International Journal of Science and Engineering Applications
Volume 15-Issue 01, 07 – 15, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1501.1002

Dynamic Multi-Scale Perception Network for Underwater Object Detection

Yitian Li*1st
School of Artificial Intelligence
Hubei University
Wuhan, China

Yuzhang Chen
School of Artificial Intelligence
Hubei University
Wuhan, China

Zihao Wang
School of Artificial Intelligence
Hubei University
Wuhan, China

Dehua Zhong
School of Naval Architecture and Ocean Transportation
Guangdong Ocean University
Zhanjiang, China

Abstract: Underwater optical imaging suffers from severe degradation due to light absorption, scattering, and suspended particles, leading to low contrast, blurred details, and color distortion, which significantly impairs the accuracy of object detection. To address these challenges, we propose a lightweight yet powerful Dynamic Multi-Scale Perception Network (DMSPNet) for underwater object detection. DMSPNet integrates a Spectral-Guided Feature Extraction Backbone and a Task-Aligned Decomposition Detection Head, enabling coordinated optimization between multi-scale feature representation and detection task alignment. The backbone employs ReefBlock_SFS and AquaMamba_SFS blocks to enhance feature discriminability under degraded conditions. The detection head introduces a task decomposition mechanism with temperature-gated classification and offset-limited regression, improving detection precision for underwater targets. Extensive experiments on the URPC2019 dataset demonstrate that DMSPNet achieves superior performance with 78.2% mAP@0.5 and 50.3% mAP@0.5:0.95, outperforming several baseline detectors while maintaining a compact parameter size of only 2.74M. Ablation studies further validate the effectiveness of each proposed component. The network demonstrates remarkable efficiency and robustness in complex underwater environments, offering a practical solution for real-time underwater perception systems.

Keywords: Underwater Object Detection, Spectral-Spatial Convolution, Mamba, Task-Aligned Decomposition, Dynamic Perception, Marine Vision

1. Introduction

Accurate detection of underwater targets is pivotal for marine exploration, ecological monitoring, underwater robotics, and resource development [1]. With the advancement of Autonomous Underwater Vehicles (AUVs) and Remotely Operated Vehicles (ROVs), vision-based detection systems have become indispensable for tasks such as species identification, wreck localization, and pipeline inspection [2]. However, underwater imagery presents unique challenges distinct from terrestrial computer vision applications. The underwater environment introduces complex optical degradation mechanisms [3]. Light absorption varies across wavelengths, with red light attenuated most rapidly, causing significant color distortion. Scattering effects from suspended particles create haze-like artifacts, reducing contrast and blurring object boundaries. Non-uniform illumination, often from artificial sources, creates challenging lighting conditions with strong shadows and highlights. These factors collectively result in low visibility, blurred edges, and diminished color fidelity, critically hindering the performance of conventional object detectors typically designed for clear terrestrial scenes [4].

Recent deep learning-based detectors have revolutionized object detection in terrestrial environments. The YOLO series [5, 6, 7] pioneered real-time detection with its one-stage architecture, balancing speed and accuracy. Transformers, as introduced in DETR [8], brought attention mechanisms to detection, enabling better modeling of global context. However, these architectures show limited adaptability to underwater conditions due to the pronounced domain gap caused by unique underwater degradations. Specialized approaches for underwater detection have emerged along several directions. Enhancement-based methods [9, 10] preprocess images to improve quality before detection, but often introduce artifacts that mislead detection networks. Domain adaptation techniques [11, 12] attempt to bridge the gap between clear and underwater domains but require paired data or complex training schemes. Architecture-specific modifications [13, 14] tailor network components for underwater characteristics but typically focus on single aspects of the problem. Our approach differs by providing an integrated solution that simultaneously addresses feature degradation, multi-scale fusion, and task alignment within a unified framework.Mamba-based Long-Range Context Modeling: We integrate selective state-space models (Mamba) within our backbone to capture long-range spatial dependencies. This enables effective modeling of global scene context—crucial for understanding the complex spatial relationships in underwater environments—while maintaining computational efficiency through gated mechanisms.

