UWSegDepth: Semantic-Aware Object-Level Depth Estimation in Underwater Scenes

Pin-Chi Pan (潘品齊)[1] *and Soo-Chang Pei* (貝蘇章)[2]

[1] Graduate Institute of Communication Engineering,
National Taiwan University, Taiwan,
E-mail: r12942103@ntu.edu.tw,
[2] Department of Electrical Engineering,
National Taiwan University, Taiwan,
E-mail: peisc@ntu.edu.tw

(a) (b) (c) (d)

Fig. 1. Overview of the UWSegDepth pipeline. Given an input image **(a)**, instance masks **(b)** are generated by BARIS-ERA [10], and a depth map **(c)** is estimated using TRUDepth [9] enhanced with SADDER. UWSegDepth then computes the average depth for each segmented object to produce the final output **(d)**.

Abstract

Accurate depth estimation and object recognition are essential for underwater tasks such as navigation, habitat monitoring, and exploration. However, light scattering, color attenuation, and water turbidity make it difficult to estimate depth from a single image. We propose SADDER (Segmentation-Augmented Differential Depth Estimation Regressor), a lightweight module that improves depth estimates by correcting residual errors guided by instance segmentation. We also introduce UWSegDepth, a straightforward post-processing method that calculates the average depth of each segmented object, adding object-level structure to pixel-wise predictions. Experiments on the FLSea benchmark show clear improvements, especially in shallow and murky conditions. The proposed method provides reliable depth estimates and clear object-level information, making it suitable for practical underwater applications.

Keywords: *Underwater Scenes, Depth Estimation, Instance Segmentation, Object-Level Depth Estimation.*

1. INTRODUCTION

Accurate perception of underwater environments is critical for a wide range of marine applications, including environmental monitoring [1], autonomous navigation [2], and 3D scene reconstruction [3]. In particular, monocular depth estimation and instance segmentation are key to interpreting scene structure and identifying objects [4, 5]. However, underwater vision remains highly challenging due to environment-specific issues such as light scattering, color attenuation, and turbidity, which can severely affect both depth accuracy and object recognition.

In underwater depth estimation tasks, traditional methods based on image formation models (IFMs) [6, 7] often struggle with varying environmental conditions and depend on manually tuned physical parameters. More recent learning-based methods, such as UDepth [8], adopt classification-based depth estimation with adaptive binning strategies. These methods show promising results by using several lightweight architectures and incorporating wavelength-sensitive features. Other approaches, like TRUDepth [9], integrate sparse depth priors from visual triangulation to improve scale consistency, even when training data is limited or lacks depth calibration. Together, these developments reflect

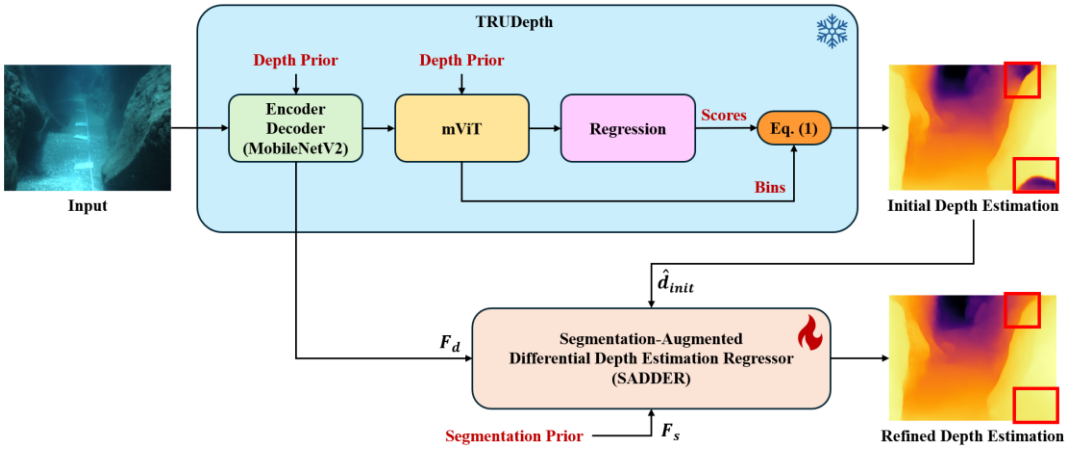


Fig. 2. Overview of the proposed Segmentation-Augmented Differential Depth Estimation Regressor (SADDER). The proposed method refines TRUDepth predictions using SADDER, which leverages segmentation features F_s and deep features F_d to correct the initial depth $\hat{d}_{init}$, yielding sharper and more accurate depth maps.

steady progress in addressing the unique challenges of underwater depth estimation.

Nevertheless, such methods still struggle in regions with low visibility or complex geometry, where monocular information alone is unreliable. While depth estimation captures scene structure, recognizing and separating objects is equally important for tasks such as object interaction and spatial reasoning. This highlights the need to incorporate segmentation information into the depth estimation process. Recent work in underwater instance segmentation, such as BARIS-ERA [10], addresses lighting variability and environmental shifts by combining structural refinement with fixed domain-adaptive features, allowing for more reliable mask predictions under challenging conditions.

