Journal of *Marine Science and Engineering*

MDPI

Article

E2E-AUD: An End-to-End Adaptive Underwater Detection Framework Integrating Physical Priors and Frequency-Adaptive Learning

Wenhao Zhou [1,2], Junbao Zeng [1,*], Shuo Li [1] and Yuexing Zhang [1]

[1] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; zhouwenhao@sia.cn (W.Z.); shuoli@sia.cn (S.L.); zhangyuexing@sia.cn (Y.Z.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: zengjb@sia.cn

Abstract

Underwater detection is crucial for the autonomous operation of Autonomous Underwater Vehicles (AUVs). However, underwater environments pose significant challenges, including severe image degradation, complex target deformation, and densely distributed small objects. Most existing methods treat image enhancement as an independent preprocessing module and rely on fixed-shape convolution kernels for feature extraction, which often leads to inconsistent optimization objectives and limited capability in handling irregular targets and fine-grained small-object details. To address these issues, we propose an End-to-End Adaptive Underwater Detection framework (E2E-AUD). Specifically, a lightweight image enhancement module, UnitModule, is embedded into the detection network so that enhancement can be jointly optimized with detection and directly serve downstream feature learning. In addition, linear deformable convolution (LDConv) is introduced into the backbone to adaptively model polymorphic targets, while Haar wavelet downsampling (HWD) is adopted to preserve boundary and texture information through frequency-domain analysis. Experiments on the DUO and URPC datasets demonstrate that E2E-AUD achieves superior performance over both general-purpose and underwater-specific detectors. Specifically, on the DUO dataset, our model reaches 86.2% mAP50 and 67.8% mAP50-95, outperforming the recent YOLOv12 by 3.0% and 2.7%, respectively. On the highly turbid URPC dataset, it achieves 84.3% mAP50 and 50.8% mAP50-95, surpassing the competitive underwater-specific detector LEFEN by notable margins in strict localization metrics. Furthermore, E2E-AUD maintains a real-time inference speed of 21.8 FPS with highly constrained computational complexity (9.4 GFLOPs), proving its exceptional balance between detection accuracy and deployment efficiency compared to previous methods.

Keywords: underwater object detection; end-to-end learning; physics-aware enhancement; wavelet-based downsampling; deformable convolution

Academic Editor: Xianbo Xiang

Received: 7 May 2026
Revised: 3 June 2026
Accepted: 5 June 2026
Published: 7 June 2026

1. Introduction

Underwater object detection (UOD) is increasingly important for marine-resource exploration, underwater robotic operations, and ecological monitoring [1]. Accurate and real-time UOD is essential not only for marine-biodiversity conservation but also for Autonomous Underwater Vehicles (AUVs) performing complex tasks [2]. However, underwater scenes are substantially more challenging than terrestrial environments. Although deep convolutional neural network (CNN)-based detectors have achieved strong performance

on terrestrial datasets, their accuracy often deteriorates when they are directly applied to underwater imagery. This degradation arises primarily from three characteristics of the underwater environment. First, wavelength-dependent absorption, together with forward and backward scattering caused by suspended particles, produces color casts, low contrast, haze, and blurred details [3,4]. Second, many underwater targets, including sea cucumbers and starfish, have irregular and deformable morphologies that are difficult to represent using fixed geometric priors [5]. Third, underwater targets are frequently small and densely distributed. Their fine-grained textures can therefore be lost during CNN downsampling, resulting in missed detections [6,7].

Recent studies on underwater optical wireless communication (UOWC) and the Internet of Underwater Things (IoUT) further demonstrate that aquatic optical propagation is strongly affected by water type, wavelength-dependent absorption and scattering, turbulence, and ambient noise [8–11]. Although these studies focus on communication-channel modeling rather than visual object detection, they provide useful physical context for understanding why underwater imagery exhibits spatially varying attenuation, contrast loss, and color distortion. These environmental variations motivate the use of physics-aware enhancement in underwater detection systems.

Existing approaches to these challenges can be broadly divided into two categories: image-enhancement methods and end-to-end detection methods. Traditional enhancement approaches restore underwater images by inverting physical degradation models, such as the Jaffe–McGlamery model [12]. However, their effectiveness depends on accurately estimating optical parameters that are difficult to obtain under real-world conditions. Data-driven enhancement methods based on CNNs or generative adversarial networks (GANs), including WaterGAN [13] and UWCNN [14], learn color correction from synthetic data. Unpaired enhancement and color-transfer methods, such as FUnIE-GAN [15] and Water-Color Transfer [16], reduce the dependence on paired training samples. Nevertheless, enhancement networks are typically optimized independently of the downstream detector. As a result, visually improved images may contain artifacts or feature distortions that are detrimental to detection [17]. A second line of research improves the detector itself through stronger backbones, attention mechanisms, and multi-scale feature fusion [18]. For example, LMFEN [19] preserves fine-grained features during downsampling, whereas ResNet-RS [20] improves the representation of blurred boundaries. However, most detectors still rely on conventional convolutions with fixed sampling grids, which cannot readily adapt to irregular target contours [21]. Moreover, strided convolution and pooling inevitably discard information from small objects during downsampling [22].

These limitations reveal two closely related gaps. First, image enhancement and object detection are usually optimized as separate stages, even though the restored image should ultimately support the downstream recognition task. Second, conventional detection networks remain limited in their ability to preserve small-object details and model deformable underwater targets.

To address these issues, we propose an End-to-End Adaptive Underwater Detection framework (E2E-AUD). The main contributions of this work are as follows:

(1) We embed a lightweight differentiable enhancement unit, termed UnitModule, at the front end of the detector. By combining a physics-inspired underwater degradation model with data-driven learning, UnitModule is jointly optimized with the detection objective, allowing enhancement to directly support discriminative feature extraction.
(2) We replace conventional stem downsampling with Haar wavelet downsampling (HWD). By exploiting multi-resolution wavelet decomposition, HWD preserves and separates high-frequency boundary and texture information, thereby enriching the low-level representation of small targets.

(3) We incorporate linear deformable convolution (LDConv) into the network. LDConv learns dynamic sampling offsets and adapts the receptive pattern to irregular target geometries, improving the representation of underwater organisms with diverse poses and deformable structures.

Experiments on the DUO and URPC datasets show that E2E-AUD improves detection accuracy while maintaining competitive inference efficiency under challenging underwater conditions, including color distortion, target deformation, and dense small-object distributions.

In summary, this paper proposes E2E-AUD, an End-to-End Adaptive Underwater Detection framework designed to simultaneously overcome complex optical degradation and non-rigid target deformation. By unifying physics-aware image enhancement (UnitModule), frequency-domain detail preservation (HWD), and dynamically adaptive feature extraction (LDConv), our approach effectively bridges the gap between image restoration and object detection, achieving state-of-the-art accuracy on multiple challenging underwater datasets. The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed framework and its main components. Section 4 presents the experimental settings and results. Section 5 concludes the paper and discusses future directions.

2. Related Work

2.1. Underwater Image Enhancement and Restoration

The goal of underwater image enhancement and restoration is to recover clear visual information from degraded observations. Over time, they have been deeply transformed in their approach from physical-model-based methods to data-based frameworks and, most recently, to task-based methods.

Physical Model-based Methods: The basis of these methods is the mathematical modeling of the optical imaging process occurring in the underwater environment. Other than the popular Jaffe–McGlamery model or any other simplified version, more significant works like Du et al. [23] proposed an approach, which is based on the theory of Retinex, to separate the illumination and reflectance information; Zhuang et al. [24] applied physical absorption and scattering models in combination to improve the contrast. Furthermore, various variants that achieve restoration by estimating background light and transmission maps [25] have also demonstrated clear physical interpretability. Although these methods have explicit physical interpretation, their performance is strongly influenced by the precision of parameter estimation, which restricts their applicability to complex and dynamic underwater conditions.

Related research on UOWC and IoUT channel modeling has systematically investigated the effects of absorption, scattering, water type, optical turbulence, background noise, and transmitter configuration on underwater light propagation [5–8]. These studies primarily address communication-link performance, including received power, signal-to-noise ratio, and bit error rate. Nevertheless, the reported environmental dependencies are also relevant to underwater visual perception because they explain the physical origins of attenuation, haze, and color distortion. In this work, we use this broader physical context to motivate task-oriented image enhancement, while limiting the scope of the proposed framework to underwater object detection.

Deep Learning-based Methods: Due to the emergence of the data-driven paradigm, CNNs and GANs have been rising in popularity over time. UWCNN [14] was trained with enhancement networks on synthetic datasets and Water-GAN [13] and FDCE-Net [26] were optimized to improve visual quality with the help of adversarial learning. Nevertheless, most of them require either synthetic data or very carefully matched real-world

data, whereas it is very hard to obtain a high-quality pair of underwater images. More significantly, the approaches usually consider pixel-level (e.g., PSNR, SSIM) or perceptual measures as optimization goals. The improved outcomes might look appealing to the eye but be mismatched in terms of the discriminative feature representation needed to perform downstream high-level vision tasks [17].

Task-Oriented Enhancement Methods: Scientists have recently started using the approach of task-oriented enhancement methods. The strategies may consist of joint training of the module of enhancement (task network [27]) or addition of constraints specific to particular tasks to the loss function [28,29]. Nevertheless, current joint optimization models typically lack lightweight modeling of underwater color degradation mechanisms. In addition, more than half of the models incorporate the enhancement module as a relatively intricate sub-network, which adds to the model complexity and the complexity of training.

2.2. Object Detection and Its Underwater Applications

The two-stage paradigm (Faster R-CNN series) and the one-stage paradigm (YOLO series, SSD) form the latest mainstream detectors. Of these, the most famous of all, the YOLO series, with its outstanding performance on the task of achieving a good balance between accuracy and speed [30,31], has become a model of choice in a computationally restricted environment like underwater robots. Furthermore, DETR [32] and its variants have also entered the detection domain as Transformer architectures with strong global modeling ability. Still, they are not used in practice due to their great computational complexity.

To overcome the issues of the underwater setting, scientists have made multi-fold advances in detection networks. They involve creating a multi-scale feature fusion structure to address the scale changes (e.g., S-FPN [33,34]), adding an attention mechanism to concentrate on blurred objects [35], and taking advantage of data augmentation to create artificial degraded environments with enhanced robustness [36]. However, most of these enhancements have been limited to the backend of the detection network. Regarding input image quality degradation, they typically adopt an attitude of "offline preprocessing" or "passive adaptation," failing to achieve a deep coupling of enhancement and detection at the architectural level.

2.3. Network Design for Deformation Adaptation and Detail Preservation

Two significant challenges in the current detection problems are variable body postures of non-rigid underwater organisms and the decay of boundaries of small things. The ability to model geometrically complicated deformations is greatly improved with Deformable Convolutional Networks (DCN), which can introduce learnable offsets at every sample point of the convolution kernel to allow the receptive field to adaptively fit into the target structures [37]. The latest DCNv4 [38] has achieved further breakthroughs in computational efficiency. In order to overcome information loss due to downsampling, scientists have started taking advantage of the multi-resolution nature of wavelet transformations to break down and save details of high frequencies in frequency space, as it was used in super-resolution [39,40] and downsampling layers that detect objects [41]. These techniques are efficient to protect edge textures of small objects against total removal when contracting the feature maps.

3. Methodology

3.1. Overall Architecture

Underwater object detection can be formulated as a mapping learning problem within a degraded domain D. Given a degraded underwater image $X \in \mathbb{R}^{H \times W \times 3}$, the image not only contains the target objects O but is also subject to nonlinear interference from depth-

dependent transmission rates t and background light A. Traditional approaches attempt to address this via independent enhancement networks $E(\cdot)$ and detection networks $D(\cdot)$, formulated as $\hat{y} = D(E(I))$.

