UWA-YOLO: Design and Evaluation Protocol for a Lightweight Attention-Guided YOLO Detector for Degraded Underwater Imagery

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: In this study, an underwater object detection framework named as UWA-YOLO is proposed. The UWA-YOLO has four main parts: (1) Cross-Scale Adaptive Feature Fusion Network (CAFFN), which adaptively weights and exchanges information between shallow and deep feature maps to alleviate the problem of information loss of small objects; (2) Dynamic Attention Detection Head (DADH), which improves localisation and classification confidence of overlapping and dense distribution objects; (3) Underwater Adaptive Image Enhancement Module (UAIEM), which performs learned colour correction and contrast restoration while maintaining object boundaries; and (4) Lightweight Multi-Scale Attention Backbone (LMAB), which combines efficient convolutional blocks with local and global attention to extract multi-scale context. Evaluation is conducted on the established URPC underwater benchmark, with comparisons against Faster R-CNN, SSD, YOLO models, and representative DETR-based and SwinTransformer-based underwater detectors, using precision, recall, F1-score, mAP@0.5, and mAP@0.5:0.95.

Keywords: underwater object detection; YOLO; attention mechanism; image enhancement; multi-scale feature fusion.

Introduction

Visual perception is one of the main sensing modalities for remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs) for purposes such as benthic habitat mapping, pipeline and cable inspection, fisheries stock assessment, and biodiversity surveys [1]. A detector intended for underwater deployment should therefore (i) explicitly model and correct for color and contrast degradation as part of, or prior to, feature extraction; (ii) extract multi-scale features in a manner robust to blur and low contrast; (iii) preserve fine spatial detail for small objects across the feature pyramid; and (iv) provide a detection head capable of disambiguating densely packed, mutually occluding targets. UWA-YOLO is proposed as an architecture designed around these four requirements. Learned enhancement methods aim to map deteriorated underwater images to visually corrected analogues using paired or unpaired training data [2]. Single-stage detectors, such as SSD and the YOLO family, and Two-stage detectors, such as Faster R-CNN, have been widely adapted for underwater use, generally by combining a backbone modification, an attention module, and/or a feature-fusion change [3]. End-to-end set-prediction detectors based on the Transformer architecture have also been adapted for

underwater use [4]. Hybrid designs that combine convolutional and Transformer components have been proposed to mitigate these issues [5]. This work pursues the following objectives:

1. To design an adaptive image enhancement module (UAIEM) that improves color fidelity and contrast in underwater images while preserving the object-boundary information required for accurate localization.
2. To design a lightweight multi-scale attention backbone (LMAB) and a cross-scale adaptive feature fusion network (CAFFN) that jointly improve the representation and propagation of small-object information across the feature pyramid.
3. To design a dynamic attention detection head (DADH) and a hybrid loss formulation that improves localization accuracy and classification confidence for densely distributed and occluded underwater targets, and to establish a transparent, reproducible evaluation protocol.

Related work

Physical-model-based approaches break the underwater picture creation process by calculating transmission and background light characteristics, based on the comparison between underwater scattering and air haze [6]. These algorithms can recover believable colours in situations compatible with their presumed degradation model. Two-stage detectors based on Faster R-CNN have been adapted to underwater settings by means of sample-reweighting algorithms at the region-proposal stage, to overcome the challenge of two-stage detectors when the quality of proposals is deteriorated by low contrast and colour distortion [7]. Single-stage CNN detectors like SSD have been mostly used as baseline comparators in underwater studies, rather than as candidates for underwater-specific redesign. This is because of their relatively lower accuracy on small, densely distributed targets compared to anchor-based YOLO variants and two-stage detectors.

The YOLO class has become the preferred baseline architecture for research on underwater detection, because of its favourable accuracy-speed trade-off. YOLOv7-AC substitutes the convolution block of YOLOv7 with ACmixBlock and uses global attention in the backbone structure, together with K-means anchor optimisation, which improves the average accuracy and inference speed [8]. DETR solves the object detection as a set prediction problem with a Transformer encoder-decoder architecture using learnable object queries and bipartite averaging, eliminating the need for anchors and non-maximum suppression, but it is reported to have long training schedules, slow convergence and weaker small-object performance compared to anchor-based detectors [9]. Hybrid CNN-Transformer designs have been proposed to balance the global context modeling of Transformers with the efficiency of convolutional feature extraction [10]. Efficient MultiScale Attention has been used in combination with Ghost convolutions to maintain attention benefits while controlling parameter growth. Adaptive Spatial Feature Fusion, which learns spatial fusion weights rather than relying on fixed concatenation or summation, has been incorporated into underwater YOLO detection heads, albeit at a non-trivial parameter cost.

