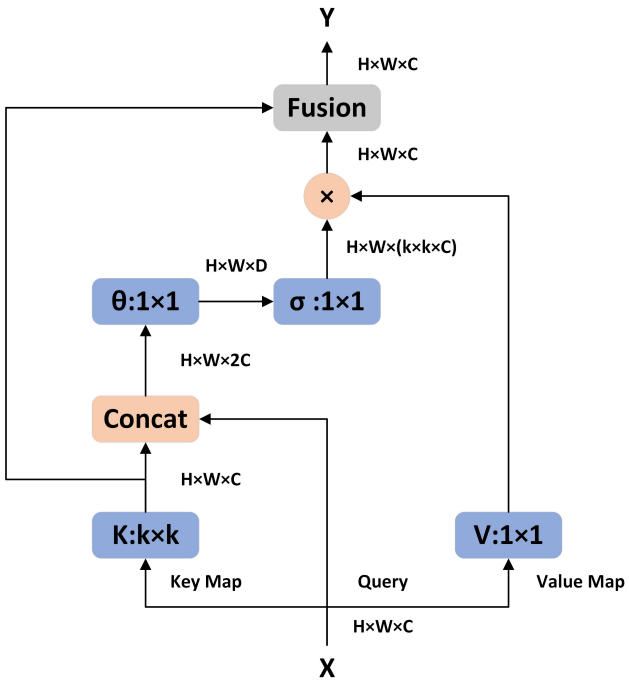


Figure 3. Network structure of the CoT module, where '×' denotes the local matrix multiplication.

2.4. BiFPN Structure

Effectively representing and processing multi-scale features in a network poses a challenging problem in object detection. Initially, object detection networks directly utilized features extracted from the backbone network for direct prediction. Subsequently, the Feature Pyramid Network (FPN) [11] introduced a top-down approach to fuse multi-scale features, establishing the groundwork for multi-scale feature fusion. Building upon this, PANet [29] augmented the basic FPN architecture with an additional bottom-up feature fusion network. PANet has gained significant prominence in recent years, and has been employed as the neck component in YOLOv5 and YOLOv7. However, improved performance inevitably leads to an increase in parameters and computational complexity.

Different from the above two approaches, BiFPN [18] proposes a simple and efficient weighted bidirectional feature pyramid network that can easily and quickly carry out multi-scale feature fusion, allowing the accuracy and efficiency of the detection network to be improved at the same time. First, BiFPN removes nodes that contribute less to the network in order to achieve greater simplification. Second, an additional feature connection is added between the input node and the output node on the same scale to fuse more features without adding much cost, as shown in Figure 4.

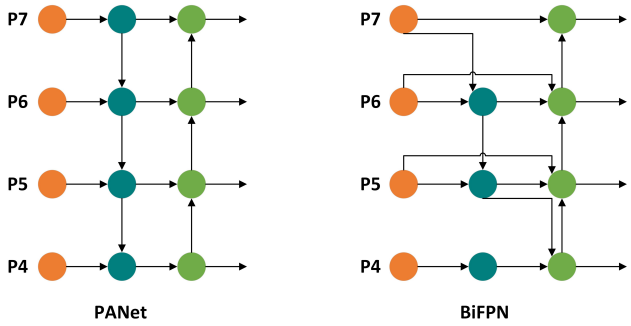


Figure 4. Structural diagrams of PANet and BiFPN; the left side shows PANet and the right side shows BiFPN. This structure is employed in the subsequent sections to enhance the YOLOv7 network.

Moreover, due to the variation in the scales of the fused features, achieving balanced fusion becomes challenging. To address this issue, BiFPN introduces learnable weights for each feature requiring fusion. This enables the model to intelligently utilize features from different scales, which can be mathematically represented as Equation (7):

$$O = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot I_i \tag{7}$$

where $w_i \geqslant 0$, $\epsilon = 0.0001$ is a small value chosen to avoid numerical instability. As a concrete example, the fusion features of the P_6 scale shown in Figure 4 can be formulated as Equation (8):

$$\begin{aligned} P_6^{mid} &= Conv\left(\frac{w_1 \cdot P_6^{in} + w_2 \cdot Resize\left(P_7^{in}\right)}{w_1 + w_2 + \epsilon}\right) \\ P_6^{out} &= Conv\left(\frac{w_1^{'} \cdot P_6^{in} + w_2^{'} \cdot P_6^{mid} + w_3^{'} \cdot Resize\left(P_5^{out}\right)}{w_1^{'} + w_2^{'} + w_3^{'} + \epsilon}\right) \end{aligned} \tag{8}$$

where P_6^{mid} is the intermediate feature of scale 6 in the top-down path, P_6^{out} is the output feature of scale 6 in the bottom-up path, *Resize* is the upsampling or downsampling operation used to unify the feature scale, and *Conv* denotes the convolution operation. Other scale features are fused in the same way.

3. Enhanced YOLOv7-Based Approach

In this section, an enhanced object detection method for SSS (Side Scan Sonar) images based on YOLOv7 is proposed. The method comprises model pretraining, an improved YOLOv7 network, and a modified loss function.

