frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 15 April 2025
DOI 10.3389/fmars.2025.1542832

OPEN ACCESS

EDITED BY
Muhammad Yasir,
China University of Petroleum (East China), China

REVIEWED BY
Nuno Pessanha Santos,
Portuguese Military Academy, Portugal
Martin Aubard,
University of Porto, Portugal

*CORRESPONDENCE
Jing Han
hanj@nwpu.edu.cn

RECEIVED 10 December 2024
ACCEPTED 11 March 2025
PUBLISHED 15 April 2025

CITATION
Cui X, Zhang J, Zhang L, Zhang Q and Han J (2025) Small object detection in side-scan sonar images based on SOCA-YOLO and image restoration. *Front. Mar. Sci.* 12:1542832. doi: 10.3389/fmars.2025.1542832

Small object detection in side-scan sonar images based on SOCA-YOLO and image restoration

Xiaodong Cui, Jiale Zhang, Lingling Zhang, Qunfei Zhang and Jing Han*

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

Although side-scan sonar can provide wide and high-resolution views of submarine terrain and objects, it suffers from severe interference due to complex environmental noise, variations in sonar configuration (such as frequency, beam pattern, etc.), and the small scale of targets, leading to a high misdetection rate. These challenges highlight the need for advanced detection models that can effectively address these limitations. Here, this paper introduces an enhanced YOLOv9(You Only Look Once v9) model named SOCA-YOLO, which integrates a Small Object focused Convolution module and an Attention mechanism to improve detection performance to tackle the challenges. The SOCA-YOLO framework first constructs a high-resolution SSS (sidescan sonar image) enhancement pipeline through image restoration techniques to extract fine-grained features of micro-scale targets. Subsequently, the SPDConv (Space-to-Depth Convolution) module is incorporated to optimize the feature extraction network, effectively preserving discriminative characteristics of small targets. Furthermore, the model integrates the standardized CBAM (Convolutional Block Attention Module) attention mechanism, enabling adaptive focus on salient regions of small targets in sonar images, thereby significantly improving detection robustness in complex underwater environments. Finally, the model is verified on a public side-scan sonar image dataset Cylinder2. Experiment results indicate that SOCA-YOLO achieves Precision and Recall at 71.8% and 72.7%, with an mAP50 of 74.3%. It outperforms the current state-of-the-art object detection method, YOLO11, as well as the original YOLOv9. Specifically, our model surpasses YOLO11 and YOLOv9 by 2.3% and 6.5% in terms of mAP50, respectively. Therefore, the SOCA-YOLO model provides a new and effective approach for small underwater object detection in side-scan sonar images.

KEYWORDS

side-scan sonar, image restoration, YOLOv9, attention mechanism, Space-to-Depth Convolution

1 Introduction

Side-scan sonar (何勇光, 2020) is an extensively utilized underwater sensing technology, mainly applied in underwater terrain mapping, object detection, and exploration tasks. In contrast to conventional downward-looking sonar, side-scan sonar transmits acoustic waves at horizontal or inclined angles, thereby covering a larger area of seabed features and improving detection performance. As a result, side-scan sonar is widely utilized in areas such as maritime archaeology, submerged pipeline monitoring, and wreck exploration (Gomes et al., 2020; Tian et al., 2007; Fengchun et al., 2002; Sun et al., 2021; Jinhua et al., 2016). Nevertheless, the intricate underwater environment often introduces multiple sources of noise and blurring in side-scan sonar images, including scattering noise, multipath artifacts, noise streaks, and acoustic shadow distortions. Furthermore, instrumental noise arises from the sensor's inherent electronic noise and the transducer's non-ideal properties, potentially leading to image signal degradation. The interaction of these noise factors results in considerable difficulties in processing side-scan sonar images for real-world applications.

The unique properties of side-scan sonar images introduce significant difficulties in target detection. Firstly, sonar imagery often exhibits considerable background noise and spurious objects, including natural seabed formations and acoustic backscatter from sediment particles, which frequently resemble real targets and result in an elevated false alarm rate in detection models. Secondly, targets in sonar images generally manifest as small, diffuse high-intensity reflections with vague edges and uneven signals, making them indistinguishable from surrounding textures and increasing the difficulty of segmentation from the background. Furthermore, side-scan sonar image data exhibit substantial distribution discrepancies across different scenarios. Given the high cost and inefficiency of underwater data acquisition, labeled datasets are often scarce. This non-uniformity and data insufficiency severely hinder the generalization capability of algorithms, posing a formidable challenge for achieving accurate target detection in complex underwater settings. The rapid progress in artificial intelligence and machine learning has facilitated the fusion of advanced image processing techniques with target detection models, substantially enhancing side-scan sonar image quality and improving the precision of seabed target detection (Yasir et al., 2024; Cheng et al., 2023; Wen et al., 2024; Fan et al., 2022; Yu et al., 2021; Fayaz et al., 2022).

Among existing underwater target detection methods for side-scan sonar images, some object detection models have become relatively outdated and struggle to meet the current diversified underwater application requirements. Although some studies have improved traditional deep learning models, these enhancements often fail to adequately consider the inherent structural characteristics of side-scan sonar images. This neglect of sonar image characteristics makes targeted model optimization challenging, resulting in subpar detection performance in practical applications. Furthermore, while some modified models have enhanced detection capabilities to some extent, their parameter counts have also increased substantially, leading to higher computational costs. Therefore, developing an effective underwater target detection method tailored to the specific requirements of side-scan sonar images is particularly crucial. As shown in Figure 1, by improving existing object detection models with greater emphasis on the structure and characteristics of side-scan sonar images, we can significantly enhance detection performance while effectively controlling model parameters and computational complexity, thereby providing more reliable metrics for underwater detection tasks.

This paper is structured into four main sections: The first section provides a literature review, systematically summarizing the current research status in underwater side-scan sonar image target detection. The second section focuses on methodology, providing a detailed explanation of the proposed detection model and its theoretical framework. The third section presents experimental validation, where multiple comparative and ablation experiments empirically analyze the performance advantages of the proposed model. The fourth section provides conclusions and future perspectives, discussing in depth the future research directions and trends in underwater side-scan sonar image processing based on an evaluation of the model's practical performance.

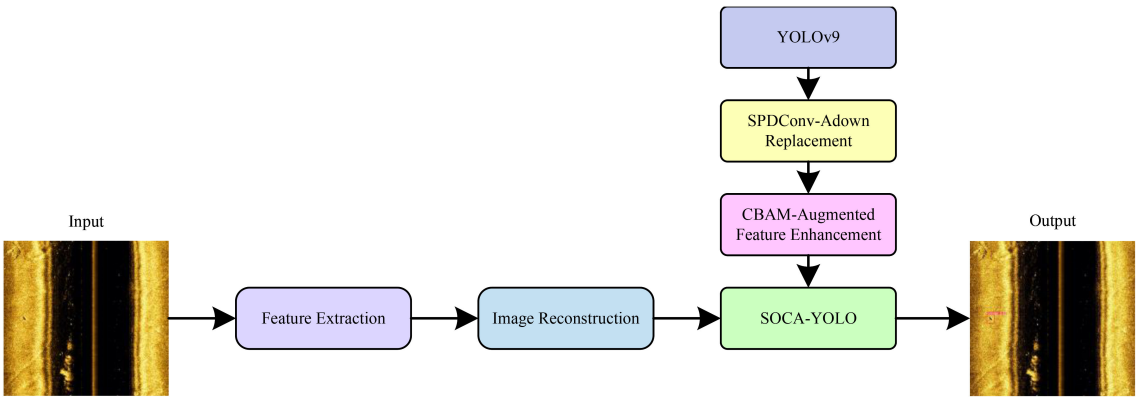

FIGURE 1
SSS object detection architecture.

The main contributions of this paper are as follows:

1. SwinIR-based sonar image enhancement method: To address the issues of low quality and high noise interference in traditional sonar images, the SwinIR super-resolution reconstruction network is introduced into the field of sonar image preprocessing. This method can more effectively enhance image quality, providing clearer input features for subsequent target detection.
2. Optimal model selection for small target detection in images: In the task of small target detection in side-scan sonar images, a comparison of existing object detection network models reveals that YOLOv9, through its auxiliary reversible branch, retains critical feature information, significantly enhancing the model's ability to detect small targets, particularly improving target recognition accuracy in complex backgrounds.
3. CBAM-enhanced detection model: Building upon the standard YOLOv9 network, the convolutional block attention module (CBAM) is innovatively incorporated. Unlike the original model's reliance solely on convolutional feature extraction, this method adaptively focuses on key target features, significantly improving target detection accuracy in complex underwater environments.
4. SPDConv replacement for ADown downsampling scheme: To address the challenges of small target detection in sonar images, the original ADown module in YOLOv9 is replaced with the SPDConv (Space-to-Depth Convolution) module. Compared to traditional downsampling methods, this improvement effectively mitigates the issue of small target feature loss.
5. Sonar image dataset reconstruction and evaluation: Existing public sonar datasets are systematically restructured, and a data partitioning standard more aligned with practical application scenarios is proposed. Experimental results demonstrate that the proposed improvements outperform traditional methods across all metrics.

