Journal of *Marine Science and Engineering*

MDPI

Article

UWS-YOLO: Advancing Underwater Sonar Object Detection via Transfer Learning and Orthogonal-Snake Convolution Mechanisms

Liang Zhao *, Xu Ren, Lulu Fu, Qing Yun and Jiarun Yang

College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China; renxu@stu.haut.edu.cn (X.R.); 2022930852@stu.haut.edu.cn (L.F.); yunqing@stu.haut.edu.cn (Q.Y.); yangjiarun@stu.haut.edu.cn (J.Y.)

* Correspondence: liangzhao@haut.edu.cn; Tel.: +86-186-2371-8213

Abstract

Accurate and efficient detection of underwater targets in sonar imagery is critical for applications such as marine exploration, infrastructure inspection, and autonomous navigation. However, sonar-based object detection remains challenging due to low resolution, high noise, cluttered backgrounds, and the scarcity of annotated data. To address these issues, we propose UWS-YOLO, a novel detection framework specifically designed for underwater sonar images. The model integrates three key innovations: (1) a C2F-Ortho module that enhances multi-scale feature representation through orthogonal channel attention, improving sensitivity to small and low-contrast targets; (2) a DySnConv module that employs Dynamic Snake Convolution to adaptively capture elongated and irregular structures such as pipelines and cables; and (3) a cross-modal transfer learning strategy that pre-trains on large-scale optical underwater imagery before fine-tuning on sonar data, effectively mitigating overfitting and bridging the modality gap. Extensive evaluations on real-world sonar datasets demonstrate that UWS-YOLO achieves a mAP@0.5 of 87.1%, outperforming the YOLOv8n baseline by 3.5% and seven state-of-the-art detectors in accuracy while maintaining real-time performance at 158 FPS with only 8.8 GFLOPs. The framework exhibits strong generalization across datasets, robustness to noise, and computational efficiency on embedded devices, confirming its suitability for deployment in resource-constrained underwater environments.

Keywords: underwater object detection; sonar image analysis; cross-modal transfer learning; channel attention; dynamic snake convolution; real-time embedded systems

check for updates

Academic Editor: Philippe Blonde

Received: 10 August 2025
Revised: 7 September 2025
Accepted: 18 September 2025
Published: 24 September 2025

Citation: Zhao, L.; Ren, X.; Fu, L.; Yun, Q.; Yang, J. UWS-YOLO: Advancing Underwater Sonar Object Detection via Transfer Learning and Orthogonal-Snake Convolution Mechanisms. *J. Mar. Sci. Eng.* **2025**, *13*, 1847. https://doi.org/10.3390/jmse13101847

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Sonar technology is a cornerstone of ocean exploration, playing a pivotal role in underwater navigation, hydrographic surveying, resource assessment, and maritime security. Among imaging sonar modalities, forward-looking sonar, side-scan sonar, and synthetic aperture sonar are widely deployed, each offering distinct advantages for underwater observation [1]. While sonar imaging provides invaluable insights into seabed morphology and submerged objects, the underwater acoustic environment—characterized by complex propagation behaviors, reverberation, and scattering—introduces substantial noise and artifacts. These degradations not only impair image quality but also hinder accurate target detection and interpretation.

Recent advances in deep learning have revolutionized image analysis, delivering state-of-the-art performance in object detection, segmentation, and classification across diverse visual domains [2]. However, directly applying these approaches to sonar imagery remains challenging due to (1) high levels of sensor-induced noise and speckle artifacts, (2) cluttered and heterogeneous seabed backgrounds, and (3) large intra-class variability in target shapes, scales, and orientations [3]. These factors collectively reduce detection robustness and generalization.

1.1. Research Gap and Motivation

Despite increasing research attention, sonar-target detection using deep neural networks still faces persistent bottlenecks:

- Most existing methods are directly adapted from optical image detection frameworks without exploiting sonar-specific spatial–spectral characteristics, resulting in suboptimal feature representation.
- The scarcity of large-scale annotated sonar datasets hinders effective training from scratch, leading to overfitting and poor cross-domain generalization.
- Standard convolution operators struggle to accurately model elongated, low-contrast targets such as pipelines, cables, or debris.
- Cross-modal transfer learning between optical and acoustic imagery remains underexplored, leaving untapped potential for leveraging abundant optical underwater datasets.

These gaps highlight the need for sonar-tailored network architectures and learning strategies that directly address noise robustness, data scarcity, and shape-specific target modeling.

1.2. Contributions

In summary, the main contributions of this work are as follows:

- We propose UWS-YOLO, a sonar-specific object detection framework that addresses the dual challenges of detecting small, blurred, and elongated targets and maintaining real-time performance under limited annotated sonar datasets.
- We introduce the C2F-Ortho module in the backbone to enhance fine-grained feature representation by integrating orthogonal channel attention, improving sensitivity to low-contrast and small-scale targets.
- We design the DySnConv module in the detection head, which leverages Dynamic Snake Convolution to adaptively capture elongated and contour-aligned structures such as underwater pipelines and cables.
- We propose a cross-modal transfer learning strategy that pre-trains the network on large-scale optical underwater imagery and fine-tunes it on sonar data, effectively mitigating overfitting and bridging the modality gap.
- Extensive experiments on the public UATD sonar dataset show that UWS-YOLO achieves a +3.5% mAP@0.5 improvement over the YOLOv8n baseline and outperforms seven state-of-the-art detectors in both accuracy and recall while retaining real-time performance (158 FPS) with lightweight computational complexity (8.8 GFLOPs).

These contributions collectively advance the state of the art in sonar-based underwater target detection, providing a practical and efficient framework for deployment in real-time, resource-constrained marine applications.

1.3. Impact Statement

Beyond the demonstrated methodological advances, the proposed UWS-YOLO framework holds substantial practical significance for a wide range of real-world maritime

applications. Its ability to robustly detect small, elongated, and low-contrast targets in noisy sonar imagery directly benefits autonomous underwater vehicle (AUV) navigation, subsea infrastructure inspection, environmental monitoring, and maritime defense. Furthermore, the cross-modal transfer learning paradigm establishes a scalable pathway for leveraging abundant optical underwater datasets to accelerate progress in other resource-constrained sensing domains. These contributions collectively advance the state of the art in underwater perception and can serve as a cornerstone for next-generation intelligent ocean exploration systems.

The remainder of this paper is organized as follows: Section 2 reviews the state of the art in sonar target detection. Section 3 details the proposed methodology, including network architecture and transfer learning strategy. Section 4 presents experimental results and ablation studies. Section 5 concludes the paper and outlines future research directions.

2. Related Work

2.1. Object Detection

Object detection aims to automatically localize and classify objects within visual data, defining both their positions and categories. Methodological developments have evolved along two principal paradigms: traditional feature-engineering approaches and modern deep learning techniques.

Early traditional methods relied on handcrafted descriptors such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), followed by conventional classifiers. While effective under constrained conditions, they exhibited limited robustness to variations in scale, rotation, illumination, and background clutter.

The emergence of convolutional neural networks (CNNs) marked a paradigm shift, enabling hierarchical feature learning directly from raw image data. The early deep learning era was dominated by two-stage detectors, most notably the R-CNN family, which separated region proposal generation from object classification. Faster R-CNN [4] represented a seminal breakthrough by introducing the Region Proposal Network (RPN), enabling shared convolutional features and dramatically improving processing speed.

The demand for real-time detection led to the development of single-stage detectors such as YOLO [5] and SSD [6], which predict bounding boxes and class probabilities in a unified forward pass, achieving significantly faster inference. Despite these advances, challenges remain in detecting small, occluded, or densely packed targets. Recent research addresses these issues through multi-scale feature fusion, context-aware modeling, and attention mechanisms to capture fine-grained cues. Domain shift remains another critical bottleneck, motivating the adoption of unsupervised and semi-supervised domain adaptation. For latency-sensitive applications, model compression techniques such as pruning, quantization, and knowledge distillation have been actively explored to balance accuracy with computational efficiency.

2.2. Underwater Object Detection in Sonar Imagery

Sonar-based object detection has progressed through modality-specific innovations aligned with three major sensor types: forward-looking sonar (FLS), side-scan sonar (SSS), and synthetic aperture sonar (SAS).

For FLS, early work by Galceran et al. [7] established efficient feature computation using integral images. This was extended by Fan et al. [8], who introduced deep residual networks with adaptive optimization. More recently, Lu et al. [9] proposed AquaYOLO, an enhanced version of YOLOv8 designed to achieve more robust underwater sonar object detection.

In SSS, Manik et al. [10] pioneered multibeam–SSS data fusion, laying the groundwork for improved spatial coverage. Furthermore, Li et al. [11] advanced the field with real-time detection frameworks, while Wang et al. [12] achieved hardware–software co-design breakthroughs for embedded intelligent sonar platforms.

For SAS, Williams [13] developed unsupervised high-resolution detection algorithms, followed by Galusha et al. [14] who proposed adaptive anomaly detection strategies. Williams and Brown [15] further enhanced efficiency through 3D processing pipelines, enabling large-scale seabed surveys.

2.3. Current Challenges

Despite significant progress in sonar-based object detection, the field continues to face several persistent and interrelated challenges.

First, achieving effective multi-scale feature extraction remains critical for addressing the resolution variability inherent in different imaging ranges, sonar configurations, and seabed conditions. Inadequate multi-scale representations can lead to missed detections of small, distant, or partially occluded targets.

Second, attention-guided clutter suppression is essential for mitigating interference from complex underwater backgrounds, where acoustic backscatter, reverberation, and environmental noise can easily obscure or distort true target signatures. Robust suppression mechanisms are particularly important in shallow-water or highly heterogeneous seabed scenarios.

Third, ensuring robust cross-environment generalization is imperative for transferring models between diverse marine environments without substantial performance degradation. Challenges stem from domain shifts caused by variations in sonar type, operating frequency, platform motion, and environmental factors such as turbidity or thermoclines.

