Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060547

ARTICLE

ProNet: Underwater Forward-Looking Sonar Images Target Detection Network Based on Progressive Sensitivity Capture

Kaiqiao Wang[1,2], Peng Liu[1,2,*] and Chun Zhang[1,2]

[1]Hainan Acoustics Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Haikou, 570105, China
[2]Marine Information, Hainan Observation and Research Station, Lingshui, 572423, China

*Corresponding Author: Peng Liu. Email: liup@dsp.ac.cn

Received: 04 November 2024; Accepted: 25 December 2024; Published: 06 March 2025

ABSTRACT: Underwater target detection in forward-looking sonar (FLS) images is a challenging but promising endeavor. The existing neural-based methods yield notable progress but there remains room for improvement due to overlooking the unique characteristics of underwater environments. Considering the problems of low imaging resolution, complex background environment, and large changes in target imaging of underwater sonar images, this paper specifically designs a sonar images target detection Network based on Progressive sensitivity capture, named ProNet. It progressively captures the sensitive regions in the current image where potential effective targets may exist. Guided by this basic idea, the primary technical innovation of this paper is the introduction of a foundational module structure for constructing a sonar target detection backbone network. This structure employs a multi-subspace mixed convolution module that initially maps sonar images into different subspaces and extracts local contextual features using varying convolutional receptive fields within these heterogeneous subspaces. Subsequently, a Scale-aware aggregation module effectively aggregates the heterogeneous features extracted from different subspaces. Finally, the multi-scale attention structure further enhances the relational perception of the aggregated features. We evaluated ProNet on three FLS datasets of varying scenes, and experimental results indicate that ProNet outperforms the current state-of-the-art sonar image and general target detectors.

KEYWORDS: Forward-looking sonar image; target detection; subspace decomposition; progressive sensitivity capture

1 Introduction

Underwater target detection aims to identify targets in underwater environments. This study has consistently attracted interest because of its broad applications in areas like oceanography [1], underwater navigation [2], and aquaculture [3]. Compared to light and electromagnetic waves, the attenuation of acoustic waves during observation and measurement in water is much smaller. Therefore, sonar devices that utilize the propagation characteristics of sound waves in water have become some of the most effective underwater detection equipment available today. Among these, forward-looking sonar (FLS) is widely used because of its advantages such as fast imaging speed and independence from platform movement. FLS emits sound waves and receives echoes to map the pixel intensity values in sonar images. Strong echoes correspond to high pixel intensity, while weak echoes exhibit low pixel intensity. By processing these FLS sonar images, rapid localization and detection of underwater targets can be achieved.

A conventional approach to target detection in sonar imaging involves the manual design of template matching rules that are informed by prior knowledge [4–7]. The matching methods include setting grayscale threshold, brightness, shadow areas and combining multiple templates, etc. This process necessitates the pre-designation of the most effective and sensitive features pertinent to the specific task at hand. While such methods can yield satisfactory detection performance in particular scenarios, they are inherently constrained by low transferability and generalization across varied contexts. Moreover, the challenge of acquiring sufficiently accurate prior knowledge exacerbates these issues, as it remains theoretically unattainable to achieve a perfect understanding of all possible variations within sonar data. Consequently, this gap underscores a significant potential for performance enhancement in existing methodologies.

As the field evolves, researchers are increasingly recognizing the necessity to move beyond traditional template-based approaches. Recent advancements in deep learning techniques, particularly those employing neural networks, have shown great promise in improving the robustness and adaptability of target detection systems. By harnessing large datasets and leveraging complex algorithms, deep learning models can autonomously learn intricate patterns and features from raw sonar data, thereby mitigating the dependency on predefined templates. Notably, in visual target detection tasks, both one-stage algorithms (such as SSD [8] and YOLO [9]) and two-stage algorithms (such as Faster R-CNN [10] and Mask R-CNN [11]) have drawn significant attention from researchers due to their respective advantages in computational complexity and detection accuracy.

Inspired by this, many scholars have transferred relevant ideas from vision-based target detection methods to sonar image target detection tasks, achieving certain results. For example, References [12–15] has combined YOLOv2 [16], YOLOv3 [17], and YOLOv5 [18] with other neural structures to automatically learn effective high-dimensional features in different environments. However, this paper argues that the aforementioned deep learning methods still have room for performance improvement because they do not fully consider the characteristics of underwater tasks. Compared to general vision-based target detectors, there are several significant issues specific to sonar image-based target detectors.

- **Low imaging resolution:** The resolution of underwater sonar imaging is generally low, which results in the loss of detail and structural texture information, making it very difficult to distinguish between different targets. Meanwhile, due to the unique nature of the underwater environment, the boundaries of the targets are often not clear, affecting the accuracy of target recognition.
- **Complex background environment:** The current underwater environmental background is influenced by various factors such as water temperature, water quality, and depth. These factors can increase the reflection noise during the sound wave propagation process, adding extra difficulties to detection. This noise not only obscures the target signal but may also disguise itself as a target, further increasing the complexity of detection.
- **Impact of incident angle on imaging effect:** Due to the limitations of the beam's incident angle, even in the same hydrological environment, the imaging effect of the same target can vary dramatically under different orientations. This variation may lead to significant differences in the target's shape, size, and brightness, posing challenges to the stability and accuracy of automated detection systems.

To solve the above problems as much as possible, this paper specifically designed a sonar image target detection network named ProNet. The ProNet adopts a progressive way to gradually focus on the sensitive features of each stage of the backbone network, thereby improving the adaptability of the network to different target features and noise environments. Specifically, the network first decomposes the input features of each stage into multiple heterogeneous subspaces, processes these subspaces in parallel using different local receptive fields, and finally aggregates the effective information obtained from different subspaces. By repeatedly performing this operation in the network, the capture of sensitive areas of features is gradually achieved. The primary contributions of this paper are outlined below:

- We propose a basic block for the backbone network, which can decompose input features into multiple heterogeneous subspaces, process these subspaces in parallel using different local receptive fields, and finally aggregate the effective information obtained from different subspaces.
- Based on the proposed basic blocks, we designed a sonar image target detection network named ProNet. We conducted experiments on different FLS image datasets MDD and WHD, and the experimental results showed that our proposed network performed better than existing state-of-the-art methods.

2 Related Work

Early sonar image target detection methods used manually designed features for template matching to achieve target localization and segmentation [19], by comparing whether the pixel grayscale of sonar images exceeds a set detection threshold to implement a constant false alarm rate algorithm [20]. For instance, Dobeck et al. [21] utilized prior knowledge to create a target template with high brightness areas, shadow areas, and front and rear background areas, and conducted matched filtering detection on regions of interest (ROI) in large-scale side view sonar images. Subsequently, utilizing the intensity, shape, and size of target highlights and shadows, design features were created, and K-Nearest Neighbor classification was employed for secondary confirmation of ROI, leading to the recognition of mine-like objects. Williams [22] developed an unsupervised rapid target detector for large-scale synthetic aperture sonar images, utilizing a cascaded architecture and integral images to accelerate detection speed, achieving near real-time detection. Hurtós et al. [23] proposed a model suitable for cascading front view acoustic images. Each layer of the model creates a template and uses template matching method to achieve target discrimination of front view acoustic images. Although these methods can effectively identify targets under certain conditions, their limitations are evident. For example, manual parameter adjustments are often required, relying on researcher's prior knowledge, and their adaptability to complex scenes is limited, making it difficult to achieve automatic learning and optimization.