2 Related work

As deep learning technology advances, various effective approaches have been introduced in image enhancement. The goal of image enhancement is to enhance image visual quality and interpretability using different algorithms, spanning from basic filtering to sophisticated color adjustment and detail refinement. Methods based on Convolutional Neural Networks (CNNs), such as Du (Du et al., 2023), employ four conventional CNN models for training and predicting on the same submarine SSS dataset. A comparative analysis was conducted on the predictive accuracy and computational efficiency of the four CNN models. Generative Adversarial Networks (GANs) employ adversarial learning between a generator and a discriminator to produce highly detailed images. Jiang (Jiang et al., 2020), for example, introduced a GAN-based semantic image synthesis model that can efficiently generate high-quality SSS images with reduced computational cost and time. Swin Transformer (Liu et al., 2021) serves as a versatile vision model designed mainly for image classification, object detection, and semantic segmentation (Lin et al., 2022; Gao et al., 2022; He et al., 2022; Jannat and Willis, 2022), with potential applications in image enhancement and video processing. It is specifically designed for efficient high-resolution image processing and has demonstrated superior performance in multiple visual tasks. SwinIR (Liang et al., 2021), built upon Swin Transformer, is a deep learning framework tailored for image restoration, encompassing super-resolution, noise reduction, and deblurring, among other tasks. Retaining the strengths of Swin Transformer, it integrates task-specific optimizations for image restoration, leading to improved processing efficiency and output quality. SwinIR has demonstrated significant performance improvements across various fields. For instance, in medical imaging, its application in low-dose PET/MRI restoration achieves a mean SSIM of 0.91 at a 6.25% dose level, substantially enhancing image quality (Wang et al., 2023b). In the domain of remote sensing, experiments on benchmark datasets show that SwinIR can enhance the resolution of satellite and aerial images—at a 2× scaling factor, its PSNR reaches 35.367dB and its SSIM increases to 0.9449, thereby facilitating more accurate topographic monitoring and mapping (Ali et al., 2023). Moreover, in video enhancement and facial recognition (Zheng et al., 2022; Lin, 2023), SwinIR's robust feature extraction and reconstruction capabilities significantly improve detail recovery and overall performance, as evidenced by its competitive results in multiple top-tier challenges. These advancements in deep learning have propelled significant innovations in image enhancement techniques.

In the field of computer vision, object detection and image enhancement are two complementary and important research directions. Image enhancement techniques aim to improve image quality, providing more accurate inputs for object detection, while object detection techniques focus on identifying and localizing objects of interest within images. Deep learning-based object detection methods are primarily divided into two categories: one-stage methods and two-stage methods. One-stage detection models directly predict target locations and categories through a single network forward pass, offering faster speed but potentially slightly lower accuracy. Representative works include the SSD (Single Shot Detector) series (Liu et al., 2016) and the YOLO (You Only Look Once) family (Redmon, 2016; Redmon and Farhadi, 2017; Redmon, 2018; Bochkovskiy et al., 2020; Li et al., 2022; Wang et al., 2023a, 2025, 2024). Two-stage detection models first generate candidate regions and then classify and regress these regions, achieving higher accuracy but at a relatively slower speed. Representative works include the R-CNN family (Girshick et al., 2014; Ren et al., 2016; He et al., 2017). Currently, these methods have been widely applied in underwater object detection tasks using sonar images and have achieved significant results (Heng et al., 2024; Yang et al., 2025; Ma et al., 2024; Yulin et al., 2020; Polap et al., 2022).

Deep learning-based side-scan sonar image enhancement and object detection technologies have achieved significant progress in both theoretical research and practical applications. Burguera et al

(Burguera and Oliver, 2016) employed a probability model-based high-resolution seabed mapping method, correcting sonar data using physical models to generate high-precision images surpassing the device's resolution, laying the foundation for scientific applications. Tang et al. (2023) proposed a deep learning-based real-time object detection method, incorporating lightweight network design to address the challenges of detection efficiency and accuracy in complex underwater terrains. Li et al. (2024) designed an image generation algorithm for zero-shot and few-shot scenarios by combining UA-CycleGAN and StyleGAN3, significantly enhancing the generalization performance of deep learning-based object detection models. Yang et al. (2023) employed diffusion models to generate high-fidelity sonar images and validated the effectiveness of these enhanced data in practical object detection tasks. Zhu et al. (2024) significantly improved the stability and global information extraction capabilities of generative models by introducing CC-WGAN and CBAM modules, while also enhancing the accuracy of object detection. Yang et al. (2024) generated full-category sonar image samples using diffusion models combined with transfer learning, and trained object detection and semantic segmentation models with these samples, significantly improving model performance and data diversity. Aubard et al. (2024) proposed the YOLOX-ViT model, effectively compressing the model size using knowledge distillation while maintaining high detection performance, particularly reducing false alarm rates in underwater environments. Peng et al. (2024) designed a single-image enhancement method based on the CBL-sinGAN network, incorporating CBAM modules and L1 loss functions to enhance the construction capability of small-sample object detection models while preserving sonar image style.

3 Method

This section introduces the proposed SOCA-YOLO model, which integrates the image restoration model SwinIR, the CBAM (Woo et al., 2018) attention mechanism, the SPDConv (Sunkara and Luo, 2022) convolution module, and the YOLOv9 object detection model.

3.1 SwinIR

Image restoration is the process of transforming low-quality images into high-quality versions. SwinIR, a model based on the Swin Transformer, is primarily used for image super-resolution, denoising, and JPEG compression artifact reduction.

SwinIR combines the strengths of both Transformers and CNNs, outperforming traditional CNNs in handling large images due to its local attention mechanism. SwinIR employs a sliding window approach, dividing the input image into multiple small windows and processing each window separately, while retaining the Transformer's ability to manage long-range pixel relationships within the image. As illustrated in Figure 2, SwinIR is designed based on the Swin Transformer and comprises three modules: Shallow Feature Extraction, Deep Feature Extraction, and High-Quality Image Reconstruction.

The Shallow Feature Extraction module extracts initial features through convolutional layers, preserving lowfrequency information and passing it to the reconstruction module. The Deep Feature Extraction module incorporates Residual Swin Transformer Blocks (RSTB), which achieve local attention and cross-window interactions through multiple Swin Transformer layers. Residual connections provide a shortcut for feature aggregation, and convolutional layers further enhance the features. Finally, the High-Quality Image Reconstruction module combines shallow and deep features to produce high-quality images. Each module is detailed below.

Shallow Feature Extraction Module: This module uses a 3×3 convolution to extract shallow features. The main purpose of this process is to retain low-frequency information, leading to better and more stable results. A low-quality imageis $I_L$input at the input stage, and after passing through the shallow feature extraction module H_S, the shallow feature F_0 is obtained as shown in Equation 1:

$$F_0 = H_S(I_L) \tag{1}$$

Deep Feature Extraction Module: This module consists of several RSTBs (Residual Swin Transformer Blocks) and a 3×3 convolution. Each RST is composed of an even number of Swin Transformer Layers (STL) and a convolution layer. This module further processes the shallow features, resulting in its deep feature F_D, as shown in Equation 2.

$$F_D = H_D(F_0) \tag{2}$$

Here, H_D represents the deep feature extraction module.

High-Quality Image Reconstruction: The shallow and deep features are aggregated, transferring both the lowfrequency and high-frequency information of the image to the reconstruction layer. The high-quality image reconstruction module uses a sub-pixel convolution layer to upsample the feature map, resulting in the reconstructed high-quality image I_H, as shown in Equation 3:

$$I_H = H_{RE}(F_0 + F_D) \tag{3}$$

Here, H_{RE} represents the high-quality image reconstruction module.

3.2 CBAM

The Convolutional Block Attention Module (CBAM) is an efficient attention module for feedforward convolutional neural networks, proposed by Sanghyun Woo et al, as illustrated in Figure 3a. CBAM enhances the model's perceptive capability by incorporating a Channel Attention Module (CAM) (Figure 3b) and a Spatial Attention Module (SAM) (Figure 3c) into CNNs, thereby improving performance without adding significant network complexity. As a lightweight and versatile module, CBAM can be seamlessly integrated into any CNN architecture, adding minimal parameters and enabling end-to-end training with YOLOv9 models.

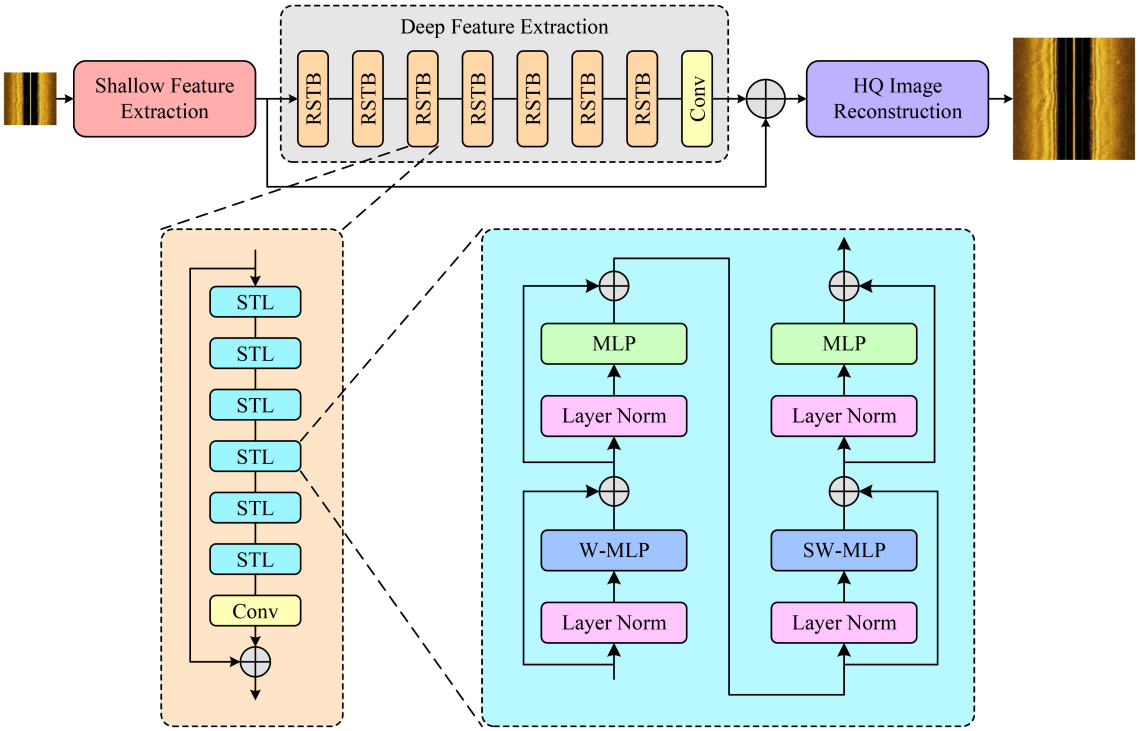

FIGURE 2
SwinIR transformer architecture.