3.1. Model Pretraining

Due to the difficulty, low efficiency, and data sparsity of SSS image acquisition, it is difficult to train high-performance models. Through pretraining on large-scale data, the model learns general feature representations, which can be used as initial parameters or feature extractors for fine-tuning of specific tasks. This can accelerate the training process of the model and improve the generalization ability and performance of the model. Hence, the approach of model migration is adopted, which involves transferring the weight parameters of a model trained on different data for utilization by the object network. To enhance the network's convergence, pretraining weights obtained from widely used optical datasets such as Pascal VOC [30] and MS COCO [31] are employed. This approach aims to leverage potential similarities between acoustic and optical images. Although they represent different perceptual modes, they share a number of different feature representations, making it reliable to use fitted models of optical images as pretrained models of acoustic images.

3.2. Improved YOLOv7 Network

In order to take into account both the general performance and real-time performance of small object detection in SSS images, in this paper we selected the YOLOv7 network as the base network, then made the following improvements. The improved YOLOv7 network is shown in Figure 5.

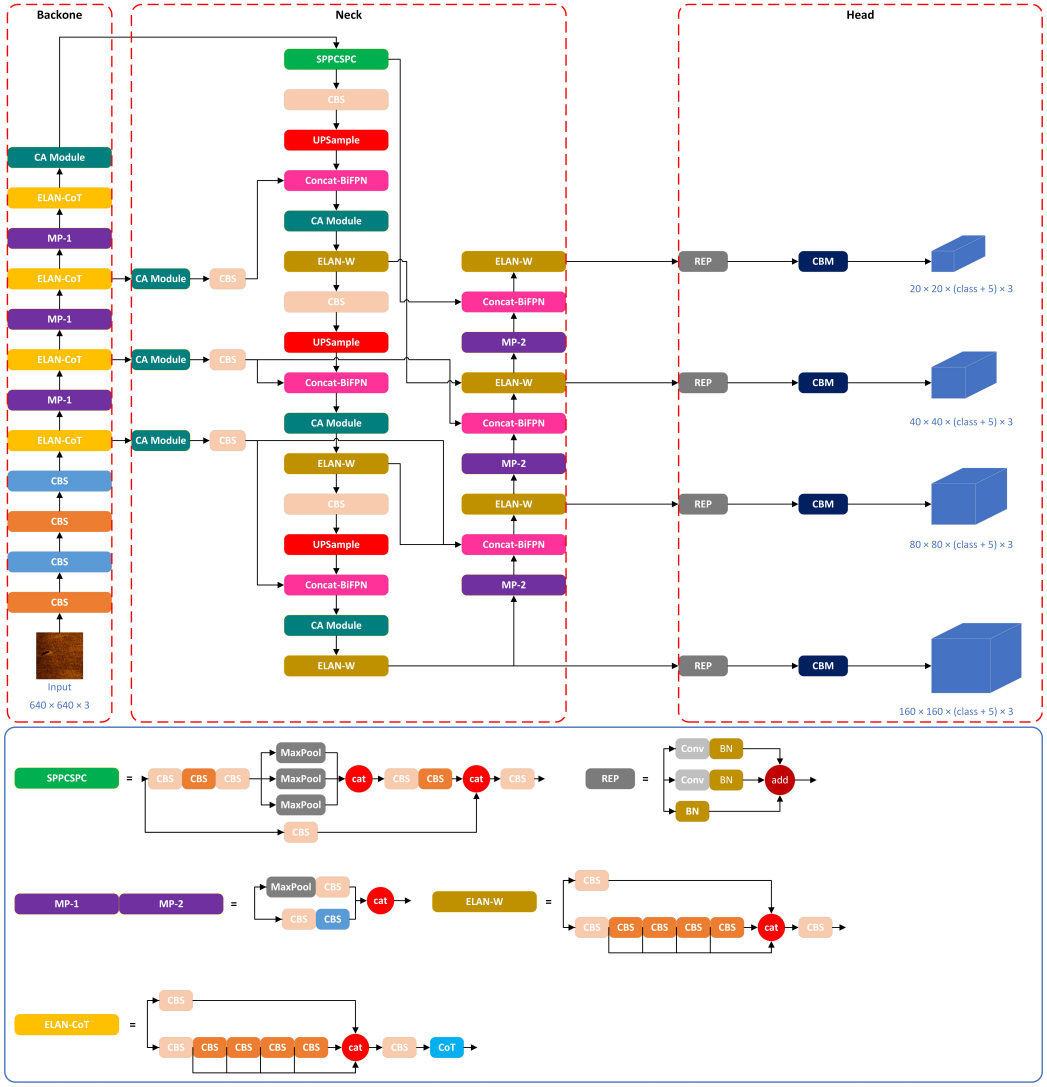


Figure 5. Improved YOLOv7 network. On the basis of the original YOLOv7 network, we add a small object detection layer. Two attention mechanisms are added, one to the backbone part and one to the neck part, and features of different scales are re-fused. The input image size of the network is $640 \times 640 \times 3$. A partial structural diagram of the modules is shown below the figure; only those modules that were changed in the improved network are shown here.

