remote sensing

MDPI

Article

Ultra-Fast Object Detection for Side-Scan Sonar Images via Target Presence Awareness

Guoqing Xie [1,2], Guang Pan [1], Ju He [1], Hu Xu [3,*] and Yang Yu [1]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; xieguoqing@mail.nwpu.edu.cn (G.X.); hj@mail.nwpu.edu.cn (J.H.); nwpuyuy@nwpu.edu.cn (Y.Y.)
[2] Yichang Testing Technique Research Institute, Wuhan 430010, China
[3] State Key Laboratory of Submarine Geoscience, Shanghai Jiao Tong University, Shanghai 200240, China
* Correspondence: xuhu@mail.nwpu.edu.cn

Highlights

What are the main findings?

- A lightweight coarse-to-fine SSS object detection framework with Target Presence Analysis (TPA) module
- Achieved an ultra-fast inference speed of 174.74 FPS on embedded platforms while maintaining higher detection accuracy and recall compared to state-of-the-art models like YOLOv11m.

What is the implication of the main finding?

- The target presence-aware strategy significantly reduces computational redundancy and energy consumption, enabling real-time ocean observation for resource-constrained AUVs during long-duration missions.

Abstract

Side-scan sonar (SSS) imaging plays a critical role in underwater perception for autonomous underwater vehicles (AUVs). However, the spatial sparsity of targets and the limited computational resources remain challenging for real-time object detection. Existing methods typically adopt dense inference strategies, leading to substantial computational redundancy and limited deployment feasibility. In this work, we propose a lightweight and ultra-fast SSS object detection framework based on target presence awareness. The proposed framework follows a coarse-to-fine inference paradigm, in which a target presence analysis module is first employed to rapidly filter out target-absent image patches, and only target-positive patches are forwarded to an Object Forward Detection (OFD) module for fine-grained detection. The TPA module integrates spatial–frequency convolution to efficiently capture both local structural cues and global contextual information with minimal computational overhead. Furthermore, an AttnConv-enhanced detection module is introduced in the OFD stage to strengthen high-frequency target features and improve fine-grained detection performance. Extensive experiments on public SSS datasets demonstrate that the proposed method achieves an mAP of 74.63% on the AI4Shipwrecks dataset and 63.02% on the SSS-Mine dataset. Notably, the framework delivers an ultra-fast inference speed of 174.74 FPS on embedded hardware, representing a 5.2× speedup over conventional dense-processing detection methods.

Keywords: object detection; side-scan sonar; lightweight detection; target presence awareness

Academic Editor: Filiberto Chiabrando

Received: 11 March 2026
Revised: 29 April 2026
Accepted: 20 May 2026
Published: 22 May 2026

1. Introduction

Side-scan sonar (SSS) has become a fundamental sensor for autonomous underwater vehicles in a wide range of underwater applications, including seabed mapping, marine debris inspection, and target recognition [1–3]. As illustrated in Figure 1, automatic and real-time object detection from SSS images plays a crucial role in enabling reliable underwater perception and decision-making for long-duration autonomous missions. Despite its advantages, efficient object detection in high-resolution SSS images remains highly challenging due to several intrinsic characteristics of sonar images [4].

The imaging mechanism of SSS differs fundamentally from that of conventional optical CCD sensors. Unlike optical cameras, which capture discrete area-based frames, SSS operates using a line-by-line scanning strategy. As the Autonomous Underwater Vehicle (AUV) advances, the transducer continuously emits acoustic pulses and records the corresponding backscattered intensity along a narrow swath perpendicular to the direction of motion. These successive acoustic returns are then sequentially aggregated along the vehicle's trajectory to generate a continuous waterfall image. As a result, SSS imagery typically exhibits elongated, strip-like structures, enabling coverage of large-scale seabed regions. To accommodate memory and computational constraints, these images are commonly divided into multiple patches for sequential processing. Moreover, underwater environments are dominated by background regions such as sandbeds, mudflats, and uniform seabed textures, while meaningful targets such as shipwrecks, pipelines, or man-made objects appear sparsely and occupy only a small fraction of the image area [5]. As a result, the majority of neural network computations are expended on background regions that do not contribute to effective target recognition.

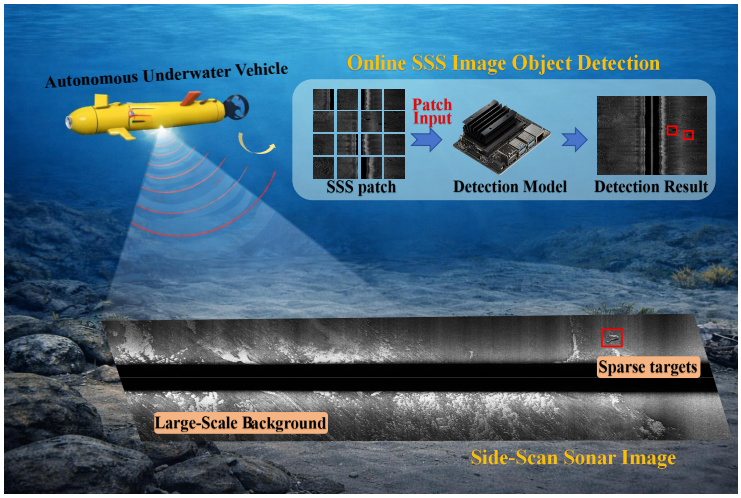

Figure 1. The overview of real-time SSS image detection in resource-constrained AUV environments.

Early underwater object detection approaches primarily relied on handcrafted feature extraction and statistical modeling. Among these, the Constant False Alarm Rate (CFAR) detection is a representative technique, which identifies targets by comparing pixel intensities against adaptive thresholds derived from local background statistics [6]. However, these traditional methods possess inherent limitations. They rely heavily on manually selected parameters and heuristic thresholds, which lack adaptability to varying underwater environments. In the presence of complex seabed reverberations, low signal-to-noise ratios (SNR), and non-uniform intensity distributions, traditional detectors often suffer from high false alarm rates and poor robustness, failing to capture the high-level semantic features of diverse targets.

To overcome these bottlenecks, traditional machine learning (ML) methods were introduced, shifting the focus from heuristic rules to statistical classification based on handcrafted features. These approaches typically involve a two-stage pipeline: feature extraction followed by a supervised classifier. Researchers have explored various texture and geometric descriptors, such as Haar-like features [7], Local Binary Patterns (LBP) [8], and Histograms of Oriented Gradients (HOG) [9], to represent sonar targets. For instance, support vector machines (SVM) have been widely employed to distinguish targets from complex seabed clutter [10]. While traditional ML methods offer better generalization than simple signal processing, they still rely heavily on manual feature engineering, which is labor-intensive and often fails to capture the subtle, non-linear acoustic characteristics of diverse underwater objects in dynamic environments.

Driven by the need for more robust representations, deep learning (DL)-based methods have emerged as the dominant paradigm. Most current SSS object detection methods [11,12] adopt a one-stage or dense-processing paradigm, where all image patches are processed uniformly by deep neural networks regardless of whether targets are present. Song et al. [13] introduced a real-time detection framework for AUVs based on self-cascaded convolutional neural networks, aiming to improve inference efficiency. Le et al. [14] explored the use of deep neural networks for underwater object detection by replacing conventional convolutional layers with Gabor-based filters, which enhanced sensitivity to small targets. Cao et al. [15] employed the single-stage detector to extract candidate obstacle regions from sonar imagery, while Zhang et al. [16] adopted transfer learning to construct a forward-looking sonar target detection model based on YOLOv10. To further enhance feature representation, Yu et al. [17] combined YOLOv11 with transformer modules, introducing attention mechanisms and multi-scale down-sampling operations. In addition, Zhang et al. [16] developed a self-training sonar target detection framework using automated deep learning, incorporating a memory-efficient differentiable architecture search strategy to support large input sizes and flexible network configurations. Although these methods achieve promising recognition accuracy, they suffer from substantial computational redundancy and are ill-suited for real-time deployment on resource-constrained AUV platforms.

As illustrated in Figure 2, the majority of sonar image patches contain no targets of interest. Based on this observation, we propose an ultra-fast SSS object detection framework driven by target presence awareness. Instead of directly applying computationally expensive fine-grained recognition to all image patches, a coarse-to-fine inference strategy is adopted. Specifically, a lightweight target presence analysis module (TPA) is first employed to rapidly determine whether a given image patch potentially contains a target. Only patches identified as target-positive are subsequently forwarded to a fine-grained recognition network for accurate target identification, while target-absent patches are discarded at an early stage. The TPA module is designed to be highly efficient and recall-oriented, ensuring that potential targets are preserved while significantly reducing unnecessary computation. Quantitative evaluations on representative public SSS datasets indicate that the proposed framework exhibits a favorable balance between detection performance and computational efficiency. Specifically, the target presence-aware strategy demonstrates the potential to significantly improve inference throughput for high-resolution SSS imagery, achieving an inference speed of 174.74 FPS on embedded hardware while maintaining accuracy levels competitive with state-of-the-art dense-processing models. Experimental results suggest that by effectively bypassing target-absent regions, the method can reduce inference latency and energy footprints. This makes it a promising candidate for real-time underwater target recognition, particularly for autonomous missions where computational

resources and power budgets are constrained. The main contributions of this paper are summarized as follows:

(1) We propose a lightweight target presence-aware recognition framework for large-scale high-resolution SSS images, which explicitly exploits the sparsity of underwater targets to reduce redundant computation;
(2) A fast target presence analysis module is designed to efficiently filter out target-absent sonar patches before fine-grained recognition, thereby enabling a coarse-to-fine inference strategy;
(3) Ultra-fast and energy-efficient target recognition performance is achieved while maintaining competitive accuracy, making the proposed method suitable for real-time deployment on resource-constrained underwater platforms.

