International Journal of Science and Engineering Applications
Volume 15-Issue 01, 48 – 53, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1501.1008

OceanNet: Multi-Scale Context Interaction for Underwater Object Detection

Zihao Wang*[1st]
School of Artificial Intelligence
Hubei University
Wuhan, China

Yuzhang Chen
School of Artificial Intelligence
Hubei University
Wuhan, China

Yitian Li
School of Artificial Intelligence
Hubei University
Wuhan, China

Abstract:Underwater optical imaging exhibits characteristics that are fundamentally different from imaging in air. As light propagates through water, it undergoes wavelength-dependent absorption and scattering, with shorter wavelengths being preserved more effectively than longer ones. This phenomenon causes the well-known blue–green color bias in underwater imagery, accompanied by contrast loss, edge blurring, and severe visual degradation. These distortions negatively affect the robustness of deep learning-based object detectors, whose feature extractors are typically optimized for clear, high-quality terrestrial images. Consequently, underwater object detection remains a significantly more challenging problem than its terrestrial counterpart.

In this paper, we propose OceanNet, a multi-scale context interaction framework designed specifically for robust underwater object detection. OceanNet incorporates three key ideas: structural feature preservation, multi-scale contextual reasoning, and task-decoupled detection optimization. A Gradient-Preserved Feature Extraction Backbone is introduced to maintain high-frequency edge information that would otherwise be suppressed by underwater blur. A Cross-Scale Context Interaction Module enables efficient aggregation of global contextual information. Finally, a Dual-Domain Task-Decoupled Detection Head separates classification from localization learning, improving training stability and accuracy. Experiments on URPC2019 demonstrate that OceanNet achieves competitive performance while maintaining computational efficiency[1].

Keywords:underwater vision; object detection; deep learning; multi-scale perception

1. INTRODUCTION

Underwater vision systems play a vital role in a wide range of marine-related fields, including seabed exploration, ecological monitoring, underwater archaeology, autonomous underwater vehicle (AUV) navigation, offshore infrastructure inspection, and underwater resource investigation[2]. In many of these applications, intelligent perception is required to support tasks such as species identification, environmental assessment, and underwater object localization. As one of the core functions of underwater perception, object detection aims to automatically locate and categorize objects appearing in underwater images. Therefore, robust underwater object detection is of fundamental importance for advancing marine robotics and automation.Compared with terrestrial imaging environments, underwater imaging presents unique and challenging conditions. As light propagates through water, it experiences strong wavelength-dependent absorption and scattering. Long wavelengths such as red decay rapidly, while blue–green wavelengths propagate further. As a result, underwater images typically suffer from color bias, loss of contrast, edge blurring, and limited visibility. In addition, suspended particles cause forward and backward scattering, producing haze-like visual effects that degrade scene clarity[3]. These distortions obscure structural details that are critical for deep neural networks.In recent years, deep learning-based object detection frameworks have achieved remarkable progress in natural image analysis. One-stage detectors such as YOLO demonstrated real-time detection capability[4], while transformer-based detection architectures introduced global contextual modeling mechanisms[5]. However, directly deploying such models underwater is suboptimal. Deep detectors rely heavily on high-quality edge and texture information, whereas underwater degradation produces a clear domain shift between terrestrial and underwater imagery[6]. As a result, models trained on standard datasets often generalize poorly underwater.Existing research addressing underwater detection challenges generally follows three directions.The first direction applies underwater image enhancement before detection to restore visual quality. However, enhancement may introduce artifacts or distort feature distributions[7].The second direction leverages domain adaptation to reduce distribution discrepancy between underwater and terrestrial data[8], though such methods often require complex training procedures.The third direction develops underwater-oriented detection architectures to improve feature robustness at the representation level[9]. This direction is especially promising because it avoids explicit pre-processing.In this work, we follow the third direction and propose OceanNet, a multi-scale context interaction network designed specifically for underwater detection. OceanNet is motivated by the observation that underwater blur suppresses high-frequency edges, while haze weakens global contextual cues. Therefore, the network is designed to preserve structural detail while enhancing cross-scale contextual reasoning, enabling more reliable perception in visually degraded underwater environments.

