electronics

Article

Multi-Dimensional Collaborative Enhancement Method for Small-Object Detection in Complex Underwater Scenarios

Sisi Zhu [1,2], Xinyu Li [2], Yu Zhang [3,*], Yuanyuan Liu [3], Xiaotian Wang [4] and Xinnan Fan [4]

[1] College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213000, China; zhu_sisi@ctg.com.cn
[2] China Yangtze Power Co. Ltd., Yichang 443000, China; li_xinyu2@ctg.com.cn
[3] College of Artificial Intelligence and Automation, Hohai University, Changzhou 213000, China; lyy7535@hhu.edu.cn
[4] College of Information Science and Engineering, Hohai University, Changzhou 213000, China; wxthhu@hhu.edu.cn (X.W.); fanxn@hhu.edu.cn (X.F.)
* Correspondence: 241314010011@hhu.edu.cn

Abstract

Underwater object detection is challenged by scale variation, light attenuation, scattering, and non-uniform illumination, which weaken texture and boundary cues and increase background interference, particularly for small and distant objects. To address these issues, this study proposes a multi-dimensional collaborative enhancement framework. A Feature Enhancement Module (FEM) integrates local and contextual information through heterogeneous convolutional branches and residual learning. A Spatial-Attention-Based Spatial Pyramid Pooling (SPP_SA) module combines multi-scale pooling with spatial attention to suppress irrelevant background responses. A Soft Nearest-Neighbor Interpolation (SNI) module introduces response-scaled nearest-neighbor upsampling to regulate feature magnitude during resolution recovery and support cross-level fusion. Experiments on UTDAC2020, Brackish, and RUOD demonstrate consistent improvements under different underwater conditions. Compared with the YOLOv11 baseline, the proposed method improves $mAP_{50:95}$ by 5.56%, 3.37%, and 0.81% on UTDAC2020, Brackish, and RUOD, respectively. These results support the effectiveness of the proposed framework for mixed-scale underwater object detection, including small-scale objects.

Keywords: underwater object detection; feature enhancement; spatial attention; feature upsampling; multi-dimensional collaborative enhancement

Check for updates

Academic Editor: Manohar Das

Received: 30 June 2026
Revised: 10 August 2026
Accepted: 12 August 2026
Published: 14 August 2026

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

1. Introduction

Underwater object detection is a core component of intelligent marine perception. However, wavelength-dependent light attenuation, backscattering caused by suspended particles, and non-uniform illumination can substantially degrade underwater imaging quality, resulting in color distortion, weak texture and boundary responses, reduced contrast, and complex background interference [1,2]. These degradations increase missed detections and localization errors in practical underwater scenes. In addition, object scale varies naturally with imaging distance and viewpoint. Distant aquatic organisms often occupy only a small spatial region and contain limited discriminative information, making small-object detection particularly vulnerable to underwater degradation. Therefore, this study considers small-object detection as an important challenge within mixed-scale

underwater object detection rather than assuming that all objects in underwater scenes are small.

Although Convolutional Neural Networks (CNNs) and Transformer-based detectors have achieved substantial progress in object detection, directly applying them to degraded underwater scenes still faces difficulties in feature preservation, background discrimination, and object localization. Existing studies have introduced multi-scale feature pyramids and self-attention mechanisms to enhance detail representation and contextual perception [3], multi-path and attention-based strategies to improve feature selection under background interference [2,4], soft nearest-neighbor interpolation to alleviate feature-fusion and spatial-alignment problems [5], and feature-enhancement or context-aware modules to strengthen small-object representation [6]. However, these strategies generally emphasize individual stages of the detection pipeline rather than explicitly coordinating feature enhancement, background-interference suppression, and resolution recovery.

Accordingly, three interrelated problems remain in complex underwater object detection. First, low contrast and scattering degradation weaken object texture and boundary responses; small and distant objects are particularly susceptible because their limited spatial extent makes discriminative information more likely to be attenuated during successive downsampling. Second, suspended particles and non-uniform illumination produce spatially irregular background responses that may be retained or propagated during multi-scale contextual aggregation, while excessive suppression may also weaken useful object structures. Third, spatial information is progressively lost during downsampling, whereas conventional upsampling may not simultaneously preserve boundary information and cross-level spatial consistency during feature fusion. These problems are coupled: strengthening weak object responses can also amplify background interference, contextual filtering may suppress useful boundaries, and inappropriate resolution recovery may further degrade localization.

To address these coupled challenges, this paper proposes a multi-dimensional collaborative enhancement framework for underwater object detection. The framework coordinates feature enhancement, contextual filtering, and response-regulated resolution recovery, with particular attention to small and distant objects whose feature responses are more vulnerable to degradation. The main contributions are summarized as follows:

(1) A Feature Enhancement Module (FEM) is designed using three heterogeneous convolutional branches composed of conventional, asymmetric, and dilated convolutions to capture complementary local details and broader contextual information. Learnable multi-branch fusion and residual information preservation improve feature representation for underwater objects at different scales and are particularly beneficial to small and distant objects whose weak texture and boundary information is more susceptible to loss during deep feature extraction.

(2) To reduce the influence of scattering-induced background interference during multi-scale contextual aggregation, a spatial-attention-based spatial pyramid pooling (SPP_SA) module is constructed by coupling multi-scale pooling with dual-channel spatial attention. Spatial weights generated from channel-wise average- and max-pooling statistics recalibrate the 5×5, 9×9, and 13×13 pooling branches, suppressing irrelevant background responses while preserving spatially consistent object structures. This process also helps prevent weak responses from small objects from being overwhelmed by complex underwater backgrounds.

(3) A Soft Nearest-Neighbor Interpolation (SNI) module is introduced into the feature-fusion network. A response-scaling coefficient is applied after nearest-neighbor interpolation to regulate aggregate feature-response magnitude during resolution recovery. In collaboration with FEM and SPP_SA, SNI supports stable cross-level feature fusion and preserves

boundary and texture information without introducing learnable upsampling parameters, thereby improving localization for objects that are sensitive to spatial-information loss.

The remainder of this paper is organized as follows. Section 2 reviews related studies from three perspectives: underwater object detection, attention mechanisms, and feature upsampling and nearest-neighbor interpolation, and summarizes the strengths and limitations of representative methods through comparative analysis. Section 3 presents the proposed multi-dimensional collaborative enhancement framework and describes the structures and computational processes of FEM, SPP_SA, and SNI. Section 4 introduces the datasets, training settings, and evaluation metrics, and presents comparative, complexity, and ablation experiments. Section 5 discusses the main findings, practical implications, and scientific limitations of this study. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1. Underwater Object Detection

In general object detection, one-stage detectors represented by SSD [7], FSAF [8], and YOLOv7 [9] have been widely used in real-time tasks because of their efficient inference speed. However, their feature representations may degrade when they are directly applied to low-contrast underwater scenes with strong scattering interference. Two-stage methods such as Faster R-CNN [10], Cascade RPN [11], and Cascade R-CNN [12] improve localization accuracy through region proposal networks or cascade refinement. RetinaNet [13] alleviates foreground–background imbalance in dense detection through focal loss, but its performance in degraded underwater scenes still depends on the quality of multi-scale feature representations. PAFPN [14], Libra R-CNN [15], and DetectoRS [16] further strengthen the multi-scale representation capability of feature pyramids and improve detection performance for objects at different scales. In recent years, Transformer-based architectures such as Deformable DETR [17], Swin Transformer [18], and DAB-DETR [19] have demonstrated strong performance in general object detection through long-range dependency modeling. For underwater or other complex scenes, methods such as ERL [4], Boosting R-CNN [20], D2Det [21], and SABL [22] improve detection accuracy through error reweighting, feature refinement, or boundary-aware regression strategies.

Compared with general scenes, underwater object detection is constrained not only by object-scale and category variations but also by the optical propagation characteristics of water. Wavelength-dependent attenuation changes the color and brightness distributions of objects, while scattering by suspended particles reduces contrast and produces irregular bright responses. Non-uniform illumination, water-current disturbances, and partial occlusion further increase intra-class variation. The UTDAC2020 dataset represents turbid nearshore conditions and contains echinus, holothurian, scallop, and starfish categories, whereas the Brackish dataset represents fixed underwater monitoring conditions involving dynamic illumination and geometric deformation. These characteristics require underwater detectors to enhance multi-scale representations while preventing background suppression from damaging useful texture and boundary information.

Dai et al. proposed edge-guided representation learning to strengthen object contours under underwater degradation [4]. Li and Shi proposed YOLO-GE, which enhances underwater object features and suppresses background interference through attention fusion [2]. Hu et al. improved feature representation through detail enhancement and multi-scale feature fusion [23]. Wang and Zhu combined dynamic convolution with cascaded group attention to improve adaptive multi-scale feature selection [24]. Zhang et al. proposed a YOLOv11-based underwater detector that optimizes feature fusion and attention modeling [25]. More recently, Luo et al. proposed MAS-YOLOv11 to improve underwater

object detection within a YOLOv11-based framework [26]. Zhang et al. developed FSD-Net to enhance degraded underwater representations jointly in the frequency and spatial domains [27]. Gang et al. introduced YOLO-MFD, which combines multi-scale feature modeling with a dynamic detection head for shoreline underwater object detection [28]. For related aquatic-scene detection, Shi et al. proposed LMF-Net, which combines spatial-edge fusion, frequency-decomposed attention, and adaptive downsampling to enhance fine-grained feature representation and suppress water-surface background interference [29]. These recent studies further demonstrate the importance of degradation-aware feature enhancement, multi-scale representation, and adaptive detection mechanisms in complex underwater scenes.

Although these studies have achieved progress, three related issues remain in complex underwater environments. First, scattering and low contrast weaken texture and boundary responses, making discriminative information vulnerable to loss during successive feature extraction. Second, strong local responses caused by suspended particles may be retained or propagated during multi-scale pooling and feature fusion, while excessive background suppression may also weaken object structures. Third, conventional upsampling does not explicitly regulate feature-response changes during resolution recovery, potentially aggravating boundary degradation and cross-level misalignment. Existing methods generally focus on individual stages such as feature extraction, attention recalibration, or localization refinement. The sequential relationship among feature enhancement, background-interference suppression, and stable resolution recovery has received comparatively less attention. Therefore, the proposed method constructs a collaborative processing pipeline across these three stages.