This paper introduces the Dynamic Multi-Scale Perception Network (DMSPNet), a comprehensive solution for underwater object detection with the following key contributions: Spectral-Guided Feature Extraction Backbone: We propose a novel backbone architecture that integrates ReefBlock_SFS with explicit high-frequency enhancement and AquaMamba_SFS for long-range context modeling. This component actively restores edge information and fine textures degraded by underwater scattering while capturing global dependencies. HydroAttn Multi-Scale Fusion Neck: We design a lightweight attention module that captures contextual dependencies along horizontal and vertical axes using 1D depthwise convolutions. This efficient attention mechanism enhances salient features while suppressing noise across different scales, improving detection of faint underwater targets. AbyssDetect Task-Aligned Decomposition Head: We introduce a detection head that explicitly decomposes classification and regression tasks using separate pathways with temperature-gated confidence smoothing and bounded offset regression. This addresses the inherent conflict between classification confidence and localization accuracy in underwater conditions. Comprehensive Evaluation: We conduct extensive experiments on the URPC2019 benchmark, demonstrating state-of-the-art performance. Detailed ablation studies validate the contribution of each component, and visual analysis provides insights into the network's behavior in challenging underwater scenarios.

2. METHOD

2.1 Overall Architecture

The overall architecture of DMSPNet follows a single-stage detection paradigm, comprising three principal components: a Spectral-Guided Backbone, a Multi-Scale Feature Fusion Neck, and a Task-Aligned Detection Head, as illustrated in Figure 1.

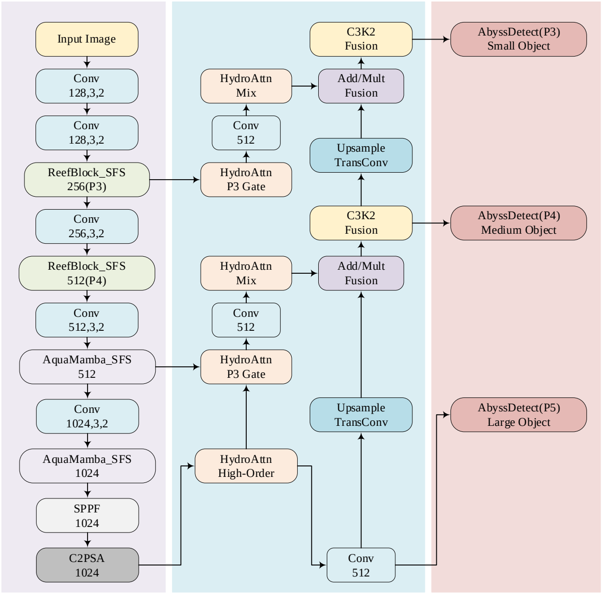

Figure 1 DMSPNet Overall Architecture

The network processes underwater images through: (a) Spectral-Guided Backbone with ReefBlock_SFS and AquaMamba_SFS blocks for enhanced feature extraction; (b) Multi-Scale Fusion Neck with HydroAttn modules for contextual feature integration; (c) Task-Aligned Detection Head with AbyssDetect for explicit decomposition of classification and regression tasks.

The network ingests an underwater image $I \in \mathbb{R}^{3\times H\times W}$ and outputs bounding boxes alongside class probabilities for identified targets. The backbone is constructed upon a modified YOLOv11-nano framework [7], where standard convolutional and C3k blocks are substituted with specialized modules engineered for underwater conditions. The neck incorporates HydroAttn modules for adaptive cross-scale feature fusion. The head is our novel AbyssDetect, which performs dynamic, task-aligned feature decomposition and prediction.

2.2 Spectral-Guided Backbone with AquaMamba Enhancement

2.2.1 *Problem Formulation*

Underwater image degradation can be modeled as:

$$I_{observed} = J\cdot t + A\cdot(1-t) + \eta \quad (1)$$

where J is the ideal scene radiance, t is the transmission map representing light attenuation, A is the ambient light, and η represents additive noise [3]. This degradation particularly affects high-frequency components crucial for object boundaries and fine details.

2.2.2 *Spectral-Spatial Feature Enhancement*

To mitigate high-frequency loss, we introduce Spectral-Spatial Convolution (SFS_Conv), which processes features with explicit frequency awareness. Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, SFS_Conv operates as:

$$\begin{aligned} SFS\_Conv(X) &= DWConv(GroupNorm(X)) \\ &+\alpha\cdot Lap(DWConv(GroupNorm(X))) \end{aligned} \quad (2)$$

where DWConv denotes depthwise convolution, GroupNorm provides channel-wise normalization, and Lap represents a Laplacian operator. The scaling factor α is learned during training, allowing adaptive enhancement strength determination for varying underwater conditions.

The Laplacian kernel L is defined as:

$$L = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

This kernel acts as a high-pass filter, emphasizing regions of rapid intensity change (edges).