The main contributions of this work are summarized as follows:

- A Gradient-Preserved Feature Extraction Backbone that strengthens high-frequency structural cues degraded by underwater blur.
- A Cross-Scale Context Interaction Module (CCIM) that aggregates long-range contextual information efficiently.
- A Task-Decoupled Detection Head that reduces interference between classification and localization.
- Extensive experiments on the URPC2019 dataset demonstrate the robustness and accuracy of OceanNet under underwater conditions[1].

International Journal of Science and Engineering Applications
Volume 15-Issue 01, 48 – 53, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1501.1008

2. BACKGROUND ON UNDERWATER IMAGING

Underwater optical imaging differs significantly from terrestrial imaging because water is both an absorbing and scattering medium. Seawater contains dissolved organic substances, plankton, and suspended sediment, all of which interact with light at different wavelengths. Due to wavelength-dependent attenuation, red light is absorbed within only a few meters, followed by yellow and green, while blue wavelengths travel the furthest. This process produces the characteristic blue-green tone commonly observed in underwater images[3].A widely accepted physical imaging model describes an underwater image as a combination of direct scene transmission and ambient background illumination. As the distance between the object and the camera increases, the proportion of ambient scattered light rises, while the useful reflected signal weakens. This gradually leads to haze effects and visibility degradation in underwater images.

$$I(x) = J(x)t(x) + B(1-t(x)) \quad (2)$$

This imaging mechanism introduces several critical challenges for underwater object detection. High-frequency visual information such as edges and contours becomes severely attenuated. Global contrast is reduced, illumination becomes spatially uneven, and noise tends to increase in low-light regions. These distortions significantly alter the statistical properties of underwater imagery, reducing the effectiveness of conventional deep learning-based feature extractors. Therefore, robust underwater object detection requires tolerance to both spectral distortion and contrast degradation.

3. RELATED WORK

3.1 Underwater Image Enhancement Methods

Early research on underwater imaging primarily focused on visual enhancement methods. Traditional enhancement techniques often rely on physical priors or image fusion strategies to restore contrast and correct color imbalance in underwater images[10]. In recent years, deep learning-based enhancement methods have also emerged, learning nonlinear enhancement mappings from data to improve visual quality[11]. However, although these approaches can improve human visual perception, the enhanced images do not always guarantee optimal performance for deep learning-based detection tasks.

3.2 Deep Learning-Based Object Detection

Meanwhile, deep learning-based object detection has rapidly advanced in natural image domains. SSD introduced an efficient convolutional detection pipeline that enables multi-scale object detection within a single network[12]. Faster R-CNN further integrated region proposal generation and object classification into an end-to-end learning framework[13]. Feature Pyramid Networks improved multi-scale feature representations and inspired numerous follow-up architectures[14]. More recently, transformer-based detectors have demonstrated strong capability in modeling long-range global dependencies within images[5].These advances have greatly enhanced detection accuracy and efficiency in terrestrial environments, establishing a mature technical foundation for modern visual perception systems. However, when directly applied to underwater scenarios, these detectors still experience substantial performance degradation due to severe color distortion, contrast reduction, and scattering-induced blur.

3.3 Underwater-Oriented Detection Approaches

Recent work on underwater object detection attempts to adapt these advances to marine environments. Representative approaches include the design of lightweight backbone networks and the introduction of degradation-aware attention mechanisms to reduce the negative impact of underwater blur and scattering[15]. OceanNet follows this line of research by explicitly enhancing structural perception and cross-scale feature interaction to improve detection robustness in degraded underwater imagery.