The mainstream sonar image target detection DL-based methods can be divided into two-stage methods and one-stage methods. The two-stage method typically first generates a series of candidate regions in the first stage, which may contain the target to be detected; Then, in the second stage, these candidate regions are classified and the bounding boxes are fine-tuned. For instance, Jiang et al. [24] proposed an active target detection method for sunken ships detection that selects key images as training samples. This improves performance and reduces labeling costs, but is time-consuming. Furthermore, Zhang et al. [25] proposed a self-training method that builds and optimizes a backbone network on the classification dataset. Using this network, they created an automated detector that achieved good results on a small sonar image dataset. Fan et al. [26] designed a feature extraction network that uses residual blocks instead of the backbone in mask-based convolutional neural networks, reducing network parameters without affecting accuracy. Long et al. [27] developed a denoising auxiliary network to reduce speckles in FLS images. They also introduced a feature selection strategy using scene priors to eliminate features that don't match the target size, reducing redundant anchor points and improving detection speed. Although two-stage detection methods can handle targets of different scales and shapes with high accuracy, their detection speed is usually inferior to that of one-stage methods.

To further meet the real-time requirements of sonar detection, researchers focus on exploring one-stage methods to achieve sonar image detection, which considers the detection task as a whole regression problem. Kong et al. [12] proposed a dual path feature fusion network for feature extraction, which achieves robust and real-time sonar target detection. Gao et al. [28] embedded a coordinate attention mechanism in the YOLOv5 backbone network and used transposed convolution in the neck network to achieve higher upsampling performance, thereby improving detection accuracy. Qin et al. [29] introduced an attention mechanism into the backbone network of the YOLOV7 model and integrated Multi-GnBlock blocks in the Neck, improving the model's ability to handle complex backgrounds in sonar images. As shown in Table 1, we organized the relevant methods mentioned above.

Table 1: Summary of related works

Method	Type	Features	Limitations
[19]	Handcrafted sonar image target detection	Template matching based on pixel grayscale threshold comparison	These methods depend on prior knowledge, which limits their adaptability to complex scenes. As a result, achieving automatic learning and optimization becomes challenging.
[20]		Target detection using constant false alarm rate algorithm	
[21]		Creating a target template with high brightness areas, shadow areas, and front and rear background areas	
[22]		Utilizing a cascaded architecture and integral images to enable real-time detection	
[23]		Using model suitable for cascading images and creating a multi-layer template	
[24]	Two-stage detection method based on deep learning	Introducing an iterative training mechanism to improve detection performance	Although two-stage detection methods can handle targets of different sizes and shapes with high accuracy, their detection speed is usually inferior to that of one-stage methods
[25]		Proposing a self-training strategy that automatically constructs and optimizes a backbone network	
[26]		Designing a feature extraction network based on residual blocks to reduce network parameters without affecting accuracy	
[27]		Introducing a denoising auxiliary network and filing out the size-mismatched feature levels	
[12]	One-stage detection method based on deep learning	Proposing a dual path feature fusion network to reach robust and real-time detection	The detection accuracy is slightly inferior to the two-stage methods
[28]		Combining transposed convolution and YOLOv5 for higher up-sampling performance	
[29]		Combining Multi-GnBlock blocks and YOLOv7 to handle complex backgrounds in sonar images	
[30]		Lightweight downsampling and feature extraction, focusing on high-resolution shallow layers, enhance detection accuracy and speed.	
Ours	Pronet	Proposing a novel sonar image target detection method based on progressive sensitivity capture	Improving the performance of existing one-stage methods

On the one hand, the recent YOLOv10 [31] has achieved excellent performance in general target detection tasks through various component optimization strategies, providing a new foundational technology for sonar image target detection tasks. On the other hand, the recent although DL-based networks have made great progress in sonar image target detection, the neglect of effectively utilizing rich spatial information and their interaction hinders further improvement of these methods. Therefore, this paper builds a novel target detection network for FLS sonar images based on YOLOv10. The network decomposes the features of each stage into multiple heterogeneous subspaces, uses multiple kernel sizes to process the subspace features separately, and obtains multi-scale spatial features with rich spatial information. Then, a convolutional modulation network [32] is used to execute the relevant interaction of these subspaces, thereby achieving higher accuracy in FLS image target detection.

3 Proposed Method

3.1 Overall Architecture

Our proposed underwater forward-looking sonar image target detection network based on progressive sensitivity capture, named ProNet, has its overall architecture as shown in Fig. 1. The design of the architecture is inspired by the ViT [33] and YOLOv10 [31]. We designed a backbone for feature extraction and a detection head for outputting detection results, where an input FLS image is processed to produce an output FLS image with target classes and detection boxes. Specifically, the backbone network we designed is composed of basic blocks in four stages, with adoption rates of $\{4, 8, 16, 32\}$ and layer depths as $\{L_1, L_2, L_3, L_4\}$. The detailed structure of each basic block is shown in Fig. 2, which will be discussed in the following sections.

