Appendix

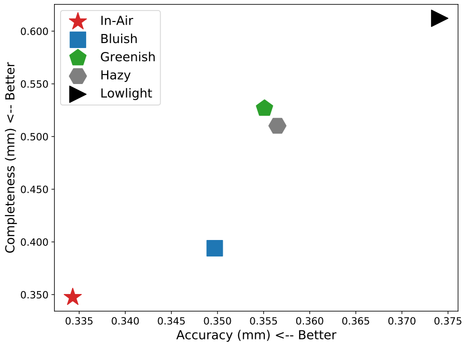
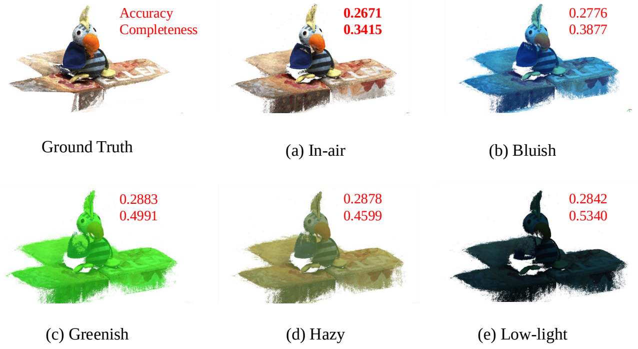


Fig. 1. Domain gap between the in-air and underwater domains.



Ground Truth

(a) In-air

(b) Bluish

(c) Greenish

(d) Hazy

(e) Low-light

Fig. 2. Point cloud reconstruction results of the in-air MVS model [1] on the *scan 4* of the (a) in-air evaluation set [2], underwater (b) bluish, (c) greenish, (d) hazy, and (e) low-light evaluation set, respectively. The number in top and bottom row denote the reconstruction accuracy and completeness error in mm, **lower the better**.

A. *In-air and Underwater Domain Gap*

As shown in Fig 1, there exists a reconstruction performance gap between the in-air and underwater domains. We adopt representative TransMVSNet [1] trained on the in-air MVS training set [2] as the baseline model and evaluate its point cloud reconstruction performance on the **in-air** MVS evaluation set [2], our synthesized underwater **bluish**, **greenish**, **hazy**, and **low-light** version of MVS evaluation set, respectively. The experimental results in Fig. 1 and Fig. 2 show that both the reconstruction accuracy and completeness drop significantly when applying the in-air model to the underwater domain, verifying the necessity of constructing a large-scale underwater multi-view stereo (UwMVS) dataset to shrink this domain gap and facilitate end-to-end UwMVS learning.

B. *Visualization of the t-SNE embeddings*

To validate the effectiveness of the proposed physically-guided approach for synthesizing multi-view underwater images, we visualize the t-SNE [3] embeddings of both the synthesized UwMVS dataset and a self-collected dataset of 2000 real-world underwater images. As shown in Fig. 3, the synthetic images from our UwMVS dataset exhibit significant overlap with the real-world images, demonstrating the effectiveness and superiority of the proposed synthesis approach.

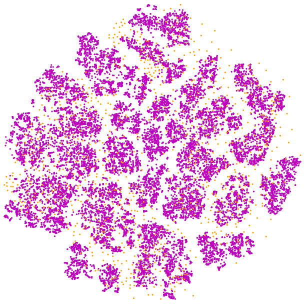


Fig. 3. Visualization of t-SNE embeddings of our synthesized UwMVS dataset (purple dots) and real-world underwater dataset (orange dots).

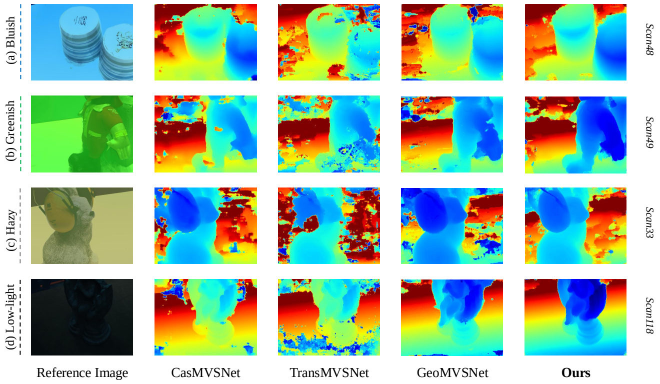


Fig. 4. Depth estimation results of recent methods [1], [4], [5] and ours for underwater scenes: (a) bluish, (b) greenish, (c) hazy, and (d) low-light.