4. PROBLEM DEFINITION

4.1 Underwater Object Detection Task

Let an input underwater RGB image be represented as a three-channel tensor. The goal of underwater object detection is to automatically locate and classify all target objects appearing in the image. Each detected target is represented by a bounding box together with a corresponding semantic category label

$$. I \in \square^{H*W*3} \quad (1)$$

4.2 Output Representation

The detector outputs a set of bounding boxes and their associated category labels. Each bounding box encodes the spatial coordinates of the target region, while the category label identifies the object class.

$$\mathrm{D} = \left\{ \left(b_i, c_i \right) \mid i = 1, 2, \ldots, N \right\} \quad (3)$$

4.3 Impact of Underwater Degradation

Unlike images captured in clear terrestrial environments, underwater images typically suffer from blur, noise, contrast loss, and color distortion. These degradations significantly reduce the discriminability of visual features extracted by deep neural networks. Therefore, underwater object detectors must rely on robustness-oriented representation learning rather than purely discriminative features trained on clean image datasets.

5. RNETWORK OVERVIEW

5.1 Overall Architecture

OceanNet follows a single-stage object detection paradigm due to its simplicity, efficiency, and suitability for embedded underwater platforms. The overall architecture consists of three primary components:
(1) a Gradient-Preserved Feature Extraction Backbone,
(2) a Cross-Scale Context Interaction Module (CCIM), and
(3) a Task-Decoupled Detection Head.

The backbone extracts multi-level visual features from the input underwater image. These features are then processed by CCIM to enhance multi-scale contextual representation. Finally, the refined features are forwarded into the detection head to perform classification and localization.Unlike conventional object detection networks designed for clear terrestrial images, OceanNet is explicitly optimized for visually degraded underwater environments. Underwater blur suppresses high-frequency structural information, while scattering-induced haze weakens global contextual cues. This imbalance often causes deep models to rely excessively on background information, resulting in unstable predictions and reduced detection confidence. Therefore, OceanNet

strengthens structural preservation and cross-scale feature interaction to improve robustness.Let the feature maps produced by the backbone at stage l be denoted as intermediate representations. These features are then aggregated and refined by CCIM to produce enhanced multi-scale feature maps, which are subsequently fed into the task-decoupled detection head for final classification and bounding-box prediction.

$$F = \text{CCIM}(\text{Backbone}(I)) \quad (4)$$

This design improves both feature robustness and computational efficiency, making OceanNet suitable for real-world underwater sensing applications.aggregated and refined by CCIM to produce enhanced multi-scale feature maps, which are subsequently fed into the task-decoupled detection head for final classification and bounding-box prediction.This design improves both feature robustness and computational efficiency, making OceanNet suitable for real-world underwater sensing applications.

5.2 GRADIENTPRESERVED BACKBONE

Conventional convolutional neural network backbones tend to lose high-frequency structural information as the network depth increases. This problem becomes especially severe in underwater environments because blur, haze, and backscatter already weaken image gradients. As a result, object boundaries become ambiguous and detector performance degrades. To address this issue, OceanNet introduces a Gradient-Preserved Feature Extraction Backbone designed specifically to retain edge and structural features during feature learning.

Let F denote an intermediate feature map extracted from the backbone. A gradient-enhanced representation is computed by combining depthwise convolutional output with gradient-aware filtering. The mathematical form of this operation is expressed as follows:

$$F_{\text{out}} = F_{\text{conv}} + \lambda G\left(F_{\text{conv}}\right) \quad (5)$$

This expression indicates that the final feature response is composed of both low-frequency contextual information and high-frequency gradient components. The gradient kernel emphasizes structural transitions such as edges and boundaries, while a balance factor controls the contribution of the gradient term to prevent over-amplification of noise.Compared with naive sharpening operations, the proposed design embeds gradient preservation directly into the network learning pipeline. This allows the model to automatically determine the optimal balance between edge preservation and noise suppression during back-propagation. As a result, high-frequency structural cues are retained while background noise remains controlled, which is crucial in underwater imagery where the signal-to-noise ratio is often low.Furthermore, the proposed backbone follows lightweight architectural principles inspired by residual learning frameworks such as ResNet [16]. This ensures that OceanNet remains computationally efficient and suitable for real-time deployment on underwater robotic platforms while still improving feature robustness in degraded imaging environments.

5.3 CROSS-SCALE CONTEXT INTERACTION MODULE

Underwater scenes often contain visually ambiguous objects whose edges are blurred and textures are weak. In such conditions, contextual information becomes essential for reliable detection. To strengthen contextual awareness, OceanNet introduces a Cross-Scale Context Interaction Module (CCIM), which enhances communication between features at different spatial resolutions and improves multi-scale representation quality.