2.2.3 *ReefBlock_SFS Module*

The ReefBlock_SFS module builds upon the standard bottleneck architecture, replacing conventional convolutions with SFS_Conv operations. The module enhances local feature representations by processing features through two sequential SFS_Conv layers with residual connections. The first SFS_Conv reduces channel dimensionality by a factor e , while the second restores the original channel count. This design enables efficient computation while preserving critical spatial information.

Mathematically, the transformation can be expressed as:

$$\begin{aligned} X_{hidden} &= SFS\_Conv_1(X_{in}) \\ X_{out} &= SFS\_Conv_2(X_{hidden}) + X_{in}(if \quad shortcut \quad enabled) \end{aligned} \quad (3)$$

The detailed structure of the AquaMamba_SFS block is illustrated in Figure 2. It visualizes the three parallel pathways—Gating, Identity, and Spectral-Spatial—along with the key operations of gated fusion, high-frequency enhancement via Laplacian operator, and the residual connection, providing a clear blueprint of its long-range context modeling mechanism.

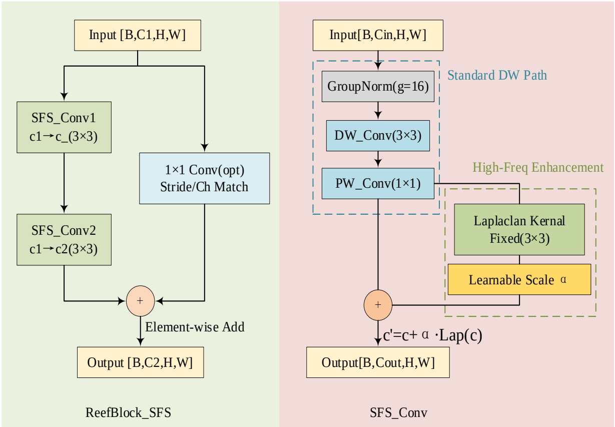

Figure 2 Network diagram of ReefBlock_SFS & SFS_Conv

2.2.4 AquaMamba_SFS Module

For capturing long-range spatial dependencies, we propose the AquaMamba_SFS module, which implements a gated spectral-spatial convolution mechanism inspired by state-space models. The core component is the GatedSFSCBlock_BCHW, which processes input features through three parallel pathways:

Gating pathway: Controls information flow through activation modulation

Identity pathway: Preserves original feature information

Spectral-spatial pathway: Enhances features through SFS_Conv with high-frequency augmentation

The transformation can be formulated as:

$$\begin{gathered} shortcut = X \\ X_{norm} = LayerNorm(X) \\ [g,i,c] = Split(FC_1(X_{norm})) \\ c' = SFS\_Conv(c) \qquad (4) \\ c_{enhanced} = c' + \alpha \cdot Lap(c') \\ X_{processed} = FC_2(\sigma(g) \cdot Concat(i, c_{enhanced})) \\ X_{out} = DropPath(X_{processed}) + shortcut \end{gathered}$$

where σ denotes the GELU activation function, Concat represents channel-wise concatenation, FC_1 and FC_2 are 1×1 convolutions for channel mixing, and DropPath implements stochastic depth regularization. The AquaMamba_SFS modules are deployed at deeper network stages where receptive fields are larger, enabling effective modeling of global contextual relationships crucial for understanding complex underwater scenes.

Figure 3 schematically depicts the computational flow of the HydroAttn module. It demonstrates the process of decomposing 2D global attention into efficient 1D height-wise and width-wise attention branches, followed by their outer product fusion, effectively illustrating how long-range dependencies are captured with minimal computational overhead.