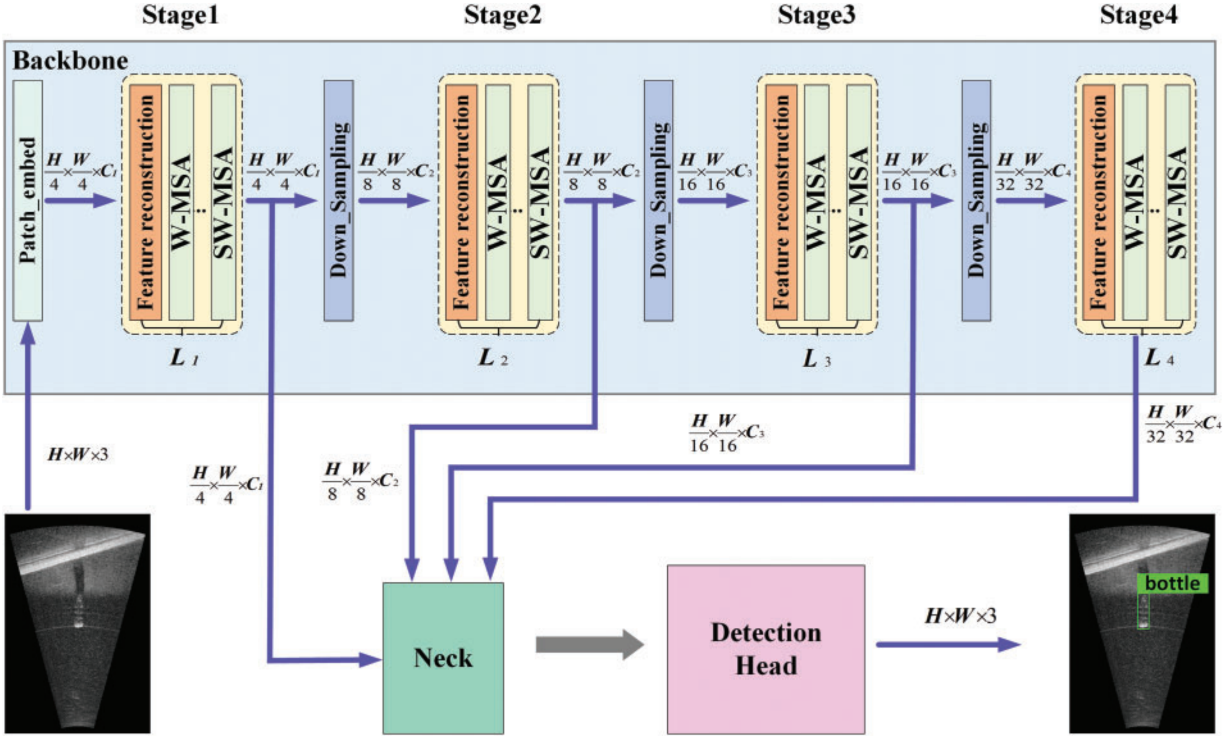

Stage1
Stage2
Stage3
Stage4
Backbone
Patch_embed
$\frac{H}{4}\times\frac{W}{4}\times C_1$
Feature reconstruction
W-MSA
SW-MSA
$\frac{H}{4}\times\frac{W}{4}\times C_1$
Down_Sampling
$\frac{H}{8}\times\frac{W}{8}\times C_2$
Feature reconstruction
W-MSA
SW-MSA
$\frac{H}{8}\times\frac{W}{8}\times C_2$
Down_Sampling
$\frac{H}{16}\times\frac{W}{16}\times C_3$
Feature reconstruction
W-MSA
SW-MSA
$\frac{H}{16}\times\frac{W}{16}\times C_3$
Down_Sampling
$\frac{H}{32}\times\frac{W}{32}\times C_4$
Feature reconstruction
W-MSA
SW-MSA
L_1
L_2
L_3
L_4
$H\times W\times 3$
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
Neck
Detection Head
$H\times W\times 3$
bottle

Figure 1: Overall architecture of our proposed underwater forward-looking sonar images target detection network, called ProNet. It includes a backbone for feature extraction and a detection head for generating output detection results, transforming input FLS images into output images with target classes and detection boxes. The backbone has four stages, each made up of a basic block. The sampling rates for these stages are $\{4, 8, 16, 32\}$, with corresponding layer depths of $\{L_1, L_2, L_3, L_4\}$

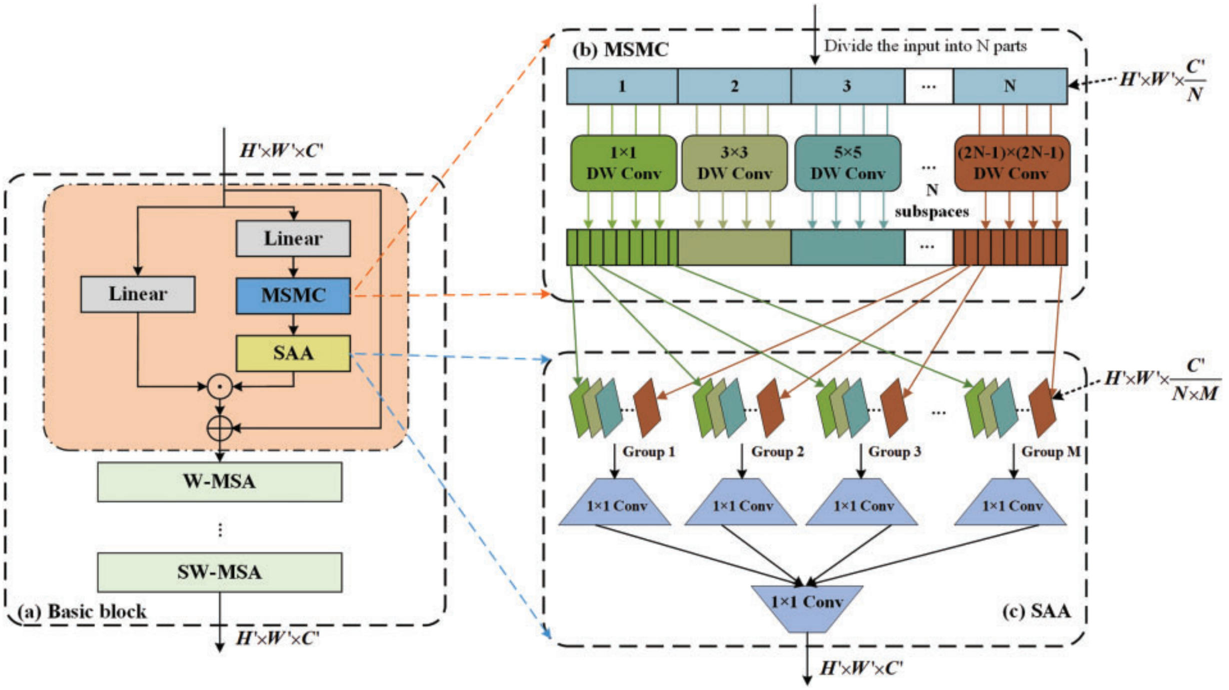


Figure 2: The detailed of the basic block proposed in this paper. (a) The workflow of the proposed basic block. (b) and (c) are the module descriptions of multi-subspace mixed convolution (MSMC) and scale-aware aggregation (SAA), respectively. In the figure, W-MSA denotes Window Multi-Head Self Attention and SW-MSA represents Shifted Window Multi-Head Self Attention. DW-Conv means depth-wise convolution

3.2 Basic Blocks

The basic block structure we proposed is shown in Fig. 2, each basic block is composed of the feature reconstruction proposed in this paper and the functional modules W-MSA and SW-MSA of Swin Transformer [34]. The basic idea of the feature reconstruction process in this paper is to highlight the sensitive regions within the features that may contain targets while filtering out some noise that interfere with subsequent target detection. In terms of specific implementation, this paper constructs a dual-stream network, where branch one consists of a linear layer aimed primarily at accelerating the convergence of the network during the training phase. Branch two is composed of a linear layer, a Multi-Subspace Mixed Convolution (MSMC) module, and a Scale-Aware Aggregation (SAA) module. We have provided detailed descriptions of the structures of the MSMC and SAA modules in Fig. 2. The design idea of this branch is to first use the MSMC module to map the input features into N different subspaces, where convolutional filters with various local receptive fields are employed to extract local associative features at different scales from these N subspaces. Subsequently, for the features obtained from these N subspaces, we utilize the SAA module for group cross-fusion, and finally, the resulting features are element-wise multiplied with the features obtained from the first branch. The subsequent MSA structure is used for further relational perception of the aggregated features.

As shown in Fig. 2a, after capturing multi-scale spatial features using MSMC and aggregating them with SAA, we obtain an output feature map M, which we call a reconstructor. Then, through the scalar product, we use this reconstructor to reconstruct the value V. For the input feature $X \in \mathbb{R}^{H' \times W' \times C'}$, we calculate the output $Z \in \mathbb{R}^{H' \times W' \times C'}$ as follows:

$$Z = M \odot V \tag{1}$$

$$V = W_v X \tag{2}$$

$$M = SAA(MSMC(W_s X)) \tag{3}$$

where ⊙ is element level multiplication, W_v and W_v are the weight matrices of the linear layer. The value of the reconstructor M is determined by MSMC and SAA, and it dynamically adjusts with different inputs to achieve adaptive reconstruction, which focuses on the sensitive areas of the feature maps.

3.2.1 MSMC

We proposed the MSMC to obtain various spatial features across multiple scales. Furthermore, MSMC can expand the receptive field using a large convolutional kernel, enhancing its ability to model long-range dependencies. As shown in Fig. 2b, MSMC divides the input channel into N subspaces and introduces multiple convolutions of different kernel sizes to process these subspace features in parallel. We set an initial kernel size of 3 × 3 and gradually increase it by 2 in the subsequent subspaces. In this way, the feature maps within each subspace adaptively filter background information and focus on sensitive features of different granularities. This process can be expressed as follows:

$$MSMC(X) = Concat(DW_{k_1 \times k_1}(x_1), \ldots, (DW_{k_n \times k_n}(x_n)) \tag{4}$$

where $x = [x_1, x_2, \ldots, x_n]$ represents split up the input feature x into multiple subspaces in the channel dimension, and $k_i \in \{3, 5, \ldots, 2N - 1\}$ represents a monotonic increase of 2 in kernel size.

