Perception Models in Harsh Domains: Detection and Depth in Underwater Images

Authors[1]
[1]University
Email: {authors}@uni.edu

***Abstract*—Depth estimation and object detection are relevant tasks in computer vision. This work presents a comparative analysis of the perception models for the mentioned tasks in harsh underwater scenarios. In underwater applications such as autonomous navigation, environmental monitoring, and infrastructure inspection, the image degradation can result from light attenuation, scattering, and turbidity. Addressing these challenges requires robust perception models that operate in the constrained conditions, motivating in the evaluation of state-of-the-art approaches. Depth estimation is evaluated using the Marigold, Depth Anything V2, and Depth Anything V3 models. Furthermore, for object detection, architectures such as YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv26, and RF-DETR are utilized, as well as specialized approaches, FeatEnHancer, AMSP-UOD, AquaFeat, and AquaFeat+. The quantitative and qualitative analysis of models' performance and of insights, integrated geometric and semantic information for the perception of robotic systems in ocean exploration.**

***Index Terms*—Harsh scenarios, Underwater images, Robotics perception, Object Detection, Depth-Estimation, Computer vision tasks**

I. Introduction

Monocular depth estimation and object detection are pillars of computer vision, enabling systems to infer three-dimensional geometry and semantic information from two-dimensional images [1], [2]. In underwater environments, these tasks are essential for applications such as autonomous navigation, marine ecosystem monitoring, and infrastructure inspection [3], [4]. However, the underwater medium presents unique challenges: light attenuation, scattering, and turbidity result in low contrast, severe color distortion, and loss of detail [5], [6]. These degradations not only hinder the extraction of reliable depth-inference features [7], [8] but also significantly degrade the performance of object detection models by obscuring key visual cues.

Underwater perception systems must be capable of discerning two fundamental aspects of a scene: what is present and where it is located. In this work, we address these challenges through object detection and monocular depth estimation. Rather than coupling these tasks into a single pipeline, we evaluate them independently to enable a clearer, more controlled assessment of foundation model capabilities. Our goal is to investigate the presence of a domain gap when these models are applied to harsh underwater environments, while comparing them with methods specifically designed for underwater scenarios.

In this context, this work performs a dual comparative study, evaluating state-of-the-art methods for both depth estimation and object detection in challenging underwater scenarios. Regarding depth estimation, we compare three distinct frontier approaches. Depth Anything V2 (DA2) [9] is based on a supervisory-driven large-scale learning strategy, and Depth Anything V3 (DA3) [10] uses a single transformer backbone and a depth-ray prediction target to achieve high generalization and spatial consistency. In contrast, Marigold [11] adopts an innovative strategy, Stable Diffusion, by repurposing diffusion-based generative models [12] to leverage vast visual priors for zero-shot depth estimation, which is particularly promising for underwater environments where ground-truth depth data is scarce.

In parallel, for the object detection task, we evaluated the performance of the YOLOv8m [13] and YOLOv10s [14] models under different enhancement regimes. Additionally, we include other representative detectors, such as YOLOv26 [15] and RF-DETR [16], to provide a broader perspective on recent advances in detection architectures. This study compares these reference models with state-of-the-art underwater methods, including FeatEnHancer [17], AMSP-UOD [18], AquaFeat [19], and AquaFeat+ [20]. These approaches aim to improve detection performance under challenging underwater conditions by enhancing images. The comparison focuses on a comprehensive set of metrics, including Precision, Recall, processing speed (FPS), and mean Average Precision at different thresholds ($mAP_{0.5}$ and $mAP_{0.5:0.95}$). The main contributions of this work are:

- Comparative evaluation of models in harsh underwater environments: this work investigates the depth estimation and object detection models under harsh underwater conditions, including low visibility, turbidity, scattering, and color distortion, and highlights their robustness in real-world scenarios.
- Quantitative and qualitative analysis in harsh underwater environments: evaluation by metrics, allowing the identification of strengths, limitations, and model behavior under these conditions.
- Geometric and semantic information for perception in harsh underwater environments: the study provides insights of depth estimation and object detection in such environments, contributing to the development of more robust and efficient perception systems for robotics and

ocean exploration.

II. Related Work about Underwater Perception

In this section, we review two perception tasks in computer vision applied to harsh domains, such as underwater environments. First, we explore advances in monocular depth estimation, fundamental for robotics navigation and orientation in subaquatic environments where real-world data are scarce. Furthermore, we discuss state-of-the-art detection models that handle the various visual degradations common to oceanic scenarios, ensuring accurate target identification in monitoring and inspection missions.