3.2.1. Expanding Detection Scale

In SSS object detection, smaller objects usually require larger detection scale, which is because small objects occupy only a few pixels in the image and require higher spatial resolution to accurately detect, while large objects can be effectively detected at lower spatial resolution. Therefore, the maximum detection scale of 80×80 in the YOLOv7 network cannot meet the requirements of small objects in SSS images. To enhance the network's ability to detect small objects, a detection layer of 160×160 was introduced. This addition aims to improve the network's performance in accurately detecting smaller objects within the given context.

3.2.2. Incorporating Attention Mechanisms

In order for the model to automatically learn the correlation between channels and pay more attention to the important channels, the CA module was incorporated into the neck part of the YOLOv7 network to improve its feature discrimination and generalization abilities. In particular, the CA module was added after each scale feature of the backbone part's output and after the feature fusion module in the top-down process. The CA module takes into account both channel information and orientation-related location information while being flexible and lightweight enough to be easily inserted into the core module of the YOLOv7 network.

Meanwhile, the CoT module was integrated into the ELAN module of the backbone part, which is referred to as ELAN-CoT. This inclusion aims to further enhance the feature extraction capability of the backbone network. The ELAN-CoT module blends context encoders and context attention mechanisms into the transformer model, and can be used as an alternative to standard convolution in the backbone part. This allows the network to capture long-distance dependencies in the sequence, helps the model better understand the relationship between the object and its surroundings, and improves the accuracy of object localization.

3.2.3. BiFPN Feature Fusion

To enhance the efficiency and performance of multi-scale feature fusion, the BiFPN structure was employed to optimize the feature fusion process. BiFPN incorporates an adaptive fusion strategy that dynamically adjusts feature weights across different levels to effectively convey relevant information. In comparison to traditional pyramid structures, BiFPN adopts a more compact design that reduces parameters and computational requirements by weight sharing and feature reuse. This results in improved computational efficiency without compromising performance. Consequently, BiFPN can be seamlessly integrated into the YOLOv7 network, enhancing stability and efficiency during the optimization and training processes while providing researchers with greater flexibility to adapt and enhance specific tasks.

3.3. Loss Function

The loss function of our method consists of three parts: object confidence loss, class confidence loss, and coordinate regression loss. The object confidence loss is used to measure the prediction accuracy of the model for the existence of the object, while the class confidence loss is used to measure the classification accuracy of the model for the object class. Both are calculated using the binary cross-entropy loss, as shown in Equation (9):

$$L^{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{9}$$

where $N = 2$, y_i is the class of sample i and p_i is the predicted value of sample i.

To mitigate the impact of highly competitive anchor frames as well as the adverse gradients caused by low-quality examples, an alternative loss function called the WIOU loss [32] is employed instead of the original CIOU loss for calculating coordinate regression losses. This substitution allows the network to prioritize anchor frames of average quality, leading to potential improvements in the overall performance of the detector. The WIOU loss can be mathematically expressed as Equation (10):

$$\begin{aligned} L^{WIOU} &= \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right) L^{IOU} \\ L^{IOU} &= 1 - \frac{W_i H_i}{wh + w_{gt}h_{gt} - W_i H_i} \end{aligned} \tag{10}$$

where W_g and H_g represents the size of the smallest union box, W_i and H_i are the size of the intersection box, (x, y, w, h) is the center coordinate and size of the prediction box, and $(x_{gt}, y_{gt}, w_{gt}, h_{gt})$ is the center coordinate and size of the ground truth box. Therefore, the final loss function is as follows:

$$\mathcal{L} = w_{obj} \cdot L_{obj}^{BCE} + w_{cls} \cdot L_{cls}^{BCE} + w_{box} \cdot L_{box}^{WIOU} \tag{11}$$

where w_{obj}, w_{cls}, and w_{box} are the loss weight coefficients.

4. Experiments and Analysis

In this section, the proposed method is utilized for conducting experiments and analysis, including the selection of datasets, training procedures, and performance evaluation.

4.1. Datasets

The dataset used in this section was obtained from repeated voyages by our laboratory using a small ship fitted with SSS along a pre-placed object during sea trials. The SSS we used had a frequency of 450 kHz and was able to effectively detect up to 150 m; the installation angle was 20° horizontal downward tilt. The equipment used in the experiment is shown in Figure 6.

Figure 6. The small ship and the SSS used for acquiring the dataset.

The sonar image obtained directly from the SSS is a waterfall stream image, and the resolution of a single image is as high as 1386 × 63,000. To streamline the process of image training and detection, the object portion is extracted from the high-resolution sonar image and used as the input image, then each image is resized to a resolution of 400 × 400.

In this experiment, we placed two kinds of objects in advance, one a cylinder with a base circle diameter of 60 cm and height of 60 cm, and the other a cone with a base circle diameter of 40 cm and height of 30 cm. If the pixel area of an object is less than 1% of the image area, the object can be defined as a small object. Specifically, the SSS image we used had an area of 160,000 pixel^2; thus, when the object pixel area is less than 1600 pixel^2, it is considered a small object in the context of this article. Among the three types of objects in the SSS dataset, the length of three-quarters of the objects is no more than 40 pixels, the width is no more than 20 pixels, and the area of target pixels is no more than 800 pixel^2, allowing the detection performance of small objects in the model to be measured to a certain extent. The graphical representation of this process is depicted in Figure 7.