C. *Depth Estimation Comparison*

We qualitatively compare the depth estimates obtained from our method with those of recent learning-based approaches [1], [4], [5] in Fig. 4. Our method achieves more accurate and complete depth estimation in challenging underwater scenarios, including bluish, greenish, hazy, and low-light degradations with complex geometries. This enhanced depth estimation performance leads to significant improvements in reconstruction quality, as demonstrated in Table II and Fig. 5 of the manuscript.

D. *Evaluation Metrics*

We quantify underwater point cloud reconstruction using reconstruction accuracy and completeness. Let $\mathcal{R}$ denote the reconstructed point cloud and $\mathcal{G}$ the ground-truth point cloud.

Accuracy measures the distance from $\mathcal{R}$ to $\mathcal{G}$, reflecting the quality of the reconstructed points, i.e., how closely $\mathcal{R}$ aligns with $\mathcal{G}$. Specifically, for each point $\mathbf{r}$ in $\mathcal{R}$, we compute its Euclidean distance to the nearest point in $\mathcal{G}$ and iterate this computation over all points in $\mathcal{R}$ to obtain a distance distribution. The mean of this distribution is defined as the accuracy:

$$e_{\mathbf{r}\to\mathcal{G}} = \min_{\mathbf{g}\in\mathcal{G}} \|\mathbf{r}-\mathbf{g}\|_2, \tag{1}$$

$$\mathit{Accuracy} = \frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}} [e_{\mathbf{r}\to\mathcal{G}} < d] \cdot e_{\mathbf{r}\to\mathcal{G}}, \tag{2}$$

where $\|\cdot\|_2$ denotes the Euclidean distance, $|\cdot|$ represents the number of points, $[\cdot]$ is the Iverson bracket, and d denotes the outlier rejection threshold.

Completeness measures the distance from $\mathcal{G}$ to $\mathcal{R}$, indicating the extent to which the ground truth $\mathcal{G}$ is restored. Specifically, for every point $\mathbf{g}$ in $\mathcal{G}$, we compute its Euclidean distance to the nearest point in $\mathcal{R}$ and perform this computation for all points in $\mathcal{G}$ to obtain a distance distribution. The mean of this distribution is defined as completeness:

$$e_{\mathbf{g}\to\mathcal{R}} = \min_{\mathbf{r}\in\mathcal{R}} \|\mathbf{g}-\mathbf{r}\|_2, \tag{3}$$

$$\mathit{Completeness} = \frac{1}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}} [e_{\mathbf{g}\to\mathcal{R}} < d] \cdot e_{\mathbf{g}\to\mathcal{R}}, \tag{4}$$

where $\|\cdot\|_2$ denotes the Euclidean distance, $|\cdot|$ represents the number of points, $[\cdot]$ is the Iverson bracket, and d denotes the outlier rejection threshold.

Overall Score There is a trade-off between reconstruction accuracy and completeness. Accuracy can be maximized with a sparse but precisely localized point cloud, while completeness can be maximized with a dense point cloud that covers the entire space. Therefore, the UwMVS method might achieve high accuracy with low completeness, or high completeness with low accuracy.

Notably, the accuracy and completeness metrics described above are defined for a single reconstruction scenario. For quantitative benchmarking on the test set of our proposed UwMVS dataset, accuracy and completeness are averaged over 22 underwater scenes. To balance accuracy and completeness, the overall score is computed as the arithmetic mean of the mean accuracy and mean completeness. A lower overall score indicates better reconstruction performance.

References

[1] Y. Ding, W. Yuan, Q. Zhu, H. Zhang, X. Liu, Y. Wang, and X. Liu, “Transmvsnet: Global context-aware multi-view stereo network with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 8585–8594.

[2] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, “Large-scale data for multiple-view stereopsis,” *International Journal of Computer Vision*, vol. 120, pp. 153–168, 2016.

[3] D. M. Chan, R. Rao, F. Huang, and J. F. Canny, “Gpu accelerated t-distributed stochastic neighbor embedding,” *Journal of Parallel and Distributed Computing*, vol. 131, pp. 1–13, 2019.

[4] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, “Cascade cost volume for high-resolution multi-view stereo and stereo matching,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2495–2504.

[5] Z. Zhang, R. Peng, Y. Hu, and R. Wang, “Geomvsnet: Learning multi-view stereo with geometry perception,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 508–21 518.