remote sensing

MDPI

Article

Object Detection in Small-Sample Underwater Sonar Images Using Improved YOLOv11

Md Nur Islam, Zhen Cheng and Guanying Huo *

College of Information Science and Engineering, Hohai University, Changzhou 213200, China
* Correspondence: huoguanying@163.com

Highlights

What are the main findings?

- The improved YOLOv11 coordinates hierarchical feature extraction, local contextual processing, and scale-adaptive detection for small-sample underwater sonar imagery.
- On the SOD dataset, the proposed model achieves 99.0% precision, a 92.61% F1-score, 77.4% mAP@0.5, and 59.3% mAP@0.5:0.95.

What are the implications of the main findings?

- The coordinated framework improves the detection and localization of small, low-contrast sonar targets in cluttered seabed environments.
- The method offers a promising approach for sonar object detection when labeled training data are limited, particularly when detection accuracy is prioritized over computational efficiency.

Abstract

Underwater object detection in sonar images is challenging because of low image resolution, acoustic noise, seabed clutter, weak target boundaries, substantial target-scale variation, and limited labeled data. To address these challenges, we develop an improved *You Only Look Once version 11* (YOLOv11)-based framework for small-sample sonar object detection. Rather than introducing new individual network components, the proposed framework coordinates hierarchical feature extraction, local contextual feature processing, and scale-adaptive detection according to the characteristics of sonar imagery. Specifically, a ResNet50 backbone is used to obtain deeper multilevel representations, a RepViT-enhanced C3k2 module is incorporated to strengthen the representation of small and low-contrast targets in cluttered backgrounds, a dual-pooling operation is employed to preserve prominent local echo responses and broader contextual information, and Dynamic Head is used to improve feature aggregation across different target scales and spatial locations. In addition, data augmentation and image synthesis techniques are applied to increase sample diversity and mitigate overfitting under limited-data conditions. Experiments on the *Sonar Object Detection* (SOD) and *Underwater Acoustic Target Detection* (UATD) datasets demonstrate the effectiveness of the proposed framework. On the SOD dataset, the model achieves an F1-score of 92.61%, an mAP@0.5 of 77.4%, and an mAP@0.5:0.95 of 59.3%. On the UATD dataset, it achieves an F1-score of 84.0%, an mAP@0.5 of 85.1%, and an mAP@0.5:0.95 of 38.1%. The experimental results confirm the model's robustness in challenging underwater scenarios and its strong potential for high-precision applications in oceanic exploration, maritime rescue, and surveillance.

Keywords: underwater object detection; sonar imagery; side-scan sonar; deep learning; YOLOv11; small object detection

Check for updates

Academic Editors: Francesco Serafino, Gianmarco Lunghi and José Carlos Nieto-Borge

Received: 6 July 2026
Revised: 4 September 2026
Accepted: 29 September 2026
Published: 9 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Sonar imaging is a widely used underwater acoustic detection technology for underwater object detection, search and rescue operations, and sediment detection [1,2]. Among these applications, the detection of tiny underwater targets has remained a key focus of research for scholars worldwide. Traditional underwater object detection techniques typically include two main components: feature extraction and detection using a predefined classifier [3]. However, these methods often involve complex processing steps, and the effectiveness of manually extracted features is inherently limited, which constrains their overall performance.

Therefore, light attenuation underwater significantly restricts traditional optical imaging techniques, making optical cameras challenging [4] to see. As a result, most optical imagery is only appropriate for short-range underwater object detection. On the other hand, sonar systems provide significant advantages for underwater measurements with long-range capability and high efficiency. They have effectively mapped seabed topography, conducted underwater sediment searches, and explored seabed minerals [5–8]. Among these, sonar offers superior imaging capabilities, allowing it to approach the seabed via towing and capture high-resolution seabed images used in underwater search operations, as illustrated in Figure 1a. This provides observers with the shape of object contours and is one of the essential methods for detecting and identifying seabed objects [9–12].

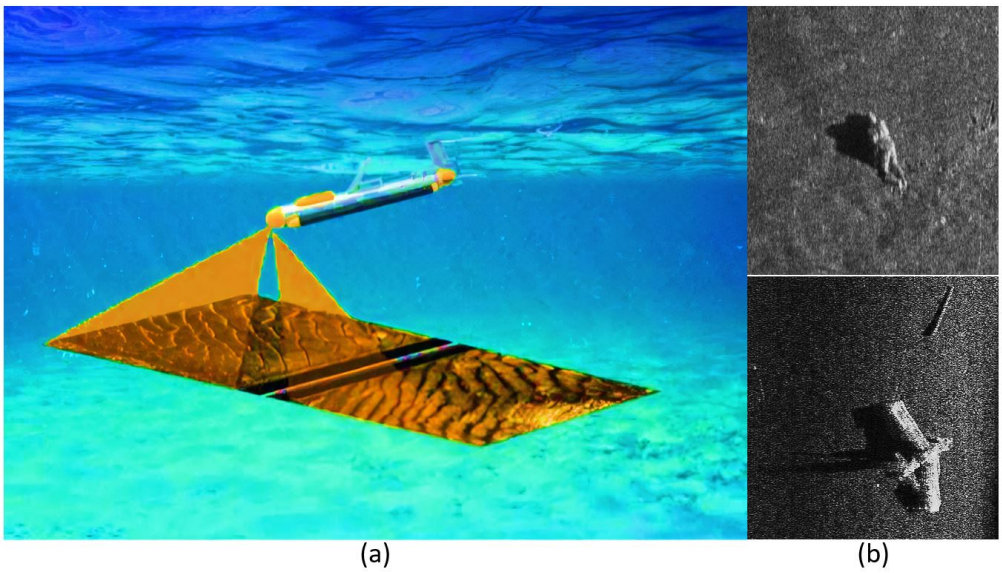


Figure 1. (**a**) Underwater Side-scan Sonar (SSS) working principle and (**b**) collected sonar images.

Sonar target detection methods primarily relied on digital signal processing techniques to analyze target contours. Common methods included texture recognition using Fourier transforms, K-means clustering, and Markov Random Field (MRF) models [13]. This model is characterized by high reliability and strong stability. However, its heavy reliance on manual expertise, coupled with the inherent complexity of sonar imaging mechanisms, renders it vulnerable to various environmental factors such as reverberation, multipath interference, and noise [14,15]. These issues result in reduced detection accuracy and limited generalizability, particularly in the context of poor-quality sonar imagery, ultimately hindering its effectiveness in meeting the demands of modern underwater target detection.

Consequently, in the last several years, the rapid invention and widespread application of deep learning technologies [16] have led to significant progress in addressing many challenges associated with using conventional techniques for object detection. Deep learning methods have demonstrated considerable promise in automating the feature

extraction process, consequently enhancing the accuracy of sonar detection, especially in intricate and noisy backgrounds. Recent studies have further enhanced underwater sonar object detection by incorporating attention mechanisms, multi-scale feature fusion, and transformer-based architectures to improve robustness against noisy backgrounds and small target objects [17–20]. These technological advancements have made sonar detection systems more dependable and less dependent on manual intervention, offering a more efficient and scalable solution for underwater object detection [21–23].

Traditional approaches to sonar image analysis typically rely on pixel-level features [24], grayscale thresholding techniques [25,26], prior knowledge of object characteristics [27,28], and handcrafted filtering methods [29]. However, the inherently low quality and high noise levels of marine sonar images hinder the extraction of reliable pixel features and the determination of robust grayscale thresholds [27]. In contrast, deep learning-based techniques have recently demonstrated significant potential and have been progressively applied in underwater object detection tasks [30,31].

Recent studies have also explored advanced underwater image enhancement to address degradation and limited training data. Wang et al. proposed a reinforcement-learning-based meta underwater camera for adaptive enhancement [32], followed by a self-organized enhancement framework for handling diverse underwater degradation [33]. More recently, WaterCycleDiffusion employed visual–textual fusion and diffusion priors to improve underwater image quality while reducing dependence on paired training data [34]. In maritime sensing, Li et al. developed a graph representation learning framework for reliable target association in high-frequency surface wave radar (HFSWR) [35]. However, these methods mainly focus on image enhancement or target tracking rather than small-sample underwater sonar object detection, which remains challenging due to noise, clutter, and limited annotated data.Liu et al. demonstrated that increasing the training dataset can improve detection generalization and reduce overfitting [36]. Zou et al. developed YOLOv11-SDC to enhance small-object detection using sparse features and attention-based feature fusion [37]. Zhang et al. proposed ESL-YOLO to address blurred boundaries, noise, and scale variations, while Zhang and Gao introduced RCDI-YOLO to improve detection in cluttered and complex SSS environments [38,39]. However, jointly addressing limited annotated data, sonar noise and clutter, multi-scale feature representation, and accurate target localization remains challenging.

Thus, *SeabedObjects-KLSG Detection* (SOD) [40] and *Underwater Acoustic Target Detection* (UATD) collect corresponding images based on the intensity of sound wave reflections. In side-scan sonar operations, the towfish (or towed body) containing the transducer array is towed behind a surface vessel to image a swath of the seafloor. The interaction of acoustic waves with structurally complex targets generates enhanced backwards scattering, leading to a high-intensity return. Consequently, the pixel intensity values in the region of the sonar image corresponding to the target are markedly elevated. The object blocks the sound waves near the bright object area, creating a corresponding shadow region shown in Figure 1b. The remaining areas are background regions caused by seabed noise, with the brightness in the sonar image gradually decreasing from the highlighted object areas to the shadow regions.

Therefore, due to the sonar system's ability to image based on different echo signals, it holds significant application value in underwater operational environments. In recent years, several successful cases have emerged, where sonar images have aided in identifying underwater objects during emergency search and rescue operations. For example, the Shenzhen Maritime Safety Administration used sonar to quickly locate the approximate position of a sunken ship in a short period [41]. In a rescue operation following a bridge collapse, where vehicles fell into the water, sonar was also successfully used to find two submerged

vehicles. In addition to emergency rescue jobs, sonar has been effectively employed in search and salvage activities, such as collecting fallen anchors from ships [42,43].

Current underwater sonar object detection methods face several interconnected challenges. First, collecting and annotating sonar images is expensive, resulting in limited labeled training data. Variations in sonar viewing angle, sensing distance, and target orientation also cause considerable changes in the appearance and scale of the same object, making it difficult to construct a sufficiently diverse training dataset. Second, side-scan sonar targets are generally represented by a high-intensity echo highlight and an associated acoustic shadow. The echo highlight corresponds to the reflected acoustic response from the target surface, whereas the shadow region is formed when the target blocks the incident acoustic signal. These two regions provide complementary information about the presence, position, and geometric characteristics of the target. Because seabed clutter may generate responses similar to target highlights, reliable detection requires the joint use of the target echo, its associated acoustic shadow, and the surrounding spatial context. In the proposed framework, the dual-pooling operation preserves prominent local echo responses while retaining broader contextual information around the target. This operation is intended to improve target-context representation; however, it does not constitute an explicit physical model of acoustic-shadow formation.

Third, low resolution, reverberation noise, weak target boundaries, and complex seabed backgrounds make the extraction of small-target features difficult and frequently cause missed detections and false positives. Finally, detection accuracy must be balanced against computational complexity to support the processing of large volumes of sonar data and potential real-time deployment.

These challenges affect multiple stages of the detection pipeline and cannot be adequately addressed by modifying only one component. Limited labeled data require augmentation to increase sample diversity, whereas weak target responses and seabed clutter require stronger hierarchical and contextual feature representations. Furthermore, variations in viewing geometry and sensing distance produce substantial changes in target scale, position, and highlight-shadow separation, requiring adaptive multiscale feature aggregation. Accordingly, the proposed framework integrates established components in a task-driven manner: ResNet50 strengthens hierarchical feature extraction, the RepViT-enhanced C3k2 module refines local features and suppresses clutter, dual pooling preserves complementary local and contextual information, and Dynamic Head performs adaptive multiscale feature aggregation.

The key contributions of this research are summarized as follows:

1. **Sonar-oriented YOLOv11 framework:** We develop a task-driven YOLOv11 framework for small-sample underwater sonar object detection. The framework integrates established network components according to the characteristics of sonar imagery, including low resolution, acoustic noise, seabed clutter, weak target boundaries, target-scale variation, and limited labeled data.
2. **Enhanced feature extraction:** The standard backbone is replaced with ResNet50 to obtain deeper multilevel representations. In addition, the C3k2 module is enhanced with RepViT to improve the representation of small and low-contrast targets in cluttered seabed backgrounds.
3. **Sonar-oriented feature processing and detection:** A dual-pooling operation is used to preserve prominent local echo responses and broader contextual information associated with target and acoustic-shadow regions. Dynamic Head then performs scale, spatial, and channel-aware feature aggregation to accommodate variations in target size, position, viewing angle, and appearance.

4. **Dynamic detection head:** The conventional detection head is replaced with Dynamic Head, which uses channel-, spatial-, and scale-aware attention to improve feature aggregation and adaptability to targets with different sizes and locations.
5. **Data augmentation and evaluation:** A data augmentation strategy is applied to increase sample diversity under limited-data conditions. Extensive experiments and ablation studies are conducted on the SOD dataset to evaluate the contribution of the proposed design.

2. Materials and Methods

Automatic recognition of objects in sonar imagery is a key component in underwater perception, supporting applications such as oceanographic survey and subsea target detection. In practical underwater search missions, sonar systems are typically mounted on vessels to scan designated sea areas. The acquired data are then manually inspected to determine the presence of potential targets. However, this process is both time-consuming and inefficient, highlighting the need for accurate and automated sonar image object detection methods.

A number of recent works have explored sonar and underwater remote sensing detection from different perspectives, each addressing specific limitations in existing pipelines. Yu et al. [23] proposed Transformer-YOLOv5 by integrating transformer blocks into the YOLOv5 backbone, enabling the model to capture global contextual information and long-range dependencies that are often limited in conventional convolutional architectures. Wang et al. [13] introduced AGFEN, an adaptive global feature enhancement network designed to mitigate noise interference and improve feature representation under degraded and low-contrast sonar conditions. Miao et al. [17] addressed the small-object detection problem by proposing IMFIN, which enhances multi-layer feature interaction to preserve fine-grained information of tiny targets in deep feature hierarchies. Zhang et al. [18] developed UUVDNet, a lightweight detection framework for unmanned underwater vehicles, employing attention mechanisms to balance detection accuracy and computational efficiency. More recently, Zhao et al. [44] proposed UWS-YOLO, which combines transfer learning with an Orthogonal-Snake Convolution mechanism to further improve sonar image recognition performance.