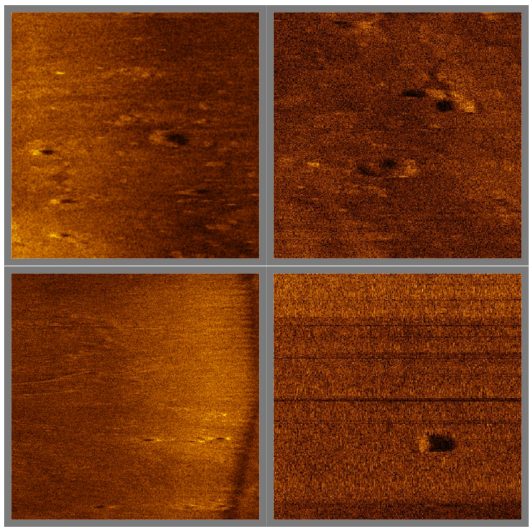


Figure 7. Partial images from the SSS dataset.

After analysis and comparison, the final dataset consisted of 800 SSS images with three types of objects, including 355 cones, 314 cylinders, and 312 non-preset targets for interference. Subsequently, the dataset was divided into training, testing, and validation sets in a ratio of 6:2:2. The dataset created according to this distribution was then used for training the network.

4.2. Training

The experiments are conducted under the PyTorch 1.12.1 framework. The CPU is an Intel(R) Xeon(R) Silver 4210R CPU@2.40 GHz, and four NVIDIA GeForce RTX 3090 (24 GB) are used for the experiment.

During the training phase, the training data were initially augmented using the built-in data augmentation method provided by YOLOv7. Subsequently, both the original YOLOv7 model and the improved YOLOv7 model proposed in this work were trained separately. The input image size was resized to 640 × 640, the batch-size was set to 64, the initial learning rate was set to 0.01 and the the number of training epochs was set to 1000. The SGD optimizer [33] was used, and the warm-up and cosine annealing learning strategies were adopted.

4.3. Performance Evaluation

In this section, the performance of the proposed enhanced YOLOv7-based approach is evaluated. This evaluation encompasses criteria for performance assessment, presentation of experimental results, and an ablation study.

4.3.1. Evaluation Criteria

To assess the performance enhancement achieved by the proposed method, Precision (P), Recall (R), and mean average precision (mAP) metrics were utilized as evaluation criteria, as introduced in PASCAL VOC 2010. The calculation formulas for these metrics are as follows:

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$AP = \int_0^1 P(R)dR \tag{14}$$

$$mAP = \sum_{i=1}^{N} \frac{AP_i}{N} \tag{15}$$

where TP is the number of positive samples that are correctly predicted, FP is the number of samples that are incorrectly predicted as positive, FN is the number of samples that are incorrectly predicted as negative, and N is the number of detected categories.

4.3.2. Experimental Results

To demonstrate the superior performance of the proposed enhanced YOLOv7-based approach, a comparative analysis was conducted with several prevailing methods in the field. The compared methods include SSD [16], Faster R-CNN [10], EfficientDet [18], YOLOv5, and YOLOv7 [19]. The above algorithms were tested on an NVIDIA GeForce RTX 3090 GPU (24 GB). Table 1 shows the performance of different mainstream object detection algorithms on the acquired SSS images dataset.

Table 1. The performance of current mainstream object detection algorithms and our proposed method tested on the SSS dataset, showing the results of each experiment.

Methods	Precision/%	Recall/%	mAP@.5/%	mAP@.5:.95/%	Average Speed/ms
SSD	82.8	79.3	78.3	46.1	59.6
Faster R-CNN	82.5	77.8	78.6	46.5	142.8
EfficientDet	89.5	82.6	81.3	50.1	18.4
YOLOv5	89.8	82.1	80.3	49.8	8.7
YOLOv7	90.7	81.9	80.2	50.3	**7.3**
Our approach	**95.5**	**87.0**	**86.9**	**55.1**	63.1

According to Table 1, the SSD algorithm exhibits the lowest performance among the five mainstream object detection algorithms, with an mAP value of 46.1%. In contrast, YOLOv7 demonstrates the best performance among these algorithms, with an mAP value of 50.3%. These findings further validate the rationale behind our improvements based on YOLOv7. In comparison to YOLOv7, the proposed method showcases notable advancements. Notably, the precision (P) increases by 4.8%, reaching 95.5%. The recall (R) demonstrates an improvement of 5.1%, reaching 87.0%. Additionally, the mean Average Precision (mAP) sees a substantial 4.8% rise, reaching 55.1%.

In terms of detection speed, YOLOv7 achieves the highest detection speed, detecting an image in 7.3 ms, while Faster R-CNN has the slowest detection speed of 142.8 ms. The method proposed in this paper achieves a detection speed of 63.1 ms, which represents a trade-off between efficiency and accuracy. However, it is important to note that in terms of real-time performance, the generation time for each data ping from the side-scan sonar ranged from 0.6 s to 0.8 s. Consequently, it takes at least 240 s to generate the image size required for detection according to the methodology proposed in this paper. This time requirement significantly surpasses the 63.1 ms achieved by the proposed method. Nevertheless, the method outlined in this paper adequately fulfills the real-time demands of the project.