Addressing these challenges is central to advancing the reliability and operational applicability of sonar-based object detection systems, and they directly inform the methodological choices presented in our work (see Table 1).

Table 1. Summary of existing approaches, their limitations, and our proposed solutions in sonar-based object detection.

Existing Approach/Trend	Identified Limitation (Gap)	Our Proposed Solution
Handcrafted feature-based methods (e.g., SIFT, HOG) for sonar target detection	Sensitive to scale, rotation, and noise; poor performance in cluttered environments	Replaces handcrafted features with deep CNN-based hierarchical feature learning tailored to sonar characteristics (OrthoNet)
Two-stage deep detectors (e.g., Faster R-CNN) applied to sonar imagery	High computational cost; difficulty meeting real-time AUV constraints	Adopts a lightweight YOLO-based single-stage architecture optimized for low-latency sonar detection
Single-stage detectors (YOLO, SSD) directly adapted from optical imagery	Lack of sonar-specific feature enhancement; limited robustness to acoustic noise and clutter	Integrates orthogonal channel attention to selectively amplify discriminative features while suppressing irrelevant noise
Training with limited labeled sonar datasets	Overfitting; poor cross-domain generalization	Employs cross-modal transfer learning from large-scale optical underwater datasets with domain adaptation for sonar
Standard convolution operators in detection heads	Limited ability to model elongated, low-contrast structures (e.g., cables, pipelines)	Introduces Dynamic Snake Convolution (DySnConv) to adapt receptive fields to target geometry
Modality-specific research (FLS, SSS, SAS)	Lack of unified frameworks; innovations remain siloed	Develops a flexible detection framework applicable across sonar modalities via shared architecture components

3. Methods

3.1. Overall Architecture of UWS-YOLO

Conventional object detection architectures such as YOLO and SSD exhibit suboptimal performance in sonar-based target detection, primarily due to their limited adaptability to

the unique statistical and structural characteristics of acoustic imagery. To address these limitations, we build upon the robust YOLOv8 framework and introduce UWS-YOLO, a detector specifically tailored for underwater sonar imagery. As illustrated in Figure 1, the proposed architecture comprises four main components: (1) an input module for data preprocessing, (2) a backbone network for hierarchical feature extraction, (3) a head network for multi-scale feature fusion and refinement, and (4) a prediction module for target localization and classification.

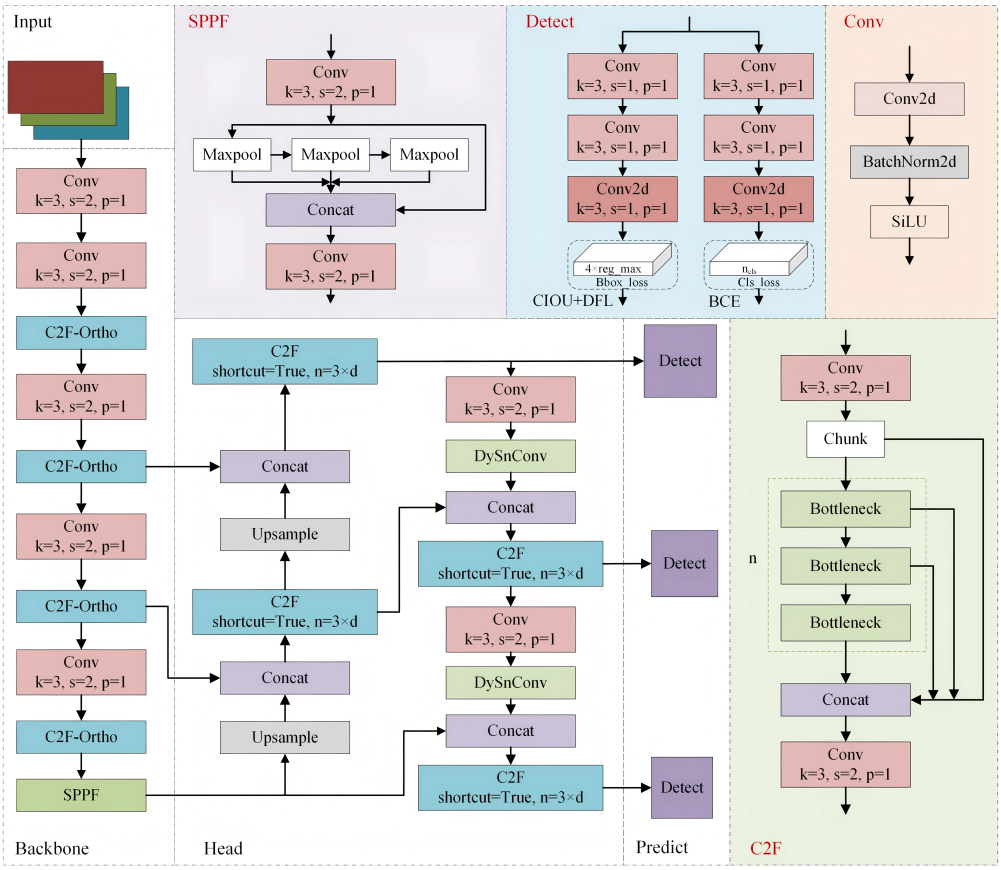


Figure 1. Schematic representation of the UWS-YOLO network architecture.

3.1.1. Input Module

This module is dedicated to input data preprocessing and augmentation, tailored for the challenges of underwater sonar imaging. All sonar images are first resized to a standardized resolution of 640 × 640 pixels to ensure consistency in spatial dimensions. Pixel intensity values are then normalized to a fixed range, facilitating stable convergence during training. To improve robustness against variations in target scale, orientation, and background complexity inherent in underwater acoustic environments, we incorporate mosaic augmentation during training. This augmentation strategy combines multiple images into a single composite input, effectively enriching contextual diversity and enabling the model to better generalize across heterogeneous sonar scenes.

3.1.2. Backbone Network

The backbone is based on an enhanced CSPDarknet53 structure, in which input features undergo five consecutive downsampling operations to produce multi-scale feature maps. Each convolutional block (Conv) consists of a convolution layer, batch normalization (BN), and a SiLU activation function. A key modification is the replacement of the original C2F module with the proposed C2F-Ortho module, which substantially improves feature representation capacity (detailed in Section 3.2). Furthermore, we employ the Spatial Pyra-

mid Pooling-Fast (SPPF) module to aggregate multi-scale context into fixed-size outputs, thereby enabling scale-invariant processing. Compared with the traditional SPP [16], SPPF significantly reduces computational overhead by sequentially applying three max-pooling layers, achieving lower latency while preserving contextual aggregation capability.

3.1.3. Head Network

For feature fusion, we adopt a Path Aggregation Network combined with a Feature Pyramid Network (PAN-FPN) structure, enabling both top-down and bottom-up information pathways. This design facilitates the integration of shallow, position-sensitive features with deep, semantic-rich features, thereby enhancing feature diversity and completeness. The Concat module merges feature maps of varying resolutions and semantic levels, while the Upsample module restores spatial granularity. Notably, we insert the DySnConv module after convolutional layers to adaptively emphasize tubular and elongated target structures common in sonar imagery, thus improving detection accuracy (detailed in Section 3.3).

3.1.4. Prediction Module

UWS-YOLO employs a decoupled detection head, with separate branches for classification and bounding box regression, each optimized with dedicated loss functions. Classification uses binary cross-entropy (BCE) loss, whereas bounding box regression is optimized via a combination of Distribution Focal Loss (DFL) [17] and Complete IoU loss (CIoU) [18], improving both convergence speed and localization precision. As an anchor-free model, UWS-YOLO dynamically assigns positive and negative samples through the Task-Aligned Assigner [19], enhancing robustness and generalization.

This holistic architectural design enables UWS-YOLO to achieve superior performance in extracting discriminative and semantically rich features from complex and noise-prone sonar imagery while simultaneously delivering accurate object predictions in real time. By integrating precision, robustness, and computational efficiency within a unified framework, the proposed architecture attains an optimal trade-off that is well-suited for the operational demands of underwater sonar object detection.

3.2. C2F-Ortho Module

Conventional channel attention mechanisms, such as the Squeeze-and-Excitation (SE) block, often learn channel descriptors that exhibit high inter-channel correlation. Such redundancy can limit the representational capacity of a neural network and impair its ability to distinguish salient targets from backgrounds, particularly in complex data such as sonar imagery.

To address this limitation, we introduce an Orthogonal Channel Attention (OCA) mechanism, where the projection filters are explicitly enforced to be mutually orthogonal. This design ensures that each filter responds to a distinct directional component within the channel-feature space, thereby maximizing descriptor diversity. By decorrelating these learned channel embeddings, the OCA mechanism offers two primary advantages: (1) it significantly enhances discriminative power, particularly in the presence of noise or clutter, and (2) it improves training stability by mitigating gradient instability (e.g., vanishing or explosion) often associated with highly correlated filters.

To further enhance feature representation in sonar imagery while maintaining computational efficiency, we propose the C2F-Ortho module. This module augments the standard C2F (Cross-Stage Partial-Net with Two-Branch Fusion) structure with our Orthogonal Channel Attention. As illustrated in Figure 2, the input tensor $\mathbf{X}$ is first processed through the C2F-Concat stage to generate an intermediate representation $\mathbf{X}_1$. The core of the C2F-Ortho module is the OCA mechanism, which extracts compressed channel descriptors via or-

thogonal projection, followed by a non-linear excitation function to generate channel-wise attention weights.