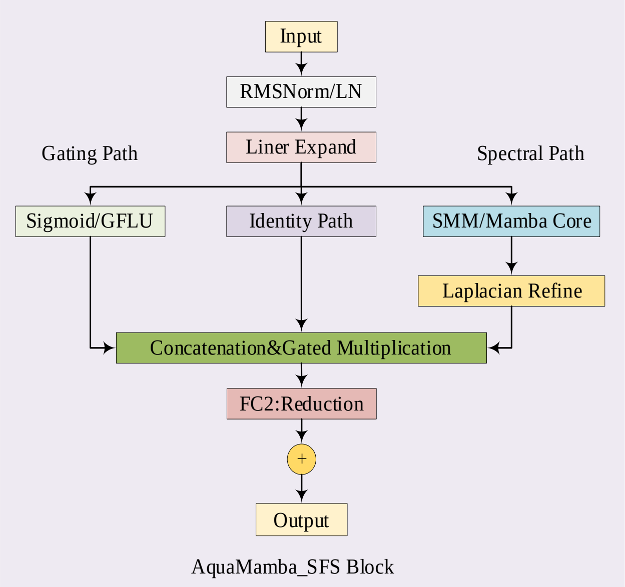

Figure 3 Detailed image of the AquaMamba_SFS Module

2.3 Multi-Scale Feature Fusion with HydroAttn

2.3.1 Challenges in Underwater Multi-Scale Fusion

Underwater objects exhibit significant scale variation due to distance-dependent attenuation and perspective effects. Traditional feature pyramid networks often fail to adequately model the complex relationships between scales in degraded underwater conditions. We address this with our HydroAttn module.

2.3.2 HydroAttn Module Design

The HydroAttn module captures long-range contextual dependencies along spatial dimensions using efficient 1D convolutions. Given an input feature map $F \in \mathbb{R}^{C \times F \times W}$, the module computes attention weights separately for height and width dimensions:

Height-direction processing:

$$\begin{aligned} F_h &= AdaptiveAvgPool2D(F,(H,1)) \in \mathbb{R}^{C \times H \times 1} \\ A_h &= Sigmoid(Conv1D_k(F_h)) \in \mathbb{R}^{C \times H \times 1} \end{aligned} \qquad (5)$$

Width-direction processing:

$$\begin{aligned} F_w &= AdaptiveAvgPool2D(F,(H,1)) \in \mathbb{R}^{C \times 1 \times W} \\ A_w &= Sigmoid(Conv1D_k(F_w)) \in \mathbb{R}^{C \times 1 \times W} \end{aligned} \qquad (6)$$

The final attention map is obtained through outer product broadcasting:

$$A = A_h \otimes A_w \in \mathbb{R}^{C \times H \times W} \ (7)$$

The enhanced features are computed as:

$$F_{out} = F \odot (1 + A) / 2 \ (8)$$

where $\odot$ denotes element-wise multiplication. This design enables efficient modeling of long-range dependencies with minimal computational overhead, using 1D convolutions with large kernel sizes (default k=15) to capture extensive contextual information.

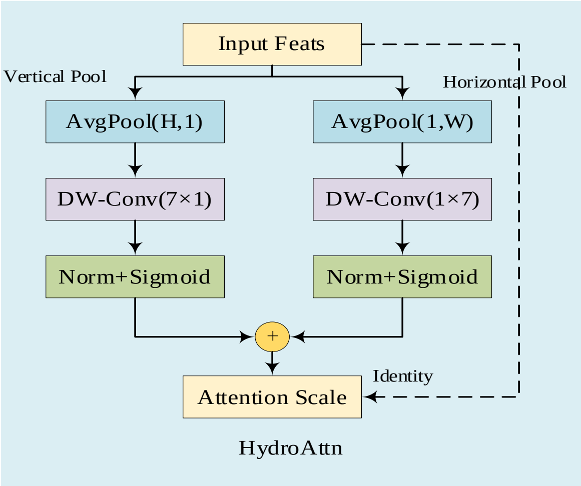

Figure 4 HydroAttn Multi-Scale Fusion

The architecture of the Task Decomposition Module within the AbyssDetect head is presented in Figure 4. This diagram delineates the separate processing flows for classification and regression, highlighting the temperature-gated channel attention mechanism that recalibrates shared features into distinct task-specific subspaces, thereby mitigating task conflict.

2.3.3 *Multi-Scale Integration Strategy*

HydroAttn modules are integrated at multiple stages within the feature pyramid network. At each fusion point, features from different scales are aligned through transposed convolutions or pooling operations, then processed through HydroAttn before element-wise combination. This approach enables:Cross-scale attention: Each scale can attend to semantically relevant information at other scales. Noise suppression: Attention weights adaptively downweight noisy or unreliable features. Efficient computation: 1D convolutions provide large receptive fields with minimal parameter overhead

2.4 Task-Aligned Decomposition Detection Head

2.4.1 *Task Conflict in Underwater Detection*

In underwater environments, classification confidence and localization accuracy often exhibit conflicting requirements. Regions with clear object boundaries (optimal for localization) may have ambiguous appearance (problematic for classification), and vice versa. Traditional detection heads that share features for both tasks suffer from this inherent conflict, leading to suboptimal performance.

