MFE-Net: A Hybrid Perception and Manifold-Preserving Framework for Robust Fine-Grained Underwater Object Detection

Jiaxin Chen[1], Xin Wang[1], Yuxu Peng[1], Dengyong Zhang[1(✉)], Lei Wang[1], and Yongjie Zhang[1]

[1] School of Computer Science and Technology, Changsha University of Science & Technology, China
zhdy@csust.edu.cn

Abstract. High-performance underwater object detection zconstitutes a pivotal technological foundation for advancing marine intelligent computing and facilitating the autonomous deployment of subsea platforms. Nevertheless, underwater visual perception is inherently constrained by non-linear optical degradation, manifesting as an ill-posed inverse problem that induces "semantic confusion" and the "feature annihilation" of sub-pixel entities. Conventional deep learning architectures, predicated on strided downsampling, frequently function as irreversible low-pass filters, precipitating a significant degradation of information entropy regarding sparse high-frequency geometric manifolds during hierarchical transmission. To address these challenges, this study formulates MFE-Net, a hierarchical hybrid perception and manifold-preserving framework. The Hybrid Perception Feature Extraction (C3CF) module harmonizes local convolutional inductive biases with global attention recalibration to effectively suppress non-structured scattering noise and enhance deep feature representation. Building upon this, the Multi-Scale Feature Enhancement Neck (MFE-Neck) incorporates a lossless Fine-grained Feature Enhancement Branch and an Efficient Multi-Scale Feature Integration (EMFI) mechanism, leveraging Soft Nearest-neighbor Interpolation (SNI) to ensure signal fidelity and manifold consistency across heterogeneous scales. Extensive quantitative evaluations on specialized benchmarks (RUOD, DUO, and DeepFish) demonstrate that MFE-Net achieves 86.4% mAP50 and 63.9% mAP50-95 on RUOD with a balanced computational load of 12.8 GFLOPs, surpassing contemporary state-of-the-art (SOTA) architectures within the evaluated scope.

Keywords: Underwater object detection, Manifold preservation, Feature annihilation, Hybrid perception, Multi-scale feature integration.

1 Introduction

The collaborative evolution of seafloor observation networks and autonomous underwater vehicles has driven a fundamental paradigm shift in seafloor environmental data collection, leading to an unprecedented growth in the scale of underwater multimedia

data [1,2]. These massive visual materials contain key information on marine biodiversity and seafloor geomorphology, and are high-value digital resources in the field of marine intelligent computing. There is still a significant imbalance between current computing processing capabilities and data generation speed, which leads to a prominent dilemma of "plenty of data but little information" [3]. In such random degradation scenarios, how to bridge the semantic gap and enable static multimedia data to have scientific interpretability remains a core challenge.

From a computational perspective, underwater imaging is a typical ill-posed inverse problem, which mainly stems from the deep coupling between the nonlinear optical properties of water and the morphological distribution of marine organisms [4,5]. Most early studies in this field mitigated these effects through physical and non-physical restoration frameworks, and offset optical degradation problems with dedicated priors [6]. These methods can effectively improve the subjective visual clarity, but the inherent contradiction between signal enhancement and feature retention remains unresolved. Most restoration solutions prioritize perceptual aesthetics at the expense of the integrity of high-frequency structures [7]. Such deviations will deviate from the original feature distribution and may also affect the performance of subsequent machine perception models [8]. Research in the field of intelligent computing is increasingly shifting toward perceptual detectors, which directly process degraded signal streams without a dedicated signal recovery step.

Modern detectors such as the YOLO series have made considerable progress, but they still have obvious limitations: they rely excessively on traditional stride convolution or pooling to achieve spatial downsampling [9]. From the perspective of digital signal processing, such operations are equivalent to irreversible low-pass filters, which prioritize the extraction of global semantic features but sacrifice local high-frequency details [10]. This mechanism often leads to the phenomenon of "feature annihilation". The geometric manifolds and structural integrity of tiny biological entities are easily discarded as random noise. During the abstraction process, systematic loss of information entropy occurs, which causes a mismatch between the internal representations of the model and the actual physical signals of tiny targets [11]. Consequently, we identify this information loss during downsampling as a critical bottleneck limiting the performance ceiling of intelligent perception in stochastically degraded underwater environments.

To surmount these limitations, we formulate MFE-Net, a hierarchical hybrid perception and manifold-preserving framework. Unlike conventional detectors that suffer from irreversible information loss due to standard strided downsampling, MFE-Net is engineered to reconcile the inherent contradiction between noise-robust deep abstraction and the preservation of sparse high-frequency geometric cues. By shifting the focus from external signal restoration or purely abstractive feature extraction to the internal reactivation of salient features, MFE-Net incorporates a Hybrid Perception Feature Extraction (C3CF) module. Distinct from standard backbones that rely on fixed receptive fields, C3CF harmonizes local convolutional inductive biases with global attention recalibration to enhance the signal-to-noise ratio (SNR) in latent representations. Furthermore, where traditional neck architectures incur significant signal dilution during multi-scale fusion, the Multi-Scale Feature Enhancement Neck (MFE-Neck) establishes a

lossless Fine-grained Feature Enhancement Branch and Efficient Multi-Scale Feature Integration (EMFI) to ensure manifold consistency across heterogeneous scales. By optimizing the feature transmission mechanism rather than merely increasing architectural depth, MFE-Net provides a robust and interpretable perception engine. Given its well-balanced computational overhead, it demonstrates significant potential for future real-time edge deployment on power-constrained underwater platforms while fundamentally mitigating the feature annihilation of minute biological targets. The primary contributions of this study are summarized as follows:

- **Hybrid Perception Feature Extraction Architecture (C3CF):** We propose the C3CF module to counteract non-structured scattering noise by integrating local convolutional priors with global attention modeling, enabling the backbone to dynamically recalibrate its receptive field.
- **Multi-Scale Feature Enhancement Neck (MFE-Neck):** We construct a manifold-preserving neck featuring a Fine-grained Feature Enhancement Branch and an EMFI mechanism to maximize the retention of high-frequency geometric cues typically lost during downsampling.
- **Intelligent Computing Benchmarking:** Extensive quantitative evaluations on specialized benchmarks (RUOD, DUO, DeepFish) demonstrate that MFE-Net achieves a competitive balance between computational efficiency and detection fidelity.

