remote sensing

MDPI

Article

A Spatial Distribution Probability-Guided Detection Framework for Underwater Sonar Imagery

Dayu Jia [1], Yan Huang [1], Jianan Qiao [1,2], Zhenyu Wang [1], Hao Feng [1] and Jiancheng Yu [1,*]

[1] State Key Laboratory of Robotics and Intelligent Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; jiadayu@sia.cn (D.J.); huangy@sia.cn (Y.H.); qiaojianan@sia.cn (J.Q.); wangzhenyu@sia.cn (Z.W.); fenghao@sia.cn (H.F.)
[2] University of Chinese Academy of Sciences, Beijing 101408, China
* Correspondence: yjc@sia.cn

Highlights

What are the main findings?

- A real sonar dataset containing practice mines and practice subsurface buoys is created. The dataset is used to train and validate a deep neural network.
- A Spatial Distribution Probability-Guided Detection Framework is designed, focusing on submarine point target detection, which achieves superior detection performance.

What are the implications of the main findings?

- Spatial distribution probability provides reliable prior knowledge which enables the deep neural network to focus on the target part.
- The proposed framework achieves robust object detection in data-scarce scenarios, demonstrating generalizability beyond underwater sonar datasets.

Abstract

Underwater target detection via side-scan sonar is vital for defense and economy but hindered by sparse targets, high data costs, and feature extraction difficulties due to textureless acoustic data and limited samples. To overcome these limitations, particularly for few-shot, small-object detection, we propose a Spatial Distribution Probability-Guided Detection Framework to aid Unmanned Underwater Vehicles (UUVs) in precise localization and clustering. The framework features a novel module that leverages a pre-trained Vision Foundation Model (DINOv3) to generate spatial distribution probability maps, guiding a Transformer-based network for accurate detection with scarce data. Additionally, it incorporates a Target Position Calculation Module and a DBSCAN-based post-processing module to determine global geographic coordinates and cluster discrete points, respectively. Experiments were conducted on both a Public Mine Detection Dataset and a self-collected dataset containing simulated mines and buoys. Ablation studies and comparison experiments demonstrated that the proposed guidance mechanism significantly improves detection performance. Furthermore, two comb-search missions verified that the system could accurately locate and cluster targets, distinguishing real targets from false detections (noise). These results confirm the framework's efficacy in enabling high-precision perception and autonomous operations for complex underwater inspection tasks.

Keywords: underwater target detection; side-scan sonar; spatial distribution probability guidance; target localization; DBSCAN clustering; Unmanned Underwater Vehicle (UUV)

Check for updates

Academic Editor: Marios Anagnostou

Received: 21 April 2026
Revised: 3 June 2026
Accepted: 6 June 2026
Published: 9 June 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

Target detection based on underwater acoustic imagery has emerged as a focal point in the field of underwater target perception. Operations involving the deployment or sabotage of seabed assets—such as mines, pipelines, and cables—have a direct impact on the international situation. For instance, during the 1990 Gulf War, multiple U.S. warships were struck by Iraqi mines. Similarly, in 2026, Iran's use of Maham mines to blockade the Strait of Hormuz precipitated an international energy crisis. In 2019, five undersea oil pipelines in Syria were suspected to have been sabotaged. In 2021, Norway's underwater surveillance network was suspected to have been compromised. In 2022, Russia's "Nord Stream" gas pipelines were suspected to have been deliberately damaged, directly affecting the energy ties between Russia and Europe. In 2024, four submarine cables in the Red Sea were deliberately damaged, impacting 25% of the telecommunications traffic between Asia and Europe. From the examples above, it can be concluded that underwater target security is closely linked to national defense, economic development, and environmental protection. Side-scan sonar is a specialized underwater acoustic imaging system which is widely used in submarine object detection tasks for its advantages of wide-area view and low cost. However, submarine target detection based on side-scan sonar suffers difficulties in data acquisition and feature extraction (the challenges are illustrated in Figure 1). Data collection is challenging due to the sparse distribution of underwater targets, the high cost of marine data acquisition, and the high cost of manual annotation. Feature extraction is difficult owing to the lack of texture information in acoustic data, for it is based on the backscattering intensity of acoustic waves, as opposed to capturing light reflections, color, and surface details like an optical camera does. Feature extraction is further complicated by the insufficiency of samples to cover class features and the imbalanced sample distribution. Therefore, investigating detection methods focusing on side-scan sonar images is a crucial research direction.

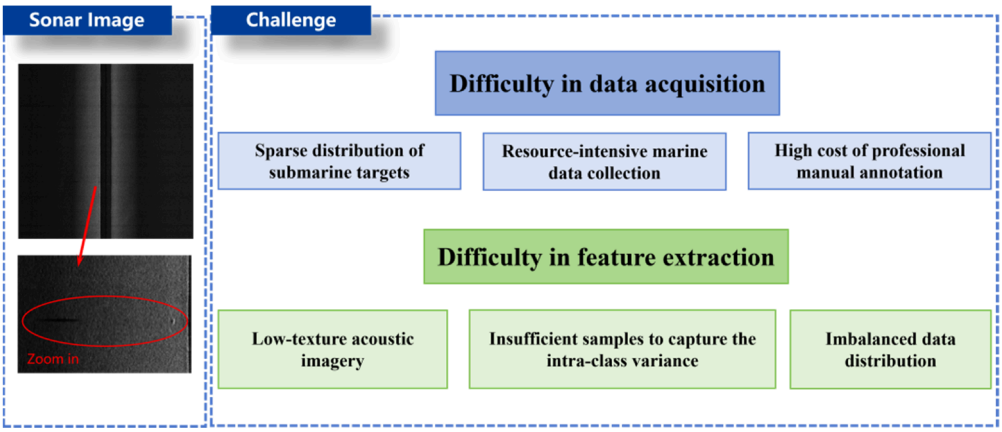


Figure 1. Challenges in submarine target detection.

With the development of Graphics Processing Units (GPUs), deep learning-based target detection technology has emerged as a research hotspot which enables the joint perception of target categories and spatial positions. Current submarine detection research primarily focuses on improvements in two major directions: First, focusing on the design of training strategies, introducing methods such as transfer learning, sonar image generation, few-shot learning, and small object learning to enhance the model's generalization ability and detection performance in complex underwater environments. Second, focusing on the optimization of network structures, improving the feature extraction backbone networks and multi-scale feature fusion modules to enhance the model's ability to express target semantic and spatial information, thereby achieving more precise detection and localization.

The transfer learning strategy utilizes simulated acoustic data or optical images for pre-training, followed by model fine-tuning to transfer learned features to real acoustic scenarios, achieving efficient target recognition. McKay et al. [1] demonstrated a simple and flexible Convolutional Neural Network feature extraction strategy, achieving impressive recognition results. Subsequently, they proposed a technical scheme utilizing robust transfer learning methods to achieve multi-instance target detection and recognition on given synthetic aperture sonar datasets. To avoid overfitting, Zhu et al. [2] used a Gaussian Mixture Model (GMM) to model the statistical features of the acoustic shadow area and extracted the shadow regions accordingly; on this basis, they constructed measured and simulated datasets. Subsequently, the simulated dataset was input into a Convolutional Neural Network (CNN) for training, retaining the feature extraction part to extract features from the measured dataset; then, the feature vectors of the measured dataset were used to reconstruct and train the classification part, realizing MLO recognition of real sonar images. Cheng et al. [3] proposed a Multi-Domain Collaborative Transfer Learning (MDCTL) method combined with a Multi-Scale Recurrent Attention Mechanism (MSRAM) to improve the accuracy of underwater sonar image classification. In the MDCTL method, the low-level feature similarity between SSS images and Synthetic Aperture Radar (SAR) images, as well as the high-level representation similarity between SSS images and optical images, were jointly utilized to enhance the feature extraction capability of deep learning models. By leveraging the different characteristics of multi-domain data to efficiently capture features useful for sonar image classification, MDCTL provides a new pathway for transfer learning. Meanwhile, MSRAM was introduced to effectively fuse multi-scale features, enabling the model to focus more on target shape details while excluding noise interference. Xu et al. [4] proposed a novel sonar image classification method (SonarNet) based on parameter transfer learning and deep learning. First, based on the VGG16 network, a new sonar image classification network—Global Average Pooling VGG (GAPVGG)—was designed. This network effectively solved the overfitting problem caused by directly using small-sample sonar images to train the original VGG16 network. Compared with the original VGG16 network, the parameter count of the GAPVGG network was reduced by nearly 90%. Subsequently, based on the parameter-based transfer learning strategy, source domain knowledge was transferred to the target domain, overcoming the problem of insufficient sonar image samples.

Few-shot learning is highly consistent with transfer learning in its underlying mechanism; its essence lies in achieving a significant leap in neural network performance under data-scarce scenarios through the reuse and adaptation of pre-trained models. Ochal et al. [5] evaluated and compared various supervised and semi-supervised few-shot learning (FSL) methods using underwater optical images and side-scan sonar images. The results showed that compared to the traditional transfer learning method of fine-tuning pre-trained models, few-shot learning methods have significant advantages. Chen et al. [6] proposed a few-shot sonar image classification method based on multi-strategy optimization fusion. First, an improved label smoothing regularization technique with category preference was employed to optimize training data labels and reduce network overconfidence; subsequently, drawing on fine-tuning strategies in transfer learning, partial parameters obtained from pre-trained models in the optical image domain were utilized to assist in improving performance in the sonar image domain; finally, the above three optimization strategies were fused to achieve few-shot learning. Xu et al. [7] designed a Deep Adaptive Sonar Image Classification Network (DASCN) based on deep learning and domain adaptation. The feature extraction module in DASCN is responsible for extracting multi-scale features from images; the attention module is used to learn the importance of different chan-

nel features; and the domain adaptation module aims to reduce the discrepancy between the source domain and the target domain.