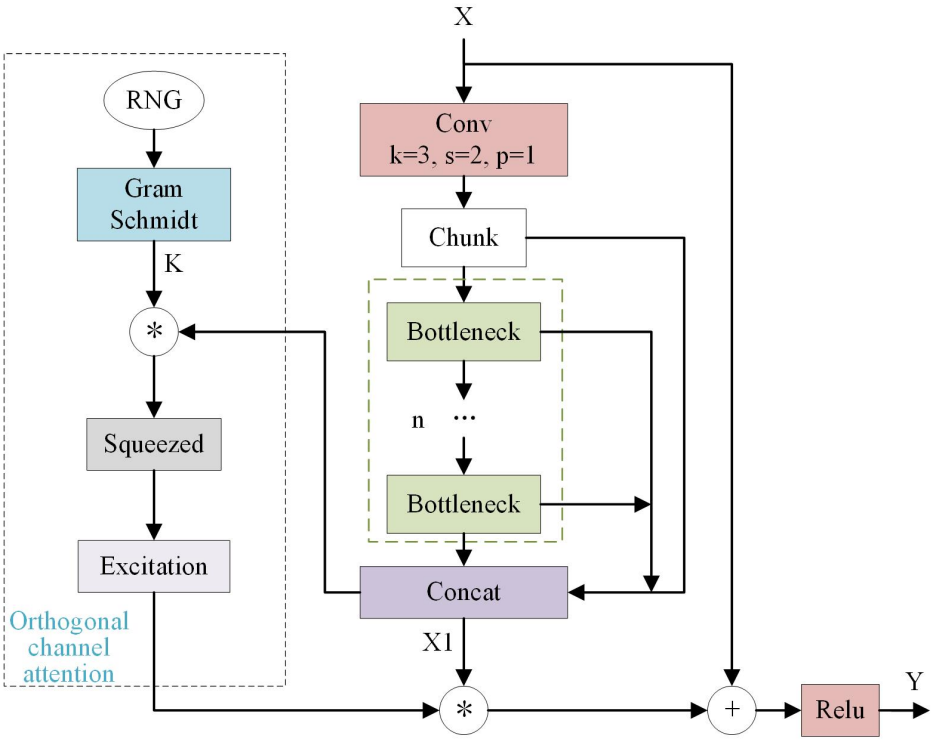


Figure 2. Architecture of the proposed C2F-Ortho module.

3.2.1. Orthogonal Projection

The core operation of the proposed OCA mechanism is an orthogonal linear projection within the channel–feature space. Let $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ denote the input feature map, where C is the number of channels and H and W are the spatial dimensions. First, we reshape $\mathbf{X}$ into

$$\mathbf{V} \in \mathbb{R}^{C\times M}, \quad M = H \times W, \tag{1}$$

where each row of $\mathbf{V}$ corresponds to the flattened spatial responses of a single channel.

The orthogonal compression is then performed according to

$$\mathbf{Z} = \mathbf{K}\mathbf{V}, \tag{2}$$

where $\mathbf{K} \in \mathbb{R}^{C\times C}$ is a learnable orthogonal projection matrix satisfying

$$\mathbf{K}^{\mathsf{T}}\mathbf{K} = \mathbf{I}_C, \tag{3}$$

with $\mathbf{I}_C$ denoting the C-dimensional identity matrix.

The c-th output descriptor is

$$Z_c = \mathbf{k}_c^{\mathsf{T}}\mathbf{V}, \tag{4}$$

where $\mathbf{k}_c^{\mathsf{T}}$ denotes the c-th row of $\mathbf{K}$. This guarantees that each channel descriptor is computed from a distinct orthogonal basis vector, ensuring statistical decorrelation and maximizing representational diversity.

3.2.2. Orthogonality Regularization

Algorithm 1 presents the detailed procedure for initializing the filter $\mathbf{K}$. Although $\mathbf{K}$ is initialized as orthogonal, standard gradient-based optimization may break this property during training. To address this, we introduce a soft orthogonality regularization term:

$$\mathcal{L}_{\text{ortho}} = \lambda \left\| \mathbf{K}^{\mathsf{T}}\mathbf{K} - \mathbf{I}_C \right\|_F^2, \tag{5}$$

where $\lambda > 0$ is a regularization coefficient and $\| \cdot \|_F$ denotes the Frobenius norm. The total loss function becomes

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{ortho}}, \tag{6}$$

where $\mathcal{L}_{\text{det}}$ is the task-specific detection loss (e.g., BCE, CIoU).

Algorithm 1 Orthogonal Channel Attention (OCA) Filter Initialization

Require: Input feature dimensions (C, H, W)
Ensure: Orthogonal projection matrix $\mathbf{K} \in \mathbb{R}^{C \times HW}$
if $HW < C$ **then**
  $n \leftarrow \lfloor \frac{C}{HW} \rfloor$
  Initialize empty list L
  for $i = 0$ **to** n **do**
    Initialize HW random filters $\{\mathbf{F}_j\} \in \mathbb{R}^{HW}$
    Orthogonalize $\{\mathbf{F}_j\}$ via Gram–Schmidt to obtain $\{\mathbf{F}'_j\}$
    Append $\{\mathbf{F}'_j\}$ to L
  end for
  Concatenate all L to form $\mathbf{K}$
else
  Initialize C random filters $\{\mathbf{F}_j\} \in \mathbb{R}^{HW}$
  Orthogonalize to obtain $\{\mathbf{F}'_j\}$
  Stack to form $\mathbf{K}$
end if

3.2.3. Excitation Step

After orthogonal projection, the excitation operation is applied:

$$\mathbf{A} = \sigma(\mathbf{W}_2\, \delta(\mathbf{W}_1 \mathbf{Z})), \tag{7}$$

$$\mathbf{X}_{\text{out}} = \mathbf{A} \odot \mathbf{X}, \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^C$ is the channel attention vector, $\delta(\cdot)$ is the ReLU, $\sigma(\cdot)$ is the sigmoid, and $\odot$ denotes channel-wise multiplication. $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are the MLP parameters, with r being the reduction ratio.

3.2.4. Theoretical Advantages

The integration of the orthogonality constraint into OCA provides several key benefits:

- Maximized Feature Diversity: Orthogonal rows of $\mathbf{K}$ form an orthonormal basis, minimizing redundancy and ensuring descriptors are statistically independent.
- Improved Gradient Flow: Orthogonal matrices are norm-preserving, stabilizing gradients and mitigating vanishing/exploding issues.
- Noise Robustness: Decorrelated descriptors suppress noise co-adaptation, reducing spurious activations in cluttered sonar data.

These advantages stem from established deep learning principles of orthogonality but are here innovatively applied to channel attention for enhanced representation and robustness.

Embedding OCA into the C2F-Ortho module produces diverse, decorrelated channel descriptors, enhancing robustness in complex sonar environments without increasing computational cost.

3.3. DySnConv Module

To enhance the network's ability to model elongated and curved structures prevalent in sonar imagery, we introduce the Dynamic Snake Convolution (DSConv) within a novel DySnConv module. Unlike conventional convolutions optimized for rectangular and rigid shapes, DSConv adaptively adjusts its sampling locations along a continuous, geometry-aware path, making it particularly effective for detecting tubular targets such as submarine pipelines and marine engineering structures.

3.3.1. Module Architecture

As illustrated in Figure 3, the DySnConv module processes the input feature tensor **X** via one standard convolution (Conv) layer and two parallel DSConv layers, thereby extracting features across three complementary receptive fields. The resulting multi-scale feature maps are concatenated (Concat) to form an enriched representation, followed by a 1×1 convolution to fuse information and adjust dimensionality. This design yields robust descriptors that simultaneously capture global context and fine-grained structural details of tubular targets.

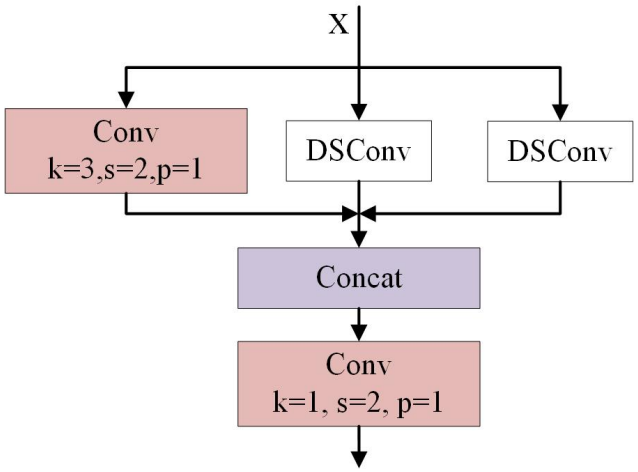


Figure 3. Structure of the proposed DySnConv module.

3.3.2. Dynamic Snake Convolution Principle

In standard convolution, kernel sampling positions follow a fixed grid pattern, which limits their ability to represent slender or curved geometries common in sonar scenes. The proposed DSConv addresses this limitation by learning deformation offsets that relocate the kernel's sampling points along a snake-like trajectory, conforming to the target's geometry.

As shown in Figure 4, for elongated objects such as underwater pipelines, DSConv dynamically adjusts its sampling path (red curve) to align with the object's central axis. This geometry-aware alignment enables more effective aggregation of semantically consistent features along the principal direction, outperforming the rigid fixed-grid sampling of standard convolution (blue grid).