2.2. Attention Mechanisms in Object Detection

Attention mechanisms are intended to guide a network to focus on object regions of interest while suppressing background interference. In object detection tasks, spatial and channel attention modules are commonly integrated into feature extraction networks. Considering the high turbidity and abundant suspended particles in underwater scenes, existing methods usually introduce attention modules into the neck or feature pyramid network to filter noise [2].

From the perspective of the reweighting dimension, channel attention determines the relative importance of different semantic channels, spatial attention identifies spatial locations requiring stronger responses, and self-attention supplements the limited receptive field of local convolutions through long-range dependency modeling. Du et al. combined self-attention with an improved feature pyramid to enhance detail representation and multi-scale perception for underwater small-object detection [3]. YOLO-GE improves object-region responses through attention fusion [2]. Dynamic convolution and cascaded group attention improve feature selection for objects at different scales through content-adaptive convolution and group-wise recalibration [24]. In underwater sonar imagery, Shi et al. further combined cross-scale channel attention with a CNN-based feature-fusion-aware aggregation network to suppress background noise and strengthen multi-scale feature interaction [30].

The above studies demonstrate that attention-based reweighting can reduce interference from large and complex background regions, but two limitations remain. First, generic attention is commonly applied directly to individual or fused feature maps and does not explicitly consider the relationship between spatial pooling and scattering-induced background responses. Second, complex global attention or multi-stage recalibration may introduce additional computational overhead and can assign high weights to irrelevant regions under strong interference. In the proposed framework, channel-wise average- and

max-pooling statistics are used to generate a dual-channel spatial descriptor that jointly recalibrates multiple spatial-pyramid branches. Spatial attention is therefore not treated as an independent methodological innovation; instead, it performs a scenario-oriented feature-selection function after multi-scale pooling and complements the feature enhancement provided by FEM.

2.3. Feature Upsampling and Nearest-Neighbor Interpolation

In the feature fusion stage, upsampling is critical for recovering spatial resolution. Common upsampling methods include nearest-neighbor interpolation and bilinear interpolation. However, underwater features generally suffer from weakened texture and boundary information. Traditional interpolation operations may introduce block artifacts or smooth local response differences, thereby affecting object-boundary localization.

Nearest-neighbor interpolation is computationally simple and directly preserves existing feature values, but it may produce discontinuous spatial responses. Bilinear interpolation obtains smoother outputs through neighborhood-weighted averaging, but this process may weaken local activation differences around blurred boundaries. Learnable upsampling is content-adaptive, but it usually introduces additional parameters and computational overhead. For underwater features already affected by scattering, low contrast, and blur, further smoothing or uncontrolled response changes during upsampling may reduce the distinction between object regions and complex backgrounds.

Li et al. revisited the feature-fused pyramid neck, designed an independent hierarchical pyramid, and introduced soft nearest-neighbor interpolation together with a weighted downsampling factor to alleviate cross-level misalignment and preserve spatial features [5]. SABL improves bounding-box regression accuracy through side-aware boundary localization [22]. These studies indicate that neighborhood modeling and boundary-aware mechanisms contribute to localization performance. Nevertheless, the former mainly addresses feature-pyramid alignment in general object-detection scenes, whereas the latter focuses on boundary regression rather than the upsampling process itself.

Building on the soft-neighbor interpolation mechanism, the proposed SNI module introduces a response-scaling coefficient after nearest-neighbor interpolation to regulate feature-response magnitude during resolution recovery. Rather than replacing the interpolation operator in isolation, SNI receives features processed by FEM and SPP_SA and performs spatial-resolution recovery at the final stage of the collaborative framework. This design links feature enhancement, contextual filtering, and stable cross-level feature fusion.

To provide a consolidated comparison of representative studies and clarify the distinction of the proposed framework, Table 1 summarizes their application scenarios, main strategies, strengths, and principal differences from this study.

Table 1. Comparison of representative related work and the proposed method.

Representative Study	Application Scenario	Main Strategy	Main Strength	Principal Limitation or Difference from This Study
Du et al. [3]	Underwater small-object detection	Self-attention and improved feature pyramid	Enhances detail representation and multi-scale perception	Focuses mainly on attention and pyramid enhancement without jointly considering background-interference suppression and resolution recovery.
Dai et al. [4]	Underwater object detection	Edge-guided representation learning	Strengthens contour representation in degraded scenes	Primarily focuses on edge representation rather than the collaborative relationship among pooling, background interference, and upsampling.

Table 1. *Cont.*

Representative Study	Application Scenario	Main Strategy	Main Strength	Principal Limitation or Difference from This Study
Li and Shi [2]	Underwater object detection	Attention–fusion enhancement	Enhances object responses and reduces background interference	Employs general attention fusion without explicitly coupling spatial-pyramid recalibration with subsequent resolution recovery.
Li et al. [5]	General object detection	Independent hierarchical pyramid, soft nearest-neighbor interpolation, and weighted downsampling	Reduces cross-level misalignment and preserves spatial features	Primarily targets general detection scenes, whereas this study integrates response-scaled interpolation with underwater feature enhancement and contextual filtering.
Zhang et al. [6]	Remote-sensing small-object detection	Feature enhancement, feature fusion, and spatial-context modeling	Balances local perception, multi-scale fusion, and global correlation	Designed for remote-sensing scenes without explicitly considering underwater scattering and suspended-particle interference.
Hu et al. [23]	Underwater object detection	Detail enhancement and multi-scale feature fusion	Improves the representation of weak and degraded object features	Mainly optimizes feature extraction and fusion without jointly investigating spatial-pooling interference and upsampling-response regulation.
Wang and Zhu [24]	Underwater object detection	Dynamic convolution and cascaded group attention	Improves adaptive multi-scale feature selection	Multi-stage feature recalibration may introduce additional complexity and does not explicitly regulate resolution recovery.
Proposed method	Complex underwater object detection	FEM, SPP_SA, and response-scaled SNI	Forms a collaborative pipeline for feature enhancement, contextual filtering, and stable resolution recovery	Its distinction lies in the complementary collaboration of three modules targeting interconnected stages; spatial attention itself is not claimed as an independent innovation.

3. Materials and Methods

The proposed algorithm improves upon the YOLOv11 object detection model, which is used as the baseline. While inheriting the efficient and real-time detection characteristics of previous YOLO models, YOLOv11 introduces C3K2 modules into the backbone and neck for feature extraction and adopts C2PSA, a cross-stage partial structure containing a spatial attention mechanism, to enhance global feature dependencies. Nevertheless, in complex underwater environments characterized by low contrast, scattering-induced background interference, and blurred object boundaries, the baseline model may inadequately preserve discriminative texture and spatial information during feature extraction and multi-scale fusion [23]. This problem is particularly pronounced for small and distant objects because their limited spatial extent makes their texture and boundary information more vulnerable to degradation and repeated downsampling. Therefore, this study proposes a multi-dimensional collaborative enhancement framework for underwater object detection with particular attention to small-object representation and localization in complex scenes.

As shown in Figure 1, the proposed framework introduces three complementary modules into the YOLOv11 architecture. FEM combines local and contextual information through heterogeneous convolutional branches; SPP_SA performs multi-scale contextual aggregation and spatial response recalibration; and SNI module regulates feature responses during resolution recovery. Together, these modules improve feature representation, background-interference suppression, and object localization.

To facilitate understanding of the proposed framework and maintain consistency in the mathematical notation, the principal variables, tensor representations, and structural hyperparameters used in this section are summarized in Table 2.

Table 2. Description of the principal variables and hyperparameters used in the proposed method.

Category	Variable/Parameter	Description
Tensors and maps	$I, \mathbf{X}_0$	Input underwater image and preprocessed input tensor
	$\mathbf{X}$	Input feature map of an individual module
	B_1, B_2, B_3	Outputs of the three heterogeneous FEM branches
	$\widetilde{\mathbf{X}}$	Channel-transformed input feature in SPP_SA
	$\mathbf{Q}_5, \mathbf{Q}_9, \mathbf{Q}_{13}$	Multi-scale pooled feature maps in SPP_SA
	$\mathbf{Z}$	Concatenated multi-scale feature in SPP_SA
	A_{spa}	Spatial attention map generated from $\mathbf{Z}$
	$F_{\text{in}}, F_{\text{out}}$	Input and output feature maps of FEM or SPP_SA
	$\mathbf{P}_3^b, \mathbf{P}_4^b, \mathbf{P}_5^b$	Backbone feature maps at different spatial scales
	$\mathbf{P}_3^n, \mathbf{P}_4^n, \mathbf{P}_5^n$	Intermediate feature maps in the top-down neck path
	$\mathbf{P}_4^o, \mathbf{P}_5^o$	Feature maps generated during bottom-up neck aggregation
	$\mathbf{U}_3, \mathbf{U}_4$	Upsampled features produced by the two SNI operations
Predictions	$\mathbf{Y}$	Three-scale prediction output of the detection head
	$\mathcal{D}$	Final object-detection result set after post-processing
Dimensions	H, W, C	Feature-map height, width, and channel number
	k	Convolution or pooling kernel size
	S	Network input size, set to 640 in this study
Hyperparameters	d	Dilation rate in FEM
	β	Residual scaling coefficient in FEM
	r	Spatial upsampling ratio in SNI
	α	Response-scaling coefficient in SNI

To further clarify the computational workflow, Table 3 presents the forward-inference procedure at the tensor level. The revised pseudocode explicitly identifies the intermediate feature maps, module inputs and outputs, top-down and bottom-up fusion paths, detection-head inputs, and predictor-level post-processing operations.

The tensor-level workflow in Table 3 is consistent with the module computations described in Sections 3.1–3.3 and Equations (1)–(8). Equation (9) further explains the aggregate-response property of the default SNI scaling coefficient.

Table 3. Tensor-level computational pseudocode of the proposed framework.