The most direct and efficient strategy to address the few-shot challenge is data augmentation, which mainly covers two paths: sonar image simulation and generation. Of these, sonar image simulation technology relies on physical modeling mechanisms and integrates computer graphics algorithms to reproduce the sonar imaging process with high fidelity. Bell et al. [8] proposed a computer model for simulating the side-scan sonar process, covering the main deterministic aspects of the underlying physical processes leading to side-scan sonar image generation: the propagation of acoustic pulses in the water column, subsequent interaction with the rough seabed, and scattering, directly outputting synthetic side-scan sonar images. On this basis, Bell et al. [9] optimized and upgraded the simulation model from the two key dimensions of seabed texture characterization and sonar physical modeling. Cerqueira et al. [10] developed a GPU- and OpenGL-based system to simulate mechanically scanned imaging sonars and forward-looking sonars. Furthermore, Cerqueira [11] utilized rasterization techniques to calculate primary intersections, performing ray tracing only on reflective areas. Compared to methods based entirely on ray tracing, the number of rays emitted by this system is significantly reduced, thereby achieving a substantial performance boost without compromising the final rendering quality. The emergence of Generative Adversarial Networks (GANs) has opened a new path for the realistic generation of acoustic images. Li [12] and Li et al. [13] utilized a modified CycleGAN to achieve cross-modal generation from optical guidance maps to high-quality sonar images. Hu et al. [14] further narrowed the visual gap between synthetic images and real sonar images through a GAN strategy of injecting acoustic global features into guidance images, generating highly realistic acoustic images.

Due to the essential differences in imaging mechanisms and data distribution between sonar and natural images, directly applying deep neural networks designed for natural images often fails to achieve ideal performance. To this end, researchers have conducted targeted structural improvements and optimizations on traditional deep convolutional networks to enhance the model's perception ability of underwater targets and achieve more precise sonar image target detection.

Du et al. [15] preprocessed sonar images based on the region-dominant principle and threshold segmentation method; subsequently, they utilized the EfficientNet backbone network to extract preliminary effective features from the preprocessed data, achieving a balance between algorithm efficiency and power consumption. In the feature fusion stage, an improved weighted Bidirectional Feature Pyramid (BiFPN) structure was adopted to aggregate global information and enhance the feature representation capability of shallow feature maps, effectively optimizing the problem of unsatisfactory small-object feature extraction caused by low resolution and lack of detail information in sonar images. Lei et al. [16] proposed the SI-GAT model to simultaneously extract features from the target's highlight and shadow regions to obtain more comprehensive target features. Subsequently, weighting and metric functions were defined to calculate feature distances and correlation matrices, capturing the spatial relationships of targets. By introducing the K-Nearest Neighbors (KNN) algorithm and attention mechanisms, aggregation coefficients within the neighborhood were adaptively assigned, promoting feature aggregation and propagation. Finally, features of all nodes in the aggregated graph were combined to extract global features. Ruan et al. [17] proposed a Dual-Path Deep Residual "Shrinkage" Network (DP-DRSN) module for side-scan sonar image classification. This module can extract background information and feature texture information of input feature maps through different scales (e.g., global average pooling and global max pooling); subsequently, this scale information passes through two layers of 1×1 convolutions to increase non-linearity. This process

helps achieve cross-channel information interaction and fusion before outputting threshold parameters through the Sigmoid layer.

In traditional underwater target recognition research, detection usually refers to determining the presence or absence of targets in an image, mainly serving as a precursor to classification and recognition to lock onto Regions of Interest (ROI). In contrast, the deep learning-based target detection paradigm has shifted; it can directly output the spatial position and category information of targets synchronously through bounding boxes. Tang et al. [18] proposed a lightweight side-scan sonar shipwreck detection model based on DETR-YOLO. First, a multi-scale feature composite fusion module was introduced to enhance the detection capability for small objects; second, the SENet (Squeeze-and-Excitation Network) attention mechanism was integrated to strengthen the model's sensitivity to important channel features; finally, the WBF (Weighted Box Fusion) strategy was adopted to improve the localization accuracy and confidence of detection boxes. Experimental results showed that the DETR algorithm requires only a small number of anchor boxes to complete underwater target detection tasks; the model achieved AP_0.5 and AP0.5:0.95 values of 84.5% and 57.7% respectively on the test set. Steiniger et al. [19] applied YOLOv2 and YOLOv3 to side-scan sonar image target detection, verifying that improvements to the YOLO architecture developed for RGB images brought significant gains on sonar data. Yu et al. [20] proposed a novel TR-YOLOv5s network architecture and downsampling principle, introducing attention mechanisms to meet the dual requirements of accuracy and efficiency in underwater target recognition. Experiments verified that the proposed method achieved a mean Average Precision (mAP) of 85.6% and a macro F2 score of 87.8%, representing improvements of 12.5% and 10.6% respectively compared to the YOLOv5s network trained from scratch, and possessing a real-time recognition speed of approximately 0.068 s/frame. Li et al. [21] addressed the problem that the lightweight model SSD-MV3 is limited by input resolution and struggles to effectively detect tiny objects directly in large-size high-resolution synthetic aperture sonar (SAS) images, proposing a detection scheme based on a "Divide and Conquer Strategy." First, a redundant cutting algorithm was used to divide the original large image into several sub-images suitable for model input; subsequently, the lightweight model was called to perform parallel detection on each sub-image; finally, the dispersed detection results were fused and deduplicated through a maximum suppression algorithm. This method successfully achieved precise localization and recognition of small-scale objects of interest in high-resolution SAS images. Yang et al. [22] proposed a Foreground Semantic Enhancement Module, innovatively correlating semantic maps with features at various levels, thereby increasing the discrepancy between foreground and background and significantly highlighting target information. Secondly, to address the issue of insufficient high-frequency information, a Foreground Edge Enhancement Module was proposed, creatively combining Recurrent Neural Networks (RNNs) and utilizing spatial semantic information from different directions to enhance edge features, thereby improving the feature representation of foreground targets. The core modules achieved a performance improvement of up to 10% in mAP. Tang et al. [23] proposed a side-scan sonar underwater target segmentation model based on a UNet algorithm with Hybrid Dilated Convolution and Pyramid Partition Attention mechanisms (BHP-UNet). First, the model adopts a Hybrid Dilated Convolution module to expand the receptive field while improving the model's ability to learn deep semantics and shallow features. Secondly, a Pyramid Partition Attention module was introduced to handle multi-scale spatial features while establishing long-range dependencies between global and local information. This method achieved the highest Dice value and Intersection over Union (IoU) on the test set, at 78.31% and 77.71% respectively. Er et al. [24] believe that real-time performance is the primary constraint for underwater target detection, where anchor-free

detectors demonstrate significant advantages. This class of methods abandons the traditional Non-Maximum Suppression (NMS) post-processing step and eliminates the huge computational overhead brought by preset anchors; by directly regressing the category and coordinates of targets through convolutional networks, inference latency is significantly reduced, thereby achieving superior real-time detection efficiency. Wang et al. [25] trained on hundreds of representative manually annotated segmentation images, subsequently utilizing the U-Net network to perform real-time segmentation on sonar images to detect any pipe sections present in the image. Then, by skeletonizing the detected pipe segments and fitting parametric curves, the geometric shape of the pipe was estimated. Based on this geometric estimation result, the system can repeatedly issue waypoint commands to the underwater robot performing inspection tasks, enabling the AUV or ROV to perform fully automated fly-over inspections of seabed-visible pipes. This method achieved a Dice score of 0.73 and an IoU of 0.60.

However, existing methods fail to meet the requirements of few-shot small underwater object detection. In summary, this paper makes the following key contributions:

- A real sonar dataset containing practice mines and practice subsurface buoys is created. The dataset is used to train and validate the deep neural network.
- The core novel spatial distribution probability-guided network is designed, focusing on submarine point target detection. Detection results of targets, including location and confidence, are generated by the proposed network for further processing.
- A post-processing method, which is composed of the DBSCAN algorithm, is designed to cluster coordinate points.
- Experiments on public datasets and our self-collected dataset are designed to verify the effectiveness of the proposed framework.

The remainder of this paper is structured as follows. Section 1 provides a review of side-scan sonar target detection algorithms, while Section 2 introduces the framework of the proposed submarine target inspection system. Section 3 details the methodology for our spatial distribution probability-guided network and post-processing based on DBSCAN. Section 4 introduces the data collection for deep learning-based method training and validation. Section 5 presents the experimental results. Finally, Section 6 concludes the paper and outlines future work.

2. Overviews of Our Proposed Submarine Target Inspection System

The "Sea-Whale" UUV independently developed by the Center for Innovation Marine Robotics of Shenyang Institute of Automation, Chinese Academy of Sciences, which is a stable platform consisting of a navigation and positioning system, a control system, a propulsion system, a communication system and a target detection system. The framework of "Sea-Whale" UUV is shown in Figure 2. The core sensors include a side-scan sonar (for detecting target acoustic information), an altimeter (for acquiring the UUV's altitude), and an integrated navigation system combined with a GPS (Global Positioning System), INS (Inertial Navigation System), and DVL (Doppler Velocity Log) (for obtaining the UUV's coordinates). A Spatial Distribution Probability-Guided Detection Module is employed to detect targets within the sonar data acquired by the side-scan sonar. Detection results are transformed from image coordinates to world coordinates by the Target Position Calculation Module. Candidate target locations are fed into a DBSCAN-based post-processing module to aggregate them into a single point. This facilitates path planning and intelligent navigation, thereby enabling the online inspection of targets.