3.2.2 SAA

We have introduced a SAA module for information interaction between multiple subspace features. As shown in Fig. 2c, we select a channel from each subspace to construct a group and then use convolution operations to perform feature fusion within the group, thereby increasing the diversity of multi-scale spatial features. Furthermore, we use additional convolutions to perform cross group fusion. This process can be expressed as follows:

$$M = W_{inter}([G_1, G_2, \ldots, G_M]) \tag{5}$$

$$G_i = W_{intra}([H_1, H_2, \ldots, H_M]) \tag{6}$$

$$H_j^i = DWConv_{k_i \times k_j}(Hx_j^i) \in \mathbb{R}^{H \times W \times 1} \tag{7}$$

where W_{inter} and W_{intra} are the weight matrices of point-wise convolution. $j \in \{1, 2, \ldots, N\}$ and $i \in \{1, 2, \ldots, M\}$, where N and $M = \frac{C}{N}$ represent the number of subspaces and groups, respectively. Here, $H_j \in \mathbb{R}^{H \times W \times M}$ denotes the j−th subspace with depth-wise convolution, and H_j^i represents the i−th channel in the j−th subspace.

3.2.3 W-MSA and SW-MSA

The output feature maps Z obtained from the feature reconstruction of the first part of the basic block is input into the second part, which consists of W-MSA and SW-MSA modules connected head-to-tail. W-MSA and SW-MSA are the main functional modules of the Swin Transformer [34]. Unlike traditional Transformer that compute attention globally, resulting in high computational complexity, the Swin Transformer limits attention within each window through W-MSA and SW-MSA. It also introduces operations such as windows shift for inter-window information interaction, which can improve computational efficiency and feature extraction capability. Typically, W-MSA and SW-MSA appear in pairs. As shown in Fig. 3, the difference between them lies in the absence of windows shift and reverse operations in W-MSA. Fig. 4 illustrates the windows partition and shift processes, assuming that layer l is W-MSA (as shown in Fig. 4a), and layer $l+1$ is SW-MSA (as shown in Fig. 4b). By comparing the two images in (a) and (b), it can be observed that the divided windows first move O pixels to the left and up, and then the pixels within the windows move P pixels to the left and up. Performing windows shift operation followed by MAS operation enables information exchange between windows.

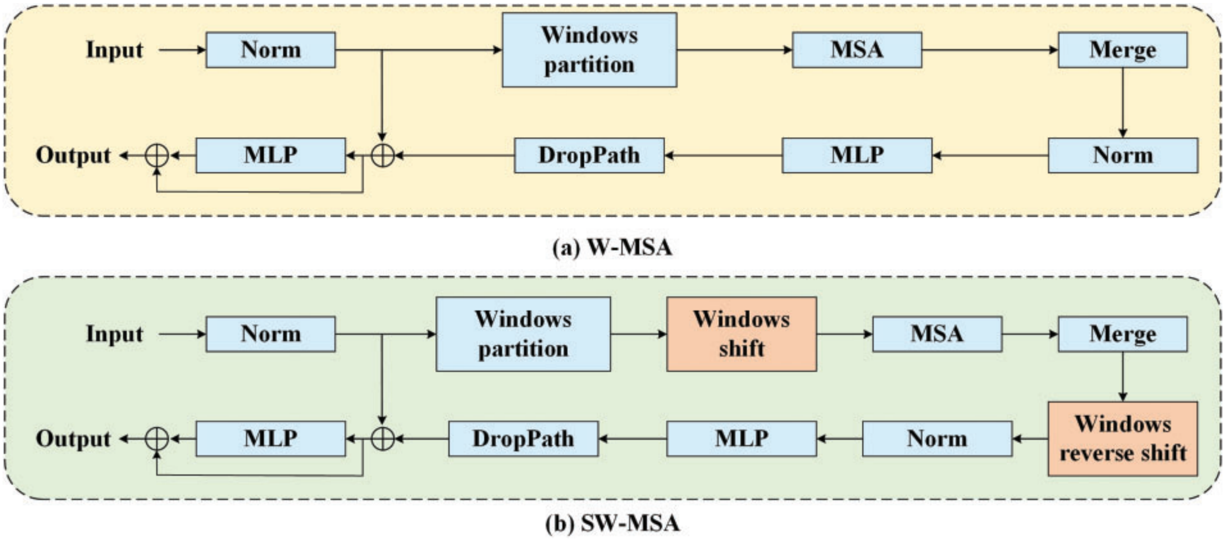


Figure 3: The flowcharts of W-MSA (Window Multi-Head Self Attention) and Shifted Window Multi-Head Self Attention (SW-MSA). The difference between them is that SW-MSA has windows shift and reverse operations, while W-MSA does not

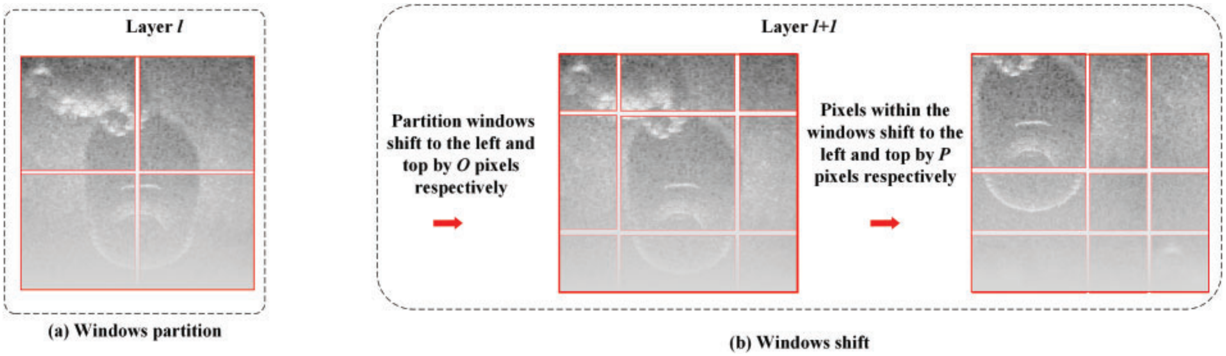

Layer l
(a) Windows partition
Layer l+1
Partition windows shift to the left and top by O pixels respectively
Pixels within the windows shift to the left and top by P pixels respectively
(b) Windows shift

Figure 4: Schematic diagram of SW-MSA windows partition and shift operations. The divided windows first move O pixels to the left and up, and then the pixels within the windows move P pixels to the left and up

Next, the Merge windows layer reduces the resolution while increasing the number of channels, so that the size of the feature maps in each stage gradually decreases and the number of channels increases. Reverse rotation is the inverse operation of performing shift, aiming to restore the original order of the feature maps; Norm is layer normalization; MLP is a multi-layer perceptron; DropPath is used to prevent overfitting. Finally, the final F_{W-MSA} and F_{SW-MSA} are obtained through residual calculation. The entire backbone network processing procedure is shown in Algorithm 1.