Line	Tensor-Level Computation
Input	Underwater image I; input size $S = 640$; FEM residual coefficient $\beta = 0.1$; SNI upsampling ratio $r = 2$; and response-scaling coefficient $\alpha = 1/r^2 = 0.25$.
M1	$\mathrm{FEM}(\mathbf{X})$: compute the three heterogeneous branches B_1, B_2, and B_3, followed by multi-branch fusion, residual addition, and ReLU activation according to Equations (1)–(4).
M2	$\mathrm{SPP_SA}(\mathbf{X})$: perform channel transformation, multi-scale pooling, spatial-attention generation, residual recalibration, and output projection according to Equations (5)–(7).
M3	$\mathrm{SNI}(\mathbf{X}; r) = \alpha u_{\mathrm{nearest}}(\mathbf{X}; r)$, as defined in Equation (8).
1	Preprocess the input image: $\mathbf{X}_0 = \mathrm{Preprocess}(I; S)$, including letterbox resizing, channel conversion, tensor formatting, and normalization.
2	Compute the shallow backbone features: $\mathbf{E}_2 = \mathrm{FEM}_1\big(\mathrm{Conv}_{s=2}(\mathrm{Conv}_{s=2}(\mathbf{X}_0))\big)$, $\mathbf{P}_3^b = \mathrm{FEM}_2\big(\mathrm{Conv}_{s=2}(\mathbf{E}_2)\big)$.
3	Generate the remaining backbone features: $\mathbf{P}_4^b = \mathrm{C3K2}\big(\mathrm{Conv}_{s=2}(\mathbf{P}_3^b)\big)$, $\mathbf{P}_5^b = \mathrm{C3K2}\big(\mathrm{Conv}_{s=2}(\mathbf{P}_4^b)\big)$.
4	Process the deepest backbone feature: $\mathbf{P}_5^n = \mathrm{C2PSA}\big(\mathrm{SPP_SA}(\mathbf{P}_5^b)\big)$.
5	Perform the first top-down fusion: $\mathbf{U}_4 = \mathrm{SNI}(\mathbf{P}_5^n; 2)$, $\mathbf{P}_4^n = \mathrm{C3K2}\big(\mathrm{Concat}[\mathbf{U}_4, \mathbf{P}_4^b]\big)$.
6	Perform the second top-down fusion: $\mathbf{U}_3 = \mathrm{SNI}(\mathbf{P}_4^n; 2)$, $\mathbf{P}_3^n = \mathrm{C3K2}\big(\mathrm{Concat}[\mathbf{U}_3, \mathbf{P}_3^b]\big)$.
7	Perform bottom-up aggregation at stride 16: $\mathbf{P}_4^o = \mathrm{C3K2}\big(\mathrm{Concat}[\mathrm{Conv}_{s=2}(\mathbf{P}_3^n), \mathbf{P}_4^n]\big)$.
8	Perform bottom-up aggregation at stride 32: $\mathbf{P}_5^o = \mathrm{C3K2}\big(\mathrm{Concat}[\mathrm{Conv}_{s=2}(\mathbf{P}_4^o), \mathbf{P}_5^n]\big)$.
9	Generate the three-scale detection predictions: $\mathbf{Y} = \mathrm{Detect}([\mathbf{P}_3^n, \mathbf{P}_4^o, \mathbf{P}_5^o])$, where the detection head performs bounding-box decoding and category-score prediction during inference.
10	Apply predictor-level post-processing, including confidence filtering, non-maximum suppression, and coordinate rescaling: $\mathcal{D} = \mathrm{PostProcess}(\mathbf{Y})$.
Output	$\mathcal{D} = \{(b_i, c_i, s_i)\}_{i=1}^{N}$, where b_i, c_i, and s_i denote the bounding box, category label, and confidence score, respectively.

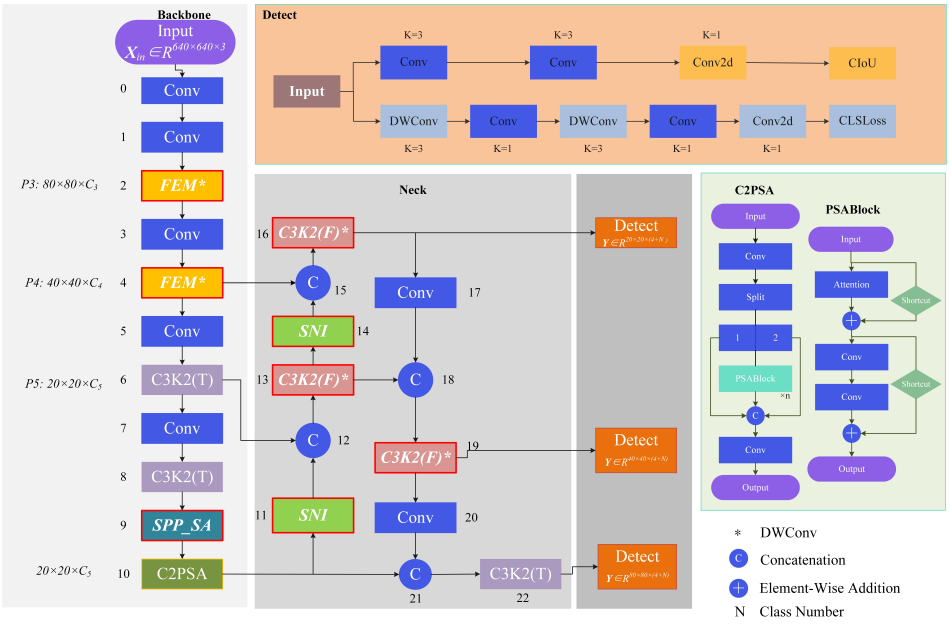


Figure 1. Architecture of the proposed multi-dimensional collaborative enhancement framework.

3.1. Feature Enhancement Module

In the original YOLOv11 framework, the C3K2 module is a key component of the backbone and neck feature fusion stages, and is responsible for multi-scale feature optimization and adjustment. As shown in Figure 1, C3K2 can be divided into C3K2(T) and C3K2(F) according to the setting of the internal C3K parameter. C3K2(T) indicates that the C3K mechanism is enabled, and a more complex cross-stage feature fusion bottleneck is used internally. By contrast, C3K2(F) indicates that this mechanism is disabled, and the internal structure degenerates into a conventional bottleneck stack, as shown in Figure 2. In complex underwater scenes, the conventional bottleneck stack may inadequately preserve local texture and boundary information during feature propagation, particularly for objects occupying small spatial regions or exhibiting blurred appearances. Therefore, this paper designs a Feature Enhancement Module (FEM), whose structure is shown in Figure 3. Rather than relying on repeated convolutions with a single receptive field, FEM employs three heterogeneous convolutional branches to capture complementary local and contextual information.

The module extracts multi-scale features through three heterogeneous convolutional branches. Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ denote the input feature map, where H, W, and C represent its height, width, and number of channels, respectively. The calculation process is as follows:

$$B_1 = C_{3\times3}(C_{1\times1}(\mathbf{X})) \tag{1}$$

where B_1 denotes the output of Branch 1, $C_{1\times1}(\cdot)$ denotes a 1×1 convolution for channel transformation, and $C_{3\times3}(\cdot)$ denotes a standard 3×3 convolution for extracting local texture and boundary information.

$$B_2 = C_{3\times3}^{d=5}(C_{3\times1}(C_{1\times3}(C_{1\times1}(\mathbf{X})))) \tag{2}$$

where B_2 denotes the output of Branch 2, $C_{1\times3}(\cdot)$ and $C_{3\times1}(\cdot)$ denote asymmetric convolutions for modeling directional feature dependencies, and $C_{3\times3}^{d=5}(\cdot)$ denotes a 3×3 dilated convolution with a dilation rate of $d = 5$.

$$B_3 = C_{3\times3}^{d=5}(C_{1\times3}(C_{3\times1}(C_{1\times1}(\mathbf{X})))) \tag{3}$$

where B_3 denotes the output of Branch 3. Branches 2 and 3 employ the asymmetric convolutions in different orders to capture complementary horizontal and vertical feature dependencies. The subsequent dilated convolution enlarges the receptive field without reducing the spatial resolution. Consequently, the three branches jointly preserve fine-grained local information and introduce broader contextual cues.

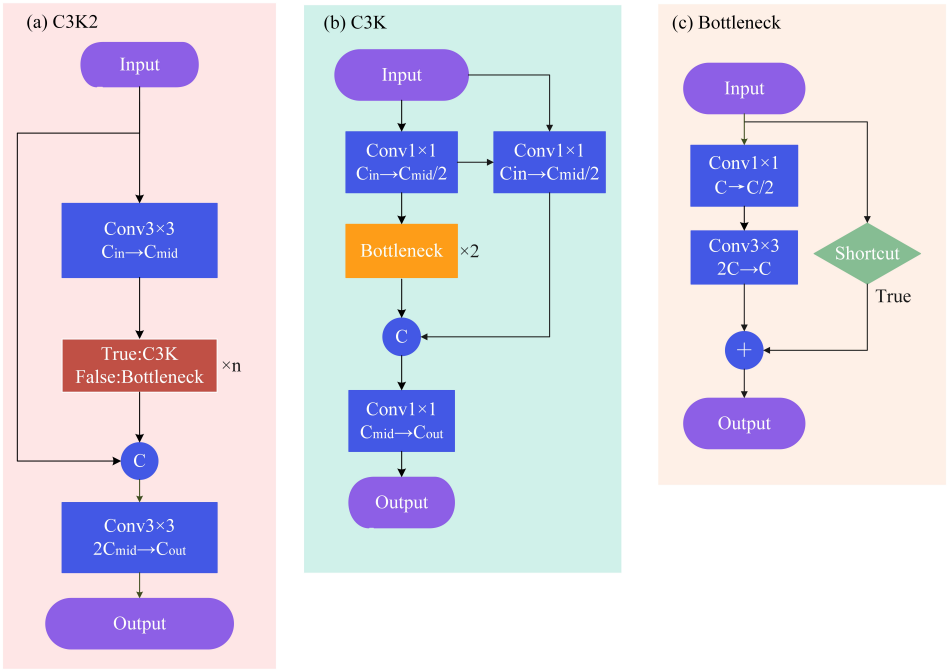


Figure 2. Architecture diagram of the C3K2 module.

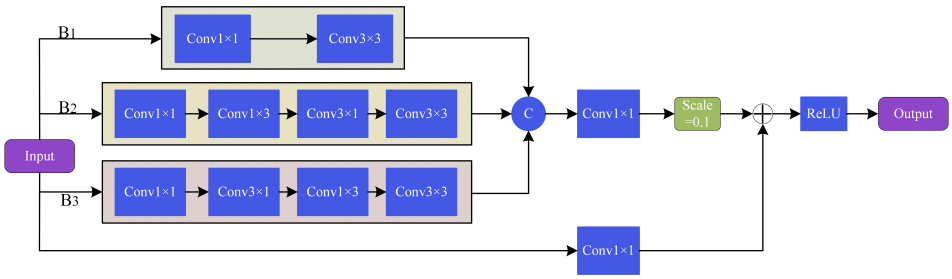


Figure 3. Schematic diagram of the FEM module.