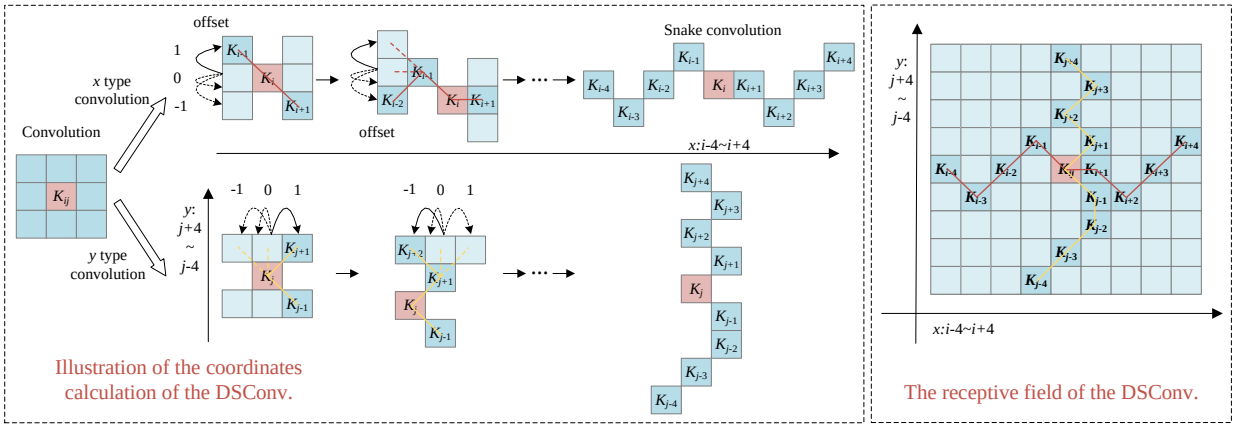


Figure 4. Conceptual illustration of Dynamic Snake Convolution.

3.3.3. Offset Learning and Coordinate Transformation

Given an input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, a parallel lightweight convolutional branch $B(\cdot)$ (e.g., a 3×3 depthwise convolution) predicts a two-dimensional offset field:

$$\mathbf{\Delta} \in \mathbb{R}^{H \times W \times 2N}, \tag{9}$$

where N denotes the number of sampling locations in the kernel (e.g., $N = 9$ for a 3×3 kernel).

Let $\mathbf{p}_n \in \mathbb{R}^2$ denote the fixed coordinate of the n-th grid sampling point. The dynamically adjusted coordinate is given by

$$\hat{\mathbf{p}}_n = \mathbf{p}_n + \mathbf{\Delta}(\mathbf{p}_n), \tag{10}$$

where $\mathbf{\Delta}(\mathbf{p}_n) \in \mathbb{R}^2$ is the learned offset vector.

For an output location $\mathbf{p}_0$, the DSConv feature response is computed as

$$Y(\mathbf{p}_0) = \sum_{n=1}^{N} w_n \cdot \mathbf{X}(\mathbf{p}_0 + \hat{\mathbf{p}}_n), \tag{11}$$

where w_n is the convolution weight and $\mathbf{X}(\cdot)$ is sampled at non-integer coordinates via bilinear interpolation.

To stabilize learning and prevent irregular sampling—particularly in low-contrast sonar imagery—we impose a continuity constraint that biases offsets to follow the target's principal axis. This preserves structural coherence and mitigates erratic displacement, which is crucial for weak-boundary objects in underwater scenes.

3.3.4. Advantages

The DySnConv module offers four key advantages:

- Geometric Adaptability: Learns spatially flexible sampling patterns that conform to elongated or curved target contours, ensuring precise structural alignment.
- Multi-Scale Feature Fusion: Combines standard convolution and DSConv outputs to capture both contextual information and fine structural cues.
- Enhanced Tubular Target Detection: Demonstrates superior accuracy for slender objects (e.g., pipelines) where fixed-grid kernels often underperform.
- Computational Efficiency: Achieves high geometric adaptivity with minimal additional computational overhead, making it viable for real-time sonar applications.

By jointly leveraging spatial adaptability and multi-scale feature integration, DySnConv substantially improves the robustness and accuracy of elongated-object detection in challenging underwater environments.

3.4. Transfer Learning

Training high-performance detection models directly from scratch on sonar imagery is challenged by two major factors: (1) the scarcity of large-scale, annotated sonar datasets and (2) the modality-specific characteristics of underwater acoustic sensing, such as speckle noise, low contrast, and acoustic shadowing. To address these constraints, we adopt a cross-modal transfer learning strategy in which knowledge learned from underwater optical imagery (source domain) is transferred to the sonar imagery domain (target domain). This leverages the model's ability to generalize high-level visual concepts across heterogeneous sensing modalities, thereby boosting detection accuracy in data-limited sonar scenarios.

3.4.1. Method Overview

We employ the proposed UWS-YOLO as the base architecture and implement transfer learning via a two-stage sequential process. Although advanced domain adaptation techniques—such as adversarial alignment or statistical distribution matching (e.g., MMD, CORAL)—could further reduce the modality gap, we select a fine-tuning–based approach for its computational simplicity, proven effectiveness, and compatibility with real-time constraints. This choice avoids additional architectural complexity and preserves inference speed.

- Source-domain pre-training: UWS-YOLO is first trained on a large-scale underwater optical image dataset. This stage enables the model to learn generic, modality-independent visual representations, including shape primitives, boundary structures, and semantic context.
- Target-domain fine-tuning: The pre-trained weights are then used to initialize the model for sonar imagery. Fine-tuning adapts these generic representations to sonar-specific signal patterns, accommodating domain characteristics such as high-intensity backscatter, granular texture, and acoustic shadow geometry.

The overall workflow of the proposed cross-modal transfer learning strategy is illustrated in Figure 5.

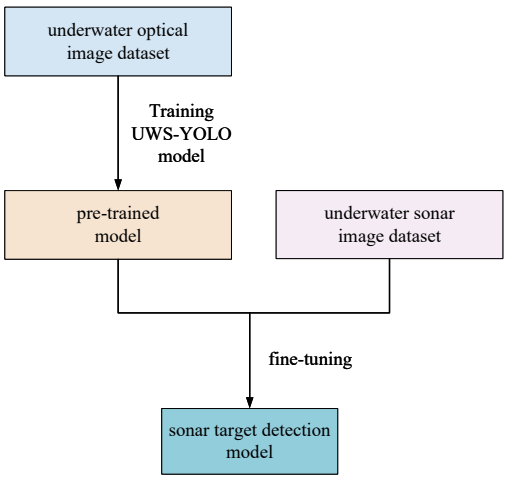


Figure 5. Schematic of the proposed cross-modal transfer learning framework.

3.4.2. Domain Gap Analysis

While the two-stage transfer paradigm is conceptually straightforward, the critical question is What form of knowledge is effectively transferred between optical and sonar domains? These modalities differ markedly in low-level attributes—optical imagery contains rich color, texture, and reflectance information, whereas sonar imagery is grayscale and dominated by speckle noise and signal-dependent artifacts.

We hypothesize that the transfer process operates primarily at a higher level of abstraction, leveraging shared geometric and semantic structures—such as edge continuity, contour geometry, and object–context relationships—rather than low-level appearance statistics. To test this hypothesis, we conduct a comparative analysis of feature embeddings and clustering quality under two training regimes: (1) learning exclusively from sonar data initialized from scratch and (2) pre-training on underwater optical imagery followed by fine-tuning on sonar data.

Our analysis yields two key observations:

- Evidence of domain adaptation: Models trained from scratch exhibit entangled feature spaces with poor inter-class separation. In contrast, cross-modal transfer produces more compact, well-separated clusters in the embedding space, as reflected by higher silhouette scores and lower within-class variance. This indicates that pre-training suppresses modality-specific noise while structuring features according to semantic similarity.
- Invariance of transferred knowledge: Consistent gains in both discrimination and generalization suggest that the transferred knowledge encodes modality-invariant object properties—e.g., cylindrical curvature of pipes, axial symmetry of divers, or composite geometry of ROVs. Pre-training instills a strong morphological prior that is subsequently adapted to acoustic signal characteristics, enabling accurate mapping from sonar imagery to abstract object representations.

These findings confirm that cross-modal transfer promotes a semantically structured, modality-agnostic feature space. Optical pre-training thus serves as a rich supervisory signal for universal object attributes, mitigating overfitting from sparse sonar labels and directly addressing the data scarcity problem.

3.4.3. Rationale and Benefits

Underwater optical and sonar imagery differ fundamentally in sensing physics—light-based versus sound-based—but share structural and semantic consistencies in object shapes, spatial arrangements, and geometric relationships. Pre-training in the optical domain enhances sonar detection through

- Improved generalization: Rich source-domain supervision alleviates overfitting on small sonar datasets.
- Feature alignment: Fine-tuning adapts generic visual priors to sonar-specific noise and texture distributions, bridging the perceptual gap between modalities.
- Accelerated convergence: Initialization from pre-trained weights reduces training time relative to random initialization.
- Enhanced detection performance: Morphological priors improve object localization and classification, particularly under complex underwater clutter and occlusion.

In summary, the proposed cross-modal transfer learning framework establishes an efficient and effective knowledge adaptation pathway from optical to sonar imagery. By coupling large-scale optical pre-training with targeted sonar fine-tuning, it narrows the domain gap, boosts robustness, and maintains real-time performance in challenging underwater environments.

