# End-to-End Underwater Multi-View Stereo for Dense Scene Reconstruction

Guidong Yang†, Junjie Wen†, Benyun Zhao, Qingxiang Li, Yijun Huang, Xi Chen, Alan Lam, and Ben M. Chen*

***Abstract*—Recent advancements in learning-based multi-view stereo (MVS) have demonstrated significant improvements over traditional counterpart, primarily due to the extensive availability of multi-view training images with ground-truth metric depths in the terrestrial in-air domain. However, underwater multi-view stereo (UwMVS) faces substantial challenges arising from the domain gap between in-air and underwater environments, leading to degraded performance when applying in-air MVS models to underwater scenarios. Furthermore, the progress of learning-based UwMVS methods has been hindered by the scarcity of underwater multi-view images with ground-truth depth maps and point clouds. In this paper, we address these challenges by introducing a physically-guided approach for synthesizing underwater multi-view images and present the first large-scale UwMVS dataset for end-to-end training and evaluation of learning-based UwMVS methods. Furthermore, we propose a novel UwMVS network that enhances geometric cue encoding to achieve more accurate and complete point cloud reconstruction. Extensive experiments on our dataset and real-world underwater scenes demonstrate that our dataset enables the trained models for underwater dense reconstruction and that our method achieves state-of-the-art performance in underwater reconstruction. Dataset, code and appendix are available at: *https://cuhk-usr-group.github.io/UwMVS/***

## I. Introduction and Related Work

Underwater reconstruction is critical for underwater archaeology [1], exploration [2], and navigation [3]. While acoustic sensors [4]–[7] have achieved decent underwater reconstruction, they do not capture the visual details achievable with low-cost cameras. Multi-view stereo (MVS) reconstructs dense point cloud of the scene from multi-view calibrated images by multi-view correspondence matching. Recently, learning-based MVS [8]–[11] significantly outperforms its traditional counterpart [12]–[15] concerning reconstruction accuracy and completeness by separating MVS into two stages: 1) learning-based multi-view depth estimation and 2) multi-view depth filtering & fusion for dense point cloud reconstruction. Benefiting from in-air MVS datasets [16]–[18], the performance of the learning-based MVS has been progressively improved.

However, for underwater reconstruction, existing underwater multi-view stereo (UwMVS) methods [19]–[23] predominantly rely on traditional MVS techniques, with a limited focus on learning-based MVS methods. The main dilemma is the difficulty of underwater data collection, which leads to the scarcity of underwater multi-view images with ground-truth metric depths for end-to-end training, as well as the absence of ground-truth point clouds for quantitative evaluation. While deploying in-air MVS models for underwater reconstruction presents a potential alternative, the significant domain gap caused by underwater degradation—such as color cast, limited visibility, blur, and illumination variation—results in a substantial drop in reconstruction performance when these models are directly applied underwater. A straightforward remedy is to perform underwater image enhancement (UIE) as a preprocessing step and then adopt in-air MVS models to reconstruct from enhanced images (UIE-then-MVS). However, for extreme underwater degradation, UIE encounters extra color deviation [24], low contrast [25], low saturation [26], low sharpness [27], edge blur [28], or noise interference [29]. Consequently, the enhanced images often fail to satisfy the feature requirements for high-level MVS and achieve multi-view photometric consistency. In contrast, UwMVS methods that train end-to-end models directly from underwater degraded multi-view images have shown significantly superior reconstruction accuracy and completeness compared to the UIE-then-MVS, as illustrated in Fig. 1.

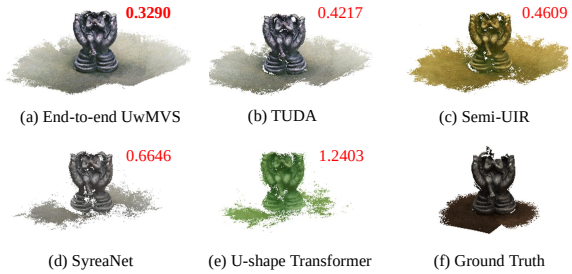


Fig. 1. With the same MVS method, end-to-end UwMVS (a) outperforms UIE-then-MVS (b)-(e) by $\mathbf{28.18}\%$, $\mathbf{40.09}\%$, $\mathbf{102.01}\%$, $\mathbf{276.99}\%$ for underwater greenish scene concerning overall point cloud reconstruction performance in $mm$ (**lower the better**), respectively. (a) adopts MVS method [10] trained on the proposed UwMVS training set and reconstructs from underwater degraded multi-view images. The color is recovered [28] during point cloud fusion. (b)-(e) follow the UIE-then-MVS, where TUDA [28], Semi-UIR [30], SyreaNet [27], and U-shape Transformer [31] are utilized for underwater image enhancement, respectively. Then, the in-air MVS model [10] is adopted to reconstruct from enhanced multi-view images. Note that the in-air MVS model is fine-tuned on the enhanced UwMVS training set for fair comparison.