2.1. YOLOv11

On 30 September 2024, Ultralytics announced their most recent object detection algorithm, as shown in Figure 2 [45]. The accuracy and speed of YOLOv11 are much better than those of previous YOLO series versions. The framework of YOLOv11 comprises three key modules: a backbone for extracting features, a neck for aggregating them, and a detection head for producing the final predictions. These improvements enhance its performance in several object detection tasks.Recent underwater sonar detection frameworks have integrated transformer-based architectures and adaptive feature enhancement modules to improve detection accuracy in complex marine environments [20,44].

YOLOv11 adopts a multi-stage downsampling backbone to progressively extract hierarchical features at different spatial resolutions. The backbone integrates improved C3k2 blocks as a replacement for the C2f module, enhancing feature representation and computational efficiency. Through convolution, batch normalization, and SiLU activation (CBS) operations, the Spatial Pyramid Pooling Fast (SPPF) module increases receptive field diversity via multi-scale feature aggregation. In addition, the C2PSA module incorporates a pyramid slice attention mechanism to enhance contextual feature representation. The neck network adopts a PAN-FPN architecture, enabling effective fusion of low-level spatial

features and high-level semantic information to improve multi-scale object detection and address localization challenges in complex scenes.

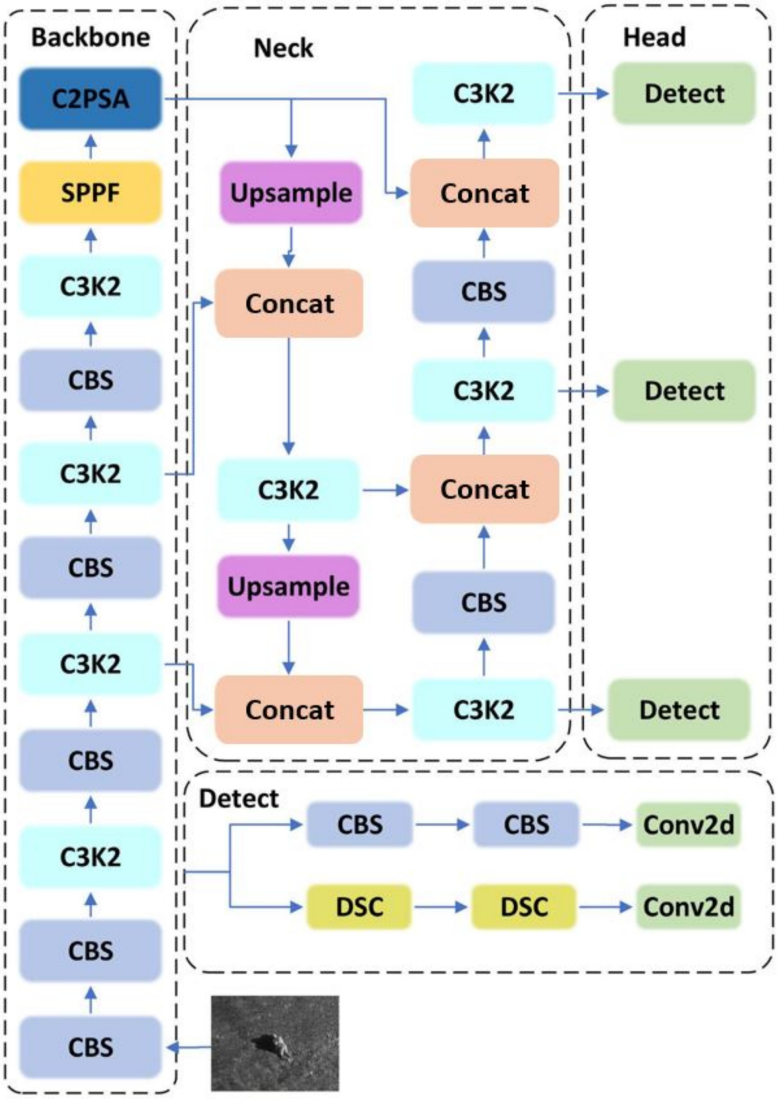


Figure 2. Structure of YOLOv11.

YOLOv11's detection head has a disjointed structure with distinct branches to predict location and category data. Depending on the task, it identifies suitable loss functions: Binary Cross-Entropy (BCE) Loss is commonly applied in object tasks.

2.2. Building the Dataset

This chapter will explore techniques to augment and expand sonar image data for improved analysis. Data augmentation was applied during training to overcome the limited dataset size and enhance the model's generalization capability. The datasets was divided into training and testing with 4:1 ratio.

The manual annotation was carried out using the LabelImg tool, in which each object was assigned a unique class ID number corresponding to its specific location and identification. To enhance generalization, we employed a number of data augmentation methods, among which are brightness adjustment, horizontal flipping, scaling, and rotation, presented in Figure 3. These augmentations provide the model with a diverse set of visual variations to process, hence promoting robustness and enhancing the accuracy level.

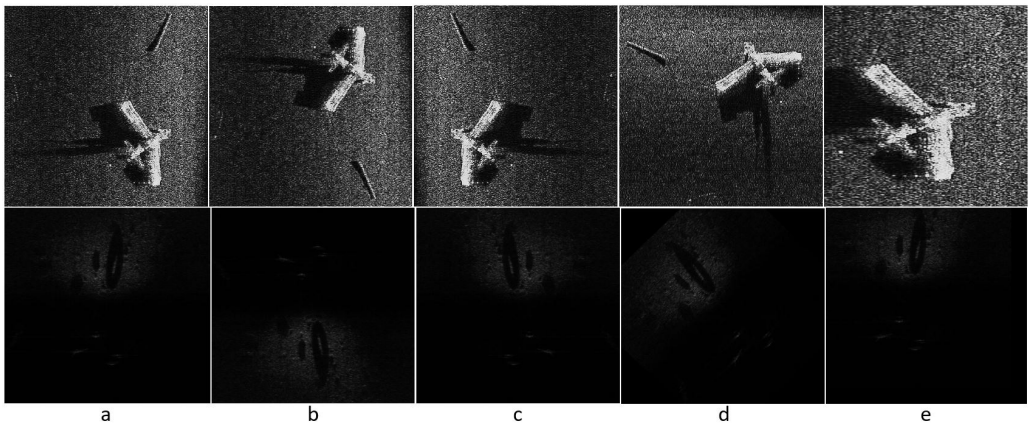


Figure 3. Example of SOD and UATD data augmentation techniques. (**a**) Original image. (**b**) Vertical flip. (**c**) Horizontal flip. (**d**) Rotation. (**e**) Pixel Translation.

2.3. Enhanced YOLOv11 Object Detection Method

This deep learning architecture is specifically designed for advanced image detection tasks, particularly within domains such as sonar imagery analysis shown in Figure 4. At its core, the model utilizes ResNet50, a convolutional neural network widely recognized for extracting detailed characteristics from input imagery while keeping critical spatial information by using residual connections. The architecture integrates specialized components like the C3K2_RVB (RipViT Block) shown in Figure 5, which combines convolutional layers with transformer-based mechanisms, thus enhancing its ability to represent complex features. The extracted features are subsequently processed through modules such as Concat, Upsample, and CBS, which enable the retention of high-resolution data and facilitate multi-scale feature extraction. The Dynamic Head (DyDetect) shown in Figure 6 module leverages these refined features to perform the final detection tasks, focusing on object localization and classification. Moreover, additional elements like Spatial Pyramid Pooling (SPPF) and C2PSA contribute to the model's effectiveness by handling varying object scales, thereby improving overall detection accuracy. This modular and dynamic approach equips the architecture to excel in tasks that require intricate spatial understanding and robust feature contextualization.

Sonar photography, characterized by its vast fields of vision and extended shooting distances, often features dense object distributions, numerous small objects, intricate backdrops, and frequent object occlusions. Despite this, the YOLOv11 detector has commendable detection accuracy on standard datasets. These characteristics make it more likely that the YOLOv11 model may overlook or misidentify tiny, far-off objects. We enhanced the feature fusion procedure, added a practical detection head, and tuned the backbone network to overcome the original YOLOv11 model's shortcomings in comprehending complex, long-range connections in sonar-captured picture data.

The proposed modifications are motivated by the physical and visual characteristics of sonar imaging. Unlike natural images, sonar images are formed through acoustic propagation, reflection, and scattering. Their appearance is influenced by the operating frequency, incidence angle, sensing distance, target geometry, seabed properties, reverberation, and acoustic shadowing. ResNet50 is therefore used to extract hierarchical representations of weak and low-contrast echo responses. The C3k2_RVB module refines local target features and suppresses clutter-related activations. The dual-pooling operation preserves prominent local echoes and broader contextual information associated with the target and its surrounding acoustic-shadow region. Dynamic Head further adapts feature aggregation to variations in target scale and spatial appearance caused by changes in sensing distance and viewing geometry.

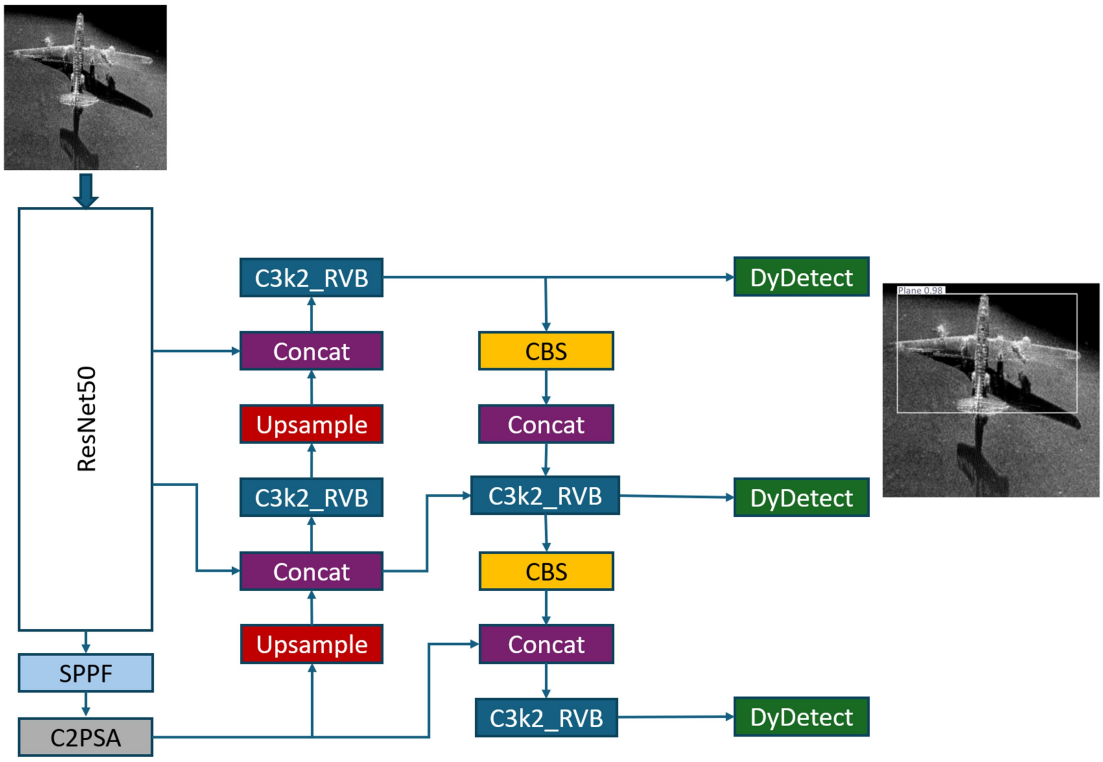


Figure 4. The framework diagram of the improved YOLOv11 detection model.

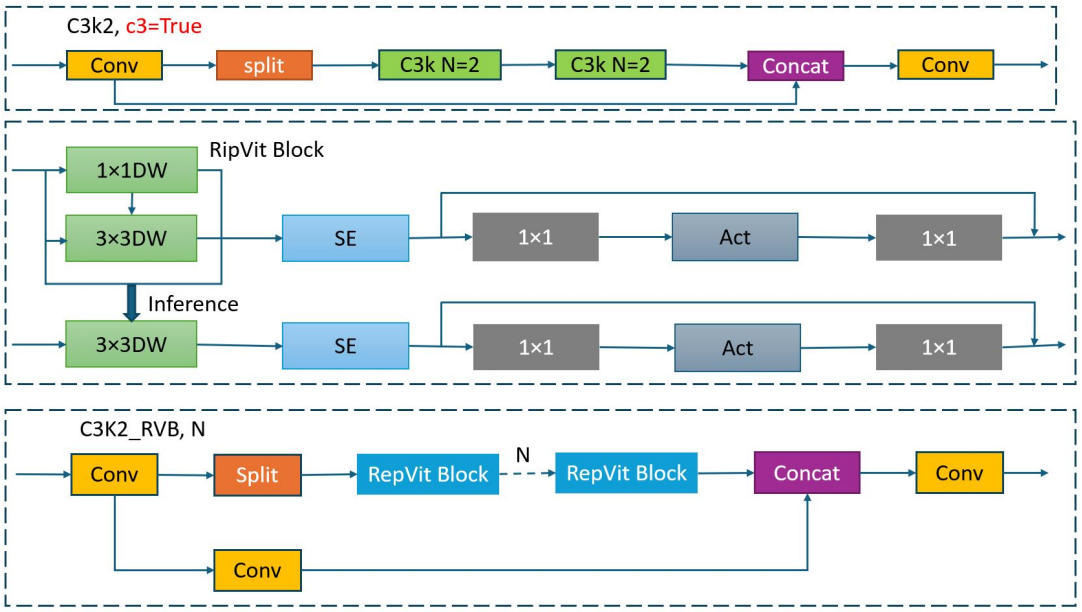


Figure 5. Configuration of the C3K2 feature, RipViT Block, and C3K2_RVB.