To address the limitations of these methods and bring together geometric and semantic information, we propose a two-stage framework for underwater object-level depth estimation (Fig. 1). First, we introduce the Segmentation-Augmented Differential Depth Estimation Regressor (SADDER), a lightweight refinement module that uses instance segmentation features to correct residual errors in initial depth maps. Built on TRUDepth, SADDER improves results in occluded areas and along object boundaries by applying detailed corrections guided by segmentation priors. Second, we present UWSegDepth, an object-level method that combines SADDER-refined depth maps with instance masks from BARIS-ERA, assigning an average depth value to each segmented object. We evaluate our approach on the FLSea dataset [11], a diverse benchmark of shallow-water scenes. Our method consistently performs better than existing baselines in both shallow and full-depth ranges, producing accurate depth estimates and clear object separation from a single underwater image. The main contributions of this work are summarized as follows:

1) We propose SADDER, a segmentation-augmented depth refinement module that improves the depth estimates produced by TRUDepth under challenging underwater conditions.
2) We introduce UWSegDepth, a lightweight method that combines SADDER-refined depth with BARIS-ERA instance masks to estimate object-level depth without additional training.
3) We demonstrate significant improvements on the FLSea benchmark, both quantitatively and qualitatively, under varying levels of visibility and scene complexity.

2. RELATED WORK

The integration of depth estimation and instance segmentation has demonstrated strong potential for enhancing scene understanding, especially in challenging environments. However, underwater applications remain relatively underexplored due to the unique visual degradation caused by light attenuation, scattering, and low texture. In this section, we review two key areas of related work: underwater depth estimation, and object-level depth estimation in degraded visual environments.

2.1. Underwater Depth Estimation

Monocular depth estimation in underwater scenes presents persistent challenges due to limited visibility, optical distortions, and a lack of ground-truth data. Traditional methods based on image formation models are sensitive to physical assumptions and often fail under changing environmental conditions. While recent learning-based methods have shown notable progress, accurate prediction remains difficult in complex or degraded settings. For instance, classification-based models incorporating spectral sensitivity and adaptive discretization, such as UDepth [8] have improved robustness in many scenarios. TRUDepth [9] further addresses scale ambiguity by introducing sparse depth constraints based on visual triangulation. Despite these advances, most existing approaches rely solely on image features without integrating object-level information, which often leads to blurred depth boundaries and poor generalization in scenes with occlusion or fine-grained structure.

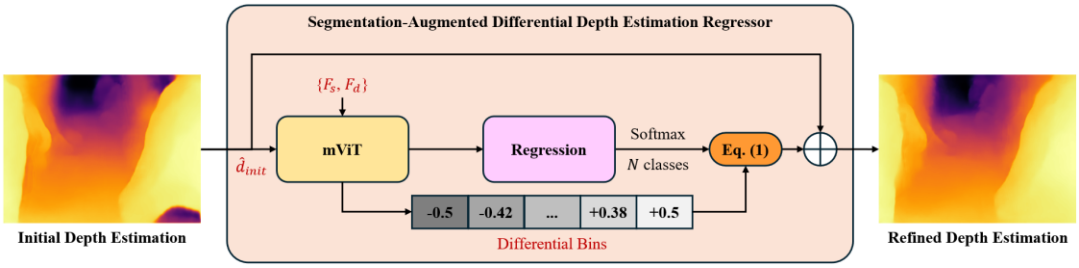


Fig. 3. Architecture of the proposed Segmentation-Augmented Differential Depth Estimation Regressor (SADDER). The model refines an initial depth map by predicting a residual correction using segmentation features.

2.2. Object-Level Depth Estimation

To bridge semantic and geometric understanding, recent studies in terrestrial and low-light vision have explored the integration of depth estimation and instance segmentation. Methods such as Panoptic-DepthLab [4] utilize a unified architecture to jointly predict segmentation masks and depth maps, facilitating more accurate 3D scene reconstruction and foreground-background separation. In low-light environments, models like Panoptic-LMFFNet [5] incorporate image enhancement techniques and domain adaptation to remain effective under varying illumination. These methods often apply straightforward strategies, such as averaging depth values within instance masks, to assign object-level depth estimates. The results suggest that integrating segmentation with geometric data enhances occlusion handling, object-level distinction, and spatial understanding. However, their application to underwater environments remains underexplored.

3. PROPOSED METHOD

This section presents a framework for pixel-wise and object-level depth estimation in underwater environments. The method aims to improve monocular depth estimation through semantic information and extend depth inference to the object level. The framework comprises two stages: segmentation-augmented depth refinement using SADDER, and object-level depth estimation based on dense predictions.

3.1. Segmentation-Augmented Depth Estimation

We build upon TRUDepth [9], a classification-based monocular depth estimation model that employs adaptive binning and a compact vision transformer architecture. Although TRUDepth achieves competitive performance under challenging underwater conditions, it often fails to preserve geometric boundaries, particularly in the presence of light scattering or occlusion.