To address the above issues, we first propose a physically-guided synthesis approach to synthesize underwater multi-view images guided by the physical degradation properties of real-world underwater images. With our synthesis method, we then construct the first large-scale UwMVS dataset for end-to-end training and evaluation of learning-based UwMVS. Afterward, we propose a novel UwMVS network with geometric encoding strategies, termed GE-UwMVS, to improve the

* Corresponding author.
† Equal contribution.
This work was supported by the InnoHK of the Government of the Hong Kong Special Administrative Region via the Hong Kong Center for Logistics Robotics (*Corresponding author: Ben M. Chen*).
The authors are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong (CUHK), Shatin, NT, Hong Kong (e-mail: {gdyang, jjwen, byzhao, yjhuang, xichen, alam, bmchen}@mae.cuhk.edu.hk, qingxiang.li@polimi.it)

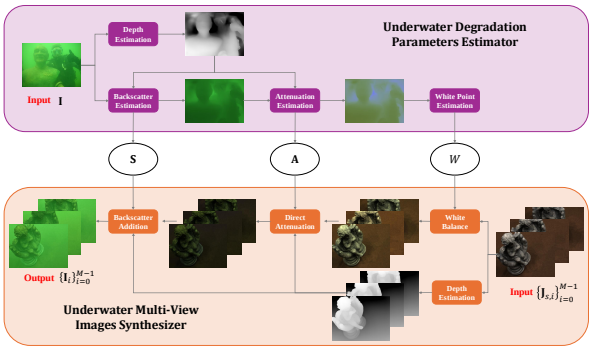


Fig. 2. Our proposed two-stage physically-guided underwater multi-view images synthesis approach: 1) Underwater degradation parameters estimator and 2) Underwater multi-view images synthesizer.

underwater reconstruction accuracy and completeness, where the depth consistency optimization is directly performed in the 3D point space during learning by invoking ground-truth depth cues from adjacent views and the surface normal geometries are explicitly encoded to refine the sampled depth hypotheses to be consistent in the local neighbor regions.

Extensive experiments validate the effectiveness of the proposed synthesis approach and the UwMVS dataset, enabling trained models to perform dense reconstruction from underwater images with degradation. Both quantitative and qualitative evaluations on the dataset and real-world underwater scenes demonstrate that GE-UwMVS achieves state-of-the-art performance in underwater point cloud reconstruction.

## II. Underwater Multi-View Images Synthesis

In this section, we first propose a physically-guided underwater multi-view images synthesis method based on the revised underwater image formation model [32]. Then, we construct the first large-scale UwMVS dataset for end-to-end training and evaluation of learning-based UwMVS.

### A. Revised Underwater Image Formation Model

Compared to the simplified underwater image formation model [33], [34], the more precise revised underwater image formation model [32] is formulated as:

$$\mathbf{I}(\mathbf{p}) = \underbrace{\mathbf{J}(\mathbf{p}) \cdot e^{-\beta^D \mathbf{Z}(\mathbf{p})}}_{\mathbf{D}(\mathbf{p})} + \underbrace{B^{\infty} \cdot (1 - e^{-\beta^B \mathbf{Z}(\mathbf{p})})}_{\mathbf{B}(\mathbf{p})}, \tag{1}$$

where $\mathbf{I}(\mathbf{p})$, $\mathbf{D}(\mathbf{p})$, and $\mathbf{B}(\mathbf{p})$ denote the underwater degraded image intensity, direct signal intensity, and backscattered signal intensity at pixel $\mathbf{p}$, respectively. $\mathbf{J}(\mathbf{p})$ denotes the underwater unattenuated image intensity and $B^{\infty}$ is the homogeneous background light. $e^{-\beta \mathbf{Z}(\mathbf{p})}$ denotes the Lambert-Beer empirical law of exponential decay, where $\mathbf{Z}(\mathbf{p})$ stands for the camera-object distance (depth) at pixel $\mathbf{p}$, $\beta^D$ and $\beta^B$ signify the wideband attenuation and backscatter coefficient.