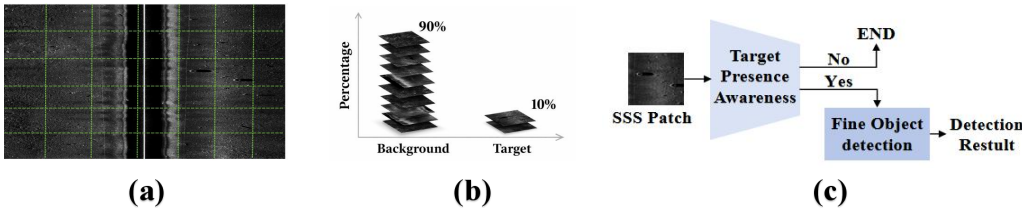

Figure 2. Patch-based SSS Image Object detection. (**a**) Original SSS Image. (**b**) Statistical distribution of background and target patches (**c**) The proposed target presence-aware coarse-to-fine detection strategy.

1.1. Traditional SSS Object Detection Related Works

Existing underwater object detection approaches can be broadly categorized into traditional signal-processing-based methods and learning-based methods. Early studies mainly focused on handcrafted feature extraction and statistical modeling techniques. Motion-based approaches, such as frame differencing and optical flow, have been widely used for detecting moving underwater targets, while morphological filtering techniques have been adopted to enhance target structures and suppress background interference.

Among these methods, constant false alarm rate (CFAR) detection has been one of the most commonly used techniques for sonar target detection, where pixel intensities are compared against adaptive thresholds derived from local background statistics [6]. To improve detection efficiency on high-resolution SSS images, ACA-CFAR [18] extends the conventional CA-CFAR scheme to a two-dimensional analysis framework inspired by integral image representations, significantly reducing computational cost. Based on the CA-CFAR paradigm, Villar et al. [19] achieved pipeline detection in sonar imagery, while Rahnemoonfar et al. [20] combined top-hat transformation and morphological operations for seagrass identification in sonar images.

Despite their simplicity and interpretability, traditional detection methods generally suffer from limited adaptability to varying underwater environments. Their performance is highly sensitive to manually selected thresholds and parameter settings, which are often tailored to specific scenarios. When applied to sonar images with low signal-to-noise ratios, non-uniform intensity distributions, and complex seabed backgrounds, these methods tend to exhibit degraded detection performance.

1.2. Deep Learning-Based SSS Object Detection Related Works

Deep learning approaches have increasingly dominated SSS image analysis due to their powerful feature learning capabilities. Early works explored bespoke convolutional architectures for sonar object detection; for instance, Song et al. [21] introduced a self-cascaded CNN for real-time underwater detection, and Palomeras et al. [22] proposed a

detection–classification framework with probabilistic grid mapping to reduce false positives in mine-like target detection.

Single-stage detectors such as YOLO variants have been adapted for sonar imagery, with Cao et al. [15] employing YOLOv8 for obstacle detection and Zhang et al. [16] using transfer learning with YOLOv10 for forward-looking sonar targets. Recently, Detection Transformers (DETR) [23] and their variants have emerged as a powerful paradigm for capturing long-range dependencies via global self-attention mechanisms. For instance, LUW-DETR [24] has been explored for sonar target recognition to overcome the limitations of local receptive fields in standard CNNs. However, while DETR-based architectures eliminate the need for hand-crafted components like non-maximum suppression (NMS), they often incur high computational costs and slow convergence, which are particularly problematic for real-time deployment on resource-constrained AUVs. Furthermore, the global attention mechanism inherent in DETR can act as a low-pass filter, potentially suppressing the high-frequency structural cues necessary for detecting small-scale mine-like objects.

Recent works extend beyond detection to segmentation and classification sub-tasks in sonar imagery, recognizing the importance of richer scene understanding. Lei et al. proposed SonarNet, a global feature-based hybrid attention network tailored for SSS image segmentation that combines dual encoders with adaptive hybrid attention for enhanced global and local feature fusion, significantly outperforming several state-of-the-art saliency detection baselines on dedicated sonar datasets [25]. Lei et al. also introduced Si-GAT, a graph structure-based network that explicitly integrates features from both bright target regions and their acoustic shadows to better capture spatial context and improve classification accuracy in SSS images [26]. Complementarily, Si et al. developed an unsupervised method to detect and segment shadow areas associated with sunken targets, addressing the challenge that shadows often carry distinct physical information yet are overlooked in typical supervised pipelines [27].

Other deep learning applications include approaches for structural inspection and target characterization, such as leveraging deep models for high-resolution sonar detection of local damage in underwater structures [28], and multi-scale fusion with efficient feature extraction to enhance sonar object detection performance [29]. Traditional methods continue to evolve, with strategies such as speckle reduction plus scene prior modeling demonstrating effectiveness for forward-looking sonar detection [30], and multilevel feature fusion networks like MLFFNet providing improved robustness on noisy sonar benchmarks [31]. Integrating attention mechanisms with YOLOv7 has also shown gains in target detection under SSS imaging conditions [32], while prior work on small target detection with forward-looking sonar has contributed foundational algorithmic insights [33,34].

Despite these advances, most existing deep learning methods still rely on dense inference paradigms that process all image regions equally and often lack explicit modeling of sonar-specific contextual priors such as acoustic shadow formation, range-dependent resolution variation, and geometric distortions intrinsic to SSS imaging. Moreover, segmentation, classification, and detection are frequently treated as separate tasks, leading to fragmented modeling of global context and target–shadow relationships. These limitations motivate the development of more efficient, physics-aware sonar vision frameworks that unify detection, segmentation, and contextual reasoning within a cohesive architecture.

2. Materials and Methods

This section presents the proposed target presence-aware SSS object detection framework in detail.

2.1. Overall Structure

The overall architecture of the proposed target presence-aware SSS object detection framework is illustrated in Figure 3. The framework is designed following a coarse-to-fine inference paradigm, explicitly exploiting the sparsity of targets in high-resolution SSS images to reduce redundant computation.

Given a high-resolution SSS image, the image is first divided into a set of local patches according to the strip-like acquisition geometry and memory constraints of AUV platforms. Instead of uniformly applying a computationally expensive detection network to all patches, the proposed framework introduces a lightweight TPA module as an early filtering stage. The TPA module rapidly evaluates each patch and estimates whether it potentially contains a target of interest. Only patches classified as target-positive by the TPA module are forwarded to the subsequent Object Forward Detection (OFD) module, which performs fine-grained object detection and localization. Patches identified as target-absent are discarded at an early stage, thereby avoiding unnecessary computation. By decoupling target presence estimation from fine-grained object detection, the proposed framework significantly reduces inference latency and energy consumption while maintaining competitive detection accuracy. Algorithm 1 presents the overall processing of the proposed target presence-aware SSS object detection framework.

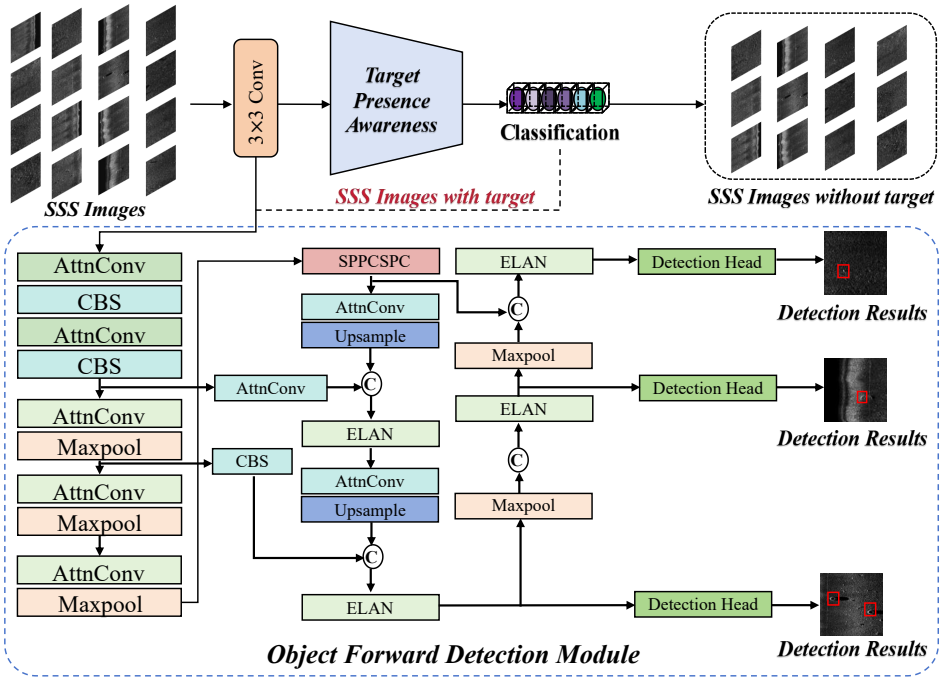

Figure 3. The overall structure of target presence-aware SSS object detection framework.