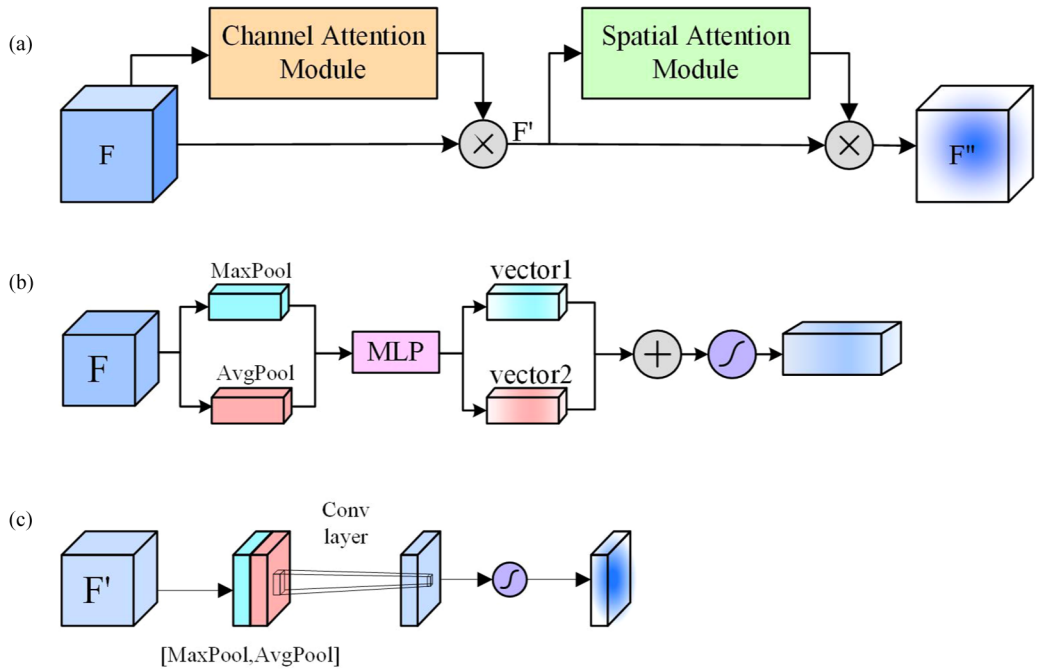

FIGURE 3
(a) Convolutional Block Attention Module (CBAM) architecture, **(b)** Channel Attention Module (CAM) architecture, **(c)** Spatial Attention Module (SAM) architecture.

The input feature map F first passes through the CAM, where the channel weights are multiplied with the input feature map to produce F'. Then, F' is fed into the SAM, where the normalized spatial weights are multiplied with the input feature map of the spatial attention mechanism, resulting in the final weighted feature map F''.

3.3 Space-to-Depth Convolution

The fundamental principle of SPDConv (Space-to-Depth Convolution) is to enhance the performance of convolutional neural networks (CNNs) when processing low-resolution images and small objects, as illustrated in Figure 4. This improvement is achieved through the following key steps:

1. Replacing Strided Convolutions and Pooling Layers: SPDConv is designed to replace traditional strided convolution and pooling layers, which often cause the loss of fine-grained information when dealing with low-resolution images or small objects.
2. Space-to-Depth (SPD) Layer: This transformation layer converts the spatial dimensions of the input image into the depth dimension, increasing the feature map depth without information loss. The SPD layer is critical for retaining spatial information, especially when processing low-resolution images and small objects. By converting spatial information into the depth dimension, the SPD layer mitigates the information loss typically associated with traditional strided convolutions and pooling operations.
3. Non-strided Convolution Layer: A convolutional layer with a stride of 1, applied after the SPD transformation, preserves fine-grained information by avoiding size reduction of the feature map. This non-strided convolution enables feature extraction while maintaining the full resolution of the feature map, which is essential for enhancing recognition performance on low-resolution images and small objects.

SPDConv effectively processes low-resolution images and small objects by combining space-to-depth transformations with non-strided convolutions. This method addresses the fine-grained information loss commonly caused by traditional strided convolutions and pooling layers during downsampling. By preserving spatial information through the SPD layer and converting it into depth features, combined with non-strided convolutions to capture finer details, SPDConv excels in small object detection tasks. It significantly enhances detection accuracy and adaptability to low-resolution images, offering a novel solution for small object detection and related tasks.

3.4 YOLOv9

Proposed in 2024, YOLOv9 is an object detection network that excels in both detection accuracy and processing speed. The model

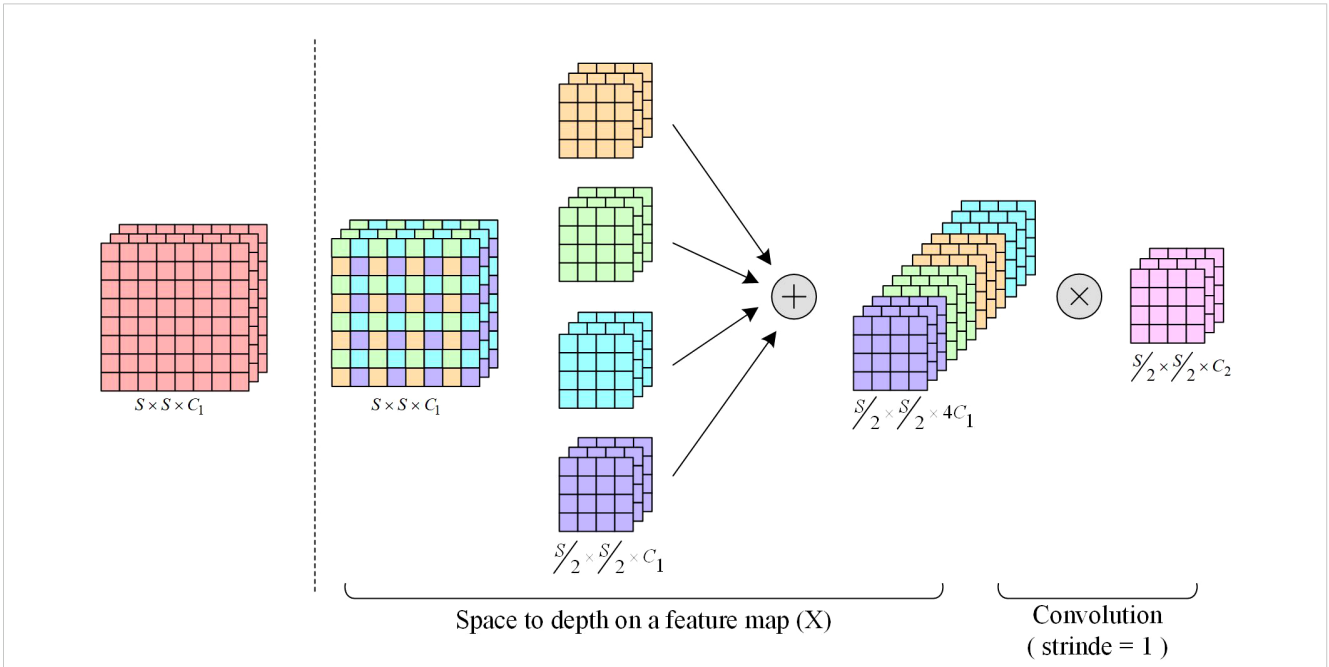

FIGURE 4
SPDConv architecture.

introduces Programmable Gradient Information (PGI), as illustrated in Figure 5. Through auxiliary reversible branches, PGI allows deep features to retain essential object characteristics, enabling the network to preserve crucial visual features of the target without sacrificing important information. This approach enhances YOLOv9's ability to maintain high performance even in complex detection scenarios.

PGI consists of three components: the main branch, multi-level auxiliary information, and the auxiliary reversible branch. Each component is detailed below:

Main Branch: The main branch includes the backbone network, neck network, and head network, which are common components in the YOLO series. The backbone network primarily uses Conv and RepNCSPELAN4 layers for feature extraction. The neck network comprises Upsample, Conv, and RepNCSPELAN4 layers, utilizing an FPN+PAN structure for multi-scale target detection. The head network processes features from the neck network to predict and classify large, medium, and small objects.

Auxiliary Reversible Branch: This branch addresses information loss that occurs as network depth increases, leading to information bottlenecks that hinder reliable gradient generation from the loss function. It introduces an additional network between the feature pyramid layers and the main branch to integrate gradient information from multiple prediction heads.

Multi-level Auxiliary Information: Multi-level auxiliary information involves inserting an integrated network between the feature pyramid's sub-layers and the main branch under auxiliary supervision. This network aggregates gradient information from various prediction heads and passes it to the main branch for parameter updates. Consequently, the feature pyramid in the main branch is not dominated by specific objects, enabling the main branch to retain comprehensive information necessary for learning target features.

3.5 SOCA-YOLO

In this study, we have improved upon the YOLOv9 object detection framework to address challenges such as noise interference, small target size, and edge blurring in side-scan sonar images. Due to the unique imaging mechanism of side-scan sonar, the images often exhibit high noise and low contrast, which can hinder traditional detection models from effectively extracting fine-grained features. Although YOLOv9 demonstrates notable advantages in real-time performance and multi-scale feature fusion, its standard convolutional layers and global feature extraction strategies still exhibit certain limitations when handling such specialized scenarios. Therefore, we propose two main improvements: the introduction of the CBAM attention mechanism into the model and the replacement of some standard convolutional layers with SPDConv modules, thereby achieving more precise feature extraction and fusion for small targets. The modified network model is illustrated in Figure 6.