### B. Underwater Multi-View Images Synthesis Approach

**Estimator** As shown in Fig. 2, the underwater degradation parameters estimator takes as input the captured real-world underwater image $\mathbf{I}$ and applies Sea-Thru [35] to successively estimate backscatter coefficients $\mathbf{S} = \{B^{\infty}, \beta^B, J', \beta^{D'}\}$, attenuation coefficients $\mathbf{A} = \{a, b, c, d\}$, and white point $W$ based on the revised underwater image formation model. The camera-object depth map $\mathbf{Z}$ is estimated via the monocular depth estimation network ZoeDepth [36]. The backscatter coefficients are related as:

$$\hat{\mathbf{B}}(\mathbf{p}) = B^{\infty} \cdot (1 - e^{-\beta^B \mathbf{Z}(\mathbf{p})}) + J' \cdot e^{-\beta^{D'} \mathbf{Z}(\mathbf{p})}, \tag{2}$$

where $\hat{\mathbf{B}} \approx \mathbf{B}$ and $J' \cdot e^{-\beta^{D'} \mathbf{Z}(\mathbf{p})}$ denotes the residual from the direct signal. The lower and upper bounds for the $B^{\infty}, \beta^B, J', \beta^{D'}$ are $[0, 0, 0, 0]$ and $[1, 5, 1, 5]$, respectively.

The attenuation coefficients are related as:

$$\hat{\beta}^D = a \cdot e^{b \cdot \mathbf{Z}(\mathbf{p})} + c \cdot e^{d \cdot \mathbf{Z}(\mathbf{p})}, \tag{3}$$

where $\hat{\beta}^D \approx \beta^D$ decays exponentially with the camera-object distance $\mathbf{Z}(\mathbf{p})$. The lower and upper bounds of $a, b, c, d$ are $[0, -\infty, 0, -\infty]$ and $[\infty, 0, \infty, 0]$, respectively.

The global white point $W$ is estimated based on the Gray World Hypothesis [37] to remove the diffuse attenuation of $\mathbf{J}$ induced by the downwelling light vertically penetrating the water and obtain the unattenuated image $\mathbf{J}_s$ above the water surface (i.e., the in-air image):

$$\mathbf{J}_s(\mathbf{p}) = (\mathbf{I}(\mathbf{p}) - \hat{\mathbf{B}}(\mathbf{p})) \cdot e^{\hat{\beta}^D \mathbf{Z}(\mathbf{p})} / W \tag{4}$$

**Synthesizer** As shown in Fig. 2, the underwater multi-view images synthesizer takes real-world in-air multi-view images $\{\mathbf{J}_{s,i}\}_{i=0}^{M-1}$ and performs the white balance, direct attenuation, and backscatter addition with the estimated $W$, $\mathbf{A}$, $\mathbf{S}$ to synthesize underwater multi-view images $\{\mathbf{I}_i\}_{i=0}^{M-1}$. Specifically, the $\hat{\mathbf{B}}(\mathbf{p})$ and $\hat{\beta}^D$ are first estimated via Eq. 2 and Eq. 3. Then, the multi-view images are synthesized as:

$$\mathbf{I}_i(\mathbf{p}) = \overbrace{\underbrace{\mathbf{J}_{s,i}(\mathbf{p}) \cdot W}_{\text{White Balance}} \cdot e^{-\hat{\beta}^D \mathbf{Z}(\mathbf{p})}}^{\text{Direct Attenuation}} + \underbrace{\hat{\mathbf{B}}(\mathbf{p})}_{\text{Backscatter Addition}} \tag{5}$$

### C. Underwater Multi-View Stereo Dataset

We utilize the proposed synthesis approach to construct the UwMVS dataset. The real-world underwater environment is categorized into four challenging scene types based on color distortion and visibility: **bluish**, **greenish**, **hazy**, and **low-light**.