Figure 8 shows ground truth labels from part of the SSS dataset. These images were employed to assess and verify the performance improvement of the YOLOv7 network.

Figure 9 shows partial detection results before and after the improvements made to the YOLOv7 network. It can be observed that when using the same SSS images to test the network before and after the improvement, our method can detect more objects, especially small objects, which indicates that our improvements enhance the ability of the model to detect such small objects.

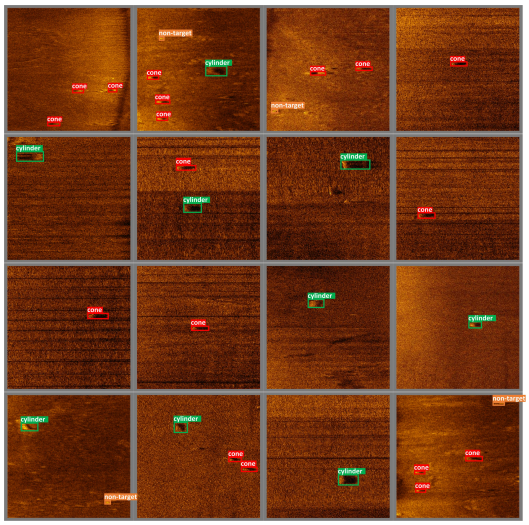
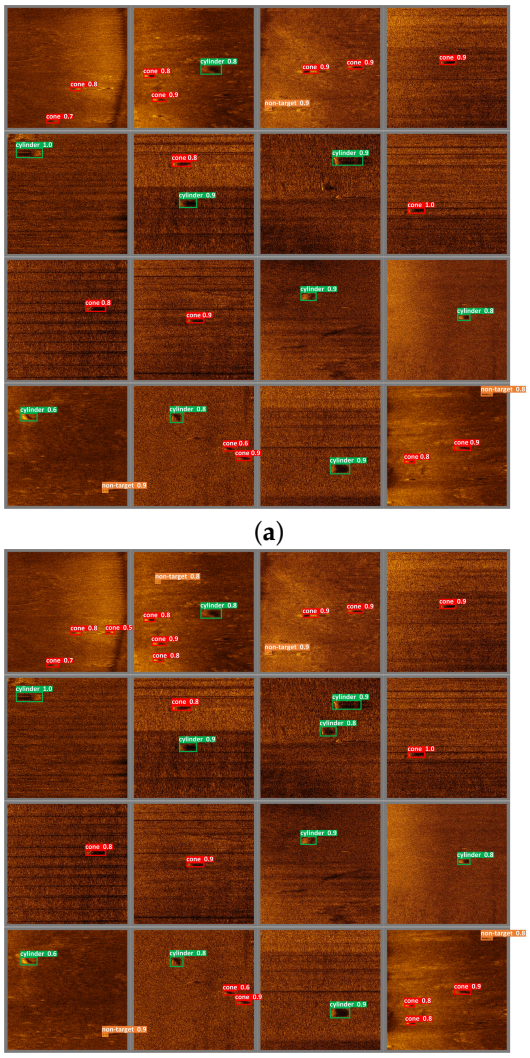


Figure 8. Ground truth labels.

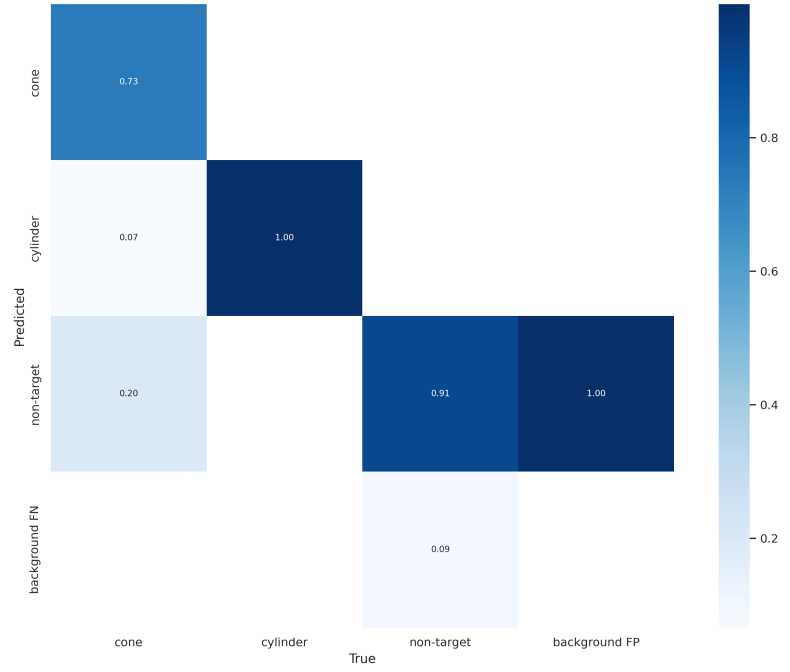


(**a**)