The hardware deployment of the submarine target inspection system is illustrated in Figure 3. The side-scan sonar receiver is responsible for ingesting sonar data, while the switch facilitates communication between the circuit boards. The core Target Detection

Module executes deep neural networks to compute target coordinates, and the Autonomous Planning Module handles decision-making and control.

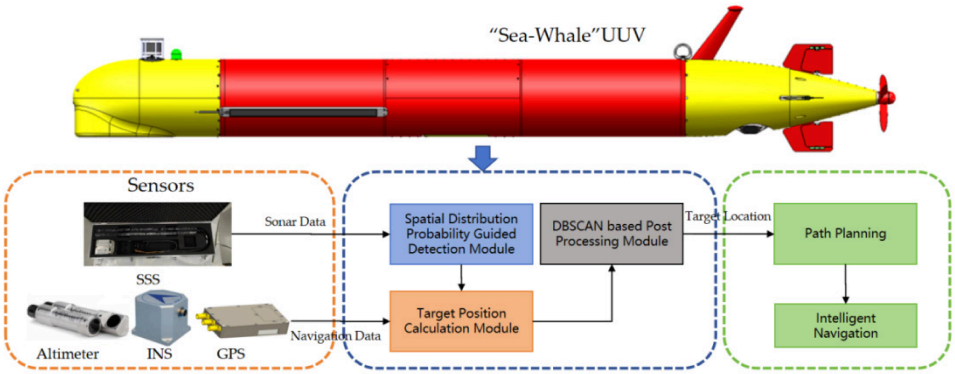


Figure 2. The framework system of the "Sea-Whale" UUV.

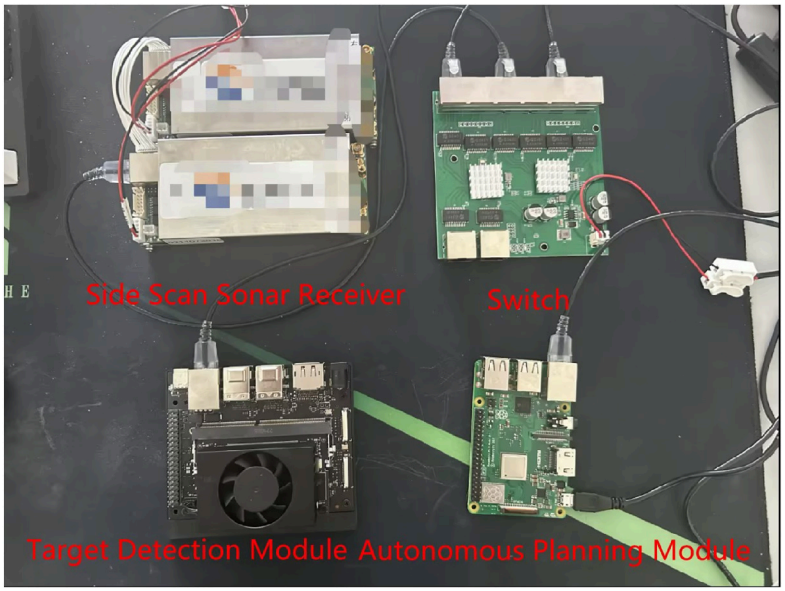


Figure 3. The hardware of the submarine target inspection system.

3. The Proposed Method

The core of the submarine target inspection system comprises three modules: the Spatial Distribution Probability-Guided Detection Module, the Target Position Calculation Module, and the DBSCAN-based post-processing module. This section details the capabilities of the Versatile Vision Foundation Model in sonar image processing, and the specifics of the three core components.

3.1. Versatile Vision Foundation Model for Sonar Image Processing

Versatile Vision Foundation Models, which are based on self-supervised learning, offer the potential to eliminate the need for manual data annotation, thereby enabling seamless scaling to massive datasets and larger architectures. A premier example of this paradigm is DINOv3 [26]. By generating high-quality dense features, DINOv3 achieves outstanding performance across a spectrum of vision tasks, significantly surpassing previous self- and weakly supervised foundation models. Notably, it outperforms specialized state-of-the-art methods in diverse settings without the need for fine-tuning.

To verify whether the DINOv3 model can extract features from sonar images, experiments are conducted to analyze the dense feature maps qualitatively. Principal Component

Analysis (PCA) is used to project the dense feature space into 3 dimensions, so that the feature maps generated by DINOv3 are visualized. The results are shown in Figure 4.

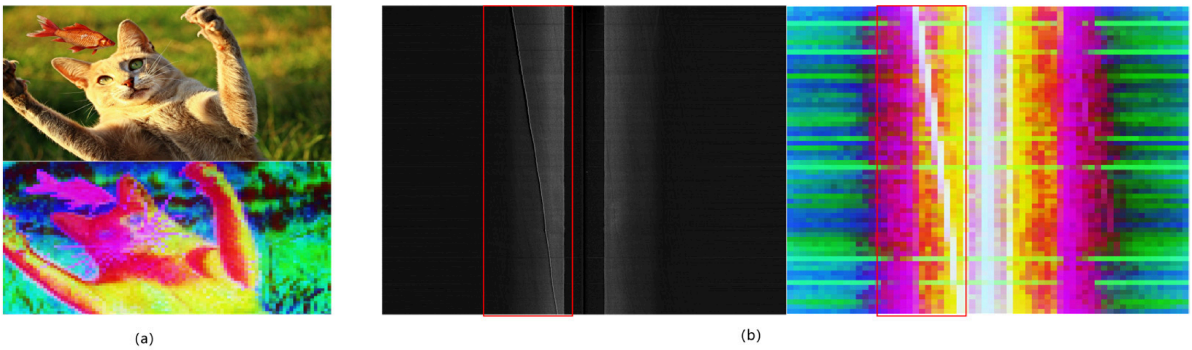


Figure 4. Visualization of DINOv3 feature maps. (**a**) Results on natural image; (**b**) results on sonar image.

In visual results of the feature maps, regions with the same semantic information share the same color. In Figure 4a, the colors of different parts of the cat, fish, and background are different, which illustrates that the DINOv3 model can extract high-quality feature from natural images. In Figure 4b, the color of the cable (marked by a red box) is different from the background, which illustrates that the DINOv3 model can be equipped with the capacity for sonar image feature extraction.

3.2. Spatial Distribution Probability-Guided Detection Module

Leveraging the high-quality feature extraction capabilities of the Versatile Vision Foundation Model, a Spatial Distribution Probability-Guided Detection Module is proposed in this paper, achieving underwater small-target detection under conditions of insufficient sonar data.

3.2.1. Spatial Distribution Probability

The pipeline for generating spatial distribution probability is illustrated in Figure 5. In the initial target feature extraction phase, a sonar image containing the target is fed into the Versatile Vision Foundation Model (loaded with pre-trained weights). Patches from the output feature map that correspond to the target are selected as the initial target features (red cuboid). Subsequently, the sonar image to be detected is input into the same model with shared weights to calculate the similarity between its feature map and the initial target features, thereby deriving the spatial distribution probability. As shown in the spatial distribution probability map in Figure 5, high responses are observed at the locations of the cable, which validates the effectiveness of the proposed method.

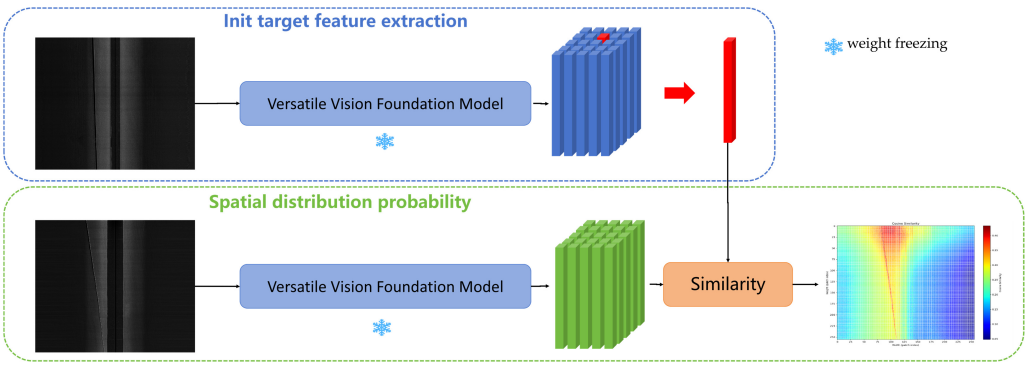


Figure 5. The pipeline for generating spatial distribution probability.

3.2.2. Framework of Spatial Distribution Probability-Guided Detection Module

The sonar image to be detected is processed by a Transformer-based feature extraction network. A key advantage of this network is its ability to capture long-range dependencies, meaning that every point in the feature map correlates with others, thereby embedding global information. Subsequently, the Feature Guidance Module fuses these deep, semantically rich feature maps with the spatial distribution probability obtained from the method in Section 3.2.1. This mechanism directs the network's attention to potential target locations, ultimately enhancing the model's object detection performance. During the training process, the weights of the Versatile Vision Foundation Model are kept frozen, while only the weights of the object detection network are updated. The architecture of the proposed Spatial Distribution Probability-Guided Detection Module is shown in Figure 6.