In our improved model, the overall architecture still adheres to the core design principles of YOLOv9, divided into three components: Backbone, Neck, and Head. However, novel

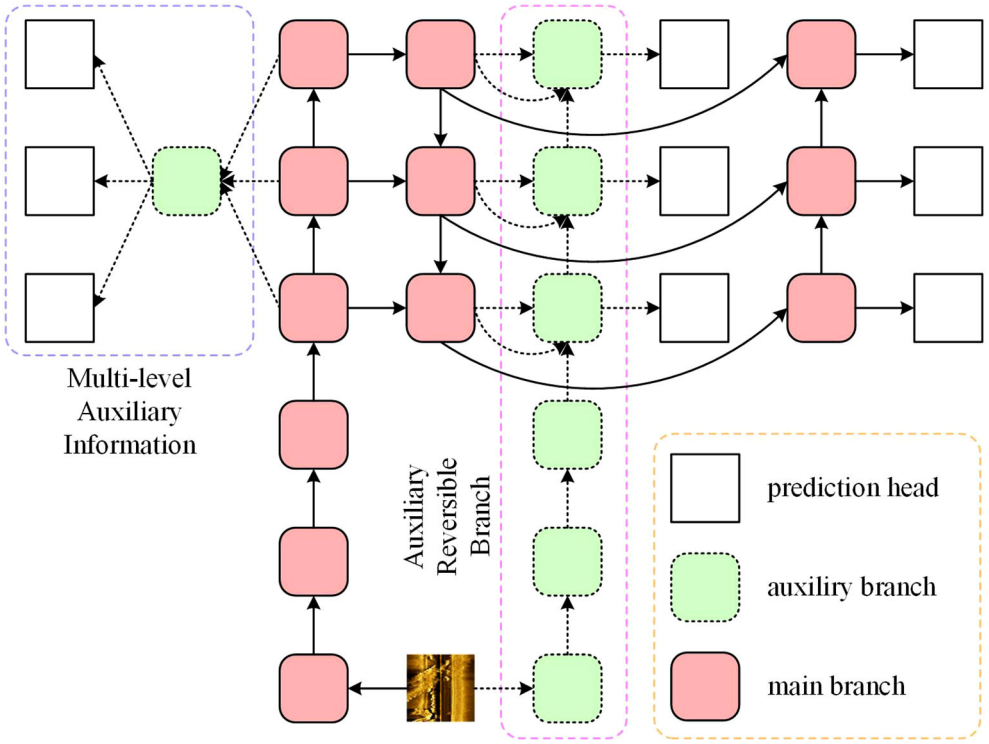

FIGURE 5
YOLOv9 architecture.

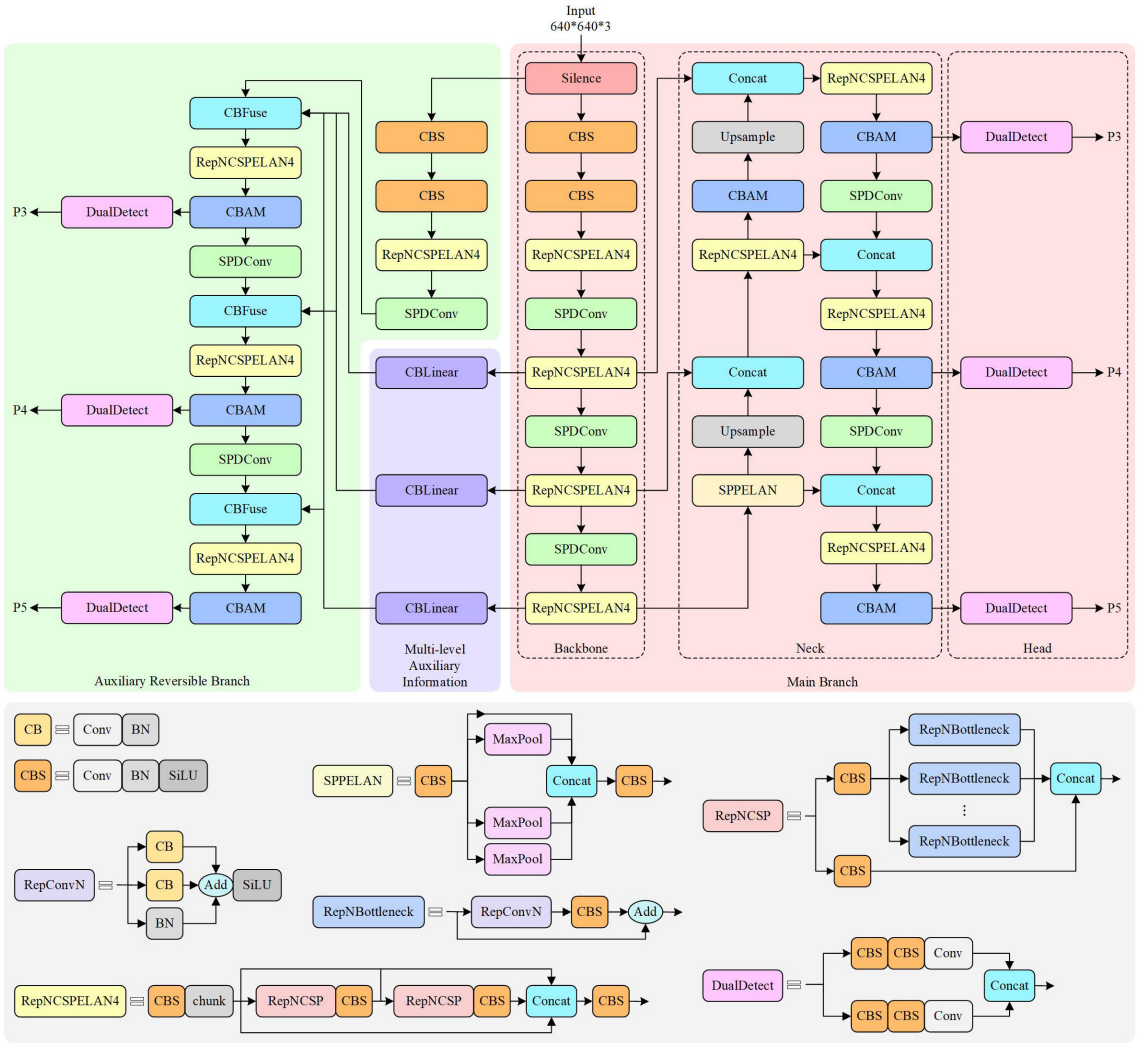

FIGURE 6
SOCA-YOLO architecture.

modules have been strategically incorporated at each stage to adapt to the characteristics of side-scan sonar images. First, the Backbone section integrates SPDConv modules alongside traditional convolutional layers to enhance multi-scale representation capabilities in feature extraction. Specifically, the SPDConv module performs spatial reorganization of input feature maps. This operation can be formally described as follows: let the input feature map be defined in Equation 4.

$$x \in \mathbb{R}^{C\times H\times W} \tag{4}$$

Initially, SPDConv samples x to derive four sub-regions, as shown in Equation Equation 5.

$$\begin{aligned} x_1 &= x[\ldots, :2, :2], \quad x_2 = x[\ldots, 1:2, :2], \\ x_3 &= x[\ldots, :2, 1:2], \quad x_4 = x[\ldots, 1:2, 1:2] \end{aligned} \tag{5}$$

The four sub-features are concatenated in the channel dimension, resulting in a new feature map, as shown in Equation 6.

$$x_{\text{SPD}} = \text{Concat}\{x_1, x_2, x_3, x_4\} \in \mathbb{R}^{4C\times\frac{H}{2}\times\frac{W}{2}}, \tag{6}$$

Subsequently, a 3×3 convolutional layer (denoted as $\text{Conv}_{3\times3}$) is employed for fusion, producing the output features, as shown in Equation 7:

$$y = \text{Conv}_{3\times3}(x_{\text{SPD}})\,. \tag{7}$$

This spatial reorganization and downsampling strategy not only reduces the size of the feature maps and computational load but also effectively captures fine-grained information through channel expansion, offering significant advantages for detecting small, blurry targets in side-scan sonar images.

In the Backbone and some Head modules, we also embed the CBAM to apply dual attention weighting to the features. Specifically, let the input feature be $F \in \mathbb{R}^{C\times H\times W}$, and first, channel statistics are computed through global average pooling and max pooling along the channel dimension, as shown in Equation 8:

$$f_{\text{avg}}(c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F(c,i,j), \quad f_{\max}(c) = \max_{i,j} F(c,i,j). \tag{8}$$

These two sets of statistics are processed through a shared multi-layer perceptron (MLP) and a Sigmoid activation to generate the channel attention vector $M_c \in \mathbb{R}^C$, which is then multiplied with the original features on a per-channel basis to obtain the intermediate feature $F' = M_c \otimes F$. Next, average and max pooling are applied along the channel dimension of F', followed by concatenation, a 7×7 convolution, and Sigmoid activation to generate the spatial attention map $M_s \in \mathbb{R}^{H \times W}$, which is then used to output the spatially weighted feature, as shown in Euqation 9:

$$F_{\text{att}} = M_s \otimes F'. \tag{9}$$

This process allows the network to automatically focus on the target regions, effectively suppress background noise, and further enhance the discriminative ability for small target features.

In the overall architecture, the multi-scale features extracted by the Backbone are strengthened by the SPDConv and CBAM modules and then passed to the Neck section. The Neck employs an FPN and PAN-style multi-scale feature fusion strategy, merging features from different levels in an abstract formulation, as shown in Euqation 10:

$$F_{\text{neck}} = \sum_{i=1}^{N} w_i \cdot f_i(F_{\text{att}}), \tag{10}$$

Here, $f_i(\cdot)$ denotes the feature transformation function for each scale branch, and w_i represents the corresponding weight. This fusion not only retains fine-grained information from each layer but also enriches the global semantics, making it particularly suitable for detecting small targets in side-scan sonar images.