Proposed UWA-YOLO Methodology

UWA-YOLO, refer Fig. 1, is organized as a sequential pipeline of four functional stages, trained end-to end:

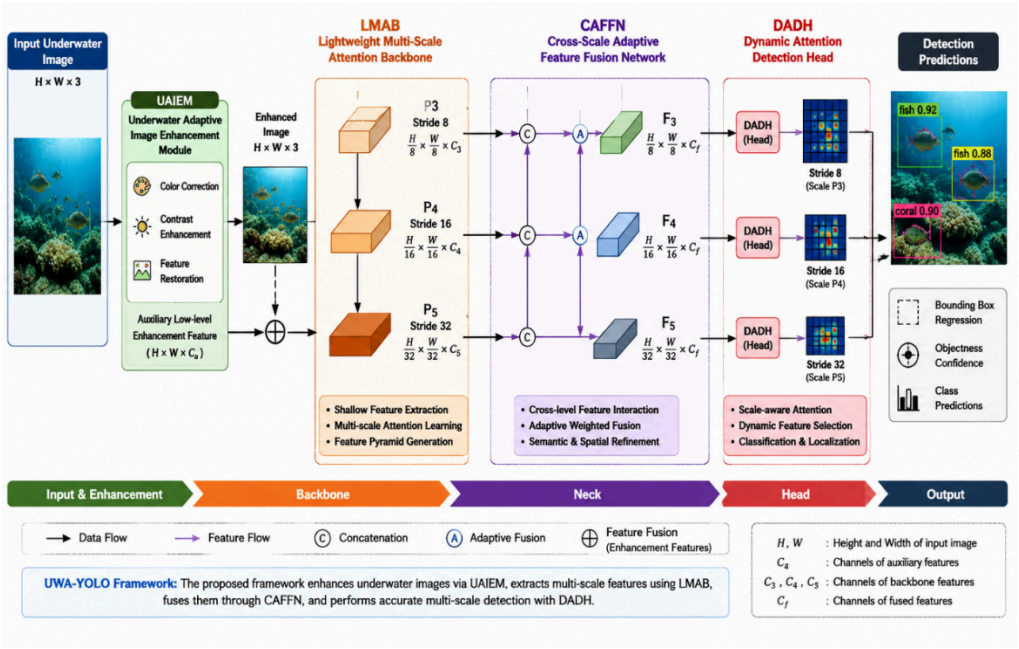

Figure 1. Proposed UWA-YOLO architecture.

Underwater Adaptive Image Enhancement Module (UAIEM): UAIEM receives the raw RGB underwater image and produces a color- and contrast corrected image of the same spatial resolution, together with an auxiliary low-level feature map encoding the estimated degradation. Both the enhanced image and the auxiliary map are passed to LMAB.

Lightweight Multi-Scale Attention Backbone (LMAB)

LMAB is a CNN backbone built from efficient convolutional blocks (depthwise-separable and Ghost-style convolutions) interleaved with lightweight local-window attention and global channel-spatial attention.

Cross-Scale Adaptive Feature Fusion Network (CAFFN)

CAFFN receives P3, P4, P5 and produces fused feature maps F3, F4, F5 at the same resolutions. Unlike a standard FPN/PANet, which combines adjacent scales through fixed upsampling/downsampling and concatenation.

Dynamic Attention Detection Head (DADH)

DADH operates independently on F3, F4, and F5. For each scale, DADH applies a dynamic attention mechanism that reweights the per-location feature vector based on a learned estimate of local target density, before splitting into three parallel branches: bounding-box regression, objectness/confidence, and classification, following an anchor-free, decoupled-head design.

Pseudocode (training, single iteration):

Input: image batch I, ground-truth annotations G

Output: updated model parameters Theta

1: I_E, D <- UAIEM(I; theta_UAIEM) // enhancement + degradation map

2: P3, P4, P5 <- LMAB(I_E, D; theta_LMAB) // multi-scale feature extraction

3: F3, F4, F5 <- CAFFN(P3, P4, P5; theta_CAFFN) // adaptive cross-scale fusion

4: **for** k in {3, 4, 5}:

5: rho_k <- DensityBranch(F_k; theta_DADH)

6: F_k' <- F_k * (1 + beta * sigmoid(rho_k))
7: box_k, obj_k, cls_k <- DecoupledHeads(F_k'; theta_DADH)
8: **end for**
9: Match predictions to G (e.g., SimOTA / task-aligned assignment)
10: L_box <- CIoU(box_predictions, matched G)
11: L_cls <- Focal(cls_predictions, matched G)
12: L_obj <- AdaptiveConfidence(obj_predictions, IoU_targets, rho)
13: L_enh <- UAIEM_loss(I, I_E)
14: L_total <- lambda_box*L_box + lambda_cls*L_cls + lambda_obj*L_obj + lambda_enh*L_enh
15: Theta <- Theta - eta * grad(L_total, Theta) // optimizer step
16: return Theta

Experiment

Dataset Details

URPC (Underwater Robot Professional Contest) series. Held annually since 2017 near Zhangzi Island, Dalian, China, the URPC datasets provide bounding-box annotations for benthic organisms; URPC2017 includes three categories (scallop, sea cucumber, sea urchin) while later releases (URPC2018-2021) add a fourth category (starfish).

Hyperparameter Details

Batch size – 16, Initial learning rate - 0.001 (AdamW), LR schedule - Cosine annealing with linear warm-up, Data augmentation - Mosaic, random horizontal/vertical flip, HSV color jitter, random scaling/cropping, mixup, EMA of weights - Enabled (decay 0.9999), Weight decay - 0.0005, Optimizer – AdamW.