However, this cascaded paradigm neglects the distributional alignment between enhanced features and detection-oriented features. In order to overcome this drawback, an integrated framework is proposed in this paper, known as E2E-AUD. As shown in Figure 1, with YOLOv13 [42] as the base, the framework reconstructs the detection flow by tightly integrating three major components:

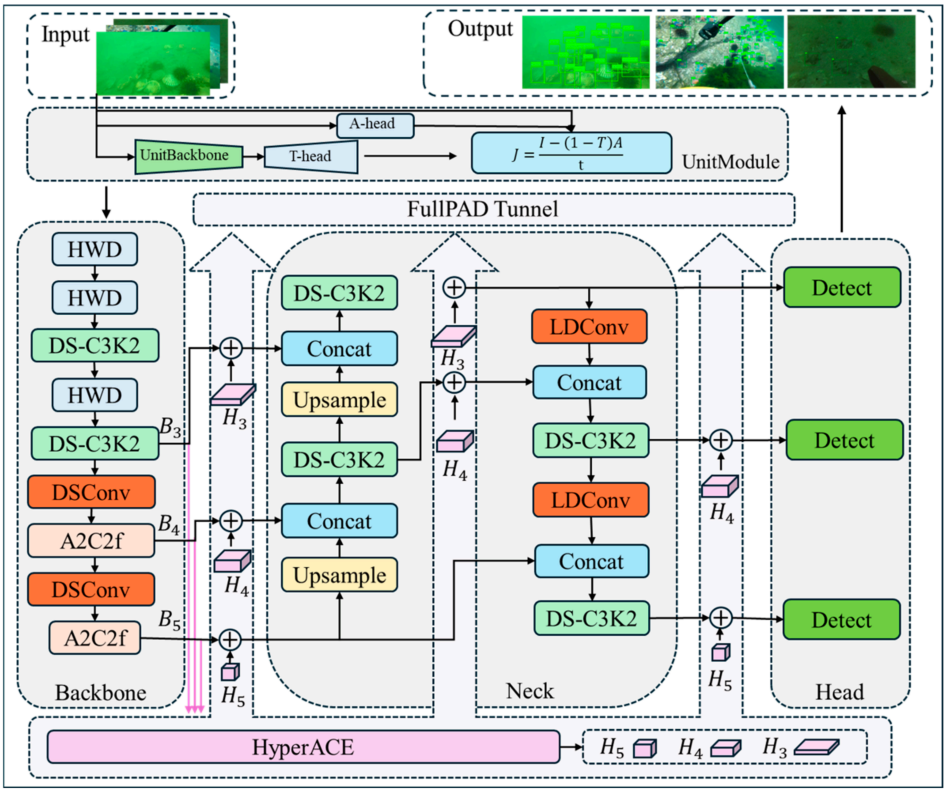


Figure 1. Overall architecture diagram of the E2E-AUD model.

Input-level: The introduced UnitModule [43] will integrate the reverse implementation of the underwater optical imaging model into the front end of the network. This module adaptively produces latent clear representations that are specific to the next detector by unsupervised joint training.

Stem-level: We are using HWD [44] instead of the conventional strided convolution. Using the orthogonal decomposition nature of the wavelet transform, HWD provides a lossless spatial high-frequency coding in frequency domain channels as opposed to high-rate downsampling, where the problem of feature dissipation of small objects is significantly addressed.

Feature-level: At the deep layers of the backbone network, we use LDConv [21]. LDConv adaptively learns how to capture complex geometric deformations of non-rigid organisms through a mechanism of linearly increasing sampling points.

The model takes raw underwater imagery as an input, where it is initially exposed to the adaptive quality improvement of the UnitModule. Afterward, important textures are saved on the Haar wavelet downsampling layer. Lastly, deformation robust features are obtained by the backbone network using the LDConv and target localization and categorization are performed through the neck and detection heads with multi-scale prediction heads.

3.2. UnitModule

UnitModule is a lightweight and differentiable image enhancement unit. The main concept of this unit is to combine a classical model of underwater optical imaging with the deep neural networks. It is able to alleviate the sub-optimal bottleneck of the conventional approach known as the enhance-then-detect paradigm by allowing joint optimization with the detector.

3.2.1. Physical Modeling and Network Implementation

The degradation process of underwater images is typically described by the Jaffe–McGlamery underwater optical imaging model. This model decomposes the total irradiance $I(x)$ received by the detector into the direct attenuation component and the background scattering component:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

As shown in Equation (1), x denotes the pixel coordinates, $J(x) \in [0, 1]$ represents the clear scene radiance to be restored, $A \in (0, 1)$ is the global background light (often referred to as background light), and $t(x) \in (0, 1)$ is the wavelength-dependent medium transmission map. By constraining $t(x)$ and A to the open interval $(0, 1)$ via Sigmoid activations in the network implementation, the model inherently avoids the numerical instability caused when $t(x) \to 0$.

To recover $J(x)$ from the degraded image $I(x)$, the theoretical inverse transformation is formulated as:

$$J(x) = \frac{I(x) - A}{t(x)} + A \tag{2}$$

Equation (2) provides an analytical inversion of the underwater imaging model. However, directly applying this inverse transformation during network training may cause numerical instability when the estimated transmission value approaches zero. Moreover, accurately estimating the physical transmission map from a single underwater image remains challenging because the degradation process is affected by complex and spatially varying environmental conditions.

To address these issues, we adopt a learnable blending mechanism inspired by the composition form of the underwater imaging model. Rather than treating Equation (3) as a mathematically equivalent reformulation of Equation (2), we use it as a numerically stable approximation:

$$J(x) = I(x) \cdot \left(1 - \hat{t}(x)\right) + \hat{A} \cdot \hat{t}(x) \tag{3}$$

In Equation (3), the network does not directly predict the physical transmission rate but instead predicts an enhancement weight map $\hat{t}(x)$ and a background light correction vector $\hat{A}$. This design enables the module to adaptively balance original pixel information with background correction information, thereby effectively removing haze while avoiding the introduction of artificial artifacts.

As shown in Figure 2, the UnitModule consists of three collaborative sub-networks: the UnitModule Backbone (UnitBackbone), the Transmission Head (T-Head), and the Atmosphere Head (A-Head).

UnitBackbone: To be more precise, the UnitBackbone consists of a lightweight architecture, namely a Stem containing two convolution layers (stride 2, consisting of channel numbers C_s1 = 32 and C_s2 = 64) and two Large Kernel Blocks (LK Blocks). The LK Block is inspired by RepLKNet [45] that uses 1×1 convolution to compress channels and then applies a K $\times$ K depth-wise separable convolution. This design improves the receptive field and nonlinear expressive ability via structural re-parameterization.

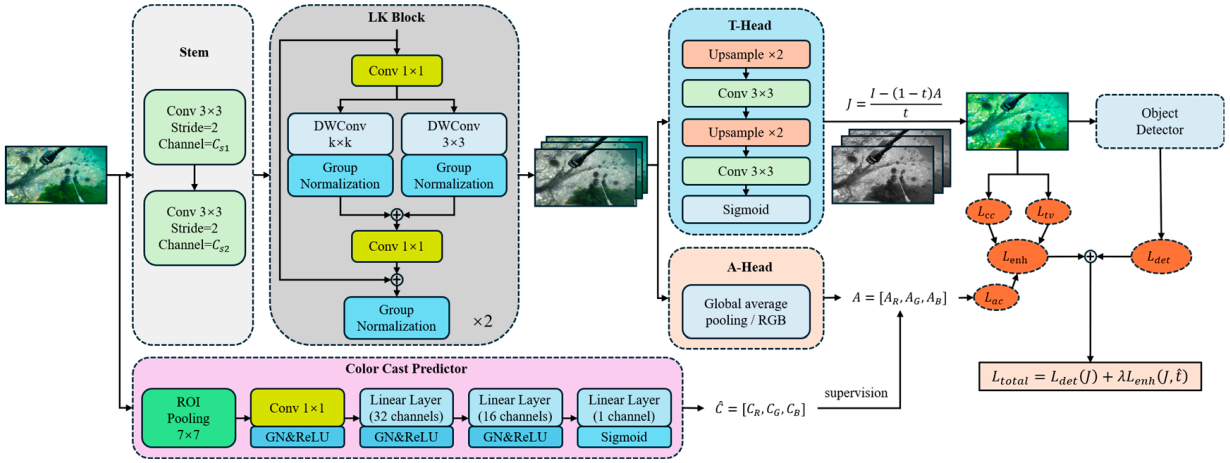


Figure 2. Architecture diagram of UnitModule.

T-Head: The objective of T-Head is to acquire the local degradation features that are related to depth. It receives the output features and forecasts the transmission map by two layers of upsampling and convolutions. A Sigmoid activation function is used in the last output layer to limit the output feature map to the (0, 1) interval.

A-Head: The A-Head is used to predict the overall optical properties of the image. First, it compresses the input features into a $1 \times 1 \times C$ global descriptor using Global Average Pooling (GAP) and then it consists of two Fully Connected (FC) layers. Lastly, it produces a 3-dimensional vector, which represents the background light parameters in the R, G, and B channels separately.

3.2.2. Unsupervised Joint Loss Function

Due to the scarcity of paired real-world "underwater-clear" datasets, supervised pixel-fidelity losses (e.g., L1 or MSE) cannot be applied. Therefore, the UnitModule adopts an unsupervised learning strategy. Its optimization objective extends beyond mere visual clarity; instead, through joint training, it aims to generate enhanced images J that minimize the task loss of the detector. The total loss function is formulated as a linear combination of the detection task loss and a physical-prior-guided unsupervised enhancement loss:

$$L_{total} = L_{det}(J) + \lambda L_{enh}(J, \hat{t}) \tag{4}$$

As shown in Equation (4), $L_{det}(J)$ represents the localization and classification loss of the detector, and $L_{enh}(J, \hat{t})$ denotes the specifically designed unsupervised enhancement loss. As formulated below, L_{enh} comprises three components that constrain the enhancement process from the dimensions of color, texture, and category, respectively:

$$L_{enh} = w_1 L_{cc} + w_2 L_{tv} + w_3 L_{ac} \tag{5}$$

As shown in Equation (5), $(w_1, w_2, w_3) = (0.1, 0.01, 0.1)$, these coefficients are non-negative balancing hyperparameters used to account for the different numerical scales of the auxiliary loss terms. They are not probability weights and are therefore not required to sum to one.

Color Consistency Loss (L_{cc}): As shown in Equation (6), this loss is fundamentally derived from the classical Gray-World Hypothesis, which posits that the average reflectance of surfaces in a natural scene tends to be achromatic (gray). By mathematically enforcing the average intensity of the RGB channels (μ_{J^r}, μ_{J^g}, μ_{J^b}) to reach equilibrium, this term effi-

ciently neutralizes the severe bluish-green color casts typical of underwater environments without requiring reference images."

$$L_{cc} = \left(\mu_{J^r} - \mu_{J^g}\right)^2 + \left(\mu_{J^r} - \mu_{J^b}\right)^2 + \left(\mu_{J^g} - \mu_{J^b}\right)^2 \tag{6}$$

Texture Smoothness Loss (L_{tv}): As shown in Equation (7), estimating the transmission map $\hat{t}(x)$ in unconstrained scenarios often leads to high-frequency noise or checkerboard artifacts. To mathematically mitigate this, we introduce a Total Variation (TV) constraint. By minimizing the gradient operators ∇_h and ∇_v, the TV loss enforces piecewise smoothness in the estimated transmission map while preserving sharp structural edges, which is critical for maintaining object contours for the downstream detector.

$$L_{tv} = \sum_{x} \left(\left| \nabla_h \hat{t}(x) \right| + \left| \nabla_v \hat{t}(x) \right| \right) \tag{7}$$

As shown in Equation (7), ∇_h and ∇_v denote the gradient operators in the horizontal and vertical directions, respectively.

Assisting Color Cast Loss (L_{ac}): As defined in Equation (8), this term utilizes the pseudo-labels y_{pseudo} generated by the Color Cast Predictor. It is calculated as the cross-entropy loss between the predicted probabilities and the pseudo-labels:

$$L_{ac} = -\sum_{k=1}^{K} y_{pseudo}^{(k)} \log\left(p_{ccp}^{(k)}\right) \tag{8}$$

In Equation (8), $p_{ccp}^{(k)}$ denotes the predicted probability output by the Color Cast Predictor (CCP), indicating the likelihood that the input image belongs to the k-th color cast category. K represents the total number of predefined color cast categories, and $y_{pseudo}^{(k)}$ is the corresponding ground-truth probability from the pseudo-label generated by the UCRT strategy.