A. Depth Estimation Task

Deep learning and neural networks in the field of computational vision have included the task of estimating monocular depth [21], [22]. Then, self-supervised monocular depth estimation models have gained attention, because they reduce the need for ground-truth data, which are difficult to obtain in underwater environments [21], [23]. Furthermore, recent approaches enable learning visual representations without supervision. Thus, increasing the model's applicability across diverse conditions [24]. Monocular depth estimation is a task that involves applications of type 3D reconstruction, robotics, and autonomous navigation. However, underwater dominion introduces challenges such as light absorption, scattering, and contrast degradation, which negatively affect the model's performance. New approaches have been proposed to address these issues, for example, the Marigold model [11], which leverages diffusion-based image generators for depth estimation. Moreover, proposals of architectures based on Vision Transformers (ViT) [25] have demonstrated strong performance in dense predictions. In parallel, advances in self-supervision [26] have enabled effective training even with limited data. Nevertheless, deploying these models on resource-constrained embedded systems, particularly in robotics, remains a challenge [27], motivating research on lightweight neural networks and optimized models. Despite progress in underwater vision and embedded systems, integrating real-time depth-estimation models into underwater robots remains an active area of research.

B. Object Detection Task

In underwater environments, object detection faces significant challenges due to low visibility, light absorption, scattering effects, and color distortions. These degradations reduce contrast and obscure object boundaries, which impact the performance of detection models. Despite these difficulties, real-time detectors of the YOLO family have been adopted due to their trade-off between accuracy and speed, which makes them suitable for autonomous underwater vehicles and marine monitoring [28]. Recent advances in object detection have focused on improving both efficiency and robustness. Models such as YOLOv8 [13], YOLOv10 [14], and the more recent variant, YOLOv26 [15], introduce architectural refinements that improve feature representation, optimize training strategies, and increase inference speed. In parallel, transformer-based detectors, like RF-DETR [16], have emerged as a robust alternative, offering enhanced global modeling. FeatEnHancer [17] was designed to improve object detection in harsh environments by enhancing features specifics for the downstream task. These improvements are specifically relevant in underwater environments, where they are necessary to detect small objects and maintain real-time performance. Furthermore, there are works that explore complementary strategies to improve detection performance under adverse conditions, including data augmentation, domain adaptation, and integration with image enhancement methods. Methods such as AMSP-UOD [18], AquaFeat [19], and AquaFeat+ [20] incorporate additional modules to mitigate the impact of underwater degradations. However, their main contributions lie in supporting the detection task rather than redefining it.

III. Methodology

The following sections present the datasets used for the tasks of depth estimation and object detection, highlighting their characteristics and relevance for evaluation in the underwater domain. Next, the models applicable for each task are described, covering foundation and advanced approaches.

A. Datasets

1) Datasets of Depth Estimation: To evaluate the performance and generalization capacity of monocular depth estimation models, three underwater datasets were selected: SUIM [29], UIEB [30], and USIS10k [31]. SUIM allows us to verify whether the depth estimation maintains geometric consistency around complex entities, such as divers and artificial structures, which exhibit depth discontinuities relative to the seabed. The UIEB dataset was chosen because it contains real images with large variations in illumination and turbidity. Given that the underwater environment imposes challenges such as selective color absorption and light scattering, UIEB enables us to test whether the foundation models perform well despite these degradations. Finally, the inclusion of the USIS10K dataset allows us to test the stability of the models across a wide range of underwater domains, from shallow waters with high sunlight incidence to deep, turbid environments, thereby mitigating the risk of overfitting to specific conditions.

2) Datasets of Object Detection: In order to test the models' limitations for object detection, we selected a processed version of the FishTrack23 dataset [19], [32] and the TrashCan dataset [33]. The first one was chosen due to its high difficulty, as the set is a compilation of different datasets and environments that include both colorful and black-and-white images. Furthermore, it contains only one class of objects (fish), which are mostly occluded and covered by extreme haze and low luminosity. On the other hand, TrashCan was chosen as a complement to the FishTrack23 set; it includes multiple object classes and less extreme, but still realistic, underwater degradation. While FishTrack23 stresses the model under severe visibility constraints and single-class ambiguity,

TrashCan evaluates its ability to generalize across diverse object categories, shapes, and appearances commonly found in underwater environments. This combination enables a more comprehensive assessment: it isolates whether performance gains stem from improved feature representation under extreme conditions or from enhanced discriminability across classes, thereby exposing limitations that would not be evident in a single-class, highly specialized dataset.