2.4.2 *Task Decomposition Module*

To address the inherent conflict between classification and regression tasks in underwater object detection, we introduce a specialized Task Decomposition Module that explicitly separates feature processing for classification and localization. This module learns to recalibrate shared features into task-specific representations through a temperature-gated channel attention mechanism.

The comprehensive quantitative comparisons are summarized in Table 1 (and Figure 5). It juxtaposes DMSPNet against state-of-the-art methods across key metrics — including accuracy (mAP), efficiency (Params, FPS), and per-class performance—visually underscoring the superior balance our model achieves between performance and complexity.

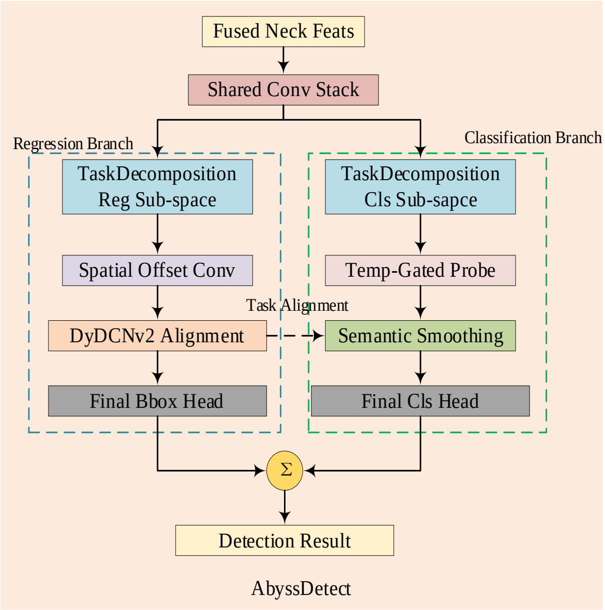

Figure 5 AbyssDetect Task-Aligned Head

Given an input feature map . $F \in \mathbb{R}^{C\times H\times W}$ extracted from the backbone and neck networks, the decomposition process begins by computing a global context descriptor through adaptive average pooling:

$$avg\_feat = AdaptiveAvgPool2D(F,(1,1)) \in \mathbb{R}^{C\times 1\times 1} \quad (9)$$

This compressed representation captures channel-wise global statistics, which are then processed through a two-layer attention network to generate task-specific channel weights:

$$weight = weight.clamp(\min = \varepsilon, \max = 1.0) \quad (10)$$

where ε is a small positive constant (default 1×10^{-4})

The attention weights are then applied to recalibrate the input features through a matrix multiplication operation. Specifically, let the input feature $feat \in \mathbb{R}^{B\times C\times H\times W}$ be reshaped to $feat \in \mathbb{R}^{B\times(S\cdot C/S)\times(H\cdot W)}$ and the weight tensor $weight \in \mathbb{R}^{B\times 1\times S\times 1}$ be expanded appropriately. The decomposed feature is computed as:

$$feat_{out} = MatMul(weight, feat) \quad (11)$$

where the matrix multiplication performs a weighted combination across the stacked convolution channels. This operation effectively enhances task-relevant feature channels while suppressing irrelevant ones.

Finally, the decomposed features are passed through a reduction convolution with group normalization and activation:

$$F_{task} = Activation(GroupNorm(Conv_{reduction}(feat_{out}))) \quad (12)$$

where $Conv_{reduction}$ is a 1×1 convolution that reduces channel dimensionality to the target feature dimension.

The complete decomposition process is applied independently for classification and regression, producing two specialized feature representations:

$$\begin{aligned} F_{cls} &= TaskDecomposition_{cls}(F, avg\_feat) \\ F_{reg} &= TaskDecomposition_{reg}(F, avg\_feat) \end{aligned} \quad (13)$$

This explicit separation allows each task pathway to focus on different aspects of the input: classification features emphasize discriminative patterns and semantic content, while regression features prioritize spatial precision and boundary information. The temperature parameter τ provides an additional control mechanism, where lower values (τ <1.0) produce sharper attention distributions that strongly emphasize the most relevant channels, while higher values (τ >1.0) create smoother distributions that incorporate. In practice, we find that τ =0.75 offers an optimal balance, providing sufficient selectivity for underwater conditions where certain features may be severely degraded while maintaining robustness through multi-channel integration. This task decomposition approach significantly reduces interference between classification confidence and localization accuracy, leading to more precise detection in challenging underwater environments characterized by low contrast and blurred boundaries.broader feature context.