TABLE I
Statistics of the Proposed UwMVS Dataset

| Attributes | Training Set | Validation Set | Test Set |
|---|---|---|---|
| # of Reconstruction Scenarios | 79 | 18 | 22 |
| # of Underwater Scene Types | 4 | 4 | 4 |
| # of Underwater Degradation Levels | 7 | 7 | 1 |
| # of Images in Total | 108,388 | 24,696 | 4,312 |
| # of Images for Each Scenario | 1,372 | 1,372 | 196 |
| Ground-Truth Depth Map | ✓ | ✓ | - |
| Ground-Truth Point Cloud | ✓ | ✓ | ✓ |

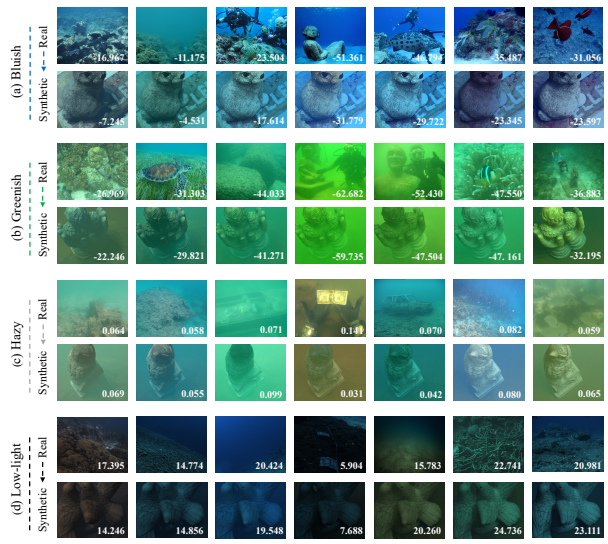


Fig. 3. Example images from our UwMVS dataset. We categorize the underwater environment into four types of challenging scenes: (a) **bluish**, (b) **greenish**, (c) **hazy**, and (d) **low-light**. For each scene type, the top row displays real-world underwater images across seven degradation levels, while the bottom row presents the corresponding synthetic underwater images. A closer match between the numbers in real and synthetic image pairs indicates more similar visual properties, see Subsection IV-C for details.

For each scene type, we collect real-world underwater images across seven different degradation levels. For each degradation level within a scene type, the synthesis approach takes as input a real-world underwater image $\mathbf{I}$ and in-air multi-view images $\{\mathbf{J}_{s,i}\}_{i=0}^{M-1}$ from a standard in-air MVS dataset [16] to produce synthetic underwater multi-view images $\{\mathbf{I}_i\}_{i=0}^{M-1}$ that inherit the real-world underwater degradation characteristics. This process is repeated for each in-air reconstruction scenario to build the UwMVS dataset. Example images from the UwMVS dataset are shown in Fig. 3.

To the best of our knowledge, our UwMVS dataset is the first large-scale dataset for end-to-end training and evaluation of learning-based UwMVS methods. Table I provides a summary of the dataset statistics. It includes 79, 18, and 22 underwater reconstruction scenarios for training, validation, and testing, respectively. Each scenario consists of 1372 images for both training and validation, and 196 images for testing. The dataset covers 4 underwater scene types per scenario and incorporates 7 levels of underwater degradation for training and validation to enhance model robustness. The most severe degradation level is reserved for the test set to evaluate the generalization capability of the learning-based UwMVS. In total, the training, validation, and test sets comprise $108,388$, $24,696$, and $4,312$ underwater multi-view images, respectively. Our synthesis approach allows for the generation of underwater multi-view images with ground-truth depth maps and point clouds directly obtained from the in-air MVS dataset, thereby eliminating the need for labor-intensive depth annotation and costly underwater laser scanning.

## III. Underwater Multi-View Stereo with Geometric Encoding

In this section, we propose the UwMVS network with geometric encoding, denoted as GE-UwMVS, designed to achieve more accurate and complete underwater point cloud reconstruction. An overview of the GE-UwMVS network is presented in Fig. 4. The network receives $N$ underwater images $\{\mathbf{I}_i \in \mathbb{R}^{3\times H\times W}\}_{i=0}^{N-1}$ as input, along with their corresponding camera intrinsics $\{\mathbf{K}_i \in \mathbb{R}^{3\times 3}\}_{i=0}^{N-1}$ and extrinsics $\{[\mathbf{R}_i \in \mathbb{R}^{3\times 3}; \mathbf{t}_i \in \mathbb{R}^{3\times 1}]\}_{i=0}^{N-1}$. The network predicts the depth map $\mathbf{D}_{0,\text{est}} \in \mathbb{R}^{H\times W}$ for the reference image $\mathbf{I}_0$. For each reconstruction scenario, every image is iteratively treated as the reference image to perform per-view depth estimation. The estimated depth maps are then filtered and fused [38] to generate a dense point cloud reconstruction.