B. Models

1) Models of Depth Estimation: The Marigold [11] model was selected to introduce a generative approach based on diffusion models, establishing itself in the literature as a reference in "fine-grained geometry." It leverages the prior knowledge of image synthesis models for depth tasks, promising spatial continuity with an unprecedented richness of detail, even in images with poor textures or severe noise. Its selection is justified by its ability to maintain structural coherence in low-visibility environments where traditional methods often fail, making it a high-fidelity benchmark for the detailed inspection of offshore infrastructures. Depth Anything V2 [9] was chosen due to its robust generalization, practical efficiency, and broad academic adoption as a fundamental vision model. Unlike models that require fine-tuning for specific domains, DA2 leverages massive training on unlabeled data across various benchmarks to offer a zero-shot capability that ignores typical visual artifacts of the underwater environment, such as haze and color loss. Its architecture allows for fast and reliable inference, meeting the low-latency requirements needed for the autonomous navigation of AUVs and ROVs in real time. Depth Anything V3 [10], the latest state-of-the-art model, evolves the previous architecture by focusing on extreme scale accuracy and edge refinement. It was selected because it claims to overcome depth resolution limitations on complex and small objects, which is critical for robotic manipulation tasks and precision isolation in underwater equipment. Together, these three models allow us to evaluate everything from the high-fidelity depth capabilities of diffusion models to the robustness and speed of next-generation transformer architectures, covering the entire spectrum of underwater robotics needs.

2) Models of Object Detection: The YOLO-based detectors [13]–[15], [34] were selected as primary baselines due to their strong and well-documented detection accuracy and real-time performance. As single-stage detectors, YOLO models unify localization and classification within a single forward pass, and enable low-latency inference that is especially suitable for robotics applications. Their widespread adoption in both academic literature and industry ensures reproducibility and comparability, while their evolution across versions (e.g., YOLOv8–YOLOv26) reflects state-of-the-art design choices in backbone, neck, and head architectures. Additionally, their robustness across diverse datasets and ease of deployment on embedded and resource-constrained platforms make them a representative and practically relevant benchmark for evaluating object detection methods in real-world robotic systems.
Additionally, the AquaFeat [19] and AquaFeat+ [20] models were chosen as they claim to improve YOLO's metrics in harsh environments when plugged into them, making them significantly better at Precision, Recall, F1-Score and mAPs, while keeping reasonable real-time efficiency. Finally, following the task-centric approach, we are also using the FeatEnHancer [17] method, as it is focused on enhancing images for downstream tasks in harsh environments.

To broaden our evaluation beyond convolutional single-stage detectors, we also included RF-DETR [16], a real-time Detection Transformer architecture. While YOLO models excel in fast, localized feature extraction, RF-DETR leverages global self-attention mechanisms and bipartite matching, offering a fundamentally different approach to object localization and classification. Although transformer-based models traditionally struggle with high latency, RF-DETR is specifically engineered to achieve state-of-the-art accuracy while maintaining real-time inference speeds compatible with underwater robotic deployments. Furthermore, its architecture is highly optimized for fine-tuning on custom datasets. Evaluating RF-DETR on the FishTrack23 and TrashCan datasets enables us to investigate whether its global context awareness provides greater resilience to severe haze, occlusion, and lighting variations than traditional detectors, thereby offering new insights into bridging the domain gap in harsh underwater perception.

The AMSP-UOD [18] model was chosen to represent specialized detection frameworks engineered specifically for underwater images. This architecture utilizes attentional multi-scale priors to address the unique challenges of subsea perception, such as wavelength-dependent light absorption and non-uniform scattering. Its inclusion in the benchmark allows for a comparison between general-purpose detectors and those featuring integrated restoration and enhancement priors. By evaluating AMSP-UOD, this study examines whether the integration of multi-scale attention mechanisms effectively mitigates the domain gap caused by underwater haze and low contrast, providing a benchmark for comparing environment-specific feature extraction against general-purpose approaches in robotic vision systems.

IV. Results and Analysis

This section presents a detailed analysis of the experimental results obtained for both monocular depth estimation and object detection tasks. To provide a comprehensive evaluation, we utilize a combination of quantitative metrics, focusing on geometric consistency for depth and localization accuracy for detection, alongside qualitative visual assessments. The experiments were designed to test the limits of current foundation models and specialized architectures across various underwater conditions, ranging from controlled visibility to extreme turbidity and light attenuation. By decoupling these tasks, we can isolate how specific environmental degradations impact spatial perception versus semantic identification in robotic systems.