To address these issues, we propose a novel depth refinement module, SADDER (Segmentation-Augmented Differential Depth Estimation Regressor), which focuses on predicting the residual between the initial depth estimate and the ground truth. As shown in Fig. 2, SADDER takes three inputs: the initial predicted depth map $\hat{d}_{init}$, segmentation features F_s, and deep features F_d from the encoder-decoder backbone. By incorporating instance segmentation features into the refinement process, the method enhances estimation accuracy, particularly in regions affected by occlusion or boundary ambiguity.

As illustrated in Fig. 3, these features are concatenated and fed into a lightweight modified vision transformer (mViT), which predicts the likelihood of each pixel belonging to one of several differential depth bins. These bins represent the deviation of depth values from the initial estimate, encompassing both positive and negative offsets. Based on the quantitative results of TRUDepth [9], we observed that the per-pixel depth error typically falls within 0.5 meters. Therefore, in our implementation, the differential depth Δd is constrained to a narrow interval of approximately $[-0.5, +0.5]$ meters. meters. This bounded range enables SADDER to refine depth estimates effectively while reducing the risk of large corrections that may lead to instability.

Subsequently, a softmax layer transforms the predicted logits into probability scores p_i over N discrete bins, and the differential depth Δd is computed as a weighted sum of differential bin centers c_i using these probabilities:

$$\Delta d(x,y) = \sum_{i=1}^{N} c_i * p_i(x,y) \qquad (1)$$

This design enables precise, localized refinement of the depth map by incorporating both segmentation priors and spatial features, particularly in regions where monocular depth estimation is intrinsically uncertain. SADDER thus generates a differential depth map Δd, which is applied as a correction to the initial estimate. The final depth output is computed as:

$$\hat{d}_{refined} = \hat{d}_{init} + \Delta d \qquad (2)$$

This formulation allows the model to concentrate on correcting systematic errors, particularly near object boundaries or in visually complex regions where monocular information may be insufficient. During training, the parameters of the original TRUDepth model are kept frozen. Only the SADDER module, including its convolutional layers and the modified vision transformer (mViT), is trained. This training strategy reduces computational cost while ensuring that SADDER refines, rather than overrides, the baseline depth estimates.

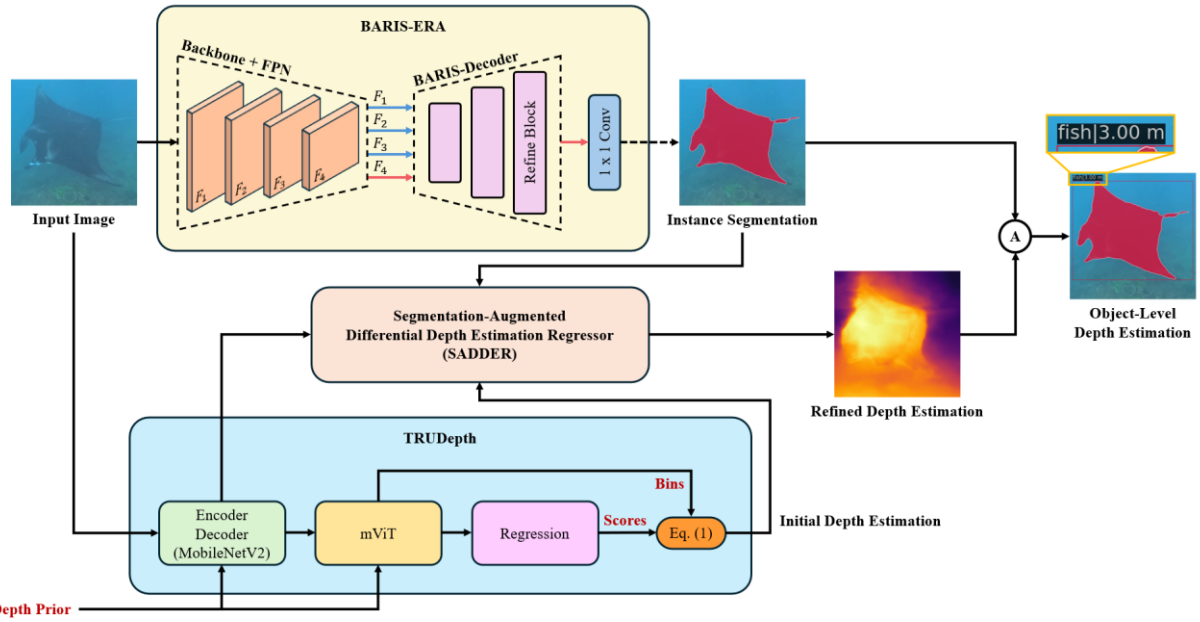


Fig. 4. Overview of the UWSegDepth pipeline. Instance masks from BARIS-ERA and refined depth from SADDER are combined to generate object-level depth estimates via spatial averaging.