2 Related Work

2.1 Paradigm Shift in Underwater Signal Processing and Perception

Underwater visual perception, which is restricted by the complex coupling of nonlinear optical properties and heterogeneous morphological distributions, therefore makes obtaining advanced semantic flows a typical ill-posed inverse problem [12]. Conventional methods are generally categorized into statistic-based enhancement (e.g., CLAHE [13] and Retinex [14]), which redistributes pixel intensities to augment local contrast, and physics-based restoration (e.g., Dark Channel Prior [15]), which inverts the underwater image formation model via atmospheric light and depth estimation. Besides these heuristic priors, the data - driven paradigm based on generative adversarial networks (GANs) has been extensively researched for cross - domain image - to - image translation, utilizing adversarial learning to map the degraded signal distributions to the clear reference spaces for synthesizing pseudo - real situations [16].

However although these generation models enhance subjective visibility they often generate random imperfections or "over-smoothing" of key high-frequency biological textures which may damage the discriminative feature representations needed for downstream machine perception. Therefore current research shifts to end-to-end perception frameworks bypassing explicit repair stages but instead focusing on inherently activating significant features directly from degraded intensities. These architectures, which aim to coordinate the anti - noise characteristics of deep semantic abstraction and the retention of potential geometric manifolds, effectively bridge that "semantic gap" and generate scientifically interpretable quiet underwater multimedia data [17].

2.2 Evolution of Object Detectors and Multi-Scale Integration Bottlenecks

The technical development path of object detection is mainly propelled by deep convolutional neural networks (CNNs), like the YOLO series which has turned into an industry benchmark as it has achieved the optimal balance between inference latency and detection fidelity [9]. As for architectures, from structures with gradient flow optimization such as CSPNet [18] and E-ELAN [19], it has developed to feature extraction layers with efficient parameters in contemporary frameworks like YOLOv11 [20], and at the same time, models based on Transformer are also exploring the potential of global long-distance dependence modeling. In order to adapt to the random characteristics within the underwater domain, specific modifications are put forward, for example, using RG-YOLO [21] to carry out manifold re - activation and using Marine YOLO [22] to implement multi - scale attention so as to strengthen the representation of degraded features.

However, there remains a formidable bottleneck. That is the widespread reliance on traditional strided convolutions or pooling functions, regarding them as irreversible low-pass filters, often resulting in "feature annihilation", during which time the geometric manifolds of tiny entities are often discarded as noise. Moreover, the traditional multi-scale integration through FPN [23] and PANet [24] structures also has serious problems, namely there is semantic misalignment between noisy shallow features and vague deep abstractions. Although contemporary methods like AGS-YOLO [25] optimize fusion weights via adaptive gating, the fundamentally remaining destructive downsampling operators are retained. On the contrary, the MFE-Net we put forward, by bringing in the lossless fine-grained feature enhancement branch and the efficient multi-scale feature fusion (EMFI), gets over these limitations and offers a powerful signal basis for the high-precision target localization in intelligent marine computing.

3 Proposed Method

3.1 Overview

To address the formidable challenges of underwater perception, such as stochastic scale variation and sub-pixel "feature annihilation," this study formulates MFE-Net, a robust intelligent framework predicated on the YOLOv11 [20] architecture, as illustrated in Fig. 1. The system is structured as a multi-stage pipeline comprising a Hierarchical Feature Extraction Backbone, a Multi-Scale Feature Enhancement Neck (MFE-Neck), and specialized detection heads, meticulously balancing representational capacity with computational efficiency. The backbone adopts a five-stage hierarchical topology (B1 through B5), utilizing 3×3 strided convolutions and C3CF modules—specifically designed to suppress non-structured scattering noise and mitigate "semantic confusion" in latent manifolds—to progressively transform stochastically degraded pixel intensities into discriminative, high-level semantic manifolds. By incorporating advanced spatial pooling and attention-driven recalibration, the framework augments the effective receptive field and ensures high-SNR (Signal-to-Noise Ratio) latent representations for downstream intelligent reasoning.

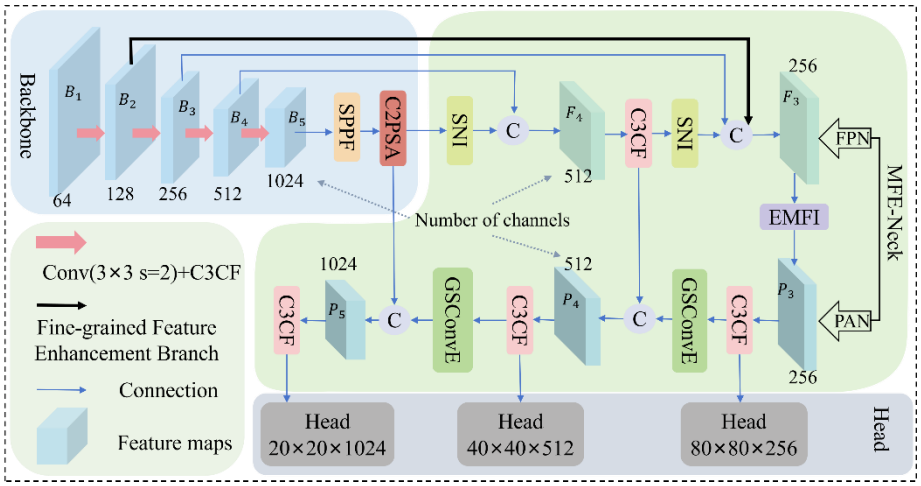

Fig. 1. The overall architecture of MFE-Net: the Backbone extracts hierarchical semantic features, the MFE-Neck performs manifold preservation and multi-scale integration, and the Detection Heads execute final target localization and classification.

As a manifold-preserving engine, the MFE-Neck fuses FPN-based top-down (F3, F4) and PAN-based bottom-up (P3, P4, P5) pathways. To counteract the "feature annihilation" of tiny entities, its Fine-grained Feature Enhancement Branch uses lossless pixel rearrangement to directly connect the B2 layer and F3 node, bypassing the entropy loss of traditional downsampling. To address random scale variations and ensure semantic alignment, it integrates an adaptive weighted sampling scheme, the EMFI module, and signal-preserving convolutions. The rectified multi-scale feature maps are then propagated to native detection heads (20×20, 40×40, 80×80) for precise identification and regression.