2.2. Target Presence Analysis Module

The Target Presence Analysis (TPA) module is designed to efficiently determine whether an input SSS image patch contains potential targets, while incurring minimal computational overhead. To enhance feature representation without sacrificing efficiency, each input patch is first processed by a 3×3 convolutional layer for preliminary feature extraction, and the resulting feature maps are subsequently fed into the TPA module.

To enable effective target presence perception under complex sonar backgrounds, as illustrated in Figure 4, the TPA module incorporates a spatial-frequency convolution fusion mechanism. By leveraging Fourier-domain operations, the module is able to efficiently capture global contextual information. According to Fourier theory, point-wise operations in the spectral domain correspond to global receptive fields in the spatial domain, allowing

long-range dependencies to be modeled with low computational cost. Specifically, given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, the TPA module processes it through two parallel branches operating in the spatial and frequency domains, respectively.

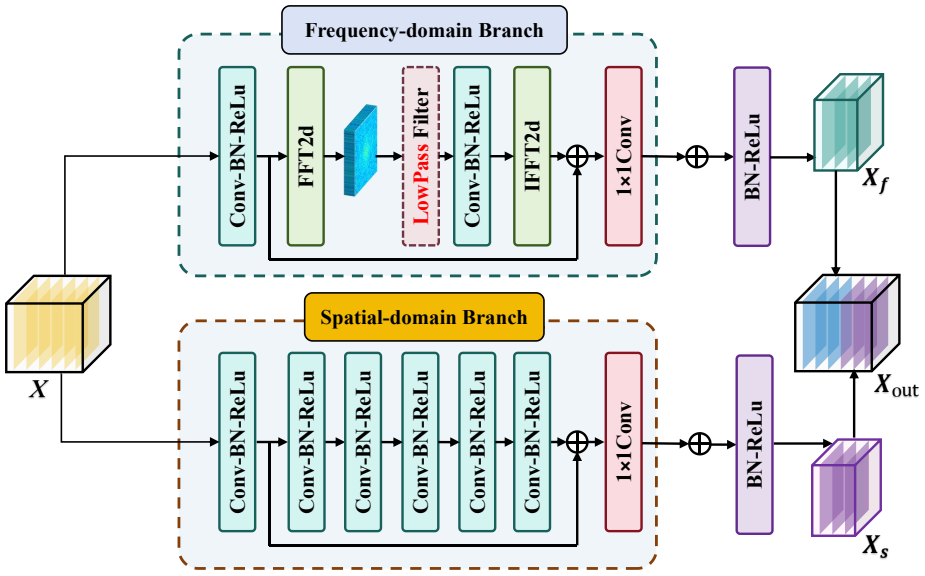

Figure 4. The detailed structure of TPA module.

Spatial-domain branch. The spatial branch aims to capture local structural cues using conventional convolutional operations, which can be formulated as:

$$X_{\mathrm{s}} = \sigma(W_s(X) + b_s), \tag{1}$$

where W_s and b_s represent the learnable convolution kernel and bias, respectively, and $\sigma(\cdot)$ denotes a nonlinear activation function.

Frequency-domain branch. To capture global contextual information efficiently, the frequency branch first transforms the input feature map into the spectral domain using the Fourier transform:

$$\hat{X} = \mathcal{F}(X), \tag{2}$$

where $\hat{X} \in \mathbb{C}^{H\times W\times C}$ denotes the complex-valued frequency representation.

Then, a convolution operation is applied in the frequency domain to modulate spectral components:

$$\hat{X}_{\mathrm{f}} = W_f \odot \hat{X}, \tag{3}$$

where $W_f \in \mathbb{C}^{H\times W\times C}$ denotes the learnable frequency-domain filter and $\odot$ represents element-wise multiplication. The modulated spectral features are subsequently transformed back to the spatial domain via the inverse Fourier transform:

$$X_{\mathrm{f}} = \mathcal{F}^{-1}(\hat{X}_{\mathrm{f}}). \tag{4}$$

Finally, the outputs of the spatial and frequency branches are fused to obtain the target presence-aware representation:

$$X_{\mathrm{out}} = X_{\mathrm{s}} + X_{\mathrm{f}}, \tag{5}$$

which integrates local spatial cues and global contextual information for efficient target presence estimation.

Algorithm 1: Target Presence-Aware Coarse-to-Fine Object Detection for Large-Scale SSS Images

Input: A large-scale SSS image $I \in \mathbb{R}^{H \times W}$

Output: Final detection set $\mathcal{B}$ in the original image coordinate system

Patch Preparation:

  Divide I into non-overlapping (or sliding) patches: $\mathcal{P} = \{P_k\}_{k=1}^{N}$, $P_k \in \mathbb{R}^{S \times S}$

  Initialize detection set $\mathcal{B} \leftarrow \emptyset$

Coarse Stage: Target Presence Analysis (TPA):

  for $k = 1$ **to** N **do**

    Shallow feature extraction: $X_k = Conv_{3\times3}(P_k)$

    Spatial branch: $F_k^{spa} = \sigma(W_s(X_k) + b_s)$

    Frequency branch: $\hat{X}_k = \mathcal{F}(X_k)$

      $\hat{F}_k^{freq} = W_f \odot \hat{X}_k$

      $F_k^{freq} = \mathcal{F}^{-1}(\hat{F}_k^{freq})$

    Spatial–frequency fusion: $F_k^{tpa} = F_k^{spa} + F_k^{freq}$

    Presence prediction: $p_k = Sigmoid(Head_{cls}(Pool(F_k^{tpa})))$

    if $p_k \geq \tau$ **then** mark P_k as target-positive

Fine Stage: Object Forward Detection (OFD):

  Collect target-positive patch set: $\mathcal{P}^+ = \{P_k \mid p_k \geq \tau\}$

  for each $P_k \in \mathcal{P}^+$ **do**

    Backbone with AttnConv: $Z_k = Backbone_{\text{AttnConv}}(P_k)$

    Neck feature aggregation: $Y_k = Neck(Z_k)$

    Detection head: $\mathcal{B}_k = Head_{det}(Y_k)$

2.3. Object Forward Detection Module

The Object Forward Detection (OFD) module is responsible for performing fine-grained object detection and localization on sonar patches identified as target-positive by the TPA module. As illustrated in the lower dashed box of Figure 3, the OFD module follows a high-performance architecture composed of an AttnConv-based backbone, a multi-scale feature fusion neck, and a decoupled detection head. To accurately model local target structures while maintaining linear computational complexity, the backbone utilizes the AttnConv module, which emulates self-attention mechanisms through a hybrid design of static and dynamic convolutions.

As detailed in Figure 5 the AttnConv module processes the input feature $X \in \mathbb{R}^{H \times W \times C}$ by first performing a Split operation along the channel dimension. The resulting features are then forwarded through two parallel paths:

$$Y = \text{Static}(X_{split}) + \text{Dynamic}(X_{split}) \tag{6}$$

Static Path. One branch focuses on capturing stable local structural information. As shown in the left part of Figure 5, it employs a sequence of 3×3 Conv, Static Large Kernel (LK) Conv, and 1×1 Conv layers, which are subsequently aggregated via a Concate operation to preserve translation equivariance and fine-grained textures.

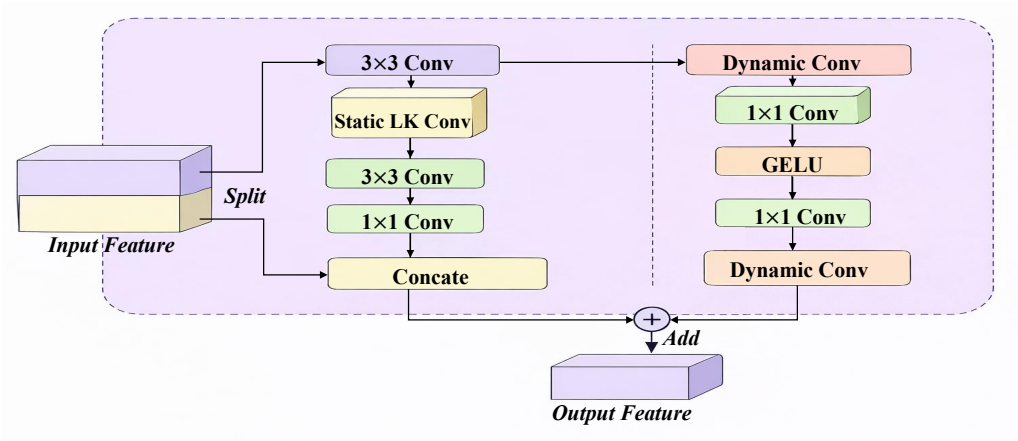

Figure 5. The detailed structure of the AttnConv module.

Dynamic Path. The other branch enables long-range dependency modeling by generating content-aware dynamic kernels. Specifically, as visualized in the right part of Figure 5, it utilizes Dynamic Conv layers interleaved with 1×1 convolutions and GELU activation functions. This modulation is expressed as:

$$X' = \mathcal{A}(X) \odot (W_{dyn} * X), \tag{7}$$

where $\mathcal{A}(X)$ represents the content-aware attention weights derived from the input context.

Through this dual-path design, AttnConv effectively approximates global representation power with $\mathcal{O}(n)$ complexity. Following the backbone, the neck network aggregates multi-scale features from various stages, including the SPPCSPC and ELAN blocks shown in Figure 3, to integrate complementary contextual information. The feature fusion process in the neck is designed to jointly encode local details and large-scale scene characteristics. By integrating this early-stage information, the neck compensates for the loss of spatial resolution in deep layers and enhances feature representation. Finally, as depicted on the right side of Figure 3, the aggregated features are forwarded to the Detection Heads to predict the final bounding boxes and class probabilities. The coarse-to-fine transition ensures that computational resources are concentrated on refining the localization of potential targets rather than processing vast, target-absent background regions.