3.3. Haar Wavelet Downsampling

Standard convolutional downsampling inevitably leads to the loss of high-frequency information exceeding the Nyquist frequency. To address the characteristic weak textures of tiny underwater objects, we introduce a HWD module based on Multiresolution Analysis (MRA), the structure of which is illustrated in Figure 3.

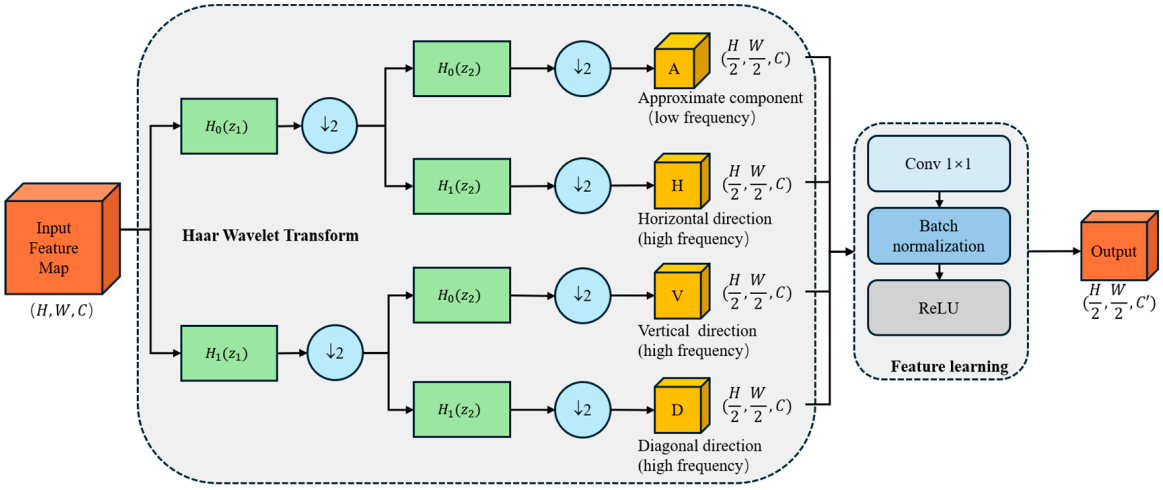


Figure 3. Framework diagram of HWD module. H_0 represents low-pass filtering, H_1 represents high-pass filtering, and z_1 and z_2 represent the two spatial dimensions in the image respectively.

The Haar wavelet transform is founded on the linear combination of the one-dimensional scaling function $\phi(x)$ and the wavelet function $\psi(x)$; it should be noted that rather than describing physical wave propagation in an aquatic medium, these are purely mathematical 1D scalar functions used as discrete filters in digital image processing. They are applied separably to the spatial dimensions of the feature map:

$$\phi_1(x) = \frac{1}{\sqrt{2}}\phi_{1,0}(x) + \frac{1}{\sqrt{2}}\phi_{1,1}(x) \tag{9a}$$

$$\psi_1(x) = \frac{1}{\sqrt{2}}\phi_{1,0}(x) - \frac{1}{\sqrt{2}}\phi_{1,1}(x) \tag{9b}$$

As shown in Equation (9a,b), $\phi_{jk}(x) = \sqrt{2^j}\phi\left(2^j x - k\right)$ defines the multi-scale basis functions. For a two-dimensional feature map $X \in \mathbb{R}^{H\times W\times C}$, the transformation generates four sub-band components:

$$\text{Haar}(X) = \{A, H, V, D\} \tag{10}$$

Here in Equation (10), $A \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ represents the low-pass filtering result, preserving the global structure, while $H, V, D \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ capture high-frequency edge textures in the horizontal, vertical, and diagonal directions, respectively. Unlike traditional image processing that outputs four independent images, we adopt a "Space-to-Channel Rearrangement" strategy, concatenating the calculated four components along the channel dimension: $X_{encoded} = \text{Concat}(A, H, V, D)$. Consequently, the dimensions of the feature map become $\frac{H}{2} \times \frac{W}{2} \times C$, and the spatial high-frequency information is explicitly encoded into channel features.

To fuse features of different frequencies and regulate the channel dimension, we introduce a 1×1 convolution layer after the encoding block, followed by Batch Normalization (BatchNorm) and the SiLU activation function, as shown in Equation (11):

$$Y_{out} = \sigma\left(\text{BN}(\text{Conv}_{1\times1}(X_{encoded}))\right) \tag{11}$$

This convolution layer acts as a "learnable frequency selector." During backpropagation, the network can adaptively adjust the attention weights assigned to the low-frequency structure (A) or high-frequency edges (H, V, D) according to the specific requirements of the underwater detection task. This mechanism effectively suppresses noise interference caused by suspended particles while simultaneously enhancing the texture response of the targets.

3.4. LDConv

Most of the targets of detection in underwater settings consist of non-rigid species and have very changeable geometric patterns and irregular shapes. The Fixed-Square Sampling Grid limits the Traditional Standard Convolution to a fixed-size square grid that would not allow efficient alignment of the effective receptive field with the actual boundary of the target. Therefore, substantial background noise is added in the feature extraction stage, which causes reduction in the accuracy of boundary location.

In order to address these geometric restrictions, we present LDConv, an innovative operator that can perfectly model an irregular target using dynamic sampling point offset, as shown in Figure 4. The most important innovation of LDConv is that it separates lies in the decoupling of the convolutional kernel size from the number of parameters. This allows a linear increase in the number of parameters rather than a quadratic increase in conventional convolutions and thus becomes very effective in capturing the intricate variation in poses of underwater creatures.

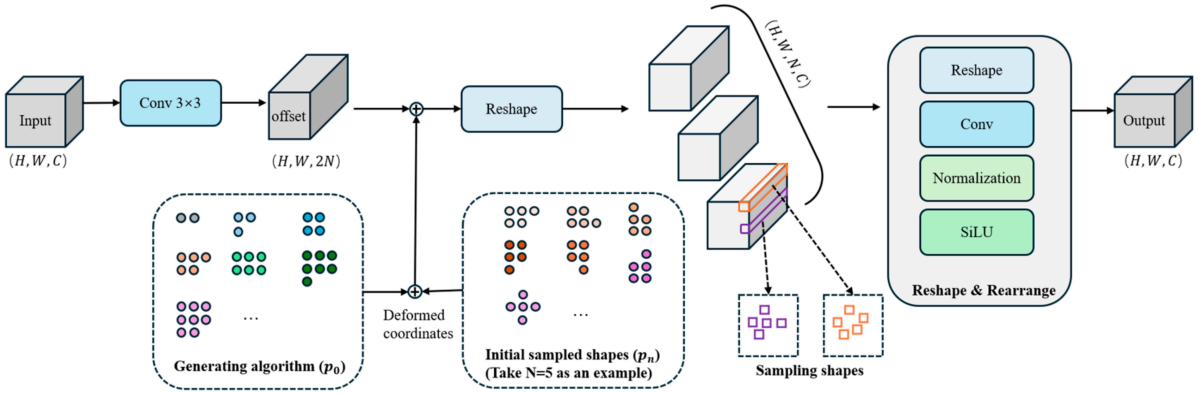


Figure 4. Framework diagram of LDConv module.

For a given input feature map $X \in \mathbb{R}^{H \times W \times C}$, the output of LDConv at position p_0 is defined as follows:

$$y(p_0) = \sum_{n=1}^{N} w_n \cdot X(p_0 + p_n + \Delta p_n) \tag{12}$$

As shown in Equation (12), N denotes the total number of sampling points, w_n represents the weight of the n-th sampling point, p_n is the predefined initial coordinate offset, and Δp_n is the learnable offset dynamically derived from the feature map.

Unlike the Deformable Convolutional Network (DCN), which fixes the number of sampling points at k^2, LDConv can have any arbitrary integer. This means that as the parameter count P increases, it does so linearly with N (P $\propto$ N). This flexibility provides the network with the ability to acquire more precise feature representations by changing N without suffering significant computational costs.

In order to allow irregular convolutional kernels to have structured sampling grids, LDConv uses an arbitrary-sized convolution algorithm to create the initial sampling coordinates P_n. This algorithm first constructs a regular sampling grid and then produces irregular sampling coordinates of other points, subsequently joining them to obtain a combined sampling grid. The initial sampling coordinates of LDConv are generated using the structured procedure provided in Appendix A. Given the number of sampling points N, the procedure constructs the initial coordinate set P_N, which is subsequently combined with the learned offsets before bilinear interpolation and feature aggregation.

In Algorithm A1, the parameter C guarantees that the generated coordinates are symmetrically distributed and normalized around the origin $(0, 0)$. Furthermore, by explicitly defining a random seed S, we ensure that the continuous uniform distribution used for the irregular offsets (x_r, y_r) produces a deterministic geometric layout across identical configurations, facilitating strict reproducibility.

The feature extraction process of LDConv involves a highly coordinated data flow, taking $N = 5$ in Figure 4 as an example. As illustrated, the process bifurcates into two parallel streams from the input feature map $X \in \mathbb{R}^{H \times W \times C}$. First, in the coordinate generation stream, initial sampling shapes (P_n) are generated based on the value of N using Algorithm A1. The focus here is to generate the corresponding sampling coordinates on the feature map for a kernel with N parameters. Second, in the offset learning stream, the learnable spatial offsets for the respective kernel are obtained via a standard 3×3 convolution operation, yielding an offset field of dimension $(H, W, 2N)$. This tensor is then reshaped to explicitly match the N spatial coordinates (x, y). These offsets are added to the original coordinates to generate new, convolution-specific sampling coordinates (Deformed coordinates in Figure 4). This step emphasizes the generation of diverse sampling shapes for convolutions at different spatial positions on the feature map. Finally, in the feature

aggregation phase, features at the corresponding deformed continuous locations are acquired through bilinear interpolation and resampling. These resampled features are then processed by a sequence of operations, including a convolution layer, Normalization, and a SiLU activation function, before being reshaped and rearranged to output the final feature map of dimension (H, W, C).

In our proposed model architecture, we implement a hierarchical substitution strategy: In the shallow layers (high resolution): we employ LDConv with $N = 7$ to preserve fine-grained detailed features. In the deep layers (low resolution): we utilize LDConv with $N = 9$ to enhance semantic perception and global modeling capabilities.

4. Experiments and Results

4.1. Implementation Details

The hardware and software configurations are essential elements during the investigation of deep learning that have a direct effect on the model performance and experimental reproducibility. To provide a clear and strict experimental setting, a summary of the detailed hardware and software specifications applied in the current paper is given in Table 1.

Table 1. Hardware and software configurations.

Configuration	Parameter
Operating system	Ubuntu 22.04
Accelerated environment	Cuda 12.1
Language	Python 3.11
Framework	Pytorch 2.3.0
GPU	RTX4090 (24 GB)

The purpose of the standard training protocol followed by us was to make it easier to compare the proposed UW-YOLO-Bio with any of the baseline approaches in a fairer and more equitable manner. This protocol was designed carefully so that it could work as per the best practices in object detection research and be optimized for use in underwater situations. The configuration includes some major hyperparameters, including the training epochs, resolution of input, data augmentation strategies, and optimizers. Through such standardization of these variables, we reduced possible confounders and ensured validity of the comparative findings. Table 2 contains a list of detailed hyperparameters used in training.

Table 2. Training hyperparameters and configurations.