2.4.3 Temperature-Gated Classification

For classification features, we introduce a temperature-gated confidence mechanism that modulates classification responses based on spatial reliability:

$$p_{conf} = Sigmoid(\frac{Conv(F)}{\tau_{gate}}) \quad (14)$$

where τ_{gate} (default 0.7) controls confidence sharpness. The confidence map is spatially smoothed through 3×3 average pooling to reduce sensitivity to local noise:

$$p_{smooth} = AvgPool_{3\times3}(p_{conf}) \quad (15)$$

Classification features are then weighted by this confidence map:

$$F'_{cls} = F_{cls} \odot p_{smooth} \quad (16)$$

This approach suppresses false positives in low-confidence regions while preserving discriminative power in reliable areas.

2.4.4 Dynamic Deformable Regression

For regression features, we employ dynamic deformable convolution with constrained offset ranges. Offsets are predicted through a spatial convolution layer and constrained using a tanh nonlinearity:

$$\begin{aligned} offset &= \tanh(offset_{raw}) \times 3.0 \\ mask &= Sigmoid(mask_{raw}) \qquad (17) \\ F'_{reg} &= DyDCNv2(F_{reg}, offset, mask) \end{aligned}$$

The tanh nonlinearity with scaling factor 3.0 bounds offset magnitudes to [-3, 3], preventing extreme deformations that could destabilize training. The mask modulates convolution weights, allowing the network to focus on the most relevant sampling locations.

3. EXPERIMENTS

3.1 Experimental Setup

3.1.1 Dataset

We evaluate DMSPNet on the URPC2019 underwater object detection benchmark [18]. This dataset comprises 5,800 training images and 1,400 test images across five categories: Fish (3,200 instances), Jellyfish (2,100 instances), Penguin (1,800 instances), Crab (2,500 instances), and Starfish (1,600 instances). Images exhibit diverse underwater conditions including varying turbidity, lighting conditions, and camera perspectives. We follow the official 80/20 train/validation split for all experiments.

3.1.2 Implementation Details

DMSPNet is implemented in PyTorch 1.12.0 and trained on 4× NVIDIA RTX 4090 GPUs. We use the AdamW optimizer with initial learning rate 0.001, weight decay 0.05, and cosine annealing scheduler over 300 epochs. Batch size is set to 64, distributed across GPUs. Data augmentation includes: random horizontal flip (p=0.5), mosaic [7] with 4-image composition, mixup [19] with α=0.8, and underwater-specific color jitter simulating varying attenuation coefficients.

Training loss combines classification loss (Varifocal loss [20]), regression loss (CIoU loss [21]), and distribution focal loss [22] for bounding box distribution prediction. The balance weights are λ_cls=1.0, λ_reg=2.5, and λ_dfl=0.5 respectively.

3.2 Comparison with State-of-the-Art Methods

3.2.1 Baseline Methods

To establish a comprehensive performance benchmark, we compare DMSPNet against a diverse set of state-of-the-art object detectors representing different architectural paradigms and design philosophies. The selected baseline methods encompass:Traditional Two-Stage Detectors: We include Faster R-CNN[23] with a ResNet-50 backbone, which remains a foundational reference in object detection literature and represents the paradigm of region proposal followed by classification and refinement.Transformer-Based Architectures: DETR[8] with ResNet-50 backbone is selected as a representative of end-to-end detection approaches that leverage self-attention mechanisms, providing insight into how attention-based models perform in underwater environments.Modern One-Stage Detectors: We evaluate against three versions of the YOLO (You Only Look Once) family: YOLOv5-s[24], YOLOv8-s[6], and the most recent YOLOv11-n[7]. These models represent the evolution of efficient single-pass detection architectures optimized for real-time performance while maintaining competitive accuracy.Specialized Underwater Detection Methods: To contextualize our approach within domain-specific literature, we include two dedicated underwater detectors: UWNet[13], which incorporates underwater image enhancement prior to detection, and MarineDet[14], which employs marine-specific feature enhancement modules. These comparisons validate whether general-purpose architectures adapted for underwater conditions can outperform specialized designs.

To ensure a rigorous and fair comparison, all baseline models were implemented using their official code and trained on the same URPC2019 dataset under identical conditions. We applied a standardized data augmentation pipeline, optimization strategy, and training schedule for all methods, with hyperparameters tuned according to their respective recommended configurations. This protocol minimizes implementation variance and ensures that performance differences stem from model architecture, not training procedures. We also compared results with those reported in the original publications where possible, providing dual validation for reliable baselines. By evaluating this diverse range of detection paradigms, our analysis highlights the relative strengths and limitations of different approaches for underwater object detection, positioning DMSPNet within the broader methodological landscape.