The current framework does not explicitly simulate acoustic propagation or incorporate a differentiable acoustic-scattering model. Such physical modeling would require acquisition metadata and environmental parameters, including sonar frequency, grazing angle, sensor configuration, target material and geometry, and seabed properties, which are unavailable in the SOD and UATD datasets. Therefore, the present method uses sonar-specific visual and contextual characteristics as architectural guidance rather than directly embedding an acoustic-scattering equation into the network.

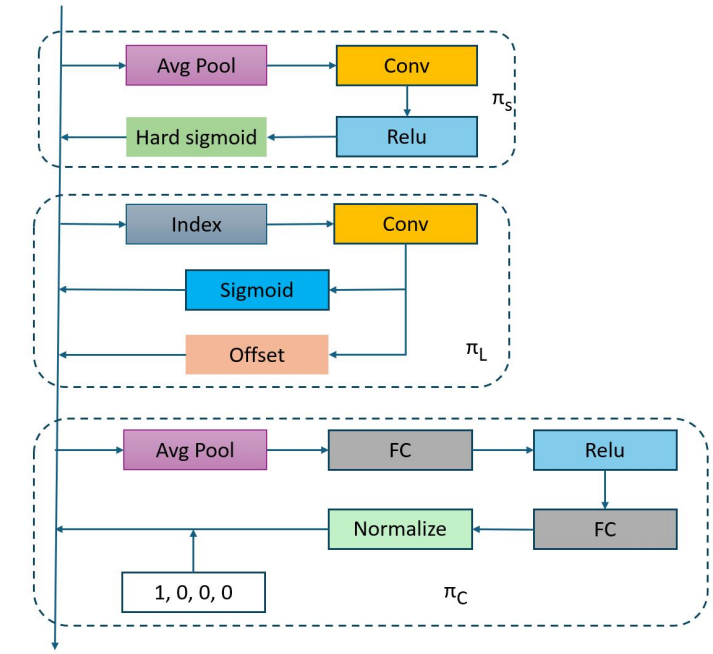


Figure 6. Structure of Dynamic Head.

2.3.1. ResNet-50 Backbone

The ResNet-50 backbone, based on residual learning principles, efficiently handles the issue of network deterioration as network depth grows. By incorporating residual connections, ResNet-50 enables gradients to flow straight through identity mappings, allowing the training of much deeper networks without the vanishing gradient problem. This approach significantly boosts the network's capacity to extract high-level semantic characteristics, making it ideal for object identification tasks involving complex scenes.

ResNet-50 is constructed using four layers of residual blocks after an initial convolutional stage or layer, as illustrated in Figure 7. The residual block consists of a set of convolutional layers and a shortcut connection that can bypass these layers to pass information. The following is a statement of the fundamental residual block computation procedure:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{1}$$

where $\mathcal{F}(\mathbf{x}, \{W_i\})$ denotes the residual mapping to be learned, and x and y represent the input and output feature maps, respectively. The shortcut link guarantees that the network may readily learn identity mappings if required, thereby optimizing intense networks.

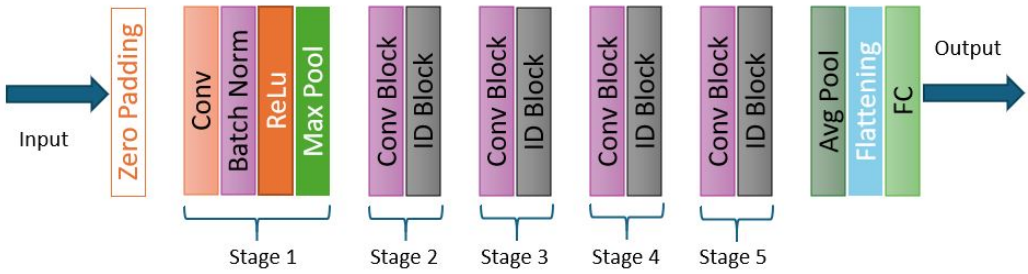


Figure 7. Structure Of ResNet50.

In the ResNet-50 design, bottleneck blocks are implemented to minimize computational costs while preserving high representational capacity. Each bottleneck block has

3 convolutional stages or layers (1×1, 3×3, and 1×1 convolutions) as shown in Figure 8, allowing for efficient feature transformation and dimensionality reduction or expansion.

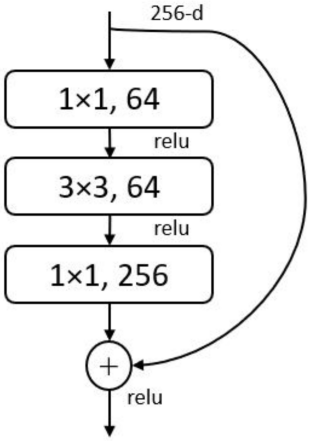


Figure 8. Bottleneck Block.

In the proposed framework, ResNet50 is not introduced as a new backbone architecture. It is selected because its residual hierarchy provides deeper multilevel representations than the original lightweight YOLOv11 backbone. Shallow features retain the spatial details of small echo highlights, whereas deeper features represent broader target and background context. Residual connections also support stable optimization when training with a limited sonar dataset. However, this improvement introduces additional computational cost, which is quantitatively evaluated in the experiments.

2.3.2. Improved C3K2 Module

The C3k2 in YOLOv11 exhibits poor feature extraction performance, particularly in recognizing tiny objects in sonar footage. As shown in Figure 5, the RepViT Block [46] was incorporated into the C3k2 feature to enhance the accuracy of tiny object detection. The result is the enhanced C3k2-RVB module. The spatial attention mechanism in this module may adaptively modify the receptive field's dimensions, expanding the processing capacity of the network to accommodate features of varying sizes.

A neural network module called the RepViT Block makes use of convolutional networks with self-attention mechanisms. It employs structural reparameterization to increase feature extraction during training. This architecture incorporates several branches, including convolutional and skip connections, which are simplified into a single-branch structure during inference. This change considerably increases computing efficiency. The design successfully uses two important advantages:

Convolutional neural networks (CNNs) are effective at capturing local details, edges, and texture patterns. These properties are particularly useful for representing the weak echo responses and fine-grained structures of small sonar targets. In the RepViT block, depthwise convolutions efficiently perform local spatial feature extraction with relatively low computational cost. Rather than relying on a full transformer self-attention mechanism, RepViT uses channel-wise feature recalibration to assign greater importance to informative feature responses and suppress irrelevant background activations. This operation helps distinguish small sonar targets from seabed clutter while preserving computational efficiency. The combination of depthwise convolution, channel attention, and structural reparameterization improves feature representation under complex sonar conditions.

During the training phase, the C3k2-RVB module forms a multi-branch structure using 2 parallel convolution layers with depthwise separability (3×3 DW and 1×1 DW) to enhance feature extraction capabilities. The reparameterization technique reduces compu-

tation costs significantly by condensing these branches to a single 3 × 3 DW convolution component or layer during the inference phase. Additionally, the quantity of feature channels and combined data is changed using a 1 × 1 convolution layer, and the global feature modeling capabilities are improved using a self-attention module (SE). To increase efficiency and feature fusion capabilities, the structure is ultimately reduced to a 3 × 3 DW convolution stage or layer during the inference phase. Two 1 × 1 convolution layers and an SE module follow. This new architecture makes the model more appropriate for realistic inference in resource-constrained settings by lowering computing costs and improving the model's accuracy at detecting tiny objects under complex conditions.

2.3.3. Improved Detection Head

The detection head of YOLOv11 may have difficulty distinguishing targets with different sizes, locations, and appearances, particularly in complex sonar backgrounds. Variations in sensing distance, viewing angle, and object dimensions cause the same target category to appear at substantially different scales and spatial positions. A fixed detection head may therefore fail to adapt its feature-processing strategy, resulting in missed detections or inaccurate localization of small sonar targets. Attention-guided detection heads and adaptive feature aggregation strategies have recently shown promise for underwater object detection under cluttered acoustic conditions [19,44].

As illustrated in Figure 8, Dynamic Head [47] is incorporated to perform channel-aware, spatial-aware, and scale-aware feature aggregation. Channel-aware attention recalibrates feature responses according to their importance for classification and localization. Spatial-aware attention emphasizes informative target regions and suppresses irrelevant seabed-background activations, whereas scale-aware attention adaptively combines information from different feature-pyramid levels. These operations improve the detector's ability to process small, low-contrast, and spatially varying sonar targets. The Dynamic Head operation can be expressed as

$$W(F) = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F \tag{2}$$

where $F \in \mathbb{R}^{L \times S \times C}$ denotes the input feature tensor, and π_L, π_S, and π_C represent the scale, spatial, and channel-attention functions, respectively. These attention operations are applied sequentially to adaptively refine the feature representation. In the proposed framework, Dynamic Head complements the hierarchical features extracted by ResNet50 and the target-feature refinement performed by C3k2_RVB.

The computational cost and inference speed of Dynamic Head are evaluated in the ablation and comparative experiments. This analysis provides a practical assessment of the trade-off between improved sonar detection performance and the additional computational complexity introduced by the enhanced detection head.

2.3.4. Improvements in the Network Feature Layer Based on Combined Pooling

The pooling layer, the down-sampling layer, is an essential and standard operation in CNNs. In a network model, the pooling layer typically follows the convolutional layer. It performs feature sampling within the region defined by the pooling window size, aiming to reduce the network's feature parameters and dimensions, which helps decrease the occurrence of overfitting. Following the execution of the pooling operation on the input feature map, it can highlight key features in the data and suppress noise interference to the network:

$$y_{i,j} = \max\{x_{p,q} \mid p \in R_i,\ q \in R_j\} \tag{3}$$

$$y_{i,j} = \frac{1}{|R|} \sum_{p \in R_i} \sum_{q \in R_j} x_{p,q} \tag{4}$$

In addition, pooling enhances the network model's scale and rotational invariance. Even if the data image undergoes pixel rotations, it can still reliably extract stable features of the object.

In CNNs, the commonly used pooling methods are mainly Max Pooling, shown in Equation (3) [48–50], and Average Pooling, shown in Equation (4) [51,52]. As shown in Figure 9, Max Pooling chooses the region's maximum value as the output for the pooling window. During backpropagation, only the gradient at the location of the maximum value is propagated backwards, while the gradients at other places are set to zero. Max Pooling is effective at preserving local texture feature information of the image. Figure 10 illustrates the forward and backward propagation processes of Average Pooling. This pooling method computes the average value within the pooling region. It assigns this average to each pixel location during gradient propagation, which is beneficial for extracting global semantic feature information from the image.

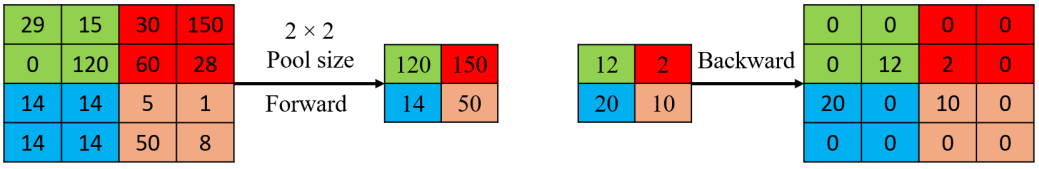


Figure 9. Forward and backward propagation process of Max Pooling.

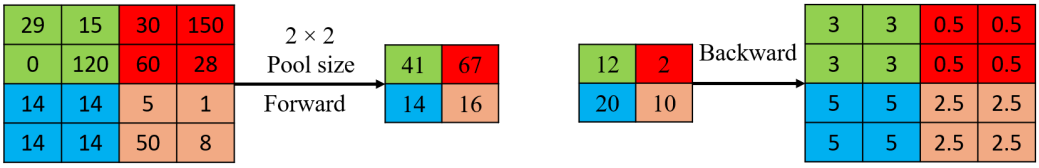


Figure 10. Forward and backward propagation process of Average pooling.

This paper combines max pooling and average pooling as a supporting feature-processing operation to preserve complementary local and contextual information in sonar images. Max pooling retains prominent local responses associated with strong acoustic echoes, whereas average pooling preserves broader contextual information related to the surrounding target and acoustic-shadow region. Their combination enriches the feature representation and supports the extraction of sonar-specific target characteristics. Figure 11 shows the network block diagram with the integrated dual-pooling operation.

2.4. Experimental Setup and Parameter Configurations

The specific experimental setup in this chapter is shown in Table 1, which outlines the setup used for the experiments and provides an overview of the hardware, software, and training parameters that were crucial for the study. The system runs on Windows 11, powered by an Intel Core i9-14900F processor with 64 GB of RAM, which provides enough processing power for the tasks. The experiments also rely on the NVIDIA GeForce RTX 4090, 24564MiB graphics card, which is great for handling the heavy computations required for deep learning and image processing. On the software side, the tools used include OpenCV 4.10, CUDA 11.7, and PyTorch 2.0.1. As for the training settings, training was conducted over 200 epochs with a batch size of 16 and a learning rate of 0.001, facilitating stable and efficient model convergence.

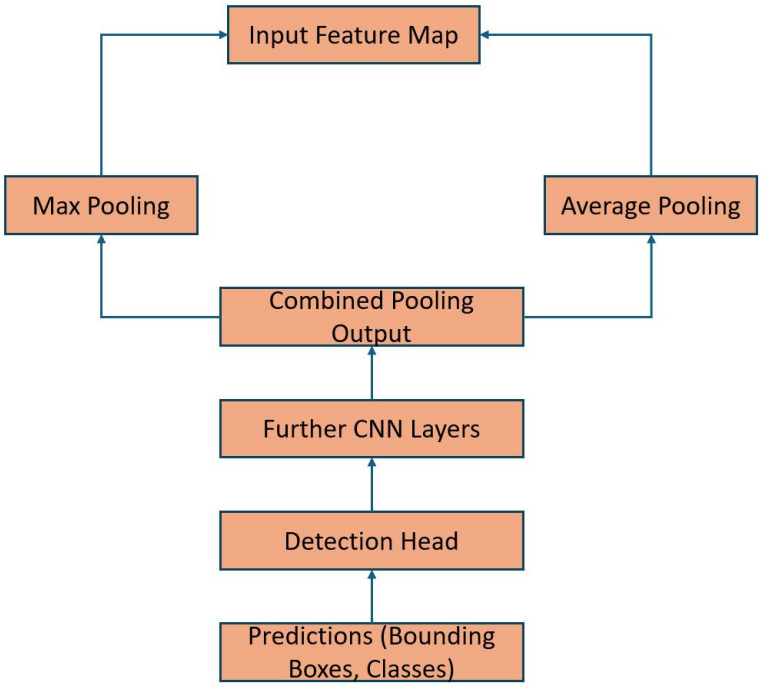


Figure 11. Improved combined dual-pooling network block diagram.