After multi-scale feature extraction, the FEM module fuses the outputs of the three branches through concatenation along the channel dimension, and then uses a 1×1 convolution for channel compression to reduce redundant information. This learnable fusion operation combines local details with contextual information obtained from different receptive fields. At the same time, the module introduces residual learning. A tunable scaling factor $\beta = 0.1$ controls the update strength of the new features, enabling progressive optimization while retaining the original feature information. The formulation is as follows:

$$F_{\text{out}} = \text{ReLU}(\beta C_{1\times1}(\text{Concat}[B_1, B_2, B_3]) + C_{1\times1}(F_{\text{in}})) \tag{4}$$

where F_{in} and F_{out} denote the input and output feature maps of FEM, respectively; $\text{Concat}(\cdot)$ denotes channel-wise concatenation; $C_{1\times1}(\cdot)$ denotes channel compression or alignment using a 1×1 convolution; β denotes the residual scaling coefficient, which is set to 0.1; and $\text{ReLU}(\cdot)$ denotes the rectified linear unit activation. Although FEM is motivated in part by the feature-loss problem of small underwater objects, it is not designed as an exclusive extractor for small objects. Its feature-enhancement capability arises from the

complementary receptive fields of the three branches, learnable multi-branch fusion, and residual information preservation. This general mechanism benefits objects at different scales, while objects occupying small spatial regions or exhibiting blurred boundaries may benefit more because their texture and boundary information is more vulnerable to loss during deep feature extraction. Similar studies have also shown that dynamic convolution and grouped attention mechanisms can enhance multi-scale feature extraction [24].

3.2. Spatial-Attention-Based SPP_SA Module

The Spatial Pyramid Pooling-Fast (SPPF) module in existing detection frameworks uses fixed-size pooling kernels, such as 5×5 kernels, for serial pooling and efficient feature acquisition. However, in underwater scenes, this design has difficulty adapting to variations in object scale and is vulnerable to complex background interference. In turbid underwater environments, light scattering by suspended particles can produce localized bright responses and irregular texture fluctuations, which appear as high-frequency background interference in the feature maps. Conventional spatial pyramid pooling aggregates multi-scale contextual information but does not explicitly distinguish object regions from these interference responses. This interference is particularly detrimental to small underwater objects because their weak and spatially limited responses can be easily overwhelmed by strong background activations. Therefore, this study integrates spatial attention with spatial pyramid pooling and proposes a spatial-attention-based spatial pyramid pooling (SPP_SA) module, whose complete structure is shown in Figure 4. Through spatial-attention-guided multi-scale pooling, the module improves feature selectivity and representation capability.

First, the input feature is transformed by a 1×1 convolution to adjust its channel representation. Three max-pooling branches with kernel sizes of 5×5, 9×9, and 13×13 are then applied to the transformed feature to aggregate contextual information over different spatial ranges. The transformed feature and the three pooled features are concatenated to obtain the multi-scale representation $\mathbf{Z}$.

$$\begin{aligned} \widetilde{\mathbf{X}} &= C^{\text{in}}_{1\times1}(\mathbf{X}), \\ \mathbf{Q}_k &= \text{MaxPool}_{k\times k}(\widetilde{\mathbf{X}}), \qquad k \in \{5, 9, 13\}, \\ \mathbf{Z} &= \text{Concat}[\widetilde{\mathbf{X}}, \mathbf{Q}_5, \mathbf{Q}_9, \mathbf{Q}_{13}]. \end{aligned} \tag{5}$$

where $\widetilde{\mathbf{X}}$ denotes the channel-transformed input feature; $C^{\text{in}}_{1\times1}(\cdot)$ denotes the input 1×1 convolution; $\mathbf{Q}_k$ denotes the feature obtained using a $k \times k$ max-pooling kernel; and $\mathbf{Z}$ denotes the concatenated multi-scale feature.

To suppress background interference, channel-wise average pooling and maximum pooling are applied to the concatenated multi-scale feature $\mathbf{Z}$. Channel-wise average pooling summarizes the distributed response at each spatial location, whereas channel-wise max pooling preserves the strongest activation. Their complementary information allows the module to distinguish spatially persistent object responses from isolated background interference. The two spatial descriptors are then concatenated and processed by a 7×7 convolution followed by a sigmoid function to generate the spatial attention map:

$$A_{\text{spa}} = \sigma(C_{7\times7}(\text{Concat}[\text{AvgPool}_c(\mathbf{Z}), \text{MaxPool}_c(\mathbf{Z})])) \tag{6}$$

where $A_{\text{spa}} \in \mathbb{R}^{H\times W\times1}$ denotes the spatial attention map; $\text{AvgPool}_c(\mathbf{Z})$ and $\text{MaxPool}_c(\mathbf{Z})$ denote average and maximum pooling of the concatenated multi-scale feature $\mathbf{Z}$ along the channel dimension, respectively; $\text{Concat}[\cdot]$ denotes channel-wise concatenation; $C_{7\times7}(\cdot)$ denotes a 7×7 convolution; and $\sigma(\cdot)$ denotes the sigmoid activation function that maps the spatial weights to the interval $(0, 1)$.

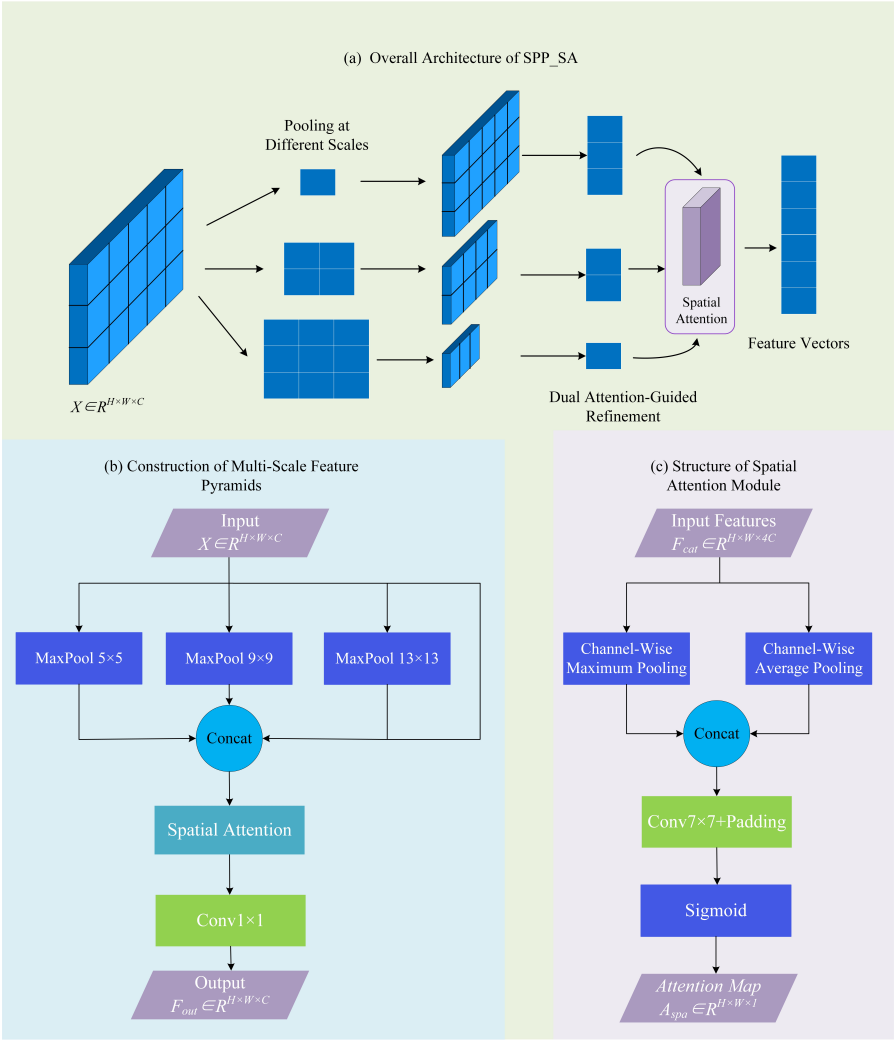


Figure 4. Flow and architecture diagram of the SPP_SA module.

Finally, the learned spatial weights recalibrate the concatenated multi-scale feature $\mathbf{Z}$. A residual connection preserves the original multi-scale response, after which a 1×1 output convolution performs channel compression and feature fusion:

$$F_{\text{out}} = C_{1\times1}^{\text{out}}(\mathbf{Z} \odot A_{\text{spa}} + \mathbf{Z}) \tag{7}$$

where F_{out} denotes the output feature map of SPP_SA; $\odot$ denotes element-wise multiplication, with A_{spa} broadcast along the channel dimension; the residual term $\mathbf{Z}$ preserves the multi-scale representation before spatial recalibration; and $C_{1\times1}^{\text{out}}(\cdot)$ denotes the output projection and channel-compression convolution.

The SPP_SA module does not suppress high-frequency background interference through a fixed smoothing operation alone. Instead, multi-scale pooling provides contextual evidence over different receptive fields, while the learned spatial attention map assigns lower weights to irrelevant background regions and retains spatially consistent object responses. Consequently, the module improves the representation of underwater objects at different scales and reduces the influence of scattering-induced background interference.

3.3. SNI Soft Nearest-Neighbor Interpolation Module

In object detection frameworks, the upsampling module recovers spatial resolution to enable multi-scale feature fusion. However, bilinear interpolation generates new feature values through weighted averaging, which may smooth local activation differences around

object boundaries. This effect can weaken edge and texture responses during feature fusion, particularly for underwater objects with low contrast or blurred appearances. For small objects, whose boundaries may span only a limited number of feature-map locations, such smoothing or response distortion can have a proportionally greater influence on localization accuracy.