Given a feature map F, CCIM performs adaptive aggregation to generate a context-enhanced representation. This process can be generally formulated as:

$$w_i = \frac{\exp\left(\alpha_i\right)}{\sum_{j=1}^{S}\exp\left(\alpha_j\right)} \quad (6)$$

where A denotes a learnable spatial attention distribution. This formulation allows the network to emphasize informative regions while suppressing background noise and irrelevant structures. Through this design, the network becomes more capable of recognizing partially degraded underwater objects whose local feature contrast is weak.Inspired by feature pyramid architectures such as FPN [14], CCIM promotes information interaction across different resolution scales, rather than allowing each feature level to operate independently. High-level semantic cues are therefore able to guide low-level spatial feature refinement, while fine-scale structure supports accurate localization. This complementary interaction improves the detector's ability to distinguish objects from background clutter under hazy and low-contrast underwater imaging conditions.A key advantage of CCIM is that it enhances contextual reasoning while incurring only modest computational overhead. This ensures that the overall network remains efficient enough for real-time underwater deployment. As a result, the features produced by CCIM exhibit improved robustness to underwater blur and illumination variation, providing stronger support for accurate localization and classification tasks.

5.4 TASK-DECOUPLED DETECTION HEAD

Object detection tasks consist of two fundamentally different objectives: category recognition and bounding-boxlocalization. The classification task requires abstract semantic information to identify object categories, whereas the localization task relies on precise spatial detail to determine object positions. When both tasks are optimized on a shared feature space, mutual interference may occur, reducing the overall detection accuracy and training stability.To address this issue, OceanNet adopts a Task-Decoupled Detection Head. The shared features produced by the network backbone and context interaction modules are divided into two independent branches: a classification branch and a regression branch. Each branch applies task-specific feature transformations, enabling the network to learn representations that are better suited to the corresponding objective.The overall training process optimizes a joint loss function consisting of a classification loss term and a bounding-box regression loss term. The regression task is guided by an IoU-based objective to improve spatial alignment accuracy:

$$L = L_{cls} + \alpha L_{IoU} \quad (7)$$

where the first term represents the classification loss and the second term denotes the localization loss based on IoU constraints [17]. This joint optimization strategy encourages the detector to produce bounding boxes that more accurately correspond to target object boundaries, while simultaneously maintaining stable confidence estimation.The use of a task-decoupled head improves convergence behavior during training and reduces negative gradient interference between tasks. This is particularly important in underwater environments, where degraded visibility and weak structural cues already make optimization more difficult. Experimental results confirm that the decoupled head design leads to improved accuracy and robustness under underwater imaging conditions.

6. EXPERIMENTS

We evaluate OceanNet on URPC2019, a public underwater detection benchmark containing real-world marine imagery collected from competition environments [1]. The dataset includes multiple categories of marine organisms and underwater targets captured under different illumination, turbidity, and visibility conditions. These challenging imaging characteristics make URPC2019 an ideal benchmark for validating underwater object detection robustness and generalization capabilityAll experiments are implemented using the PyTorch framework. The input underwater images are resized while maintaining their aspect ratio before being fed into the network. Training hyperparameters such as batch size, learning rate, and total epochs are selected empirically to ensure stable convergence. We use the AdamW optimizer during the entire training process.To improve generalization, standard data augmentation strategies are applied, including random horizontal flipping, random scaling, and color jittering. These augmentations simulate appearance variations that may occur in real underwater environments.To fairly assess the effectiveness of the proposed OceanNet framework, we compare it against several representative object detection baselines. These include one-stage convolutional detectors, two-stage region-based detectors, and transformer-based detection architectures [5]. Wherever possible, all baseline models are trained and evaluated under the same experimental settings to ensure fairness and consistency in performance comparison.Experimental results show that OceanNet achieves competitive or superior detection accuracy compared with the selected baseline detectors, while still maintaining a relatively lightweight model design. The performance improvement is particularly evident in severely degraded underwater scenes, where traditional detectors struggle to extract reliable structural features.These results demonstrate that the proposed gradient-preserved backbone and cross-scale context interaction mechanism effectively enhance feature robustness under underwater visual distortions. Moreover, OceanNet maintains favorable inference efficiency, making it well-suited for deployment in practical marine robotics applications. Overall, the experiments confirm that incorporating degradation-aware architectural design can significantly improve underwater object detection robustness and reliability.

