A Robust Physics-Aware and Attention-Driven Task-Oriented Deep Learning Framework for Underwater Object Detection

Prabu S[1,2], Abeer Aljohani[3]

[1]Postdoctoral Fellow, Lincoln University College, Petaling Jaya, Selangor, Malaysia;

[2]School of Computing, SRM Institute of Science and Technology Tiruchirappalli Campus, Tiruchirappalli, 621105, India;

[3]Department of Computer Science, Applied College, Taibah University, Al-Madinah Al-Munawwarah, Saudi Arabia

Email ID: [1]pdf.prabhu@lincoln.edu.my, [2]prabus2@srmist.edu.in, [3]aahjohani@taibahu.edu.sa

---

Abstract: Object recognition in underwater environments remains problematic because of the strong reduction of the appearance, contrast and boundary information of objects due to light absorption, wavelength-dependent attenuation and backscatter. The effects are especially challenging for spotting small and low-contrast aquatic objects under murky underwater conditions. The existing methods generally either decouple image enhancement and object detection or use unrestricted enhancement networks, which may lead to objective mismatch between perceptual restoration and detection performance. To overcome these restrictions, in this work we present a task-oriented deep learning framework, named PA-AFNet (Physics-Aware, Attention-Fused Network), which combines the underwater image generation model with the multi-scale object detection. The proposed system consists of a differentiable physics-aware enhancement module, which estimates the transmission and backscatter parameters and reconstructs a physically restricted improved image. The estimated transmission information is kept as an auxiliary signal and directly fed into a multi-scale attention fusion module that is integrated with the YOLOv11 feature pyramid. This allows the detector to adaptively refine the features according to the spatially variable degradation characteristics of the underwater images. The whole framework is optimized in an end-to-end manner under the detection objective, such that the enhancement and detection modules can be learned collaboratively rather than separately. The experiments were conducted on URPC2017 and URPC2018 underwater object detection benchmarks with the same training setting. The proposed framework is effective with 88.2% precision, 84.1% recall, 88.2% mAP@0.5, 54.8% mAP@0.5:0.95 and an F1-score of 86.10% to demonstrate the combination of physical image modeling, task-oriented enhancement and degradation-conditioned multi-scale feature attention for underwater object detection.

Keywords: Underwater object detection; physics-informed deep learning; underwater image formation model; multi-scale attention; task-oriented enhancement; YOLOv11.

Introduction

The impact of optical object detection in autonomous underwater vehicles, remotely operated vehicles and fixed underwater surveillance systems is expanding for applications such as marine ecological

monitoring and subsea infrastructure inspection. Unlike terrestrial images, the acquisition of underwater visual data involves a medium actively participating in the formation of the image: light is absorbed and scattered depending on wavelength, path length, and optical properties of the water column, leading to a systematic and physically describable degradation, rather than an arbitrary noise process [1,2]. The detectors are computationally efficient and well suited to the onboard hardware constraints of underwater platforms, but are typically applied either directly on raw or lightly pre-processed underwater imagery, ignoring the degradation physics altogether, or at the downstream of a separately trained image enhancement network optimized for restoration quality, rather than for the eventual detection task. Both design choices sacrifice accuracy for reasons that are structural rather than incidental to any particular architecture [3-7].

Three limitations recur across the underwater detection literature and motivate the present work. First, enhancement-detection decoupling introduces an objective mismatch. When an enhancement network is trained with reference-based losses and then frozen before a detector is trained on its output, the enhancement stage has no mechanism to learn which visual cues the detector actually needs. Second, purely data-driven enhancement treats color and contrast correction as an unconstrained image-to-image translation problem [3,4][6,7]. Third, existing attention mechanisms integrated into underwater detectors are typically imported from terrestrial detection literature without accounting for the fact that underwater degradation affects different spatial frequencies and object scales unevenly. Backscatter primarily degrades low-frequency global contrast and disproportionately affects small, distant objects whose signal is already weak, whereas forward scattering blurs edges at all scales but is more perceptually tolerable for large objects with strong low-frequency support. A single-scale attention mechanism, however finely tuned, cannot represent this scale-dependent degradation asymmetry, motivating a fusion mechanism that operates explicitly across the multiple feature-pyramid scales [8,9] already produced by the detection backbone.

Related work

Early underwater image correction methods were largely built on hand-crafted priors adapted from atmospheric dehazing, most notably variants of the dark channel prior adjusted for the wavelength-dependent absorption that distinguishes water from air as a scattering medium [1,2][6]. These methods produce interpretable, physically motivated corrections but rely on assumptions — such as locally constant transmission or a single global backscatter estimate — that break down in scenes with strong depth variation or artificial illumination, both common in ROV-captured footage. Subsequent work replaced hand-crafted priors with learned estimators for transmission and veiling light while retaining the same underlying formation equation, improving robustness to scene variation while retaining physical interpretability.