2.4. Loss Function

The target presence analysis and object detection tasks are optimized using separate loss functions. The TPA module is formulated as a binary classification task and trained using the binary cross-entropy loss [35]:

$$\mathcal{L}_{\text{cls}} = -(y \log p + (1 - y) \log(1 - p)), \tag{8}$$

where p and y denote the predicted target presence probability and the corresponding ground-truth label, respectively. To mitigate the severe class imbalance between target-present and target-absent patches, class-balanced weighting is applied to encourage recall-oriented target presence estimation.

Following several one-stage detection methods [36], the object detection task consists of a bounding box regression loss, an objectness loss, and a classification loss:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{box}} + \lambda_{\text{obj}} \mathcal{L}_{\text{obj}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}. \tag{9}$$

3. Results

3.1. Experimental Datasets

In this section, we evaluate the effectiveness and efficiency of the proposed target presence-aware SSS object detection framework. Experiments are conducted on two representative public SSS datasets covering different underwater target types and acquisition scenarios.

(1) SSS-Mine Dataset. The SSS-Mine dataset [37] consists of 1170 real-world SSS images collected between 2010 and 2021. The data acquisition was performed by a Teledyne Marine Gavia Autonomous Underwater Vehicle (AUV), a modular maritime platform widely deployed for mine countermeasure (MCM) and search-and-recovery operations. This dataset is specifically designed for underwater mine detection and classification tasks, providing sufficient visual information to distinguish between mine-like contacts and non-mine-like bottom objects (NOMBO). In practical mine-hunting scenarios, many target-like objects are intentionally excluded from the ground truth to minimize false alarms.

(2) AI4Shipwrecks Dataset. The AI4Shipwrecks dataset [38] focuses on shipwreck detection and underwater archaeological exploration. It contains 286 high-resolution SSS images collected from 24 distinct shipwreck sites within the Thunder Bay National Marine Sanctuary (TBNS), using the EdgeTech 2205 SSS system. A distinctive feature of the AI4Shipwrecks dataset is its emphasis on large-scale shipwreck structures with complex geometric patterns. Each image is annotated at the pixel level for segmentation tasks, enabling precise localization of shipwreck regions. In our experiments, the segmentation annotations are converted into bounding box annotations to support object detection evaluation.

For both datasets, the data samples are split into training and testing sets with a ratio of 7:3. To avoid data leakage caused by spatial correlation, the split is performed at the scene or survey-segment level rather than at the image level. After dataset splitting, each SSS image is divided into fixed-size patches of 300×300 pixels, which are used as the inputs to the detection model. The corresponding annotations are processed in the same manner, and labels are assigned to each patch according to the targets contained within it. This patch-based processing strategy is consistent with practical AUV deployment scenarios and enables efficient training and inference on large-scale SSS imagery. Table 1 presents the pixel-level object distribution of the two SSS datasets.

Table 1. Pixel-level object distribution of the SSS datasets.

Dataset	Object Category	Pixel Count	Pixel Proportion
AI4Shipwreck [38]	Wreck ship	6,428,596	4.91%
SSS-Mine [37]	Mine	76,167	0.23%

3.2. Implementation Details and Evaluation Metrics

The proposed MobileSonar network is implemented using the PyTorch 1.9 framework [39] and trained on a workstation equipped with an Intel i7-10400F CPU, an NVIDIA RTX 4090 GPU, 32 GB RAM, and the Ubuntu 16.04 operating system. All models are trained for 200 epochs using the Adam optimizer [40] with an initial learning rate of 1×10^{-3}. A cosine learning rate decay strategy [41] is adopted throughout training, and the momentum coefficient is set to 0.943. The batch size is fixed to 8. For both training and inference, input sonar image patches are resized to 300×300 pixels. To enhance detection robustness under complex seabed conditions, Mosaic data augmentation [42] is applied during detector training. For a fair comparison, all baseline methods are trained using the same input resolution and identical data augmentation strategies. For performance evaluation, different metric

sets are adopted according to task characteristics. For the object detection task [43], we report Precision (P), Recall (R), False Alarm Rate (FAR), Missed Detection Rate (MDR), and mean average precision (mAP) under multiple Intersection-over-Union (IoU) thresholds. The definitions of FAR and MDR metrics are given as follows:

$$\mathrm{FAR} = \frac{\sum_{j=0, j\neq i}^{k} p_{ij}}{\sum_{j=0}^{k} p_{ij}}, \tag{10}$$

$$\mathrm{MDR} = \frac{\sum_{j=0, j\neq i}^{k} p_{ji}}{\sum_{j=0}^{k} p_{ji}}, \tag{11}$$

where k represents the number of classes, and p_{ij} denotes the number of pixels belonging to class j that are predicted as class i. Similarly, p_{ii}, p_{ij}, p_{ji} are commonly referred to as true positive, false positive, and false negative, respectively.

In addition, the mean Average Precision (mAP) is calculated by averaging the AP values over all k object categories:

$$mAP = \frac{1}{k} \sum_{i=1}^{k} AP_i. \tag{12}$$

where AP_i denotes the average precision for each class. The metrics mAP_{35}, mAP_{50}, and mAP_{75} represent the mAP values calculated at Intersection-over-Union (IoU) thresholds of 0.35, 0.50, and 0.75, respectively.

For the target presence classification task, Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC) are employed as evaluation metrics. In addition, to rigorously assess the practical utility of the proposed method, inference efficiency is evaluated on a representative hardware platform: an NVIDIA Jetson Orin Nano embedded platform (ARM Cortex-A78AE architecture) for edge deployment. Specifically, the NVIDIA Jetson Orin series has been widely adopted for deep-learning model edge computing in contemporary autonomous underwater vehicle (AUV) systems, owing to its optimal balance between AI inference throughput and strict power constraints. By benchmarking on these platforms, we ensure that the experimental environment closely mirrors the computational capabilities and architectural constraints of actual deployed underwater robotic systems.

3.3. Quantitative Analysis for Object Detection

To demonstrate the effectiveness of the proposed target presence-aware joint detection framework, we conduct comprehensive quantitative evaluations on public SSS datasets and compare our method with several representative lightweight object detection approaches. Specifically, the compared methods include Fast R-CNN [44], YOLOv5-s [45], YOLOv8n [46], YOLOv8s [46], YOLOv10m [47], YOLOv11m [48], LUW-DETR [24], YOLO-Sonar [29], YOLO2026m [49], YOLO-Master-S [50], as well as the proposed method. Notably, all the baseline models and our model are trained with their recommended training strategies. Figure 6 presents the training accuracy curves of each model on SSS dataset.

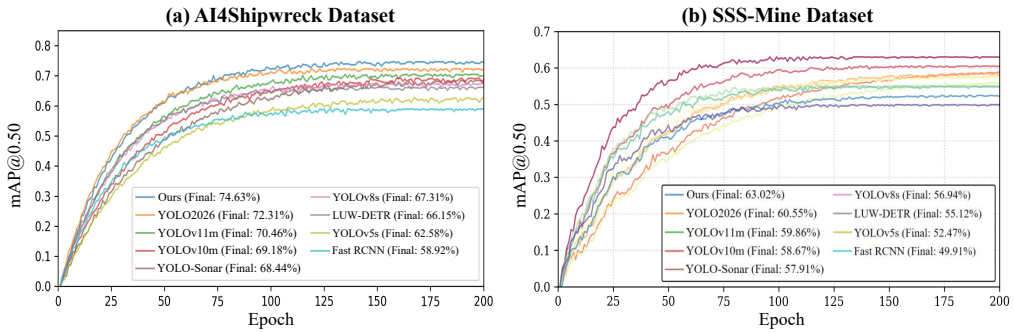

Figure 6. The training accuracy curves of each model on SSS dataset: (**a**) AI4Shipwreck dataset. (**b**) SSS-Mine dataset.

In Table 2, we present the quantitative evaluation results on the SSS-Mine dataset. As shown, the proposed method achieves the best overall performance across all evaluation metrics. In particular, our method attains the highest recall (67.18%) and the lowest missed detection rate (MDR) (32.82%), indicating a stronger ability to identify target-present regions under severe background clutter. Compared with the strongest baseline YOLOv11m, the proposed framework improves recall while simultaneously reducing both FAR and MDR, demonstrating that early-stage target presence analysis effectively suppresses background-induced false detections without sacrificing detection completeness.

Table 2. Performance comparison on SSS-Mine dataset using different detection methods.