TABLE I
QUANTITATIVE RESULTS OF DEPTH IN THE DATASET SUIM

Method	Type	Edge Align (↑)	Smoothness (↓)	Edge Aware Smoothness (↓)
Marigold [11]	Small	0.0595	0.0030	0.0029
Marigold [11]	Base	0.0620	0.0034	0.0032
Marigold [11]	Large	0.0632	0.0034	0.0032
DA2 [9]	Small	0.0783	0.0026	0.0025
DA2 [9]	Base	0.0803	0.0027	0.0026
DA2 [9]	Large	**0.0809**	0.0027	0.0026
DA3 [10]	Small	0.0298	0.0021	0.0021
DA3 [10]	Base	0.0291	**0.0020**	**0.0020**
DA3 [10]	Large	0.0362	0.0024	0.0023

TABLE II
QUANTITATIVE RESULTS OF DEPTH IN THE DATASET UIEB

Method	Type	Edge Align (↑)	Smoothness (↓)	Edge Aware Smoothness (↓)
Marigold [11]	Small	0.0567	0.0023	0.0022
Marigold [11]	Base	0.0620	0.0026	0.0024
Marigold [11]	Large	**0.0644**	0.0025	0.0024
DA2 [9]	Small	0.0563	0.0023	0.0021
DA2 [9]	Base	0.0587	0.0024	0.0022
DA2 [9]	Large	0.0590	0.0024	0.0022
DA3 [10]	Small	0.0235	**0.0018**	0.0018
DA3 [10]	Base	0.0249	**0.0018**	**0.0017**
DA3 [10]	Large	0.0266	0.0020	0.0020

TABLE III
QUANTITATIVE RESULTS OF DEPTH IN THE DATASET USIS10K

Method	Type	Edge Align (↑)	Smoothness (↓)	Edge Aware Smoothness (↓)
Marigold [11]	Small	0.0464	0.0020	**0.0019**
Marigold [11]	Base	0.0474	0.0026	0.0025
Marigold [11]	Large	0.0491	0.0026	0.0025
DA2 [9]	Small	0.0549	0.0023	0.0022
DA2 [9]	Base	0.0556	0.0024	0.0023
DA2 [9]	Large	**0.0558**	0.0024	0.0023
DA3 [10]	Small	0.0292	0.0020	**0.0019**
DA3 [10]	Base	0.0289	**0.0019**	**0.0019**
DA3 [10]	Large	0.0346	0.0021	0.0020

A. Depth Results

The results of the DA2 model consistently stood out in the Edge Align metric across all datasets. In the SUIM dataset Table I, the DA2's Large model size reached a maximum value of 0.0809, showing that this architecture has a superior ability to locate and preserve structural discontinuities in low-visibility environments, which suggests a stronger sensitivity to object boundaries and depth transitions even under scattering and color degradation. However, the DA3 model, although showing lower edge alignment values than its predecessor, dominated the smoothness metrics. In the UIEB Table II and USIS10K Table III datasets, the model obtained the lowest Smoothness indices (reaching 0.0018 and 0.0019), indicating the generation of more continuous depth maps with fewer abrupt artifacts, which is important for stability in offshore robotic systems, where noisy depth estimations can compromise navigation and control. This behavior suggests that DA3 prioritizes global consistency over fine structural detail, reducing high-frequency variations that may correspond to noise rather than meaningful geometry. Marigold model showed an intermediate and balanced performance. Notably, in the UIEB dataset Table II, Marigold Large outperformed the Depth Anything versions in the Edge Align criterion (0.0644), proving competitive in real underwater image scenarios with variations in color and turbidity, where both structural preservation and robustness to appearance shifts are required. This indicates that Marigold achieves a trade-off between edge fidelity and smoothness, avoiding excessive sharpening while still maintaining relevant geometric cues.

To illustrate these analyses, Figure 1 presents a qualitative comparison of the inferences. They show that edge preservation in DA2 yields sharper contours and more defined object boundaries, whereas the smoother predictions of DA3 produce surfaces with less visual noise and greater spatial coherence, even at smaller checkpoints, reinforcing the quantitative trends observed in the metrics.

B. Object Detection Results

The FishTrack [32] results in Table IV highlight clear differences in detection performance across models. Baseline YOLO models present limited recall (the ability to detect all relevant objects) and weaker overall detection quality, reflected in lower F1-scores (harmonic mean of precision and recall) and $\text{mAP}_{0.50:0.95}$ (mean Average Precision across multiple IoU thresholds). In contrast, RF-DETR achieves the highest precision (0.899), recall (0.629), and F1-score (0.740), indicating strong overall detection performance.