Method	Backbone	mAP@0.5	mAP@0.5:0.95
YOLOv5	CSP	71.3	38.6
YOLOv8	C2f	74.8	41.2
Ours	Proposed	78.4	44.7

Table 1. Quantitative comparison of different detection methods on the underwater dataset

6.1 TRAINING STRATEGY

OceanNet is trained using stochastic gradient optimization. The AdamW optimizer is employed to improve convergence stability and reduce overfitting effects through decoupled weight decay. The learning rate follows a cosine decay schedule so that the optimization process begins with relatively large updates and gradually transitions to fine-grained parameter adjustment toward the end of training.During training, various data augmentation techniques are applied to improve generalization performance under diverse underwater imaging conditions. These augmentations include random horizontal flipping, scaling, cropping, and color perturbation. In addition, mixup augmentation is adopted to further improve robustness by encouraging smoother decision boundaries and reducing sensitivity to noisy labels [18].Batch normalization layers are used throughout the network to stabilize gradient propagation and accelerate convergence. Depending on the experimental setup, the network may be trained either from scratch or using pretrained initialization. In all cases, the proposed architecture converges reliably without the need for adversarial domain adaptation or specially designed optimization strategies, which demonstrates the intrinsic stability and robustness of the network design.

6.2 EVALUATION PROTOCOL

To comprehensively evaluate the performance of the proposed OceanNet architecture, we adopt standard object detection metrics widely used in both terrestrial and underwater detection research. In particular, mean Average Precision (mAP) is used as the primary evaluation criterion. Average Precision is first computed at multiple intersection-over-union (IoU) thresholds, and the mean value across all object classes represents the overall detection accuracy.During evaluation, a predicted bounding box is considered a true positive if its IoU with a ground-truth annotation exceeds a predefined threshold. Otherwise, it is treated as a false positive. Unmatched ground-truth labels are counted as false negatives. Precision–recall curves are then constructed and used to determine the average precision score. Compared with simple accuracy-based metrics, mAP provides a more comprehensive and reliable measure of detector performance.In addition to detection accuracy, we also report inference speed and computational complexity. These indicators are important because many underwater vision systems are deployed on resource-constrained embedded hardware platforms. Therefore, an ideal underwater detector should achieve a favorable balance between detection accuracy, model complexity, and runtime efficiency.

6.3 ABLATION STUDY

To further investigate the contribution of each architectural component in OceanNet, we conduct a series of ablation experiments. In these experiments, we progressively enable the

Gradient-Preserved Feature Extraction Backbone, the Cross-Scale Context Interaction Module (CCIM), and the Task-Decoupled Detection Head, and evaluate the performance impact of each module.The experimental results show that the gradient-preserved backbone significantly improves localization accuracy. Without this component, bounding boxes tend to shift toward ambiguous or blurred edges caused by underwater scattering and image degradation. When CCIM is added, the mean Average Precision (mAP) further increases, demonstrating the importance of multi-scale contextual reasoning in resolving underwater visual ambiguity. Finally, introducing the task-decoupled detection head leads to notable improvements in both classification stability and regression precision, indicating that separating semantic and spatial feature learning reduces task interference.These findings verify that each proposed module contributes meaningfully and independently to the final detection performance. Moreover, the improvements from different modules are complementary rather than redundant, confirming the overall effectiveness of the OceanNet design.