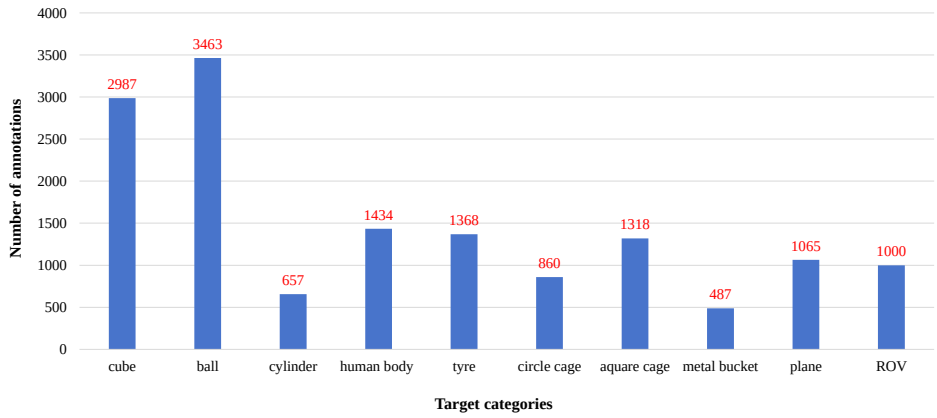
4. Experiments

4.1. Datasets

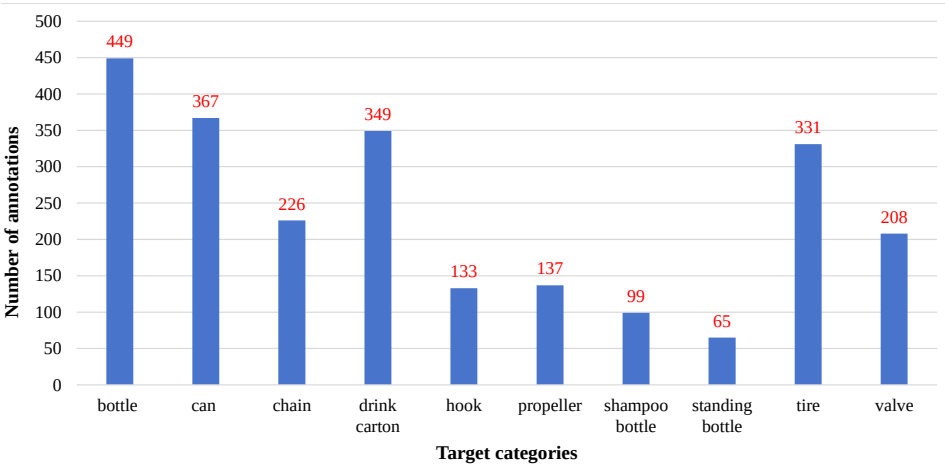
In this study, the proposed detection framework undergoes a two-stage training process, involving initial pre-training on a large-scale underwater optical image dataset, followed by fine-tuning with an underwater sonar dataset. To ensure a comprehensive evaluation of its robustness and generalization capacity across heterogeneous sensing modalities, we perform extensive experiments on two representative underwater sonar datasets—Underwater Acoustic Target Detection (UATD) [20] and Marine Debris Forward-Looking Sonar (MDFLS)—as well as a large-scale underwater optical dataset, the Real-World Underwater Object Detection (RUOD) dataset.

4.1.1. UATD Dataset

The UATD dataset contains 9200 sonar images with 14,639 annotated instances across ten target categories: cube, ball, cylinder, human body, tire, circular cage, square cage, metal bucket, plane, and remotely operated vehicle (ROV). As shown in Figure 6a, the class distribution is highly imbalanced; for example, the ball category has over 3400 instances, whereas the metal bucket category has only about 500. To address this, standard augmentation strategies (e.g., mosaic, mixup) and class-aware sampling are applied during training. The dataset is split into 7600 images for training and 800 each for validation and testing.



(**a**) UATD dataset



(**b**) MDFLS dataset

Figure 6. Instance distributions of (**a**) UATD and (**b**) MDFLS datasets.

4.1.2. MDFLS Dataset

The MDFLS dataset comprises 1868 forward-looking sonar images captured with the ARIS Explorer 3000, covering ten common marine debris categories: bottle (449 instances), can (367), drink carton (349), hook (133), propeller (137), shampoo bottle (99), tire (331), chain (226), valve (208), and standing bottle (65). As illustrated in Figure 6b, the dataset is notably imbalanced, reflecting real-world debris occurrence. Data are divided into training, validation, and test sets at an 8:1:1 ratio.

4.1.3. RUOD Dataset

The RUOD dataset g[21] contains 14,000 high-resolution underwater optical images with 74,903 annotated instances across ten categories: fish, diver, starfish, coral, turtle, holothurian, echinus, scallop, cuttlefish, and jellyfish. Fish, echinus, coral, starfish, and holothurian are the most frequent, consistent with their abundance in natural marine environments (Figure 7). The dataset is split into 9800 training images and 2100 images each for validation and testing, providing large-scale and taxonomically diverse data for learning transferable underwater features, thereby enabling robust cross-domain adaptation.

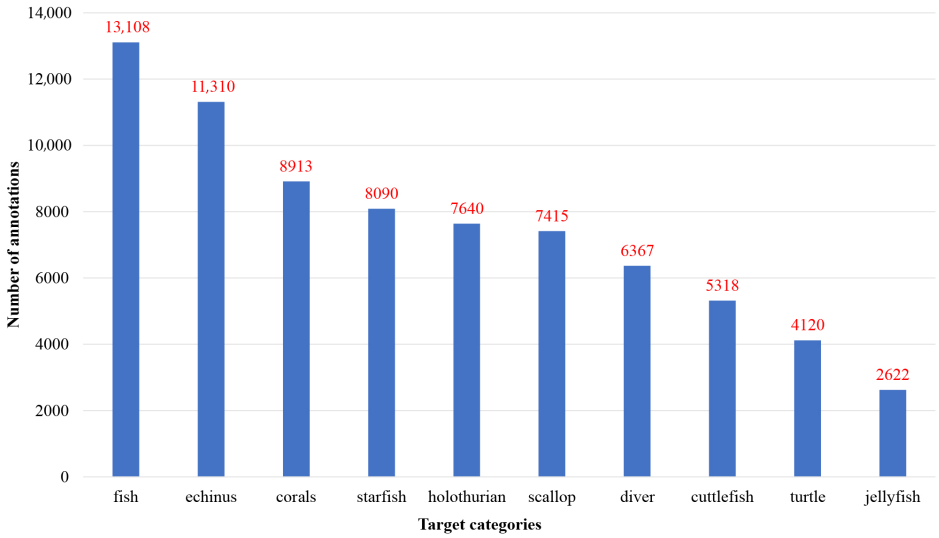


Figure 7. Instance distribution of the RUOD dataset.

4.2. Experimental Setup

All experiments were conducted under a fixed hardware and software configuration, as summarized in Table 2.

Table 2. Hardware and software environment.

Environment Type	Configuration
CPU	Intel(R) Xeon(R) Silver 4214 @ 2.20 GHz
GPU	NVIDIA GeForce RTX 2080 Ti
RAM	512 GB
Operating System	Windows 10
Deep Learning Frameworks	PyTorch 1.9.0, CUDA 10.2, cuDNN 10.2
Programming Language	Python 3.9

The hyperparameter configuration used for model training is listed in Table 3. These parameters were selected empirically to ensure a stable optimization process and effective

convergence. Note that the initial learning rate Lr_0 and final learning rate Lr_1 correspond to the starting and ending values of a cosine learning rate scheduler.

Table 3. Training hyperparameter configuration.

Hyperparameter	Value
Epochs	100
Batch Size	32
Image Resolution	640×640
Momentum	0.937
Initial Learning Rate (Lr_0)	0.001
Final Learning Rate (Lr_1)	0.01
Optimizer	Adam

4.3. Evaluation Metrics

To comprehensively evaluate model performance, we adopt multiple complementary metrics: Precision (P), recall (R), and mean average precision (mAP) to assess detection accuracy; Frames Per Second (FPS) to measure inference speed; parameter count to assess model size; and GFLOPs to quantify computational complexity.

The definitions of P, R, and mAP are given in Equations (12)–(14):

$$P = \frac{TP}{TP + FP}, \tag{12}$$

$$R = \frac{TP}{TP + FN}, \tag{13}$$

$$mAP = \frac{1}{C} \sum_{i=1}^{C} \int_0^1 P_i(r_i)\, dr_i, \tag{14}$$

where

- TP—True Positives, correctly predicted positive instances;
- FP—False Positives, incorrectly predicted positive instances;
- FN—False Negatives, actual positives missed by the model;
- $P_i(r_i)$—precision at recall r_i for class i;
- C—total number of target classes.

FPS is computed as the number of images processed per second during inference, using a fixed batch size of 1. GFLOPs represent the number of floating-point operations required to process a single image, serving as a proxy for computational cost.

These metrics collectively reflect detection accuracy, efficiency, and deployability, enabling a fair comparison between proposed and baseline models in both resource-constrained and high-performance computing scenarios.

4.4. Ablation Studies

We conduct comprehensive ablation experiments to isolate and quantify the individual and complementary effects of the C2F-Ortho module, the DySnConv module, and the cross-modal transfer learning strategy. Beyond evaluating overall detection accuracy, we further analyze the sensitivity of key hyperparameters to provide a fine-grained understanding of their respective contributions and interactions.

4.4.1. Module-Wise Ablation Analysis

Table 4 summarizes the ablation results on the UATD sonar dataset, where S, T, and C denote the average mAP for small, tubular, and common objects, respectively.

Table 4. Ablation study on the UATD sonar dataset.

Group	Components	mAP@ 0.5	mAP@ 0.5:0.95	F1	Params (M)	GFLOPs	FPS	mAP-S	mAP-T	mAP-C
1	Baseline (YOLOv8n)	83.6	54.2	0.845	3.01	8.1	212	70.1	75.3	88.5
2	+C2F-Ortho	84.0	55.0	0.839	3.05	8.1	185	72.5	76.0	88.2
3	+DySnConv	84.6	55.8	0.832	3.45	8.7	172	71.2	79.8	87.9
4	+Both Modules	86.1	57.9	0.859	3.49	8.8	166	73.8	81.1	90.1
5	+Transfer Learning	**87.1**	**59.5**	**0.863**	3.49	8.8	158	**75.9**	**82.6**	**91.5**

C2F-Ortho (Group 2): Yields a modest overall gain of $+0.4\%$ mAP@0.5 but a more notable $+2.4\%$ improvement for small objects (mAP-S), validating its role in enhancing detection sensitivity to weak targets. A slight drop in F1-score, primarily due to reduced precision, suggests a trade-off between sensitivity and false positive control.

DySnConv (Group 3): Achieves a substantial $+4.5\%$ improvement for tubular targets (mAP-T), consistent with its design objective of modeling elongated geometries. Overall mAP@0.5 improves by $+1.0\%$, though reductions in recall and F1-score indicate that, in isolation, this module may compromise detection performance on non-tubular classes.