Methods tailored for underwater degradation further improve robustness: AMSP-UOD attains the best $\text{mAP}0.50$ (0.724) (Average Precision at IoU 0.50) and $\text{mAP}0.50:0.95$ (0.46), while AquaFeat establishes a more balanced regime, consistently improving recall while maintaining high precision compared to standard YOLO models. This leads to strong F1-scores across both YOLOv8 (0.729) and YOLOv10 (0.721) backbones; notably, AquaFeat (YOLOv8) achieves competitive recall (0.624), $\text{mAP}0.50$ (0.677), and $\text{mAP}0.50:0.95$ (0.421).

The TrashCan dataset detection results in Table V indicate competitive performance among standard YOLO models, alongside a noticeable decrease in effectiveness for restoration-specialized approaches. YOLOv8 achieves the highest $\text{mAP}0.50$ (0.720) (Average Precision at IoU 0.50) within the baseline group, while YOLOv9 attains the highest precision (0.797), reflecting more accurate positive detections. YOLOv10 and YOLOv11 emphasize localization and coverage, reaching the highest $\text{mAP}0.50:0.95$ (0.545) (mean Average Precision across multiple IoU thresholds) and recall (0.656) (ability to detect all relevant objects), respectively, whereas YOLOv26 achieves the highest F1-score (0.704) (harmonic mean of precision and recall), indicating the best balance between these metrics.

AMSP-UOD remains highly competitive, achieving strong $\text{mAP}0.50$ (0.723) and $\text{mAP}0.50:0.95$ (0.553). In contrast, RF-DETR exhibits a significant performance drop in this domain, with an F1-score of only 0.209. AquaFeat variants, including AquaFeat+ (YOLOv10), underperform relative to standard backbones, with AquaFeat (YOLOv8) reaching an $\text{mAP}_{0.50}$ of 0.657. This degradation is attributed to the relatively clear water conditions in the TrashCan dataset: as AquaFeat is designed to address severe underwater degradation and low-light scenarios, its restoration mechanisms may introduce unnecessary processing artifacts when applied to

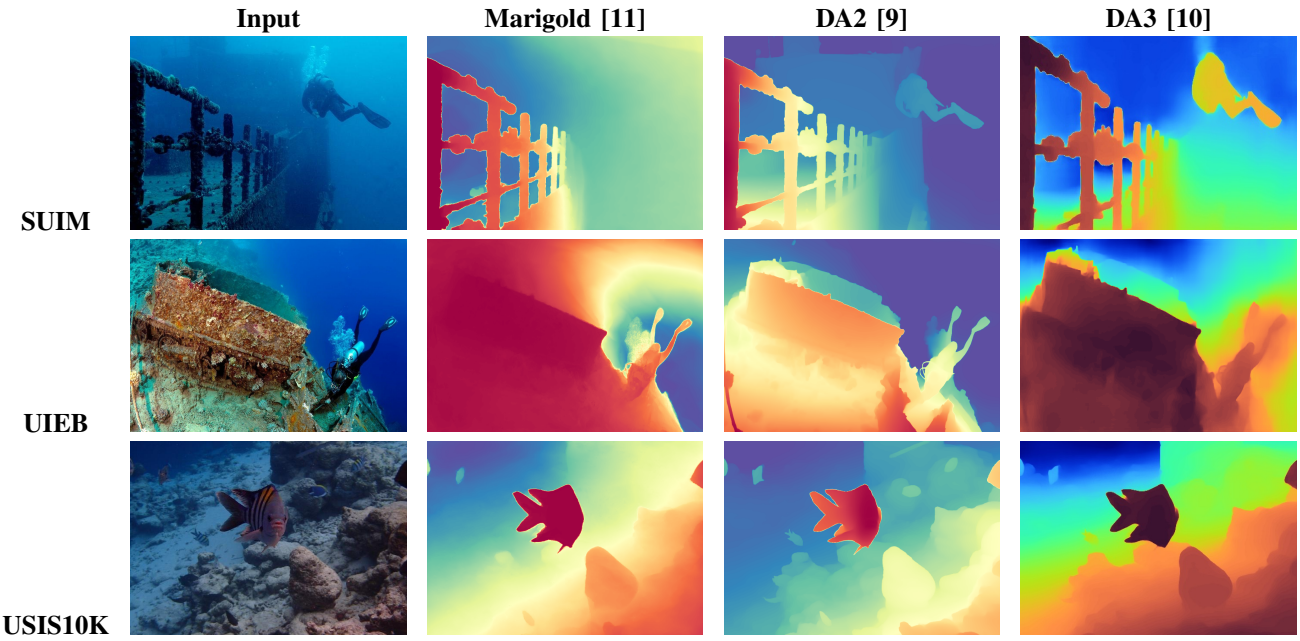

Fig. 1. Qualitative Results of Depth (Large checkpoints) using the models (columns) Marigold, DA2, and DA3 in the datasets (lines) SUIM, UIEB, and USIS10K.