Parameter	Value	Notes
Epoch	200	
Input Size	640 × 640	After resizing
Augmentation	Mosaic (dose = 10), Random Flip, HSV Hue/Saturation/Value/adjustment	Standard YOLO augmentation
Optimizer	SGD	Momentum = 0.9 Weight Decay = 0.0005
Learning Rate Scheduler	Linear decay from 0.01 to final LR = 0.01 × 0.01	
NMS IoU Threshold	0.7	
Validation IoU Threshold	0.5	For mAP calculation

4.2. Comprehensive Evaluation Framework

To rigorously benchmark the proposed detector, several evaluation metrics are utilized based on the confusion matrix. True Positives (TP), False Positives (FP), and

False Negatives (FN) serve as the fundamental variables. These metrics correspond to Equations (13)–(16), respectively.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

Precision evaluates the proportion of True Positives among all predicted positives, while Recall assesses the model's ability to retrieve all ground-truth instances.

$$AP = \int_0^1 P(R)dR \approx \sum_{i=0}^{n-1} P_i \Delta R(i) \tag{15}$$

The Precision–Recall (P-R) curve is plotted to visualize the equilibrium between P and R. The AP value, calculated as the area under the P-R curve, represents the detection accuracy for an individual class, where n represents the total number of discrete recall thresholds. The index i denotes the i-th threshold operating point on the P-R curve. P_i is the measured precision at the i-th point, and $\Delta R_i = R_i - R_{i-1}$ represents the change in Recall between consecutive points.

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{16}$$

As the primary metric for object detection, mAP is computed by taking the arithmetic mean of AP scores across all detected categories, offering a comprehensive measure of the algorithm's multi-class performance.

4.3. Datasets

The proposed method was tested on two standard underwater datasets, DUO [46] and URPC [47]. These datasets contain various underwater problems such as hazing effects, significant color distortion, light interference as well as intricate background clutter, making them good validating datasets.

DUO (Detection in Underwater Objects): This publicly accessible benchmark is particularly tailored to underwater object detection challenges. It has 7782 images, which contain 74,515 instances of annotations divided into four main marine specimens. The dataset has been divided as follows: training set (5447 images), validation set (778 images) and test set (1557 images). DUO is especially useful when it comes to evaluating model performances in relation to low contrast, multi-scale target variants, and clutter and contains multi-resolution images (ranging from 586 × 482 up to 3840 × 2160 pixels). It can be extensively applied in creating visual systems of Autonomous Underwater Vehicles (AUVs). In the context of the physical imaging model introduced in Equation (1), the images in the DUO dataset exhibit highly variable background light vectors (A), typically dominated by blue and green wavelengths due to specific depth absorption. Furthermore, the varying imaging distances result in fluctuating transmission rates ($t(x)$), leading to the diverse multi-scale target contrast and clutter observed in this dataset. The dataset is available at: https://github.com/chongweiliu/DUO (accessed on 3 June 2026).

URPC (Underwater Robot Perception Challenge): The URPC dataset has been generated based on realistic activities within the field of the Zhangzidao Marine Ranch in Dalian, China. The dataset consists of 5543 photos, which are annotated using four economically important marine animals. The information divides the training (3880 images), validation (554 images), and test (1109 images) sets. Its application includes areas such

as smart use and management of fish farms and automatic harvesting of fish. The in situ images represent serious problems, particulate interference (marine snow) and heavy object overlap. This set will be used as a major tool that would be used to train the perception systems of underwater robots in working conditions. Mapped to our physical formulas, the URPC dataset represents a scenario of extreme optical degradation. The severe particulate interference (marine snow) physically translates to an intense and unpredictable scattering component ($A(1 - t(x))$), while the heavy turbidity forces the global transmission rate ($t(x)$) to diminish sharply, thereby burying object structural details within the background noise. The dataset can be accessed at: https://github.com/mousecpn/DG-YOLO (accessed on 3 June 2026).

4.4. Ablation Study

4.4.1. Parameter Sensitivity Analysis of LDConv

In order to ensure that the proposed LDConv is effective and find the best hyperparameters, quantitative experiments were performed on the DUO dataset, with the baseline architecture being YOLOv13n. Specifically, we investigated the influence of the number of sampling points, denoted by N, on detection performance. Here, N represents the number of sampling positions used by LDConv rather than the side length of a conventional square convolution kernel. In a conventional $K \times K$ convolution kernel, the number of sampling points is fixed as K^2. In contrast, LDConv allows N to be set as an arbitrary positive integer. Therefore, the number of parameters increases approximately linearly with N rather than quadratically with the side length of a square convolution kernel. The results are summarized in Table 3. The results demonstrate a positive correlation between detection accuracy (mAP_{50}) and the number of sampling points. As N increases from 5 to 11, the model's performance shows a steady upward trend, peaking at 82.9% for both $N = 10$ and $N = 11$. This represents a significant improvement over the baseline's 81.1%. This enhancement is attributed to the expanded effective receptive field provided by a larger N, which augments the network capability to model the complex geometric deformations of irregular underwater targets. Conversely, inference speed (FPS) exhibits a strong negative correlation with N. The FPS suffers a precipitous decline from 31.8 at $N = 5$ to 17.8 at $N = 11$ due to the increased computational cost of offset generation. Notably, the parameter count (Params) increases only marginally in a linear fashion with N, validating the parameter efficiency of the LDConv design.

Table 3. Sensitivity ablation experimental data with different sampling points (N).

LDConv	Precision (%)	Recall (%)	AP50 (%)	AP50-95 (%)	Params (%)	FPS
baseline	84.1	73.3	81.1	60.8	2.57	18.8
5	81.2	69.9	79.2	58.6	2.507	31.8
6	83.2	71.2	80	59.4	2.518	29.7
7	86.8	74.5	81.5	61	2.529	29.1
8	91.1	72	82.4	60	2.541	25.3
9	90.9	75.5	82.7	62.3	2.552	23.1
10	91.1	75.6	82.9	62.5	2.565	20.5
11	91.3	75.8	82.9	62.5	2.588	17.8

The experimental data reveals a significant accuracy–speed trade-off. We identified two distinct equilibrium points at $N = 7$ and $N = 9$, where the model achieves an optimal balance between performance gains and computational overhead. Consequently, to maximize the strengths of LDConv, we adopt a Hierarchical Substitution Strategy in our final architecture:

- Shallow Layers (High Resolution): We employ LDConv with $N = 7$. Since shallow layers are responsible for extracting fine-grained spatial details, a moderate N preserves crucial texture information while ensuring high inference efficiency.
- Deep Layers (Low Resolution): We utilize LDConv with $N = 9$. In deep layers where features are semantically rich but spatially coarse, a larger N is deployed to expand the receptive field and enhance global semantic perception.

4.4.2. Ablation Experiments of Three Different Modules

In order to assess the contribution of individual elements and their synergistic interaction in the context of the proposed elements in the framework of the so-called E2E-AUD, we performed a thorough ablation experiment on the DUO and URPC datasets. The research was conducted on three main modules of the module: the physics-aware UnitModule, the frequency-domain HWD, and the geometry-adaptive LDConv.

The identical protocols were used to train all models, strictly following the hyperparameters and hardware specifications described in Section 4.1. The baseline model was established using YOLOv13n and then each successive integration of respective proposed modules. In order to make sure that the results were statistically reliable, the results presented are averages of several independent experiments. The quantitative comparison results are listed in Table 4.

Table 4. Results of ablation experiment data from three different modules.

Dataset	UnitModule	HWD	LDConv	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	FPS
DUO				84.1	73.3	81.1	60.8	18.8
	√			87.2	73.5	83.5	64.1	18.2
		√		89.4	76	83.1	62.5	18.7
			√	90.9	75.5	82.7	62.3	**23.1**
	√	√		89.8	75.9	85.5	65.1	18.6
	√		√	91.1	75.8	84.9	66.5	22.1
		√	√	89.9	76	85.8	67.2	22.7
	√	√	√	**91.2**	**76.1**	**86.2**	**67.8**	21.8
URPC				81.0	75.2	77.9	45.8	18.2
	√			83.1	76.0	80.5	46.7	17.9
		√		82.5	75.9	81.1	46.2	18.1
			√	83.0	77.2	83.5	48.2	**20.6**
	√	√	√	**84.5**	**78.6**	**84.3**	**50.8**	20.2

(Note: Rows 5–7 for URPC are omitted for brevity in this snippet but strictly follow the same trend as DUO in the full analysis.) Bold indicates optimal performance. A check mark symbol indicates that the corresponding module is used.

The implementation of the UnitModule was associated with a significant increase in detection accuracy, as shown in Table 4. On DUO, UnitModule alone increased precision by 3.1 and mAP50 by 2.4 relative to the baseline. This performance improvement confirms our theoretical argument in Section 3.2 that color distortion and hazes can be effectively reduced using the inverse physical imaging model. Through the creation of the latent clear representation $J(x)$ (as formulated in Equation (3)), the UnitModule matches the distribution of features of degraded inputs with the domain of the detector. By adaptively balancing the original pixel information with the estimated transmission map $t(x)$ and background light A, it eliminates False Negatives related to low contrast by a large margin. Although a slight decrease in FPS (from 18.8 to 18.2) was observed due to the

additional computational overhead of the T-Head and A-Head, the trade-off is justified by the substantial accuracy gains.

The HWD module proved to be very effective in reconstructing fine-grained data. In the case of the URPC dataset, which has a feature set of densely spaced small targets, the HWD module autonomously increased mAP50 by 3.2% on URPC (77.9% to 81.1%). It is verified that the high-frequency directional components (H, V, D from Equation (10))—which capture crucial edges and textures—are encoded into the channel dimension without loss, thereby preventing the feature dissipation typically caused by strided convolution. It is worth noting that HWD had the largest Recall improvement rate of single modules on DUO (+2.7%), implying that the frequency-sensitive approach is the most important towards reducing the number of small marine life missed during detection.

One of the most significant findings of the experiment was that the LDConv has a double benefit. The benefits are robustness and efficiency improvement. LDConv independently improved mAP50-95 by 1.5% on DUO, proving its effectiveness in modeling non-rigid biological deformations through the learnable dynamic sampling offsets (ΔP_N defined in Equation (12)). In contrast to traditional deformable convolutions that generally lead to higher latency, our implementation of LDConv increased the inference speed. The FPS of the DUO dataset increased by 18.8 (Baseline) to 23.1 (LDConv). The empirically confirmation of this validates the efficacy of the linear parameter growth mechanism ($P \propto N$) presented in Section 3.3, which is the concept of detaching kernel size with parameter count, enabling larger receptive fields with less computational complexity.

The given modules exhibit a definite synergistic effect of $1 + 1 > 2$. UnitModule + HWD is concentrating on Quality + Detail. An mAP50 of 85.5% was attained with this model by first optimizing the image quality, followed by the preservation of the restored details as the downsampling is done. The ultimate model consists of the three modules performed at state-of-the-art (SOTA). The full model reached a precision of 91.2% and an mAP50 of 86.2 on the DUO dataset, which is an overall increase in precision and an increase in mAP50 of 7.1% and 5.1% respectively over the baseline. The improvements across the URPC data (mAP50 improved by 6.4%, from 77.9% to 84.3%) were consistent, indicating the power of generalization of the framework to various degrees of water quality and target distribution.

To sum up, the proposed method will not only greatly improve the accuracy of the detection due to physics-inspired improvement and frequency-domain analysis but also preserve a high real-time rate because of the effective construction of LDConv. It confirms that the E2E-AUD is applicable to be used in practice on AUVs.

4.5. Comparative Analysis with State-of-the-Arts

In order to benchmark the effectiveness of the E2E-AUD in a rigorous manner, we have used thorough comparative analysis with two types of baseline methods:

Generic Object Detectors: The traditional two-stage architecture (Faster R-CNN [48], Cascade R-CNN [49]) and the current state-of-the-art one-stage architecture, YOLO series (YOLOv8 [50], YOLOv10 [51], YOLOv11 [52], and the latest YOLOv12 [53]).

Underwater-Specific Detectors: These include LEFEN [19], DJL-Net [54], ERL-Net [55], Boosting R-CNN [56] and GCC-Net [57], all of which are designed to work in the underwater setting specifically.

All comparative experiments were conducted with DUO and URPC datasets using the same training protocols and hardware configurations to be able to fairly evaluate the results.

4.5.1. Performance on DUO Dataset

The DUO dataset is characterized by complex lighting and diverse target scales. The quantitative results are summarized in Table 5.

Table 5. Comparison results of different models on the DUO dataset.