3.2 Hybrid Perception Feature Extraction Module

The optical properties inherent in the underwater environment will lead to serious contrast reduction and texture blurring, often making biological targets merge into the complex unstructured background. As for the main components of traditional convolutional neural networks (CNNs), because of their fixed geometric receptive fields and limited ability to suppress scattering noise without global context guidance, it is often rather difficult to extract discriminative feature descriptors in such an environment. To get over these limitations, the devised MFE-Net has the C3CF (hybrid perception feature extraction) module as its main component unit and it is used to substitute the standard C3K2 block in the benchmark. Inspired by the design concept of the SCTNet [26], the C3CF module is developed which can offer powerful feature representations from uneven multimedia streams and ensure that prominent biological clues can be given priority even as the spatial resolution of each layer decreases.

At the macro-architectural level, as illustrated in Fig. 2, the C3CF module inherits the high-efficiency gradient flow of the Cross Stage Partial (CSP) topology. As

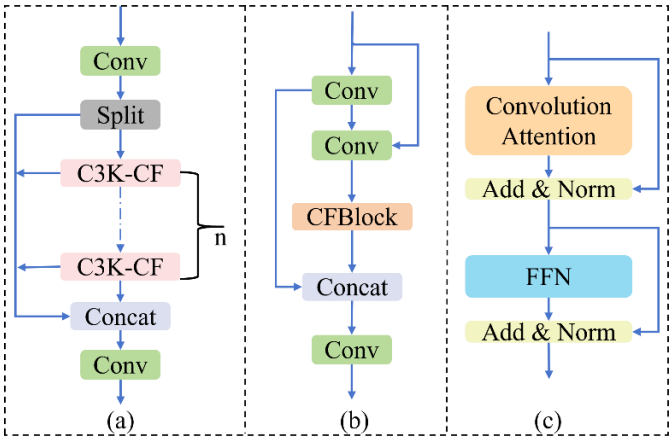

Fig. 2. The structural components of the C3CF module: (a) illustrates the macro-architectural bifurcation for gradient flow efficiency, (b) shows the C3K-CF unit for feature reactivation, and (c) details the CFBlock for capturing semantic context through convolutional attention.

illustrated in Fig. 2(a), the input feature manifold is bifurcated into two parallel computational paths: one stream remains as an identity mapping to preserve raw structural integrity, while the other undergoes intensive transformation via a series of stacked C3K-CF units. Each C3K-CF unit, as illustrated in Fig. 2(b), utilizes a partial residual structure that integrates a convolutional preprocessing stage with the CFBlock to facilitate feature reactivation while ensuring stable gradient propagation across deep networks. This bifurcated design significantly reduces computational redundancy, allowing the backbone to maintain high throughput in real-time underwater observation scenarios.

The micro-computational core of the module is the CFBlock, as illustrated in Fig. 2(c), which is designed as a serial Transformer-like CNN block to capture high-quality semantic context. The computational flow within the block follows a cascaded paradigm defined as:

$$f = Norm(x + ConvAttention(x)) \tag{1}$$

$$y = Norm(f + FFN(f)) \tag{2}$$

where Norm refers to batch normalization, and x, f, y denote the input, hidden feature, and output, respectively.

In the Convolutional Attention Layer, the framework addresses the $O(N^2)$ complexity of standard self-attention by utilizing learnable static convolution kernels as shared "external" Keys (K) and Values (V), while the input is used only to generate the Query (Q). To capture the anisotropic spatial distribution of aquatic organisms, this layer employs a Strip Convolution Decomposition strategy, approximating large-kernel 2D convolutions via cascaded $1 \times k$ and $k \times 1$ operators to achieve linear complexity O(N). The attention weights are formalized as:

$$A = Softmax(DWConv_{1\times k}(Q) \times K + DWConv_{k\times 1}(Q) \times K) \tag{3}$$

Subsequently, the signal enters the Conv-FFN for local refinement. Unlike typical Transformer FFNs that rely on vectorization, this pure convolution-based structure preserves the spatial topological coherence of sub-pixel biological textures and implicitly introduces absolute positional encoding via the Zero Padding Effect, thereby enhancing the model's discriminative power in turbid waters.

3.3 MFE-Neck: High-Fidelity Manifold-Preserving Engine

In the structural design of the advanced intelligent perception framework the neck network serves as a vital hub for hierarchical feature fusion among different heterogeneous manifolds and semantic alignment. However in the context of underwater visual data the traditional neck architecture usually faces a dual computational bottleneck: the irreversible dissipation of high-frequency geometric clues during successive downsampling processes and the serious semantic misalignment between noisy shallow features and vague deep abstractions. To surmount these challenges, this research has devised a multi-scale feature -enhanced neck (MFE-Neck) which is considered as a high - fidelity manifold - preserving engine. Different from the traditional linear summation mode the MFE-Neck adopts an innovative strategy that is composed of three parts: a fine-grained feature enhancement branch which reactivates sub-pixel target clues by making use of a lossless pixel rearrangement strategy an efficient multi-scale feature integration (EMFI) module which realizes full-scale perception and an optimized set of sampling operators which keep the change of signal strength.

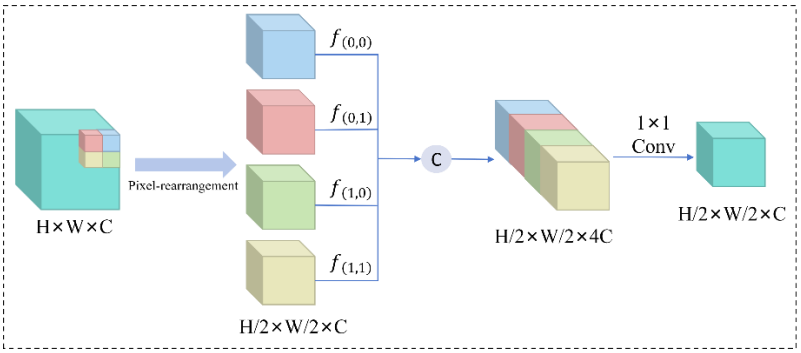

Fig. 3. The mechanism of the Fine-grained Feature Enhancement Branch: SPD-Conv transforms spatial resolution into channel depth via deterministic pixel-rearrangement to ensure lossless feature transmission.