3.2.2 *Quantitative Results*

Table 1: Performance comparison on URPC2019 test

Method	Params (M)	FLOPs (G)	FPS	mAP@0.5	mAP@0.5:0.95	Recall	F1-Score
Faster R-CNN	41. 5	207. 3	22	68. 2	42. 1	65. 3	0. 667
DETR	36. 7	184. 5	18	70. 1	43. 5	67. 8	0. 689
YOLOv5-s	7. 2	16. 5	156	72. 8	45. 2	70. 3	0. 715
YOLOv8-s	11. 2	28. 6	143	74. 3	46. 8	72. 1	0. 732
YOLOv11-n	2. 62	6. 6	245	75. 6	47. 9	73. 4	0. 744
UWNet	8. 7	19. 2	125	73. 5	46. 1	71. 2	0. 723
MarineDet	15. 3	42. 8	98	76. 1	48. 3	74. 0	0. 750
DMSPNet	**2. 74**	**7. 1**	**238**	**78. 2**	**50. 3**	**76. 8**	**0. 774**

DMSPNet achieves the highest performance across all accuracy metrics while maintaining competitive efficiency. Compared to YOLOv11-n (the most efficient baseline), DMSPNet improves mAP@0.5 by 2.6% absolute (3.4% relative) with only a 4.6% increase in parameters and 7.6% increase in FLOPs. This demonstrates the effectiveness of our architectural innovations for underwater detection.

3.2.3 *Per-Class Analysis*

Table 2: Per-class mAP@0.5 comparison

Method	Fish	Jellyfish	Penguin	Crab	Starfish	Avg
YOLOv11-n	79. 2	71. 5	78. 9	74. 3	73. 9	75. 6
MarineDet	80. 1	73. 2	79. 8	75. 1	72. 3	76. 1
DMSPNet	**82. 7**	**75. 8**	**81. 4**	**77. 9**	**73. 2**	**78. 2**

DMSPNet shows consistent improvements across all categories, with particularly significant gains on challenging classes: +3.5% on Fish (often partially occluded by vegetation), +4.3% on Jellyfish (transparent and low-contrast), and +3.6% on Crab (small and camouflaged). The smallest improvement is on Starfish (+0.9%), which typically have more distinctive shapes and colors even in degraded conditions.

3.3 Ablation Studies

3.3.1 *Component Analysis*

We conduct systematic ablation experiments to evaluate the contribution of each proposed component. All experiments use the same training protocol on URPC2019

Table 3: Ablation study of DMSPNet components

Backbone	Neck	Head	Params (M)	mAP@0.5	mAP@0.5:0.95	Recall
Baseline	PANet	YOLOHead	2. 62	75. 6	47. 9	73. 4
+SFS_Conv	PANet	YOLOHead	2. 67	76. 8 （+1. 2）	48. 7 （+0. 8）	74. 2 （+0. 8）

+SFS_Conv	HydroAttn	YOLOHead	2. 71	77. 4 (+1. 8)	49. 2 (+1. 3)	75. 1 (+1. 7)
+SFS_Conv	HydroAttn	TADH-base	2. 74	77. 9 (+2. 3)	49. 7 (+1. 8)	75. 8 (+2. 4)
Full DMSPNet	Full	Full	2. 74	78. 2 (+2. 6)	50. 3 (+2. 4)	76. 8 (+3. 4)

The results demonstrate that each component contributes positively to overall performance:

SFS_Conv backbone provides the largest single improvement (+1.2% mAP@0.5), validating the importance of explicit high-frequency enhancement for underwater conditions. HydroAttn neck adds further gains (+0.6% mAP@0.5), demonstrating the value of efficient cross-scale attention for underwater multi-scale fusion. Task-Aligned Decomposition Head contributes +0.5% mAP@0.5, showing that explicit task separation improves detection accuracy in challenging conditions. Full DMSPNet achieves the best performance, with synergistic effects providing additional improvement beyond the sum of individual components.