TABLE IV
QUANTITATIVE RESULTS OF DETECTION IN FISHTRACK23 DATASET.

Method	Type	Precision↑	Recall↑	F1-Score↑	$mAP_{0.50}$ ↑	$mAP_{0.50:0.95}$ ↑
YOLOv8 [13]	medium	0.847	0.584	0.691	0.647	0.387
YOLOv10 [14]	small	0.777	0.549	0.643	0.592	0.325
YOLOv26 [15]	medium	0.801	0.519	0.630	0.583	0.338
FeatEnhancer (YOLOv8) [17]	medium	0.838	0.593	0.695	0.649	0.384
AMSP-UOD [18]	-	0.866	0.578	0.693	**0.724**	**0.46**
RF-DETR [16]	basic	**0.899**	**0.629**	**0.740**	0.674	0.404
AquaFeat (YOLOv8) [19]	medium	0.877	0.624	0.729	0.677	0.421
AquaFeat (YOLOv10)	small	0.859	0.621	0.721	0.676	0.421

TABLE V
QUANTITATIVE RESULTS OF DETECTION IN TRASHCAN DATASET.

Method	Type	Precision↑	Recall↑	F1-Score↑	$mAP_{0.50}$ ↑	$mAP_{0.50:0.95}$ ↑
YOLOv8 [13]	medium	0.795	0.630	0.703	0.720	0.538
YOLOv9 [14]	medium	**0.797**	0.623	0.700	0.710	0.531
YOLOv10 [14]	medium	0.777	0.637	0.700	0.714	0.545
YOLOv11 [14]	medium	0.747	**0.656**	0.699	0.683	0.517
YOLOv26 [15]	medium	0.789	0.636	**0.704**	0.706	0.538
AMSP-UOD [18]	Small	0.767	0.632	0.693	**0.723**	**0.553**
RF-DETR [16]	basic	0.222	0.203	0.209	0.244	0.145
AquaFeat (YOLOv8) [19]	medium	0.707	0.615	0.658	0.657	0.475
AquaFeat (YOLOv10)	medium	0.781	0.595	0.675	0.666	0.473
AquaFeat+ (YOLOv10) [20]	medium	0.732	0.600	0.659	0.635	0.451

high-visibility images, ultimately reducing both precision and recall compared to non-specialized models.

The qualitative analysis, presented in Figure 2, show the comparison between the Ground Truth and some of the models used in the FishTrack dataset. For instance, AquaFeat and AMSP-UOD were the only methods able to detect 4 out of the 5 objects in the scene. This pattern of AquaFeat and AMPS-UOD being the best models is also repeated in columns 2 and 3, where they were the only ones able to partially occluded or small fish, while YOLOv8m and FeatEnHancer struggled with some of the objects. The last column show that all models detected the two fish in the middle, but only AquaFeat was able to detect the one at the top.

V. CONCLUSION

Depth estimation and object detection are critical components of robotic perception in underwater applications, including autonomous navigation, environmental monitoring, and inspection. However, factors such as light attenuation, scattering, and turbidity introduce a significant domain gap that degrades visual quality and challenges model generalization. This work presented a comparative analysis of state-of-the-art methods for both tasks. The results indicate that Transformer-based models trained on large-scale data, such as the Depth Anything family, exhibit strong robustness in recovering underwater 3D structure. In particular, Depth Anything V2 achieves superior edge preservation and geometric detail, while Depth Anything V3 produces smoother and more consistent depth maps with reduced noise, highlighting a trade-off between structural fidelity and spatial coherence.

For object detection, the results show that general-purpose models remain competitive in clear-water conditions, while underwater-specialized approaches improve robustness under severe degradation but may introduce performance degradation when applied outside their target domain. These findings suggest that effective underwater perception requires balancing generalization and domain-specific adaptation, rather than relying solely on either strategy. Consequently, combining robust feature extraction with adaptive mechanisms for underwater conditions emerges as a key direction for improving perception systems in ocean exploration and offshore robotics.