Method	Precision (%)	Recall (%)	FAR (%) ↓	MDR (%) ↓	mAP_{35} (%)	mAP_{50} (%)	mAP_{75} (%)
Fast R-CNN	62.41	55.02	37.59	44.98	58.76	49.91	26.94
YOLOv5-s	65.13	57.94	34.87	42.06	61.82	52.47	30.81
YOLOv8n	66.08	59.88	33.92	40.12	63.97	54.86	33.02
YOLOv8s	67.54	61.82	32.46	38.18	65.79	56.94	35.11
YOLOv10m	68.96	63.04	31.04	36.96	66.98	58.67	37.24
YOLOv11m	69.84	64.02	30.16	35.98	67.91	59.86	38.71
LUW-DETR	66.42	60.15	33.58	39.85	64.21	55.12	34.05
YOLO-Sonar	67.88	62.34	32.12	37.66	65.92	57.34	36.12
YOLO2026m	70.21	65.11	29.79	34.89	68.45	60.55	40.23
YOLO-Master-S	68.12	61.56	31.88	38.44	66.34	57.91	35.88
Ours	**72.31**	**67.18**	**27.69**	**32.82**	**70.46**	**63.02**	**44.16**

Table 3 further compares the quantitative detection results on AI4Shipwrecks dataset. Different from SSS-Mine dataset, AI4Shipwrecks mainly consists of large-scale shipwreck structures with complex geometric patterns. As shown in the table, the proposed method consistently outperforms all baseline detectors in terms of precision, recall, and average precision metrics. Notably, our method achieves the highest AP_{75} value (51.87%), which is significantly higher than that of YOLOv11m.

Table 4 summarizes the model complexity and inference speed of different object detection methods. Although the proposed method has a moderate number of parameters, it achieves a significantly higher inference speed of 174.74 FPS, far exceeding all compared baseline models. This remarkable efficiency gain mainly stems from the target presence-aware coarse-to-fine inference strategy, which filters out a large number of target-absent patches at an early stage and avoids redundant computation in the detection module. Figure 7 presents the visualization comparison for the efficiency of detection models.

Table 3. Performance comparison on AI4Shipwrecks dataset using different detection methods.

Method	Precision (%)	Recall (%)	FAR (%) ↓	MDR (%) ↓	mAP_{35} (%)	mAP_{50} (%)	mAP_{75} (%)
Fast R-CNN	73.18	64.63	26.82	35.37	69.47	58.92	31.84
YOLOv5-s	76.54	68.27	23.46	31.73	73.41	62.58	36.03
YOLOv8n	77.42	70.08	22.58	29.92	75.06	64.83	38.47
YOLOv8s	78.96	72.61	21.04	27.39	77.18	67.31	41.26
YOLOv10m	80.21	73.97	19.79	26.03	78.72	69.18	43.76
YOLOv11m	81.03	75.09	18.97	24.91	79.61	70.46	45.42
LUW-DETR	78.15	71.34	21.85	28.66	76.02	66.15	39.82
YOLO-Sonar	79.44	73.52	20.56	26.48	77.85	68.44	42.15
YOLO2026m	81.65	76.84	18.35	23.16	80.52	72.31	48.96
YOLO-Master-S	78.62	72.18	21.38	27.82	76.54	66.89	40.54
Ours	**82.57**	**78.74**	**17.43**	**21.26**	**82.08**	**74.63**	**51.87**

Table 4. Model complexity and inference speed comparison.

Method	Params (M)	GPU Inference (FPS)	CPU Inference (FPS)
Fast R-CNN	34.21	14.6	1.2
YOLOv5-s	17.15	42.4	3.8
YOLOv8n	12.68	38.7	3.5
YOLOv8s	13.72	36.2	2.9
YOLOv10m	19.30	24.8	2.1
YOLOv11m	22.44	28.1	2.5
LUW-DETR	74.35	12.4	0.5
YOLO-Sonar	25.42	20.8	2.6
YOLO2026m	20.40	31.2	4.1
YOLO-Master-S	29.15	43.2	3.9
Ours	**24.85**	**174.74**	**15.6**

Figures 8 and 9 present the visualization results of different detection methods on the SSS-Mine and AI4Shipwrecks dataset. In the SSS-Mine dataset, characterized by intense background noise and low-contrast environments, traditional detectors such as YOLOv5-s and YOLOv8s frequently misclassify seabed rocks or shadows as targets; in contrast, our proposed method effectively captures global context through the spatial-frequency fusion mechanism of the TPA module, thereby significantly suppressing false alarms. When detecting the complex geometric structures within the AI4Shipwrecks dataset, where YOLOv11m is prone to missed detections or imprecise localization, our method leverages the high-frequency feature enhancement capabilities of the AttnConv dynamic kernels within the OFD module to achieve more complete bounding box regression and fine-grained localization.

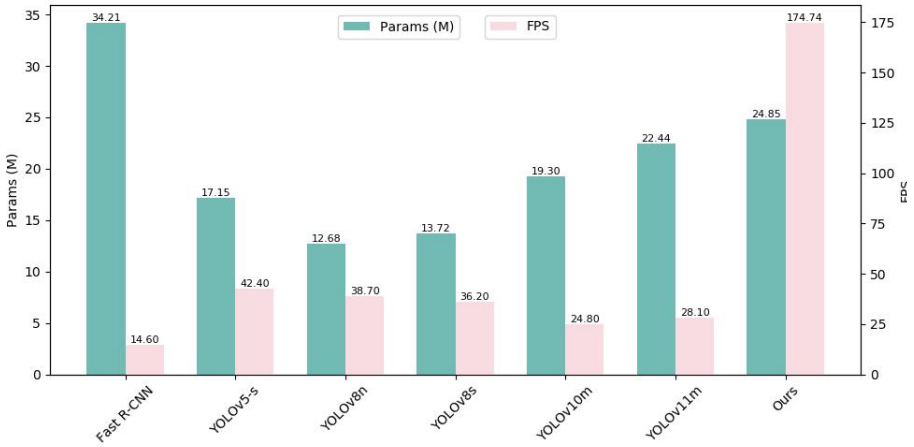

Figure 7. Visualization comparison for the efficiency of detection models.

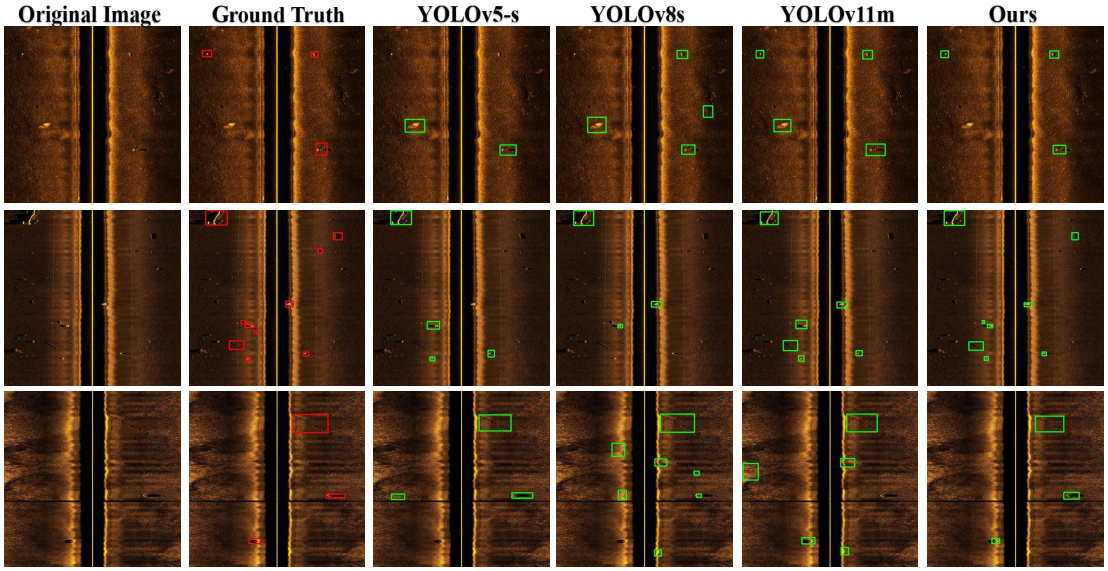

Figure 8. Visual comparison of object detection results on SSS-Mine dataset.

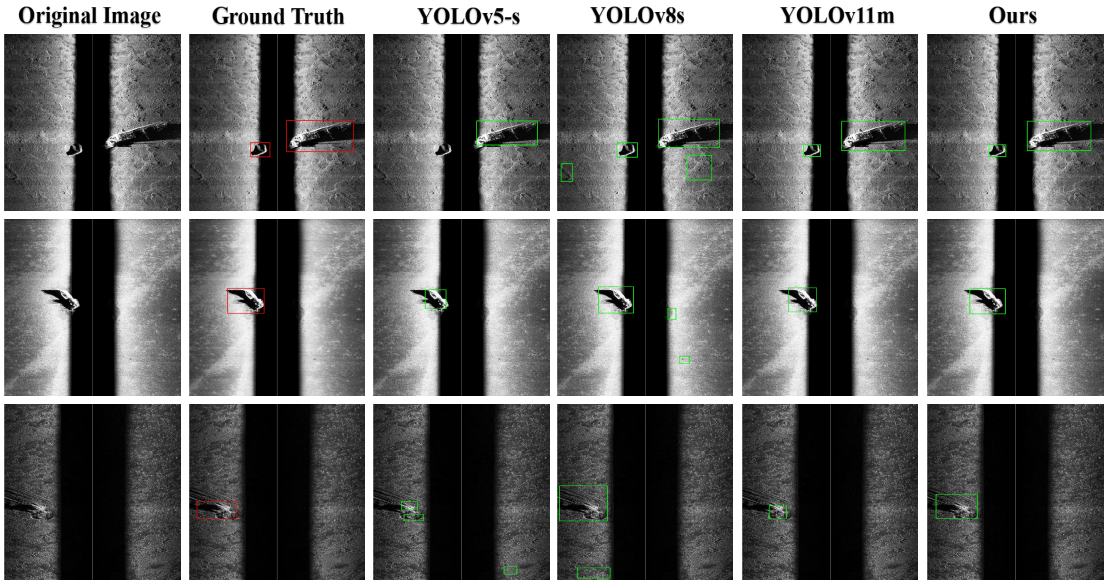

Figure 9. Visual comparison of object detection results on AI4Shipwrecks dataset.