To solve this problem, this paper introduces the SNI module, whose structure is shown in Figure 5. By introducing a response-scaling coefficient into standard nearest-neighbor interpolation, the SNI module restores spatial resolution while regulating the magnitude of the replicated feature responses. Given a spatial upsampling ratio r, the calculation is formulated as follows:

$$F_{\text{out}} = \alpha \cdot u_{\text{nearest}}(F_{\text{in}}; r), \qquad \alpha_0 = \frac{1}{r^2} \tag{8}$$

where F_{in} and F_{out} denote the input and upsampled output feature maps within the neck network, respectively; $u_{\text{nearest}}(\cdot; r)$ denotes standard nearest-neighbor interpolation with a spatial upsampling ratio of r; α denotes the response-scaling coefficient; and α_0 denotes its default value. Since nearest-neighbor interpolation replicates each input response over r^2 spatial positions, setting $\alpha_0 = 1/r^2$ compensates for the corresponding increase in the aggregate response magnitude.

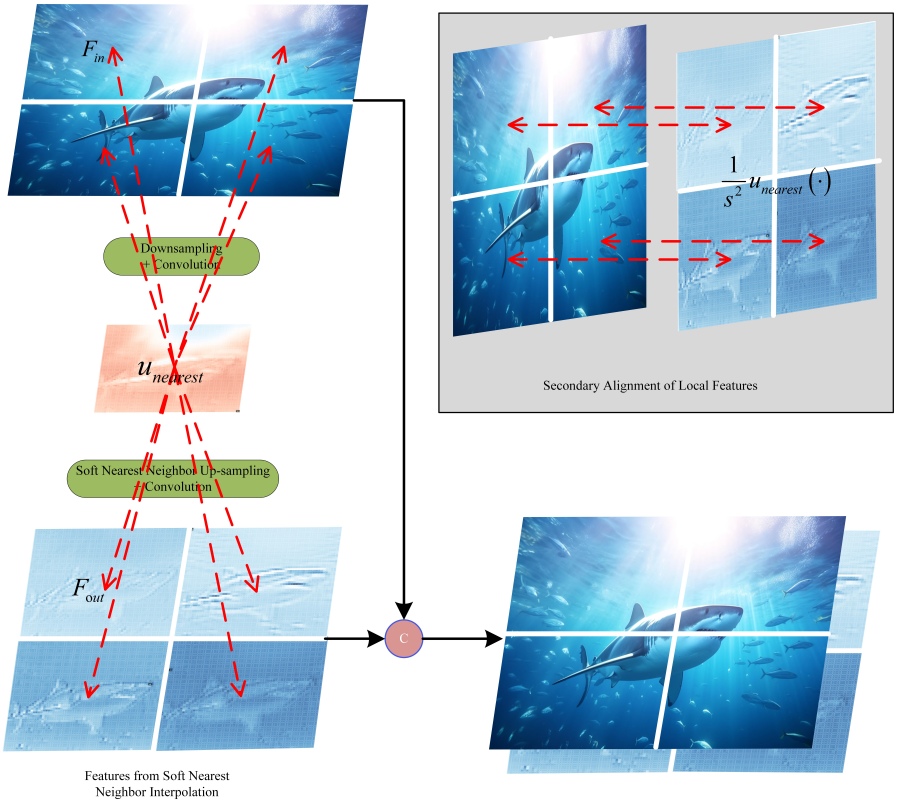


Figure 5. Architecture diagram of the SNI module.

In the proposed network, the spatial upsampling ratio is fixed at $r = 2$ to align adjacent feature-pyramid levels. Therefore, the default response-scaling coefficient is $\alpha_0 = 0.25$.

$$\sum_{i,j} \|F_{\text{out}}(i,j)\|_1 = \alpha r^2 \sum_{i,j} \|F_{\text{in}}(i,j)\|_1 \tag{9}$$

where i and j denote the spatial indices of a feature map and $\|\cdot\|_1$ denotes the L_1-norm across the channel dimension. When $\alpha = \alpha_0 = 1/r^2$, the aggregate feature-response

magnitude before and after upsampling remains unchanged. This constraint prevents the direct spatial replication operation from excessively increasing the feature magnitude during resolution recovery.

In summary, by introducing a response-scaled soft nearest-neighbor interpolation mechanism, the SNI module recovers spatial resolution while regulating the aggregate feature-response magnitude. It alleviates excessive response amplification and supports the preservation of boundary and texture information during multi-scale feature fusion, thereby improving object localization in complex underwater environments, particularly for small objects whose localization is sensitive to feature-resolution loss.

4. Experimental Results and Analysis

4.1. Datasets and Training Details

The experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 graphics processing unit (GPU), an Intel Core i5-13490K central processing unit (CPU), and 64 GB of random-access memory (RAM), running Ubuntu 22.04 Long-Term Support (LTS), using the PyTorch 1.12 framework and the Ultralytics YOLO framework (version 8.3.57). The model was trained using stochastic gradient descent (SGD) with a cosine annealing learning-rate scheduler. The initial learning rate was set to 1×10^{-2}, the weight decay coefficient was 0.0005, and the batch size was 64. All input images were resized to 640×640, and the native data augmentation strategy of the framework was integrated. The maximum number of training epochs was set to 800. For datasets with explicit validation subsets, validation was performed after each epoch, and the checkpoint with the best validation performance was retained for evaluation. The same training settings were applied to the baseline model, the proposed model, and all ablation variants. The RUOD setting is described separately because the public split used in this study contains only training and evaluation partitions.

The automatic optimizer-selection strategy of the Ultralytics framework selected SGD because the training schedule involved more than 10,000 optimization iterations. Rather than manually replacing it with Adam, SGD was retained and consistently applied to all model variants to ensure a fair comparison.

Three public underwater object detection datasets were selected for evaluation: UTDAC2020, Brackish, and RUOD. These datasets cover different object-scale distributions, imaging conditions, object categories, and underwater environments. The UTDAC2020 dataset covers turbid nearshore waters and contains 6460 underwater images, including four categories: echinus, holothurian, scallop, and starfish. The dataset contains objects with substantial scale variation, including small-scale instances, making it suitable for evaluating detection under weak-object-feature conditions. The dataset was divided into 5167 training images, 1193 validation images, and 100 test images.

To characterize the object-scale distribution rather than assuming that all underwater objects are small, the annotations in UTDAC2020 were further analyzed at the 640×640 input resolution used in this study. For this network-oriented scale analysis, objects with bounding-box area $A < 64^2$ were categorized as small-scale instances, those with $64^2 \leq A < 192^2$ as medium-scale instances, and those with $A \geq 192^2$ as large-scale instances. The thresholds were selected with reference to the spatial reduction caused by successive backbone downsampling. As shown in Table 4, UTDAC2020 exhibits a mixed-scale distribution and contains 2053 small-scale and 9814 medium-scale instances.

Table 4. Object-scale distribution of UTDAC2020 at the 640 × 640 input resolution.

Category	Small	Medium	Large	Total
Echinus	1305	5535	17,661	24,501
Holothurian	235	1167	4740	6142
Scallop	192	1382	5676	7250
Starfish	321	1730	6736	8787
Total	2053	9814	34,813	46,680

The Brackish dataset contains 14,674 underwater images captured in fixed scenes and includes six categories, such as fish, small fish, and crab. Its main challenges are geometric deformation and dynamic illumination variation. The dataset was divided into 11,739 training images, 1467 validation images, and 1468 test images. Fixed dataset splits were used in all comparative and ablation experiments. To further evaluate robustness under different underwater conditions, the RUOD dataset was additionally introduced. RUOD contains 14,000 underwater images and ten object categories, including fish, echinus, coral, starfish, holothurian, scallop, diver, cuttlefish, turtle, and jellyfish. The dataset includes images collected from diverse sources and contains substantial variations in background, object category, scale, color cast, visibility, and illumination. Following the public RUOD split used in previous studies, 9800 images were used for training and the remaining 4200 images were used as an evaluation split. Since this split does not provide a separate validation subset, the 4200-image evaluation split was assigned to both the validation and test fields in the Ultralytics configuration and was therefore used for checkpoint selection and metric reporting. Accordingly, the RUOD results are treated as a supplementary cross-dataset robustness evaluation rather than an independent test-set evaluation.

To evaluate the robustness of the proposed method with respect to training randomness, six independent training runs were conducted using different random seeds. All runs used the same dataset split, number of training epochs, input resolution, optimizer and learning-rate schedule, data-augmentation strategy, hardware and software environment, and model configuration described above. For each run, the checkpoint with the highest validation $mAP_{50:95}$ was selected for evaluation. Because repeated baseline runs were unavailable, the YOLOv11 results reported in the main comparison experiment were treated as fixed reference values. Two-sided one-sample exact Wilcoxon signed-rank tests were then conducted separately for mAP_{50} and $mAP_{50:95}$, with the significance level set to 0.05.

4.2. Evaluation Metrics

To objectively and comprehensively evaluate the detection performance and computational characteristics of the proposed multi-dimensional collaborative enhancement framework, this paper adopts the following standard evaluation metrics commonly used in object detection: Average Precision (AP), mean Average Precision (mAP), number of parameters (Params), giga floating-point operations (GFLOPs), and frames per second (FPS).

Precision (P) and Recall (R) are calculated as follows:

$$P = \frac{TP}{TP + FP}, \tag{10}$$

$$R = \frac{TP}{TP + FN}, \tag{11}$$

where TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative detections, respectively. Precision and Recall are first calculated for each object category, and the reported overall values summarize the performance across all categories.

AP measures the detection performance of the model for an individual category and is equal to the area under the precision–recall curve, while mAP is the mean AP over all object categories. In the metric notation, the subscript denotes the intersection-over-union (IoU) threshold used to determine a correct detection. Specifically, mAP_{50} represents mAP at an IoU threshold of 0.50, whereas $mAP_{50:95}$ represents mAP averaged over IoU thresholds from 0.50 to 0.95 at intervals of 0.05.

To provide an additional accuracy-oriented measure for fixed-threshold comparison, a detection-oriented Accuracy (*Acc*) is introduced. Since true negatives are not naturally defined in object detection, *Acc* is calculated using true-positive, false-positive, and false-negative detections as follows:

$$Acc = \frac{TP}{TP + FP + FN}. \tag{12}$$

For all datasets, final detections were generated using a confidence threshold of 0.25 and a non-maximum suppression (NMS) IoU threshold of 0.70. A prediction was counted as a true positive only when it matched a ground-truth box of the same category with an IoU of at least 0.50 under one-to-one matching. Unmatched predictions and unmatched ground-truth boxes were counted as false positives and false negatives, respectively. The resulting *Acc* is reported as a complementary fixed-threshold metric alongside the standard mAP metrics.