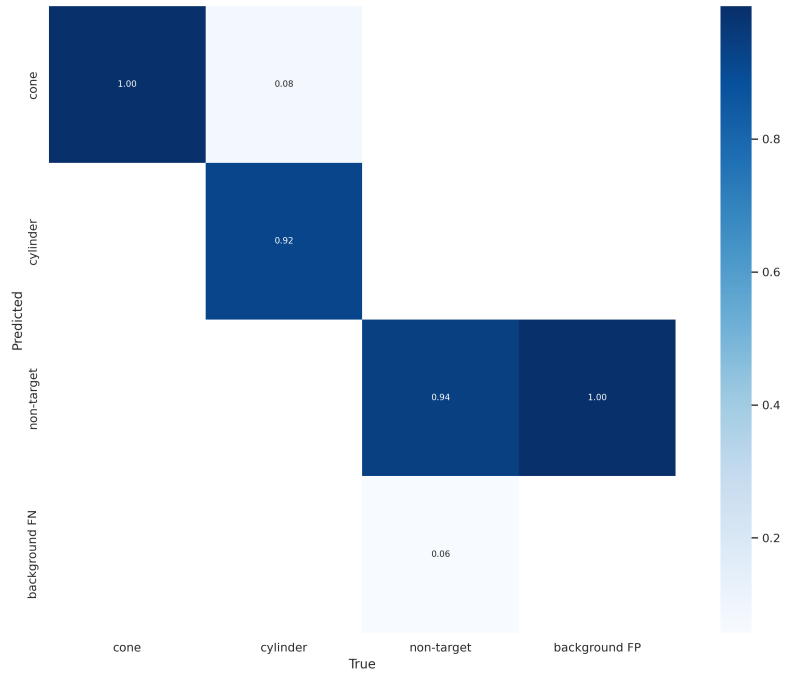
(**b**)

Figure 9. Comparison of detection results between YOLOv7 and our method on the test set: (**a**) object detection test results with YOLOv7 and (**b**) object detection test results with our method.

Furthermore, to conduct a comprehensive comparison between YOLOv7 and our method, the confusion matrices and PR curves of both models were compared. As shown in Figure 10, our method exhibits higher accuracy and better balance in detecting both the object class and the background class.



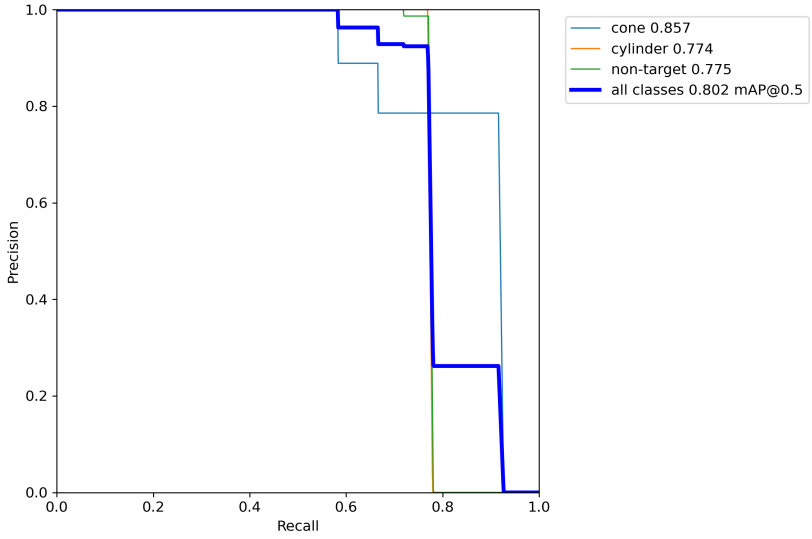
(**a**)



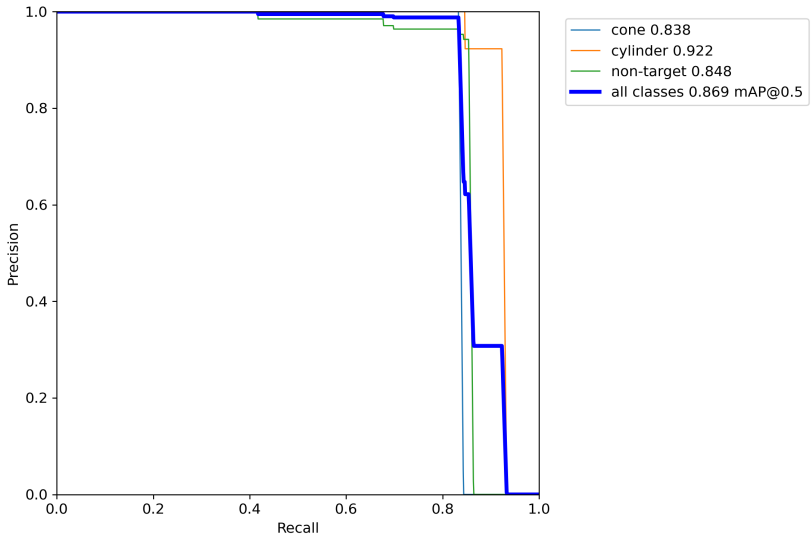
(**b**)

Figure 10. Comparison of confusion matrices for YOLOv7 and our method: (**a**) confusion matrix with YOLOv7 and (**b**) confusion matrix with our method.