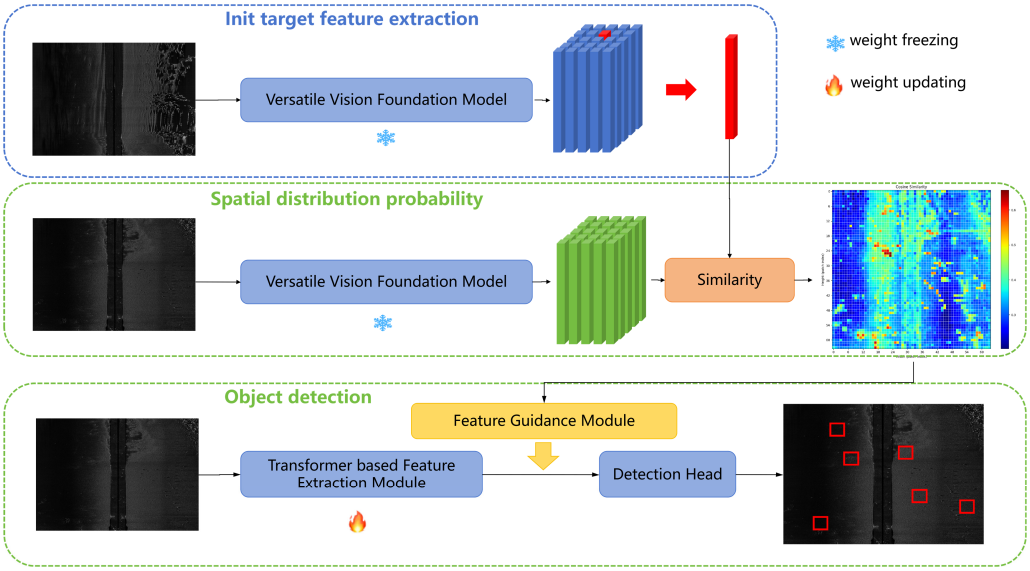


Figure 6. The architecture of the Spatial Distribution Probability-Guided Detection Module.

The details of the Transformer Blocks, Patch Merging, and Feature Guidance Module in the Spatial Distribution Probability-Guided Detection Module are presented as follows.

- **Transformer Blocks:** The Transformer-based feature extraction module employs four Transformer Blocks to capture the long-range dependencies inherent in linear target features. The architecture of a single Transformer block is illustrated in Figure 7a. Each block incorporates a pre-normalization design. First, the input feature maps are normalized using Layer Normalization (LayerNorm) and then processed by a Multi-Head Attention (MHA) mechanism. The output from the MHA is combined with the original input via a residual connection. This summed tensor is then passed through a second LayerNorm layer before entering a Feed-Forward Network (FFN), which produces the final output of the block. The core of the Transformer block is the Multi-Head Attention mechanism. This mechanism projects input vectors into multiple distinct subspaces, enabling each "head" to independently compute attention weights from different representational perspectives. The outputs from all heads are concatenated and subsequently integrated via a final linear projection. This process allows the final representation to synthesize rich information from diverse dimensions. Mathematically, the scaled dot-product attention is defined as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (1)$$

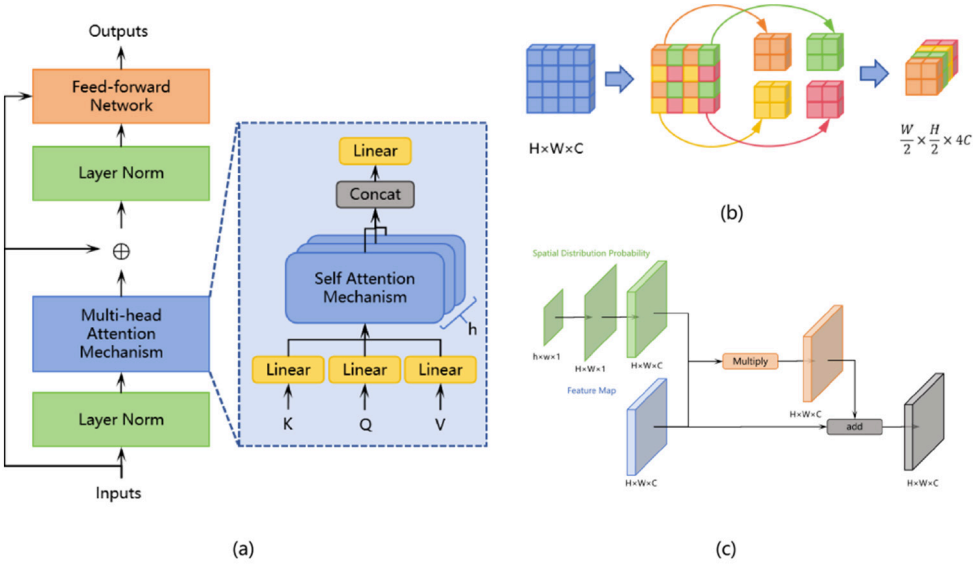


Figure 7. Sub-modules of the proposed network. (**a**) Transformer block. (**b**) Patch Merging. (**c**) Feature Guidance Module.

The Multi-Head Attention mechanism is formulated as

$$Multi\text{-}headAttention(Q, K, V) = Concat(head_1, \ldots, head_n)W^{out}, \quad (2)$$

where $head_i$ = *Attention* (QW_iQ,KW_iK,VW_iV); Q, K, V are the query, key and value matrices; d is the query/key dimension; W_iQ, W_iK, W_iK, and W_{out} are learnable weight matrices; and Concat denotes concatenation operation along the channel dimension.

- **Patch Merging:** The proposed network incorporates Patch Merging modules to perform spatial downsampling. The detailed operation is illustrated in Figure 7b. Given an input feature map of size $H \times W \times C$, the module partitions the spatial dimensions into non-overlapping 2×2 patches. Each group consists of four adjacent pixels: the top-left (orange), top-right (green), bottom-left (yellow), and bottom-right (red). These corresponding patches are then rearranged based on their relative positions. Subsequently, the rearranged patches are concatenated along the channel dimension, yielding a downsampled feature map of size $\frac{H}{2} \times \frac{W}{2} \times 4C$. A key advantage of Patch Merging is that it achieves spatial compression and preserves more detailed information.
- **Feature Guidance Module:** To achieve feature guidance via spatial distribution probability, we designed the Feature Guidance Module, as detailed in Figure 7c. First, the spatial distribution probability map with dimensions $h \times w \times 1$ is upsampled to $H \times W \times 1$ to align with the spatial resolution of the input feature map. It is then replicated along the channel dimension to match the channel count C of the input. Finally, utilizing a residual structure, the input feature map is multiplied by the transformed probability map to emphasize target regions. This result is then added to the original input feature map to yield the final feature-guided output.

3.3. Target Position Calculation Module

Geometric triangulation is employed to achieve precise spatial localization of the target, details are shown in Figure 8. By integrating the measurable UUV-to-seabed altitude with the acquired Slant Range to the target pixel, the global coordinates of the target are computed. This derivation operates under the assumption of a planar seabed and relies

on known UUV navigation states—specifically position and heading—as illustrated in the figure below.

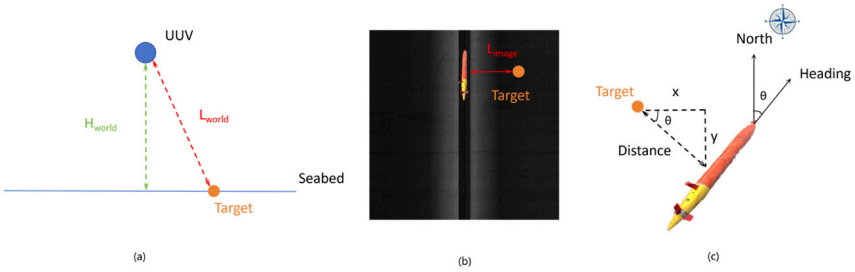


Figure 8. Schematic diagram of spatial localization. (**a**) Side-scan sonar imaging principle. (**b**) Location in image coordinate system. (**c**) Location in global coordinate system.

The distance between the UUV and target along longitude and latitude are described as follows.

$$\begin{aligned} L_{image} &= j - \frac{W}{2}, \\ L_{world} &= L_{image} \times ImageResolution, \\ Distance &= \sqrt{{L_{world}}^2 - {H_{world}}^2}, \\ &\begin{cases} x = Distance \times cos(\theta) \\ y = Distance \times sin(\theta) \end{cases}, \end{aligned} \tag{3}$$

where j is the pixel horizontal location of the target, W is the horizontal resolution of the sonar image, L_{image} is the Slant Range in the image coordinate system, L_{world} is the Slant Range in the global coordinate system, *ImageResolution* represents the real distance between adjacent pixels, Distance is the distance between the UUV and the target, H_{world} is the height of the UUV, θ is the heading angle of the UUV, and x, y are the distance of the UUV and target along longitude and latitude.