Table 1. Configuration table for software and hardware environment.

	Specific Configuration Information
Hardware	System: Windows 11 CPU: Intel(R) Core (TM) i9-14900F RAM: 64 G Graphics Card: NVIDIA GeForce RTX 4090
Software Environment	OpenCV 4.10, CUDA 11.7, PyTorch 2.0.1.
Training parameters	Epoch = 200, Batch = 16, learning rate = 0.001

2.5. Performance Assessment Metrics

In sonar object recognition tasks, false positives (FP) and negatives can arise due to background noise, low-resolution imagery, and object resemblance. Prevalent measures to assess the efficacy of detection algorithms include Mean Average Precision (mAP@0.5 and mAP@0.5:0.95). These metrics provide an extensive assessment of a model's proficiency in correctly identifying and localizing objects.

Precision: Precision quantifies the fraction of successfully recognized objects among all detections produced by the model. It indicates how reliable the detections are. Precision is computed as:

$$P = \frac{TP}{TP + FP} \tag{5}$$

where TP stands for true positives (correct detections), while FP for false positives (incorrect detections).

Recall: Recall assesses how effectively a model can detect every relevant element in an image. Formally, it is the proportion of true positives to all actual objects. The formula is:

$$R = \frac{TP}{TP + FN} \tag{6}$$

where FN indicates false negative cases when the model fails to identify an object.

Mean Average Precision: Mean Average Precision [53] is a critical assessment metric in object detection. The mAP@0.5 measure computes the average accuracy for every class using 0.5 as the Intersection over Union (IoU) threshold. Simultaneously, it mAP@0.5:0.95 offers a comprehensive performance assessment by averaging accuracy over several IoU thresholds ranging from 0.5 to 0.95, incremented by 0.05. An elevated mAP value indicates superior overall detection efficacy. The formula for mAP is:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{7}$$

where AP_i denotes the average precision of class i, and N represents the total number of object classes.

F1-Score: The F1-score serves as a complete assessment measure that combines both Precision and Recall into a single score, thus demonstrating the balance between the model's capacity to identify important occurrences and its inclination to avoid inaccurate detections. This metric is especially informative in scenarios involving class imbalance, where high precision or recall alone may not adequately represent detection performance. The F1-Score is formulated as the equilibrium value of (P) and (R), as follows:

$$F_1 = 2 \times \frac{PR}{P + R} \tag{8}$$

A higher F1-score implies a more favorable conflict between false positives and false negatives, indicating a more accurate and effective detection method.

3. Results

3.1. Object Detection Experiment Results and Analysis for Each Category in SSS Images

3.1.1. SOD Dataset

The raw SSS imagery used in this investigation is primarily sourced from collaborating laboratory units and internet platforms.

The photographs include four categories: human body, ship, plane, and small mine, as shown in Figure 12. Various circumstances categorize 638 unprocessed SSS imagery. To strengthen the detection network and mitigate the impact of seabed background on object feature extraction, the unprocessed SOD dataset includes 130 background imagery devoid of objects, which function as negative samples.

These four categories vary in scale. The first two categories (ship and plane) have larger sizes in SSS images, but there is some background interference due to sedimentation. The latter two categories (human bodies and small objects) are smaller, representing typical minor object detection problems. We improved YOLOv11 combined with the algorithm proposed in the methodology, which was first compared with the original YOLOv11 detection performance. The detection results for each category are shown in Figure 13.

Since this involves a deep comparison between YOLOv11 and our method, we examined both the precision-recall curves and confusion matrices for the two models. By clarifying class-wise separability, the confusion matrix emphasizes how efficiently each model separates the target from the background, including its handling of false positives and false negatives. Unlike the ROC curve, which evaluates the model's performance globally, the precision-recall curve focuses on the model's proficiency to classify instances of the target class correctly (true positives) and maintain a low level of false positives and false negatives at different classification thresholds.

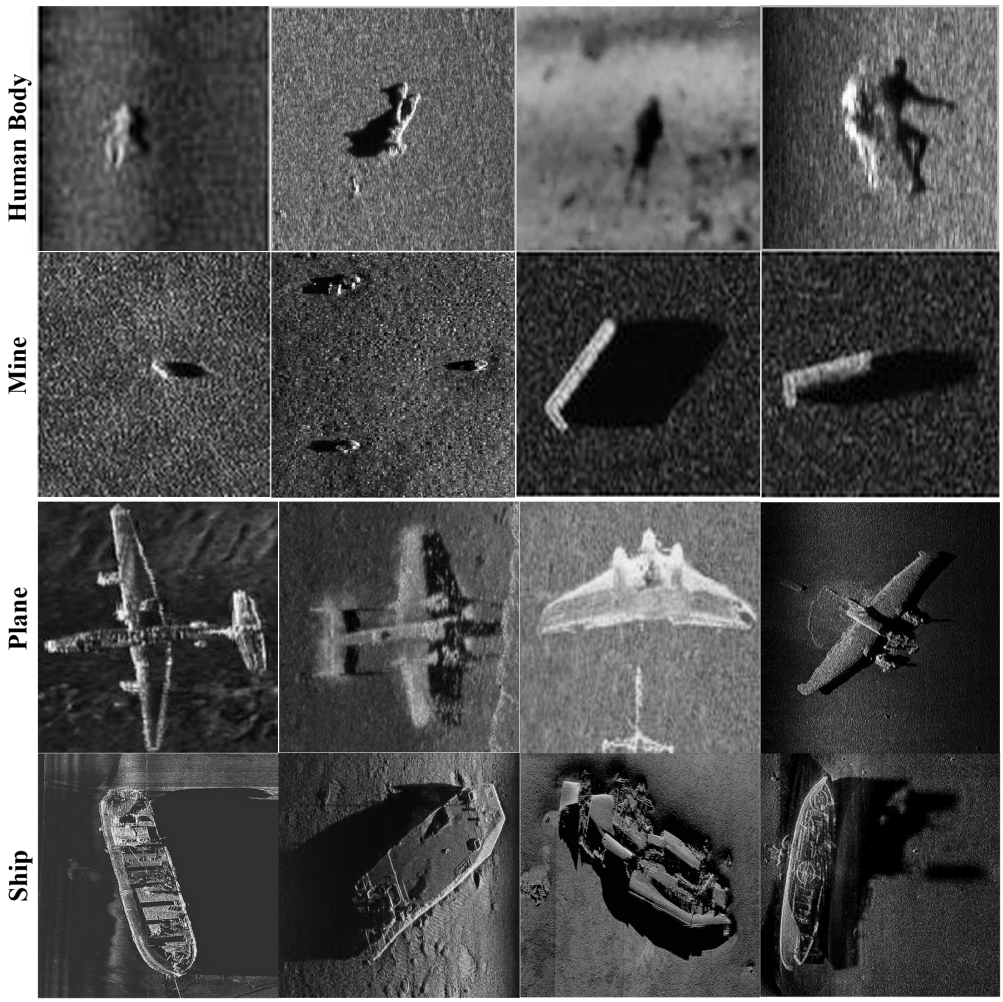


Figure 12. Distribution of images in the original SOD dataset across four object categories.

In Figure 14, the research results suggest that our approach performs better than YOLOv11 in several key aspects. Not only does our approach achieve higher overall accuracy, but it also exhibits a more balanced performance across both the target and background classes. This indicates that our model is more adept at distinguishing between the relevant objects and the irrelevant background, leading to fewer missed detections. In contrast, YOLOv11, while strong, shows a less favorable balance between correctly identifying the target objects and avoiding false detections in the background, particularly in cases with challenging class separations.

Figure 13 shows the original SSS imagery, the original YOLOv11 algorithm's detection results, and the enhanced algorithm's detection results for four categories. As shown in Figure 13b, the original YOLOv11 detection method is 70% for recognizing human objects, and the original SSS picture includes much noise. In Figure 13b, for small mine objects that are pretty similar to the background, the original YOLOv11 algorithm detections have a high detection accuracy of 85%. Similarly, the original YOLOv11 algorithm displays some missed detections for sunken ships and aircraft categories. In summary, the original YOLOv11 method, when used to identify objects in SSS imagery results, is not perfect for determining the locations of objects based on their attributes appropriately. It does not correctly detect the objects in the SSS picture, resulting in missed detections and poor performance.

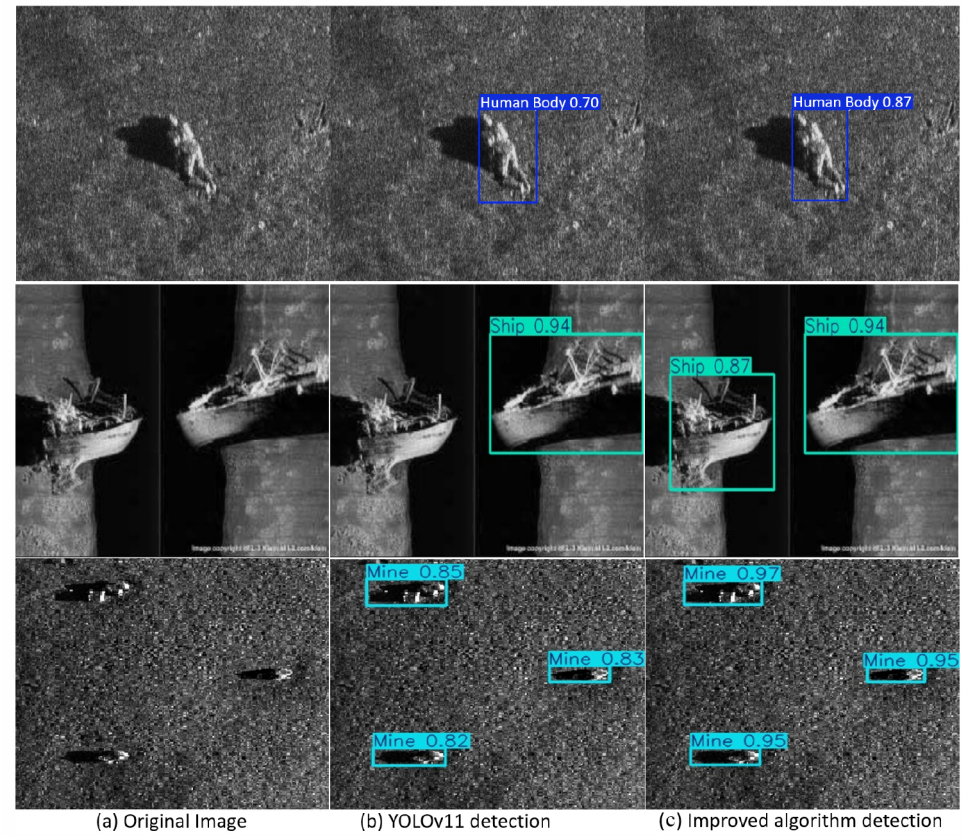


Figure 13. Comparison of experimental results on SOD images.

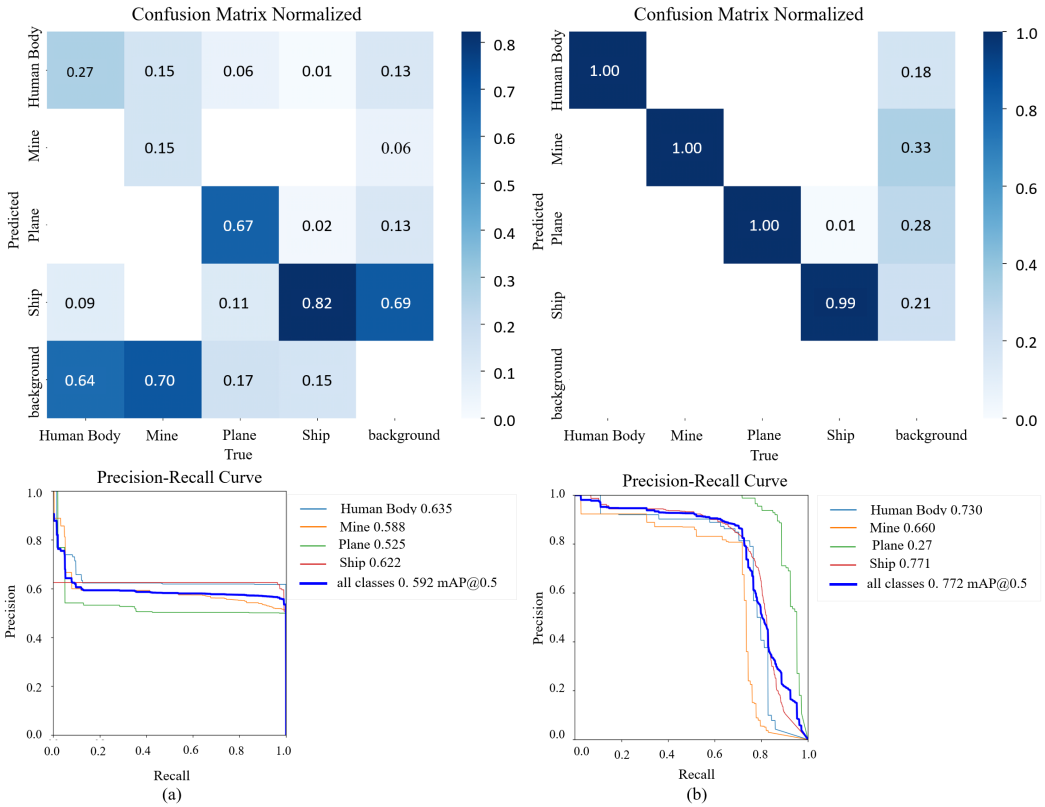


Figure 14. (**a**) Original YOLOv11 and (**b**) our Proposed YOLOv11 confusion matrix normalized and precision-recall from the SOD dataset.