Method	Type	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	FPS
Faster R-CNN	Generic	79.1	68.2	80.5	59.2	11.4
Cascade R-CNN		82.4	71.9	80.9	60.2	11.6
YOLOv8		85.7	74.4	82.9	62.7	**21.9**
YOLOv10		85.9	74.3	83.2	62.9	20.1
YOLOv11		83.4	75.7	83.7	64.1	20.5
YOLOv12		84.9	75.4	83.2	65.1	19.5
LEFEN	Underwater	89.2	75.1	86.0	66.3	19.4
DJI-Net		88.7	75.6	84.2	65.5	18.1
ERL-Net		86.5	74.0	81.4	67.5	19.1
Boosting R-CNN		90.9	75.1	80.6	66.9	12.5
GCC-Net		89.1	74.2	81.6	66.3	16.4
ours		**91.2**	**76.1**	**86.2**	**67.8**	**21.8**

Bold indicates optimal performance.

As we have presented in Table 5, our model exhibits a significant advantage over generic detectors. To be more precise, it reaches an mAP50 of 86.2% and an mAP50-95 of 67.8%, which is 3.0% and 2.7% better than the most recent YOLOv12, respectively. Although two-stage models such as Cascade R-CNN may have good accuracy (in terms of mAP50 80.9%), they cannot perform at real-time speeds (11.6 FPS). By comparison, our model has an extremely high inference rate of 21.8 FPS, similar to YOLOv8 (21.9 FPS) but with much greater precision (+3.3% on mAP50). It means that our approach strikes a better balance between accuracy and latency.

When compared to the domain-specific algorithms, we have demonstrated that our E2E-AUD has greater robustness and precision. While the LEFEN may be a tough competitor with an mAP50 of 86.0%, our model has outperformed it with all the main metrics, with 2.0-percentage points higher precision (91.2% versus 89.2%) and 1.5-percentage points higher mAP50-95 (67.8% compared to 66.3%). This means that our physics-aware UnitModule can also efficiently mitigate the domain shift due to changing water turbidity, and allows the detector to more confidently recognize objects among background noise. It is of note that ERL-Net demonstrates very high localization ability (mAP50-95 of 67.5). Nevertheless, it shows much worse performance in mAP50 (81.4), which is 4.0 percentage points lower than our model. This indicates the fact that while ERL-Net fits well to easy targets, it fails to find many of the harder ones. On the other hand, our model is the best performer in terms of both measures, which confirms that the geometry-adaptive LDConv accurately identifies non-rigid biological targets, which are often not detected by conventional convolution kernels. Moreover, our model does it without sacrificing speed. At an inference rate of 21.8 FPS, it is significantly faster than Boosting R-CNN (12.5 FPS) and GCC-Net (16.4 FPS), which makes it suitable to use with latency-sensitive underwater robots.

4.5.2. Performance on URPC Dataset

The URPC dataset presents challenges such as dense small object distribution and heavy turbidity. The results are presented in Table 6.

As shown in Table 6, our model has very good generalization abilities, with optimal results across all the available metrics: precision (84.5%), Recall (78.6%), mAP50 (84.3%) and mAP50-95 (50.8%). It is also interesting that our approach is significantly better (+3.1%) than YOLOv12 in terms of mAP50 and is slightly superior (+2.7%) in the case of mAP50-95. This large discrepancy indicates the weakness of standard CNN designs to manage extreme underwater degradation. We have an effective UnitModule solution to counteract such environmental disruptions so that the detector can perceive finer details.

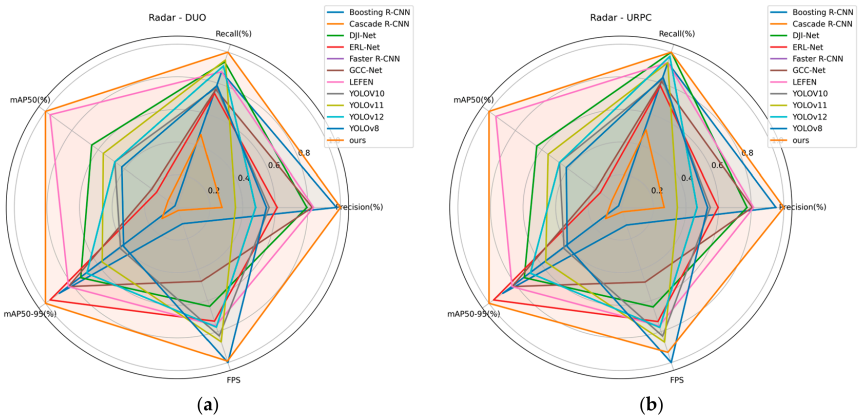
Table 6. Comparison results of different models on the URPC dataset.

Method	Type	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	FPS
Faster R-CNN	Generic	72.1	71.2	78.5	42.2	11.3
Cascade R-CNN		75.4	74.9	78.9	43.2	11.6
YOLOv8		78.7	77.4	80.9	45.7	**21.9**
YOLOv10		78.9	77.3	81.2	45.9	20.1
YOLOv11		76.4	78.1	81.7	47.1	20.5
YOLOv12		77.9	78.4	81.2	48.1	19.5
LEFEN	Underwater	82.2	78.1	84.0	49.3	19.4
DJI-Net		81.7	78.6	82.2	48.5	18.1
ERL-Net		79.5	77.0	79.4	50.5	19.1
Boosting R-CNN		83.9	78.1	78.6	49.9	12.5
GCC-Net		82.1	77.2	79.6	49.3	16.4
ours		**84.5**	**78.6**	**84.3**	**50.8**	21.2

Bold indicates optimal performance.

In comparison to specialized models such as GCC-Net (mAP50 79.6%) and DJI-Net (mAP50 82.2%), our architecture has a clear performance benefit of +4.7% and +2.1% respectively. In comparison with even the very competitive LEFEN (mAP50 84.0%), our model achieves a significant advantage in terms of mAP50-95 (+1.5%). This measure is especially vital when performing robotic-grasping problems. The repeated dominance of our DUO and URPC datasets guarantees that our "E2E-AUD" is not tailored to some particular water state but has high resilience to different marine conditions.

In order to visually compare the differences between our model and other models on different indicators, we have plotted radar charts on the test data from two datasets, as shown in Figure 5.



(**a**) (**b**)

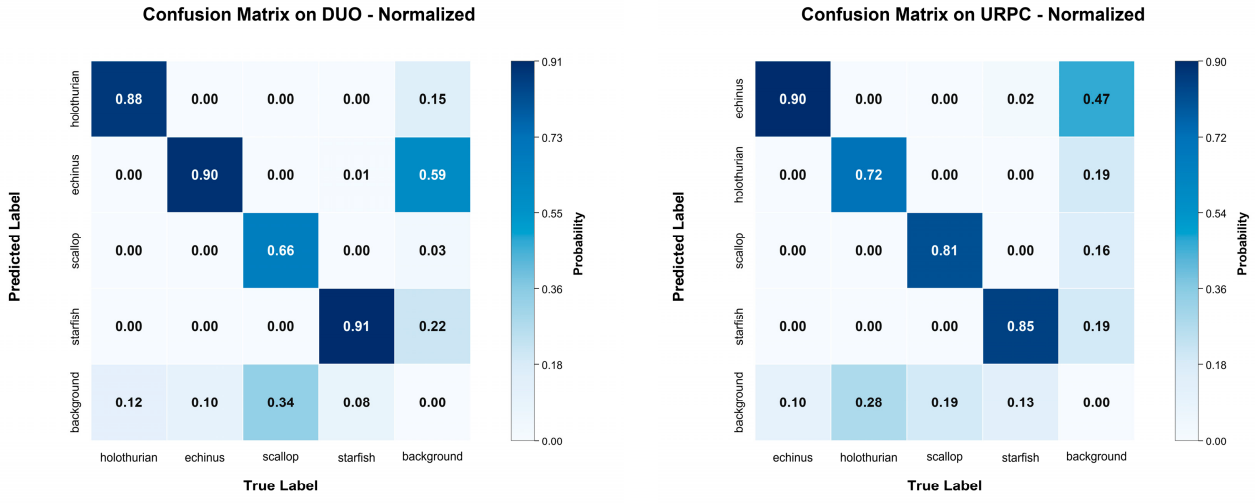
Figure 5. Radar plots comparing the performance of different models on the two datasets: (**a**) DUO and (**b**) URPC.

4.5.3. Discussion on Marine Physical and Biological Interpretability

The comparative results presented above demonstrate that E2E-AUD consistently outperforms both generic state-of-the-art detectors and previous underwater-specific models (e.g., LEFEN, GCC-Net). This superiority is not merely a result of increased parameter capacity, but is fundamentally rooted in the strong physical and biological meanings embedded within our architecture. From the perspective of marine optical physics, traditional purely data-driven models often struggle with domain shifts caused by varying water types. By explicitly integrating the Jaffe–McGlamery degradation model into the UnitModule, our framework dynamically simulates and reverses the depth-dependent transmission attenuation ($t(x)$) and atmospheric backscattering (A) at the feature level. This physical prior allows the network to maintain high discriminability in highly turbid conditions (as seen in URPC), where models lacking explicit optical physics constraints suffer severe performance drops. From the perspective of marine biological morphology, benthic organisms of high economic value (such as holothurians and starfish) are predominantly soft-bodied invertebrates characterized by extreme non-rigid deformations. Previous architectures rely on fixed geometric convolution kernels, which inherently contradict the polymorphic nature of marine life. The LDConv module addresses this biological reality by adaptively simulating the irregular contours of these organisms through dynamically learned sampling offsets. Consequently, E2E-AUD not only achieves higher quantitative localization accuracy (mAP50-95) but also offers a highly interpretable solution tailored specifically for the complex realities of marine robotic perception.

4.6. Confusion Matrix

In order to visualize the capabilities of the model in its classification performance and determine particular failure modes, we drew the confusion matrices in their normalized form on the DUO and URPC datasets, as it is depicted in Figure 6. The columns in such matrices are the Ground Truth labels, whereas the rows are the Predicted labels. Further, the diagonal elements denote the Recall of any given category, as the lowest row signifies the False Negative rate and the final column depicts the False Positive distribution.



(**a**) (**b**)

Figure 6. Normalized confusion matrices of the proposed algorithm on the test sets. (**a**) Confusion matrix on the DUO dataset. (**b**) Confusion matrix on the URPC data.

Figure 6a shows the classification performance on the DUO dataset. The starfish and Echinus models have excellent performance, with Recall rates of 0.91 and 0.90 respectively.

Off-diagonal cell values of marine species are nearly all 0.00, suggesting that the proposed approach has acquired highly discriminative characteristics and has reduced inter-class confusion to negligible levels. It confirms the efficacy of the HWD block in maintaining different frequency-domain textures. The primary problem is the scallop group; the Recall is quite low, 0.66, and the False Negative rate is also high, 0.34. It is due to the natural behavior of scallops that tend to bury themselves in sand or blend into their surroundings and they are comparatively very small. By examining the final column, we see that 59% of the wrong positive results were wrongly categorized as Echinus. It implies that complex backgrounds on the DUO set have great morphological resemblance to sea urchins, at times misleading the detector.

Figure 6b shows the findings of the URPC dataset, which has greater turbidity. The model achieves high Recall values concerning Echinus (0.90) and starfish (0.85), and scallop (0.81). It is interesting to note that the scalability of the module is much better in URPC than in DUO; hence, it can be indicated that the UnitModule is effective in increasing the contrast level of scallop shells in turbid water, which makes them easier to identify. The smallest value of Recall is seen with respect to the Holothurian at 0.72, with a miss value of 0.28. There are no stiff geometric structures in Holothurians and their soft, pliable bodies can easily assimilate into the surroundings. Although our LDConv module makes deformation modeling better, the highly developed camouflage of sea cucumbers is still a significant problem. As with DUO, the majority of errors involve Echinus as the misclassified category (0.47). This consistent trend across datasets indicates that distinguishing sea urchins from background clutter remains a key direction for future optimization, potentially requiring more aggressive hard-negative mining strategies.