In addition, the number of parameters and GFLOPs are used to evaluate the spatial and temporal complexity of the model, while FPS is used to assess the real-time inference performance on actual hardware.

4.3. Comparative Experiments

The comparison methods in Table 5 are accompanied by their corresponding references for method identification and verification. The YOLOv11 baseline and the proposed method were evaluated under the unified experimental settings described above.

The proposed method was first compared with the baseline YOLOv11 model on the UTDAC2020 dataset. It improves mAP_{50} from 0.814 to 0.829 and $mAP_{50:95}$ from 0.468 to 0.494. The smaller gap between mAP_{50} and $mAP_{50:95}$ indicates that the proposed method retains better performance under stricter localization requirements, particularly for objects with blurred boundaries.

Table 5. Comparative results of same-domain methods on the UTDAC2020 dataset.

Method	Echinus	Holothurian	Scallop	Starfish	mAP_{50}	$mAP_{50:95}$
SSD [7]	0.419	0.308	0.412	0.462	0.775	0.400
FSAF [8]	0.402	0.318	0.389	0.479	0.758	0.397
NAS-FCOS [31]	0.460	0.370	0.475	0.525	0.828	0.458
YOLOv7 [9]	0.431	0.331	0.466	0.483	0.786	0.427
Faster R-CNN [10]	0.449	0.355	0.448	0.498	0.812	0.438
Cascade RPN [11]	0.462	0.361	0.473	0.522	0.804	0.454
PAFPN [14]	0.449	0.354	0.450	0.504	0.808	0.439
Cascade R-CNN w/FPN [12]	0.456	0.360	0.460	0.516	0.806	0.448
Libra R-CNN [15]	0.451	0.356	0.454	0.506	0.804	0.442
DetectoRS [16]	0.464	0.368	0.465	0.516	0.808	0.453
Deformable DETR [17]	0.437	0.328	0.441	0.471	0.771	0.419
Swin Transformer [18]	0.457	0.372	0.462	0.521	0.828	0.453
DAB-DETR [19]	0.432	0.336	0.465	0.485	0.788	0.429
ERL w/Faster R-CNN [4]	**0.498**	0.405	0.488	0.541	**0.832**	0.483
ERL w/RetinaNet [4]	**0.498**	**0.406**	0.488	**0.543**	0.828	0.484
YOLOv11 [32]	0.490	0.354	0.512	0.516	0.814	0.468
Ours	**0.515**	0.390	**0.531**	0.541	0.829	**0.494**

Note: The four category columns report class-wise $AP_{50:95}$. Bold values indicate the best performance, and underlined values indicate the second-best performance.

Compared with the other methods in Table 5, the proposed method achieves the second-best mAP_{50} of 0.829, which is only 0.003 lower than the best result obtained by ERL w/Faster R-CNN. More importantly, it achieves the highest $mAP_{50:95}$ of 0.494, exceeding the second-best ERL w/RetinaNet result of 0.484. These results demonstrate that the proposed method provides competitive overall detection performance and more accurate localization under stricter IoU thresholds.

The class-wise $AP_{50:95}$ results further show that the proposed method improves the detection performance for echinus and scallop from 0.490 and 0.512 to 0.515 and 0.531, respectively. These improvements indicate that the heterogeneous convolutional structure enhances the representation of object textures and blurred boundaries in complex underwater scenes.

A comprehensive analysis of recall and precision shows that the improved algorithm achieves an overall enhancement in positive-sample detection capability. The recall and precision of the algorithms before and after improvement are plotted as precision–recall (PR) curves in Figure 6. The PR curve of the improved algorithm is closer to the upper-right corner of the coordinate axis, indicating that it achieves a better balance between recall and precision at different thresholds. Thus, it can detect more objects while maintaining higher accuracy. To intuitively demonstrate the detection advantages of the proposed method in complex underwater environments, Figure 7 further presents qualitative detection comparisons between the baseline model and the improved algorithm under multi-scale object, low-contrast, and occlusion scenarios.

As shown in Table 6, the proposed method exhibited stable performance across the six independent runs. The median mAP_{50} and $mAP_{50:95}$ values were 0.82920 and 0.49356, respectively, with narrow interquartile ranges. All six runs achieved higher values than the fixed YOLOv11 reference results of 0.814 and 0.468. The two-sided one-sample exact Wilcoxon signed-rank tests yielded $W = 0$ and $p = 0.03125$ for both metrics. Therefore, the repeated-run results were significantly higher than the corresponding baseline reference values at the 0.05 significance level. These results indicate that the performance improvements are consistent across different random initializations rather than being attributable to a single favorable training run.

The proposed method was subsequently compared with the baseline YOLOv11 model on the Brackish dataset. It improves mAP_{50} from 0.987 to 0.991 and $mAP_{50:95}$ from 0.830 to 0.858, indicating improved detection and bounding-box localization performance.

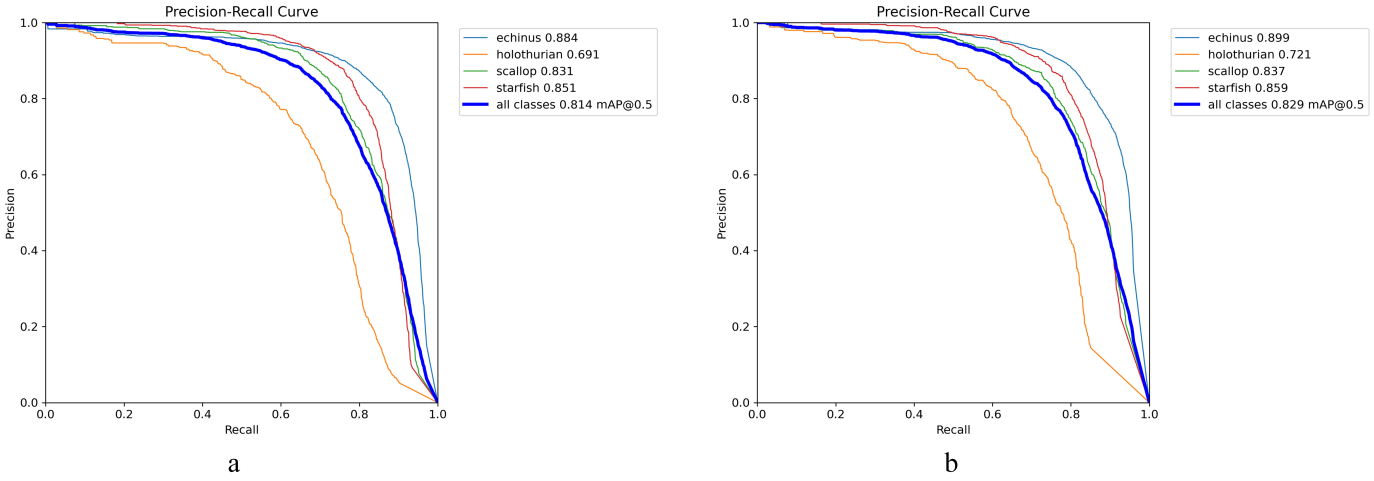


Figure 6. Precision–recall curves on the UTDAC2020 dataset: (**a**) baseline model; (**b**) improved method.

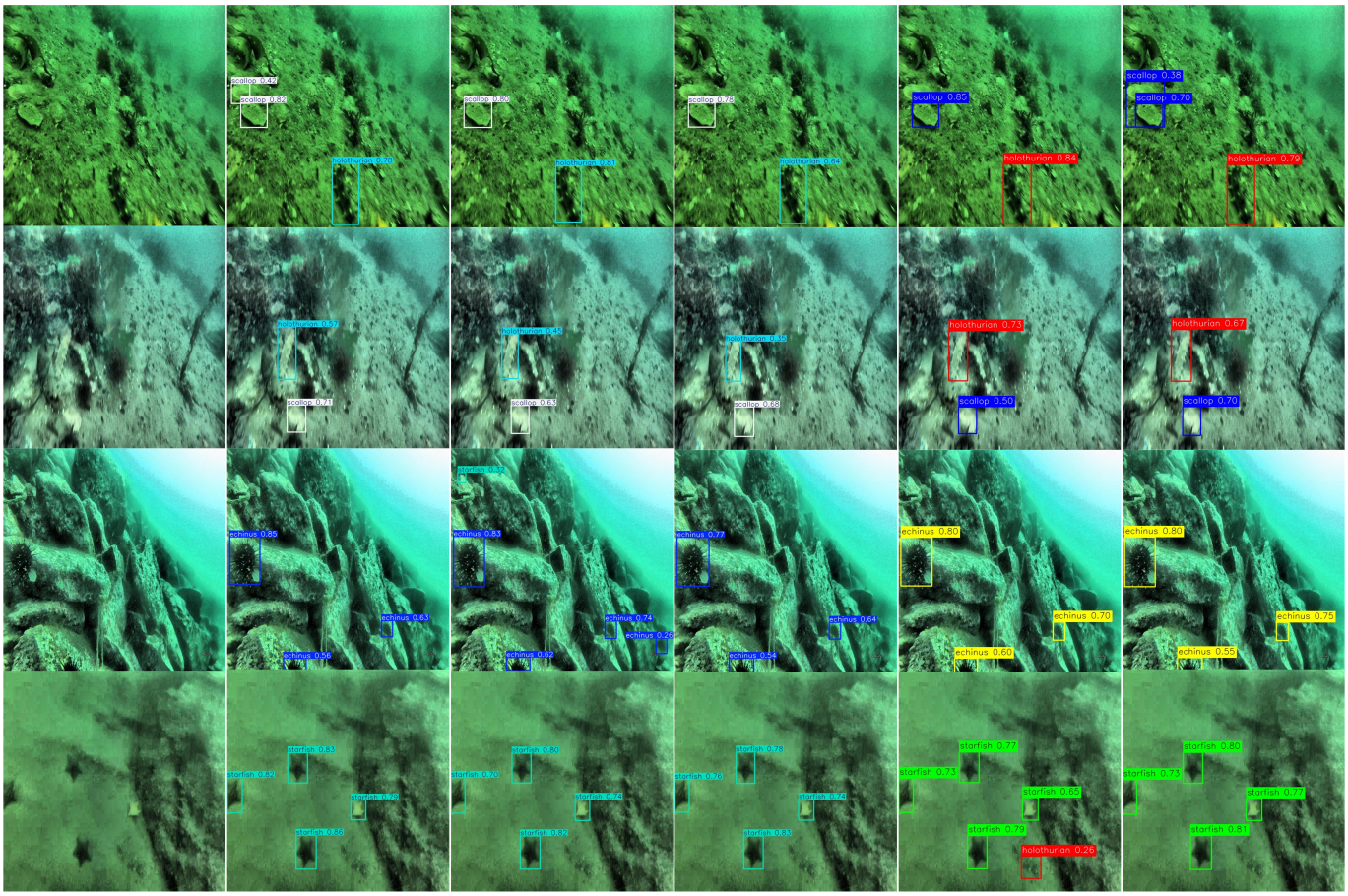


Figure 7. Detection results compared with other methods.