To train SADDER effectively, we adopt a two-part loss function that balances metric precision and structural consistency, following the design in TRUDepth [9]. The total loss is:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{\text{RMSE}} + \lambda_2 \cdot \mathcal{L}_{\text{SILog}}\,, \tag{3}$$

where $\lambda_1 = 0.3$ and $\lambda_2 = 0.6$ control the contribution of each term. The RMSE term encourages accurate per-pixel predictions:

$$\mathcal{L}_{\text{RMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{d}_i - d_i\right)^2}, \tag{4}$$

while the SILog term ensures scale-invariant consistency:

$$\mathcal{L}_{\text{SILog}} = \beta \cdot \sqrt{\frac{1}{N}\sum_{i=1}^{N} {g_i}^2 - \frac{\lambda}{N^2}\left(\sum_{i=1}^{N} g_i\right)^2}\,, \tag{5}$$

where $g_i = \log(\hat{d}_i) - \log(d_i)$, $\beta = 10$, $\lambda = 0.85$. This formulation is particularly effective in underwater environments, where absolute depth may be unreliable but structural information remains valuable.

3.2. UWSegDepth: Object-Level Depth Estimation

While SADDER improves pixel-wise depth estimation, many downstream applications require object-level spatial understanding rather than per-pixel depth alone. Inspired by recent advances in terrestrial vision [4, 5], we adopt a straightforward aggregation scheme that combines refined depth maps with instance segmentation results to estimate object-level depth. This approach facilitates coherent spatial interpretation of individual objects and their relative positions, which is particularly advantageous for underwater tasks such as autonomous navigation, ecological monitoring, and marine robotics.

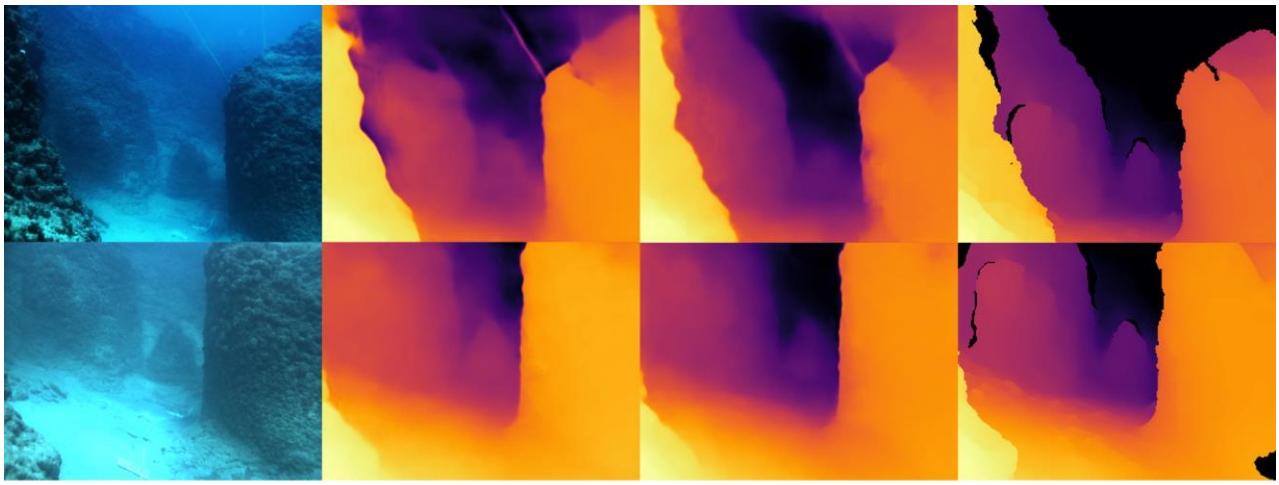
To achieve object-level depth estimation in underwater scenes, we propose UWSegDepth, an object-level depth estimation framework that integrates the outputs of two previously introduced models: the instance segmentation model BARIS-ERA [10] and the depth estimation model TRUDepth refined with SADDER. Instead of modifying existing architectures or employing joint training, the proposed method applies a post-processing procedure that combines semantic and geometric information at the instance level.

As illustrated in Fig. 4, given an input image, instance masks are obtained from BARIS-ERA, and a dense depth map is generated by TRUDepth with SADDER. For each object, its depth is computed by averaging the pixel-wise depth values within its corresponding segmentation mask:

$$d_k = \frac{1}{|\mathcal{M}_k|}\sum_{(x,y)\in\mathcal{M}_k} \hat{d}(x,y)\,, \tag{6}$$

Here, $\mathcal{M}_k$ denotes the binary mask of the k-th instance and $\hat{d}(x,y)$ is the predicted depth at pixel (x,y). Instances with insufficient valid pixels are excluded to ensure numerical stability. This averaging process is both computationally efficient and less sensitive to local noise, producing spatially consistent and metrically reliable object-level depth estimates.

The final output provides each segmented object with both semantic labels and estimated distances, facilitating structured scene interpretation. This modular design allows for straightforward integration with other segmentation or depth estimation methods, offering a practical approach for object-level depth analysis in underwater environments.



(a) Input (b) TRUDepth **(c) Ours** (d) Ground Truth

Fig. 5. Qualitative comparison of depth estimation results on the FLSea dataset, showing that our method yields more reliable predictions under challenging underwater conditions.

4. EXPERIMENTS

For underwater depth estimation, we conduct a series of experiments on the FLSea dataset [11] to evaluate the effectiveness of our proposed SADDER module and UWSegDepth framework. This section presents our experimental setup, quantitative and qualitative results.