Figure 11 showcases the enhanced detection performance of diverse objects accomplished through our method; it can be seen that the mAP@0.5 value achieves a remarkable growth of 6.7%.



(**a**)



(**b**)

Figure 11. Comparison of PR curves for YOLOv7 and our method: (**a**) PR curves with YOLOv7 and (**b**) PR curves with our method.

Based on the experimental results presented above, it can be concluded that our method greatly enhances the detection performance of SSS images, making it particularly well-suited for detecting small underwater objects. However, the inherent nature of deep learning heavily relies on extensive data, and the limited availability of large-scale SSS image datasets poses a challenge for training our method on a sufficient number of samples. This limitation inevitably impacts the scope and generalizability of our experimental results. For instance, Figure 12 shows that our method exhibits false detections in comparison to the original YOLOv7 method. Although our method improves the detection performance of the model, it brings about new special cases of instability. This situation can be attributed to the limited size of the dataset, which results in the attention mechanism excessively prioritizing the shadow portion of the object. As a consequence, false detections may occur. The study of sonar image data expansion is a hot topic today, which can provide a way to further optimize our methods. Our future research will focus on the expansion and enhancement of sonar images to make it easier for the network to learn the features of the object.

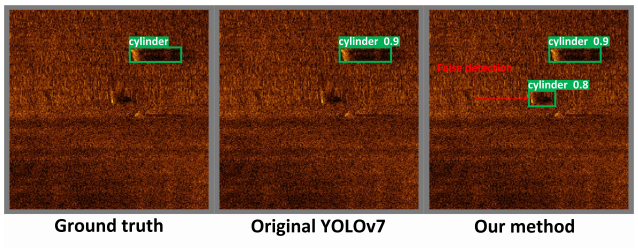


Figure 12. False detection of objects in SSS images.

4.3.3. Ablation Study

To assess the effectiveness and individual impact of each of the proposed innovations on network performance, an ablation study was conducted. This study mainly makes four improvements to the original YOLOv7 network, including expanding the detection scale, adding a CoT module, adding a CA module, and using BiFPN feature fusion. The four proposed improvements were incrementally incorporated into the original YOLOv7 network. The SSS image dataset was then used to conduct experiments while evaluating the impact of each improvement in a step-by-step manner. The results of each experiment are shown in Table 2.

Table 2. Results of the ablation study. The improvements from top to bottom are added to the original YOLOv7 network, and the results are recorded each time.The innovation points proposed in this paper are added successively from top to bottom, and the order and content of addition are expressed in the form of checkmark.

Improvements	i	ii	iii	iv	v	Precision/%	Recall/%	mAP@.5/%	mAP@.5:.95/%
Original YOLOv7 (i)	✓					90.7	81.9	80.2	50.3
Expanding detection scale (ii)	✓	✓				92.3	82.1	82.9	51.8
CoT module (iii)	✓	✓	✓			93.1	84.1	84.1	53.5
CA module (iv)	✓	✓	✓	✓		95.1	87.0	86.6	54.2
BiFPN feature fusion (v)	✓	✓	✓	✓	✓	95.5	87.0	86.9	55.1

As depicted in Table 2, each proposed improvement integrated into the YOLOv7 model contributes to the enhancement of the network's detection performance to a different extent. Specifically, after expanding the detection scale, the mAP value sees a 1.5% increase, validating the introduction of a 160 × 160 scale detection layer to effectively enhance the network's recognition capability. After adding the CoT module, the mAP value increases by 1.7%, which is the largest single improvement, while after adding the CA module the mAP value increases by 0.7%. These results show that the improvement of the attention mechanism on the network performance is closely related to the location of the addition. Lastly, the incorporation of BiFPN feature fusion contributes to a 0.9/% increase in the mAP value, affirming that the utilization of learnable weights enhances the reasonableness and reliability of feature fusion. Overall, the mAP value experiences a noteworthy 4.8% improvement, confirming the effectiveness of each proposed enhancement.

5. Conclusions

In this paper, we have introduced a novel object detection method for small objects in SSS images, which we refer to as the enhanced YOLOv7-based approach. Specifically, our approach makes the following major improvements. (1) An additional detection layer with a scale of 160 × 160 is incorporated into the existing three detection layers of YOLOv7. This enhancement aims to specifically improve the detection capability for small objects. (2) A CoT module is integrated into the ELAN module to enhance the network's feature representation. The self-attention mechanism of the CoT module is leveraged for this purpose. (3) In the neck section, an additional CA module is introduced to guide the network's attention towards the essential features present in the image, thereby promoting effective learning.