However, as this research explains, the enhanced YOLOv11 algorithm significantly raises the overall detection accuracy. The enhanced algorithm combines object feature information in the SSS images with corresponding shadow features.

As shown in Figure 13c, the improved algorithm effectively suppresses background interference, enabling more accurate and reliable object detection across all four categories. Figure 15 further separately lists the detection results of the enhanced YOLOv11 framework for each object category, demonstrating that the improved detection algorithm performs much better across all categories and demonstrates good detection performance for objects of various scales in SSS images. This makes the improved model highly suitable for SSS object detection tasks for objects easily missed in images containing small objects and sunken ships.

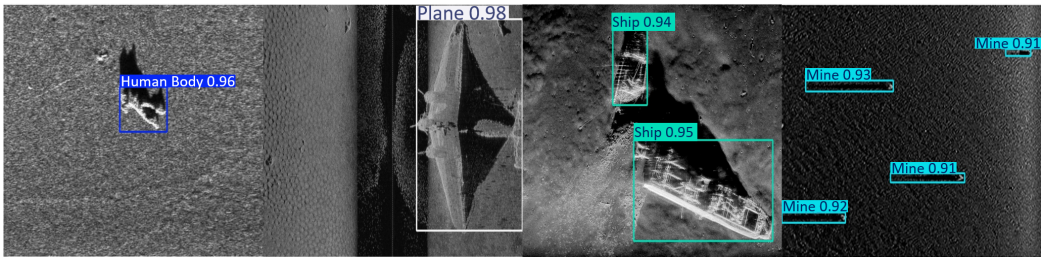


Figure 15. Example of the detection outcomes of different categories of objects in SOD images using the improved algorithm.

3.1.2. Underwater Acoustic Target Detection (UATD) Dataset

The open-source Underwater Acoustic Target Detection (UATD) dataset is used in the research, which contains sonar images processed through the Tritech Gemini 1200iK sonar (Tritech International Ltd., Aberdeenshire, UK) a popular device for MFLS data collection. The dataset, which features aerial coverage of lakes and shallow-water areas, comprises over 9000 images in raw sonar format. It provides a wealth of material for study, including photos of ten different target objects, such as ball, circle cage, cube, cylinder, human body, metal bucket, plane, rov, square cage, and tyre. All these elements provide a rich and diverse dataset that can be utilized for various kinds of scientific investigations. Figure 16 shows a sample of sonar images from the UATD dataset, which shows that the sonar imagery achieves a comparable level of complexity. Additionally, Figure 17b,c highlight that YOLOv11 and our proposed method lead to an improvement in the results.

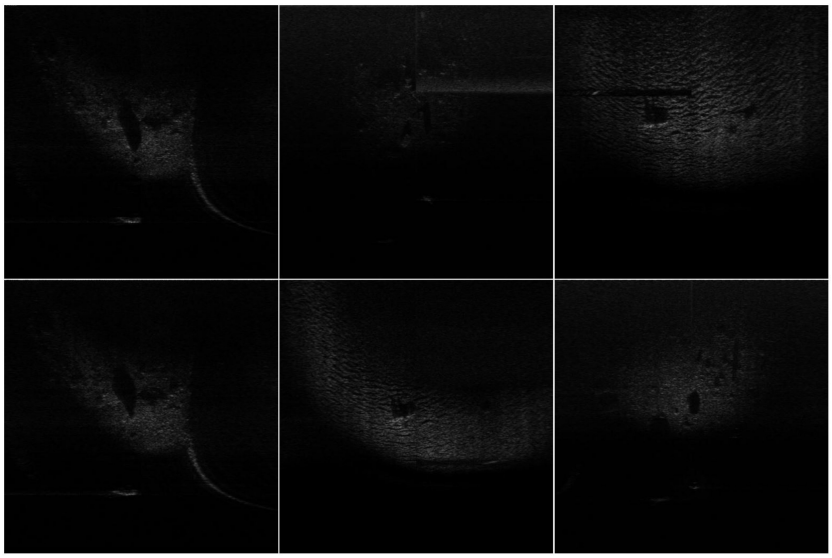
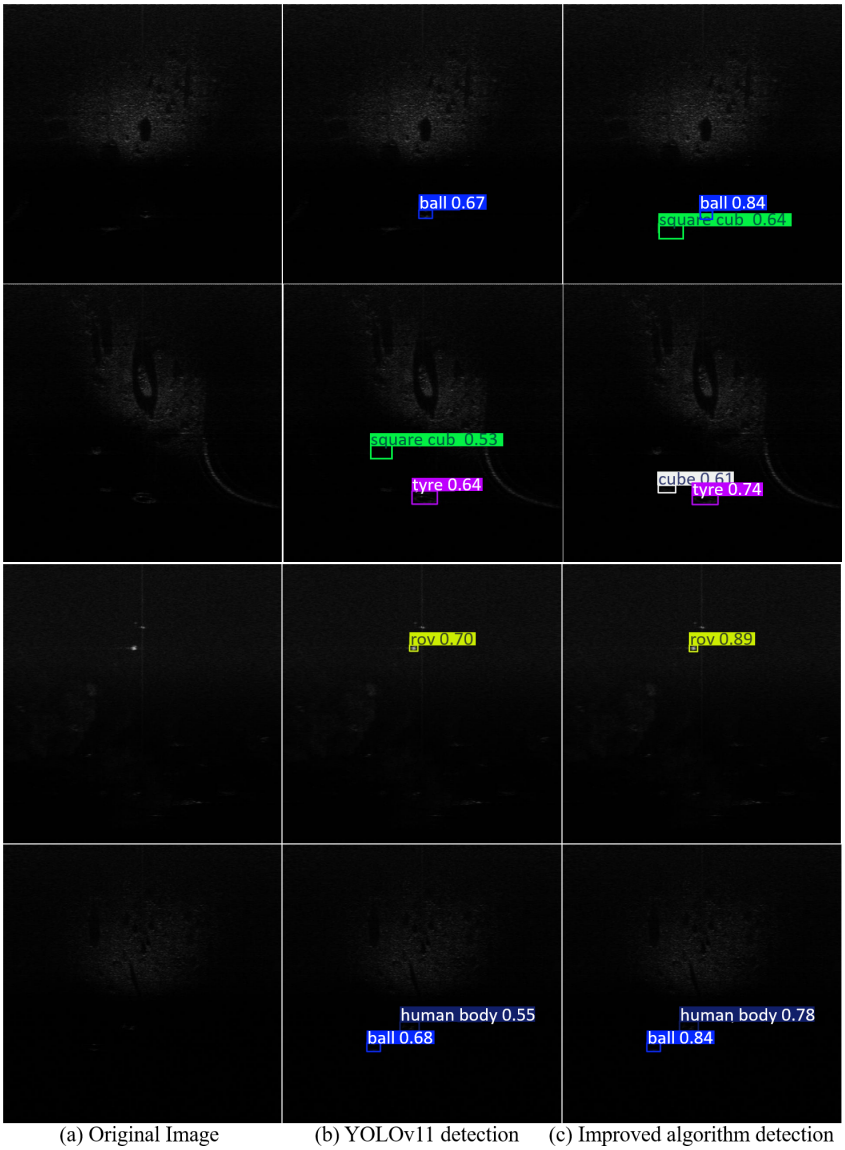


Figure 16. Sample images from the UATD dataset collection.



(a) Original Image (b) YOLOv11 detection (c) Improved algorithm detection

Figure 17. Comparison of experimental results of UATD images.

Precision of 87.1%, recall rate of 91%, mAP@0.5 of 83.6%, mAP@0.5:0.95 of 37.9% and F1-score of 84.0% were achieved by the baseline YOLOv11 model. However, the YOLOv11 module was significantly enhanced, resulting in a significant improvement in all metrics. mAP@0.5 increased to 85.1%, recall improved to 93%, and mAP@0.5:0.95 reached 38.1%. The F1-score remained strong at 84.0%, consistent with the baseline, while precision improved to 94.0%. These results suggest that the improved YOLOv11 module significantly boosts both feature extraction and detection performance, leading to better target detection accuracy across various confidence thresholds.

To thoroughly compare YOLOv11 with our method, we analyzed the precision-recall (PR) curves and confusion matrices of the models. As shown in Figure 18, our method outperforms YOLOv11, showcasing improved accuracy and a more balanced ability to recognize both target and background classes.

Experiments were conducted using the YOLOv11 and the proposed enhanced model to evaluate object detection performance on SSS and UATD images. Detection outcomes of the original YOLOv11 and the improved model are shown in Figures 19 and 20, with quantitative comparisons provided in Table 2.

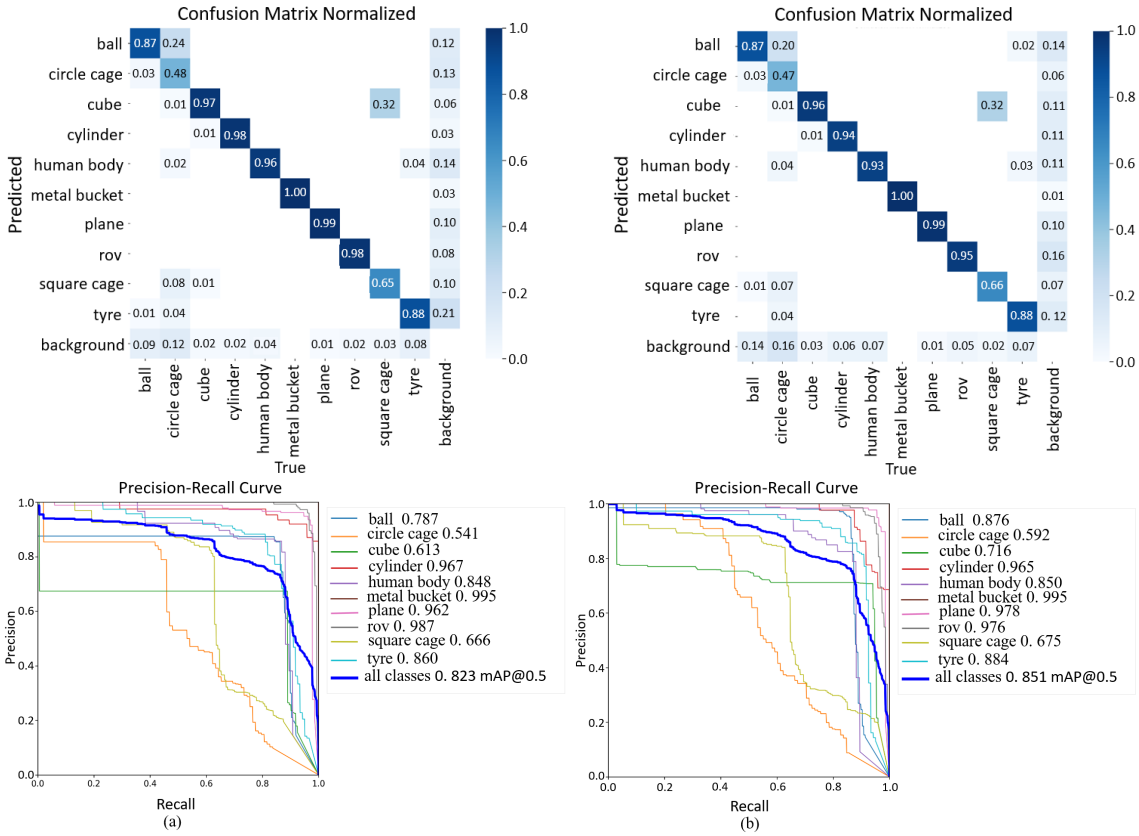


Figure 18. (**a**) Original YOLOv11 and (**b**) our proposed YOLOv11 confusion matrix normalized and precision-recall on the UATD dataset.

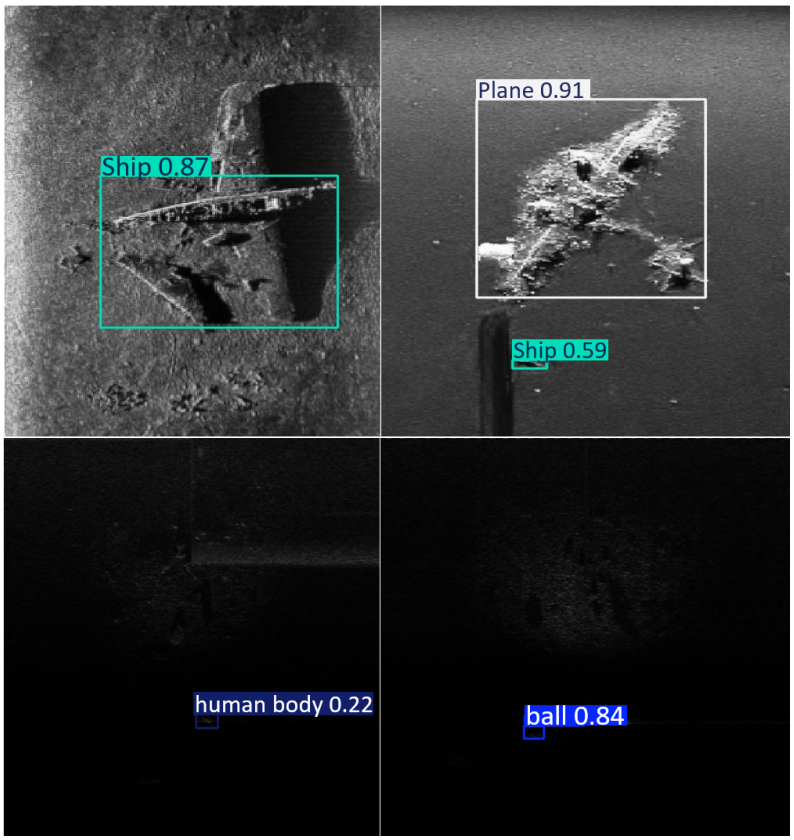


Figure 19. Incorrect detections by the original YOLOv11 model on SOD and UATD.