Given the Earth's equatorial radius a = 6,378,136, polar radius b = 6,356,751, and the longitude and latitude coordinates of UUV (*UUV_Longitude* and *UUV_Latitude*), the longitude and latitude coordinates of the target can be calculated as follows.

$$\begin{aligned} \alpha &= \arctan\left(\frac{b^2}{a^2} \times \tan\left(UUV_Latitude \times \frac{\pi}{180}\right)\right), \\ R &= \frac{1}{\sqrt{\left(\frac{\cos\alpha}{a}\right)^2 + \left(\frac{\sin\alpha}{b}\right)^2}}, \\ distin_1 &= \frac{\pi}{180} \times R \times \cos\alpha, \\ temppara &= \sqrt{\left[\left(\frac{R}{a}\right)^2 - \left(\frac{R}{b}\right)^2\right]^2 \times \left(\sin\frac{\alpha}{2}\right)^2 + 1}, \\ distin_2 &= temppara \times R \times \frac{\pi}{180}, \\ &\begin{cases} Longitude = UUV_Longitude + \frac{y}{distin_1} \\ Latitude = arctan\left(\frac{a^2}{b^2} \times \tan\left(\alpha + \frac{x}{distin_2} \times \frac{\pi}{180}\right)\right) \times \frac{180}{\pi} \end{cases}. \end{aligned} \tag{4}$$

3.4. DBSCAN-Based Post-Processing Module

Due to inertial navigation errors, the localization of the same target varies across different survey lines. To satisfy the requirements of subsequent path planning, this

paper employs the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [27] to cluster these discrete position coordinates.

The algorithm defines two key parameters: ε, which represents the radius of the neighborhood around a data point, and *MinPts*, the minimum number of points required to form a dense region. The points are categorized as follows.

- **Core Point:** A point is defined as a Core Point if its ε-neighborhood contains at least MinPts points (including itself). Core Points serve as the "seeds" of a cluster.
- **Border Point:** A point that is not a Core Point but lies within the ε-neighborhood of a Core Point is classified as a Border Point. Border Points belong to the cluster of the associated Core Point but cannot expand the cluster further.
- **Noise Point:** Points that are neither Core Points nor Border Points are considered Noise Points. These are treated as outliers or background noise and do not belong to any cluster.

The workflow of the DBSCAN algorithm is described as follows in Algorithm 1.

Algorithm 1 Density-Based Spatial Clustering of Applications with Noise(DBSCAN).

1. **Initialization**: Mark all points as "unvisited".
2. **Iteration**: Randomly select an unvisited point p from the dataset.
3. **Check**:
   - If the number of points in the ε-neighborhood of p is less than *MinPts*, then p is temporarily labeled as a Noise Point.
   - If the number of points in the ε-neighborhood of p is greater than or equal to *MinPts*, then p is identified as a Core Point. A new cluster is created, and p along with all points in its neighborhood are added to this cluster.
4. **Expansion**: For each Core Point newly added to the cluster, the algorithm recursively examines its ε-neighborhood. If a point within this neighborhood is also identified as a Core Point, all points in its neighborhood are incorporated into the current cluster. This process continues iteratively until no new points can be added to any cluster.
5. **Repetition**: Repeat steps 2–4 until all points have been visited. Ultimately, every point is either assigned to a cluster or labeled as Noise Point.

Unlike the traditional K-Means algorithm, which clusters data based on the distance to cluster centroids, DBSCAN groups data points based on "density connectivity," allowing it to handle arbitrary data distributions without being constrained by shape. Furthermore, DBSCAN does not require the number of clusters to be specified in advance. Consequently, the DBSCAN algorithm is well-suited for underwater target clustering scenarios.

4. Dataset

In order to train and validate the proposed Spatial Distribution Probability-Guided Detection Module, an actual collection of side-scan sonar datasets for submarine targets is carried out.

4.1. Data Acquisition

A proprietary "Sea-Whale" UUV equipped with ES4590 side-scan sonar systems from Hytratron is utilized as the data collection platform for the experiment, which is shown in Figure 9a. ES4590 side-scan sonar systems possess both high- and low-frequency modes, and the low-frequency mode is used in our experiments. The parameter details are shown in Table 1. We deployed two simulated moorings, two gas cylinders, and one simulated mine as the detection targets, which is shown in Figure 9b.

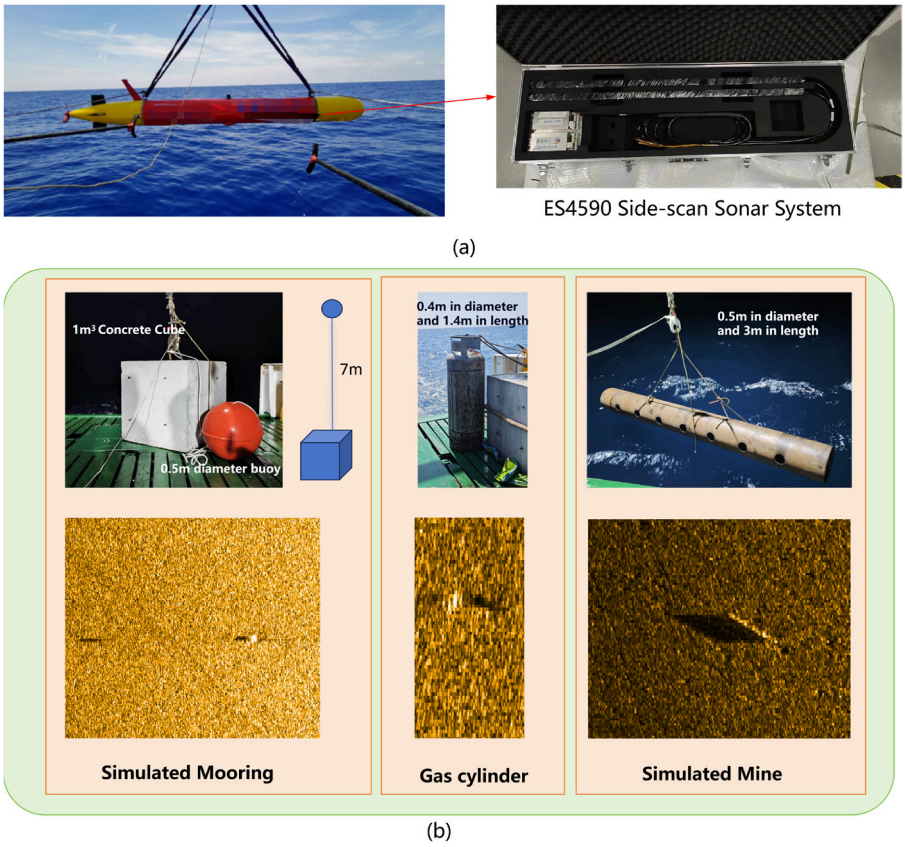


Figure 9. Data acquisition. (**a**) Data collection platform. (**b**) Deployed target and sonar image.

Table 1. Sonar parameters for the ES4590.

Parameter	High Frequency	Low Frequency
Frequency	900 kHz	450 kHz
Range	75 m	150 m
Horizontal beam width	0.2°	0.2°
Vertical beam width	50°	50°
Horizontal resolution	1 cm	2 cm
Vertical resolution	0.07 m@20 m 0.17 m@50 m 0.26 m@75 m	0.17 m@50 m 0.34 m@100 m 0.51 m@150 m

4.2. Data Processing

The acoustic intensity from raw sonar data could not be used for deep learning training; therefore, a pipeline of data processing is designed to generate sonar images and pixel-level annotation. Firstly, the acoustic intensity data is fed into a denoise module to remove environment noise and white noise. Then a normalization module is used to linearly normalize the acoustic intensity from 1 to 255. After that, an equalization module is used to deal with the horizontal value imbalance caused by acoustic attenuation. The horizontal acoustic intensity resolution provided by the side-scan sonar (SSS) is 10,800. Due to constraints on detection time and the GPU memory of the online processing board (NVIDIA Jetson Orin Nano), we downsampled it to 1024. Given that the dual-swath range is 300 m, the pixel spacing is approximately 0.3 m, which satisfies the detection requirements. The vertical resolution is determined by the accumulation of sonar data from different pings. The SSS transmits 3 pings of data per second, and we selected 1024 pings for accumulation, which corresponds to approximately 341 s of sonar data. Finally, we used LabelImg to annotate sonar images accurately and efficiently. The proposed dataset

contains 250 real side-scan sonar images. The training set and validation set are divided into the proportions of 80% and 20%.

5. Experiments and Analysis

To validate the performance of the proposed method, ablation studies and comparison experiments on a public dataset and the proposed dataset were conducted, with a specific focus on the Spatial Distribution Probability-Guided Detection Module and the DBSCAN-based post-processing module.

5.1. Experimental Setup

Training and validation of the deep neural network were conducted using an NVIDIA RTX 4090 (24 GB) GPU with PyTorch framework (version 2.4.0). The input image resolution was set to 1024 × 1024, with a batch size of 8. The model was trained for 1000 epochs using the AdamW optimization strategy, with an initial learning rate of 0.001. In terms of data augmentation, geometric transformations such as random translation, scaling, and horizontal flipping were applied to enhance the model's robustness to variations in linear target size and orientation.

Precision, Recall, mAP_{50}, and $mAP_{50–95}$ were used as evaluation metrics to evaluate the performance of the models.

Precision is the number of real targets out of all the detected targets. The mathematical formulation is expressed as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{5}$$

where TP is true positive and FP is false positive.

Recall is the number of targets detected out of the total. The mathematical formulation is expressed as follows:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{6}$$

where TP is true positive and FN is false negative.

mAP is the abbreviation for mean Average Precision, which is a widely adopted metric for evaluating accuracy in object detection tasks. mAP is calculated by integrating the area under the Precision–Recall curve. mAP_{50} is the mean Average Precision (mAP) over all classes at an IoU threshold of 0.5, which is used to evaluate object detection performance. $mAP_{50–95}$ is the mean Average Precision (mAP) over all classes at an IoU threshold from 50 to 95, which is a more strict metric used to evaluate object location performance.