Table 6. Results of six independent runs and exact one-sample Wilcoxon signed-rank tests on UTDAC2020.

Independent Run	mAP_{50}	$mAP_{50:95}$
Run 1	0.82842	0.49366
Run 2	0.82915	0.49346
Run 3	0.82979	0.49415
Run 4	0.83010	0.49213
Run 5	0.82914	0.49315
Run 6	0.82925	0.49406
Median	0.82920	0.49356
Interquartile range	[0.82914, 0.82966]	[0.49323, 0.49396]
YOLOv11 reference	0.81400	0.46800
Wilcoxon statistic W	0	0
Two-sided exact p-value	0.03125	0.03125

As shown in Table 7, the proposed method achieves the highest mAP_{50} of 0.991 among the compared methods. It outperforms ERL w/Faster R-CNN, YOLOv11, and the other representative detection models, confirming the effectiveness of the proposed multi-dimensional collaborative enhancement framework in complex underwater scenes.

As a supplementary cross-dataset robustness evaluation under more diverse underwater conditions, the baseline YOLOv11 and the proposed method were evaluated on the RUOD evaluation split using the same training settings. As shown in Table 8, the proposed method improves all four evaluation metrics. The mAP_{50} increases from 0.847 to 0.856, while $mAP_{50:95}$ increases from 0.620 to 0.625. Recall also increases from 0.772 to 0.781, indicating that the proposed framework maintains consistent improvements in a dataset containing more diverse categories, backgrounds, and imaging conditions.

Table 7. Comparative results of same-domain methods on the Brackish dataset.

Method	mAP_{50}
Faster R-CNN w/FPN [10]	0.629
Deformable DETR [17]	0.823
Swin Transformer [18]	0.852
RetinaNet w/FPN [13]	0.937
Boosting R-CNN [20]	0.974
D2Det [21]	0.981
SABL w/Faster R-CNN [22]	0.986
ERL w/Faster R-CNN [4]	<u>0.988</u>
YOLOv11 [32]	0.987
Ours	**0.991**

Note: Bold values indicate the best performance, and underlined values indicate the second-best performance.

Table 8. Comparison with the YOLOv11 baseline on the RUOD dataset.

Method	P	R	mAP_{50}	$mAP_{50:95}$
YOLOv11	<u>0.866</u>	<u>0.772</u>	<u>0.847</u>	<u>0.620</u>
Ours	**0.867**	**0.781**	**0.856**	**0.625**

Note: Bold values indicate the best performance, and underlined values indicate the second-best performance.

To further provide a fixed-threshold accuracy comparison, the detection-oriented Accuracy defined in Equation (12) was evaluated for YOLOv11 and the proposed method on all three datasets. The corresponding TP, FP, FN, and Accuracy values are reported in Table 9.

Table 9. Detection-oriented Accuracy of YOLOv11 and the proposed method on the three underwater datasets.

Dataset	Method	TP	FP	FN	Acc (%)
UTDAC2020	YOLOv11	293	102	94	59.918
UTDAC2020	Ours	255	27	132	**61.594**
Brackish	YOLOv11	3389	134	77	94.139
Brackish	Ours	3417	94	49	**95.983**
RUOD	YOLOv11	17,220	2769	5749	66.905
RUOD	Ours	17,552	3041	5417	**67.482**

Note: Bold values indicate the higher Accuracy value for each dataset.

As shown in Table 9, the proposed method achieves detection Accuracy values of 61.594%, 95.983%, and 67.482% on UTDAC2020, Brackish, and RUOD, respectively, compared with 59.918%, 94.139%, and 66.905% for YOLOv11. These correspond to improvements of 1.676, 1.844, and 0.577 percentage points, respectively. The results provide an additional fixed-threshold comparison that is complementary to the threshold-integrated mAP evaluation.

4.4. Computational Complexity Analysis

To quantify the additional computational cost introduced by the proposed modules, the static model complexity and measured inference speed of the YOLOv11n baseline and the proposed model were evaluated under the same 640×640 input resolution on a single RTX 4090 GPU. Params and GFLOPs are used to characterize model size and computational demand, respectively, while FPS measures actual inference throughput on the experimental platform. The results are shown in Table 10.

As shown in Table 10, the proposed framework increases the number of parameters from 2.6 M to 3.2 M, corresponding to a 23.08% increase, while GFLOPs increase from 6.3 to 10.7, corresponding to a 69.84% increase. Therefore, the proposed method does not aim to reduce theoretical computational complexity; instead, it introduces additional computation to improve feature representation, contextual recalibration, and feature recovery. Despite

this increase in static complexity, the measured inference speed remains comparable, changing from 127.83 to 128.80 FPS on the same RTX 4090 platform. These results indicate an accuracy–complexity trade-off rather than a lightweight-design advantage.

Table 10. Comparison of computational complexity and inference speed.

Model	Params (M)	GFLOPs	FPS
YOLOv11n	2.6	6.3	127.83
Ours	3.2	10.7	128.80
Relative change	+23.08%	+69.84%	+0.76%

Note: Params denotes the number of model parameters. GFLOPs are reported for a single 640 × 640 forward pass, and FPS is measured on the same RTX 4090 platform for both models.

4.5. Ablation Study

To verify the contribution of each module, systematic ablation experiments were conducted on the UTDAC2020 dataset. The results are shown in Table 11.

Table 11. Ablation study on the UTDAC2020 dataset.

FEM	SPP_SA	SNI	Echinus	Holothurian	Scallop	Starfish	mAP_{50}	$mAP_{50:95}$
			0.885	0.690	0.831	0.851	0.814	0.468
✓			0.896	**0.728**	0.831	**0.862**	**0.829**	0.489
	✓		0.892	0.706	0.829	0.850	0.819	0.469
		✓	0.892	0.702	0.819	0.852	0.816	0.475
✓	✓		<u>0.897</u>	0.715	0.832	<u>0.860</u>	0.826	<u>0.491</u>
✓		✓	0.890	<u>0.722</u>	<u>0.834</u>	**0.862**	<u>0.827</u>	0.489
	✓	✓	0.889	0.710	0.830	0.856	0.823	0.486
✓	✓	✓	**0.899**	0.721	**0.837**	0.859	**0.829**	**0.494**

Note: Bold values indicate the best performance, and underlined values indicate the second-best performance. A checkmark indicates that the corresponding module is included.

The ablation results show that FEM provides the largest single-module improvement, increasing mAP_{50} from 0.814 to 0.829 and $mAP_{50:95}$ from 0.468 to 0.489. At the category level, FEM increases the AP_{50} of holothurian from 0.690 to 0.728 and improves the AP_{50} values of echinus and starfish from 0.885 and 0.851 to 0.896 and 0.862, respectively. When applied independently, SPP_SA increases mAP_{50} to 0.819, whereas SNI increases $mAP_{50:95}$ to 0.475. The pairwise combinations FEM+SPP_SA, FEM+SNI, and SPP_SA+SNI achieve mAP_{50} values of 0.826, 0.827, and 0.823, respectively. The complete model integrating all three modules achieves a joint-best mAP_{50} of 0.829 and the highest $mAP_{50:95}$ of 0.494. These results demonstrate that the three modules provide complementary improvements in feature representation, background-interference suppression, and localization performance.

FEM enhances local texture, boundary, and contextual representations through three heterogeneous convolutional branches. However, stronger feature responses may also retain irrelevant activations generated by suspended particles and non-uniform illumination. SPP_SA complements FEM by aggregating contextual information over multiple spatial ranges and using learned spatial-attention weights to reduce the influence of irrelevant background regions while preserving spatially consistent object responses. During subsequent feature-pyramid fusion, SNI applies response-scaled nearest-neighbor interpolation to recover spatial resolution while regulating the aggregate feature-response magnitude. The contextual features processed by FEM and SPP_SA therefore provide more reliable information for cross-level resolution recovery. Together, the three modules form a sequential collaborative pipeline consisting of feature enhancement, spatial contextual recalibration, and response-regulated upsampling, thereby improving detection robustness in complex underwater scenes.

In addition, to investigate the sensitivity of the SNI module to the response-scaling coefficient, additional ablation experiments were conducted by varying α while keeping

the spatial upsampling ratio fixed at $r = 2$. The fixed upsampling ratio ensures the spatial alignment of adjacent feature-pyramid levels. According to the SNI formulation, the default coefficient is $\alpha = 1/r^2 = 0.25$. Therefore, four values, $\{0.125, 0.25, 0.5, 1.0\}$, were selected in a geometric progression with a factor of two to cover different scaling strengths. Here, $\alpha = 0.125$ represents stronger feature scaling, $\alpha = 0.25$ is the default setting, and $\alpha = 0.5$ represents weaker scaling, while $\alpha = 1.0$ indicates that no additional response scaling is applied. The results are shown in Table 12.

Table 12. Ablation experiments on the response-scaling coefficient α of the SNI module.

α	P	R	mAP_{50}	$mAP_{50:95}$
0.125	<u>0.829</u>	0.746	0.822	0.479
0.25	**0.831**	0.745	**0.829**	**0.494**
0.5	0.819	**0.755**	0.823	0.478
1.0	0.828	<u>0.747</u>	<u>0.824</u>	<u>0.481</u>

Note: Bold values indicate the best performance, and underlined values indicate the second-best performance.