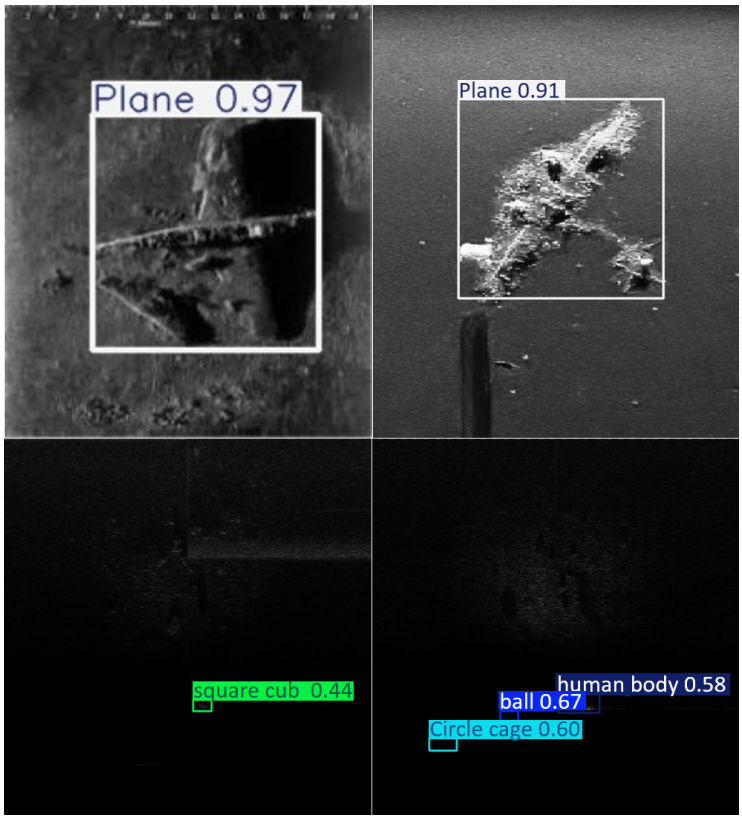


Figure 20. Comparison of detection effects of improved algorithms.

Table 2. Performance Comparison of Original and Improved YOLOv11 on SOD and UATD Datasets.

Metric	SOD		UATD	
	Original YOLOv11 (%)	**Improved (%)**	**Original YOLOv11 (%)**	**Improved (%)**
Precision	96.6	99.0	87.1	94.0
Recall	69.0	87.0	91.0	93.0
mAP@50	59.7	77.4	83.6	85.1
mAP@50:95	56.8	59.3	37.9	38.1
F1	80.5	92.6	84.0	84.0
Params (M)	2.18	28.04	2.58	28.04
FPS	798.0	422.40	974.4	417.6

As illustrated in Figure 19, the original YOLOv11 model exhibits significant detection failures on sonar imagery, including both false positives and missed detection. The model struggles to accurately localize small objects and frequently confuses background clutter with target objects, resulting in unreliable detection performance across all four categories. However, as shown in Figure 20, the improved detection algorithm successfully addresses the above false and missing object detection issues. It accurately detects object categories and effectively resists background noise interference during object detection.

From the comparative analysis of various metrics in Table 2, it is clear that the original YOLOv11 method achieves an accuracy of only 96.6% and 87.1% on the test set, failing to perform high-precision object detection tasks for SSS and UATD images. However, the YOLOv11 algorithm model was enhanced by using better-integrated features across the network's high and low layers, more successfully incorporating the impact of the shadow area characteristics on object location determination.

Finally, the proposed method, dual pooling, achieves a test set detection accuracy of 99.0% and 94.0%, demonstrating competitive performance compared with existing methods for SSS and UATD. This fully demonstrates boosting the network's generalization capacity and highlights the outstanding performance of the improved algorithm provided in this chapter regarding accuracy and algorithm adaptability.

Moreover, introducing the average recall rate for each category further evaluates each algorithm's ability to resolve missed and false detections. The comparative results presented in Table 2 demonstrate that, relative to the original algorithm, the detection algorithm based on dual pooling proposed achieves the highest average recall rate across all categories, reaching 87% and 93%. This indicates that for the vast majority of SSS and UATD imagery in the dataset, the improved algorithm proposed in this chapter can accurately detect objects in the images, effectively avoiding missed and false detections caused by background interference.

3.2. Comparative Experiment

The results of the comparative experiments on the SOD dataset are presented in Table 3. Among the general-purpose YOLO models, YOLOv8 and YOLOv11 achieve the highest precision values, reaching 96.0% and 96.6%, respectively. YOLOv10 provides the highest inference speed at 1143.6 FPS, while maintaining competitive detection performance and computational efficiency. In addition to the mainstream YOLO models, several recent sonar-oriented detectors, including YOLO-CSA, RCDI-YOLO, YOLOv11-SDC, and ESL-YOLO, are included to provide more relevant comparisons for underwater sonar object detection.

Table 3. Comparison of Results from different Experiments on SOD Dataset.

Models	F1 (%)	P (%)	mAP@50 (%)	mAP@50:95 (%)	Params (M)	FLOPS (G)	FPS
YOLOv5	71.0	90.0	59.7	56.80	**2.18**	**5.80**	973.8
YOLOv6	72.0	89.0	59.9	57.60	4.23	11.6	940.4
YOLOv7	71.0	91.0	60.8	58.15	6.20	**5.80**	860.60
YOLOv8	72.0	96.0	60.4	57.40	2.68	6.80	906.20
YOLOv9	71.0	94.0	59.8	58.18	16.71	60.9	244.40
YOLOv10	71.0	89.0	61.1	59.12	2.67	8.20	**1143.6**
YOLOv11 (Baseline)	80.5	96.6	59.7	56.80	2.58	6.50	798.0
YOLOv12	80.9	95.8	62.7	52.40	2.57	6.48	234.8
YOLO-CSA [36]	79.0	83.7	75.1	55.0	2.63	7.16	287.66
RCDI-YOLO [39]	75.2	82.1	72.0	51.9	2.92	7.73	189.32
YOLOv11-SDC [37]	69.2	70.2	66.8	46.3	2.22	6.50	611.5
ESL-YOLO [38]	72.9	72.9	69.7	49.5	3.86	10.14	260.97
Improved (our)	**92.6**	**99.0**	**77.4**	**59.30**	28.04	72.50	422.40

The proposed improved model achieves the best overall detection performance among all compared methods. It obtains an F1-score of 92.6%, a precision of 99.0%, an mAP@50 of 77.4%, and an mAP@50:95 of 59.30%. These results demonstrate that the proposed sonar-oriented integration improves both detection accuracy and the balance between precision and recall. Compared with the YOLOv11 baseline, the proposed model improves the F1-score from 80.5% to 92.6% and the mAP@50 from 59.7% to 77.4%.

This performance improvement comes at the cost of increased computational complexity. The proposed model requires 28.04 million parameters and 72.50 GFLOPs, compared with 2.58 million parameters and 6.50 GFLOPs for the YOLOv11 baseline. Its inference speed is 422.40 FPS, which is lower than that of the lightweight YOLO models but remains sufficient for many accuracy-oriented sonar detection applications. Therefore, the pro-

posed model is more suitable for scenarios in which detection reliability and accuracy are prioritized over minimum computational cost.

3.3. Comparison of Feature-Extraction and Attention Modules

To further evaluate the effectiveness of the proposed C3k2_RVB module, additional experiments were conducted using alternative feature-extraction and attention mechanisms. Each alternative module was inserted into the same C3k2 position and trained under identical conditions on the SOD dataset. The same dataset split, input resolution, number of epochs, optimizer, augmentation strategy, and evaluation metrics were used for all configurations.

As shown in Table 4, C3k2_RVB clearly outperforms both the original C3k2 baseline and the attention Transformer-augmented variants, delivering the highest F1-score (90.59%), Recall (86.0%), and mAP@50 (75.4%) of all the modules tested. This is a large jump over the baseline C3k2, which reaches only 80.50% F1 and 69.0% Recall, and it also outpaces the SE, CBAM, and Lightweight Transformer variants by a wide margin. Interestingly, although SE, CBAM, and the Lightweight Transformer all push mAP@50 somewhat higher than the baseline, this comes with a noticeable drop in F1-score and Precision, and CBAM in particular suffers a sharp fall in mAP@50:95, down to 43.7%. C3k2_RVB avoids this trade-off: its mAP@50:95 (56.20%) stays close to the original C3k2 (56.80%) even as Recall and mAP@50 rise substantially. This suggests that combining depthwise convolution, channel-wise feature recalibration, and structural reparameterization gives the model a more reliable way to represent small, low-contrast sonar targets against cluttered backgrounds, rather than simply trading one metric for another. Efficiency-wise, C3k2_RVB is also the strongest option overall: it adds only a small parameter increase over the baseline (2.51M vs. 2.18M), leaves FLOPs unchanged at 6.50G, and keeps inference speed nearly identical (796.8 vs. 798.0 FPS). The SE, CBAM, and Transformer variants, by comparison, either cost more computation, run slower, or both (324.8–847.5 FPS across 1.85–4.07M parameters and 4.1–11.4 GFLOPs), making C3k2_RVB the best-balanced choice between accuracy and computational cost.

Table 4. Comparison of C3k2_RVB with Alternative Feature-Extraction and Attention Modules on the SOD Dataset.

Module	F1 (%)	P (%)	R (%)	mAP@50 (%)	mAP@50:95 (%)	Params (M)	FLOPs (G)	FPS
Original C3k2	80.50	**96.6**	69.0	59.7	**56.80**	**2.18**	6.50	798.0
+SE	71.4	79.4	64.9	65.8	52.0	4.07	11.40	324.8
+CBAM	71.1	75.9	66.9	67.5	43.7	1.85	**4.10**	**847.5**
+Lightweight Transformer	75.0	76.5	73.4	72.2	48.9	3.38	9.56	821.4
C3k2_RVB	**90.59**	95.7	**86.0**	**75.4**	56.20	2.51	6.50	796.8

3.4. Ablation Experiment

Table 5 presents an ablation study of the principal architectural components on the SOD dataset. The evaluated components are ResNet50, C3k2_RVB, and DyDetect.

The baseline YOLOv11 obtains an F1-score of 80.50%, with a precision of 96.6%, recall of 69.0%, mAP@50 of 59.7%, and mAP@50:95 of 56.80%. The relatively low recall indicates that the baseline misses a considerable number of targets, particularly small or low-contrast objects in cluttered sonar backgrounds.

When ResNet50 is used as the only modified component, the F1-score increases to 91.67%, with a precision of 97.0% and recall of 86.9%. The deeper residual hierarchy improves multilevel feature extraction; however, the mAP@50 is 60.3%, indicating that stronger feature extraction alone does not fully resolve localization errors. This configura-

tion also increases the computational cost to 73.6 GFLOPS and reduces the inference speed to 148.0 FPS.

The C3k2_RVB-only configuration achieves an F1-score of 90.59%, a precision of 95.7%, a recall of 86.0%, an mAP@50 of 75.4%, and an mAP@50:95 of 56.20%. These results indicate that the RepViT-enhanced feature-extraction module improves the representation of small targets and helps distinguish target responses from seabed clutter. It also maintains low computational complexity, with 2.51M parameters, 6.50 GFLOPS, and 796.8 FPS.

The DyDetect-only configuration obtains an F1-score of 91.39%, a precision of 97.5%, a recall of 86.0%, an mAP@50 of 75.5%, and an mAP@50:95 of 56.70%. This result supports the role of scale, spatial, and channel-aware feature aggregation in handling sonar targets with different sizes and locations. Its computational cost is 5.40 GFLOPS, with an inference speed of 523.2 FPS.

The combination of ResNet50 and C3k2_RVB achieves an F1-score of 90.52%, an mAP@50 of 77.0%, and an mAP@50:95 of 56.72%. Although this configuration improves mAP@50, its computational cost increases to 75.3 GFLOPS, showing the trade-off between deeper feature extraction and computational efficiency.

Finally, the complete model containing ResNet50, C3k2_RVB, and DyDetect achieves the highest F1-score of 92.61%, precision of 99.0%, mAP@50 of 77.4%, and mAP@50:95 of 59.30%. These results indicate that the three components provide complementary functions: ResNet50 strengthens hierarchical feature extraction, C3k2_RVB refines small-target and clutter-related features, and DyDetect performs adaptive multiscale detection. The complete model therefore provides the best overall detection and localization performance, although its increased computational cost remains a limitation. The dual-pooling operation is retained as an integrated supporting feature-processing operation and is not treated as an independently ablated factor in this table.

Table 5. Comparison of Results from Ablation Experiments on SOD Dataset.

1	2	3	F1 (%)	P (%)	R (%)	mAP@50 (%)	mAP@50:95 (%)	Params (M)	FLOPS (G)	FPS
-	-	-	80.50	96.6	69.0	59.7	56.80	**2.18**	6.50	**798.0**
✓	-	-	91.67	97.0	86.90	60.3	56.90	28.23	73.60	251.8
-	✓	-	90.59	95.7	86.0	75.4	56.20	2.51	6.50	796.8
-	-	✓	91.39	97.5	86.0	75.5	56.70	2.45	**5.40**	530.4
✓	✓	-	90.52	96.8	85.0	77.0	56.72	28.49	75.30	697.6
✓	✓	✓	**92.61**	**99.0**	**87.0**	**77.40**	**59.30**	28.04	72.50	422.4

Note: '-' means the module is not used; and '✓' means the module is used. ① ResNet50, ② C3k2_RVB, ③ DyDetect.

Overall, the ablation results demonstrate that the full framework provides the most favorable accuracy and localization performance among the evaluated configurations. Compared with the baseline, it improves the F1-score from 80.50% to 92.61%, mAP@50 from 59.7% to 77.4%, and mAP@50:95 from 56.80% to 59.30%. These improvements support the complementary role of the integrated modules for challenging small-sample sonar object detection.