The number of parameters is used to evaluate the size of the models. It is obtained by calculating the weights and biases of each layer in the neural network. In this experiment, we utilized the thop package in PyTorch to perform the calculation.

Inference time is used to evaluate the efficiency of the models.

5.2. Experiments on Public Dataset

The Public Mine Detection Dataset [28] used in this study contains 1170 real sonar images taken between 2010 and 2021 using a Teledyne Marine Gavia Autonomous Underwater Vehicle (AUV), which includes enough information to classify its content objects as NOn-Mine-like BOttom Objects (NOMBO) and MIne-Like COntacts (MILCO). The dataset is annotated and can be quickly deployed for object detection, classification, or image segmentation tasks. The dataset was randomly split into a training set of 1158 images and a validation set of 12 images.

An ablation study was conducted on the Public Mine Detection Dataset. The proposed Spatial Distribution Probability-Guided Detection Module could be modified to any object detection network. To validate the efficacy of our approach, we selected the classic YOLOv8 [29] as the baseline for experiments. To ensure a fair comparison, all methods were trained using identical training parameters. The results of the ablation study are presented in Table 2. The baseline YOLOv8 model achieves a solid performance. Upon integrating the Transformer architecture, both precision and mAP_{50} are significantly enhanced, indicating that global feature modeling effectively reduces false positives. The introduction of the Guidance Module significantly improves the Recall, thereby minimizing the occurrence of missed detections. Notably, the incorporation of the proposed Guidance Module and Transformer architecture leads to a substantial improvement in the $mAP_{50–95}$ metric (from 0.56403 to 0.88613) and achieves a satisfying recall rate. This demonstrates that our method not only reduces missed detection but also ensures highly accurate localization.

Table 2. Results of ablation study on Public Mine Detection Dataset.

Methods	Precision	Recall	mAP_{50}	$mAP_{50–95}$	Parameters	Time(ms)
YOLOv8	0.81920	0.85510	0.90701	0.56403	3.0 M	3.9
YOLOv8 + Transformer	0.95707	0.88889	0.96115	0.59123	74.0 M	20.3
YOLOv8 + Guidance	0.99186	1	0.995	0.69136	314.4 M	32.4
YOLOv8 + Transformer + Guidance (ours)	0.99335	1	0.995	0.88613	387.4 M	46.6

The introduction of the Transformer architecture and the Guidance Module leads to an increase in both the number of parameters and the inference time. The number of parameters of the proposed method is 387.4 M, which can be easily deployed on the Target Detection Module introduced in Section 2. The SSS transmits 3 pings of data per second and the Target Detection Module performs the algorithm every 10 pings. The inference time 46.6 ms is far less than the interval of data reception 3333 ms.

A visual comparison of the detection results on the Public Mine Detection Dataset is presented in Figure 10. Both methods successfully detect the mine targets. Specifically, for the top sonar image, the confidence score of the target detected by the baseline YOLOv8 is only 0.3. In contrast, the integration of the Transformer and the proposed method significantly improves this score to 0.9 and 0.8, respectively. This further validates the effectiveness of the proposed approach.

Since the Public Mine Detection Dataset does not provide AUV status information, it is impossible to calculate the geodetic coordinates of the targets. Therefore, the DBSCAN-based post-processing module was not validated on this dataset.

A comparison experiment is designed to prove the advantage of the proposed method. We compare the evaluation metrics of the proposed method with YOLOv4 [28] (which is reported in the paper proposing the Public Mine Detection Dataset), YOLOv8 (the baseline), and YOLO26 [30] (the latest version of YOLO). Results are shown in Table 3. The experimental results demonstrate that our proposed YOLOv8 + Transformer+ Guidance method achieves a satisfying performance in mine detection. Compared to the classic YOLOv4, the mainstream YOLOv8, and even the state-of-the-art YOLO26, our approach not only attains a 100% Recall (ensuring zero missed detections within the validation set), but also elevates the most rigorous comprehensive localization metric, $mAP_{50–95}$, to 0.88613 while maintaining an exceptional Precision of 99.335%. The $mAP_{50–95}$ of the proposed method outperforms both the baseline YOLOv8 (0.56403) and the current leading YOLO26 (0.74311), fully validating that the integration of the Transformer and Guidance mechanisms effectively enhances feature extraction, anti-interference capability, and bounding box localization accuracy under complex scenarios.

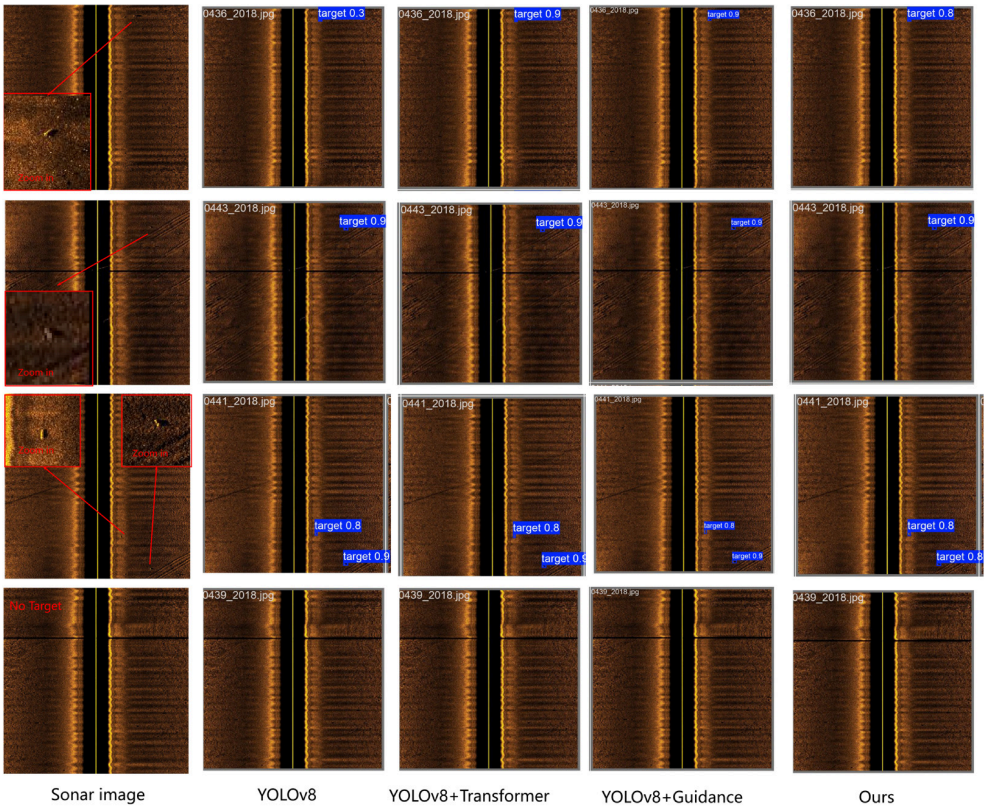


Figure 10. Visualization results on Public Mine Detection Dataset.

Table 3. Comparison results on Public Mine Detection Dataset.

Methods	Precision	Recall	mAP_{50}	$mAP_{50–95}$
YOLOv4	0.82	0.64	0.75	-
YOLOv8	0.81920	0.85510	0.90701	0.56403
YOLO26	0.99751	1	0.995	0.74311
YOLOv8 + Transformer + Guidance (ours)	0.99335	1	0.995	0.88613

5.3. Experiments on Our Dataset

To further validate the detection capability of the proposed Spatial Distribution Probability-Guided Detection Module for underwater targets, validation experiments were conducted on our proprietary dataset. In the experiments, the Target Position Calculation Module and the DBSCAN-based post-processing module were utilized to achieve target localization and clustering, respectively.

We adopted the same comparative methodology described in Section 5.2, and the results are presented in Table 4. Due to the extremely small scale of the targets and the limited number of samples in the dataset, the baseline YOLOv8 achieved a mAP_{50} of only 0.50996. This indicates that the standard YOLOv8 fails to effectively perform target detection on our dataset. The integration of the Transformer architecture significantly improves the model's recall (increasing to 0.615) and localization precision, as evidenced by the substantial jump in mAP scores. The mAP_{50} rises to 0.679 and the $mAP_{50–95}$ reaches 0.285, showcasing the Transformer's effectiveness in capturing complex feature dependencies. With the help of the Guidance Module, the Recall improves from 0.45332 to 0.63546, which proves that the incorporation of spatial distribution probability contributes to a reduction in missed detections. Our proposed method, YOLOv8 + Transformer + Guidance, achieves the most balanced and superior performance. It attains the highest Precision (0.813) and Recall (0.692) among all models, indicating both high reliability in predictions and a reduced

rate of missed detections. This results in a mAP_{50} score of 0.715. Although the $mAP_{50–95}$ (0.250) is slightly lower than that of the Transformer-only variant, this discrepancy may be attributed to the quality of data annotation. Nevertheless, the significant gain in strict precision metrics confirms the effectiveness of the proposed guidance mechanism.

Table 4. Results of ablation study on our dataset.

Methods	Precision	Recall	mAP_{50}	$mAP_{50–95}$	Parameters	Time(ms)
YOLOv8	0.74618	0.45332	0.50996	0.18642	3.0 M	4.5
YOLOv8 + Transformer	0.68769	0.61538	0.67939	0.28468	74.0 M	23.0
YOLOv8 + Guidance	0.67336	0.63546	0.65109	0.20411	314.4 M	56.9
YOLOv8 + Transformer + Guidance (ours)	0.81269	0.69231	0.71465	0.24956	387.4 M	70.0

The number of parameters has no relation to the hardware or image resolution, so the results are the same as those in Section 5.2. The inference time is 70.0 ms, which demonstrates that the proposed method can be applied to object detection tasks for UUVs.