Combined Modules (Group 4): Integrating both C2F-Ortho and DySnConv produces a $+2.5\%$ mAP@0.5 improvement over the baseline while recovering the F1-score to 0.859. Category-specific gains for small ($+3.7\%$) and tubular ($+5.8\%$) targets are nearly additive, underscoring their complementary effects.

Transfer Learning (Group 5): Augmenting the combined model with cross-modal transfer learning further improves mAP@0.5 by $+1.0\%$ and attains the highest F1-score (0.863). The performance uplift is uniform across categories, with the most pronounced benefits on small and tubular classes, highlighting its utility for robust generalization to challenging targets.

C2F-Ortho primarily enhances small-target sensitivity, DySnConv specializes in tubular-object discrimination, and their joint use yields complementary performance gains. Cross-modal transfer learning further strengthens generalization across all categories while preserving computational efficiency.

4.4.2. Hyperparameter Sensitivity Analysis

We analyze the sensitivity of two key hyperparameters to assess their impact on detection accuracy and training stability.

Orthogonal Regularization Weight (λ). The coefficient λ controls the strength of the orthogonality constraint in the C2F-Ortho module, modulating the trade-off between feature decorrelation and representational flexibility. Experiments show that $\lambda = 0.01$ yields optimal performance. Smaller values (e.g., 0.001) fail to sufficiently enforce orthogonality, leading to feature redundancy and reduced inter-class separability. Larger values (e.g., 0.1) over-constrain the weight matrix, impeding its capacity to encode discriminative features and degrading detection accuracy.

DSConv Offset Learning Rate Multiplier. The learning rate multiplier for the deformable offset branch in DySnConv governs the interplay between the adaptive sampling process and the main network. A reduced multiplier of 0.1—lower than the global learning rate—markedly improves training stability and final accuracy. This slower update rate allows offset predictions to evolve progressively, giving the backbone time to adapt to changing receptive fields without destabilization.

Careful calibration of λ and the offset learning rate multiplier is critical to harmonizing module interactions and fully exploiting their performance potential. Such tuning

ensures robust optimization dynamics while maximizing the benefits of the proposed architectural innovations.

4.5. Comparative Experiments with Other Cutting-Edge Approaches

To comprehensively validate the superiority of the proposed UWS-YOLO model, we conducted comparative experiments against seven state-of-the-art object detection frameworks: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv7, YOLOv7-CHS [22], YOLOv5n, and YOLOv5s. The evaluation covers detection accuracy (precision P, recall R, and mAP@0.5), inference speed (FPS), and model complexity (parameters and GFLOPs). The results are summarized in Table 5.

Table 5. Performance comparison of underwater object detection models. The best results are highlighted in bold.

Model	P (%)	R (%)	mAP@0.5 (%)	Parameters	GFLOPs	FPS
YOLOv5n	84.2	74.2	81.0	**1,772,695**	**4.2**	270
YOLOv5s	83.5	77.2	82.8	7,037,095	15.8	227
YOLOv7	**87.8**	83.6	82.5	36,490,443	103.2	81
YOLOv7-CHS	87.0	82.1	81.9	33,585,246	40.3	111
YOLOv8n	86.8	82.5	83.6	3,007,598	8.1	212
YOLOv8s	87.4	80.5	**86.0**	11,129,454	28.6	**285**
YOLOv8m	86.0	82.2	83.0	25,845,550	78.7	123
UWS-YOLO	87.3	**84.1**	**87.1**	3,491,638	8.8	158

As shown in Table 5, YOLOv5n achieves the lowest computational complexity (1.77 M parameters, 4.2 GFLOPs) but also the lowest mAP@0.5 (81.0%), reflecting a clear trade-off between efficiency and accuracy. YOLOv8s delivers the highest inference speed (285 FPS) but incurs a significantly higher computational cost (28.6 GFLOPs), with its mAP@0.5 (86.0%) still lower than that of UWS-YOLO. YOLOv7 and YOLOv7-CHS achieve high precision (87.0–87.8%) but suffer from excessive parameter counts (33–36M), high FLOPs (40.3–103.2), and relatively low speeds (81–111 FPS), limiting their suitability for real-time embedded applications.

In comparison, UWS-YOLO achieves the highest mAP@0.5 (87.1%) and recall (84.1%) among all tested models, while maintaining moderate complexity (3.49 M parameters, 8.8 GFLOPs) and real-time speed (158 FPS). This balanced performance highlights the effectiveness of our architectural innovations, including the C2F-Ortho and DySnConv modules, as well as the cross-modal transfer learning strategy. For a clear presentation and comparison of the experimental results, the following two scenarios are selected as representative cases for detailed analysis.

Qualitative Analysis in Scenario 1. The original images were enhanced to improve the visibility of the detection results, as illustrated in Figure 8. UWS-YOLO (Figure 8h) detects all relevant targets with tight bounding boxes and fewer missed detections compared to YOLOv5n (Figure 8a) and YOLOv8n (Figure 8e), which tend to miss low-contrast or small objects. These improvements stem from C2F-Ortho's enhanced shallow-layer feature extraction, which boosts sensitivity to weak targets.

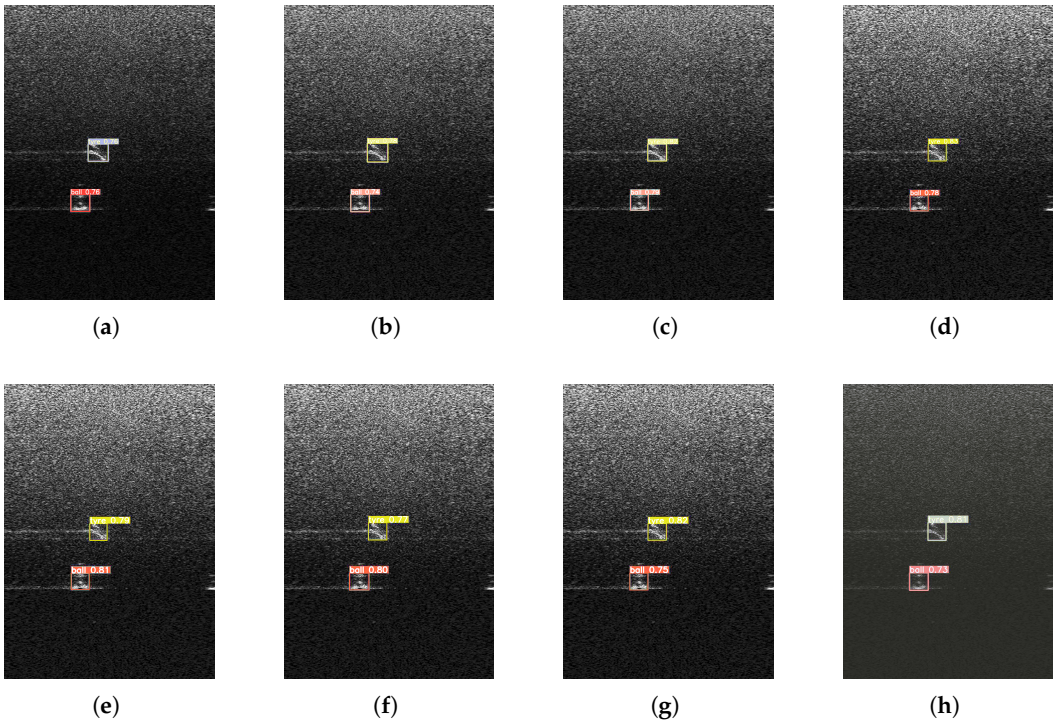


Figure 8. Visual comparison of detection results in scenario 1: (**a**) YOLOv5n; (**b**) YOLOv5s; (**c**) YOLOv7; (**d**) YOLOv7-CHS; (**e**) YOLOv8n; (**f**) YOLOv8s; (**g**) YOLOv8m; (**h**) UWS-YOLO.

Qualitative Analysis in Scenario 2. In a more complex experimental setting, the preprocessed detection results are presented in Figure 9. YOLOv5s (Figure 9b) and YOLOv8s (Figure 9f) exhibit more false negatives and localization errors in such conditions. In contrast, UWS-YOLO (Figure 9h) maintains high detection accuracy, successfully separating adjacent objects with minimal overlap errors. This robustness results from DySnConv's adaptive receptive fields, which better capture object geometry.