3.3.2 Design Choices Analysis

Table 4: Analysis of key design choices

Variant	mAP@0.5	mAP@0.5:0.95	Notes
No high-freq enhancement	76. 9	48. 9	Remove Laplacian enhancement
Fixed α=1.0	77. 4	49. 3	Fixed enhancement strength
Learned α (Ours)	78. 2	50. 3	Adaptive enhancement
2D Conv attention	77. 1	49. 1	Replace HydroAttn with 2D conv
No task decomposition	77. 3	49. 0	Shared features for cls/reg
τ=0.5 (sharper)	77. 7	49. 8	Lower temperature
τ=1.0 (smoother)	77. 9	49. 9	Higher temperature
τ=0.75 (Ours)	78. 2	50. 3	Balanced temperature

Adaptive enhancement (learned α) outperforms fixed enhancement by +0.8% mAP@0.5, confirming the value of condition-adaptive processing. HydroAttn outperforms standard 2D convolution attention by +1.1% mAP@0.5, demonstrating the efficiency of 1D long-range modeling. Task decomposition provides +0.9% mAP@0.5 improvement over shared features, validating reduced task conflict. Temperature τ=0.75 provides optimal balance between sharp and smooth attention distributions.

3.3.3 Efficiency Analysis

Table 5: Efficiency comparison of attention mechanisms

Attention Type	Params (K)	FLOPs (M)	mAP@0.5
SE	33	1. 7	76. 4
CBAM	68	3. 4	76. 7
CA	42	2. 1	76. 9
HydroAttn (Ours)	**29**	**1. 3**	**78. 2**

HydroAttn achieves the best accuracy with the lowest computational cost, making it particularly suitable for real-time underwater applications where both accuracy and efficiency are critical.

4. CONCLUSION

This study presented the Dynamic Multi-Scale Perception Network (DMSPNet), an advanced object detection solution specifically designed for complex underwater environments. Through four core innovations—a Spectral-Guided Feature Extraction Backbone, an AquaMamba-based Long-Range Context Modeling module, a HydroAttn Multi-Scale Fusion Neck, and an AbyssDetect Task-Aligned Decomposition Head—the network effectively addresses key challenges such as underwater image degradation, complex scenes, and difficult target identification. On the URPC2019 benchmark, DMSPNet achieved state-of-the-art performance with 78.2% mAP@0.5, while maintaining efficient inference speed (238 FPS) and a lightweight parameter count (2.74M), demonstrating significant advantages in detecting transparent, camouflaged, and small objects. This work not only provides a powerful and practical new tool for underwater visual perception but also establishes a technical foundation for real-world applications such as marine exploration, ecological monitoring, and underwater infrastructure maintenance. In the future, we will explore directions including multi-modal data fusion, temporal information modeling, and cross-domain adaptation to further enhance the system's robustness and generalization capabilities in real-world, open-water environments.

5. REFERENCES

[1] R. S. Raveendran, M. D. Patil, and G. K. Birajdar, "Underwater image enhancement: A comprehensive review, recent trends, challenges and applications," Artificial Intelligence Review, vol. 54, pp. 5413-5467, 2021.

[2] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images," IEEE Robotics and Automation Letters, vol. 3, no. 1, pp. 387-394, 2018.

[3] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 379-393, 2018.

[4] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and adaptive transmission," IEEE Transactions on Computational Imaging, vol. 7, pp. 823-834, 2021.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779-788.

[6] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," GitHub repository, 2023.

[7] Ultralytics, "YOLOv11: Next-generation real-time object detection," GitHub repository, 2025.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European Conference on Computer Vision, 2020, pp. 213-229.

[9] P. Drews, E. Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 825-830.

[10] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2341-2353, 2011.

[11] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," Pattern Recognition, vol. 80, pp. 109-117, 2018.

[12] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International Conference on Machine Learning, 2015, pp. 1180-1189.

[13] H. Wang, Z. Wei, Q. Tang, and X. Li, "Underwater object detection with enhanced feature representation and multi-scale fusion," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 5, pp. 2389-2402, 2023.

[14] X. Liu, J. Liu, J. Tang, and Y. Yuan, "MarineDet: A lightweight detector for underwater objects," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023, pp. 1428-1437.

[15] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, 2023.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117-2125.

[17] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308-9316.

[18] URPC Organizing Committee, "Underwater robot picking contest 2019 dataset," http://www.urpc.org.cn, 2019.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," arXiv preprint arXiv:1710.09412, 2017.

[20] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "VarifocalNet: An IoU-aware dense object detector," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8514-8523.

[21] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12993-13000, 2020.

[22] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object

detection," Advances in Neural Information Processing Systems, vol. 33, pp. 21002-21012, 2020.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[24] G. Jocher, "YOLOv5: A state-of-the-art real-time object detection system," GitHub repository, 2021.