Table 6, the progressive introduction of different data augmentation operations generally improves detection performance on the SOD dataset. Compared with the baseline, the data augmentation strategy increases the F1-score from 62.20% to 92.61%, recall from 56.4% to 87.0%, mAP@50 from 58.6% to 77.4%, and mAP@50:95 from 33.90% to 59.30%, corresponding to improvements of 30.41, 30.6, 18.8, and 25.4 percentage points, respectively. These results indicate that our model, incorporating the combined augmentation strategy,

increases sample diversity and reduces overfitting under limited-data conditions, thereby improving the model's generalization ability for small-sample sonar object detection. Moreover, the subsequent inclusion of scaling, translation, and brightness adjustment produces substantial overall improvements.

Table 6. Ablation Study of Different Data Augmentation Strategies on the SOD Dataset.

Setting	F1 (%)	P (%)	R (%)	mAP@50 (%)	mAP@50:95 (%)
Baseline	62.20	69.5	56.4	58.6	33.90
Baseline + Data Augmentation	63.70	76.0	54.8	62.6	38.50
Improved	80.50	96.6	69.0	59.7	56.80
Improved + Data Augmentation	**92.61**	**99.0**	**87.0**	**77.4**	**59.30**

3.5. Heatmap Visualization

To further investigate the interpretability of the proposed model, EigenGradCAM was employed to visualize and compare the feature responses generated by the baseline YOLOv11 and the proposed model. Figure 21 presents representative sonar images from the Human, Mine, Plane, and Ship categories, together with their corresponding activation maps.

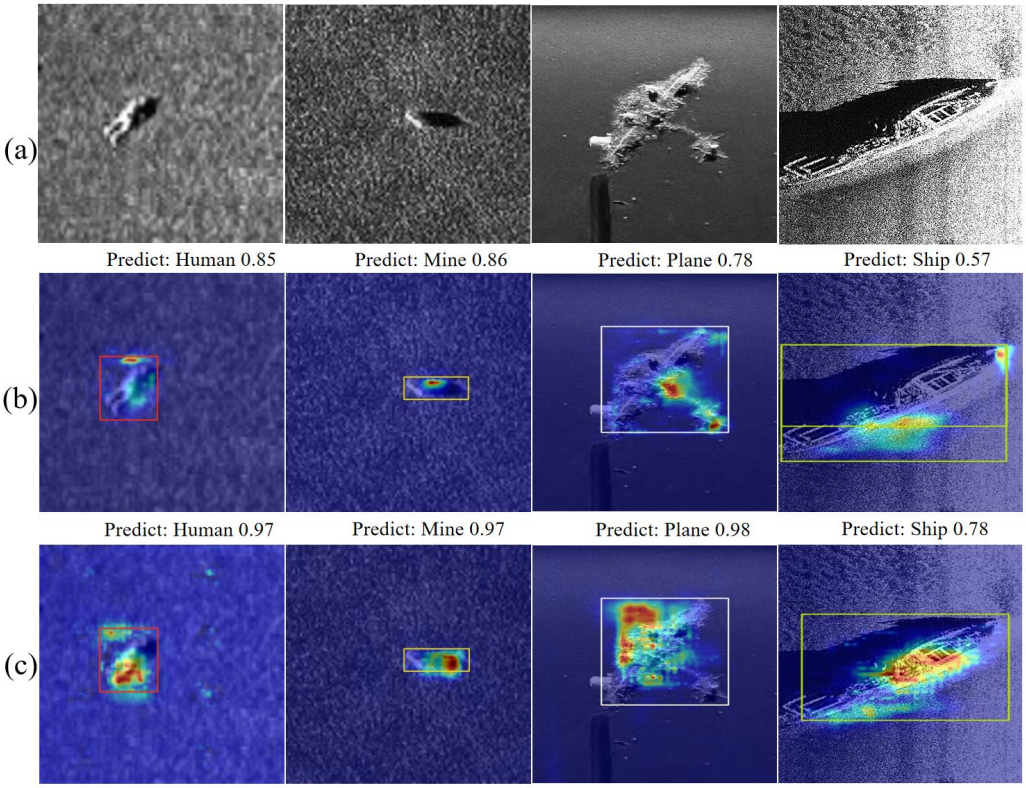


Figure 21. EigenGradCAM visualization for representative sonar targets. (**a**) Original sonar images, (**b**) activation maps generated by the baseline YOLOv11 model, and (**c**) activation maps generated by the proposed model.

As shown in Figure 21b, the baseline YOLOv11 produces relatively weak and spatially scattered responses in several challenging cases. Some activation regions extend toward irrelevant seabed textures and do not consistently cover the complete target region. This behavior is particularly evident for the small Human and Mine targets, for which the baseline model generates incomplete responses and shows limited alignment with the annotated bounding boxes.

In contrast, the proposed model, shown in Figure 21c, produces stronger, more concentrated, and more spatially coherent activations. Its responses cover the target regions more completely and are better aligned with the annotated bounding boxes. For the Ship and Plane examples, the proposed model also responds to relevant contextual regions surrounding the targets, whereas the baseline model exhibits stronger activation on isolated background areas or unrelated high-intensity pixels.

These observations suggest that the proposed feature-extraction and feature-fusion modules improve the representation of target regions and their surrounding context. The proposed model suppresses responses to irrelevant seabed textures and provides more coherent localization cues for small, low-contrast, and visually ambiguous sonar targets.

Figure 22 presents a qualitative comparison of EigenGradCAM visualizations generated by the baseline YOLOv11 model and the proposed improved model. For each example, the left image represents the baseline model and the right image represents the proposed model.

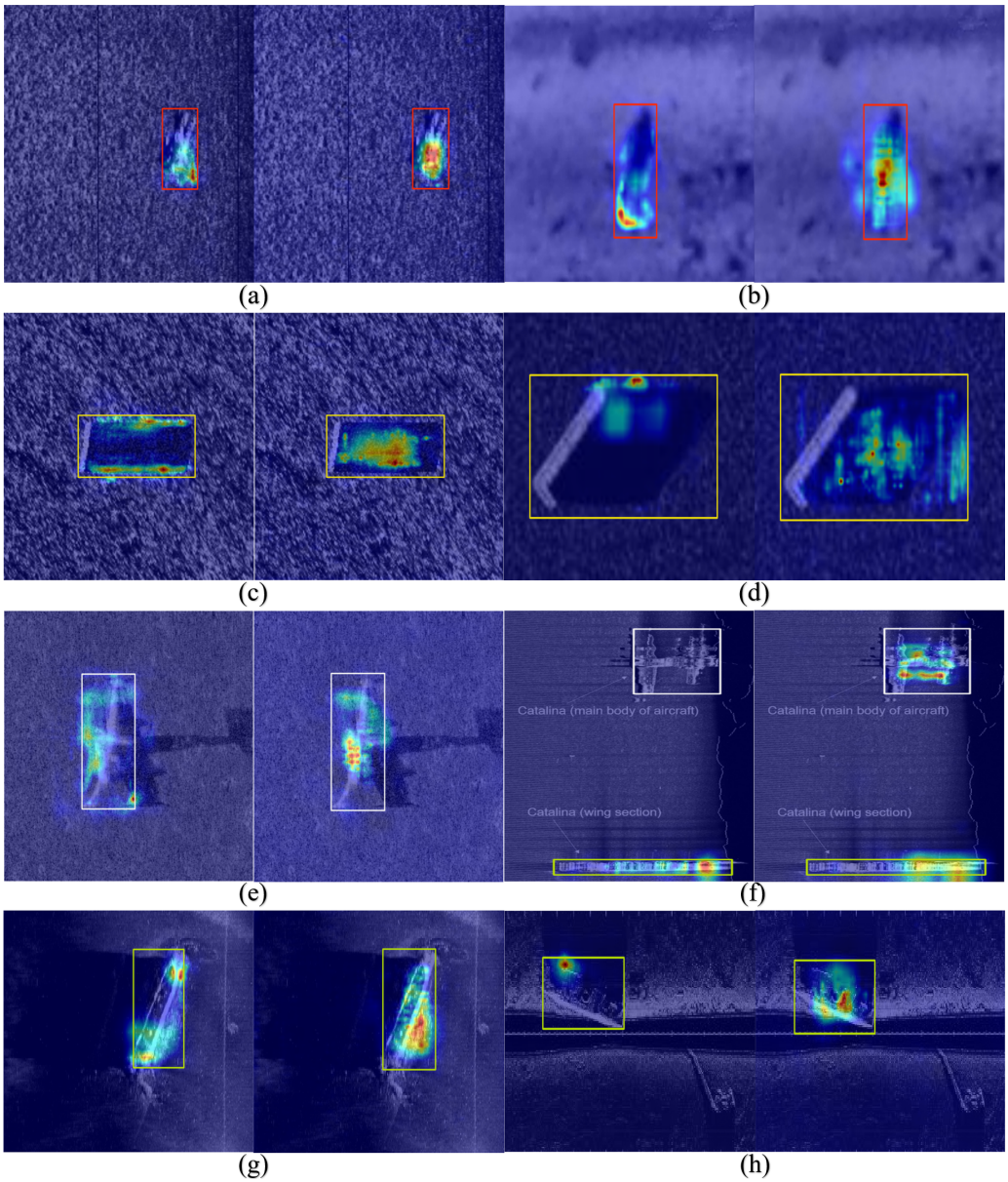


Figure 22. Qualitative comparison of EigenGradCAM visualizations. For each pair, the left image shows the baseline YOLOv11 model and the right image shows the proposed improved model. Warmer colors indicate stronger contributions to the prediction. (**a**,**b**) Human targets; (**c**,**d**) Mine targets; (**e**,**f**) Plane targets; (**g**,**h**) Ship targets.

The comparison in Figure 22a shows that the right image produces deeper and more concentrated activation within the human target, whereas the baseline response is weaker and incomplete. In Figure 22b, the enhanced feature-extraction and sonar-oriented feature-processing components highlight both the target and its nearby acoustic shadow more clearly, while the baseline model activates only a limited portion of the target. Figure 22c,d show that the proposed feature-processing strategy helps the model focus more strongly on the small target shadow structure, although some background responses and fragmented activation remain in Figure 22d. In Figure 22e, the baseline response is weak and incomplete, with part of the activation extending beyond the object boundary; in comparison, the enhanced model produces clearer activation over the plane and its associated shadow. The qualitative comparison in Figure 22f shows that the baseline model fails to highlight one of the targets completely, whereas the enhanced model produces a stronger and more concentrated response, although a small amount of activation still extends beyond the annotated region. Figure 22g illustrates that the combined feature-extraction and feature-processing improvements transform the limited and incomplete baseline response into a clearer and more concentrated activation pattern over the ship. Similarly, Figure 22h shows stronger and more coherent activation over both the ship and its associated shadow than the baseline model. Overall, these observations suggest that the improved model pays more targeted attention to small-object and acoustic-shadow features and is generally better able to suppress irrelevant seabed noise, although some background interference remains in difficult cases.

4. Discussion

The experimental results indicate that the proposed framework improves the detection of small and low-contrast sonar targets by coordinating complementary feature-processing mechanisms rather than relying on a single architectural modification. The ablation study shows that ResNet50, C3k2_RVB, and Dynamic Head provide complementary contributions. ResNet50 strengthens hierarchical feature extraction, C3k2_RVB improves local feature representation with limited additional computational cost, and Dynamic Head provides adaptive aggregation across spatial locations and feature scales. The complete configuration achieves the strongest overall performance, with an F1-score of 92.61%, precision of 99.0%, mAP@50 of 77.4%, and mAP@50:95 of 59.30%.

The data augmentation experiments further demonstrate the importance of increasing sample diversity under limited-data conditions. The combination of augmentation operations improves F1-score, recall, and mAP compared with the non-augmented baseline. This result indicates that augmentation can partially compensate for the limited number of labeled sonar images and reduce overfitting. However, the gains should be interpreted together with the dataset size and composition, because augmentation cannot fully replace a larger and more diverse collection of independently acquired sonar observations.

The proposed framework also introduces a clear accuracy complexity trade-off. The complete model contains 28.04 million parameters and requires 72.50 GFLOPs, compared with 2.58 million parameters and 6.50 GFLOPs for the YOLOv11 baseline. Its inference speed is 422.4 FPS, which remains high in the reported experimental setting but is substantially lower than the lightweight baseline. Therefore, the proposed framework is more appropriate for applications in which detection reliability is prioritized over minimum computational cost.

The improvement is particularly relevant to sonar imagery because target highlights can be weak, small, or partially confused with seabed clutter. The EigenGradCAM visualizations provide qualitative evidence that the proposed model produces more concentrated and spatially coherent responses over target regions and relevant surrounding context.

These observations are consistent with the quantitative improvements in recall and F1-score, suggesting that the enhanced feature representation helps reduce missed detections of small and low-contrast targets.

Several limitations should also be considered. First, the SOD dataset contains a limited number of labeled sonar images, and the reported results may not represent all operating conditions, sonar systems, or seabed environments. Second, the framework uses sonar-specific visual and contextual characteristics but does not explicitly model acoustic propagation, scattering, or shadow formation. Third, the increased computational cost may limit deployment on resource-constrained underwater platforms. Future work should therefore evaluate the framework on larger and more diverse datasets and investigate lightweight and physics-informed approaches that incorporate available sonar acquisition metadata.

5. Conclusions

This study addresses underwater sonar object detection under challenging conditions, including low image quality, acoustic noise, seabed clutter, weak target boundaries, small target sizes, scale variation, and limited labeled data. The contribution of this work is a sonar-oriented integration framework based on established network components, rather than the introduction of fundamentally new individual modules. The proposed framework coordinates hierarchical feature extraction, local-contextual feature processing, scale-adaptive detection, and data augmentation according to the characteristics of sonar imagery.

Experimental results on the SOD and UATD datasets demonstrate the effectiveness of the integrated framework. On the SOD dataset, the proposed model achieves an F1-score of 92.61%, an mAP@0.5 of 77.4%, and an mAP@0.5:0.95 of 59.3%. On the UATD dataset, it achieves an F1-score of 84.0%, an mAP@0.5 of 85.1%, and an mAP@0.5:0.95 of 38.1%. The ablation results further show that the principal modules provide complementary benefits for feature extraction and multiscale detection.