In the Head section, the improved features are processed through a series of modules such as SPPELAN, RepNCSPELAN4, and CBAM, and then further integrated using upsampling and cross-layer concatenation (Concat) to merge multi-scale information. It is worth mentioning that in the subsequent design of the Head, we also introduce multi-level reversible auxiliary branches (through CBLinear and CBFuse modules), which re-fuse features from different levels of the Backbone, providing stronger discriminative signals for final target detection. Finally, after passing through the DualDDetect module, the network outputs detection results containing target categories, bounding box coordinates, and confidence scores, as shown in Equation 11:

$$\hat{Y} = f_{\text{head}}(F_{\text{neck}}), \tag{11}$$

The network is then trained end-to-end using a multi-task loss function, composed of localization loss, classification loss, and confidence loss, as shown in Equation 12:

$$L = \lambda_{\text{loc}} L_{\text{loc}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{conf}} L_{\text{conf}}. \tag{12}$$

This improvement strategy fully integrates the advantages of SPDConv for spatial reorganization and downsampling, CBAM's dual attention weighting ability for features, and the overall design of multi-scale fusion. It significantly enhances the model's detection performance for small targets in side-scan sonar images, while balancing real-time processing and efficiency, providing a solid theoretical and technical foundation for future practical deployment.

During model training, the original images are first uniformly resized to a standard dimension of 640×640×3. This standardization ensures consistency in input data. Subsequently, the images undergo a series of convolution and pooling operations, through which the network generates feature maps of varying scales. Shallow feature maps retain finer details for detecting small targets, while deep feature maps capture global information for large target detection. This multi-scale feature extraction mechanism effectively enhances the network's capability to detect targets of varying sizes.

4 Experiments and analysis

To validate that our SOCA-YOLO network achieves superior results on public side-scan sonar images compared to other methods, we designed the following experiments. First, we applied SwinIR to preprocess the original dataset, generating a high-resolution dataset. We then compared various object detection models, demonstrating that our network exhibits a certain level of superiority. Additionally, we conducted comparative experiments using different convolution modules and attention mechanisms to verify the effectiveness of the SPDConv module and the CBAM attention mechanism. Finally, ablation experiments confirmed that each of our proposed improvements contributes positively to the overall performance.

4.1 Environment and dataset

To comprehensively assess the effectiveness of the proposed method, we conduct experiments in a high-performance computing environment and evaluate the model on a publicly available side-scan sonar image dataset. This section provides a detailed description of the experimental setup and dataset used in our study.

4.1.1 Environment

To ensure the reproducibility of experiments and the efficiency of computational performance, the experimental environment in this study is built on the mainstream deep learning framework PyTorch, fully meeting the computational requirements for model training and inference. Detailed configuration information is presented in Table 1.

4.1.2 Dataset

The experimental dataset used in this paper is the publicly available Cylinder2 ([Dataset] yeesonmin@naver.com, 2022), utilized to evaluate the model's performance. Released in 2022, this dataset contains 478 side-scan sonar images categorized into two classes: cylinders and manta rays, with each image containing exactly one object. Each object occupies a relatively small pixel area compared to the full image, making this dataset suitable for

TABLE 1 System configuration.

Name	Configuration
Python	3.9.18
PyTorch	1.12.0
CUDA	11.3
CPU	Intel(R) Core(TM) i5-13600KF@3.50GHz
GPU	NVIDIA GeForce RTX 4070Ti (12GB)

underwater small object detection tasks. We excluded the portion containing manta rays (140 images), retaining only the 338 cylinder images. The dataset was subsequently split into training, validation, and test sets in an 8:1:1 ratio, which was then used to train the network. The basic configuration of the dataset is shown in Table 2.

4.2 Evaluation metrics

During the image restoration stage using SwinIR, the image quality was evaluated using standard metrics, including Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM).

PSNR: Given a clean image and a noisy image of size m×n, the Mean Squared Error (MSE) is defined, as shown in Equation 13:

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{13}$$

At this point, PSNR is defined as shown in Equation 14:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \tag{14}$$

Here, MAX_I^2 represents the maximum possible pixel value in the image. If each pixel is represented by 8-bit binary, then the maximum value is 255. Typically, if the pixel value is represented in B-bit binary, then $\mathrm{MAX}_I = 2^B - 1$.

SSIM: The SSIM formula is based on three comparison measures between samples x and y: luminance (Equation 15), contrast (Equation 16), and structure (Equation 17).

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{15}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{16}$$

TABLE 2 Dataset split settings.

Dataset	Images
Train	270
Val	34
Test	34

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{17}$$

Typically, $c_3 = \frac{c_2}{2}$, where μ_x represents the mean of x, μ_y represents the mean of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, and σ_{xy} is the covariance between x and y. Thus SSIM can be expressed, as shown in Equation 18:

$$\mathrm{SSIM}(x,y) = [l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma] \tag{18}$$

Setting, $\alpha = \beta = \gamma = 1$ we obtain Equation 19:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{19}$$

During the training and testing phases, the model's performance is evaluated according to the PASCAL VOC 2010 standard, which includes Precision (P), Recall (R), and Mean Average Precision (mAP). P represents the proportion of samples correctly predicted as positive out of all samples predicted as positive by the model. R represents the proportion of correctly predicted positive samples out of all true positive samples. mAP is used to comprehensively assess the model's performance across all categories by calculating the average precision at various recall thresholds. Since this paper focuses on detecting a single target type, the AP value is equivalent to the mAP value. Ideally, a higher mAP value indicates better model performance. The formulas for calculating P, R, and mAP are provided in equations Equations 20–23.

$$P = \frac{TP}{TP + FP} \tag{20}$$

$$R = \frac{TP}{TP + FN} \tag{21}$$

$$AP = \int_0^1 P(R)\,dR \tag{22}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{23}$$

Here, TP represents true positives, where positive samples are correctly predicted as positive; FP represents false positives, where negative samples are incorrectly predicted as positive; and FN represents false negatives, where positive samples are incorrectly predicted as negative.

4.3 Experimental results

To validate the effectiveness of the proposed method, we conduct a series of comparative experiments. First, we apply SwinIR for image restoration and analyze its impact on the quality of side-scan sonar images. Then, we perform multiple comparative studies, including object detection model comparison, attention mechanism comparison, and convolution module comparison. These experiments provide a comprehensive evaluation of the contributions of different components to the overall detection performance.

4.3.1 Using SwinIR for image processing

In this paper, we employ the SwinIR model as a preprocessing step to enhance the quality of original side-scan sonar images. The enhanced images are subsequently used to train and validate the SOCA-YOLO model, which is designed for small object detection. Pretrained weights from the official SwinIR GitHub repository (Liang et al., 2021) are utilized to leverage the architecture's robust super-resolution capabilities. The application of SwinIR results in processed side-scan sonar images with sharper edges, reduced noise, and improved fine details—key factors for accurate detection. Figure 7 presents comparative examples of the original and enhanced images, illustrating the effectiveness of this preprocessing step.

To intuitively assess the effectiveness of SwinIR in enhancing image clarity, we used PSNR and SSIM to compare the experimental results. The findings indicate that, compared to the original images, the processed images achieved average PSNR and SSIM values of 36.14 and 0.9807, respectively. These results demonstrate that SwinIR not only improves the visual quality and resolution of the images but also yields higher PSNR and SSIM values. Consequently, this enhancement facilitates more accurate detection of small objects, with notable improvements across various detection metrics.

4.3.2 Comparative experiment

1. Comparison of SOCA-YOLO with mainstream object detection networks.

To verify the performance of this method, we conducted comparative experiments with several mainstream object detection models, including SSD, Faster R-CNN, and various YOLO series models. Table 3 presents the experimental results of each model on the side-scan sonar dataset.

As shown in Table 3, the proposed method outperforms the original YOLOv9 and other object detection algorithms across multiple metrics. Specifically, compared to the original YOLOv9, P increases by 4.2%, R by 7.2%, and mAP50 by 6.5%. In comparison with SSD, Faster R-CNN, and the latest YOLO models, the proposed algorithm demonstrates superior performance in metrics such as P, R, and mAP. Although YOLO11 achieves a higher P of 75.8% compared to SOCA-YOLO's 71.8%, SOCA-YOLO surpasses YOLO11 in both recall and mAP50, highlighting its balanced and robust detection capabilities.

These results indicate that the algorithm significantly enhances the detection capability for small underwater targets. Figure 8 displays sample results of SOCA-YOLO target detection, illustrating noticeable improvements in both detection metrics and practical detection outcomes. However, some instances of missed and false detections remain in the detection process.

Furthermore, to provide a more comprehensive comparison of our model's superiority, we also compared the P-R curves. Figure 9 presents the P-R curve of the original YOLOv9 and the P-R curve of SOCA-YOLO.

In summary, for small object detection in underwater side-scan sonar images, the proposed method significantly outperforms mainstream object detection algorithms. Figure 10 compares the detection results of SOCA-YOLO with other models for the same target. As shown in the Figure 10, while other models produce false positives and missed detections, SOCA-YOLO accurately identifies the target, demonstrating its robustness and precision.