4.1. Dataset and Evaluation Metrics

Dataset. The FLSea dataset serves as the benchmark for our evaluation. It contains 22,451 RGB-depth image pairs collected from 12 shallow-water environments, with depth typically ranging up to 10 meters. We follow the standard split introduced in prior work [9], using 10 sites for training and 2 unseen sites (*u_canyon* and *sub_pier*) for testing to ensure spatial separation. Depth maps are generated using Agisoft Metashape, a photogrammetric tool that reconstructs metric depth from multi-view underwater footage. To preserve natural underwater degradation characteristics, we use the raw RGB images without color correction.

Evaluation Metrics. To evaluate the quality of the predicted depth maps, we adopt a set of standard quantitative metrics, following prior work [9]. All metrics are computed over the test set and reported as mean or median accuracy. The error metrics between the predicted $\hat{d}$ and ground truth depth d are calculated as follows:

$$MARE(\hat{d}_i, d_i) = \frac{1}{N}\sum_{i}^{N}\frac{|\hat{d}_i - d_i|}{|d_i|}, \quad (7)$$

$$RMSE_{Linear}(\hat{d}_i, d_i) = \sqrt{\frac{1}{N}\sum_{i}^{N}(\hat{d}_i - d_i)^2}, \quad (8)$$

Table 1. Quantitative comparison with state-of-the-art methods on the FLSea test set, reported using **MEAN** accuracy.

Range	Method	RMSE (Linear)	RMSE (Log)	RMSE (SILog)	MARE
Full range	TRUDepth [9]	0.4711	0.0939	0.0924	0.0414
	+ **Ours**	**0.4424**	**0.0885**	**0.0866**	**0.0396**
	Δ_{Ours}	**6.09%**	**5.75%**	**6.28%**	**4.35%**
$d < 5\,m$	TRUDepth [9]	0.2108	0.0678	0.0667	0.0339
	+ **Ours**	**0.1877**	**0.0635**	**0.0612**	**0.0334**
	Δ_{Ours}	**10.96%**	**6.34%**	**8.25%**	**1.47%**
$d < 1\,m$	TRUDepth [9]	0.1119	0.0943	0.0584	0.0861
	+ **Ours**	**0.0895**	**0.0833**	**0.0510**	**0.0736**
	Δ_{Ours}	**20.02%**	**11.66%**	**12.67%**	**14.52%**

$$RMSE_{Log}(\hat{d}_i, d_i) = \sqrt{\frac{1}{N}\sum_{i}^{N}(\log(\hat{d}_i) - \log(d_i))^2}, \quad (9)$$

$$RMSE_{SILog}(\hat{d}_i, d_i) = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(\log(\hat{d}_i) - \log(d_i) + \alpha(\hat{d}_i, d_i)\right)^2}, \quad (10)$$

where $\alpha(\hat{d}_i, d_i) = \frac{1}{N}\sum_{i}^{N}(\log(d_i) - \log(\hat{d}_i))$ is the term that makes the error scale invariant [17].

4.2. Implementation Details

We build upon the TRUDepth architecture using a MobileNetV2 [12] backbone and a pretrained vision transformer. The SADDER module is trained using PyTorch [13] on an NVIDIA Titan RTX GPU. Following [14], we initialize SADDER with a zero-initialization strategy to stabilize learning. The AdamW optimizer [15] is used with a base learning rate of 10^{-5} and exponential decay (factor 0.9). Batch size is set to 6. Data augmentations include horizontal flips, brightness scaling, color jitter, and depth scaling. For supervision, 200 sparse prior points are randomly sampled from multi-view consistency to guide depth learning, accounting for 0.26% of pixels at 320×240 resolution.

Table 2. Quantitative comparison with state-of-the-art methods on the FLSea test set, reported using **MEDIAN** accuracy.

Range	Method	RMSE (Linear)	RMSE (Log)	RMSE (SILog)	MARE
Full range	TRUDepth [9]	0.4235	0.0835	0.0826	0.0387
	+ **Ours**	**0.3608**	**0.0771**	**0.0757**	**0.0365**
	Δ_{Ours}	**14.80%**	**7.66%**	**8.35%**	**5.68%**
$d < 5\,m$	TRUDepth [9]	0.1933	0.0610	0.0605	**0.0306**
	+ **Ours**	**0.1658**	**0.0578**	**0.0557**	**0.0306**
	Δ_{Ours}	**14.23%**	**5.25%**	**7.93%**	**0%**
$d < 1\,m$	TRUDepth [9]	0.0416	0.0442	**0.0326**	0.0298
	+ **Ours**	**0.0392**	**0.0419**	**0.0326**	**0.0289**
	Δ_{Ours}	**5.77%**	**5.20%**	**0%**	**3.02%**

4.3. Quantitative Comparison

We evaluate our method on the FLSea test set against the baseline TRUDepth [9] across three depth intervals: *Full range*, $d < 5\,m$, and $d < 1\,m$. We adopt four standard metrics: RMSE (Linear), RMSE (Log), RMSE (SILog), and MARE. Both mean and median performance are reported in Tables 1 and 2, respectively.