Generally speaking, the confusion matrices demonstrate the fact that the "E2E-AUD" has a relatively high level of classification accuracy. The errors can mostly be found within the domain of Background–Foreground confusion (missing out the camouflaged targets or detecting the background clutter), not inter-class confusion. It is affirming that the feature extractor of the model is robust and semantically correct, which confirms the soundness of the design rationale of combining frequency-domain analysis with physics-based enhancement. Furthermore, the confusion matrices implicitly reveal the operational boundaries and failure cases of the E2E-AUD framework under extreme underwater scenarios. First, under extreme occlusion—such as scallops buried deeply in the sandy seabed (Figure 6a)—the lack of visible geometric contours causes the LDConv to fail to generate effective sampling offsets, leading to False Negatives. Second, under severe turbidity and camouflage—such as holothurians in the URPC dataset (Figure 6b)—the excessive suspended particles can disrupt the physical priors of the UnitModule, causing soft-bodied targets to completely blend into the background. Finally, while the model demonstrates strong robustness across DUO and URPC, testing on unseen-domain conditions (e.g., highly anomalous artificial lighting or unexplored deep-sea topographies) remains a challenge, as extreme domain shifts can bias the global atmospheric light estimation (A) in the A-head. Addressing these specific failure modes through multi-modal sensing and domain-adaptive learning will be a primary focus of our future work.

4.7. Vision Comparison and Qualitative Analysis

4.7.1. Comparison of SOTA Models

To intuitively validate the superiority of the proposed framework in complex underwater environments, we selected five representative challenging scenarios for qualitative analysis, as shown in Figure 7. The comparison includes the general-purpose detector YOLOv13, the underwater-specific model LEFEN, and the proposed E2E-AUD.

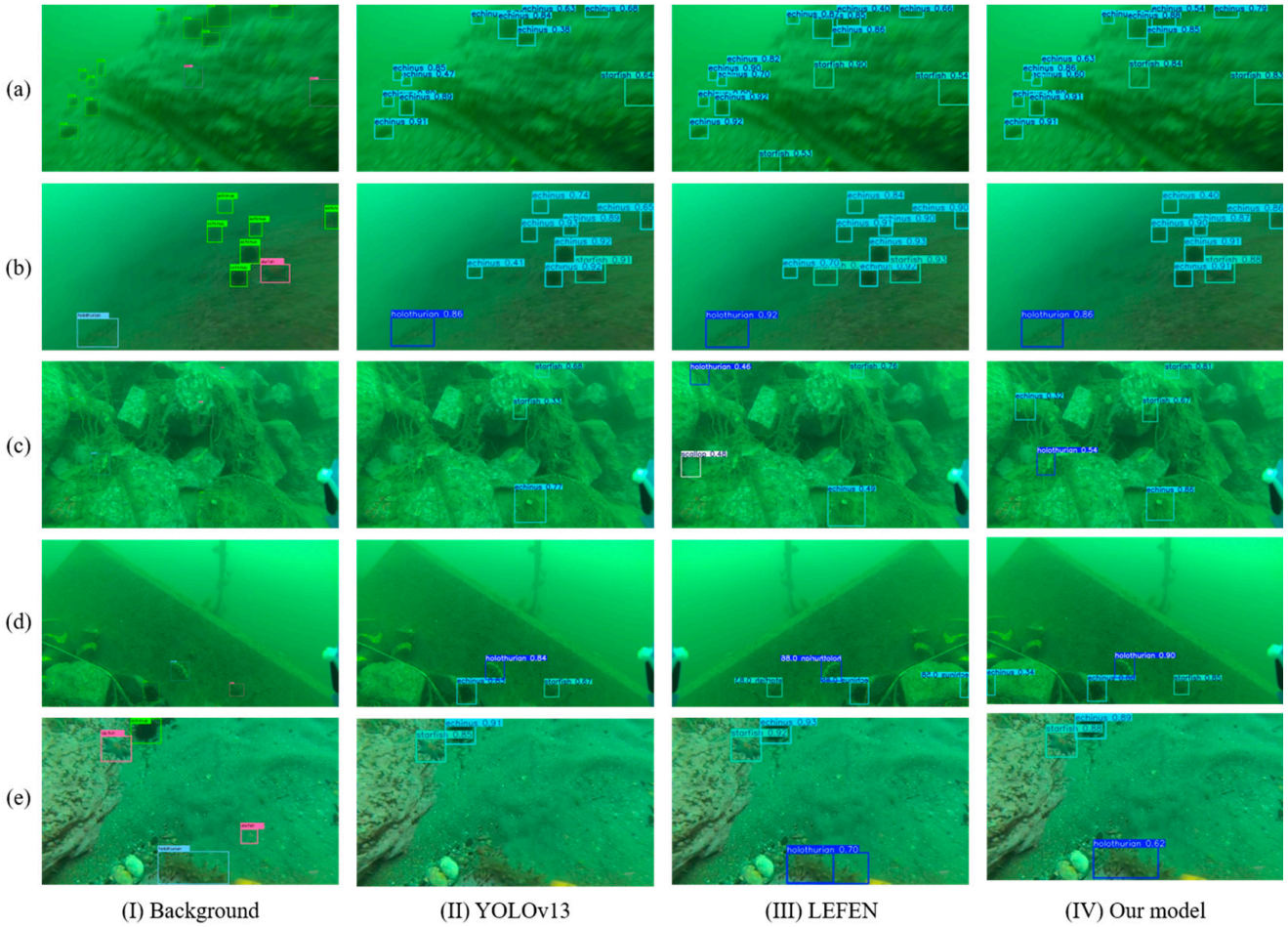


Figure 7. Qualitative comparison of detection results across five representative challenging underwater scenarios (**a**–**e**). The four columns from left to right represent the Ground Truth, YOLOv13, LEFEN, and the proposed "End-to-End Underwater Specialist", respectively. Scenarios (**a**,**b**): Turbid Environments and Small Object Detection. Scenarios (**c**,**d**): Complex Background Interference and Edge Target Capture. Scenario (**e**): Extreme Camouflage and Severe Occlusion.

When there are heavy turbidity and faint minor objects, the robustness of standard and current underwater models is obviously inadequate. Based on Scenario (a), the YOLOv13 has a relatively large number of False Negatives (missed detections) of starfish. At the same time, the LEFEN model provides a False Positive because of the interference with the background noise and it mistakes a non-existing starfish. On the contrary, our model, using the UnitModule to invert the aquatic degradation procedure adaptively, and thereby improve image contrast, attains zero misses and false alarms. Due to the blurring of the background that causes feature dissipation, Scenario (b) makes both YOLOv13 and LEFEN misidentify the background noise as an Echinus. Moreover, LEFEN identifies a nonexistent starfish falsely. We use the HWD module to provide lossless detail of texture in the frequency domain, and our model shows better discrimination ability and high detection performance.

Irrelevant background factors often cause feature confusion. Scenario (c) shows that a lot of False Negatives are generated by YOLOv13 with dense seaweed, whereas LEFEN is prone to category misclassification, although it produces the detection boxes. The proposed model does not just obtain the right classification of all the targets, but its localization accuracy is much better. It can be explained by the LDConv module, which adjusts the receptive field to the real morphology of objects automatically. Moreover, the model can also be seen to capture parts of targets that are at the edge of the field of view, as shown by

Scenario (d). As a result of poor modeling of incomplete features, YOLOv13 cannot detect the partially visible target; however, both our model and LEFEN can capture edge target, which confirms the strong feature perception of the proposed framework.

Scenario (e) simulates the most severe underwater-detection circumstances: intense target-camouflage and mutual occlusion. While all models show vulnerabilities in confrontation with starfish that are nearly indistinguishable to the environment, there is a difference between the performances of the models with regard to the camouflaged holothurian. YOLOv13 does not detect the target at all, and LEFEN incorrectly categorizes it. Incorporating the physics-aware improvement of the UnitModule and dynamic modeling of non-rigid objects under LDConv, our model is able to correctly identify the holothurian by bounding boxes, which tightly cover the target contour.

As shown by the visualization outcome, the presented model has particular benefits regarding background noise suppression, visual degradation reduction, and deformation irregularity sensing. This achievement can be explained majorly by the fact that these three elements were tightly coupled in the physical domain, frequency domain, and geometric domain, respectively. It provides strong precision of detection with a powerful performance even under very extreme conditions of the underwater environment.

4.7.2. Interpretability Analysis via Grad-CAM Visualization

To further investigate the internal decision-making process and the ability to perceive the proposed model within the complex underwater environment, we have used Gradient-weighted Class Activation Mapping (Grad-CAM) to represent the feature activation maps of the last backbone layer. As seen in Figure 8, there was comparative analysis between the original images, our proposed E2E-AUD and the baseline YOLOv13. These results indicate that there are great differences in the distribution of attention and features that are selected.

YOLOv13 has diffuse and fragmented activation patterns across all test samples (a–c). As depicted in (III) of Figure 8a,c, the high-activation areas (red regions) of YOLOv13 tend to wander out of the center of the target and spread to the neighboring background like a reef edge or turbid water. Our model, on the other hand, is extremely densely activated, with its characteristics being very compact. Geometric centroids are tightly contained within the core red zones. It gives a very strong empirical confirmation that LDConv module, through learning dynamic offsets, allows the field of reception to dynamically adjust itself to the irregular outlines of marine creatures, which leads to considerable improvement in the precision of the feature localization.

Suspended particles and fluctuating light conditions in water can provoke inaccurate responses in the presence of features. Under Scenarios (b) and (c), YOLOv13 generates significant pseudo-activations (distributed extensively as bluish and greenish noise) on background reef areas and turbid waters. However, our model has a very low level of activation (deep blue), indicating its higher ability to suppress noise. This is mainly due to physics-based reversal of degradation indicators by the UnitModule and filtering of high-frequency noise through the HWD, which enables the network to retrieve highly discriminative target features under environmental disturbance.

In the situation where the targets are weakly textured due to the irregular illumination, such as Scenario (c), our model has a strong activity crest (dark red) in the target area, which is not as present in YOLOv13. Such indication indicates that by the lossless image detail encoding made available to HWD, the proposed framework does well to maintain the fundamental image texture information when downsampling and thus enhances the semantic perception of small and blurry objects in the network.

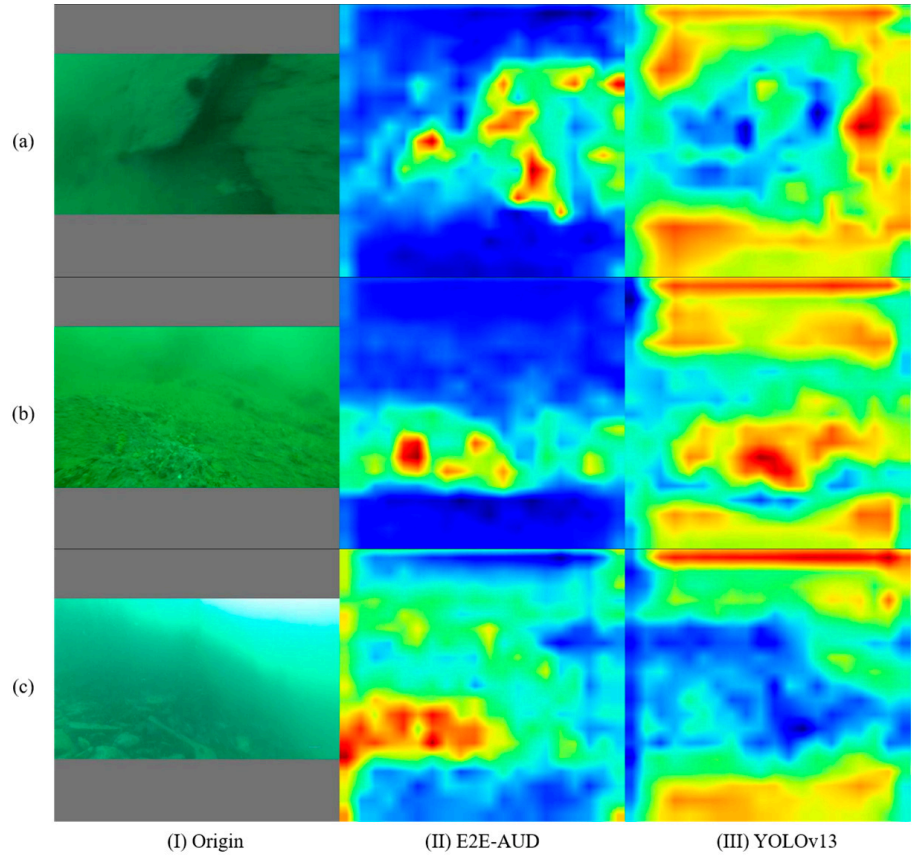


Figure 8. Grad-CAM visualization of the feature activation maps for the final backbone layer across different underwater scenes (**a**–**c**). (**I**). Original input images with varying levels of turbidity and lighting interference. (**II**). Heatmaps generated by the proposed E2E-AUD. (**III**). Heatmaps generated by the baseline YOLOv13.