The experimental results show that $\alpha = 0.25$ provides the best overall performance, achieving an mAP_{50} of 0.829 and an $mAP_{50:95}$ of 0.494. Although $\alpha = 0.5$ produces the highest Recall of 0.755, its Precision and mAP values are lower, indicating that the increased number of detected responses does not result in more accurate overall detection. Reducing α to 0.125 also decreases both mAP metrics, whereas removing the additional response scaling by setting $\alpha = 1.0$ produces inferior results to the default setting. Therefore, $\alpha = 0.25$ was selected as the final coefficient because it achieves the best balance between Precision, Recall, and localization performance.

5. Discussion

The experimental results demonstrate consistent improvements of the proposed framework under different underwater conditions. Compared with the YOLOv11 baseline, the proposed method improves mAP_{50} and $mAP_{50:95}$ on UTDAC2020 from 0.814 to 0.829 and from 0.468 to 0.494, corresponding to relative improvements of 1.84% and 5.56%, respectively. On Brackish, the corresponding relative improvements are 0.41% and 3.37%. The supplementary evaluation on RUOD further increases mAP_{50} from 0.847 to 0.856 and $mAP_{50:95}$ from 0.620 to 0.625, corresponding to relative improvements of 1.06% and 0.81%, respectively. These results indicate that the proposed collaborative enhancement strategy provides consistent performance gains across datasets with different object-scale distributions and underwater imaging characteristics.

The UTDAC2020 scale analysis further shows that the evaluated scenes contain a mixed distribution of object sizes, including small-scale instances. Such objects are particularly sensitive to texture attenuation, background interference, and spatial-information loss during feature extraction and fusion. The ablation results also clarify the complementary roles of the three proposed modules. FEM provides the largest single-module improvement in $mAP_{50:95}$, increasing it from 0.468 to 0.489 by enhancing local texture, boundary, and contextual representations. SPP_SA complements this enhancement by using multi-scale contextual aggregation and spatial recalibration to reduce the influence of irrelevant underwater background responses. During subsequent feature fusion, SNI regulates feature-response magnitude while recovering spatial resolution. Although the individual modules produce different levels of improvement, their joint use yields the highest $mAP_{50:95}$ of 0.494, supporting the design motivation of coordinating feature enhancement, contextual filtering, and resolution recovery rather than treating these stages independently.

The complementary fixed-threshold Accuracy evaluation provides additional evidence from a different evaluation perspective. Compared with YOLOv11, the proposed method improves detection Accuracy by 1.676, 1.844, and 0.577 percentage points on UTDAC2020,

Brackish, and RUOD, respectively. Together with the mAP results, these findings indicate that the proposed method maintains improved detection performance under both threshold-integrated and fixed-threshold evaluation criteria.

The improved detection performance is accompanied by additional computational cost. The proposed framework increases the number of parameters from 2.6 M to 3.2 M and GFLOPs from 6.3 to 10.7, while the measured inference speed remains comparable on the same RTX 4090 platform. Therefore, the proposed method should be interpreted as an accuracy–complexity trade-off rather than a complexity-reduction or lightweight solution. The increased theoretical computational demand may become more relevant when the detector is transferred from a desktop GPU to resource-constrained underwater platforms.

This study also has several limitations and opportunities for further improvement. First, although UTDAC2020, Brackish, and the supplementary RUOD evaluation cover different underwater imaging conditions and object distributions, broader validation across additional water types, imaging devices, and application domains is still required. Second, extremely low illumination, severe occlusion, and strong motion blur may continue to weaken discriminative visual information and reduce detection performance. Third, the current efficiency evaluation was conducted on a desktop GPU, and the computational characteristics of the proposed framework have not yet been verified on embedded underwater platforms. Future work will therefore focus on broader cross-domain evaluation, model lightweighting and compression, and practical deployment on resource-constrained underwater devices.

6. Conclusions

This paper addresses object detection in complex underwater scenes, with particular attention to small and distant objects whose weak texture and boundary information is more susceptible to optical degradation, background interference, and spatial-information loss. A multi-dimensional collaborative enhancement framework is proposed to coordinate feature enhancement, contextual filtering, and resolution recovery within a YOLOv11-based detector.

The Feature Enhancement Module (FEM) employs heterogeneous convolutional branches to strengthen local texture, boundary, and contextual representations. The spatial-attention-based spatial pyramid pooling (SPP_SA) module combines multi-scale contextual aggregation with spatial recalibration to reduce the influence of irrelevant underwater background responses. The Soft Nearest-Neighbor Interpolation (SNI) module applies response-scaled nearest-neighbor interpolation to regulate feature-response magnitude during resolution recovery and support stable cross-level feature fusion. Together, the three modules form a complementary processing pipeline for feature enhancement, background-interference suppression, and object localization.

On UTDAC2020, the proposed method achieves an mAP_{50} of 0.829 and an $mAP_{50:95}$ of 0.494, compared with 0.814 and 0.468 for the YOLOv11 baseline, respectively. On Brackish, the proposed method achieves an mAP_{50} of 0.991. Supplementary evaluation on RUOD yields an mAP_{50} of 0.856 and an $mAP_{50:95}$ of 0.625, compared with 0.847 and 0.620 for YOLOv11. These results demonstrate consistent improvements under different underwater conditions and support the effectiveness of the proposed framework for mixed-scale underwater object detection, including small-scale objects.

The proposed improvements are accompanied by increased theoretical computational cost, and broader validation under additional underwater conditions is still required. Future work will therefore focus on evaluation across more diverse water types and imaging devices, model lightweighting and compression, and deployment on resource-constrained underwater platforms.

https://doi.org/10.3390/electronics15163634

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

Funding: This research was funded by the China Three Gorges Corporation (Grant No. 2324020012).

Data Availability Statement: The datasets used in this study are publicly available from the original sources cited in the article. The source code is not publicly available. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors S.Z. and X.L. were employed by China Yangtze Power Co., Ltd., a subsidiary of China Three Gorges Corporation. This study received funding under a project of China Three Gorges Corporation (Grant No. 2324020012). The funder was not involved in the study design, data collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication. The remaining authors declare no commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liu, T.L.; Chen, X.Z.; Lu, L.Q.; Zhang, Y.T.; Zhang, W.Z. Underwater biological detection model based on improved YOLOv8 and multi-scale fusion. *J. Zhongkai Univ. Agric. Eng.* **2026**, 1–16.
2. Li, Q.; Shi, H. YOLO-GE: An attention fusion enhanced underwater object detection algorithm. *J. Mar. Sci. Eng.* **2024**, *12*, 1885. [CrossRef]
3. Du, R.S.; Wang, Z.S.; Meng, L.D.; Jing, Y. Underwater small target detection based on self-attention and improved pyramid. *J. Shenzhen Univ. (Sci. Eng. Ed.)* **2025**, *42*, 428–436. [CrossRef]
4. Dai, L.; Liu, H.; Song, P.; Tang, H.; Ding, R.; Li, S. Edge-guided representation learning for underwater object detection. *CAAI Trans. Intell. Technol.* **2024**, *9*, 1078–1091. [CrossRef]
5. Li, H. Rethinking features-fused-pyramid-neck for object detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 74–90.
6. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for small object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [CrossRef]
7. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision–ECCV 2016, Proceedings, Part I*; Springer: Amsterdam, The Netherlands, 2016; pp. 21–37.
8. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.
9. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
11. Vu, T.; Jang, H.; Pham, T.X.; Yoo, C.D. Cascade RPN: Delving into high-quality region proposal network with adaptive convolution. In *Advances in Neural Information Processing Systems*; ACM Digital Library: New York, NY, USA, 2019.
12. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
13. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
14. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
15. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.
16. Qiao, S.; Chen, L.-C.; Yuille, A. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 10208–10219.
17. 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, Vienna, Austria, 4 May 2021.

18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
19. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
20. 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]
21. Cao, J.; Cholakkal, H.; Anwer, R.M.; Khan, F.S.; Pang, Y.; Shao, L. D2Det: Towards high quality object detection and instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11482–11491.
22. Wang, J.; Zhang, W.; Cao, Y.; Chen, K.; Pang, J.; Gong, T.; Shi, J.; Loy, C.C.; Lin, D. Side-aware boundary localization for more precise object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 403–419.
23. Hu, Y.J.; Yang, Y.; Wu, X.N.; Wang, M.X.; Pan, Z.P. Underwater object detection based on detail enhancement and multi-scale feature fusion. *Comput. Syst. Appl.* **2025**, *34*, 69–78. [CrossRef]
24. Wang, Y.B.; Zhu, W.C. Underwater object detection algorithm based on dynamic convolution and cascaded group attention. *Laser Optoelectron. Prog.* **2026**, *63*, 0637008.
25. Zhang, H.R.; Feng, W.M.; Yang, L.X.; Ma, Y.J. CSAF-YOLO: An improved underwater small-object detection algorithm based on YOLOv11. *Comput. Appl.* **2026**, *46*, 1578–1585. [CrossRef]
26. Luo, Y.; Wu, A.; Fu, Q. MAS-YOLOv11: An Improved Underwater Object Detection Algorithm Based on YOLOv11. *Sensors* **2025**, *25*, 3433. [CrossRef] [PubMed]
27. Zhang, C.; Wu, S.; Huang, B.; Zhao, B.; Cui, F.; Li, X. FSD-Net: Underwater Object Detection Based on Frequency and Spatial Domain Feature Enhancement. *Front. Artif. Intell.* **2026**, *9*, 1770342. [CrossRef] [PubMed]
28. Gang, Y.; Li, T.; Li, S.; Shang, Y. YOLO-MFD: A Multi-Scale Feature and Dynamic Head Framework for Prefabricated Shoreline Underwater Object Detection. *Sci. Rep.* **2026**, *16*, 10971. [CrossRef] [PubMed]
29. Shi, P.; Wu, H.; Gu, Y.; Xin, Y. LMF-Net: Lightweight multidimensional feature fusion network for floating debris detection on water surfaces. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 4206111. [CrossRef]
30. Shi, P.; Wang, H.; Zhang, Q.; Xin, Y. Cross-scale channel attention and feature fusion-aware aggregation for sonar images object detection. *Knowl.-Based Syst.* **2026**, *337*, 115371. [CrossRef]
31. Wang, N.; Gao, Y.; Chen, H.; Wang, P.; Tian, Z.; Shen, C.; Zhang, Y. NAS-FCOS: Fast neural architecture search for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11940–11948.
32. Jocher, G.; Qiu, J. *Ultralytics YOLO11*, version 11.0.0; Ultralytics: Frederick, MD, USA, 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 7 August 2026).

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