Author Contributions: Conceptualization, M.N.I. and G.H.; methodology, M.N.I.; software, M.N.I.; validation, M.N.I. and Z.C.; formal analysis, M.N.I.; investigation, M.N.I.; resources, G.H.; data curation, M.N.I. and Z.C.; writing original draft preparation, M.N.I.; writing review and editing, M.N.I., Z.C. and G.H.; visualization, M.N.I.; supervision, G.H.; project administration, G.H.; funding acquisition, G.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, Grant Nos. U2441254 and 62571179.

Data Availability Statement: Publicly available datasets were analyzed in this study. The SeabedObjects-KLSG sonar object detection (SOD) dataset is available at https://github.com/nurislam690206/SonarObjectDetection (accessed on 28 September 2026). The Underwater Acoustic Target Detection (UATD) dataset is available at https://universe.roboflow.com/itch/underwater-acoustic-target-detection-uatd/dataset/1 (accessed on 28 September 2026).

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

References

1. Dura, E.; Zhang, Y.; Liao, X.; Dobeck, G.; Carin, L. Active learning for detection of mine-like objects in side-scan sonar imagery. *IEEE J. Ocean. Eng.* **2005**, *30*, 360–371. [CrossRef]
2. Qin, X.; Luo, X.; Wu, Z.; Shang, J. Optimizing the Sediment Classification of Small Side-Scan Sonar Images Based on Deep Learning. *IEEE Access* **2021**, *9*, 29416–29428. [CrossRef]
3. Abu, A.; Diamant, R. A Statistically-Based Method for the Detection of Underwater Objects in Sonar Imagery. *IEEE Sens. J.* **2019**, *19*, 6858–6871. [CrossRef]

4. Kim, H.G.; Seo, J.; Kim, S.M. Underwater Optical-Sonar Image Fusion Systems. *Sensors* **2022**, *22*, 8445. [CrossRef] [PubMed]
5. Ohta, Y.; Yoshida, H.; Ishibashi, S.; Sugesawa, M.; Fan, F.H.; Tanaka, K. Seabed resource exploration performed by AUV "Yumeiruka". In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 September 2016; pp. 1–4. [CrossRef]
6. Rumson, A.; Tvedt, A.B.M.; Aas, C. An efficient coring technology for deep sea mineral exploration. In Proceedings of the OCEANS 2023—Limerick, Limerick, Ireland, 5–8 June 2023; pp. 1–11. [CrossRef]
7. Specht, M. Application of UAV and USV Vehicles for Bathymetric Monitoring of Waterbodies with High Variability of the Seabed Topography. In Proceedings of the 2023 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea), Valletta, Malta, 4–6 October 2023; pp. 143–147. [CrossRef]
8. Silva, R.; Pereira, P.; Matos, A.; Pinto, A. Volumetric Gradient-Aware Methodology for the Exploration of Foreign Objects in the Seabed. In Proceedings of the OCEANS 2024, Halifax, NS, Canada, 23–26 September 2024; pp. 1–4. [CrossRef]
9. Zhu, B.; Wang, X.; Chu, Z.; Yang, Y.; Shi, J. Active Learning for Recognition of Shipwreck Target in Side-Scan Sonar Image. *Remote Sens.* **2019**, *11*, 243. [CrossRef]
10. Abu, A.; Diamant, R. Feature Set for Classification of Man-Made Underwater Objects in Optical and SAS Data. *IEEE Sens. J.* **2022**, *22*, 6027–6041. [CrossRef]
11. Russell, G.; Bell, J.; Holt, P.; Clarke, S. Sonar image interpretation and modelling. In Proceedings of the Symposium on Autonomous Underwater Vehicle Technology, Monterey, CA, USA, 4–6 June 1996; pp. 317–324. [CrossRef]
12. Ye, X.; Yang, H.; Jia, Y.; Liu, J. Geometric Correction Method of Side-scan Sonar Image. In Proceedings of the OCEANS 2019—Marseille, Marseille, France, 17–20 June 2019; pp. 1–7. [CrossRef]
13. Wang, Z.; Zhang, S.; Huang, W.; Guo, J.; Zeng, L. Sonar Image Target Detection Based on Adaptive Global Feature Enhancement Network. *IEEE Sens. J.* **2022**, *22*, 1509–1530. [CrossRef]
14. Zhang, S.; Wu, J.; Yin, T. Simulation of sonar reverberation signal considering the ocean multipath and Doppler effect. *Front. Mar. Sci.* **2023**, *10*, 1279693. [CrossRef]
15. Lyons, A.; King, J.; Brown, D. Effects of reverberation and noise on the estimation of synthetic aperture sonar multi-look coherence. In Proceedings of the 4th International Conference on Synthetic Aperture Sonar and Synthetic Aperture Radar 2018, United Kingdom, Lerici, Italy, 5–7 September 2018; Proceedings of the Institute of Acoustics; pp. 91–98. [CrossRef]
16. Qin, X.; Luo, X.; Wu, Z.; Shang, J.; Zhao, D. Deep Learning-Based High Accuracy Bottom Tracking on 1-D Side-Scan Sonar Data. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
17. Miao, Y.; Sun, S.; Li, Y.; Sun, H.; Huang, Z. An inter-layer multi-scale feature interconstruction network for small target detection in underwater sonar images. *Ocean Eng.* **2025**, *341*, 122525. [CrossRef]
18. Zhang, X.; Pan, H.; Jing, Z.; Ling, K.; Peng, P.; Song, B. UUVDNet: An efficient unmanned underwater vehicle target detection network for multibeam forward-looking sonar. *Ocean Eng.* **2025**, *315*, 119820. [CrossRef]
19. Li, X.; Zhao, Y.; Su, H. Efficient underwater object detection based on feature enhancement and attention detection head. *Sci. Rep.* **2025**, *15*, 5973. [CrossRef] [PubMed]
20. Gu, X.; Tang, S.; Cao, Y.; Yu, C. Underwater object detection in sonar imagery with detection transformer and Zero-shot neural architecture search. *arXiv* **2025**, arXiv:2505.06694.
21. Li, L.; Li, Y.; Yue, C.; Xu, G.; Wang, H.; Feng, X. Real-time underwater target detection for AUV using side scan sonar images based on deep learning. *Appl. Ocean Res.* **2023**, *138*, 103630. [CrossRef]
22. Wu, M.; Wang, Q.; Rigall, E.; Li, K.; Zhu, W.; He, B.; Yan, T. ECNet: Efficient Convolutional Networks for Side Scan Sonar Image Segmentation. *Sensors* **2019**, *19*, 2009. [CrossRef] [PubMed]
23. Yu, Y.; Zhao, J.; Gong, Q.; Huang, C.; Zheng, G.; Ma, J. Real-Time Underwater Maritime Object Detection in Side-Scan Sonar Images Based on Transformer-YOLOv5. *Remote Sens.* **2021**, *13*, 3555. [CrossRef]
24. Williams, D.P. Fast Target Detection in Synthetic Aperture Sonar Imagery: A New Algorithm and Large-Scale Performance Analysis. *IEEE J. Ocean. Eng.* **2015**, *40*, 71–92. [CrossRef]
25. Villar, S.A.; Acosta, G.G.; Solari, F.J. OS-CFAR process in 2-D for object segmentation from Sidescan Sonar data. In *Proceedings of the 2015 XVI Workshop on Information Processing and Control (RPIC)*; IEEE: New York, NY, USA, 2015; pp. 1–6. [CrossRef]
26. Zhou, T.; Wang, Y.; Zhang, L.; Chen, B.; Yu, X. Underwater Multitarget Tracking Method Based on Threshold Segmentation. *IEEE J. Ocean. Eng.* **2023**, *48*, 1255–1269. [CrossRef]
27. Mukherjee, K.; Gupta, S.; Ray, A.; Phoha, S. Symbolic Analysis of Sonar Data for Underwater Target Detection. *IEEE J. Ocean. Eng.* **2011**, *36*, 219–230. [CrossRef]
28. Tian, W.; Mu, Z.; Lu, Y.; Cai, S.; Xiong, Y.; Huang, Y. A fast level set segmentation algorithm for sonar images driven by Gaussian Mixture Model probability. *Alex. Eng. J.* **2025**, *129*, 16–28. [CrossRef]

29. Midtgaard; Hansen, R.E.; Sæbø, T.O.; Myers, V.; Dubberley, J.R.; Quidu, I. Change detection using Synthetic Aperture Sonar: Preliminary results from the Larvik trial. In Proceedings of the OCEANS'11 MTS/IEEE KONA, Kona, HI, USA, 19–22 September 2011; pp. 1–8. [CrossRef]
30. 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. [CrossRef]
31. Szymak, P.; Piskur, P.; Naus, K. The Effectiveness of Using a Pretrained Deep Learning Neural Networks for Object Classification in Underwater Video. *Remote Sens.* **2020**, *12*, 3020. [CrossRef]
32. Wang, H.; Sun, S.; Ren, P. Meta underwater camera: A smart protocol for underwater image enhancement. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 462–481. [CrossRef]
33. Wang, H.; Zhang, W.; Ren, P. Self-organized underwater image enhancement. *ISPRS J. Photogramm. Remote Sens.* **2024**, *215*, 1–14. [CrossRef]
34. Wang, H.; Zhang, W.; Xu, Y.; Li, H.; Ren, P. WaterCycleDiffusion: Visual–textual fusion empowered underwater image enhancement. *Inf. Fusion* **2026**, *127*, 103693. [CrossRef]
35. Li, X.; Sun, W.; Ji, Y.; Huang, W. A Plot-to-Track Association Framework Based on Graph Representation Learning for Compact HFSWR. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 12742–12760. [CrossRef]
36. Liu, D.; Chen, H.; Chen, J.; Ding, Y.; Xu, Z. Side-Scan Sonar Image-Based Object Detection With YOLOv8. *IET Radar Sonar Navig.* **2026**, *20*, e70209. [CrossRef]
37. Zou, C.; Yu, S.; Yu, Y.; Gu, H.; Xu, X. Side-Scan Sonar Small Objects Detection Based on Improved YOLOv11. *J. Mar. Sci. Eng.* **2025**, *13*, 162. [CrossRef]
38. Zhang, Z.; Shuai, C.; Yuan, C.; Li, B.; Ma, J.; Shang, X. ESL-YOLO: Edge-Aware Side-Scan Sonar Object Detection with Adaptive Quality Assessment. *J. Mar. Sci. Eng.* **2025**, *13*, 1477. [CrossRef]
39. Zhang, J.; Gao, B. RCDI-YOLO: A target-detection method for complex environment side-scan sonar images based on improved YOLOv8. *Front. Mar. Sci.* **2025**, *12*, 1679077. [CrossRef]
40. Huo, G.; Wu, Z.; Li, J. Underwater Object Classification in Sidescan Sonar Images Using Deep Transfer Learning and Semisynthetic Training Data. *IEEE Access* **2020**, *8*, 47407–47418. [CrossRef]
41. Plets, R.; Quinn, R.; Forsythe, W.; Westley, K.; Bell, T.; Benetti, S.; Mcgrath, F.; Robinson, R. Using Multibeam Echo-Sounder Data to Identify Shipwreck Sites: Archaeological assessment of the Joint Irish Bathymetric Survey data. *Int. J. Naut. Archaeol.* **2011**, *40*, 87–98. [CrossRef]
42. Grządziel, A. Results from Developments in the Use of a Scanning Sonar to Support Diving Operations from a Rescue Ship. *Remote Sens.* **2020**, *12*, 693. [CrossRef]
43. Cai, C.; Chen, J.; Yan, Q.; Liu, F.; Zhou, R. A prior information-based coverage path planner for underwater search and rescue using autonomous underwater vehicle (AUV) with side-scan sonar. *IET Radar Sonar Navig.* **2022**, *16*, 1225–1239. [CrossRef]
44. Zhao, L.; Ren, X.; Fu, L.; Yun, Q.; Yang, J. UWS-YOLO: Advancing Underwater Sonar Object Detection via Transfer Learning and Orthogonal-Snake Convolution Mechanisms. *J. Mar. Sci. Eng.* **2025**, *13*, 1847. [CrossRef]
45. Jocher, G.; Qiu, J. Ultralytics YOLO11. 2024. Available online: https://docs.ultralytics.com/models/yolo11 (accessed on 28 September 2026).
46. Wang, A.; Chen, H.; Lin, Z.; Han, J.; Ding, G. Rep ViT: Revisiting Mobile CNN From ViT Perspective. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 15909–15920. [CrossRef]
47. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic Head: Unifying Object Detection Heads with Attentions. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7369–7378. [CrossRef]
48. More, Y.; Dumbre, K.; Shiragapur, B. Horizontal Max Pooling a Novel Approach for Noise Reduction in Max Pooling for Better Feature Detect. In Proceedings of the 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Vienna, Austria, 28–29 October 2023; pp. 1–5. [CrossRef]
49. Nguyen, D.A.; Tran, X.T.; Dang, K.N.; Iacopi, F. A lightweight Max-Pooling method and architecture for Deep Spiking Convolutional Neural Networks. In Proceedings of the 2020 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Ha Long Bay, Vietnam, 8–10 December 2020; pp. 209–212. [CrossRef]
50. Yu, C.; Hung, P.H.; Hong, J.H.; Chiang, H.Y. Efficient Max Pooling Architecture with Zero-Padding for Convolutional Neural Networks. In Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Pune, India, 1–3 March 2023; pp. 747–748. [CrossRef]
51. Kuan, K.; Manek, G.; Lin, J.; Fang, Y.; Chandrasekhar, V. Region average pooling for context-aware object detection. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Nara, Japan, 10–13 October 2017; pp. 1347–1351. [CrossRef]

52. Jiang, P.; Zou, C. Convolution neural network with multiple pooling strategies for speech emotion recognition. In Proceedings of the 2022 6th International Symposium on Computer Science and Intelligent Control (ISCSIC), Beijing, China, 11–13 November 2022; pp. 89–92. [CrossRef]
53. Yue, Y.; Finley, T.; Radlinski, F.; Joachims, T. A support vector method for optimizing average precision. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 23–27 July 2007; Volume SIGIR '07, pp. 271–278. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.