3.4. Quantitative Analysis for Target Presence

To quantitatively evaluate the effectiveness of the proposed TPA module, we conduct a dedicated target presence classification experiment. The proposed TPA module is compared with commonly used convolutional classification backbones, including MobileNetv3 [51] and ResNet18 [52]. For a fair comparison, all models are trained and evaluated on the same patch-based dataset split.

Table 5 presents the quantitative target presence classification results on different datasets. As shown, the proposed TPA module consistently outperforms MobileNet and ResNet-based baselines across all evaluation metrics. On AI4Shipwrecks dataset, the TPA module achieves the highest accuracy, recall, and AUC, indicating superior target presence discrimination capability. On more challenging SSS-Mine dataset, the proposed TPA module maintains higher recall and F1-score than the compared methods, demonstrating improved robustness under strong background clutter and class imbalance. The proposed TPA module provides a favorable balance between classification performance and efficiency, making it well-suited as an early-stage filtering component in the proposed coarse-to-fine SSS object detection framework. Figure 10 presents the visualization results of feature map for target classification.

Table 5. Classification performance comparison on different datasets.

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
AI4Shipwrecks	MobileNetv3	96.84	93.52	95.63	95.53	96.2
	ResNet18	97.47	96.61	96.28	96.12	98.1
	TPA (Ours)	**98.85**	**97.14**	**97.92**	**96.51**	**99.3**
SSS-Mine	MobileNet	93.26	94.03	93.91	93.92	94.8
	ResNet18	94.73	94.65	93.14	95.36	96.7
	TPA (Ours)	**95.62**	**94.87**	**95.48**	**96.16**	**98.5**

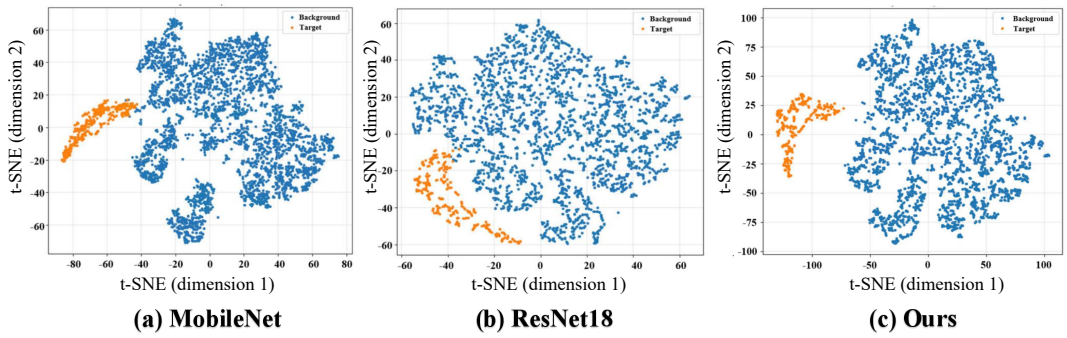

Figure 10. Visualization results of feature map for target classification.

3.5. Ablation Analysis

To verify the effectiveness of each component in the proposed object detection method, we conducted ablation experiments with different configurations.

(1) **Component contribution of TPA Module**. Table 6 reports the ablation results of the TPA module with different combinations of spatial and frequency convolution. When only the frequency convolution branch is used, the performance is limited, indicating insufficient discriminative capability without local spatial modeling. In contrast, using only spatial convolution yields significantly better results, demonstrating the importance of local structural cues for target presence estimation. By jointly integrating spatial and frequency convolution, the full TPA module achieves the best performance across all metrics on both datasets. In particular, improvements in F1-score and AUC indicate that spatial–frequency

fusion effectively enhances target presence discrimination and robustness under complex sonar backgrounds.

Table 6. Ablation study on spatial and frequency convolution modules for target classification.

Method	AI4Shipwrecks			SSS-Mine		
	Accuracy (%)	F1-Score (%)	AUC (%)	Accuracy (%)	F1-Score (%)	AUC (%)
Frequency Baseline	94.42	93.08	0.946	92.13	91.62	94.8
Spatial Baseline	98.01	95.67	0.989	94.88	95.21	98.1
TPA (Ours)	**98.85**	**96.51**	**0.993**	**95.62**	**96.16**	**0.985**

In addition, Table 7 reports the ablation results of the TPA module under different branch configurations. When only the frequency convolution branch is used (Frequency Baseline), the model exhibits the highest False Alarm Rate (FAR) and Missed Detection Rate (MDR) across both datasets. On the SSS-Mine dataset, the AP_{50} drops to 54.12%, indicating that relying solely on spectral information is insufficient for discriminating targets from complex seabed clutter without local structural modeling. The Spatial Baseline yields significantly better results than the frequency-only variant, with AP_{50} increasing by approximately 3.42% and 5.23% on the AI4Shipwrecks and SSS-Mine datasets, respectively. This confirms that local geometric features, such as target edges and acoustic shadows, are primary descriptors for SSS object detection. Effectiveness of Spatial-Frequency Fusion: By jointly integrating spatial and frequency convolutions, the full TPA module achieves the state-of-the-art performance across all metrics. Specifically, the fused model reduces the FAR to 17.43% on AI4Shipwrecks and the MDR to 32.82% on SSS-Mine.

Table 7. Ablation study on spatial and frequency convolution modules for target detection.

Method	AI4Shipwrecks			SSS-Mine		
	FAR (%) ↓	MDR (%) ↓	mAP_{50}(%) ↑	FAR (%) ↓	MDR (%) ↓	mAP_{50}(%) ↑
Frequency Baseline	22.15	28.34	68.52	35.42	42.18	54.12
Spatial Baseline	19.86	23.55	71.94	31.08	36.47	59.35
TPA (Ours)	**17.43**	**21.26**	**74.63**	**27.69**	**32.82**	**63.02**

(2) **Component contribution of OFD module.** Table 8 presents the cross-dataset ablation results of the Object Forward Detection (OFD) module with and without the AttnConv component. As shown, removing AttnConv leads to consistent performance degradation on both AI4Shipwrecks and SSS-Mine datasets. Specifically, the full model achieves lower FAR and MDR while yielding higher mAP_{50} compared with the variant without AttnConv, indicating improved robustness against false alarms and missed detections. The performance gains are observed consistently across datasets with different target scales and scene characteristics, demonstrating that AttnConv effectively enhances fine-grained feature representation and localization capability.

Table 8. Ablation study on Attn modules for target classification.

Method	AI4Shipwrecks			SSS-Mine		
	FAR (%) ↓	MDR (%) ↓	mAP_{50} (%) ↑	FAR (%) ↓	MDR (%) ↓	mAP_{50} (%) ↑
Removing AttnConv	19.08	23.14	72.81	29.96	35.27	60.74
Ours	**17.43**	**21.26**	**74.63**	**27.69**	**32.82**	**63.02**

To further validate the enhancement of feature representation capabilities provided by the AttnConv module, we visualized the feature maps at various stages of the Object Forward Detection (OFD) module using heatmaps (Figure 11). By comparing panels (b) with (c) and (d) with (e), it can be observed that prior to the incorporation of AttnConv, the responses within the feature maps (b, d) appear relatively diffuse, with target signals often becoming obscured amidst the complex sonar background noise. Enhanced Target Response: Following the introduction of the AttnConv module, the feature maps (c, e) exhibit stronger and more concentrated activation responses within target regions (e.g., shipwreck structures or the edges of mines). This demonstrates that AttnConv, through its content-aware dynamic kernels, effectively simulates a self-attention mechanism, thereby reinforcing the extraction of high-frequency target features.

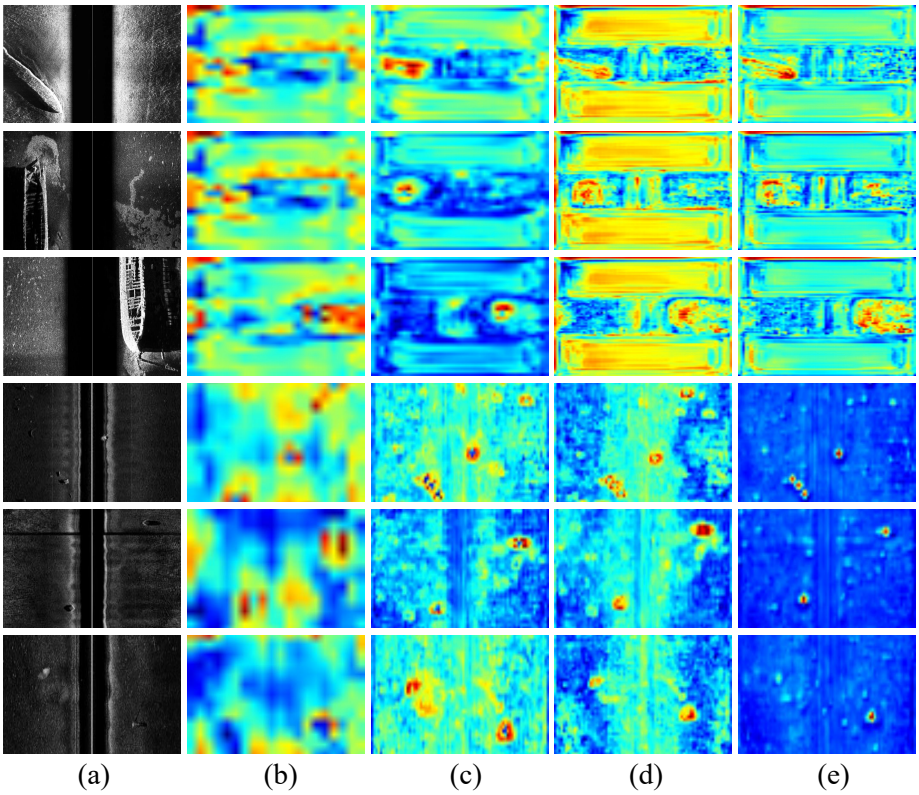

Figure 11. Visualization heat map of SSS feature. (**a**) SSS image inputs, (**b**) OFD feature 1, (**c**) OFD feature 1 With AttnConv, (**d**) OFD feature 2, (**e**) OFD feature 2 With AttnConv.