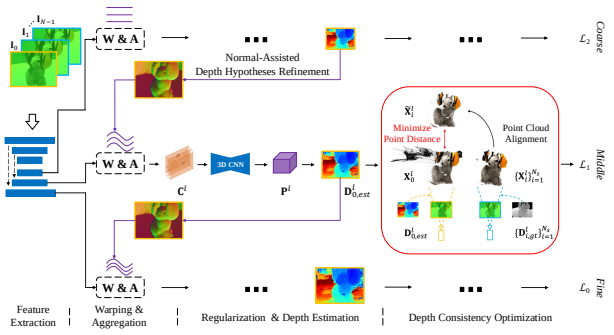


Fig. 4. Network overview of GE-UwMVS, where $L$ is set to 3 to form a three-stage coarse-to-fine network. The red box highlights the proposed depth consistency optimization module (the point clouds with original colors are demonstrated for clarity), while the violet line indicates the proposed normal-assisted depth hypotheses refinement module.

### A. Adaptive Cost Volume Aggregation

Given $N$-view underwater images $\{\mathbf{I}_i\}_{i=0}^{N-1}$, we utilize a Feature Pyramid Network (FPN) [39] to extract feature pyramids at $L$ scales, denoted as $\{\mathbf{F}_i^l\}_{i=0}^{N-1} \in \mathbb{R}^{C_l\times \frac{H}{2^l}\times \frac{W}{2^l}}$. Here, $l \in \{0, 1, \ldots, L-1\}$ indexes the feature level, and $C_l$ represents the number of channels at level $l$. For each feature level $l$, the depth range of the reference view, $[\mathbf{D}_{\text{min}}^l, \mathbf{D}_{\text{max}}^l]$, is uniformly discretized into $M_l$ discrete depth hypotheses:

$$\mathbf{D}_{\text{ini},m}^l = \mathbf{D}_{\text{min}}^l + m\left(\frac{\mathbf{D}_{\text{max}}^l - \mathbf{D}_{\text{min}}^l}{M_l - 1}\right), \tag{6}$$

where $\{\mathbf{D}_{\text{max}}^l, \mathbf{D}_{\text{min}}^l, \mathbf{D}_{\text{ini},m}^l\} \in \mathbb{R}^{\frac{H}{2^l}\times \frac{W}{2^l}}$ denote the maximum, minimum, and initial depth hypotheses at feature level $l$, respectively. The index $m \in \{0, 1, \ldots, M_l - 1\}$ corresponds to the depth hypothesis plane. While the depth range at coarse levels is predefined, finer-level ranges are dynamically adjusted based on coarser estimates. The initial depth hypotheses $\mathbf{D}_{\text{ini},m}^l$ are further refined using surface normal geometries to obtain the final depth hypotheses $\mathbf{D}_m^l$ for more accurate and complete depth estimation and reconstruction, as detailed in Subsection III-D.

We establish the pairwise pixel coordinate mapping between the reference-view feature map $\mathbf{F}_0^l$ and the adjacent source-view feature maps $\{\mathbf{F}_i^l\}_{i=1}^{N-1}$ at depth $\mathbf{D}_m^l$ using differentiable homography:

$$\mathbf{p}_i = \mathbf{K}_i\left[\mathbf{R}_{0\to i}\left(\mathbf{K}_0^{-1}\mathbf{p}_0\mathbf{D}_m^l(\mathbf{p}_0)\right) + \mathbf{t}_{0\to i}\right], \tag{7}$$

where $\mathbf{p}_0$ and $\mathbf{p}_i$ represent the pixel coordinates in the reference view and the $i$-th source view, respectively. We

compute the relative rotation matrix $\mathbf{R}_{0\to i}$ and the translation vector $\mathbf{t}_{0\to i}$ between the reference view and the $i$-th source view as $\mathbf{R}_i\mathbf{R}_0^{-1}$ and $\mathbf{t}_0 - \mathbf{R}_i\mathbf{R}_0^{-1}\mathbf{t}_i$, respectively. Here, $\mathbf{K}_0$ and $\mathbf{K}_i$ denote the camera intrinsics for the reference view and the $i$-th source view, respectively. Given $\mathbf{p}_i$ from $\mathbf{F}_i^l$, we use differentiable bilinear interpolation to interpolate the $i$-th source-view feature map $\widetilde{\mathbf{F}}_i^l$ aligned to the reference view.