2. Comparison of SPDConv with other convolutional methods.

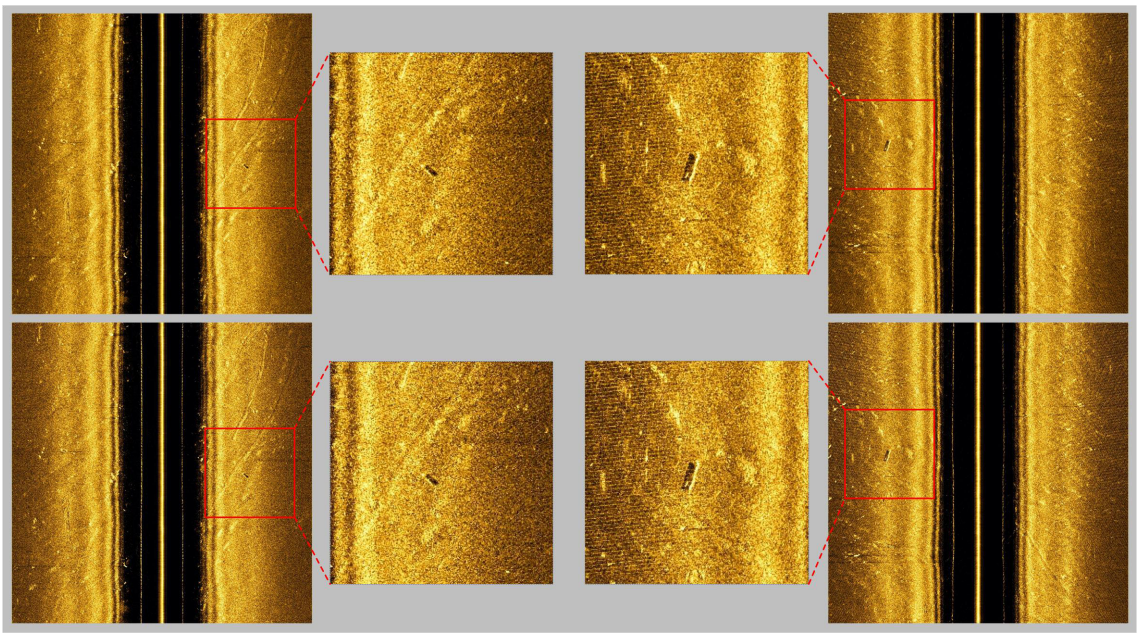

FIGURE 7
Partial results of SwinIR preprocessing, with the first row showing the original images, the second row showing the restored images, and the red boxes indicating a zoomed-in view of the target region.

TABLE 3 Comparison of SOCA-YOLO with mainstream object detection networks.

Methods	Precision / %	Recall / %	mAP50 / %
SSD	48.6	51.5	44.8
Faster-RCNN	42.4	52.9	45.5
YOLOv9	42.4	52.9	45.5
YOLOv10	70.6	65.3	71.4
YOLO11	75.8	66.7	72.0
SOCA-YOLO	71.8	72.7	74.3

To verify the contribution of the introduced convolution module SPDConv to our model's improvements, we replaced the original YOLOv9 convolution module ADown with AConv, AKConv, and SPDConv, respectively. ADown is the default convolution module in YOLOv9; AConv is a standard convolution module consisting of a pooling layer and a convolution layer; AKConv (Zhang et al., 2023) is a variable kernel convolution module that allows the kernel to dynamically adjust its shape and size based on target characteristics; SPDConv is the proposed convolution module in our SOCA-YOLO network, designed for superior detection capability on low-resolution images and small objects. We tested each module replacement on side-scan sonar images without SwinIR preprocessing. The experimental results are shown in Table 4.

As show in Table 4, SPDConv demonstrates significant advantages in object detection tasks, outperforming other convolutional modules across all key metrics. Specifically, SPDConv achieves a P of 70.4%, a R of 71%, and a mAP50 of 72.6%. These results represent improvements of 2.8%, 5.5%, and 4.8%, respectively, compared to those obtained using the original YOLOv9 convolution module, ADown. Compared to traditional convolutional modules, these improvements are particularly important for enhancing the overall performance of the YOLOv9 network. SPDConv not only improves precision but also significantly enhances the network's detection consistency (i.e., the balanced performance of P and R), making it especially suitable for small object detection in side-scan sonar images.

3. Comparison of CBAM with other attention mechanisms.

To validate the effectiveness of the attention mechanism in our network model, we conducted comparative experiments incorporating various popular attention modules, including the SE module (Hu et al., 2018), CA module (Hou et al., 2021), ECA module (Wang et al., 2020), CBAM module, and the baseline YOLOv9 network without any attention mechanism. Each attention module was integrated into the same position within the YOLOv9 network to ensure the comparability of results. Consistent training and validation datasets were used throughout the experiments to maintain fairness. The experimental results are presented in Table 5.

The results demonstrate that the performance improvements provided by attention mechanisms depend on the specific module design. Among these, CBAM achieved the best performance, significantly enhancing both detection P and R. This outcome highlights the effectiveness of CBAM's dual-branch design in capturing feature correlations at multiple levels, thereby

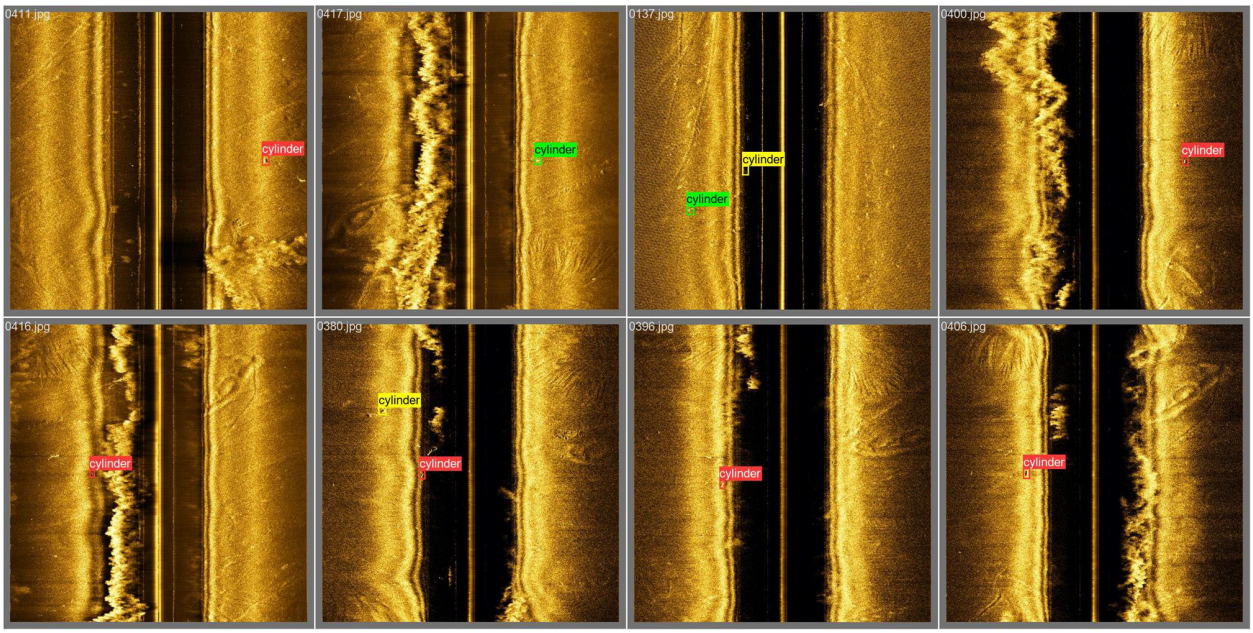

FIGURE 8
Partial results of SOCA-YOLO detection, with red boxes representing correctly detected targets, yellow boxes representing false detections, and green boxes representing missed detections.

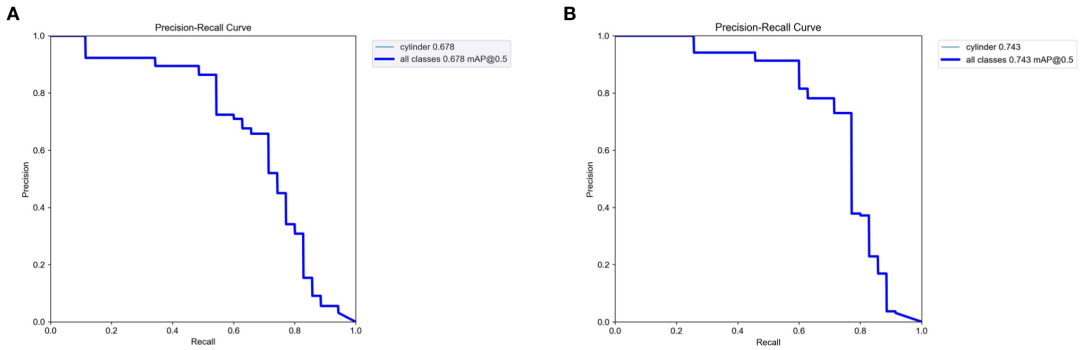

FIGURE 9
P-R curves. **(A)** shows the P-R curve of the original YOLOv9. **(B)** shows the P-R curve of SOCA-YOLO.

improving the model's ability to locate and classify targets. In comparison, the SE module, which focuses on channel attention, shows notable classification improvements in specific scenarios but offers relatively limited gains in complex environments. The CA module, by incorporating coordinate information, improves the locality of feature representations and performs well in scenarios with targets of varying aspect ratios. The ECA module strikes a balance by reducing the computational cost of attention but delivers limited improvements in small object detection.