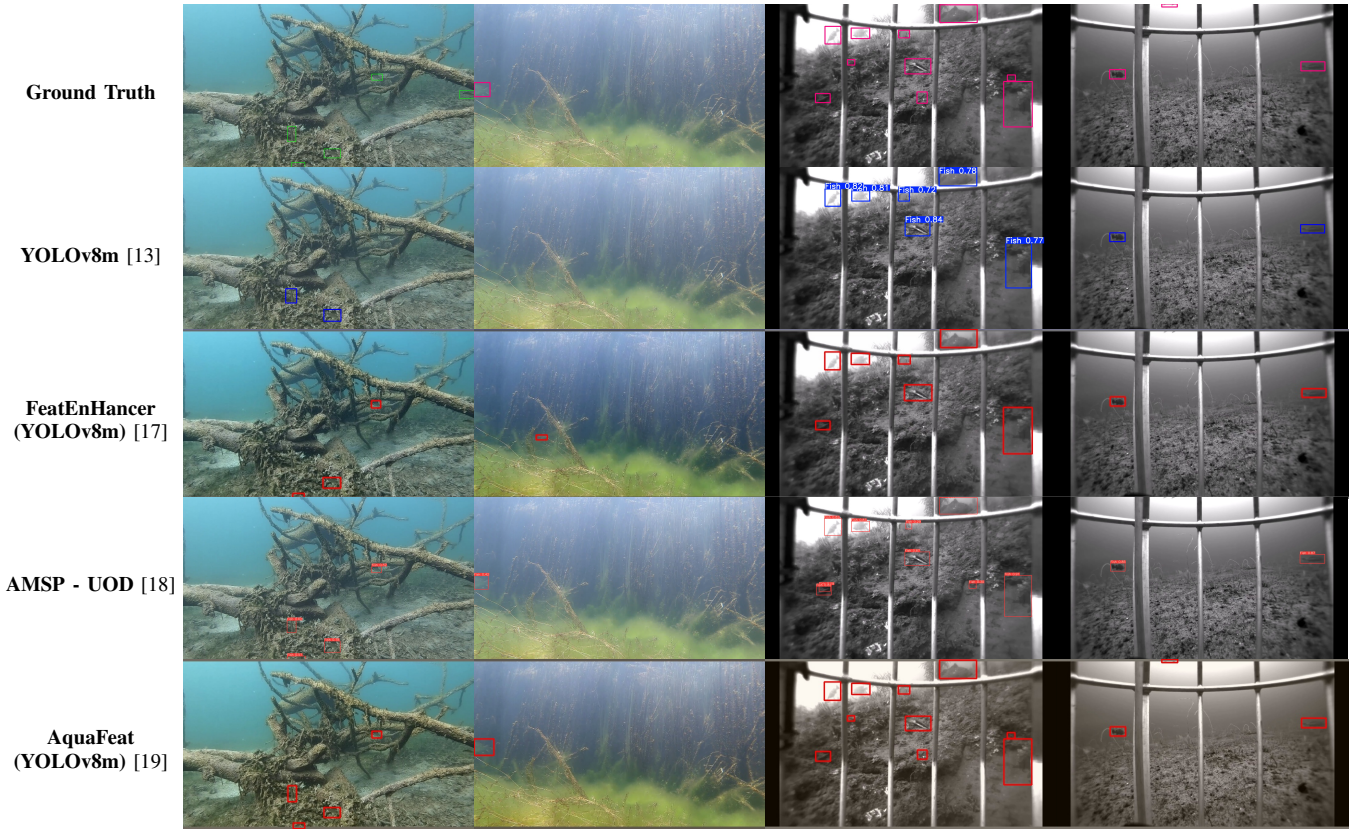

Fig. 2. Qualitative object detection comparison on FishTrack23 dataset [32]. Each column shows a different scene. The rows display the ground truth (top), results from competing methods (middle rows), and our proposed AquaFeat model (bottom).

REFERENCES

[1] U. Rajapaksha, F. Sohel, H. Laga, D. Diepeveen, and M. Bennamoun, "Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey," *ACM computing surveys*, vol. 56, no. 12, pp. 1–51, 2024.

[2] J. Liu, H. Ma, Y. Guo, Y. Zhao, C. Zhang, W. Sui, and W. Zou, "Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 11 162–11 168.

[3] M. A. Q. Quispe, J. D. G. Ramos, S. L. Brião, J. A. Díaz-Amado, and P. L. J. Drews-Jr, "Ssmss: A model of semantic segmentation for matching of satellite and sonar images," in *2025 Brazilian Symposium on Robotics (SBR) and 2025 Workshop on Robotics in Education (WRE)*. IEEE, 2025, pp. 237–242.