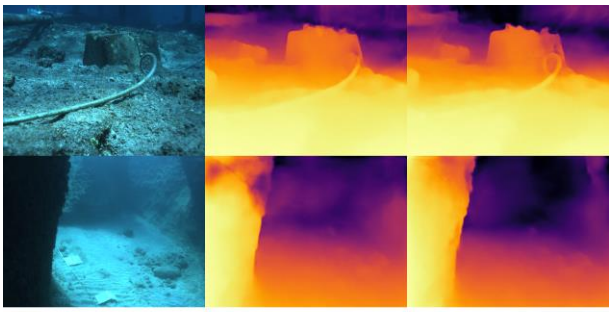
As shown in Table 1, our method significantly improves upon TRUDepth, particularly in shallow regions where accurate estimation is most challenging. We **achieve up to 20.02% improvement in RMSE (Linear)** for depths less than 1 meter. The results indicate that our segmentation-augmented refinement strategy effectively corrects boundary errors and enhances robustness in turbid conditions. Table 2 confirms the stability of our approach. Our model **improves the median RMSE (Linear) by 14.80%** in the *Full range*, with consistent gains across all metrics.

A simple approach to improving underwater depth estimation is to cascade an underwater image enhancement (UIE) model with a depth estimation network. To further examine the effectiveness of our SADDER module, we compare it with this UIE-based strategy using CCL-Net [16]. As shown in Table 3, SADDER consistently outperforms the UIE baseline across all depth intervals. While UIE may enhance image appearance, it does not reliably improve depth accuracy. In contrast, SADDER directly refines depth predictions using segmentation priors, leading to more precise and metrically consistent results.

These results demonstrate that SADDER is a lightweight and effective enhancement for underwater monocular depth estimation, achieving substantial improvements with minimal increase in model complexity and without requiring joint training.

4.4. Qualitative Results

Fig. 5 and Fig. 6 present visual comparisons of depth predictions from TRUDepth and our SADDER-augmented model. Our method yields sharper boundaries and more structurally consistent depth transitions, especially in regions affected by occlusion, scattering, or low visibility. Compared to the baseline, SADDER better preserves geometric contours and improves prediction stability in degraded underwater environments, demonstrating the effectiveness of segmentation-augmented refinement.



(a) Input (b) TRUDepth **(c) Ours**

Fig. 6. More qualitative comparison of depth estimation results on the FLSea dataset.

Table 3. Comparison between our SADDER and UIE-based strategy for enhancing underwater depth estimation on the FLSea test set, reported using **MEAN** accuracy.

Range	Method	RMSE (Linear)	RMSE (Log)	RMSE (SILog)	MARE
Full range	CCL-Net [16]	0.5015	0.1001	0.0985	0.0448
	SADDER	**0.4424**	**0.0885**	**0.0866**	**0.0396**
$d < 5\,m$	CCL-Net [16]	0.2249	0.0731	0.0716	0.0372
	SADDER	**0.1877**	**0.0635**	**0.0612**	**0.0334**
$d < 1\,m$	CCL-Net [16]	0.1119	0.0943	0.0584	0.0861
	SADDER	**0.0895**	**0.0833**	**0.0510**	**0.0736**

4.5. Object-Level Depth Estimation Results

We evaluate our UWSegDepth framework by assigning mean depth values to segmented objects from BARIS-ERA, enabling object-level spatial understanding without additional training. As shown in Fig. 7 and Fig. 8, this lightweight fusion yields interpretable depth annotations across diverse underwater targets such as fish, reefs, and divers. Compared to raw depth maps, our object-level outputs provide clearer semantic boundaries and structured spatial information. Despite the simplicity of mean pooling, it proves effective for most scenarios and robust to noise. This integration of SADDER and instance fusion enhances both metric accuracy and semantic interpretability, supporting downstream tasks like obstacle assessment and marine habitat analysis.

4.6. Computational Efficiency

We evaluate the runtime of each component in the UWSegDepth pipeline to assess its practicality for deployment. As shown in Table 4, all measurements were conducted on an NVIDIA Titan RTX GPU with a batch size of 1, and results are reported in frames per second (FPS). The full pipeline, including instance segmentation, sparse depth prior generation, TRUDepth with SADDER refinement, and object-level post-processing, runs at 0.327 FPS. While the current implementation is not suited for real-time applications, it remains feasible for offline processing or tasks where frame-wise analysis is