As can be observed from the visualization, shown in Figure 11, the baseline YOLOv8 fails to detect the targets. While YOLOv8 + Transformer is capable of detecting targets, it still suffers from missed detections. In contrast, the proposed method, incorporating the Spatial Distribution Probability-Guided Detection Module, achieves high-quality target detection.

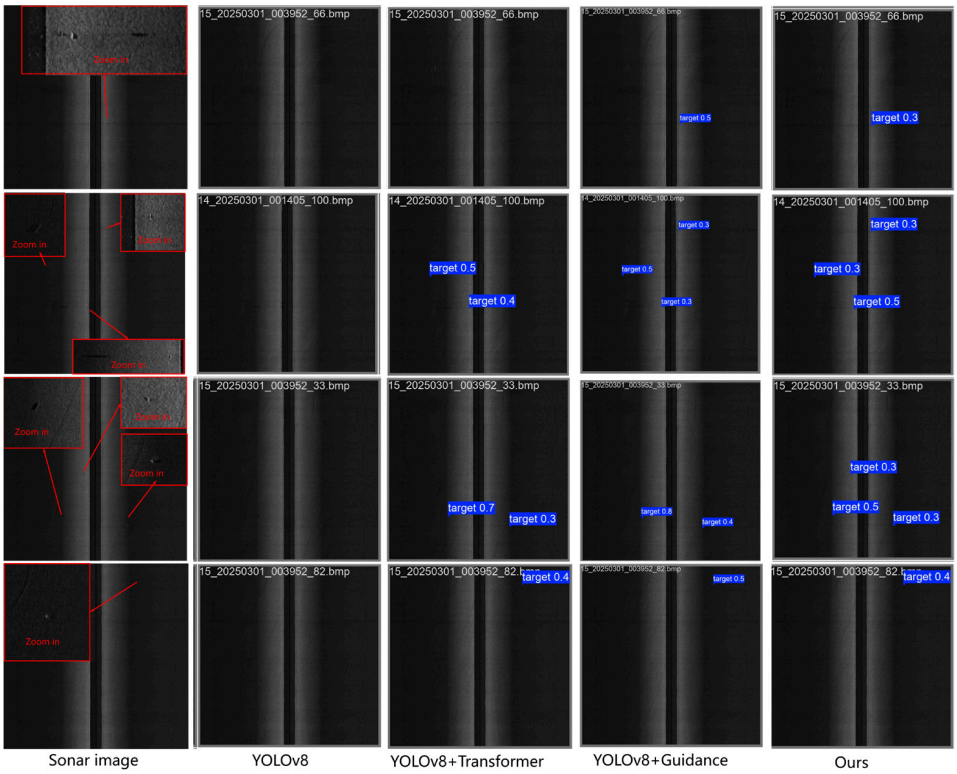


Figure 11. Visualization results on our dataset.

A comparison experiment is conducted to prove the advantage of the proposed method. We compare the evaluation metrics of the proposed method with YOLOv8 (the baseline) and YOLO26 (the latest version of YOLO). Results are shown in Table 5. Experimental results demonstrate that our improved model (YOLOv8 + Transformer + Guidance) outperforms both the original YOLOv8 and YOLO26 across all metrics. Compared to the baseline YOLOv8, the proposed method not only significantly boosts mAP_{50} from 0.50996 to 0.71465, but also successfully breaks the traditional trade-off between precision and recall in object detection, achieving simultaneous and substantial increases in both Precision

(0.81269) and Recall (0.69231). This fully validates the effectiveness of incorporating a Transformer to enhance global feature extraction and a Guidance mechanism to optimize the detection head, enabling the model to achieve an improvement in overall detection accuracy and robustness while drastically reducing false positives and missed detections.

Table 5. Comparison results on our dataset.

Methods	Precision	Recall	mAP_{50}	$mAP_{50–95}$
YOLOv8	0.74618	0.45332	0.50996	0.18642
YOLO26	0.74263	0.46154	0.59807	0.20426
YOLOv8 + Transformer + Guidance (ours)	0.81269	0.69231	0.71465	0.24956

To validate the effectiveness of the proposed target localization and clustering methods, we conducted two comb-search missions in the target deployment area. The localization and clustering results are presented in Figure 12. The red pentagrams represent the ground truth positions of the deployed targets, obtained via GPS coordinates recorded during deployment.

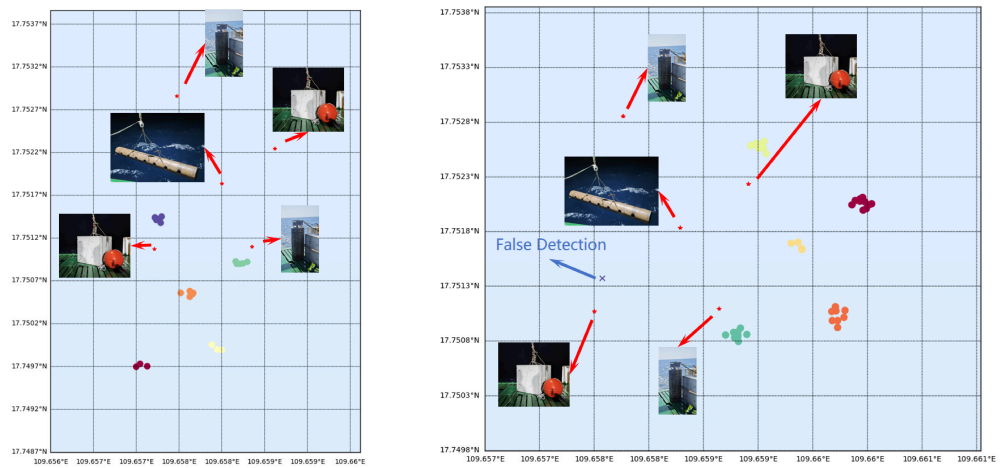


Figure 12. Localization and clustering results. (**a**) Result of the first mission. (**b**) Result of the second mission.

In the first mission, 24 detections (dots) of the 5 deployed targets were achieved across survey lines. After processing by the DBSCAN-based post-processing module, analysis confirmed that all suspected points corresponded to the deployed targets. The different colors in the figure indicate the clustering results.

In the second mission, 41 detections (dots) of the 5 deployed targets were achieved across survey lines. Following clustering by the DBSCAN-based post-processing module, analysis revealed that 40 detections were actual deployed targets (dots), while one was a false detection (blue cross). Different colors again represent the clustering results.

It can be observed that the relative spatial distribution of targets detected in different missions matches the actual distribution of the deployed targets, albeit with a positional offset. This discrepancy is attributed to inertial navigation errors, a result that underscores the necessity of online target detection and localization.

6. Discussion

This section aims to deeply explore the effectiveness, limitations, and physical implications of the proposed framework in addressing the challenges of underwater side-scan sonar (SSS) image detection.

6.1. Generalization Capability in Few-Shot Scenarios

Underwater object detection has long been constrained by data scarcity (few-shot) and textureless problems. Traditional deep learning models (such as standard YOLOv8) often perform poorly in small-sample, small-object detection, prone to overfitting or missed detections. The Spatial Distribution Probability Guidance mechanism proposed in this study effectively addresses this pain point by introducing a pre-trained Vision Foundation Model (DINOv3). Experimental results indicate that on the Public Mine Detection Dataset, the $mAP_{50–95}$ metric surged from 0.564 to 0.886 after introducing the Transformer feature extractor and guidance mechanism. This demonstrates that prior features extracted by foundation models can significantly enhance the network's feature extraction capability under data-scarce conditions, enabling the model to learn robust sonar image features even with extremely few training samples. Although in the experiments on our dataset the $mAP_{50–95}$ (0.250) is slightly lower than that of the Transformer-only variant, this discrepancy may be attributed to the quality of data annotation. Nevertheless, the significant gain in strict precision metrics confirms the effectiveness of the proposed guidance mechanism.

6.2. Localization Accuracy and Robustness in Complex Environments

In experiments on the self-constructed dataset (containing simulated mines, buoys, etc.), the framework achieved a mAP_{50} of 0.715, despite the extremely small target sizes and complex backgrounds. Notably, during the field "comb-search" mission, the system successfully detected and clustered discrete target points. However, we also observed a certain degree of offset between the detection results and the ground truth. This offset is mainly attributed to the inertial navigation drift of the UUV. This finding underscores the necessity of online object detection and post-processing clustering (such as DBSCAN) in underwater robotic systems.

6.3. Noise Suppression and False Alarm Handling

In the second field mission, the DBSCAN post-processing module successfully clustered 41 suspected target points into 40 real targets and 1 noise point. This indicates that the framework possesses good noise suppression capabilities, distinguishing real targets from false alarms caused by sonar reverberation or complex seabed topography. This is crucial for subsequent path planning and manual review, as it directly reduces the false alarm rate of underwater operations.

6.4. Limitations of the Method

Although the framework has achieved significant improvements in detection accuracy, there is still room for improvement. Firstly, the current inference speed has not yet fully met the requirements for large-scale UUV onboard real-time deployment (limited by the Transformer structure and high-resolution input); secondly, the current localization calculation assumes a flat seabed, which may introduce geometric distortion errors in complex terrains (such as steep slopes or deep trenches).