The introduction of deep convolutional and, more recently, transformer-based architectures shifted the field toward end-to-end image-to-image translation, in which a network is trained to map a degraded underwater image directly to a clean or reference-corrected counterpart using paired or unpaired training data [3-7]. These approaches can produce strong perceptual quality gains, particularly under generative adversarial or diffusion-based objectives. A smaller body of work has attempted to reintroduce physical structure into deep enhancement networks by using the formation equation as a differentiable layer rather than a preprocessing step, which is the design lineage most closely related to

the enhancement component proposed here; our contribution departs from this lineage principally in the choice of training signal, discussed next. The majority of underwater object detection systems adapt terrestrial single-stage detectors, most prominently successive iterations of the YOLO family, to underwater imagery either by direct fine-tuning on underwater datasets or by prepending a fixed or jointly-tuned enhancement stage [8][10-14]. Two-stage and query-based detectors, including DETR-style architectures, have also been explored for underwater settings, generally trading inference speed for a global receptive field that can help disambiguate cluttered benthic scenes.

Attention mechanisms in detection architectures [14,15] range from channel-recalibration modules such as squeeze-and-excitation blocks to spatial and self-attention mechanisms adapted from vision transformers. Within underwater detection specifically, attention has predominantly been applied as a generic feature-refinement block inserted into an existing backbone or feature pyramid, without an explicit mechanism for conditioning attention on the estimated severity or spectral character of the underwater degradation present in a given image.

Proposed Methodology

The proposed framework, denoted PA-AFNet for reference within this manuscript (Physics-Aware, Attention-Fused Network), consists of three stages trained end-to-end, (see Figure 1):

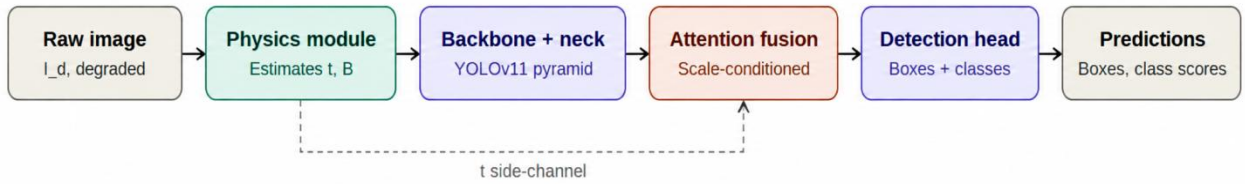


Figure 1. Pipeline architecture of the proposed system.

End-to-End Data Flow

Formally, let Φ_{phys} denote the physics-aware enhancement module, Φ_{bb} the YOLOv11 backbone (including its neck, i.e., the PAN/FPN-style feature aggregation), Φ_{attn} the attention fusion module, and Φ_{head} the detection head. The forward pass is,

$$\hat{I}_c, \hat{t} = \Phi_{phys}(I_d) \tag{1}$$

$$\{F_l\}_{l=1}^{L} = \Phi_{bb}(\hat{I}_c) \tag{2}$$

$$\{\tilde{F}_l\}_{l=1}^{L} = \Phi_{attn}(\{F_l\}_{l=1}^{L}, \hat{t}) \tag{3}$$

$$\hat{Y} = \Phi_{head}(\{\tilde{F}_l\}_{l=1}^{L}) \tag{4}$$

where F_l is the backbone/neck feature map at pyramid level l (with $L = 3$ for YOLOv11's default configuration), $\tilde{F}_l$ is the corresponding attention-refined feature map, and $\hat{Y}$ is the set of predicted bounding boxes, objectness/class scores. Notably, $\hat{t}$, the transmission map produced by the physics-aware module, is passed forward not only for use inside Equation (1)'s inversion but also directly into

the attention fusion module as a conditioning signal — this side-channel is what allows Φ_{attn} to be physics-conditioned rather than a generic learned attention block

The physics-aware module Φ_{phys} is realized as a lightweight CNN estimator E_θ operating directly on I_d at a reduced spatial resolution (downsampled by a factor of 4, then bilinearly upsampled at output) to keep its computational cost small relative to the detection backbone, consistent with the efficiency requirements (GFLOPs, parameters, FPS) that motivated the backbone choice in Section 4.1. E_θ consists of a small number of strided convolutional blocks followed by two parallel prediction heads:

$$\hat{t}(x) = \sigma\left(g_t\left(E_\theta(I_d)\right)(x)\right), \qquad \widehat{B} = \sigma\left(g_B\left(\mathrm{GAP}\left(E_\theta(I_d)\right)\right)\right) \tag{5}$$

where $g_t(\cdot)$ is a 1×1 convolutional head producing a three-channel (per-wavelength) transmission map at the estimator's internal resolution, $g_B(\cdot)$ is a small fully connected head applied after global average pooling (GAP) to produce a single global backscatter estimate per channel, and $\sigma(\cdot)$ is the logistic sigmoid enforcing the range constraints.