[4] M. M. dos Santos, G. C. de Oliveira, P. J. D. de Oliveira Evald, P. L. J. Drews-Jr, and S. S. da Costa Botelho, "Underwater robots localization using multi domain images: A survey," *Journal of Intelligent & Robotic Systems*, vol. 111, no. 2, p. 52, 2025.

[5] T. T. Schein, G. P. De Almeira, S. L. Brião, R. A. De Bem, F. G. De Oliveira, and P. L. Drews-Jr, "UDBE: Unsupervised diffusion-based brightness enhancement in underwater images," in *2024 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2024, pp. 664–670.

[6] F. Zhang, S. You, Y. Li, and Y. Fu, "Atlantis: Enabling underwater depth estimation with stable diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 11 852–11 861.

[7] J. Wang, J. Liu, D. Tang, W. Wang, W. Li, D. Chen, J. Chen, and J. Wu, "Scalable autoregressive monocular depth estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6262–6272.

[8] A. Obukhov, M. Poggi, F. Tosi, R. S. Arora, J. Spencer, C. Russel, S. Hadfield, R. Bowden, S. Wang, Z. Ma *et al.*, "The fourth monocular depth estimation challenge," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6182–6195.

[9] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," 2024. [Online]. Available: https://arxiv.org/abs/2406.09414

[10] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," *arXiv preprint arXiv:2511.10647*, 2025.

[11] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 9492–9502.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[13] Ultralytics, "YOLO-v8," GitHub repository, 2023. [Online]. Available: https://github.com/autogyro/yolo-V8

[14] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han *et al.*, "Yolov10: Real-time end-to-end object detection," *NeurIPS*, vol. 37, pp. 107 984–108 011, 2024.

[15] Ultralytics, "Ultralytics yolov26," https://github.com/ultralytics/ultralytics, 2025.

[16] I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri, "Rf-detr: Neural architecture search for real-time detection transformers," 2025. [Online]. Available: https://arxiv.org/abs/2511.09554

[17] K. A. Hashmi, G. Kallempudi, D. Stricker, and M. Z. Afzal, "Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision," in *IEEE/CVF ICCV*, 2023, pp. 6725–6735.

[18] J. Zhou, Z. He, K.-M. Lam, Y. Wang, W. Zhang, C. Guo, and C. Li, "AMSP-UOD: When vortex convolution and stochastic perturbation

meet underwater object detection," in *AAAI*, vol. 38, no. 7, 2024, pp. 7659–7667.

[19] E. d. C. Silva, T. T. Schein, S. L. Brião, G. L. M. Costa, F. G. Oliveira, G. P. Almeida, E. L. Silva, S. d. S. Devincenzi, K. d. S. Machado, and P. L. J. Drews-Jr, "Aquafeat: A features-based image enhancement model for underwater object detection," in *SIBGRAPI*, 2025. [Online]. Available: http://urlib.net/ibi/8JMKD2USNRW34M/4E8QRCB

[20] E. C. Silva, T. T. Schein, J. D. G. Ramos, E. L. Silva, S. L. Brião, F. G. Oliveira, and P. L. J. Drews, "Aquafeat+: an underwater vision learning-based enhancement method for object detection, classification, and tracking," in *ICAR*, 2025, pp. 432–437.

[21] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[22] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," pp. 611–629, 8 2018.

[23] D. Wofk, R. Ranftl, M. Müller, and V. Koltun, "Monocular visual-inertial depth estimation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 6095–6101.

[24] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[25] A. Radford *et al.*, "Vision transformers for dense prediction tasks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 5396–5406.

[26] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[27] X. Zhou, H. Liu, C. Shi, and J. Liu, "Model design and compression," *Deep Learning on Edge Computing Devices*, pp. 39–58, 2022.

[28] J. Chen and M. J. Er, "Dynamic yolo for small underwater object detection," *UMT-AIR*, vol. 57, no. 7, pp. 1–23, 2024.

[29] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2020.

[30] 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.

[31] S. Lian, Z. Zhang, H. Li, W. Li, L. T. Yang, S. Kwong, and R. Cong, "Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset," *arXiv preprint arXiv:2406.06039*, 2024.

[32] M. Dawkins, J. Prior, B. Lewis *et al.*, "Fishtrack23: An ensemble underwater dataset for multi-object tracking," in *IEEE/CVF WACV*, 2024, pp. 7167–7176.

[33] J. Hong, M. Fulton, and J. Sattar, "Trashcan: A semantically-segmented dataset towards visual detection of marine debris," *arXiv preprint arXiv:2007.08097*, 2020. [Online]. Available: https://arxiv.org/abs/2007.08097

[34] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *European conference on computer vision*. Springer, 2024, pp. 1–21.