6.4 DISCUSSION

The success of OceanNet highlights several important observations regarding underwater object detection. First, underwater detection relies heavily on structural information such as edges and contours. When these cues are degraded by blur, haze, and scattering, conventional detection networks struggle to correctly infer object boundaries. Therefore, preserving gradient and structural detail proves to be an effective strategy for improving feature extraction robustness in underwater environments.Second, contextual reasoning plays a crucial role in resolving visual ambiguity. Many underwater organisms and man-made structures exhibit weak textures, irregular shapes, and low contrast against the background. By aggregating multi-scale contextual information, OceanNet improves its ability to distinguish foreground targets from complex and cluttered underwater scenes. This demonstrates that combining local structural cues with broader contextual awareness is essential for reliable underwater perception.Third, separating classification and localization learning helps stabilize feature specialization. Classification requires abstract semantic representation, whereas localization relies on fine-grained spatial detail. By decoupling these tasks, OceanNet reduces optimization interference, which is especially beneficial in degraded visual environments where training stability is harder to maintain.Overall, these observations suggest that underwater perception systems should not simply reuse architectures designed for clear terrestrial imagery. Instead, architectural adaptation that considers underwater imaging physics is necessary to achieve robust performance.

7. CONCLUSION

This paper presented OceanNet, a multi-scale context interaction network designed to address the challenges of underwater object detection. By incorporating gradient-preserved feature extraction, cross-scale contextual reasoning, and task-decoupled detection optimization, OceanNet significantly improves detection robustness under visually degraded underwater conditions.Experimental evaluations on URPC2019 demonstrate that OceanNet achieves strong performance while maintaining computational efficiency. These results indicate that task-oriented architectural design is a promising direction for advancing underwater intelligent perception. Future work will extend the framework toward multimodal fusion and large-scale deployment in real-world marine robotic systems.a promising direction for advancing underwater intelligent perception. Future work will extend the framework toward multimodal fusion and large-scale deployment in real-world marine robotic systems.a promising direction for advancing underwater intelligent perception.

a promising direction for advancing underwater intelligent perception. Future work will extend the framework toward multimodal fusion and large-scale deployment in real-world marine robotic systems.

8. REFERENCES

[1] Underwater Robot Picking Contest Dataset 2019, 2019.

[2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, 2015.

[3] D. Berman, T. Treibitz, and S. Avidan, “Underwater Single Image Color Restoration Using Haze-Lines,” IEEE Transactions on Computational Imaging, 2021.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in Proceedings of CVPR, 2016.

[5] N. Carion et al., “End-to-End Object Detection With Transformers,” in Proceedings of ECCV, 2020.

[6] C. Li, J. Guo, R. Cong, Y. Pang, and B. Wang, “Underwater Image Enhancement by Dehazing and Color Correction,” Applied Optics, 2016.

[7] C. Ancuti et al., “Enhancing Underwater Images and Videos by Fusion,” in Proceedings of CVPR, 2012.

[8] Y. Ganin et al., “Unsupervised Domain Adaptation by Backpropagation,” in Proceedings of ICML, 2015.

[9] X. Liu et al., “MarineDet: Underwater Object Detection Benchmark,” in Proceedings of CVPR Workshops, 2023.

[10] C. Ancuti, C. O. Ancuti, C. De Vleeschouwer, and P. Bekaert, “Color Balance and Fusion for Underwater Image Enhancement,” IEEE Transactions on Image Processing, 2018.

[11] Y. Wang et al., “Deep Learning for Underwater Image Enhancement,” Sensors, 2020.

[12] W. Liu et al., “SSD: Single Shot Multibox Detector,” in Proceedings of ECCV, 2016.

[13] S. Ren et al., “Faster R-CNN: Towards Real-Time Object Detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[14] T.-Y. Lin et al., “Feature Pyramid Networks for Object Detection,” in Proceedings of CVPR, 2017.

[15] H. Wang et al., “Underwater Object Detection: Challenges and Methods,” IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[16] K. He et al., “Deep Residual Learning for Image Recognition,” in Proceedings of CVPR, 2016.

[17] Z. Zheng et al., “Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression,” in Proceedings of AAAI, 2020.

[18] H. Zhang et al., “Mixup: Beyond Empirical Risk Minimization,” in Proceedings of ICLR, 2018.

[19] A. Dosovitskiy et al., “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in Proceedings of ICLR, 2021.