Algorithm 1: The processing procedure of backbone network

```
  **Input:** The layer depths of basic blocks $\{L_1, L_2, L_3, L_4\}$; Input image $X_0$; Stage_num=4;
  **Output:** The extracted image feature $Z_{\text{stage\_num}}$;
$Z_0 = X_0$;
for $s = 1$ to stage_num do
    for $l = 1$ to $L_{\text{stage\_num}}$ do
        if $l = 1$ then
            Input $Z_{s-1}$ and use Eqs. (1) to (7) to calculate $Z_{inter}$;
        else
            if $l$ is an even number then
                Perform W-MSA operation on $Z_{inter}$ to obtain a new $Z_{inter}$;
            else
                Perform SW-MSA operation on $Z_{inter}$ to obtain a new $Z_{inter}$;
            end
        end
    end
    $Z_s = Z_{inter}$;
end
Obtain features $Z_{\text{stage\_num}}$ extracted from the backbone network
```

4 Experiments

4.1 Datasets

We employed three different FLS image datasets, MDD [35], WHD [36], and UATD [37], to evaluate the target detection method proposed in this paper. Among them, MDD was obtained by the Ocean Systems Laboratory (Heriot-Watt University) using ARIS Explorer 3000 FLS at a frequency of 3.0 MHz. This dataset consists of 1868 images, including ten types of targets: tire, propeller, shampoo bottle, drink carton, bottle, can, chain, valve, standing bottle, and hook. WHD was captured by Tritech 1200ik FLS in Weihai, China. This dataset contains 4000 images, divided into eight categories: ball, cylinder, tire, cube, human body, circular cage, metal bucket, and square cage. UATD was collected by the Pengcheng Laboratory using Tritech Gemini 1200ik sonar from lakes and shallow water environments in Maoming and Dalian. It includes ten types of targets: metal bucket, cylinder, ROV, cube, tyre, circle cage, plane, human body, square cage, and ball. The training set of this dataset contains 7600 images, with two test sets, TEST-1 and TEST-2, containing 800 images each. As shown in Table 2, to ensure a fair comparison with the baseline algorithm, we attempted to set the dataset partitions the same as the comparison algorithm [27,30].

Table 2: Statistical data of three FLS datasets

Set	MDD		WHD		UATD (TEST-1)		UATD (TEST-2)	
	Img count	**Obj count**	**Img count**	**Obj count**	**Img count**	**Obj count**	**Img count**	**Obj count**
Training	1345	1345	2880	5068	6756	10,935	6756	10,935
Test	149	149	320	565	800	1172	800	1160

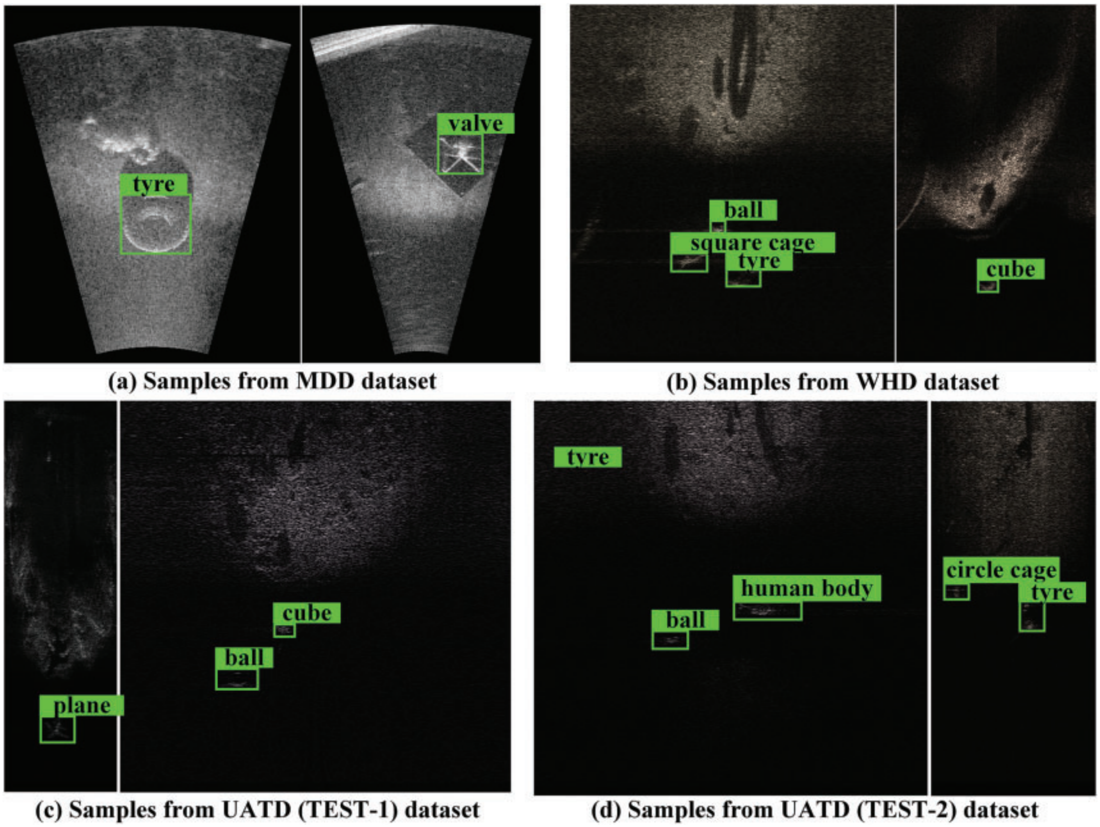
(Continued)

Table 2 (continued)

Set	MDD		WHD		UATD (TEST-1)		UATD (TEST-2)	
	Img count	Obj count	Img count	Obj count	Img count	Obj count	Img count	Obj count
Validation	374	374	800	1444	844	1372	844	1372
Total	1868	1860	4000	7077	8400	13,542	8400	13,530

Upon the release of the UATD dataset, the training set included 7600 sonar images, while the test set comprised 800 sonar images. Here, we divide the training set into a training and validation split of 8:1, while keeping the test set unchanged. Note that in the MDD dataset, the number of images is equal to the number of targets, with only one target per image, belonging to a single-target detection task. In contrast, in the WHD and UATD datasets, the number of images is less than the number of targets, with each image containing at least one target, belonging to a multi-target detection task.

The samples of the three datasets are shown in Fig. 5. These three datasets were obtained using different devices and in different scenarios. It can be seen that there is a significant difference in the imaging effect between the three. Compared with WHD and UATD samples, MDD samples have higher resolution, lower noise in the image, and clearer targets. Overall, the detection difficulty of the WHD and UATD datasets will be greater than that of MDD.



(a) Samples from MDD dataset

(b) Samples from WHD dataset

(c) Samples from UATD (TEST-1) dataset

(d) Samples from UATD (TEST-2) dataset

Figure 5: Samples from three experimental datasets, MDD, WHD, and UATD. The MDD dataset is a single-target detection task, while the WHD and UATD datasets are multi-target detection tasks

4.2 Experimental Setup

All experiments in this research were performed with the PyTorch framework and run on an NVIDIA GeForce RTX 3090. Except for the backbone network, the other structures of the network adopted the same architecture and hyperparameters as YOLOv10s. The pre-trained parameters of YOLOv10s [31] on the COCO [38] dataset were applied as the initial parameters for the detection head. Here, the values of the depths $\{L_1, L_2, L_3, L_4\}$ of the four stages' basic blocks are $\{2, 2, 6, 2\}$. To optimize the model, we used the SGD optimizer with a learning rate of 1e-2 and a momentum of 0.937. The model was trained through 500 epochs with a batch size of 32. The size of the input FLS image was adjusted to 640 × 640 × 3. The settings of other hyperparameters are shown in Table 3.