Fine-grained Feature Enhancement Branch.
In order to make sure that the sparse edge and texture clues needed for detecting tiny biological entities (such as juvenile fish or sub-pixel plankton) can be kept, a special fine-grained feature enhancement branch is designed in the MFE-Neck in this research. In the standard detection mode, shallow feature maps such as the B2 layer are usually either discarded or compressed a lot through stride convolution to cut down the

computational cost. However, from the perspective of digital signal processing, such strided convolutions are just like irreversible low-pass filters, truly having the high-frequency geometric components of small targets eliminated indeed. In order to fundamentally prevent such information entropy loss and retain the distinct manifold, this branch that is designed by us bypasses the destructive downsampling by making use of the spatial to depth convolution (SPD-convolution) module [27].

To replace the destructive weighted summation of standard convolutions with a deterministic pixel-rearrangement strategy, the SPD layer is integrated as the core operator of this branch, as illustrated in Fig. 3. Given an input manifold X of dimensions H × W × C, the SPD transformation slices the spatial grid into a series of sub-feature maps by sampling pixels at alternating indices. To achieve a downsampling factor of 2 while ensuring all spatial information is preserved, the process generates four sub-maps $f_{(0,0)}, f_{(0,1)}, f_{(1,0)}, f_{(1,1)}$, which are subsequently concatenated along the channel dimension. This transformation yields an intermediate representation of size H/2 × W/2 × 4C, defined mathematically as:

$$f_{i,j} = X[i:H:2, j:W:2, :], \ i,j \in \{0,1\} \tag{4}$$

$$X_{SPD} = Concat(f_{0,0}, f_{1,0}, f_{0,1}, f_{1,1}) \tag{5}$$

To align the channel capacity with the target dimension while strictly preserving the refined spatial manifolds, a non-strided 1 × 1 convolution is applied following this lossless spatial reduction. Unlike conventional strided downsampling which discards 75% of the spatial signal in a single step, this reversible mapping ensures that even single-pixel texture variations are fully encoded into the channel depth. By implementing this mechanism, MFE-Net can access undiluted raw high-frequency information during deep-layer inference, significantly enhancing the localization precision for dense, minute biological targets in turbid underwater multimedia streams.

Efficient Multi-Scale Feature Integration.

To address the extreme scale variation in underwater visual scenes—ranging from large-scale reef backgrounds to sub-pixel plankton—the EMFI (Efficient Multi-Scale Feature Integration) Module is integrated as the core multi-scale perception engine of the neck. Drawing inspiration from the omni-scale modeling philosophy of Omni-Kernel Network [28], the EMFI module adopts a Cross Stage Partial (CSP) topology with an asymmetric channel splitting strategy to optimize feature representation within a limited computational budget. As illustrated in Fig. 4(a), the input feature map is split along the channel dimension at a ratio of 1:3; the majority of channels (75%) pass directly through an identity mapping path as passive flow to maintain gradient stability and minimize memory overhead, while the remaining 25% of core channels serve as the active flow routed into a tripartite parallel processing unit for deep multi-scale transformation.

The EMFI core computation unit simulates the adaptive zoom mechanism of human vision through three parallel branches corresponding to local, meso-scale, and global perception scales. The Local Branch consists of a 1 × 1 depth-wise convolution

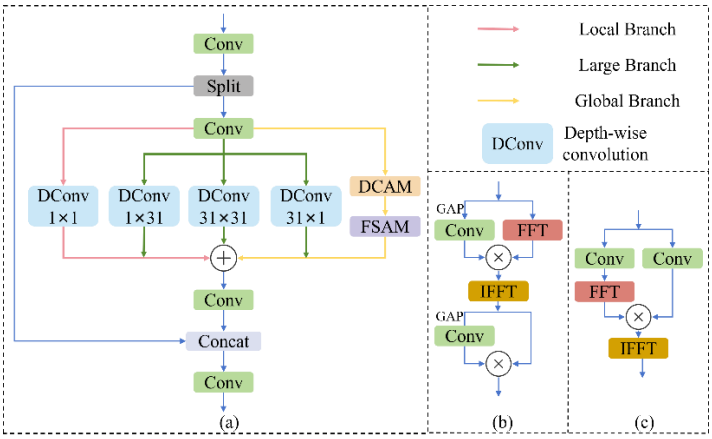

Fig. 4. The structural components of the EMFI module: (a) illustrates the CSP-based asymmetric channel bifurcation, (b) shows the DCAM for spatial feature recalibration, and (c) details the FSAM for frequency-domain noise suppression.

(DWConv) to extract high-frequency pixel-level differences, effectively preserving the edge sharpness of tiny organisms and preventing them from being assimilated by background noise. The Large-Kernel Branch aims to capture meso-to-long-range spatial context dependencies via a Large Kernel Decomposition strategy, mathematically decoupling the 2D receptive field into two cascaded 1D convolutions: a $1 \times K$ horizontal strip and a $K \times 1$ vertical strip (where $K = 31$). This decomposition reduces computational complexity from a quadratic $O(K^2)$ level to a linear $O(2K)$ level, while the resulting cross-shaped receptive field aligns with the natural swimming postures of marine organisms.

In order to suppress the all-pervasive global interference in underwater images, the third branch which is the global branch integrates a dual-domain attention mechanism to calibrate features from spatial and frequency dimensions. This branch is composed of a dual-domain channel attention (DCAM) for spatial context aggregation and a frequency selection attention (FSAM) for spectral analysis. As illustrated in Fig. 4(b) and Fig. 4(c), the FSAM utilizes the Fast Fourier Transform (FFT) to project feature maps into the frequency domain, incorporating a learnable gating mechanism to adaptively suppress spectral responses representing environmental interference, such as turbidity and noise. The global features obtained through the inverse fast Fourier transform (IFFT) are combined with the local and large branches through an element-wise addition method, and then connected with the identity mapping features finally, so as to make sure that the final output contains abundant multi-scale semantics while keeping the original low-level details.