The Grad-CAM analysis shows that our model has high interpretability and robustness. With tight correlation between the physical, frequency and geometric domains by means of the UnitModule, HWD and LDConv, the framework can successfully focus on target-related features and filter out non-trivial underwater clutter, guaranteeing consistent behavior in tough conditions of marine environments.

4.8. Computational Complexity and Embedded-Device Feasibility

A crucial requirement for deploying object detection models on Autonomous Underwater Vehicles (AUVs) is strictly bounded computational complexity. AUVs operate on limited battery power and rely on low-power embedded computing boards rather than high-end desktop GPUs. To comprehensively assess the deployment feasibility of E2E-AUD, we evaluate its complexity using hardware-independent metrics: Floating Point Operations (GFLOPs), parameter count (Params), and static Model Size.

As detailed in Table 7, the baseline YOLOv13n requires 8.1 GFLOPs and 2.6 M parameters. The integration of the physics-aware UnitModule and the frequency-domain HWD module introduces a marginal computational overhead. However, owing to the linear parameter scaling design of the LDConv module ($P \propto N$), the total complexity of the proposed E2E-AUD framework is highly constrained, demanding only 9.4 GFLOPs and 2.8 M parameters, with a total storage footprint of approximately 6.2 MB.

Mainstream AUV perception systems are typically equipped with edge AI accelerators, such as the NVIDIA Jetson Xavier NX or Jetson Orin Nano series. These embedded platforms provide compute capacities ranging from 21 to 40 TOPS (Tera Operations Per

Second) and possess 8 GB to 16 GB of memory. The 6.2 MB memory requirement of E2E-AUD is completely negligible for such devices. Furthermore, a computational load of 9.4 GFLOPs is well within the real-time processing threshold of these edge platforms. When further optimized using deployment frameworks like NVIDIA TensorRT or ONNX Runtime with FP16 precision, the E2E-AUD is theoretically capable of running at over 30 FPS on an edge device with a power consumption of under 15 W. This confirms that the proposed method not only achieves state-of-the-art accuracy in complex marine environments but also satisfies the strict size, weight, and power (SWaP) constraints required for real-world AUV operations.

Table 7. Comparison of Complexity of Different Models.

Method	Type	GFLOPs	Params (M)
Faster R-CNN	Generic	215.3	41.5
Cascade R-CNN		235.0	69.0
YOLOv8		8.7	3.2
YOLOv10		8.0	2.7
YOLOv11		**6.5**	**2.6**
YOLOv12		8.5	2.9
YOLOv13		8.1	**2.6**
LEFEN	Underwater	12.5	4.1
DJI-Net		14.2	4.5
ERL-Net		13.0	4.0
Boosting R-CNN		180.4	50.5
GCC-Net		45.0	15.2
ours		9.4	2.8

Bold indicates optimal performance.

5. Conclusions

In this paper, we proposed the E2E-AUD, a unified framework tailored for robust underwater object detection. We solved the inner contradictions of underwater visual deterioration and discriminative trait extraction through the deep integration of physics-aware enhancement, frequency-domain preservation of fine details, and geometric adaptation. The specific new results and primary contributions are summarized as follows:

- Physics-Aware Joint Enhancement: We introduced the UnitModule, which aligns the distribution of features in degraded inputs through concurrent optimization, effectively eliminating False Negatives caused by severe color casts and haze.
- Frequency-Domain Detail Preservation: By replacing traditional strided convolutions with the HWD module, the framework successfully alleviates the information loss of small targets (e.g., heavily occluded marine organisms) during downsampling operations.
- Geometric Adaptive Feature Extraction: The LDConv module was designed to adaptively capture the diverse poses and non-rigid deformations of benthic marine organisms without the quadratic parameter explosion of traditional dynamic convolutions.
- New State-of-the-Art Results: Extensive quantitative studies demonstrate that the full E2E-AUD model achieves remarkable peak performances, recording an mAP50 of 86.2% on the DUO dataset and 84.3% on the highly turbid URPC dataset, while maintaining a real-time inference speed of 21.8 FPS.

Compared with the recently published state-of-the-art literature in 2025, our E2E-AUD framework demonstrates distinct superiority. While the latest general-purpose detector YOLOv12 achieves high efficiency, our model surpasses it by 3.0% in mAP50 on the DUO dataset due to our targeted underwater optical degradation reversal. Furthermore, when compared against the recently proposed underwater-specific LEFEN model, E2E-AUD exhibits a more robust geometric modeling capability, surpassing it in strict localization metrics (mAP50-95) across both datasets without sacrificing real-time deployment feasibility (9.4 GFLOPs).

Despite the promising results, this study has several limitations that warrant future investigation. First, regarding computational cost, while E2E-AUD achieves real-time inference suitable for edge devices, the integration of UnitModule and HWD inevitably introduces a processing overhead compared to vanilla lightweight generic detectors. Second, the framework faces potential overfitting and domain generalization challenges. Since the model was primarily optimized and evaluated on specific coastal datasets (DUO and URPC), its performance may experience degradation when deployed directly in vastly different marine ecosystems with unseen background distributions. Third, the physics-aware enhancement relies on simplified underwater lighting assumptions (e.g., uniform ambient natural illumination), which may not hold in deep-sea environments or nighttime scenarios involving active, non-uniform artificial lighting from AUVs. Finally, regarding enhancement model constraints, the unsupervised task-oriented loss prioritizes detection features over pixel-perfect restoration, meaning the generated latent representations might occasionally exhibit visual artifacts or noise amplification under extreme turbidity.

Consequently, our future research will be dedicated to addressing these constraints. We plan to develop domain-adaptive learning strategies to enhance cross-domain generalization, incorporate dynamic illumination priors for non-uniform active lighting conditions, and create improved hard-negative mining methods to better differentiate highly camouflaged objects in complicated seabed environments.

Author Contributions: Conceptualization, W.Z., J.Z. and S.L.; methodology, W.Z. and J.Z.; software, W.Z.; validation, W.Z.; formal analysis, S.L.; investigation, Y.Z.; resources, J.Z.; data curation, W.Z.; writing—original draft preparation, W.Z.; writing—review and editing, J.Z.; visualization, W.Z.; supervision, J.Z.; project administration, J.Z.; funding acquisition, J.Z. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China [Grant Number 2021YFC2801100] and National Natural Science Foundation of China (NSFC) [Grant Number 42176194].

Data Availability Statement: The code is available at https://github.com/EvansChou/E2E-AUD/tree/master (accessed on 3 June 2026).

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

Nomenclature

The abbreviations and mathematical symbols used in this manuscript are summarized below. Where the same symbol appears in more than one context, the applicable section or equation is specified.

Abbreviations

Abbreviation	Definition
E2E-AUD	End-to-end adaptive underwater detection framework
AUV	Autonomous underwater vehicle
CNN	Convolutional neural network
GAN	Generative adversarial network
UnitModule	Lightweight joint image-enhancement module embedded before the detector

UnitBackbone	Feature-extraction backbone within the UnitModule
T-Head	Transmission-map prediction head within the UnitModule
A-Head	Background-light prediction head within the UnitModule
GAP	Global average pooling
FC	Fully connected layer
CCP	Color cast predictor
UCRT	Pseudo-label generation strategy used for assisting color-cast supervision
HWD	Haar wavelet downsampling
MRA	Multiresolution analysis
BN	Batch normalization
SiLU	Sigmoid linear unit activation function
LDConv	Linear deformable convolution
DCN	Deformable convolutional network
P-R	Precision–Recall
TP	True positive
FP	False positive
FN	False negative
AP	Average precision
mAP	mean average precision
mAP50	mean average precision at an intersection-over-union threshold of 0.50
mAP50-95	mean average precision averaged over intersection-over-union thresholds from 0.50 to 0.95
FPS	Frames per second
Params	Number of trainable model parameters
DUO	Detection in underwater objects dataset
URPC	Underwater robot perception challenge dataset

Mathematical Symbols

Symbol	Definition	Location
D	Degraded underwater-image domain	Problem formulation
X	Input feature map or degraded underwater image; the specific role is defined locally	Problem formulation; HWD; LDConv
O	Target objects contained in a degraded underwater image	Problem formulation
E(·)	Image-enhancement mapping	Problem formulation
D(·)	Object-detection mapping; distinguish from the degraded domain D by context	Problem formulation
x	Pixel or spatial coordinate	Equation (1)
I(x)	Observed underwater image intensity at coordinate x	Equation (1)
J(x)	Latent clear-scene radiance or enhanced image representation at coordinate x	Equations (1)–(3)
t(x)	Transmission-related map used in the UnitModule	Equations (1)–(4)
A	Background-light term used in the underwater imaging model	Equations (1)–(3)
L_total	Total training objective	Equation (4)
L_det	Detector localization and classification loss	Equation (4)
L_enh	Unsupervised enhancement-related loss	Equations (4) and (5)
λ	Balancing coefficient for the enhancement-related loss	Equation (4)
w_1, w_2, w_3	Non-negative balancing coefficients for L_cc, L_tv, and L_ac; current manuscript values: 0.1, 0.01, and 0.1	Equation (5)
L_cc	Color-consistency loss	Equations (5) and (6)

L_tv	Texture-smoothness loss based on total variation	Equations (5) and (7)
L_ac	Assisting color-cast loss	Equations (5) and (8)
μ_Jˆr, μ_Jˆg, μ_Jˆb	Mean intensities of the red, green, and blue channels of the enhanced image J	Equation (6)
∇_h, ∇_v	Horizontal and vertical gradient operators	Equation (7)
y_pseudoˆ(k)	Pseudo-label probability for the k-th color-cast category	Equation (8)
p_CCPˆ(k)	Predicted probability for the k-th color-cast category from the color cast predictor	Equation (8)
K	Total number of predefined color-cast categories	Equation (8)
k	Index of a color-cast category	Equation (8)
H_0, H_1	Low-pass and high-pass Haar filtering operators	Figure 3
z_1, z_2	Two spatial dimensions of the feature map	Figure 3
ϕ(x), ψ(x)	One-dimensional Haar scaling function and wavelet function	Equation (9a)
ϕ_j,k(x)	Scaled and shifted basis function at scale j and translation k	Equation (9b)
A, H, V, D	Approximation, horizontal-detail, vertical-detail, and diagonal-detail Haar sub-bands	Equation (10)
X_encoded	Channel-wise concatenation of the four Haar sub-bands	Equations (10) and (11)
Y_out	Output feature map after 1×1 convolution, batch normalization, and SiLU activation	Equation (11)
σ(·)	SiLU activation function	Equation (11)
p_0	Reference spatial location for LDConv sampling	Equation (12)
y(p_0)	LDConv output at reference location p_0	Equation (12)
N	Total number of LDConv sampling points	Equation (12); Section 4.4.1
n	Index of an LDConv sampling point	Equation (12)
w_n	Weight associated with the n-th LDConv sampling point	Equation (12)
p_n	Predefined initial coordinate offset of the n-th LDConv sampling point	Equation (12)
Δp_n	Learnable offset of the n-th LDConv sampling point	Equation (12)
P_n	Initial LDConv sampling-coordinate set	Section 3.4
P	LDConv parameter count; the manuscript states a linear relation $P \propto N$	Section 3.4
Precision	TP/(TP + FP)	Equation (13)
Recall	TP/(TP + FN)	Equation (14)
P_i	Precision measured at the i-th threshold operating point on the P-R curve	Equation (15)
R_i	Recall measured at the i-th threshold operating point on the P-R curve	Equation (15)
ΔR_i	Recall increment between consecutive threshold operating points	Equation (15)
i	Index of a threshold operating point or object category, as defined locally	Equations (15) and (16)
C	Total number of object categories	Equation (16)

Bold represents the header of the table.