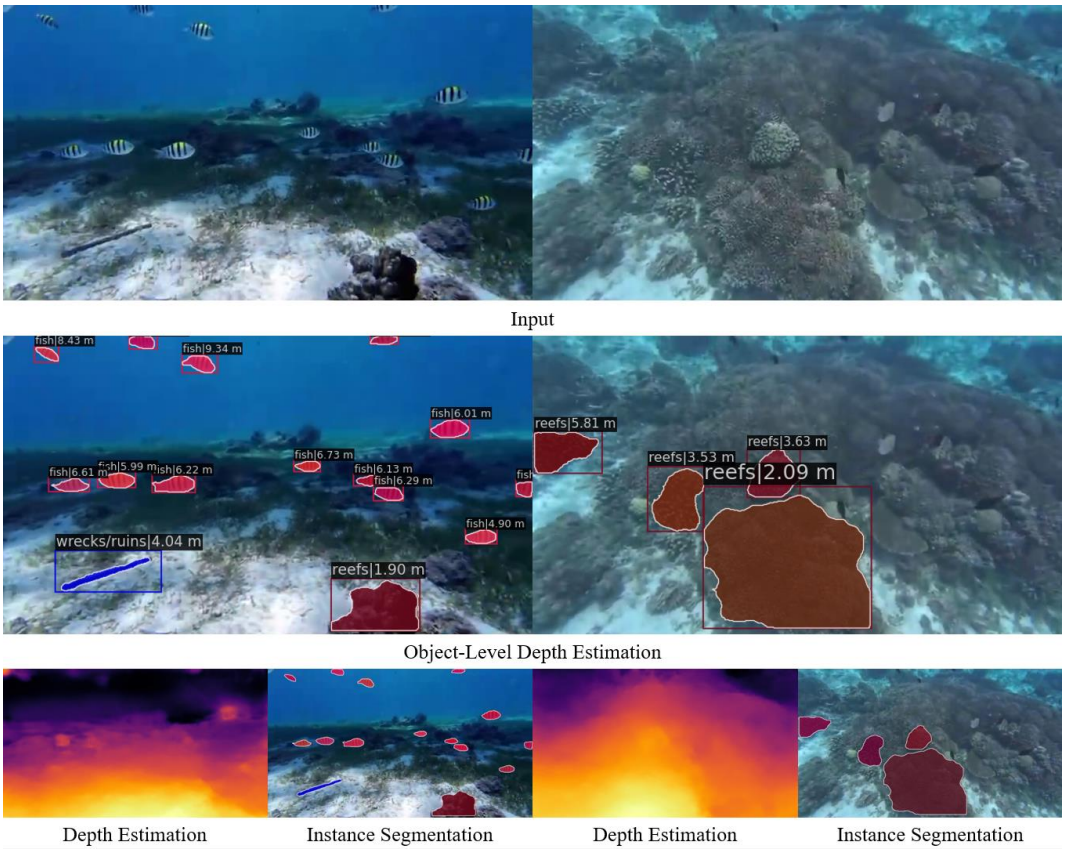


Fig. 7. Qualitative results of our UWSegDepth. The integration of instance segmentation and depth prediction allows for interpretable object-level depth estimation in underwater environments.

sufficient. To further improve computational efficiency, future work may investigate model compression techniques and optimized GPU-based implementations.

5. CONCLUSION

In this work, we present a unified framework for underwater depth estimation and object-level scene understanding. We first introduce the Segmentation-Augmented Differential Depth Estimation Regressor (SADDER), a lightweight refinement module that improves monocular depth estimation by predicting residual errors informed by instance segmentation features. Built upon TRUDepth, SADDER enhances accuracy, particularly near object boundaries and in visually degraded regions, without substantially increasing model complexity.

To bridge the gap between pixel-wise depth and object-level reasoning, we propose UWSegDepth, an object-level depth estimation framework that integrates SADDER-refined depth maps with instance masks from BARIS-ERA. By applying a simple mean aggregation within each mask, UWSegDepth assigns representative depth values to segmented objects. This efficient post-processing approach improves geometric consistency and semantic clarity, enabling accurate and structured interpretation of underwater scenes.

Table 4. Runtime Analysis of UWSegDepth Components.

Method	FPS
Perform Instance Segmentation	2.578
Get Depth Prior	0.376
Convert Depth Prior To Sparse Features	8.042
Perform Depth Estimation	66.281
UWSegDepth (Total)	0.327

Experimental results on the FLSea dataset demonstrate the effectiveness of the proposed approach. SADDER consistently improves depth accuracy under challenging conditions, and the integration with instance segmentation enables coherent object-level depth estimation. Although the current runtime limits real-time applicability, the framework remains suitable for offline or low-frame-rate scenarios. The modular design offers a robust and extensible solution for underwater perception, with future work focusing on accelerating inference while maintaining performance.

REFERENCES

[1] F. Shkurti, A. Xu, M. Meghjani, J. C. G. Higuera, Y. Girdhar, P. Giguere, B. B. Dey, J. Li, A. Kalmbach, C. Prahacs, et al. “Multi-domain monitoring of marine environments using a heterogeneous robot team,” In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 1747–1753. IEEE, 2012.