Optimized Sampling Operators and Signal Fidelity.
The performance degradation in multi-level feature fusion often stems from edge aliasing and signal dilution induced by conventional sampling operators. In underwater scenarios with low contrast, standard nearest-neighbor or bilinear interpolation can lead to

pixel-wise semantic shifts, which complicates the detector's coordinate regression. To mitigate these effects, MFE-Net reconfigures its sampling pipeline by incorporating Soft Nearest-neighbor Interpolation (SNI) and the GSConvE operator [29].

The SNI mechanism replaces the discrete "hard" operation of traditional interpolation with an adaptive weighting scheme. This approach regulates the influence of semantic features during upsampling by calculating the spatial dimension ratio between the input feature manifold F_{in} and the output feature map F_{out}:

$$F_{out} = \eta \cdot \Psi(F_{in}), \quad \eta = \frac{D_{in}}{D_{out}} \tag{6}$$

In this context, $\Psi(\cdot)$ denotes the fundamental nearest-neighbor sampling function, and η serves as a weight-downscaling factor that adaptively balances feature alignment. This mechanism effectively buffers geometric shocks during hierarchical fusion and preserves critical structural textures in degraded images.

In addition to that, for the downsampling path, this framework makes use of the GSConvE structure which optimizes the standard convolution and eliminates the batch normalization (BN) layer right after the depthwise convolution operation. Such a strategic design can stop the BN layer from over-normalizing low-variance signals, thereby maintaining the intensity variance needed to distinguish fine microbial textures in the turbid background. By means of the collaboration of SNI and GSConvE to maintain higher signal fidelity, MFE-Neck develops a robust feature transmission mechanism that guarantees high-precision localization all through the feature pyramid.

4 Experiments and Result Analysis

4.1 Experiment Setup

Implementation Details. The proposed model is implemented with the PyTorch deep learning framework with CUDA 12.4 acceleration, which is trained and tested on a high-performance Linux server cluster equipped with an NVIDIA GeForce GTX 1080 Ti GPU. The batch size is set to 16, and the input imagery was standardized to a 640 × 640 resolution. The total training budget was set to 600 epochs to ensure sufficient convergence of the hybrid perception modules under complex data distributions. We employ the Stochastic Gradient Descent (SGD) optimizer to iteratively update the model parameter weights, with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005.

Datasets. Our MFE-Net's effectiveness and computational robustness in random degradation environments are evaluated by us by using the RUOD, DUO, and DeepFish datasets. The main evaluation basis is the real underwater target detection (RUOD) dataset, which is a large-scale repository containing 14,000 high-definition images divided into a training set (9,800 images) and a test set (4,200 images). RUOD that is composed of ten different biological categories including various fish corals and sea urchins has the features of severe uneven blurriness color distortion and light interference. In order to further verify the cross - domain adaptability between different water turbidities and benthic habitats, we make use of the DUO dataset (which has 7,782

images) and the DeepFish dataset (which is approximately 40,000 images), the latter of which is specifically optimized for fine - grained species identification in tropical environments. Furthermore, in order to demonstrate that the improvement of our architecture has the general feature representation ability beyond the marine scope, we utilize the PASCAL VOC 07+12 and MS COCO 2017 datasets for land-based verification.

Evaluation Metrics. We adopt standard metrics to quantify performance: Mean Average Precision (mAP50) for foundational detection capability and mAP50-95 for high-precision localization fidelity. Additionally, parameter count (Params) and floating-point operations (FLOPs) were meticulously calculated. These metrics serve as theoretical proxies to evaluate the trade-off between model representational capacity and its real-time inference potential on edge computing devices.

4.2 Comparative Analysis on Marine Intelligent Perception

Table 1. Quantitative comparison on the RUOD dataset. Params denotes parameter count (M), and FLOPs represents floating-point operations (G).

Model	mAP50(%)	mAP50-95(%)	Params(M)	FLOPs(G)
YOLOv5n	84.1	59.7	2.2	5.9
YOLOv7-tiny [19]	85.0	57.9	6.0	13.0
YOLOv8n	84.3	60.5	2.7	6.9
YOLOv10n [30]	84.3	61.1	2.7	8.3
YOLOv11n	85.0	62.1	2.6	6.3
YOLOv12n [31]	83.9	60.2	2.6	6.4
MarineYOLO [22]	80.5	44.6	20.6	10.9
RG-YOLO [21]	85.7	60.6	10.1	31.1
DyFish-DETR [32]	83.2	57.9	23.7	61
CDM-YOLO [33]	83.9	58.6	4.4	17.7
MSFE-YOLO [34]	84.8	61.3	3.1	12.3
GAI-YOLO [35]	85.5	62.1	2.8	7.7
AGS-YOLO [25]	86.4	63.5	3.0	9.6
MFE-Net	**86.4**	**63.9**	**3.6**	**12.8**

To rigorously evaluate the robustness of MFE-Net, we conducted a systematic benchmark comparison against mainstream real-time detectors on the RUOD dataset. The experimental results, as illustrated in Table 1, reveal a theoretically significant phenomenon: although the state-of-the-art YOLOv12 (2025) innovatively introduces Area Attention and FlashAttention mechanisms to model global dependencies, its mAP50 on RUOD is only 83.9%, which is notably inferior to the convolution-based YOLOv11n (85.0%) and the proposed MFE-Net (86.4%). This performance divergence can be deeply attributed to the perspective of inductive bias. Pure attention mechanisms frequently lack the local translation invariance and neighborhood-focusing priors inherent in convolutional operators when processing unstructured underwater data. The suspended particle noise and multi-path scattering prevalent in marine imagery constitute high-frequency interference that easily disrupts the construction of global attention

maps, leading to erroneous correlations between background noise and semantic features. In contrast, MFE-Net organically integrates local convolutional inductive bias with global dynamic weights via the C3CF module, effectively suppressing the propagation of environmental artifacts at the structural level and thereby achieving significant superiority over generic terrestrial baselines.

Table 2. Quantitative comparison on the DUO and DeepFish datasets.