We perform the above coordinate mapping and interpolation process for each depth hypothesis $\mathbf{D}_m^l(\mathbf{p}_0)$ to obtain the corresponding feature map $\widetilde{\mathbf{F}}_{i,\mathbf{D}_m^l(\mathbf{p}_0)}^l$. These feature maps are then stacked along the depth dimension to construct the source-view feature volume $\{\mathbf{V}_i^l \in \mathbb{R}^{C_l\times M_l\times \frac{H}{2^l}\times\frac{W}{2^l}}\}_{i=1}^{N-1}$. The reference-view feature map $\mathbf{F}_0^l$ is repeated $M_l$ times along the depth dimension to form the reference-view feature volume $\mathbf{V}_0^l \in \mathbb{R}^{C_l\times M_l\times \frac{H}{2^l}\times\frac{W}{2^l}}$. Finally, we adaptively fuse the multi-view feature volumes $\{\mathbf{V}_i^l\}_{i=0}^{N-1}$ into the cost volume $\mathbf{C}^l \in \mathbb{R}^{M_l\times \frac{H}{2^l}\times\frac{W}{2^l}}$ by inferring per-view per-pixel visibility [10].

### *B. Cost Volume Regularization and Depth Estimation*

We apply a multi-scale 3D U-Net [9], [10], [40] to regularize the initial noise-contaminated cost volume $\mathbf{C}^l$ and to predict the probability volume $\mathbf{P}^l \in \mathbb{R}^{M_l\times \frac{H}{2^l}\times\frac{W}{2^l}}$ via the softmax operation along the depth dimension. Here, $\mathbf{P}^l$ encodes the probability maps corresponding to $M_l$ depth hypotheses. The depth estimation is framed as a pixel-wise classification problem, where the depth hypothesis with the maximum probability is taken (winner-takes-all):

$$\mathbf{D}_{0,\text{est}}^l(\mathbf{p}) = \underset{d\in\{\mathbf{D}_m^l(\mathbf{p}_0)\}_{m=0}^{M_l-1}}{\arg\max}\ \mathbf{P}_0^l(\mathbf{p}), \tag{8}$$

where $\mathbf{D}_{0,\text{est}}^l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ represents the reference-view depth estimation for feature level $l$. The depth estimation at the fine level ($l = 0$) is taken as the final output for point cloud fusion [38].

### *C. Depth Consistency Optimization*

Most existing methods [10], [11], [40], [41] perform depth inconsistency check after network learning and directly discard the inconsistent pixels during point cloud fusion, leading to incomplete reconstruction. Besides, the cost volume delivers insufficient geometric cues essential to depth estimation and subsequent reconstruction. To dynamically improve depth consistency and explicitly strengthen the geometric modeling, we perform depth consistency optimization directly in the 3D point space by encoding ground-truth depth cues from adjacent views to suppress the inconsistent pixels during learning.

**2D→3D Backward Projection** At each feature level $l$, we first back-project the pixel coordinates from the reference view and source views to the 3D point space, using the reference-view depth estimation and the source-view ground-truth depth, respectively. The backward projection is formulated as:

$$\mathbf{X}_i^l(\mathbf{p}) = (\mathbf{K}_i\mathbf{R}_i)^{-1}\mathbf{p}\mathbf{D}_i^l(\mathbf{p}) - \mathbf{R}_i^{-1}\mathbf{t}_i, \tag{9}$$