Appendix A

Algorithm A1: Initial Sampling Coordinate Generation for LDConv

Input : Number of sampling points N, Random Seed S
Output : Initial sampling coordinate set $P_{init} = \{p_1, p_2, \ldots, p_n\}$.
Initialize $P_{init} \leftarrow \emptyset$
Set random seed S to ensure reproducibility
Calculate the base grid scale $G \leftarrow \left\lfloor \sqrt{N} \right\rfloor$
Calculate the spatial center $C \leftarrow \frac{G-1}{2}$
/ Step 1: Generate regular grid coordinates centered at (0, 0) */*
for $i \leftarrow 0$ **to** $G - 1$ **do**
  for $j \leftarrow 0$ **to** $G - 1$ **do**
    if $|P_{init}| < N$ **then**
      $x \leftarrow i - C$
      $y \leftarrow j - C$
      $P_{init} \leftarrow P_{init} \cup \{(i, j)\}$
    end if
  end for
end for
/ Step 2: Supplement with irregular coordinates */*
while $|P_{init}| < N$ **do**
  Generate random offsets $\Delta x, \Delta y \sim \text{Uniform}(0, G)$
  $P_{init} \leftarrow P_{init} \cup \{(\Delta x, \Delta y)\}$
end while
return P_{init}

Bold and italics represents the key logical steps in pseudocode.

References

1. Nabahirwa, E.; Song, W.; Zhang, M.; Fang, Y.; Ni, Z. A Structured Review of Underwater Object Detection Challenges and Solutions: From Traditional to Large Vision Language Models. *arXiv* **2025**, arXiv:2509.08490. [CrossRef]
2. Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; Liotta, A. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* **2023**, *527*, 204–232. [CrossRef]
3. Wang, W.; Sun, Y.F.; Gao, W.; Xu, W.; Zhang, Y.; Huang, D. Quantitative detection algorithm for deep-sea megabenthic organisms based on improved YOLOv5. *Front. Mar. Sci.* **2024**, *11*, 1301024. [CrossRef]
4. Bakht, A.B.; Jia, Z.; Din, M.U.; Akram, W.; Saoud, L.S.; Seneviratne, L.; Lin, D.; He, S.; Hussain, I. MuLA-GAN: Multi-Level Attention GAN for Enhanced Underwater Visibility. *Ecol. Inform.* **2024**, *81*, 102631. [CrossRef]
5. Cao, D.; Chen, Z.; Gao, L. An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks. *Hum.-Centric Comput. Inf. Sci.* **2020**, *10*, 14. [CrossRef]
6. 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]
7. Mirzaei, B.; Nezamabadi-pour, H.; Raoof, A.; Derakhshani, R. Small Object Detection and Tracking: A Comprehensive Review. *Sensors* **2023**, *23*, 6887. [CrossRef]
8. Zayed, M.M.; Shokair, M.; Rosas, A.A.; Ghallab, R. Modeling and Analysis of Underwater Optical Wireless Communication Channels. *Int. J. Eng. Appl. Sci. Oct. 6 Univ.* **2025**, *2*, 32–45. [CrossRef]
9. Zayed, M.M.; Shokair, M. A Comprehensive Review of End-to-End Channel Modeling, Impairments, and Performance Analysis in Underwater Optical Wireless Communications. *Int. J. Eng. Appl. Sci. Oct. 6 Univ.* **2026**, *3*, 1–15. [CrossRef]
10. Zayed, M.M.; Shokair, M. Modeling and simulation of optical wireless communication channels in IoUT considering water types, turbulence and transmitter selection. *Sci. Rep.* **2025**, *15*, 28381. [CrossRef]
11. Zayed, M.M.; Shokair, M. Performance analysis and optimization of modulation techniques for underwater optical wireless communication in varied aquatic environments. *Sci. Rep.* **2025**, *15*, 32570. [CrossRef] [PubMed]
12. Jaffe, J.S. Computer modeling and the design of optimal underwater imaging systems. *IEEE J. Ocean. Eng.* **1990**, *15*, 101–111. [CrossRef]

13. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images. *IEEE Robot. Autom. Lett.* **2018**, *3*, 387–394. [CrossRef]
14. Li, C.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [CrossRef]
15. Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
16. Zhang, W.; Wang, H.; Ren, P.; Zhang, W. Underwater Image Color Correction via Color Channel Transfer. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
17. Awad, A.; Saleem, A.; Paheding, S.; Lucas, E.; Al-Ratrout, S.; Havens, T.C. Beneath the Surface: The Role of Underwater Image Enhancement in Object Detection. *arXiv* **2024**, arXiv:2411.14626.
18. Wang, Z.; Li, Y.; Chen, X.; Lim, S.-N.; Torralba, A.; Zhao, H.; Wang, S. Detecting Everything in the Open World: Towards Universal Object Detection. *arXiv* **2023**, arXiv:2303.11749. [CrossRef]
19. Feng, J.; Jin, T. LMFEN: Lightweight multi-scale feature enhancement network for underwater object detection in AUVs. *Ocean Eng.* **2025**, *339*, 122109. [CrossRef]
20. Shang, H.; Yu, Y.; Song, W. Underwater fish image classification algorithm based on improved ResNet-RS model. In Proceedings of the Jiangsu Annual Conference on Automation (JACA 2023), Nanjing, China, 10–12 November 2023; pp. 106–110.
21. Zhang, X.; Song, Y.; Song, T.; Yang, D.; Ye, Y.; Zhou, J.; Zhang, L. LDConv: Linear deformable convolution for improving convolutional neural networks. *Image Vis. Comput.* **2024**, *149*, 105190. [CrossRef]
22. Zhang, Y.; Liu, T.; Zhen, J.; Kang, Y.; Cheng, Y. Adaptive Downsampling and Scale Enhanced Detection Head for Tiny Object Detection in Remote Sensing Image. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. [CrossRef]
23. Du, D.; Si, L.; Xu, F.; Niu, J.; Sun, F. A Physical Model-Guided Framework for Underwater Image Enhancement and Depth Estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 1–11. [CrossRef]
24. Zhuang, P.; Li, C.; Wu, J. Bayesian retinex underwater image enhancement. *Eng. Appl. Artif. Intell.* **2021**, *101*, 104171. [CrossRef]
25. Yang, M.; Sowmya, A.; Wei, Z.; Zheng, B. Offshore Underwater Image Restoration Using Reflection-Decomposition-Based Transmission Map Estimation. *IEEE J. Ocean. Eng.* **2020**, *45*, 521–533. [CrossRef]
26. Cheng, Z.; Fan, G.; Zhou, J.; Gan, M.; Chen, C.L.P. FDCE-Net: Underwater Image Enhancement With Embedding Frequency and Dual Color Encoder. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 1728–1744. [CrossRef]
27. Liu, C.; Zhang, Y.; Xue, Y.; Qian, X. AJENet: Adaptive Joints Enhancement Network for Abnormal Behavior Detection in Office Scenario. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 1427–1440. [CrossRef]
28. Rafaely, B.; Weinzierl, S.; Berebi, O.; Brinkmann, F. Loss functions incorporating auditory spatial perception in deep learning: A review. *arXiv* **2025**, arXiv:2506.19404. [CrossRef]
29. Xu, W.; Chen, X.; Guo, H.; Huang, X.; Liu, W. Unsupervised Image Restoration With Quality-Task-Perception Loss. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5736–5747. [CrossRef]
30. Sapkota, R.; Flores-Calero, M.; Qureshi, R.; Badgujar, C.; Nepal, U.; Poulose, A.; Zeno, P.; Vaddevolu, U.B.P.; Khan, S.; Shoman, M.; et al. YOLO advances to its genesis: A decadal and comprehensive review of the You Only Look Once (YOLO) series. *Artif. Intell. Rev.* **2025**, *58*, 274. [CrossRef]
31. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
32. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 3–7 May 2021.
33. Yang, G.; Yang, J.; Fan, W.; Yang, D. Neural Network for Underwater Fish Image Segmentation Using an Enhanced Feature Pyramid Convolutional Architecture. *J. Mar. Sci. Eng.* **2025**, *13*, 238. [CrossRef]
34. Peng, F.; Miao, Z.; Li, F.; Li, Z. S-FPN: A shortcut feature pyramid network for sea cucumber detection in underwater images. *Expert Syst. Appl.* **2021**, *182*, 115306. [CrossRef]
35. Liu, K.; Peng, L.; Tang, S. Underwater Object Detection Using TC-YOLO with Attention Mechanisms. *Sensors* **2023**, *23*, 2567. [CrossRef]
36. Wang, Z.; Shen, L.; Yu, M.; Lin, Y.; Zhu, Q. Single Underwater Image Enhancement Using an Analysis-Synthesis Network. *arXiv* **2021**, arXiv:2108.09023. [CrossRef]
37. Xia, W.; Zhao, P.; Wen, Y.; Xie, H. A Survey on Data Center Networking (DCN): Infrastructure and Operations. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 640–656. [CrossRef]
38. Xiong, Y.; Li, Z.; Chen, Y.; Wang, F.; Zhu, X.; Luo, J.; Wang, W.; Lu, T.; Li, H.; Qiao, Y.; et al. Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications. *arXiv* **2024**, arXiv:2401.06197. [CrossRef]
39. Deeba, F.; Kun, S.; Dharejo, F.A.; Zhou, Y. Wavelet-Based Enhanced Medical Image Super Resolution. *IEEE Access* **2020**, *8*, 37035–37044. [CrossRef]
40. Li, Z.; Kuang, Z.-S.; Zhu, Z.-L.; Wang, H.-P.; Shao, X.-L. Wavelet-Based Texture Reformation Network for Image Super-Resolution. *IEEE Trans. Image Process.* **2022**, *31*, 2647–2660. [CrossRef]

41. Strickland, R.N.; Hahn, H.I. Wavelet transform methods for object detection and recovery. *IEEE Trans. Image Process.* **1997**, *6*, 724–735. [CrossRef]
42. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
43. Liu, Z.; Wang, B.; Li, Y.; He, J.; Li, Y. UnitModule: A lightweight joint image enhancement module for underwater object detection. *Pattern Recognit.* **2024**, *151*, 110435. [CrossRef]
44. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819. [CrossRef]
45. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling Up Your Kernels to 31 × 31: Revisiting Large Kernel Design in CNNs. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11953–11965.
46. Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; Wang, Z. A Dataset and Benchmark of Underwater Object Detection for Robot Picking. In Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; pp. 1–6. [CrossRef]
47. Liu, H.; Song, P.; Ding, R. WQT and DG-YOLO: Towards domain generalization in underwater object detection. *arXiv* **2020**, arXiv:2004.06333. [CrossRef]
48. 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.
49. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. [CrossRef]
50. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8, Version 8.0.0. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 1 June 2026).
51. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458.
52. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
53. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
54. Wang, B.; Wang, Z.; Guo, W.; Wang, Y. A dual-branch joint learning network for underwater object detection. *Knowl.-Based Syst.* **2024**, *293*, 111672. [CrossRef]
55. Wang, G.; Sun, C.; Sowmya, A. ERL-Net: Entangled Representation Learning for Single Image De-Raining. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5643–5651.
56. Song, P.; Li, P.; Dai, L.; Wang, T.; Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing* **2023**, *530*, 150–164. [CrossRef]
57. Dai, L.; Liu, H.; Song, P.; Liu, M. A gated cross-domain collaborative network for underwater object detection. *Pattern Recognit.* **2024**, *149*, 110222. [CrossRef]