	Model	mAP50	mAP50-95	Params(M)	FLOPs(G)
DUO	YOLOv5n	83.6	64.1	2.2	5.9
	YOLOv7-tiny [19]	85.0	64.7	6.0	13.0
	YOLOv8n	84.6	65.3	2.7	6.9
	YOLOv11n	83.4	64.8	2.6	6.3
	RT-DETR [36]	80.2	58.1	32.8	108.0
	AGS-YOLO [25]	85.5	67.7	3.0	9.6
	MFE-Net	**85.5**	**67.8**	**3.6**	**12.8**
DeepFish	YOLOv11n	97.8	66.8	2.6	6.3
	YOLORG-N	98.0	67.0	3.0	12.4
	MFE-Net	**98.2**	**68.0**	**3.6**	**12.8**

Comparisons with domain-specific detectors further highlight MFE-Net's superiority. Unlike CDM-YOLOv8n (inefficient parameter allocation: 17.7 GFLOPs, 83.9% mAP50) and the lightweight GAI-YOLO (accuracy constrained by excessive compression), MFE-Net achieves an optimal balance. While matching the top baseline, AGS-YOLO, in mAP50 (86.4%), MFE-Net leads by 0.4% in the stricter mAP50-95 metric (63.9% vs. 63.5%). This high-precision advantage validates that our SPD-Conv branch and SNI operator effectively mitigate multi-scale feature annihilation. Furthermore, MFE-Net's remarkable 98.2% mAP on DeepFish (Table 2) confirms that preserving signal manifolds is crucial for complex underwater vision.

4.3 Generalization Experiments Across Terrestrial Manifolds

Table 3. Generalization performance on PASCAL VOC and MS COCO datasets. Comparisons include the baseline and the latest 2025 state-of-the-art models.

	Model	mAP50	mAP50-95	Params(M)	FLOPs(G)
VOC	Gold-YOLO [38]	79.8	58.4	5.6	12.1
	YOLOv10n [30]	79.9	59.9	2.3	6.7
	YOLOv11n	80.5	60.5	2.6	6.3
	MFE-Net	**82.3**	**62.9**	**3.6**	**12.8**
COCO	Gold-YOLO [38]	55.7	39.6	5.6	12.1
	YOLOv10n	53.8	38.5	2.3	6.7
	YOLOv11n	54.2	38.6	2.6	6.5
	YOLOv12n [31]	56.0	40.1	2.6	6.5
	YOLOv13n [39]	57.8	41.6	2.5	6.4
	MFE-Net	**58.3**	**42.2**	**3.6**	**12.8**

To assess MFE-Net's universal representation, we evaluated it on PASCAL VOC 07+12 and MS COCO 2017, as illustrated in Table 3. On VOC, MFE-Net achieved 82.3% mAP50 and 62.9% mAP50-95, exceeding Gold-YOLO (79.8%/58.4%) and YOLOv11n (80.5%/60.5%). This demonstrates the C3CF module's capacity to capture complex geometric structures across domains. Unlike existing methods that often rely on fixed convolutional receptive fields—which tend to over-fit to domain-specific textures—or pure attention mechanisms that lack local structural consistency, the C3CF module harmonizes local convolutional priors with global dynamic gating. This hybrid approach enables the backbone to extract "shape-aware" features that represent the underlying geometric manifold of an object rather than its pixel-level distribution. By prioritizing these domain-agnostic geometric cues, MFE-Net effectively suppresses domain-specific noise and maintains a robust feature descriptor even when transitioning from aquatic to complex terrestrial environments, thereby significantly enhancing cross-domain generalization.

On MS COCO, MFE-Net reached 58.3% mAP50 and 42.2% mAP50-95, surpassing contemporary models such as YOLOv13 (57.8%/41.6%). These outcomes confirm that for small-object detection, resolving "feature annihilation" through the Fine-grained Feature Enhancement Branch and SNI operators is more effective than modeling high-order correlations alone. The performance margin on MS COCO suggests that existing models frequently suffer from "geometric drift" in diverse terrestrial scenes where target scales vary drastically. By ensuring the integrity of sub-pixel geometric components, MFE-Net exhibits superior structural resilience, proving that manifold preservation is a universal prerequisite for high-fidelity perception across both terrestrial and marine environments.

4.4 Ablation Studies and Architectural Synergy

To evaluate the contribution of each innovation, we conducted ablation studies on the RUOD dataset, as illustrated in Table 4, using YOLOv11n as the baseline (85.0% mAP50, 62.1% mAP50-95). Integrating the C3CF backbone alone improved mAP50 to 85.9%, confirming its capacity to filter noise from blurred textures. However, the modest mAP50-95 gain (+0.5%) suggests that while the backbone excels at target detection, high-precision localization requires further structural support.

Table 4. Ablation studies of the proposed modules on the RUOD dataset.

Model	Modules		mAP50	mAP50-95	FLOPs(G)
	C3CF	MFE-Neck			
Baseline	✗	✗	85.0	62.1	6.3
Variant A	✓	✗	85.9	62.6	7.9
Variant B	✗	✓	85.8	62.9	11.2
MFE-Net	✓	✓	**86.4**	**63.9**	**12.8**

MFE-Neck (Variant B) has significantly increased mAP50-95 to 62.9%, which emphasizes the significance of the fine-grained feature enhancement branch and also optimizes sampling alignment to re-activate the geometric manifold that is usually lost

during downsampling. The MFE- Net that has achieved 86 points. With an average precision mean of 4% at 50 (mAP50) and an average precision mean of 63.9% from 50 to 95 (mAP50 - 95), an extremely remarkable synergistic effect emerges: the backbone part provides a basis of high signal-to-noise ratio, and meanwhile the neck ensures a transmission that is lossless and precisely aligned. With an operating speed of 12.8 GFLOPs, the MFE-Net still keeps a relatively lightweight computational occupancy. The theoretical efficiency shows that this architecture is extremely beneficial to real - time edge computing on underwater platforms, having established a solid algorithmic basis for hardware deployment.