Table 3: Model parameters

Attribute	Value	Attribute	Value
Input size	640 × 640 × 3	Batch size	32
Layer depths {L1, L2, L3, L4}	{2, 2, 6, 2}	Weight decay	5e-4
Optimizer	SGD	Warmup momentum	0.8
Learning rate	1e-2	Warmup epochs	3
Momentum	0.937	Subspace N	4
Epoch	500	Group M	8

4.3 Evaluating Metrics

To assess and compare the performance and detection outcomes of our network, we utilize mean Average Precision (mAP) from COCO as the primary metric for accuracy evaluation. The method for calculating AP is defined as follows:

$$\mathrm{mAP} = \frac{1}{O} \sum_{i \in [0.50:0.05:0.95]} \mathrm{AP}_i \tag{8}$$

mAP represents the average of the AP across all categories at various IoU thresholds, where O denotes the total number of categories. The AP is determined by integrating to find the area under the precision-recall curve relative to the coordinate axes for all categories. The precision P and recall R metrics are calculated as follows:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{R} \times \mathrm{P}}{\mathrm{R} + \mathrm{P}} \tag{11}$$

where true positive (TP) and true negative (TN) refer to correct predictions, while false positive (FP) and false negative (FN) indicate incorrect outcomes. F1 takes into account both P and R, it is their harmonic mean.

4.4 Baseline System

We take recent target detection methods in the field as benchmark methods.

MBSNN: Wang et al. [39] improved the accuracy and speed of sonar target detection by introducing shortcut connections from the residual network and a multi-branch shuttle network based on the Yolov5s framework. At the same time, they changed the neck structure to a Bidirectional Feature Pyramid Network, further enhancing the connections between feature maps and improving detection performance.

UFIDNet: Long et al. [27] use supervised residual attention blocks to achieve better feature optimization for sonar image speckle reduction. In addition, the filtering strategy for scene prior detection reduces redundant anchor points, further improving detection speed.

FLSD-Net: Yang et al. [30] reduce information loss in sonar image downsampling through a lightweight initial downsampling (LID) module and a lightweight feature extraction (LFE) module. The LID module captures diverse features while minimizing information loss, and the LFE module efficiently extracts sparse target features from FLS images. By emphasizing shallow, high-resolution feature maps, FLSD-Net improves target identification.

Faster RCNN: Ren et al. [10] achieve end-to-end training and efficient candidate region generation by introducing a Region Proposal Network, significantly improving the speed and accuracy of object detection. Although not as real-time as YOLO, it still performs well in accuracy and is the preferred method for many high-precision object detection tasks.

YOLOv10s: YOLOv10 [31] introduces lightweight classification heads, spatial channel decoupled down-sampling, level guided blocks, large kernel convolution, and partial attention modules, achieving significant improvements in general object detection tasks. YOLOv10s is a compact version that balances speed and accuracy.

ER-DETR: Ref. [40] is the first real-time end-to-end object detector that surpasses existing YOLO detectors in terms of speed and accuracy, eliminating the negative impact of Non-Maximum Suppression (NMS) post-processing. By quantitatively analyzing the impact of NMS and establishing speed benchmarks, it achieves flexible speed adjustment and can adapt to object detection tasks in different scenarios.

4.5 Experimental Results

4.5.1 Experiment on MDD Dataset

The comparison results between ProNet and other methods are presented in Table 4, with the best performance highlighted in bold and the second-best in underlined. It can be seen that our ProNet outperforms existing sonar image detectors and general detectors. Specifically, ProNet performs better than the current state-of-the-art sonar image detectors RF-DETR, achieving a 2.7% higher accuracy on the overall metric $mAP_{50:95}$, as well as 0.1% and 0.3% higher accuracies on the metrics mAP_{50} and mAP_{75}, respectively. In addition, YOLOV10 is currently the most advanced method in the YOLO series, and its YOLOV10s version is our base model; The main difference between its structure and ProNet is the backbone network. It achieves the second-best performance across all previous metrics, but our ProNet surpasses it by 2.6% on the overall metric $mAP_{50:95}$ and by 0.4% and 1.8% on mAP_{50} and mAP_{75}, respectively. We attribute this improvement to the application of the backbone network proposed in this paper, which enhances the representation ability of features and focuses more on the sensitive areas of the image.

Table 4: Performance comparison on MDD

Method	**mAP$_{50}$(%)**	**mAP$_{75}$(%)**	**mAP$_{50:95}$(%)**	**P(%)**	**R(%)**	**F1(%)**
MBSNN	73.7	61.9	52.5	62.8	65.7	64.2
UFIDNet	88.0	83.4	70.5	76.9	87.7	81.9
FLSD-Net	88.2	84.1	72.7	81.4	80.9	81.1
Faster RCNN	87.6	72	60.2	77.8	79.5	78.6
YOLOv10s	90.4	85.1	73.5	**85.7**	82.6	84.1
RT-DETR	90.7	86.6	73.4	82.0	86.8	84.3
ProNet	**90.8**	**86.9**	**76.1**	80.4	**91.9**	**85.8**

Note: The underlined font in the table indicates suboptimal, while bold indicates optimal.

4.5.2 Experiment on WHD Dataset

The comparison results of the second experimental dataset, WHD, are presented in Table 5. WHD presents a more challenging scenario compared to MDD due to its higher level of ambiguity in targets and increased noise. In terms of performance metrics, ProNet, and the comparative methods exhibited lower performance on this dataset than on MDD. Despite this, ProNet emerged as the best performer. Specifically, ProNet achieved a mAP$_{50:95}$ that was 1.5% and 1.6% higher than RT-DETR and YOLOV10s, respectively. Moreover, it obtained a mAP$_{50}$ that was 2.6% and 0.4% higher than RT-DETR and YOLOV10s, respectively, and a mAP$_{75}$ that was 1.2% higher than YOLOV10s, respectively. ProNet showcased excellent adaptability across diverse scenes, demonstrating its robust performance in various real-world scenarios and strong generalization capabilities.

Table 5: Performance comparison on WHD

Method	**mAP$_{50}$(%)**	**mAP$_{75}$(%)**	**mAP$_{50:95}$(%)**	**P(%)**	**R(%)**	**F1(%)**
MBSNN	85.3	25.6	38.9	58.7	55.2	56.9
UFIDNet	90.9	47.3	36.4	85.7	83.7	84.7
FLSD-Net	92.7	46.1	53.8	85.1	80.1	82.5
Faster RCNN	82.3	24.0	37.8	79.2	76.9	78.0
YOLOv10s	94.7	47.1	54.2	**86.6**	80.0	83.2
RT-DETR	92.5	48.4	54.3	85.8	79.2	82.4
ProNet	**95.1**	**48.3**	**55.8**	86.1	**85.5**	**85.8**

Note: The underlined font in the table indicates suboptimal, while bold indicates optimal.

4.5.3 Experiment on UATD Dataset

As mentioned in the datasets section, the UATD dataset comprises two test subsets. In this section, experiments are conducted to train and optimize the model using the training and validation sets of UATD. The optimized models are subsequently utilized to evaluate performance on the two subsets, with the experimental results presented in Tables 6 and 7, respectively. Overall, the performance of our method is superior to that of six comparative algorithms. For instance, on the Test-1 dataset, our method achieves a mAP$_{50}$ of 96.5%, which exceeds the second-best algorithm, YOLOv10, by 1.5%. Other metrics, such as Precision and Recall, also outperform those of the comparative algorithms. The specialized sonar target detector, FLSD-Net, only attains a mAP$_{50}$ of 52.0%, whereas our method achieves 57.0% on the same metric, surpassing it by a significant margin of 5%. Furthermore, we can draw similar conclusions from the results on Test-2. Although our method does not always achieve the optimal performance in certain metrics—specifically, the mAP$_{50}$ is 0.4% lower than that of the best-performing FLSD-Net–our approach consistently demonstrates satisfactory detection performance overall. These results further validate the advancement of our method.