For pyramid level l, let $\hat{t}_l = \mathrm{Downsample}_l(\hat{t})$ denote the transmission map resized to F_l's spatial resolution via average pooling (chosen over strided convolution to avoid introducing additional learnable parameters into what is intended as a lightweight conditioning path). We compute a degradation-conditioned attention map

$$A_l = \mathrm{sigmoid}\left(W_l^{(2)}\,\mathrm{ReLU}\left(W_l^{(1)}[F_l \parallel \hat{t}_l]\right)\right) \tag{6}$$

where $[\cdot\|\cdot]$ denotes channel-wise concatenation, and $W_l^{(1)}, W_l^{(2)}$ are 1×1 convolutional weights specific to level l (not shared across levels, since the relationship between transmission and required correction strength differs by scale, per Section 3.3). The refined feature map is then

$$\tilde{F}_l = F_l + A_l \odot \left(W_l^{(3)} F_l\right) \tag{7}$$

where $\odot$ is element-wise multiplication and $W_l^{(3)}$ is a 1×1 convolution producing a transformed feature representation that the attention map modulates; the residual form ($F_l + \cdots$ rather than a direct replacement) is used so that Φ_{attn} can learn to apply zero correction where the transmission-conditioned signal indicates the backbone feature is already reliable, rather than being forced to always alter every feature map regardless of local degradation severity.

Experiment

Dataset Details

Evaluation is conducted on the URPC (Underwater Robot Professional Contest) benchmark series, specifically the URPC2017 and URPC2018 releases, which are among the most widely used public benchmarks for underwater object detection and contain annotated bounding boxes for benthic marine organisms.

Hyperparameter Details

The model is trained with a batch size of 16 and an initial learning rate of 0.001 using AdamW. A cosine-annealing schedule with linear warm-up is applied for learning-rate adjustment. Training uses Mosaic, flipping, HSV jitter, random scaling/cropping, and MixUp augmentation. EMA is enabled with a decay of 0.9999, while weight decay is set to 0.0005. Images are resized to 640 × 640.

Discussions

The performance comparison evaluates the proposed PA-AFNet framework against representative one-stage, convolutional, and transformer-based underwater object detectors (see Table 1). Conventional SSD achieves 75.4% mAP@0.5, while YOLO-based approaches show stronger detection performance, with YOLOv5s, YOLOv6n, and YOLOv7 reaching 83.2%, 83.5%, and 86.2%, respectively.

Table 1. Performance comparison of the proposed framework with other popular objection approaches.

Method	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1-score (%)
SSD	74.2	68.7	75.4	38.8	71.34
YOLOv5s	85.9	78.3	83.2	49.6	81.92
YOLOv6n	87.7	79.1	83.5	50.8	83.18
YOLOv7	85.9	82.2	86.2	52.1	84.01
DETR	85.2	80.9	84.6	46.6	82.99
Proposed	**88.2**	**84.1**	**88.2**	**54.8**	**86.10**

DETR obtains 84.6% mAP@0.5, demonstrating the competitiveness of transformer-based detection. PA-AFNet extends the YOLOv11 detection pipeline by jointly estimating underwater degradation and using the resulting transmission information to condition multi-scale feature attention.

Conclusions

This study presented PA-AFNet, a physics-aware and attention-driven framework for underwater object detection that jointly addresses underwater image degradation and object localization. Unlike conventional pipelines that perform enhancement independently of detection, the proposed approach integrates a differentiable physics-aware image formation module with the YOLOv11 detection architecture and retains the estimated transmission information as a conditioning signal for multi-scale attention fusion. Consequently, the enhancement process is optimized according to the requirements of the downstream detection task, while the attention mechanism can exploit degradation-related information at different feature-pyramid levels. The experimental evaluation on the URPC benchmark series demonstrates the effectiveness of the proposed design. PA-AFNet obtains 88.2% precision, 84.1% recall, 88.2% mAP@0.5, 54.8% mAP@0.5:0.95, and 86.10% F1-score. In the reported comparison, these results exceed the mAP@0.5 values of SSD (75.4%), YOLOv5s (83.2%), YOLOv6n (83.5%), YOLOv7 (86.2%), and DETR (84.6%). Results show the potential of employing physical degradation information in the detection pipeline as a valuable inductive bias for underwater visual analysis, and how transmission conditioned attention enables feature refinement adaptation across various detection scales. The end-to-end optimization technique also establishes a direct relationship between the estimated physical characteristics and the final detection objective.