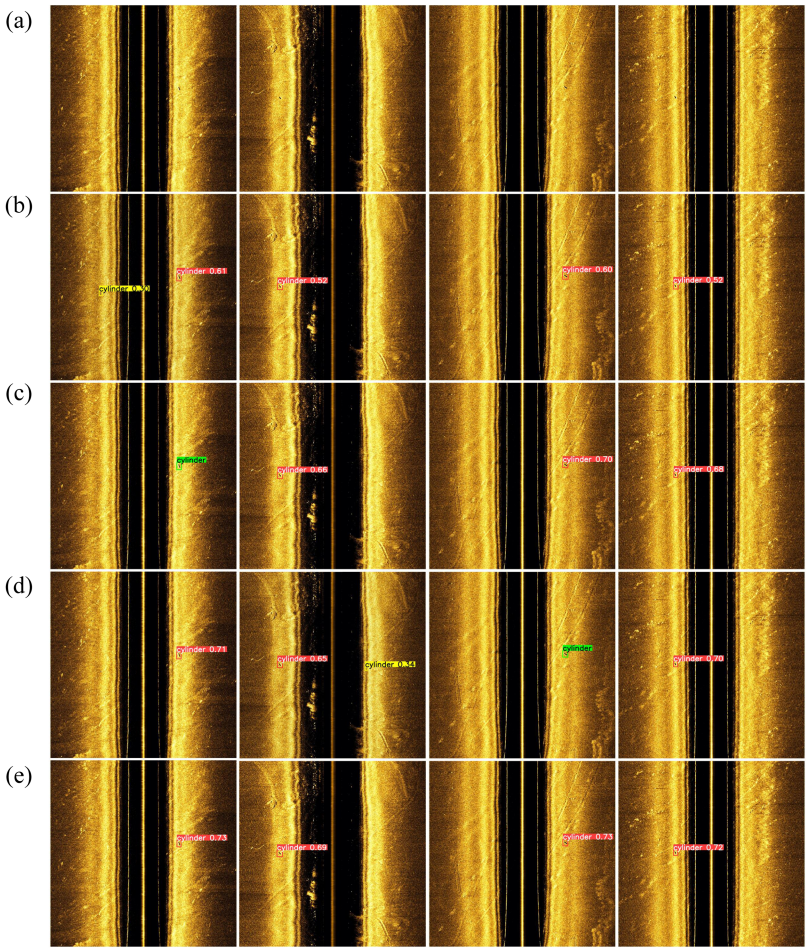

FIGURE 10
Detection results from different object detection models for four targets, with each image containing exactly one target: **(a)** Original image, **(b)** Faster-RCNN, **(c)** YOLOv9, **(d)** YOLOv10, **(e)** SOCA-YOLO. Red boxes indicate correctly detected targets, yellow boxes indicate false detections, and green boxes indicate missed detections.

TABLE 4 Comparison of SPDConv with other convolutional methods.

Model	Precision / %	Recall / %	mAP50 / %
ADwon	67.6	65.5	67.8
AConv	66.7	65.7	67.8
AKConv	68.1	69.4	69.1
SPDConv	70.4	71.0	72.6

Table 5 shows that the CBAM module achieved the best performance, with a P of 69.9%, R of 72.2%, and mAP50 of 73.3%. These values represent improvements of 2%, 6.7%, and 5.5%, respectively, compared to the baseline YOLOv9 network. However, the results also indicate that while certain attention modules provide performance enhancements, not all attention mechanisms positively impact object detection tasks. The selection and design of attention modules should be carefully adjusted and optimized to align with the specific characteristics of the task.

4.4 Ablation study

To evaluate the impact of each proposed innovation on network performance, we conducted ablation experiments on different modules. This study primarily examines the effects of using SwinIR for preprocessing the original images, replacing the original YOLOv9 convolution module with SPDConv, and adding the CBAM attention mechanism. These three enhancements were gradually incorporated into the YOLOv9 network. The experiments were conducted on the side-scan sonar image dataset, and the experimental outcomes are presented in Table 6.

TABLE 5 Comparison of CBAM with other attention mechanisms.

Model	Precision / %	Recall / %	mAP50 / %
YOLOv9	67.6	65.5	67.8
YOLOv9+SE	65.6	67.4	67.2
YOLOv9+CA	66.4	70.7	68.8
YOLOv9+ECA	69.5	70.3	66.3
YOLOv9+CBAM	69.9	72.2	73.3

As shown in Table 6, preprocessing the original dataset using SwinIR and applying the resulting high-quality images for SOCA-YOLO training and testing increased the mAP50 by 0.3%. Replacing the convolution module in the original YOLOv9 network resulted in a 5.6% increase in mAP50 compared to the original YOLOv9 results. Finally, adding the CBAM module to the YOLOv9 network with the replaced convolution module further increased the mAP50 by 0.6%. These experimental results demonstrate that each improvement is meaningful. Compared to the original network, the cumulative mAP50 increase of 6.5% significantly reduces missed detections and false detections of small objects in the original YOLOv9 network.

4.5 Generalization experiment

To validate the generalization capability of the object detection method proposed in this paper under different data distributions, we selected another publicly available side-scan sonar image dataset as the test platform (Santos et al., 2024). This dataset differs significantly from the data used during training, with marked variations in the capture environment, target types, and noise interference, thereby thoroughly assessing the model's adaptability and robustness in new scenarios. The dataset primarily comprises 1170 high-resolution side-scan sonar images and includes two types of targets—NOn-Mine-like BOttom Objects (NOMBO) and MIne-Like COntacts (MILCO)—with varying sizes and shapes. The experimental results are presented in Table 7. It can be seen that the method proposed in this paper outperforms traditional detection approaches across evaluation metrics, demonstrating strong generalization ability.

Additionally, to further analyze the detection performance across different target categories, the P-R curves for each category were plotted, as shown in Figure 11.

From the above experimental results, it is evident that the proposed method effectively adapts to noise and interference issues in public side-scan sonar image data across different marine environments, achieving high detection accuracy and recall.

5 Conclusions

In this paper, we introduced the object detection algorithm YOLOv9 with several modifications. The specific improvements are as follows: (1) Using the SwinIR model to preprocess the original

TABLE 6 Ablation study.

YOLOv9	SwinIR	SPDConv	CBAM	Precision / %	Recall / %	mAP50 / %
✓	×	×	×	67.6	65.5	67.8
✓	✓	×	×	73.5	63.4	68.1
✓	✓	✓	×	69.6	71.6	73.7
✓	✓	✓	✓	71.8	72.7	74.3

The symbol "✓" indicates that the condition was included in the experiment, while "×" signifies that the condition was not incorporated into the experimental setup.

TABLE 7 Performance of YOLO9 and SOCA-YOLO.

Method	Precision / %	Recall / %	mAP50 / %
YOLO9	82.1	65.3	74.3
SOCA-YOLO	93.7	76.2	83.7

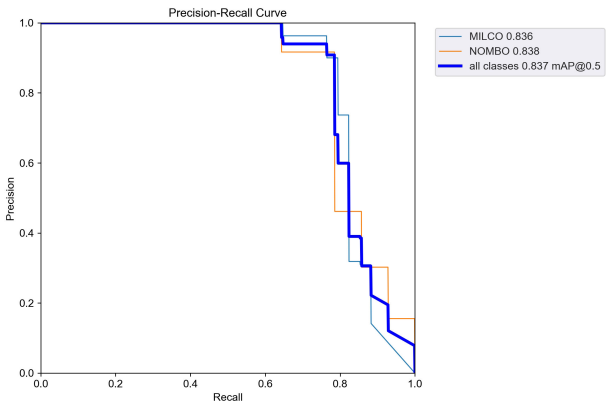

FIGURE 11
Performance of different target categories.

dataset and generate a re-divided high-resolution image dataset. (2) Adding the CBAM attention mechanism to the original YOLOv9 model to enhance focus on regions of interest. (3) Replacing the original ADown module with the SPDConv convolution module, which is more effective for small object detection. The resulting SOCA-YOLO model was applied for small object detection in underwater side-scan sonar images, achieving a Precision of 71.8%, Recall of 72.7%, and mAP50 of 74.3% on the enhanced dataset. These results indicate that the method significantly improves target detection performance in side-scan sonar images.

In future work, expanding the dataset is a crucial research direction. Although the current dataset has demonstrated the feasibility of our method, its limited scope may constrain the model's robustness and generalization ability. By incorporating additional datasets from different environments, operational conditions, and various sonar devices, we can capture a broader range of image features and noise characteristics. Such dataset expansion not only enables more comprehensive model training but also allows fine-tuning and validation of the model across various real-world scenarios. Furthermore, given the inherent unique noise characteristics of side-scan sonar images, developing specialized image processing techniques is particularly crucial. Future research can focus on designing denoising and image enhancement algorithms tailored to issues such as speckle noise and signal interference in sonar data. Exploring the integration of multimodal data is also a highly promising direction. For example, combining side-scan sonar data with optical or hyperspectral imaging data can provide complementary information, thereby improving the overall performance of detection and classification tasks. Such data fusion is expected to lead to the development of more robust and accurate models, ultimately driving new methodologies and applications in underwater imaging and analysis.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

XC: Conceptualization, Methodology, Software, Validation, Writing – review & editing, Project administration, Resources, Supervision. JZ: Validation, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. LZ: Resources, Validation, Writing – review & editing, Supervision. QZ: Project administration, Validation, Writing – review & editing. JH: Resources, Validation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This project was supported by the National Natural Science Foundation of China (Grant No.62271397 and Grant No.62171384).

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

[Dataset] yeesonmin@naver.com (2022). *cylider2 dataset*. Available online at: https://universe.roboflow.com/yeesonmin-naver-com/cylider2 (Accessed December 2, 2024).

何勇光 (2020). 海洋侧扫声呐探测技术的现状及发展[J]. 工程建设与设计. 2020 (04), 275–276. doi: 10.13616/j.cnki.gcjsysj.2020.02.328

Ali, A. M., Benjdira, B., Koubaa, A., Boulila, W., and El-Shafai, W. (2023). Tesr: two-stage approach for enhancement and super-resolution of remote sensing images. *Remote Sens.* 15, 2346. doi: 10.3390/rs15092346

Aubard, M., Antal, L., Madureira, A., and Ábrahám, E. (2024). Knowledge distillation in yolox-vit for side-scan sonar object detection. *arXiv preprint arXiv:2403.09313*. doi: 10.48550/arXiv.2403.09313

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. doi: 10.48550/arXiv.2004.10934

Burguera, A., and Oliver, G. (2016). High-resolution underwater mapping using side-scan sonar. *PloS One* 11, e0146396. doi: 10.1371/journal.pone.0146396

Cheng, C., Hou, X., Wen, X., Liu, W., and Zhang, F. (2023). Small-sample underwater target detection: a joint approach utilizing diffusion and yolov7 model. *Remote Sens.* 15, 4772. doi: 10.3390/rs15194772

Du, X., Sun, Y., Song, Y., Sun, H., and Yang, L. (2023). A comparative study of different cnn models and transfer learning effect for underwater object classification in side-scan sonar images. *Remote Sens.* 15, 593. doi: 10.3390/rs15030593

Fan, X., Lu, L., Shi, P., and Zhang, X. (2022). A novel sonar target detection and classification algorithm. *Multimedia Tools Appl.* 81, 10091–10106. doi: 10.1007/s11042-022-12054-4

Fayaz, S., Parah, S. A., and Qureshi, G. (2022). Underwater object detection: architectures and algorithms–a comprehensive review. *Multimedia Tools Appl.* 81, 20871–20916. doi: 10.1007/s11042-022-12502-1

Fengchun, L., Dianlun, Z., and Haitao, G. (2002). Image segmentation based upon bounded histogram and its application to sonar image segmentation. *J. Harbin Eng. Univ.* 2002, 1–3.