Table 6: Performance comparison on UATD (TEST-1)

Method	mAP$_{50}$(%)	mAP$_{75}$(%)	mAP$_{50:95}$(%)	P(%)	R(%)	F1(%)
MBSNN	72.1	35.9	29.8	54.1	58.8	56.4
UFIDNet	90.7	62.8	57.1	91.8	44.6	93.2
FLSD-Net	80.3	49.5	52.0	86.7	81.9	84.2
Faster RCNN	79.1	57.6	51.8	83.5	88.7	86.0
YOLOv10s	95.0	62.4	56.2	95.1	93.6	94.3
RT-DETR	94.9	63.2	55.8	94.8	**95.7**	**95.2**
ProNet	**96.5**	63.3	**57.0**	**95.6**	94.5	95.0

Note: The underlined font in the table indicates suboptimal, while bold indicates optimal.

Table 7: Performance comparison on UATD (TEST-2)

Method	mAP$_{50}$(%)	mAP$_{75}$(%)	mAP$_{50:95}$(%)	P(%)	R(%)	F1(%)
MBSNN	65.6	24.1	31.5	58.7	55.2	56.9
UFIDNet	80.3	28.8	36.4	85.7	83.7	84.7
FLSD-Net	**83.7**	30.1	37.8	86.1	84.1	85.1
Faster RCNN	80.5	27.8	35.6	79.2	76.9	78.0
YOLOv10s	80.4	30.3	36.3	**86.6**	80.0	83.2
RT-DETR	82.5	28.4	37.3	85.8	79.2	82.4
ProNet	83.3	**30.9**	**37.9**	86.1	**85.5**	**85.8**

Note: The underlined font in the table indicates suboptimal, while bold indicates optimal.

4.5.4 Experiment on Computational Complexity and Inference Speed

In this section, we embark on a comprehensive discussion regarding the computational complexity and inference speed of the target detection model. These factors play a pivotal role in determining the model's viability for practical applications, especially in real-time scenarios where timely responses are essential. By thoroughly examining these aspects, we aim to provide insights into how our model can be effectively utilized in practical settings while maintaining optimal performance levels. The experimental results are shown in Table 8. It is evident that our algorithm is not optimal in terms of computational complexity and model parameter count, but we can still achieve 24.1 FPS, which meets certain real-time detection requirements. In terms of parameter count, our model has 40.7 M parameters and can also be deployed on most existing mainstream edge computing devices, thus satisfying the requirements for lightweight deployment.

Table 8: The comparison of model computational overhead

Method	Input size	Params (M)	GFLOPs	Speed (FPS)
MBSNN	640 × 640	8.2	10.0	20.9
UFIDNet	600 × 600	37.2	75.9	3.9
FLSD-Net	640 × 640	**1.76**	**5.4**	**90.8**
Faster RCNN	600 × 600	28.3	92.5	9.9
YOLOv10s	640 × 640	8.1	24.8	56.7
RT-DETR	640 × 640	42.7	67.9	15.2
ProNet	640 × 640	40.7	87.0	24.1

Note: The underlined font in the table indicates suboptimal, while bold indicates optimal.

4.5.5 Ablation Study

In this part, we investigated the hyper-parameters selection of the extracted method ProNet and the effectiveness of the introduced feature reconstruction module. For the choice of N and M, we conducted a series of grid search experiments. Due to space limitations, we only present the two best sets of choices in Table 9: $N = 3$, $M = 4$ and $N = 6$, $M = 8$. It is evident that the experimental results for other parameter combinations are inferior to those chosen in this paper. Additionally, we replaced the introduced feature reconstruction module with the SW-MSA module to evaluate the contribution of the feature reconstruction module. As observed, the introduction of feature reconstruction led to varying degrees of improvement across all metrics. Specifically, mAP_{50}, mAP_{75} and $\text{mAP}_{50:95}$ improved by 2.9%, 1.7%, and 0.7%, respectively. Although there was a slight decrease in the p value, both the R value and F1 score achieved optimal results.

Table 9: Detection accuracy in different model variants

Model variants	mAP_{50} (%)	mAP_{75} (%)	$\text{mAP}_{50:95}$ (%)	p (%)	R (%)	F1 (%)
$N = 3$, $M = 4$, kernel sizes are 1, 3, 5	90.4	85.7	75.7	78.2	90.1	83.7
$N = 3$, $M = 4$, kernel sizes are 3, 5, 7	87.5	86.3	74.4	78.6	89.0	83.5
$N = 3$, $M = 4$, kernel sizes are 5, 7, 9	86.0	84.2	73.9	79.8	85.2	82.4
$N = 6$, $M = 8$, kernel sizes are 1, 3, 5, 7, 9, 11	90.4	86.6	75.5	80.2	90.5	85.0
$N = 6$, $M = 8$, kernel sizes are 3, 5, 7, 9, 11, 13	89.6	86.8	75.4	79.0	91.3	84.7
$N = 6$, $M = 8$, kernel sizes are 3, 5, 9, 11, 13, 15	88.2	84.7	74.5	77.8	90.2	83.5
Replace the feature reconstruction module with the SW-MSA module	87.9	85.2	75.4	**80.8**	89.3	84.8

(Continued)

Table 9 (continued)

Model variants	mAP$_{50}$ (%)	mAP$_{75}$ (%)	mAP$_{50:95}$ (%)	*p* (%)	R (%)	F1 (%)
N = 4, M = 8, kernel sizes are 1, 3, 5, 7	**90.8**	**86.9**	**76.1**	80.4	**91.9**	**85.8**

Note: The underlined font in the table indicates suboptimal, while bold indicates optimal.

5 Conclusion

In this paper, we present a novel basic block and construct a backbone network around it. By integrating this backbone network with the YOLOv10 framework, we introduce ProNet for FLS image object detection. Our evaluation of ProNet involved experiments conducted on two FLS datasets that pose distinct challenges. The results demonstrate that ProNet outperforms existing sonar image detectors as well as general detectors, highlighting its robust adaptability to diverse real-world scenarios and strong generalization capabilities.

We identify two possible directions for future research in sonar image target detection. First, there is an urgent need to enhance the robustness of detection algorithms against variations in underwater conditions, including differing water qualities, depths, and temperatures. Second, it is crucial to reduce both the model complexity and computational complexity of target detection algorithms while striving to maintain high detection accuracy. Such reductions in computational overhead will facilitate improved deployment on various edge embedded devices.

Acknowledgement: The authors express gratitude for the valuable work done by the editor and reviewers.

Funding Statement: This work was supported in part by Youth Innovation Promotion Association, Chinese Academy of Sciences under Grant 2022022, in part by South China Sea Nova project of Hainan Province under Grant NHXXRCXM202340, and in part by the Scientific Research Foundation Project of Hainan Acoustics Laboratory under grant ZKNZ2024001.

Author Contributions: The writing and research work of this paper were jointly completed by the following authors: Kaiqiao Wang was responsible for research design, experimental implementation, and paper writing, while Peng Liu and Chun Zhang were responsible for revising and polishing the paper. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials utilized in this research can be obtained by requesting them from the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Chen L, Liu Z, Tong L, Jiang Z, Wang S, Dong J, et al. Underwater object detection using invert multi-class adaboost with deep learning. In: 2020 International Joint Conference on Neural Networks (IJCNN); 2020; Glasgow, UK. p. 1–8.
2. Fan B, Chen W, Cong Y, Tian J. Dual refinement underwater object detection network. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer vision-ECCV 2020. Cham: Springer International Publishing; 2020. p. 275–91.