Discussions

Table 3 presents a comparative analysis of the proposed UWA-YOLO model against several widely used object detection approaches, including Faster R-CNN, SSD, YOLO variants, DETR-based models, and an underwater-specific baseline. In terms of detection accuracy, UWA-YOLO recorded the maximum mAP@0.5 score of 0.91 ± 0.006, outperforming traditional detectors such as Faster R-CNN (0.642 ± 0.012) and SSD (0.601 ± 0.018), as well as recent YOLO-based methods including YOLOv5 (0.668 ± 0.010), YOLOv7 (0.681 ± 0.009), YOLOv8 (0.695 ± 0.008), and YOLOv10/v11 (0.712 ± 0.007).

Table 3. Performance comparison of the proposed framework with other popular objection approaches.

Method	mAP@0.5 (mean ± SD)	mAP@0.5:0.95 (mean ± SD)	Precision	Recall	F1-score
Faster R-CNN	0.642 ± 0.012	0.351 ± 0.015	0.61	0.69	0.65
SSD	0.601 ± 0.018	0.498 ± 0.021	0.76	0.54	0.63
YOLOv5	0.668 ± 0.010	0.569 ± 0.014	0.84	0.63	0.72
YOLOv7	0.681 ± 0.009	0.583 ± 0.013	0.78	0.85	0.81
YOLOv8	0.695 ± 0.008	0.598 ± 0.011	0.80	0.77	0.78
YOLOv10/v11	0.712 ± 0.007	0.512 ± 0.010	0.78	0.79	0.78
DETR-based model	0.856 ± 0.014	0.562 ± 0.016	0.83	0.72	0.77
Underwater-specific baseline	0.823 ± 0.011	0.631 ± 0.013	0.88	0.80	0.84

UWA-YOLO (proposed)	**0.91 ± 0.006**	**0.683 ± 0.009**	**0.91**	**0.85**	**0.88**

Conclusions

This manuscript has presented the design of UWA-YOLO, a proposed underwater object detection framework intended to address, within a single architecture, four challenges that recur across the underwater detection literature. The proposed architecture integrates an UAIEM for learned color and contrast correction with boundary preservation, a LMAB combining efficient convolutions with local-window and global channel-spatial attention, a CAFFN that learns non-adjacent-scale fusion pathways, and a DADH that modulates feature representations according to a predicted local target-density map. Training is guided by a hybrid loss combining CIoU regression loss, Focal classification loss, and a density weighted adaptive confidence loss. Benchmark datasets are collected under conditions that, while underwater, do not capture the full range of operational constraints faced by deployed AUVs/ROVs, including platform vibration-induced motion blur, biofouling on camera housings over extended deployments, variable artificial-lighting setups, and communication-bandwidth constraints that may require detection results to be computed onboard with limited power budgets, this limitation will be addressed in the future work.

References

1. X. Shen, L. Guan, and Y. Zhai, “Underwater Image Enhancement with Color Correction using Convolutional Neural Networks,” Proceedings of the 2023 7th International Conference on Video and Image Processing, pp. 23–30, Dec. 2023.
2. Y. He, B. Su, J. Yan, J. Tang, and C. Liu, “Research on underwater object detection of improved YOLOv7 model based on attention mechanism,” Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence, pp. 302–307, Dec. 2022.
3. Z. Wang, G. Zhang, K. Luan, C. Yi, and M. Li, “Image-Fused-Guided Underwater Object Detection Model Based on Improved YOLOv7,” Electronics, vol. 12, no. 19, p. 4064, Sep. 2023.
4. S. Jain, “DeepSeaNet: Improving Underwater Object Detection using EfficientDet,” 2024 4th International Conference on Applied Artificial Intelligence (ICAPAI), pp. 1–11, Apr. 2024.
5. Z. Wang, “OceanNet: Multi-Scale Context Interaction for Underwater Object Detection,” International Journal of Science and Engineering Applications, Jan. 2026.
6. X. Zhao, B. Zhou, and Y. Wang, “Class-incremental few-shot underwater object detection framework,” Scientific Reports, May 2026.
7. R. Li, C. Yu, S. Bao, Z. Li, and J. Yao, “YOLO-CAB: An Efficient Deep Learning-Based Underwater Object Detection Method for Autonomous Underwater Vehicles,” Mathematics, vol. 14, no. 11, p. 1927, 2026.
8. Y. Dong and K. Kang, “EUGNet: Evidential Uncertainty-Guided Network for underwater salient object detection,” Neurocomputing, vol. 697, p. 134179, Oct. 2026.
9. H. Yu, C. H. Vo, and C. Lee, “Detection-Guided Deep Unfolding for Joint Underwater Image Enhancement and Object Detection,” IEEE Access, vol. 14, pp. 17743–17759, 2026.
10. Y. Zhao, F. Sun, and X. Wu, “FEB-YOLOv8: A multi-scale lightweight detection model for underwater object detection,” PLOS ONE, vol. 19, no. 9, p. e0311173, 2024.