(3) **Component contribution of target presence awareness**. Table 9 reports the inference efficiency of different configurations of the proposed framework. As shown, the efficiency difference is driven by the strategic bypass of redundant computations rather than model complexity. Directly forwarding all patches to the fine-grained detection stage (OURS with target) results in a significant drop in inference speed as resolution increases. At 640 × 640, the speed plunges to only 5.4 FPS, which fails to meet the real-time requirements of long-duration autonomous missions. When only the lightweight processing is involved (OURS without target), the framework maintains a high frame rate, demonstrating that the spatial-frequency fusion mechanism in the TPA module is highly efficient even at higher pixel densities. By integrating the target presence-aware inference strategy (OURS), the proposed framework achieves a substantial speed improvement compared with dense detection while maintaining full detection capability. At 640 × 640, our method maintains 126.30 FPS, providing a 23.4× speedup over the dense baseline. These results confirm that the proposed method effectively exploits the sparsity of underwater targets to eliminate

redundant computation on background-dominated regions. This makes the framework particularly suitable for real-time deployment on resource-constrained AUV platforms where low latency is critical.

Table 9. Inference efficiency comparison of different configurations under various input resolutions.

Input Resolution	OURS Without Target (FPS)	OURS with Target (FPS)	OURS (FPS)
300 × 300 pixels	201.6	21.6	174.74
416 × 416 pixels	182.4	12.8	158.20
640 × 640 pixels	145.2	5.4	126.30

(4) **Sensitivity Analysis of the TPA Threshold τ.** To investigate the influence of the TPA threshold τ on the overall framework performance, we conduct a sensitivity experiment by varying τ from 0.1 to 0.9. This hyperparameter is central to the trade-off between inference efficiency and detection accuracy. The experimental results, visualized in Figure 12, demonstrate how the threshold dictates the proportion of patches forwarded to the fine-grained OFD module. As shown in Figure 12a, when $\tau = 0.0$, the system performs dense inference on all patches, resulting in a significantly lower speed of approximately 21.6 FPS. As τ increases, the TPA module filters out more target-absent background patches, leading to a substantial boost in throughput. At the theoretical limit of $\tau = 1.0$, where all patches are discarded, the speed reaches its peak of 201.6 FPS.

Figure 12b illustrates that while efficiency improves with a higher threshold, the risk of missing potential targets also escalates. Between $\tau = 0.0$ and $\tau = 0.7$, the Missed Detection Rate (MDR) remains relatively stable, rising only marginally across both datasets. However, once τ exceeds 0.7, the MDR for the SSS-Mine dataset escalates sharply from 32.82% towards 48.22%, indicating that critical target-positive patches are being misclassified as background. Based on these trends, we identified $\tau = 0.7$ as the optimal value for the proposed framework. At this point, the system delivers ultra-fast performance (174.74 FPS) while maintaining a competitive AP_{50} and a low MDR. This setting allows the AUV to effectively exploit the spatial sparsity of underwater targets to reduce redundant computation without compromising the reliability of perception.

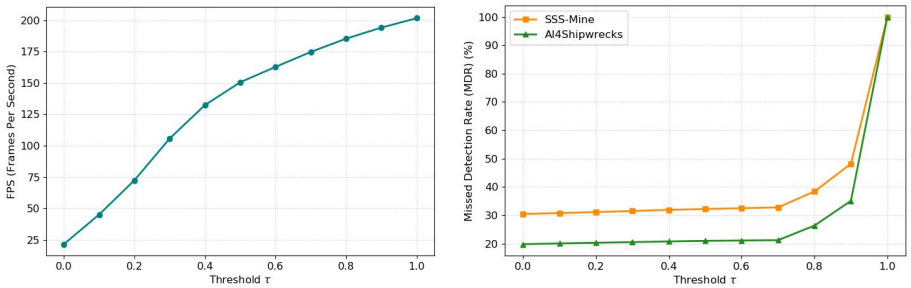

Figure 12. Sensitivity analysis of the TPA threshold τ. (**a**) Evolution of inference speed (FPS) on the NVIDIA Jetson Orin Nano platform. (**b**) Impact of varying τ on the Missed Detection Rate (MDR) for SSS-Mine and AI4Shipwrecks datasets.

4. Discussion

4.1. Addressing Computational Redundancy via Content-Dependent Inference

Real-time object detection in large-scale SSS images presents a unique set of challenges, primarily characterized by the extreme spatial sparsity of targets and the restricted computational payloads of autonomous underwater vehicles (AUVs). While contemporary

lightweight detectors, such as YOLOv10m and YOLOv11m, have made strides in balancing accuracy and speed (Table 1), they fundamentally rely on dense inference paradigms. These models process every image patch uniformly, leading to massive computational redundancy in background-dominated seabed environments where meaningful targets typically occupy less than 10% of the area. Our proposed framework elegantly circumvents this by decoupling the detection task into a coarse-to-fine pipeline, shifting the paradigm from being resolution-dependent to being content-dependent.

4.2. Exploiting Frequency-Domain Priors for Efficient Gating

The TPA module acts as an ultra-lightweight gatekeeper, utilizing a spatial-frequency fusion mechanism to filter out over 90% of target-absent regions. From a systematic perspective, the dominant trend of expanding receptive fields via Transformer-based architectures often incurs prohibitive latency on edge devices. Unlike traditional global self-attention with quadratic complexity, the frequency-domain branch in TPA models long-range contextual dependencies with linear complexity by leveraging Fourier theory. As revealed in our complexity comparison (Table 3), while YOLOv11m achieves a respectable 28.1 FPS, it remains bottlenecked by the need to execute full-network forward passes on empty background patches. By utilizing the TPA module, our framework explicitly exploits target sparsity to drastically reduce unnecessary deep neural network activations.

4.3. Synergy Between Feature Enhancement and Localization Accuracy

The synergy between the TPA and OFD modules is critical for maintaining high recall under severe background clutter. Preserving high-frequency target cues in the early stage is insufficient if the subsequent detector cannot accurately localize small structures. The AttnConv module in the OFD stage bridges this gap by emulating self-attention behavior through content-aware dynamic kernels. By combining shared-weight spatial convolutions for translation equivariance with dynamic 1×1 modulations, AttnConv strengthens high-frequency target features without the $\mathcal{O}(n^2)$ complexity overhead. Our ablation study (Table 6) confirms that this hybrid design is essential: removing the AttnConv component leads to a consistent drop in AP_{50} across both the SSS-Mine and AI4Shipwrecks datasets, particularly increasing the Missed Detection Rate (MDR) for complex geometries.

4.4. Trade-Off Analysis and Embedded Deployment Potential

A pivotal finding in our research is the sensitivity of the efficiency–accuracy trade-off to the TPA threshold τ (Figure 10). Setting $\tau = 0.7$ provides an optimal equilibrium, achieving a 23.4$\times$ speedup over dense detection while maintaining an MDR comparable to dense inference. This indicates that our framework strategically reallocates computational "attention" to high-probability regions. Regarding real-world deployment, the framework achieves ultra-fast speeds (174.74 FPS) on the NVIDIA Jetson Orin Nano and maintains 15.6 FPS on CPU-only platforms, where baselines like YOLOv11m drop to 2.5 FPS. This performance allows AUVs to process sonar patches in burst mode, reducing the average hardware duty cycle and overall energy consumption—a significant advantage for long-duration autonomous missions.

4.5. Limitations and Future Perspectives

Despite these advantages, our approach has certain limitations. First, the spatial-frequency fusion in the TPA module assumes a relatively consistent seabed texture; in highly heterogeneous environments with sudden geological transitions, the frequency-domain branch may produce transient false positives. Second, the current module does not yet incorporate cross-patch temporal consistency from sequential sonar pings. Future work will explore noise-aware dynamic thresholds and the integration of temporal motion priors

to further enhance robustness against sensor noise and non-uniform intensity distributions in extreme deep-sea conditions.

5. Conclusions

This paper presents a target presence-aware side-scan sonar object detection framework designed for real-time deployment in resource-constrained AUV environments. By explicitly exploiting the sparsity of underwater targets, the proposed framework decouples target presence analysis from fine-grained object detection and adopts a coarse-to-fine inference strategy to eliminate redundant computation on background-dominated regions. A lightweight Target Presence Analysis module is introduced to efficiently identify target-positive patches using spatial–frequency feature fusion, while an AttnConv-enhanced Object Forward Detection module is employed to improve fine-grained detection by strengthening high-frequency target cues. Experimental results on benchmark SSS datasets demonstrate that the proposed approach achieves a favorable balance between detection accuracy, inference speed, and computational efficiency, outperforming conventional dense-processing methods in terms of real-time feasibility. Future work will focus on extending the proposed framework to more complex underwater scenarios, including cross-domain generalization, as well as further optimizing the framework for real-world deployment.