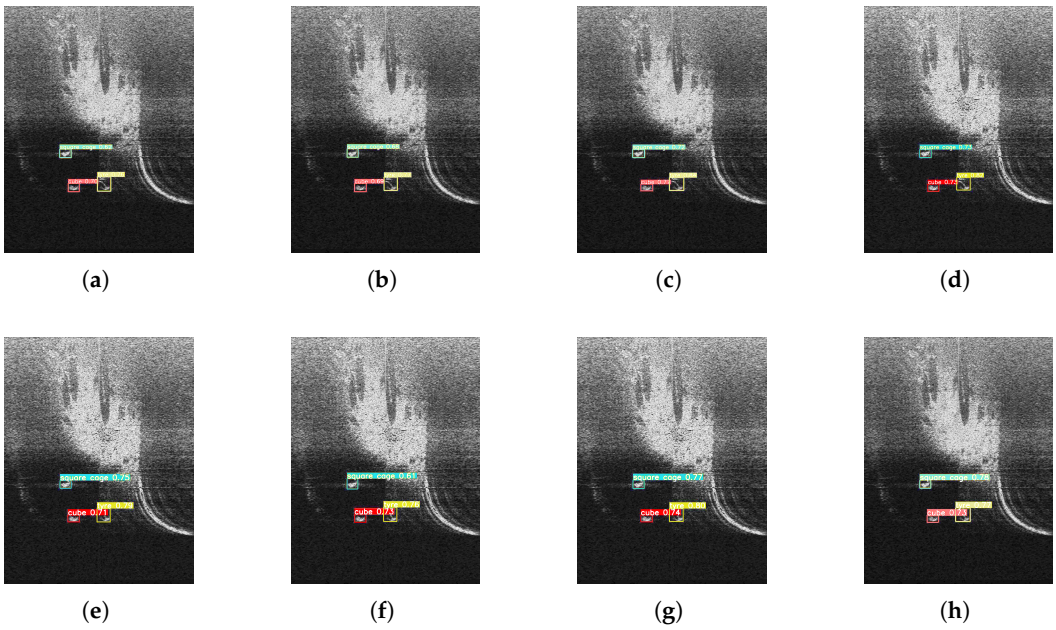


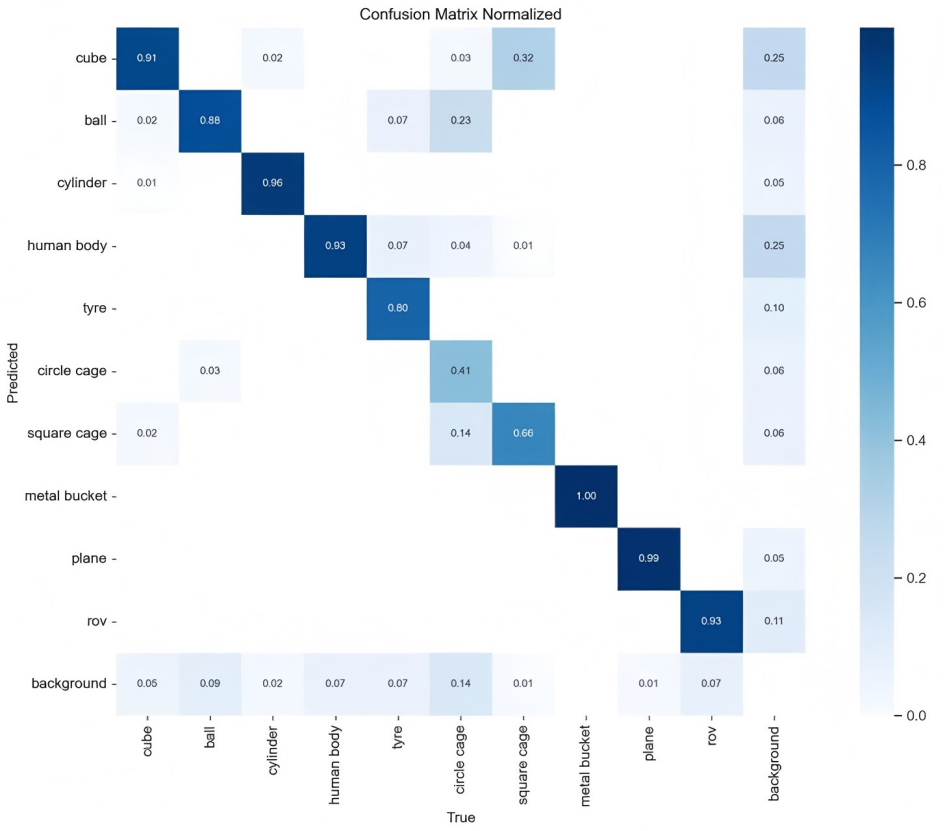
Figure 9. Visual comparison of detection results in scenario 2: (**a**) YOLOv5n; (**b**) YOLOv5s; (**c**) YOLOv7; (**d**) YOLOv7-CHS; (**e**) YOLOv8n; (**f**) YOLOv8s; (**g**) YOLOv8m; (**h**) UWS-YOLO.

We further conduct a quantitative evaluation of detection performance within the context of the two aforementioned scenarios. As presented in Table 6, the proposed UWS-YOLO consistently achieves higher mAP@0.5 scores and lower false-negative (FN) rates in both settings compared to the baseline YOLOv8n. These results demonstrate the robustness of UWS-YOLO when confronted with complex scenarios characterized by dynamic changes and its effectiveness in reducing missed detections without sacrificing precision.

Category-Level Performance via Confusion Matrices. Confusion matrices (Figure 10) provide a detailed, category-level perspective on detection performance. The YOLOv8n matrix (Figure 10a) exhibits lighter and more scattered diagonal elements, indicating uneven accuracy across categories and a higher tendency to confuse visually similar targets. In contrast, UWS-YOLO (Figure 10b) presents darker, more concentrated diagonal patterns, reflecting higher and more consistent category-level precision. Notably, the most substantial improvements are observed for small and tubular targets, which are crucial in practical underwater tasks such as subsea infrastructure inspection and marine debris monitoring. These gains can be attributed to the synergy between C2F-Ortho's multi-scale feature extraction and DySnConv's adaptive shape modeling.

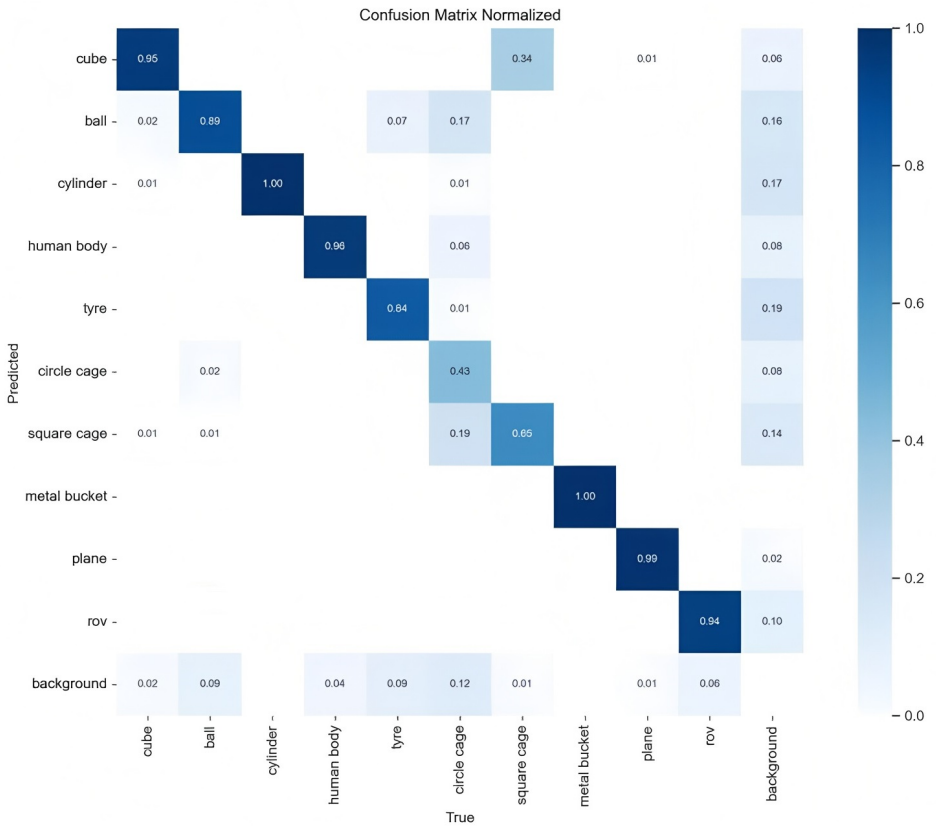
Table 6. Performance comparison in representative scenarios.

Scenario	Model	mAP@0.5	FN Rate	FP Rate
Scenario 1	YOLOv8n	82.1	0.18	0.12
Scenario 1	UWS-YOLO	**86.3**	**0.09**	**0.08**
Scenario 2	YOLOv8n	80.5	0.22	0.15
Scenario 2	UWS-YOLO	**84.7**	**0.11**	**0.10**



(**a**) YOLOv8n

Figure 10. *Cont.*



(**b**) UWS-YOLO

Figure 10. Comparison of confusion matrices: (**a**) YOLOv8n; (**b**) UWS-YOLO.

Quantitatively, UWS-YOLO attains an average diagonal value of 0.91, outperforming YOLOv8n's 0.84. Moreover, the most frequent misclassification—tire misidentified as circle cage—was reduced from 12% to 6%, underscoring the model's enhanced discriminative capability for structurally similar object categories.

4.6. Generalization and Robustness Analysis

To rigorously assess the practicality and resilience of the proposed UWS-YOLO framework beyond the standard UATD benchmark, we performed a comprehensive set of evaluations focusing on three critical aspects: (1) cross-dataset generalization, (2) robustness against noise corruption, and (3) computation on embedded hardware platforms.

Cross-Dataset Generalization: An effective underwater object detector should remain robust when applied to data from distributions different from those used for training. To examine this capability, we evaluated UWS-YOLO and the baseline YOLOv8n on the MDFLS dataset, which contains FLS imagery with visual characteristics distinct from the UATD dataset.

As reported in Table 7, UWS-YOLO achieved an mAP@0.5 of 76.4% on MDFLS, outperforming YOLOv8n (72.1%) by a substantial margin of 4.3 points. This notable gain indicates that the features learned by UWS-YOLO—reinforced through the proposed orthogonal attention mechanism and geometric adaptability—exhibit superior transferability across sonar modalities, thereby alleviating the domain shift problem to a significant extent.

Noise Robustness: Underwater acoustic imaging is inherently susceptible to various forms of noise contamination. To evaluate robustness under such degradations, we synthetically introduced two common noise types to the UATD test set: Gaussian noise ($\mu = 0$,

$\sigma^2 \in [0.01, 0.05]$) and Speckle noise (variance $\in [0.01, 0.05]$). Across all tested noise levels, UWS-YOLO consistently outperformed YOLOv8n.