4.5 Qualitative Analysis and Visual Interpretability

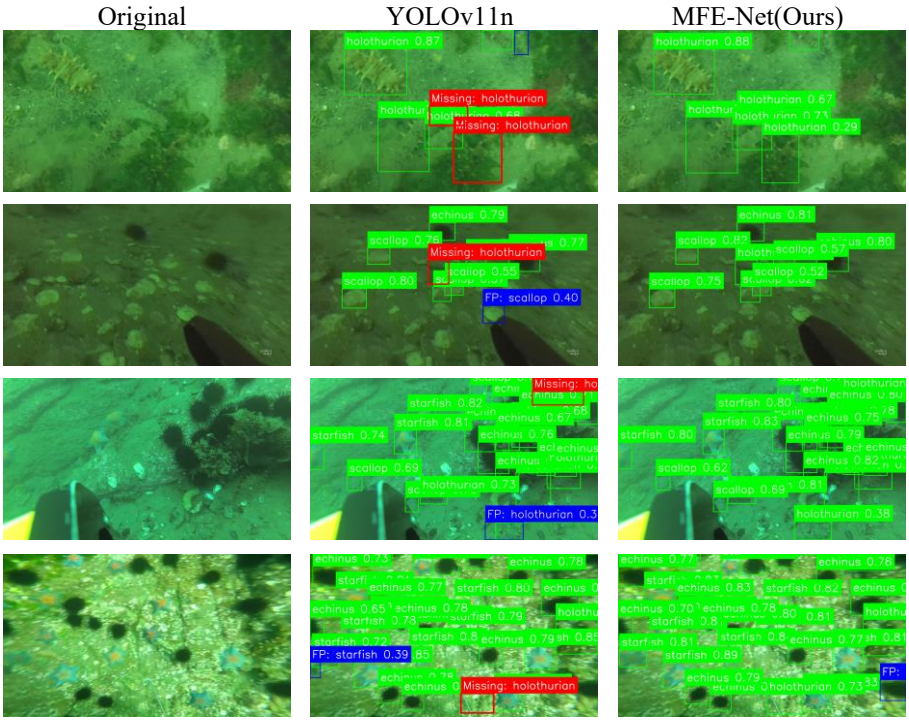

Fig. 5. Qualitative comparison of detection performance on the RUOD dataset. From left to right: Original images, YOLOv11n results, and MFE-Net (Ours). Green boxes represent successful detections, while red and blue highlights in the baseline column denote false positives and missed targets, respectively.

To qualitatively evaluate the perceptual superiority of MFE-Net, we conducted a comparative visual analysis against the baseline YOLOv11n, as illustrated in Fig. 5. In the first and second rows, where biological entities (e.g., holothurians and scallops) are small and highly integrated into the background, the baseline model exhibits significant misses and false positives. Conversely, MFE-Net successfully identifies these tiny

targets with high confidence. This visual success is primarily attributed to the Fine-grained Feature Enhancement Branch, which preserves sub-pixel geometric clues typically lost in standard architectures. Moreover, in the complex and cluttered environments of the third and fourth rows, MFE-Net maintains robust localization, demonstrating that the EMFI module effectively discriminates between foreground organisms and turbid artifacts.

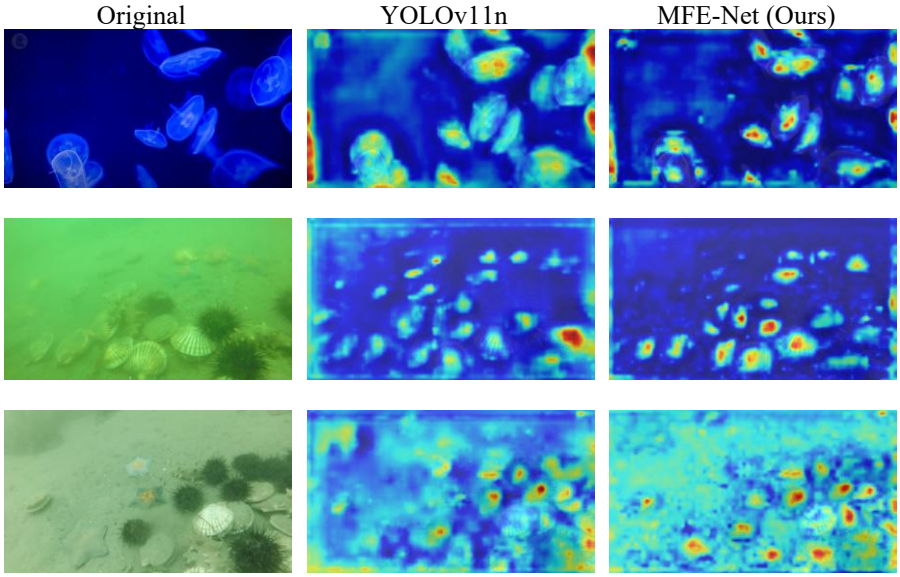

Fig. 6. Feature activation visualization via heatmaps. The comparison demonstrates that MFE-Net exhibits more precise and concentrated attention on target entities compared to the baseline, effectively suppressing background interference through spectral and spatial calibration.

In order to further explore the decision-making logic inside the model, the saliency of different architectures is visualized through feature activation mapping, as illustrated in Fig. 6. Compared with the baseline, the baseline usually presents distracted or misaligned attention - especially when there are scattered lights and suspended particles. And the heatmap generated by MFE-Net is more concentrated and precisely centered on the target manifold, just like the first row of translucent jellyfishes and the second row of benthic sea urchins. The high-intensity activated regions which closely match the morphological boundaries of organisms show that the C3CF framework has successfully filtered out random noise. This enhanced focusing confirms that our framework not only attains excellent quantitative indicators but also has high visual interpretability and focuses on the "correct" semantic features even in randomly degraded environments.

5 Conclusion

In this study, we propose MFE-Net, a robust intelligent perception framework designed to address the challenges of "feature annihilation" and "semantic confusion" in stochastically degraded underwater environments. By incorporating the Hybrid Perception Feature Extraction (C3CF) Module, the framework effectively integrates local inductive biases with global context modeling, enabling the backbone to suppress non-structured scattering noise. Furthermore, the MFE-Neck re-engineers the feature fusion pipeline through a Fine-grained Feature Enhancement Branch and Efficient Multi-Scale Feature Integration (EMFI), ensuring the preservation of high-frequency geometric manifolds and precise semantic alignment. Quantitative evaluations on the RUOD dataset demonstrate that MFE-Net achieves a favorable balance between detection fidelity and computational efficiency. Additionally, MFE-Net exhibits superior performance over several SOTA methods, particularly in high-precision localization tasks. Generalization experiments on terrestrial manifolds further confirm the structural resilience of the proposed modules across diverse visual domains.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62402062, Hunan Provincial Key Research and Development Program under Grant 2024AQ2027, 2025AQ2022, Natural Science Foundation of Hunan Province, China under Grant 2025JJ60415, Research Foundation of Education Bureau of Hunan Province, China under Grant 25B0221.