Author Contributions: Conceptualization, G.X.; methodology, G.X. and H.X.; validation, Y.Y. and H.X.; writing—original draft preparation, G.X.; writing—review and editing, G.X.; visualization, J.H.; supervision, Y.Y.; project administration, G.P. and Y.Y.; funding acquisition, G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Foundation of China under Grant 62571448, and in part by National Key Research and Development Program under Grant 2021YFC-2803000 and 2021YFC2803001.

Data Availability Statement: No new data.

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

References

1. Choi, H.-m.; Yang, H.-s.; Seong, W.-j. Compressive underwater sonar imaging with synthetic aperture processing. *Remote Sens.* **2021**, *13*, 1924. [CrossRef]
2. Li, L.; Li, Y.; Wang, H.; Yue, C.; Gao, P.; Wang, Y.; Feng, X. Side-scan sonar image generation under zero and few samples for underwater target detection. *Remote Sens.* **2024**, *16*, 4134. [CrossRef]
3. Nga, Y.Z.; Rymansaib, Z.; Anthony Treloar, A.; Hunter, A. Automated recognition of submerged body-like objects in sonar images using convolutional neural networks. *Remote Sens.* **2024**, *16*, 4036. [CrossRef]
4. Peng, Y.; Li, H.; Zhang, W.; Zhu, J.; Liu, L.; Zhai, G. Underwater sonar image classification with image disentanglement reconstruction and zero-shot learning. *Remote Sens.* **2025**, *17*, 134. [CrossRef]
5. Sun, Y.; Zheng, H.; Zhang, G.; Ren, J.; Xu, H.; Xu, C. DP-ViT: A dual-path vision transformer for real-time sonar target detection. *Remote Sens.* **2022**, *14*, 5807. [CrossRef]
6. Zheng, L.; Tian, K. Detection of small objects in sidescan sonar images based on POHMT and Tsallis entropy. *Signal Process.* **2018**, *142*, 168–177. [CrossRef]
7. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; CVPR 2001; IEEE: New York, NY, USA, 2001; Volume 1, p. I–I.
8. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*; IEEE: New York, NY, USA, 2005; Volume 1, pp. 886–893.
10. Rhinelander, J. Feature extraction and target classification of side-scan sonar images. In *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*; IEEE: New York, NY, USA, 2016; pp. 1–6.

11. 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]
12. Yang, N.; Li, G.; Wang, S.; Wei, Z.; Ren, H.; Zhang, X.; Pei, Y. SS-YOLO: A lightweight deep learning model focused on side-scan sonar target detection. *J. Mar. Sci. Eng.* **2025**, *13*, 66. [CrossRef]
13. Song, Y.; He, B.; Liu, P.; Yan, T. Side scan sonar image segmentation and synthesis based on extreme learning machine. *Appl. Acoust.* **2019**, *146*, 56–65. [CrossRef]
14. Le, H.T.; Phung, S.L.; Chapple, P.B.; Bouzerdoum, A.; Ritz, C.H.; Tran, L.C. Deep gabor neural network for automatic detection of mine-like objects in sonar imagery. *IEEE Access* **2020**, *8*, 94126–94139. [CrossRef]
15. Cao, X.; Ren, L.; Sun, C. Research on obstacle detection and avoidance of autonomous underwater vehicle based on forward-looking sonar. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 9198–9208. [CrossRef]
16. Zhang, H.; Tian, M.; Shao, G.; Cheng, J.; Liu, J. Target detection of forward-looking sonar image based on improved YOLOv5. *IEEE Access* **2022**, *10*, 18023–18034. [CrossRef]
17. 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]
18. Acosta, G.G.; Villar, S.A. Accumulated CA–CFAR process in 2-D for online object detection from sidescan sonar data. *IEEE J. Ocean. Eng.* **2014**, *40*, 558–569. [CrossRef]
19. Villar, S.A.; Acosta, G.G.; Senna, A.S.; Rozenfeld, A. Pipeline detection system from acoustic images utilizing CA-CFAR. In *Proceedings of the 2013 OCEANS-San Diego*; IEEE: New York, NY, USA, 2013; pp. 1–8.
20. Rahnemoonfar, M.; Rahman, A.F.; Kline, R.J.; Greene, A. Automatic seagrass disturbance pattern identification on sonar images. *IEEE J. Ocean. Eng.* **2018**, *44*, 132–141. [CrossRef]
21. Song, Y.; He, B.; Liu, P. Real-time object detection for AUVs using self-cascaded convolutional neural networks. *IEEE J. Ocean. Eng.* **2019**, *46*, 56–67. [CrossRef]
22. Palomeras, N.; Furfaro, T.; Williams, D.P.; Carreras, M.; Dugelay, S. Automatic target recognition for mine countermeasure missions using forward-looking sonar data. *IEEE J. Ocean. Eng.* **2021**, *47*, 141–161. [CrossRef]
23. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
24. Li, N.; Ding, B.; Yang, G.; Ni, S.; Wang, M. Lightweight LUW-DETR for efficient underwater benthic organism detection: N. Li et al. *Vis. Comput.* **2025**, *41*, 9191–9206. [CrossRef]
25. Lei, J.; Wang, H.; Fan, L.; Gu, Q.; Rong, S.; Zhang, H. SonarNet: Global Feature-Based Hybrid Attention Network for Side-Scan Sonar Image Segmentation. *Remote Sens.* **2025**, *17*, 2450. [CrossRef]
26. Lei, C.; Wang, H.; Lei, J. Si-GAT: Enhancing side-scan sonar image classification based on graph structure. *IEEE Sens. J.* **2024**, *24*, 24388–24404. [CrossRef]
27. Si, J.; Zhou, T.; Yu, X.; Du, W.; Xu, S. An unsupervised method for detecting and segmenting shadow areas of sunken targets in sonar images. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5006215. [CrossRef]
28. Tan, H.; Zheng, L.; Ma, C.; Xu, Y.; Sun, Y. Deep learning-assisted high-resolution sonar detection of local damage in underwater structures. *Autom. Constr.* **2024**, *164*, 105479. [CrossRef]
29. Shi, P.; He, Q.; Zhu, S.; Li, X.; Fan, X.; Xin, Y. Multi-scale fusion and efficient feature extraction for enhanced sonar image object detection. *Expert Syst. Appl.* **2024**, *256*, 124958. [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*, 5604413. [CrossRef]
31. Wang, Z.; Guo, J.; Zeng, L.; Zhang, C.; Wang, B. MLFFNet: Multilevel feature fusion network for object detection in sonar images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5119119. [CrossRef]
32. Wen, X.; Wang, J.; Cheng, C.; Zhang, F.; Pan, G. Underwater side-scan sonar target detection: YOLOv7 model combined with attention mechanism and scaling factor. *Remote Sens.* **2024**, *16*, 2492. [CrossRef]
33. Zhou, T.; Si, J.; Wang, L.; Xu, C.; Yu, X. Automatic detection of underwater small targets using forward-looking sonar images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4207912. [CrossRef]
34. Zheng, L.; Hu, T.; Zhu, J. Underwater sonar target detection based on improved ScEMA-YOLOv8. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1503505. [CrossRef]
35. Ruby, U.; Theerthagiri, P.; Jacob, I.J.; Yendapalli, V. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 5393–5397. [CrossRef]
36. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
37. Santos, N.P.; Moura, R.; Torgal, G.S.; Lobo, V.; de Castro Neto, M. Side-scan sonar imaging data of underwater vehicles for mine detection. *Data Brief* **2024**, *53*, 110132. [CrossRef]

38. Sethuraman, A.V.; Sheppard, A.; Bagoren, O.; Pinnow, C.; Anderson, J.; Havens, T.C.; Skinner, K.A. Machine learning for shipwreck segmentation from side scan sonar imagery: Dataset and benchmark. *Int. J. Robot. Res.* **2025**, *44*, 341–354. [CrossRef]
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 721.
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Johnson, O.V.; Xinying, C.; Khaw, K.W.; Lee, M.H. ps-CALR: Periodic-shift cosine annealing learning rate for deep neural networks. *IEEE Access* **2023**, *11*, 139171–139186. [CrossRef]
42. Dadboud, F.; Patel, V.; Mehta, V.; Bolic, M.; Mantegh, I. Single-stage uav detection and classification with yolov5: Mosaic data augmentation and panet. In *Proceedings of the 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*; IEEE: New York, NY, USA, 2021, pp. 1–8.
43. Barnes, L.R.; Schultz, D.M.; Gruntfest, E.C.; Hayden, M.H.; Benight, C.C. Corrigendum: False alarm rate or false alarm ratio? *Weather. Forecast.* **2009**, *24*, 1452–1454. [CrossRef]
44. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
45. Jocher, G. Ultralytics YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 15 April 2020).
46. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 12 August 2023).
47. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
48. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
49. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv* **2025**, arXiv:2509.25164.
50. Lin, X.; Peng, J.; Gan, Z.; Zhu, J.; Liu, J. YOLO-Master: MOE-Accelerated with Specialized Transformers for Enhanced Real-time Detection. *arXiv* **2025**, arXiv:2512.23273.
51. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 1314–1324.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.