(4) Utilizing the BiFPN structure as the foundation, a novel feature fusion approach is applied to address the challenge of balancing features across various scales. Our experiments and ablation study provide compelling evidence that the proposed method outperforms mainstream object detection algorithms. In comparison to the original YOLOv7 network, the Precision shows a remarkable improvement of 4.8%, achieving an impressive accuracy of 95.5%. Furthermore, the Recall exhibits a notable enhancement of 5.1%, reaching a commendable level of 87.0%. The mAP@.5 showcases a substantial improvement of 6.7%, resulting in an impressive mAP score of 86.9%. Moreover, the mAP@.5:.95 reaches an outstanding 55.1%, indicating a significant boost of 4.8%. Overall, our proposed method proves effective in delivering these substantial improvements. The results indicate that our method is more suitable for autonomous detection of small underwater objects, and provides a innovative approach to object detection based on SSS images.

In our future work, we intend to focus on two main areas: developing intelligent algorithms to generate high-quality sonar images in order to expand the SSS data and enrich the dataset, and exploring the influence of ocean currents on side-scan sonar images while developing advanced image processing techniques to reduce interference. These efforts have the aim of significantly improving the effectiveness and stability of detection in side-scan sonar applications.

Author Contributions: Methodology, F.Z. and W.Z.; data curation, C.C. (Chun Cao); writing advice, C.C. (Chensheng Cheng) and X.H.; device support, F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This reasearch was funded by the National Natural Science Foundation of China (52171322) and the Graduate Innovation Seed Fund of Northwestern Polytechnical University (PF2023066, PF2023067).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to acknowledge the facilities and technical assistance provided by the Key Laboratory of Unmanned Underwater Transport Technology.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SSS	Side-Scan Sonar
mAP	Mean Average Precision
CoT	Contextual Transformer
ELAN	Efficient Layer Aggregation Network
SE	Squeeze-and-Excitation
CBAM	Convolutional Block Attention Module
CA	Coordinate Attention
FPN	Feature Pyramid Network
PANet	Path Aggregation Network
BiFPN	Bidirectional Feature Pyramid Network
CIOU	Complete-IOU
WIOU	Wise-IoU

References

1. Qin, X.; Luo, X.; Wu, Z.; Shang, J. Optimizing the sediment classification of small side-scan sonar images based on deep learning. *IEEE Access* **2021**, *9*, 29416–29428. [CrossRef]
2. Yu, Y.; Zhao, J.; Gong, Q.; Huang, C.; Zheng, G.; Ma, J. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. *Remote Sens.* **2021**, *13*, 3555. [CrossRef]
3. Yu, F.; He, B.; Li, K.; Yan, T.; Shen, Y.; Wang, Q.; Wu, M. Side-scan sonar images segmentation for AUV with recurrent residual convolutional neural network module and self-guidance module. *Appl. Ocean Res.* **2021**, *113*, 102608. [CrossRef]

4. Chen, Z.; Wang, H.; Shen, J.; Dong, X. Underwater object detection by combining the spectral residual and three-frame algorithm. In *Advances in Computer Science and Its Applications: CSA 2013*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1109–1114.
5. Mukherjee, K.; Gupta, S.; Ray, A.; Phoha, S. Symbolic analysis of sonar data for underwater target detection. *IEEE J. Ocean Eng.* **2011**, *36*, 219–230. [CrossRef]
6. Yan, X.; Li, J.; He, Z. Measurement of the echo reduction for underwater acoustic passive materials by using the time reversal technique. *Chin. J. Acoust.* **2016**, *35*, 309–320.
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
9. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; Volume 28.
11. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
12. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
13. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
14. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10213–10224.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part I 14; Springer: Cham, Switzerland, 2016; pp. 21–37.
17. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
18. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
19. 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
20. Yang, Z.; Zhao, J.; Zhang, H.; Yu, Y.; Huang, C. A Side-Scan Sonar Image Synthesis Method Based on a Diffusion Model. *J. Mar. Sci. Eng.* **2023**, *11*, 1103. [CrossRef]
21. Fu, S.; Xu, F.; Liu, J.; Pang, Y.; Yang, J. Underwater small object detection in side-scan sonar images based on improved YOLOv5. In Proceedings of the 2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS), Zhoushan, China, 22–24 April 2022; pp. 446–453.
22. Wang, Z.; Zhang, S.; Huang, W.; Guo, J.; Zeng, L. Sonar image target detection based on adaptive global feature enhancement network. *IEEE Sensors J.* **2021**, *22*, 1509–1530. [CrossRef]
23. Zhang, H.; Tian, M.; Shao, G.; Cheng, J.; Liu, J. Target detection of forward-looking sonar image based on improved YOLOv5. *IEEE Access* **2022**, *10*, 18023–18034. [CrossRef]
24. Li, L.; Li, Y.; Yue, C.; Xu, G.; Wang, H.; Feng, X. Real-time underwater target detection for AUV using side scan sonar images based on deep learning. *Appl. Ocean Res.* **2023**, *138*, 103630. [CrossRef]
25. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef] [PubMed]
26. 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, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
29. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, 18–23 June 2018; pp. 8759–8768.
30. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]

31. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part V 13; Springer: Cham, Switzerland, 2014; pp. 740–755.
32. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
33. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris, France, 22–27 August 2010*; Keynote, Invited and Contributed Papers; Springer: Berlin/Heidelberg, Germany, 2010; pp. 177–186.