Gao, L., Zhang, J., Yang, C., and Zhou, Y. (2022). Cas-vswin transformer: A variant swin transformer for surface-defect detection. *Comput. Industry* 140, 103689. doi: 10.1016/j.compind.2022.103689

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA. 2014, 580–587. doi: 10.1109/CVPR.2014.81

Gomes, D., Saif, A. S., and Nandi, D. (2020). "Robust underwater object detection with autonomous underwater vehicle: A comprehensive study," in *Proceedings of the International Conference on Computing Advancements*, Dhaka, Bangladesh. (New York, NY, USA: Association for Computing Machinery), 1–10. doi: 10.1145/3377049.3377052

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2980–2988. doi: 10.1109/ICCV.2017.322

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., and Xue, Y. (2022). Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. doi: 10.1109/TGRS.2022.3230846

Heng, Z., Shuping, H., Jiaying, G., Yubo, H., and Honggang, L. (2024). "Research on the automatic detection method of side-scan sonar image of small underwater target," in *2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chongqing, China, 1567–1573. doi: 10.1109/ITNEC60942.2024.10733067

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 13708–13717. doi: 10.1109/CVPR46437.2021.01350

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 Aug. 2020, 42, pp. 2011–2023. doi: 10.1109/TPAMI.2019.2913372

Jannat, F.-E., and Willis, A. R. (2022). Improving classification of remotely sensed images with the swin transformer. *SoutheastCon* 2022, 611–618. doi: 10.1109/SoutheastCon48659.2022.9764016

Jiang, Y., Ku, B., Kim, W., and Ko, H. (2020). Side-scan sonar image synthesis based on generative adversarial network for images in multiple frequencies. *IEEE Geosci. Remote Sens. Lett.* 18, 1505–1509. doi: 10.1109/LGRS.2020.3005679

Jinhua, L., Jinpeng, J., and Peimin, Z. (2016). Automatic extraction of the side-scan sonar imagery outlines based on mathematical morphology. 海洋学报 38, 150–157. doi: 10.3969/j.issn.0253-4193.2016.05.014

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., et al. (2022). Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*. doi: 10.48550/arXiv.2209.02976

Li, L., Li, Y., Wang, H., Yue, C., Gao, P., Wang, Y., et al. (2024). Side-scan sonar image generation under zero and few samples for underwater target detection. *Remote Sens.* 16, 4134. doi: 10.3390/rs16224134

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). "Swinir: Image restoration using swin transformer," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, 2021, pp. 1833–1844. doi: 10.1109/ICCVW54120.2021.00210

Lin, H. (2023). Adversarial training of swinir model for face super-resolution processing. *Front. Computing Intelligent Syst.* 5, 87–90. doi: 10.54097/fcis.v5i1.11846

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., and Zhang, D. (2022). Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrumentation Measurement* 71, 1–15. doi: 10.1109/TIM.2022.3178991

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). SSD: Single Shot MultiBox Detector. In: B. Leibe, J. Matas, N. Sebe and M. Welling (eds) *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, 9905. Springer, Cham. doi: 10.1007/978-3-319-46448-0_2

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada., 9992–10002. doi: 10.1109/ICCV48922.2021.00986

Ma, Q., Jin, S., Bian, G., and Cui, Y. (2024). Multi-scale marine object detection in side-scan sonar images based on bes-yolo. *Sensors* 24, 4428. doi: 10.3390/s24144428

Peng, C., Jin, S., Bian, G., Cui, Y., and Wang, M. (2024). Sample augmentation method for side-scan sonar underwater target images based on cbl-singan. *J. Mar. Sci. Eng.* 12, 467. doi: 10.3390/jmse12030467

Polap, D., Wawrzyniak, N., and Włodarczyk-Sielicka, M. (2022). Side-scan sonar analysis using roi analysis and deep neural networks. *IEEE Trans. Geosci. Remote Sens.* 60, 1–8. doi: 10.1109/TGRS.2022.3147367

Redmon, J. (2016). "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779–788. doi: 10.1109/CVPR.2016.91

Redmon, J. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. doi: 10.48550/arXiv.1804.02767

Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 6517–6525. doi: 10.1109/CVPR.2017.690

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Santos, N. P., Moura, R., Torgal, G. S., Lobo, V., and de Castro Neto, M. (2024). Side-scan sonar imaging data of underwater vehicles for mine detection. *Data Brief* 53, 110132. doi: 10.1016/j.dib.2024.110132

Sun, C., Wang, L., Wang, N., and Jin, S. (2021). Image recognition technology in texture identification of marine sediment sonar image. *Complexity* 2021, 6646187. doi: 10.1155/2021/6646187

Sunkara, R., and Luo, T. (2022). No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects. In: M. R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak and G. Tsoumakas (eds) *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2022. Lecture Notes in Computer Science*. (Cham: Springer). 13715. doi: 10.1007/978-3-031-26409-2_27

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

Tian, X., Liu, Z., and Zhou, D. (2007). Mine target recognition algorithm of sonar image. *Sys. Eng. Electron* 7, 1049–1052.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023a). "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 7464–7475. doi: 10.1109/CVPR52729.2023.00721

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2024). Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems* 37, 107984–108011.

Wang, Y.-R., Wang, P., Adams, L. C., Sheybani, N. D., Qu, L., Sarrami, A. H., et al. (2023b). Low-count whole-body pet/mri restoration: an evaluation of dose reduction spectrum and five state-of-the-art artificial intelligence models. *Eur. J. Nucl. Med. Mol. Imaging* 50, 1337–1350. doi: 10.1007/s00259-022-06097-w

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 11534–11542. doi: 10.1109/CVPR42600.2020.01155

Wang, C.-Y., Yeh, I.-H., and Mark Liao, H.-Y. (2025). Yolov9: Learning what you want to learn using programmable gradient information. In: A. Leonardis, E. Ricci, S.

Roth, O. Russakovsky, T. Sattler and G. Varol (eds) Computer Vision – ECCV 2024. ECCV 2024. Lecture Notes in Computer Science. (Cham: Springer), 15089. doi: 10.1007/978-3-031-72751-1_1

Wen, X., Zhang, F., Cheng, C., Hou, X., and Pan, G. (2024). Side-scan sonar underwater target detection: Combining the diffusion model with an improved yolov7 model. *IEEE J. Oceanic. Eng.* 49, 976–991. doi: 10.1109/JOE.2024.3379481

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In: V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (eds) *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, (Cham: Springer). vol 11211. doi: 10.1007/978-3-030-01234-2_1

Yang, N., Li, G., Wang, S., Wei, Z., Ren, H., Zhang, X., et al. (2025). Ss-yolo: A lightweight deep learning model focused on side-scan sonar target detection. *J. Mar. Sci. Eng.* 13, 66. doi: 10.3390/jmse13010066

Yang, Z., Zhao, J., Yu, Y., and Huang, C. (2024). A sample augmentation method for side-scan sonar full-class images that can be used for detection and segmentation. *IEEE Trans. Geosci. Remote Sens.* 62, 1–11. doi: 10.1109/TGRS.2024.3371051

Yang, Z., Zhao, J., Zhang, H., Yu, Y., and Huang, C. (2023). A side-scan sonar image synthesis method based on a diffusion model. *J. Mar. Sci. Eng.* 11, 1103. doi: 10.3390/jmse11061103

Yasir, M., Liu, S., Pirasteh, S., Xu, M., Sheng, H., Wan, J., et al. (2024). Yoloshiptracker: Tracking ships in sar images using lightweight yolov8. *Int. J. Appl. Earth Observation Geoinformation* 134, 104137. doi: 10.1016/j.jag.2024.104137

Yu, Y., Zhao, J., Gong, Q., Huang, C., Zheng, G., and Ma, J. (2021). Real-time underwater maritime object detection in side-scan sonar images based on transformer-yolov5. *Remote Sens.* 13, 3555. doi: 10.3390/rs13183555

Yulin, T., Shaohua, J., Gang, B., Yonzhou, Z., and Fan, L. (2020). "Wreckage target recognition in side-scan sonar images based on an improved faster r-cnn model," in *2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Bangkok, Thailand, 348–354. doi: 10.1109/ICBASE51474.2020.00080

Zhang, X., Song, Y., Song, T., Yang, D., Ye, Y., Zhou, J., et al. (2023). Akconv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters. *arXiv preprint arXiv:2311.11587*. doi: 10.48550/arXiv.2311.11587

Zheng, M., Xing, Q., Qiao, M., Xu, M., Jiang, L., Liu, H., et al. (2022). "Progressive training of a two-stage framework for video restoration," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, 1023–1030. doi: 10.1109/CVPRW56347.2022.00115

Zhu, J., Li, H., Qing, P., Hou, J., and Peng, Y. (2024). Side-scan sonar image augmentation method based on cc-wgan. *Appl. Sci.* 14, 8031. doi: 10.3390/app14178031