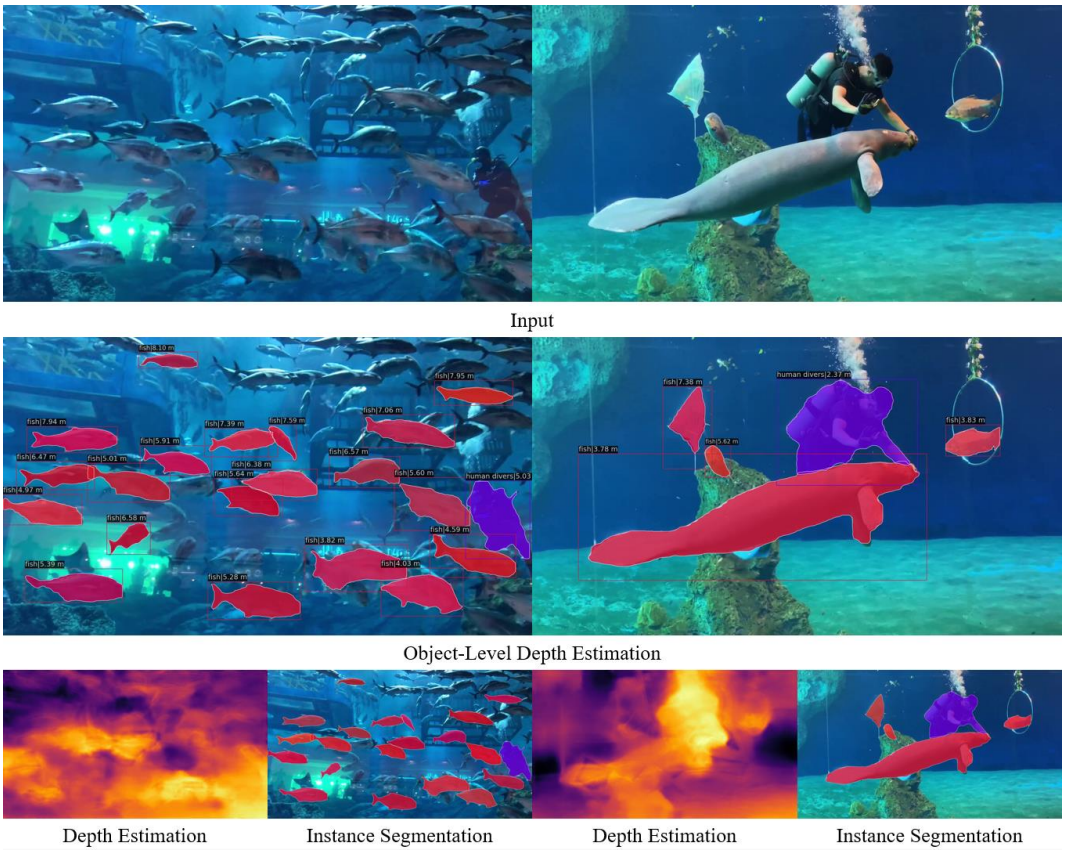


Fig. 8. More qualitative results of our UWSegDepth.

[2] Y. Gutnik, A. Avni, T. Treibitz, and M. Groper. "On the adaptation of an auv into a dedicated platform for close range imaging survey missions," Journal of Marine Science and Engineering, 10(7):974, 2022.

[3] A. Kim and R. M. Eustice. "Real-time visual slam for autonomous underwater hull inspection using visual saliency," IEEE Transactions on Robotics, 29(3):719 – 733, 2013.

[4] J.-Q. Yu. "駕駛場景中的影像辨識: 三維物件辨識與影像分割," 臺灣大學電信工程學研究所學位論文, pages 1 – 137, 2023.

[5] I.-C. Lu. "夜間暨低光源下自駕即時影像分割, 深度及優化模組," 臺灣大學電信工程學研究所學位論文, pages 1 – 109, 2024.

[6] J. Raihan A, P. E. Abas, and L. C. De Silva. "Depth estimation for underwater images from single view image," IET Image Processing, 14(16):4188 – 4197, 2020.

[7] S. Zhang, X. Gong, R. Nian, B. He, Y. Wang, and A. Lendasse. "A depth estimation model from a single underwater image with non-uniform illumination correction," In OCEANS 2017-Aberdeen, pages 1 – 5. IEEE, 2017.

[8] B. Yu, J. Wu, and M. J. Islam. "Udepth: Fast monocular depth estimation for visually-guided underwater robots," In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 3116 – 3123. IEEE, 2023.

[9] L. Ebner, G. Billings, and S. Williams. "Metrically scaled monocular depth estimation through sparse priors for underwater robots," In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 3751–3757. IEEE, 2024.

[10] P.-C. Pan, and S.-C. Pei. "BARIS: Boundary-Aware Refinement with Environmental Degradation Priors for Robust Underwater Instance Segmentation," arXiv preprint arXiv:2504.19643 (2025).

[11] Y. Randall. "Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets," Master ' s thesis, University of Haifa (Israel), 2023.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks," In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018.

[13] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in pytorch," Advances in neural information processing systems, 2017.

[14] L. Zhang, A. Rao, and M. Agrawala. "Adding conditional control to text-to-image diffusion models," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023.

[15] I. Loshchilov and F. Hutter. "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.

[16] Y. Liu, Q. Jiang, X. Wang, T. Luo, and J. Zhou. "Underwater image enhancement with cascaded contrastive learning," IEEE Transactions on Multimedia, 2024.

[17] D. Eigen, C. Puhrsch, and R. Fergus. "Depth map prediction from a single image using a multi-scale deep network," Advances in neural information processing systems, vol. 27, 2014.