For example, under severe Gaussian noise ($\sigma^2 = 0.05$), the mAP@0.5 of UWS-YOLO decreased by merely 8.7 percentage points (from 87.1% to 78.4%), whereas YOLOv8n exhibited a larger decline of 12.1 points (from 83.6% to 71.5%). This enhanced resilience is attributed to the C2F-Ortho module's capacity to suppress noise-sensitive channels while amplifying discriminative feature responses, thereby mitigating the impact of sensory corruption.

Computational Efficiency on Embedded Devices: For practical deployment on AUVs, computational efficiency is essential. We benchmarked UWS-YOLO and competing models on an NVIDIA Jetson AGX Xavier embedded platform, recording average inference speed (FPS) and power consumption (W). As shown in Table 7, UWS-YOLO achieved 37 FPS with a power draw of 22.5 W, offering an excellent balance between inference speed and detection accuracy.

While lighter architectures such as YOLOv5n achieved higher throughput (58 FPS), they incurred notable accuracy losses. In contrast, heavier models like YOLOv7 failed to meet real-time constraints, with speeds below 15 FPS. These results confirm that UWS-YOLO is well-suited for real-time applications in resource-constrained underwater environments.

Table 7. Performance comparison on cross-dataset generalization (MDFLS), noise robustness, and embedded efficiency evaluation.

Model	mAP@0.5 (MDFLS)	mAP@0.5 Under Noise ($\sigma^2 = 0.05$)	FPS (Jetson)	Power (W)
YOLOv5n	68.9	71.5	58	18.2
YOLOv8n	72.1	71.5	49	20.1
YOLOv8s	75.3	75.8	32	23.8
UWS-YOLO	**76.4**	**78.4**	**37**	22.5

4.7. Discussion

The experimental results in Sections 4.4–4.6 provide strong empirical support for the effectiveness of UWS-YOLO. Ablation studies validate the individual and synergistic contributions of the C2F-Ortho module, DySnConv operator, and the cross-modal transfer learning strategy. Comparative analyses further confirm that UWS-YOLO achieves state-of-the-art accuracy among computationally efficient detectors.

Extended evaluations in Section 4.6 address key deployment factors. UWS-YOLO demonstrates superior cross-dataset generalization across different FLS datasets, maintains stable performance under severe noise contamination, and achieves real-time inference on embedded hardware with moderate power consumption. This multi-perspective evaluation surpasses conventional benchmarks, underscoring the model's operational readiness for resource-constrained underwater environments.

Nonetheless, certain limitations remain. A residual domain gap persists due to variations in sonar hardware, acquisition parameters, and environmental conditions. In addition, qualitative inspection reveals occasional false positives in highly cluttered seabed regions, where background textures mimic target geometry or intensity patterns.

These challenges open several promising research avenues: (1) Applying domain adaptation and generalization to further reduce cross-dataset discrepancies; (2) Leveraging temporal coherence in sonar video to suppress clutter and improve detection confidence; (3) Exploring multi-modal fusion of acoustic and optical imagery to exploit complementary spatial–spectral cues; (4) Incorporating active learning to continuously refine the detector in evolving operational environments.

Overall, UWS-YOLO provides a robust and efficient foundation for underwater object detection while offering clear pathways to further enhance adaptability, precision, and reliability in real-world deployments.

5. Conclusions

Building on the preceding discussion, this study addresses two longstanding challenges in underwater sonar target detection: (1) accurately recognizing small, low-contrast, and elongated structures and (2) balancing detection performance with computational efficiency under limited annotated sonar data.

To tackle these challenges, we propose UWS-YOLO, a tailored architecture integrating three key innovations: (i) the C2F-Ortho module in the backbone, which enhances fine-grained feature extraction through orthogonal channel attention, improving the detection of small and low-contrast targets; (ii) the DySnConv module in the detection head, which adaptively models elongated and irregular geometries such as submarine pipelines; and (iii) a cross-modal transfer learning framework that pre-trains on large-scale underwater optical datasets before fine-tuning on sonar data, mitigating overfitting and narrowing the perceptual gap between acoustic and optical modalities.

Extensive experiments show that UWS-YOLO improves mAP@0.5 by 3.5% over the baseline and outperforms seven state-of-the-art detectors (YOLOv5, YOLOv7, YOLOv8 variants) in detection accuracy while sustaining real-time inference at 158 FPS. These results highlight the model's capability to meet the stringent demands of sonar-based detection tasks.

Future work will focus on (1) optimizing the architecture for greater computational efficiency without sacrificing accuracy, including hardware-aware neural architecture search and mixed-precision inference, and (2) extending the framework to more complex operational conditions, such as cluttered seabeds, multi-target occlusion, and multi-sensor fusion (e.g., sonar–optical integration). These directions aim to further enhance the adaptability and deployment readiness of UWS-YOLO for reliable, efficient underwater perception.

Author Contributions: Conceptualization, L.Z.; Methodology, X.R., L.F. and Q.Y.; Validation, J.Y.; Investigation, L.Z., X.R., L.F. and Q.Y.; Data curation, L.Z., X.R., Q.Y. and J.Y.; Writing—original draft, Q.Y.; Writing-review & editing, L.Z., X.R., L.F. and J.Y.; Funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61473114), the Key Scientific Research Projects of Higher Education Institutions in Henan Province (Grant No. 24B520006), and the Open Fund of Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education (Grant No. KFJJ2024013).

Data Availability Statement: The datasets used in the paper can be downloaded here "UATD Dataset" at https://figshare.com/articles/dataset/UATD_Dataset/21331143/3 accessed on 18 June 2024 and "Marine Debris Dataset" at https://github.com/mvaldenegro/marine-debris-fls-datasets/ accessed on 10 January 2024.

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

References

1. Hożýn, S. A Review of Underwater Mine Detection and Classification in Sonar Imagery. *Electronics* **2021**, *10*, 2943. [CrossRef]
2. Er, M.J.; Chen, J.; Zhang, Y.; Gao, W. Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review. *Sensors* **2023**, *23*, 1990. [CrossRef]
3. Sarkar, P.; De S.; Gurung, S. A Survey on Underwater Object Detection. In *Intelligence Enabled Research: DoSIER 2021*; Springer: Singapore, 2022; pp. 91–104.

4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
5. Sapkota, R.; Qureshi, R.; Calero, M.F.; Badgujar, C.; Nepal, U.; Poulose, A.; Zeno, P.; Vaddevolu, U.B.P.V.; Yan, H.; Karkee, M. YOLOv10 to Its Genesis: A Decadal and Comprehensive Review of the You Only Look Once Series. *arXiv* **2024**, arXiv:2406.19407.
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV 2016), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
7. Galceran, E.; Djapic, V.; Carreras, M.; Williams, D.P. A Real-Time Underwater Object Detection Algorithm for Multi-Beam Forward Looking Sonar. *Ifac Proc. Vol.* **2012**, *45*, 306–311.
8. Fan, Z.; Xia, W.; Liu, X.; Li, H. Detection and Segmentation of Underwater Objects from Forward-Looking Sonar Based on a Modified Mask RCNN. *Signal Image Video Process.* **2021**, *15*, 1135–1143. [CrossRef]
9. Lu, Y.; Zhang, J.; Chen, Q.; Xu, C.; Irfan, M.; Chen, Z. AquaYOLO: Enhancing YOLOv8 for Accurate Underwater Object Detection for Sonar Images. *J. Mar. Sci. Eng.* **2025**, *13*, 73. [CrossRef]
10. Manik, H.M.; Rohman, S.; Hartoyo, D. Underwater Multiple Objects Detection and Tracking Using Multibeam and Side Scan Sonar. *Int. J. Appl. Inf. Syst.* **2014**, *2*, 1–4.
11. Li, L.; Li, Y.; Yue, C.; Xu, G.; Wang, H.; Feng, X. Real-Time Underwater Target Detection for AUV Using Side Scan Sonar Images Based on Deep Learning. *Appl. Ocean Res.* **2023**, *138*, 103630. [CrossRef]
12. Wang, S.; Liu, X.; Yu, S.; Zhu, X.; Chen, B.; Sun, X. Design and Implementation of SSS-Based AUV Autonomous Online Object Detection System. *Electronics* **2024**, *13*, 1064. [CrossRef]
13. Williams, D.P. Fast Target Detection in Synthetic Aperture Sonar Imagery: A New Algorithm and Large-Scale Performance Analysis. *IEEE J. Ocean. Eng.* **2014**, *40*, 71–92. [CrossRef]
14. Galusha, A.; Galusha, G.; Keller, J.; Zare, A. A Fast Target Detection Algorithm for Underwater Synthetic Aperture Sonar Imagery. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII*; SPIE: Bellingham, WA, USA, 2018; pp. 358–370.
15. Williams, D.P., Brown, D.C. New Target Detection Algorithms for Volumetric Synthetic Aperture Sonar Data. In Proceedings of the 179th Meeting of the Acoustical Society of America, Virtually, 7–11 December 2020; AIP Publishing: Melville, NY, USA, 2020.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
17. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual, 6–12 December 2020; pp. 21002–21012.
18. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; AAAI Press: Palo Alto, CA, USA, 2020; pp. 12993–13000.
19. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-Aligned One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, 11–17 October 2021; pp. 3490–3499.
20. Xie, K.; Yang, J.; Qiu, K. A Dataset with Multibeam Forward-Looking Sonar for Underwater Object Detection. *Sci. Data* **2022**, *9*, 739. [CrossRef] [PubMed]
21. Fu, C.; Liu, R.; Fan, X.; Chen, P.; Fu, H.; Yuan, W.; Zhu, M.; Luo, Z. Rethinking General Underwater Object Detection: Datasets, Challenges, and Solutions. *Neurocomputing* **2023**, *517*, 243–256. [CrossRef]
22. Zhao, L.; Yun, Q.; Yuan, F.; Ren, X.; Jin, J.; Zhu, X. YOLOv7-CHS: An Emerging Model for Underwater Object Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 1949. [CrossRef]

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