3. Jian Z, Liang Z, Zheng L, Nan L. The design and research of intelligent search and rescue device based on sonar detection and marine battery. In: 2017 International Conference on Computer Network, Electronic and Automation (ICCNEA); 2017; Shenzhen, China. p. 383–7.
4. Bell JM, Petillot YR, Lebart K, Reed S, Coiras E, Mignotte PY, et al. Target recognition in synthetic aperture and high resolution sidescan sonar. 2006 IET Seminar on High Resolution Imaging and Target Classification; 2006; London, UK. p. 99–106.
5. Fei T, Kraus D, Zoubir AM. Ontributions to automatic target recognition systems for underwater mine classification. IEEE Trans Geosci Remote Sens. 2015;53(1):505–18. doi:10.1109/TGRS.2014.2324971.
6. Reed S, Petilot Y, Bell J. A model based approach to mine detection and classification in sidescan sonar. In: Oceans 2003. Celebrating the Past... Teaming Toward the Future (IEEE Cat. No. 03CH37492); 2003; San Diego, CA, USA. p. 1402–7.
7. Williams DP. The mondrian detection algorithm for sonar imagery. IEEE Trans Geosci Remote Sens. 2018;56(2):1091–102. doi:10.1109/TGRS.2017.2758808.
8. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision-ECCV 2016. Cham: Springer International Publishing; 2016. p. 21–37.
9. Redmon J. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, NV, USA.
10. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2016;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
11. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017; Venice, Italy. p. 2961–9.
12. Kong W, Hong J, Jia M, Yao J, Cong W, Hu H, et al. YOLOv3-DPFIN: a dual-path feature fusion neural network for robust real-time sonar target detection. IEEE Sens J. 2019;20(7):3745–56. doi:10.1109/JSEN.2019.2960796.
13. Liu K, Peng L, Tang S. Underwater object detection using TC-YOLO with attention mechanisms. Sensors. 2023;23(5):2567. doi:10.3390/s23052567.
14. Neves G, Ruiz M, Fontinele J, Oliveira L. Rotated object detection with forward-looking sonar in underwater applications. Expert Syst Appl. 2020;140:112870. doi:10.1016/j.eswa.2019.112870.
15. Tang YJS, Bian G, Zhang Y. Hipwreck target recognition in side-scan sonar images by improved YOLOv3 model based on transfer learning. IEEE Access. 2020;8:173450–60. doi:10.1109/ACCESS.2020.3024813.
16. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, HI, USA. p. 7263–71.
17. Farhadi A, Redmon J. YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767. 2018.
18. Bochkovskiy A, Wang CY, Liao HYM, Liu YJ. YOLOv5: you only look once v5: more. arXiv preprint arXiv:2007.12092. 2020.
19. Myers V, Fawcett J. A template matching procedure for automatic target recognition in synthetic aperture sonar imagery. IEEE Signal Process Lett. 2010;17(7):683–6. doi:10.1109/LSP.2010.2051574.
20. Kalyan B, Balasuriya A. Sonar based automatic target detection scheme for underwater environments using CFAR techniques: a comparative study. In: Proceedings of the 2004 International Symposium on Underwater Technology (IEEE Cat. No. 04EX869). 2004; Tokyo, Japan: IEEE. p. 33–7.
21. Dobeck GJ, Hyland JC. Automated detection and classification of sea mines in sonar imagery. In: Detection and remediation technologies for mines and minelike targets II. Arlington, Virginia, USA: SPIE; 1997. Vol. 3079, p. 90–110.
22. Williams DP. Fast target detection in synthetic aperture sonar imagery: a new algorithm and large-scale performance analysis. IEEE J Oceanic Eng. 2014;40(1):71–92. doi:10.1109/JOE.2013.2294532.
23. Hurtós N, Palomeras N, Carrera A, Marc C. Autonomous detection, following and mapping of an underwater chain using sonar. Ocean Eng. 2017;130:336–50. doi:10.1016/j.oceaneng.2016.11.072.
24. Jiang L, Cai T, Ma Q, Xu F, Wang S. Active object detection in sonar images. IEEE Access. 2020;8:102540–53. doi:10.1109/ACCESS.2020.2999341.

25. Zhang P, Tang J, Zhong H, Ning M, Liu D, Wu K. Self-trained target detection of radar and sonar images using automatic deep learning. IEEE Trans Geosci Remote Sens. 2021;60:1–14. doi:10.1109/TGRS.2021.3096011.
26. Fan Z, Xia W, Liu X, Li H. Detection and segmentation of underwater objects from forward-looking sonar based on a modified mask RCNN. Signal, Image Video Process. 2021;15(6):1135–43. doi:10.1007/s11760-020-01841-x.
27. Long H, Shen L, Wang Z, Chen J. Underwater forward-looking sonar images target detection via speckle reduction and scene prior. IEEE Trans Geosci Remote Sens. 2023;61:1–13. doi:10.1109/TGRS.2023.3248605.
28. Gao X, Zhang L, Chen X, Lin C, Hao R, Zheng J. GCT-YOLOv5: a lightweight and efficient object detection model of real-time side-scan sonar image. Signal, Image Video Process. 2024;18:1–10. doi:10.1007/s11760-024-03174-5.
29. Qin KS, Liu D, Wang F, Zhou J, Yang J, Zhang W. Improved YOLOv7 model for underwater sonar image object detection. J Vis Commun Image Rep. 2024;100:104124. doi:10.1016/j.jvcir.2024.104124.
30. Yang H, Zhou T, Jiang H, Yu X, Xu S. A lightweight underwater target detection network for forward-looking sonar images. IEEE Trans Instrum Meas. 2024;73:1–13. doi:10.1109/TIM.2024.3425490.
31. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. 2024. doi:10.48550/arXiv.2405.14458.
32. Guo M, Lu C, Liu Z, Cheng M, Hu S. Visual attention network. Comput Vis Med. 2023;9(4):733–52. doi:10.1007/s41095-023-0364-2.
33. Dosovitskiy A. An image is worth 16 × 16 words: transformers for image recognition at scale. 2020. doi:10.48550/arXiv.2010.11929.
34. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021; Montreal, QC, Canada. p. 10012–22.
35. Singh D, Matias VT. The marine debris dataset for forward-looking sonar semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021; Montreal, QC, Canada. p. 3741–9.
36. Wang Z, Zhang Z, Huang W, Guo J, Zeng L. Sonar image target detection based on adaptive global feature enhancement network. IEEE Sens J. 2021;22(2):1509–30. doi:10.1109/JSEN.2021.3131645.
37. Xie K, Yang J, Qiu K. A dataset with multibeam forward-looking sonar for underwater object detection. Sci Data. 2022;9(1):739. doi:10.1038/s41597-022-01854-w.
38. Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference; 2014 Sep 6–12; Zurich, Switzerland: Springer.
39. Wang J, Feng C, Wang L, Li G, He B. Detection of weak and small targets in forward-looking sonar image using multi-branch shuttle neural network. IEEE Sens J. 2022;22(7):6772–83. doi:10.1109/JSEN.2022.3147234.
40. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. Detrs beat yolos on real-time object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; Seattle, WA, USA. p. 16965–74.