References

1. D. Akkaynak and T. Treibitz, “A Revised Underwater Image Formation Model,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6723–6732, 2018. doi: 10.1109/CVPR.2018.00703.
2. J. Y. Chiang and Y. C. Chen, “Underwater Image Enhancement by Wavelength Compensation and Dehazing,” IEEE Transactions on Image Processing, vol. 21, no. 4, pp. 1756–1769, 2012. doi: 10.1109/TIP.2011.2179666.
3. C. Li, S. Anwar, and F. Porikli, “Underwater Scene Prior Inspired Deep Underwater Image and Video Enhancement,” Pattern Recognition, vol. 98, p. 107038, 2020. doi: 10.1016/j.patcog.2019.107038.
4. S. Anwar and C. Li, “Diving Deeper into Underwater Image Enhancement: A Survey,” Signal Processing: Image Communication, vol. 89, p. 115978, 2020. doi: 10.1016/j.image.2020.115978.
5. C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, “An Underwater Image Enhancement Benchmark Dataset and Beyond,” IEEE Transactions on Image Processing, vol. 29, pp. 4376–4389, 2020. doi: 10.1109/TIP.2019.2955241.
6. S. Fayaz, S. A. Parah, G. J. Qureshi, and V. Kumar, “Underwater Image Restoration: A State-of-the-Art Review,” IET Image Processing, vol. 15, no. 2, pp. 269–285, 2021. doi: 10.1049/ipr2.12041.
7. G. Verma and M. Kumar, “Systematic Review and Analysis on Underwater Image Enhancement Methods, Datasets, and Evaluation Metrics,” Journal of Electronic Imaging, vol. 31, no. 6, p. 060901, 2022. doi: 10.1117/1.JEI.31.6.060901.
8. J. Yan, Z. Zhou, D. Zhou, B. Su, Z. Xuanyuan, J. Tang, Y. Lai, J. Chen, and W. Liang, “Underwater Object Detection Algorithm Based on Attention Mechanism and Cross-Stage Partial Fast Spatial Pyramidal Pooling,” Frontiers in Marine Science, vol. 9, p. 1056300, 2022. doi: 10.3389/fmars.2022.1056300.
9. K. Liu, L. Peng, and S. Tang, “Underwater Object Detection Using TC-YOLO with Attention Mechanisms,” Sensors, vol. 23, no. 5, p. 2567, 2023. doi: 10.3390/s23052567.
10. K. Liu, Q. Sun, D. Sun, L. Peng, M. Yang, and N. Wang, “Underwater Target Detection Based on Improved YOLOv7,” Journal of Marine Science and Engineering, vol. 11, no. 3, p. 677, 2023. doi: 10.3390/jmse11030677.
11. X. Wang, Y. Li, Q. Zu, et al., “SFL-YOLO: An Improved YOLOv11-Based Model for Underwater Object Detection,” Scientific Reports, 2026. doi: 10.1038/s41598-026-60347-7.
12. Yongqing Shi, Wei Chen, Duo Wan and Lu Han “Underwater Robot Object Detection Algorithm Based on YOLOv11,” Sensors, vol. 26, no. 11, p. 3611, 2026. doi: 10.3390/s26113611.
13. H. Prince and B. T., “Transformer-Enhanced YOLOv5 With Cross-Convolution for Underwater Object Detection,” Journal on Advances in Signal Processing, vol. 2026, p. 51, 2026. doi: 10.1186/s13634-026-01315-1.
14. X. Chen, Q. Zeng, C. Wei, G. Cheng, and X. Li, “A Unified Underwater Object Detection Network with Image Adaptive Enhancement and Attention-Driven Feature Fusion,” Optics & Laser Technology, vol. 199, p. 115028, 2026. doi: 10.1016/j.optlastec.2026.115028.
15. Y.-H. Chen, L.-Y. Yu, and Y.-Y. Chen, “JTA-GAN: A Physics-Informed Framework for Realistic Underwater Image Generation and Improved Object Detection,” Mathematics, vol. 14, no. 4, p. 605, 2026. doi: 10.3390/math14040605.