References

1. Banesh D, Petersen M R, Ahrens J, et al. An image-based framework for ocean feature detection and analysis[J]. Journal of geovisualization and spatial analysis, 2021, 5(2): 17.
2. Sonnewald M, Lguensat R, Jones D C, et al. Bridging observations, theory and numerical simulation of the ocean using machine learning[J]. Environmental Research Letters, 2021, 16(7): 073008.
3. Wilding T A, Gill A B, Boon A, et al. Turning off the DRIP ('Data-rich, information-poor')–rationalising monitoring with a focus on marine renewable energy developments and the benthos[J]. Renewable and Sustainable Energy Reviews, 2017, 74: 848-859.
4. Wang Y, Song W, Fortino G, et al. An experimental-based review of image enhancement and image restoration methods for underwater imaging[J]. IEEE access, 2019, 7: 140233-140251.
5. Yang M, Hu J, Li C, et al. An in-depth survey of underwater image enhancement and restoration[J]. IEEE access, 2019, 7: 123638-123657.
6. Song W, Liu Y, Huang D, et al. From shallow sea to deep sea: research progress in underwater image restoration[J]. Frontiers in Marine Science, 2023, 10: 1163831.
7. Gao H, Dang D. Exploring richer and more accurate information via frequency selection for image restoration[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 35(3): 2689-2700.
8. Bayram F, Ahmed B S. A domain-region based evaluation of ML performance robustness to covariate shift[J]. Neural Computing and Applications, 2023, 35(24): 17555-17577.

9. Diwan T, Anirudh G, Tembhurne J V. Object detection using YOLO: challenges, architectural successors, datasets and applications[J]. multimedia Tools and Applications, 2023, 82(6): 9243-9275.
10. Zhang R. Making convolutional networks shift-invariant again[C]//International conference on machine learning. PMLR, 2019: 7324-7334.
11. Meni M J, White R T, Mayo M L, et al. Entropy-based guidance of deep neural networks for accelerated convergence and improved performance[J]. Information Sciences, 2024, 681: 121239.
12. Li, S., Zhang, Z., Zhang, Q., Yao, H., Li, X., Mi, J., Wang, H.: Breakthrough underwater physical environment limitations on optical information representations: an overview and suggestions. Journal of Marine Science and Engineering 12(7), 1055 (2024)
13. Musa P, Al Rafi F, Lamsani M. A Review: Contrast-Limited Adaptive Histogram Equalization (CLAHE) methods to help the application of face recognition[C]//2018 third international conference on informatics and computing (ICIC). IEEE, 2018: 1-6.
14. Petro A B, Sbert C, Morel J M. Multiscale retinex[J]. Image processing on line, 2014: 71-88.
15. He K, Sun J, Tang X. Single image haze removal using dark channel prior[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(12): 2341-2353.
16. Guo Y, Li H, Zhuang P. Underwater image enhancement using a multiscale dense generative adversarial network[J]. IEEE Journal of Oceanic Engineering, 2019, 45(3): 862-870.
17. Xue X, Yuan J, Ma T, et al. Degradation-Decoupled and semantic-aggregated cross-space fusion for underwater image enhancement[J]. Information Fusion, 2025, 118: 102927.
18. Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.
19. 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[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.
20. Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arXiv preprint arXiv:2410.17725, 2024.
21. Zheng Z, Yu W. RG-YOLO: multi-scale feature learning for underwater target detection[J]. Multimedia systems, 2025, 31(1): 26.
22. Liu L, Chu C, Chen C, et al. MarineYOLO: Innovative deep learning method for small target detection in underwater environments[J]. Alexandria Engineering Journal, 2024, 104: 423-433.
23. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
24. Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.
25. Sun W, Liu X, Hao J, et al. AGS-YOLO: An Efficient Underwater Small-Object Detection Network for Low-Resource Environments[J]. Journal of Marine Science and Engineering, 2025, 13(8): 1465.
26. Xu Z, Wu D, Yu C, et al. Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(6): 6378-6386.
27. Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects[C]//Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland, 2022: 443-459.

28. Cui Y, Ren W, Knoll A. Omni-kernel network for image restoration[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(2): 1426-1434.
29. Li H. Rethinking features-fused-pyramid-neck for object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 74-90.
30. Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in neural information processing systems, 2024, 37: 107984-108011.
31. Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors[J]. Advances in neural information processing systems, 2026, 38: 78433-78457.
32. Wang Z, Ruan Z, Chen C. DyFish-DETR: Underwater fish image recognition based on detection transformer[J]. Journal of Marine Science and Engineering, 2024, 12(6): 864.
33. Luo L, Gao Q, Liu Q, et al. Multi-scale adaptive small and overlapping target detection in underwater images[J]. The Journal of Supercomputing, 2025, 81(16): 1474.
34. Li M, Liu W, Shao C, et al. Multi-scale feature enhancement method for underwater object detection[J]. Symmetry, 2025, 17(1): 63.
35. Guo Y, Zhao X, Zhong Z, et al. GAI-YOLO: an underwater target detection algorithm based on enhanced feature representation and adaptive receptive field: Y. Guo et al[J]. The Journal of Supercomputing, 2025, 81(16): 1516.
36. Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.
37. Zhang D, Gao S, Deng B, et al. Fish detection based on Gather-and-Distribute mechanism Multi-scale feature fusion network and Structural Re-parameterization method[J]. Journal of Hydroinformatics, 2024, 26(5): 1234-1250.
38. Wang C, He W, Nie Y, et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism[J]. Advances in neural information processing systems, 2023, 36: 51094-51112.
39. Lei M, Li S, Wu Y, et al. Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception[J]. arXiv preprint arXiv:2506.17733, 2025.