7. Conclusions

This paper proposes a novel Spatial Distribution Probability-Guided Detection Framework aimed at solving the problems of target sparsity, difficulty in data acquisition, and feature extraction in underwater side-scan sonar (SSS) images. The main contributions and conclusions of this study are as follows:

- **Innovative Detection Architecture:** We designed the Spatial Distribution Probability-Guided Detection Module. This module utilizes a general-purpose Vision Foundation Model (DINOv3) to generate spatial distribution probability maps, guiding the

Transformer-based feature extraction network. This mechanism breaks the dependence of traditional Convolutional Neural Networks on large amounts of annotated data, achieving high-precision object detection under few-shot conditions.

- **Complete Perception System:** We constructed a complete system comprising an Object Position Calculation Module (converting image coordinates to global longitude and latitude) and a DBSCAN-based post-processing module (aggregating discrete detection points). This enables the UUV to perform online detection, global localization, and intelligent navigation.
- **Empirical Validity:** On the public mine dataset, the method verified its superiority in low-data regimes. On the self-constructed complex scenario dataset, the model achieved a mAP_{50} of 0.715, significantly outperforming baseline models. Field sea trials verified that the system can effectively distinguish real targets from noise and correct localization errors caused by inertial navigation drift through clustering algorithms.

In summary, the framework proposed in this study provides a robust solution for underwater sonar image analysis, successfully bridging the gap between sparse data availability and the need for high-precision perception. Future work will focus on the following aspects: introducing visual or other acoustic sensor data to enhance perception robustness in extremely turbid or acoustically complex environments; optimizing computational efficiency to achieve real-time inference on resource-constrained UUVs; and extending the framework to adapt to targets with more complex geometric shapes.

Author Contributions: Conceptualization, D.J.; Methodology, D.J. and Y.H.; Software, J.Q.; Resources, J.Y.; Writing—original draft, D.J.; Visualization, H.F.; Supervision, J.Y.; Project administration, Z.W.; Funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by State Key Laboratory of Robotics at Shenyang Institute of Automation [grant number: 2023-Z07], State Key Laboratory of Robotics and Intelligent Systems [grant number: 2025-Z02-01], State Key Laboratory of Robotics and Intelligent Systems [grant number: 2025-Z25], the National Natural Science Foundation of China (No. 52501425), and China Postdoctoral Science Foundation (No. 2025M780273).

Data Availability Statement: The Public Mine Detection Dataset is available at https://doi.org/10.1016/j.dib.2024.110132, reference number [28]. Our proposed data is not publicly available due to restrictions.

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

References

1. McKay, J.; Gerg, I.; Monga, V.; Raj, R.G. What's mine is yours: Pretrained CNNs for limited training sonar ATR. In *Proceedings of the OCEANS 2017-Anchorage, Anchorage, AK, USA, 18–21 September 2017*; IEEE: New York, NY, USA, 2017; pp. 1–7.
2. Zhu, K.; Tian, J.; Huang, H. Underwater objects classification method in high-resolution sonar images using deep neural network. *Acta Acust.* **2019**, *44*, 595–603.
3. Cheng, Z.; Huo, G.; Li, H. A multidomain collaborative transfer learning method with multiscale repeated attention mechanism for underwater sidescan sonar image classification. *Remote Sens.* **2022**, *14*, 355. [CrossRef]
4. Xu, H.; Yang, L.; Long, X. Underwater sonar image classification with small samples based on parameter-based transfer learning and deep learning. In *Proceedings of the Global Conference on Robotics, Artificial Intelligence and Information Technology (GCRAIT), Chicago, IL, USA, 30–31 July 2022*; IEEE: New York, NY, USA, 2022; pp. 304–307.
5. Ochal, M.; Vazquez, J.; Petillot, Y.; Wang, S. A comparison of few-shot learning methods for underwater optical and sonar image classification. In *Proceedings of the Global Oceans 2020, Biloxi, MS, USA, 5–30 October 2020*; IEEE: New York, NY, USA, 2020; pp. 1–10.
6. Chen, Y.; Li, B.; Liang, H.; Yang, C. Research on sonar image few-shot classification based on deep learning. *J. Northwest. Polytech. Univ.* **2022**, *40*, 739–745.

7. Xu, H.; Yang, L.; Zhang, M. Unsupervised classification based on deep adaptation network for sonar images. *J. Electron. Imaging* **2023**, *32*, 013029. [CrossRef]
8. Bell, J.M. A Model for the Simulation of Sidescan Sonar. Ph.D. Thesis, Heriot-Watt University, Edinburgh, UK, 1995.
9. Bell, J.M.; Darlington, D.J.; Elston, G.R. Techniques for the physical modelling of the sonar image generation process. In *Proceedings of the SEE International Conference on Physics in Signal and Image Processing (PSIP'99)*; SEE: Paris, France, 1999; pp. 66–72.
10. Cerqueira, R.; Trocoli, T.; Neves, G.; Joyeux, S.; Albiez, J.; Oliveira, L. A novel GPU-based sonar simulator for real-time applications. *Comput. Graph.* **2017**, *68*, 66–76. [CrossRef]
11. Cerqueira, R.; Trocoli, T.; Albiez, J.; Oliveira, L. A rasterized ray-tracer pipeline for real-time, multi-device sonar simulation. *Graph. Model.* **2020**, *111*, 101086. [CrossRef]
12. Li, X. Research on Target Sample Generation and Classification of Side Scan Sonar Image. Master's Thesis, Harbin Engineering University, Harbin, China, 2020.
13. Li, B.; Huang, H.; Liu, J.; Li, Y. Optical image-to-underwater small target synthetic aperture sonar image translation algorithm based on improved CycleGAN. *Acta Electron. Sin.* **2021**, *49*, 1746–1753.
14. Hu, Y.; Zhang, W.; Li, B.; Liu, J.; Huang, H. Self-perceptual generative adversarial network for synthetic aperture sonar image generation. In *Proceedings of the Fourteenth International Conference on Graphics and Image Processing (ICGIP 2022)*; SPIE: Bellingham, WA, USA, 2023; Volume 12705, pp. 864–872.
15. Du, Y.; Lin, W.; Zhong, W.; Yuan, Y. An effective approach for sonar image recognition with improved efficientdet and ensemble learning. *J. Phys. Conf. Ser.* **2022**, *2258*, 012038.
16. Lei, C.; Wang, H.; Lei, J. Enhancing side-scan sonar image classification based on graph structure. *IEEE Sens. J.* **2024**, *24*, 24388–24404. [CrossRef]
17. Ruan, F.; Dang, L.; Ge, Q.; Zhang, Q.; Qiao, B.; Zuo, X. Dualpath residual "shrinkage" network for side-scan sonar image classification. *Comput. Intell. Neurosci.* **2022**, *2022*, 6962838. [CrossRef] [PubMed]
18. Tang, Y.; Li, H.; Zhang, W.; Bian, S.; Zhai, G.; Liu, M.; Zhang, X. Lightweight DETR-YOLO method for detecting shipwreck target in side-scan sonar. *Syst. Eng. Electron.* **2022**, *44*, 2427–2436.
19. Steiniger, Y.; Groen, J.; Stoppe, J.; Kraus, D.; Meisen, T. A study on modern deep learning detection algorithms for automatic target recognition in sidescan sonar images. In *Proceedings of the Meetings on Acoustics*; Acoustical Society of America: Melville, NY, USA, 2021; Volume 44, p. 070010.
20. Yu, Y.; Zhao, J.; Gong, Q.; Huang, C.; Zheng, G.; Ma, J. Realtime underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. *Remote Sens.* **2021**, *13*, 3555.
21. Li, B.; Huang, H.; Liu, J.; Wei, L. Underwater Small Target Detection Method and System for Synthetic Aperture Sonar Image. CN202311062705.1, 5 January 2024.
22. Yang, C.; Li, Y.; Jiang, L.; Huang, J. Foreground enhancement network for object detection in sonar images. *Mach. Vis. Appl.* **2023**, *34*, 56. [CrossRef]
23. Tang, Y.; Wang, L.; Li, H.; Bian, S. Side-scan sonar underwater target segmentation using the BHPUNet. *EURASIP J. Adv. Signal Process.* **2023**, *2023*, 76. [CrossRef]
24. Er, M.J.; Chen, J.; Zhang, Y. Marine robotics 4.0: Present and future of real-time detection techniques for underwater objects. In *Industry 4.0—Perspectives and Applications*; IntechOpen: Rijeka, Croatia, 2022; p. 8.
25. Wang, J.; Shan, T.; Muthukumaran, C.; Osedach, T.; Englot, B. Deep learning for detection and tracking of underwater pipelines using multibeam imaging sonar. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop, Montreal, QC, Canada, 20–24 May 2019*; IEEE: New York, NY, USA, 2019.
26. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F. DINOv3. *arXiv* **2025**, arXiv:2508.10104.
27. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **2017**, *42*, 1–21.
28. Nuno, P.S.; Ricardo, M.; Gonçalo, S.T.; Lobo, V.; de Castro Neto, M. Side-scan sonar imaging data of underwater vehicles for mine detection. *Data Br.* **2024**, *53*, 110132.
29. Glenn, J.; Ayush, C.; Qiu, J. Ultralytics YOLOv8, version 8.0.0; Ultralytics: Frederick, MD, USA. Available online: https://github.com/ultralytics/ultralytics (accessed on 14 April 2026).
30. Ranjan, S.; Rahul, H.C.; Ajay, S.; Manoj, K. YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection. *arXiv* **2025**, arXiv:2509.25164.

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.