where $\{\mathbf{X}_i^l \in \mathbb{R}^{3\times\frac{H}{2^l}\times\frac{W}{2^l}}\}_{i=0}^{N_s}$ denotes the back-projected point coordinates in the world space. Here, $\mathbf{X}_0^l$ and $\{\mathbf{X}_i^l\}_{i=1}^{N_s}$ represent the point coordinates for the reference view and $N_s$ adjacent source views involved in the depth consistency optimization, respectively. The sets $\{\mathbf{K}_i\}_{i=0}^{N_s}$ and $\{[\mathbf{R}_i|\mathbf{t}_i]\}_{i=0}^{N_s}$ denote the scaled camera intrinsics and extrinsics, respectively. Additionally, $\{\mathbf{D}_i^l(\mathbf{p})\}_{i=0}^{N_s}$ represents the depth at pixel $\mathbf{p}$, where $\mathbf{D}_0^l(\mathbf{p}) = \mathbf{D}_{0,\text{est}}^l(\mathbf{p})$ is the depth estimation of the reference view, and $\{\mathbf{D}_i^l(\mathbf{p}) = \mathbf{D}_{i,\text{gt}}^l(\mathbf{p})\}_{i=1}^{N_s}$ are the ground-truth depths from the $N_s$ adjacent source views.

**Point Cloud Alignment** As illustrated in Fig. 4, after the backward projection, the reference-view point cloud $\mathbf{X}_0^l$ is partially corrupted with noisy points due to inaccurate and incomplete depth estimation $\mathbf{D}_{0,\text{est}}^l$, while the source-view point cloud $\mathbf{X}_i^l$ is complete and clean benefiting from ground-truth depth map $\mathbf{D}_{i,\text{gt}}^l$. Additionally, there is a misalignment between $\mathbf{X}_0^l$ and $\mathbf{X}_i^l$ due to cumulative errors in the camera intrinsics and extrinsics. To address this issue, we establish pairwise coordinate mapping to align the source-view point cloud with the reference view. The coordinate mapping between a reference-view pixel $\mathbf{p}_0$ and a source-view pixel $\mathbf{p}_i$ is formulated as:

$$\mathbf{p}_i = \mathbf{K}_i[\mathbf{R}_{0\to i}(\mathbf{K}_0^{-1}\mathbf{p}_0\mathbf{D}_{0,\text{est}}^l(\mathbf{p}_0)) + \mathbf{t}_{0\to i}]. \tag{10}$$

Given $\mathbf{p}_i$ from $\mathbf{X}_i^l$, we regard $x$, $y$, $z$ point coordinates as channel features and utilize the differentiable bilinear interpolation to obtain the source-view point cloud $\widetilde{\mathbf{X}}_i^l$ aligned to the reference view.

**Depth Inconsistency Check** We then compute the pointwise distance error as the Euclidean norm between reference-view points $\mathbf{X}_0^l$ and aligned source-view points $\widetilde{\mathbf{X}}_i^l$:

$$\mathbf{E}_{0\leftrightarrow i}^l(\mathbf{p}) = \left\|\mathbf{X}_0^l(\mathbf{p}) - \widetilde{\mathbf{X}}_i^l(\mathbf{p})\right\|_2, \tag{11}$$

where $\mathbf{E}_{0\leftrightarrow i}^l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ represents the error map, and $\|\cdot\|_2$ denotes the $L_2$-norm. If the pointwise distance error at pixel $\mathbf{p}$ exceeds a certain threshold, we consider the reference-view depth estimation $\mathbf{D}_{0,\text{est}}^l(\mathbf{p})$ as inconsistent. We perform the 2D→3D backward projection for $N_s$ source views for each reference view and then accumulate and average the depth inconsistency of the reference view with respect to all source views as:

$$\mathbf{M}_0^l(\mathbf{p}) = \frac{1}{N_s}\sum_{i=1}^{N_s}\left[\mathbf{E}_{0\leftrightarrow i}^l(\mathbf{p}) > \epsilon_l\right], \tag{12}$$

where $\mathbf{M}_0^l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ denotes the depth inconsistency mask of the reference view at feature level $l$. The notation $[\cdot]$ indicates the Iverson bracket, and $\epsilon_l$ represents the per-level point distance threshold.

**Loss Function** We adopt the cross-entropy loss to optimize this pixel-wise depth classification problem and supervise the difference between the predicted probability volume $\mathbf{P}^l(\mathbf{p})$ and the ground-truth probability volume $\mathbf{P}_{\text{gt}}^l(\mathbf{p})$, where $\mathbf{P}_{\text{gt}}^l(\mathbf{p})$ is obtained by one-hot encoding the depth hypotheses closest to the ground-truth depths. We weight the depth inconsistency mask $\mathbf{M}_0^l(\mathbf{p})$ over the cross-entropy loss to geometrically optimize depth consistency, providing a per-pixel penalty for reference-view depth inconsistencies with $N_s$ source views. The per-level loss $\mathcal{L}_l$ for depth optimization is defined as:

$$\mathcal{L}_l = \sum_{\mathbf{p}\in\{\mathbf{p}_v\}} \sum_{m=0}^{M_l-1} -\left(1+\mathbf{M}_0^l(\mathbf{p})\right)\mathbf{P}_{\text{gt},m}^l(\mathbf{p})\log\left(\mathbf{P}_m^l(\mathbf{p})\right), \tag{13}$$

where $\{\mathbf{P}_{\text{gt},m}^l, \mathbf{P}_m^l\} \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ denote the ground-truth and estimated probability maps for the $m$-th depth hypothesis. The set $\{\mathbf{p}_v\}$ contains pixels with valid ground-truth depth. We add 1 to retain the original cross-entropy loss and prevent excessive correction of depth consistency. The total loss for optimization is the weighted sum of the per-level loss:

$$\mathcal{L} = \sum_{l=0}^{L-1} \lambda_l \mathcal{L}_l, \tag{14}$$

where $\mathcal{L}$ represents the total loss for depth optimization, $L$ denotes the total number of feature levels, and $\lambda_l$ is the loss weight for level $l$.

### D. Normal-Assisted Depth Hypotheses Refinement

Initial depth hypotheses introduce learning ambiguity into the cross-entropy loss, as the ground-truth probability volume is obtained by one-hot encoding the depth hypotheses that are closest to the ground-truth depths. Additionally, these initial hypotheses often result in inconsistent depth estimates within local neighbor regions, particularly under challenging multi-view matching conditions. To address these issues, we encode surface normal geometries [42]–[45] to refine the depth hypotheses, ensuring geometric smoothness and consistency in local regions. We use the monocular normal estimation network Omnidata [46] to predict the normal map, which helps to resolve multi-view matching ambiguities, especially under challenging underwater conditions.

Given the reference-view normal map $\mathbf{N} \in \mathbb{R}^{3\times H\times W}$ and the $m$-th initial depth hypothesis $\mathbf{D}_{\text{ini},m}^l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ from the coarse-to-fine framework, we first interpolate $\mathbf{D}_{\text{ini},m}^l$ to the original resolution and back-project the reference-view pixel coordinates to camera coordinates:

$$\mathbf{X}(\mathbf{p}) = \mathbf{K}^{-1}\mathbf{p}\mathbf{D}_{\text{ini},m}^l(\mathbf{p}), \tag{15}$$

where $\mathbf{X}(\mathbf{p}) \in \mathbb{R}^{3\times1}$ denotes the back-projected camera coordinates at pixel $\mathbf{p}$, and $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the camera intrinsics of the reference view. For each pixel $\mathbf{p}$, we then search its $n$ square neighboring pixels $\mathbf{p}_i$ centered at $\mathbf{p}$ , where $i \in \{0, 1, \ldots, n-1\}$, and perform the same back-projection process to obtain corresponding camera coordinates $\mathbf{X}(\mathbf{p}_i) \in \mathbb{R}^{3\times1}$. We impose normal constraints with local planar priors:

$$\mathbf{N}(\mathbf{p}_i)\cdot(\mathbf{X}(\mathbf{p}_i)-\mathbf{X}(\mathbf{p})) = 0, \tag{16}$$

where $\mathbf{N}(\mathbf{p}_i) \in \mathbb{R}^{3\times1}$ is the normal vector at pixel $\mathbf{p}_i$, and $i \in \{0, 1, \ldots, n-1\}$. We then refine the neighboring initial depth hypotheses by encoding these normal geometries, as described in Eq. 16:

$$\mathbf{D}_m^l(\mathbf{p}_i) = \frac{\mathbf{N}(\mathbf{p}_i)\cdot(\mathbf{K}^{-1}\mathbf{p})}{\mathbf{N}(\mathbf{p}_i)\cdot(\mathbf{K}^{-1}\mathbf{p}_i)}\mathbf{D}_{\text{ini},m}^l(\mathbf{p}), \tag{17}$$

where $\mathbf{D}_m^l(\mathbf{p}_i)$ denotes the refined depth hypothesis at pixel $\mathbf{p}_i$. We perform this refinement for all $M_l$ depth hypotheses to obtain the refined depth hypotheses $\mathbf{D}_m^l \in \mathbb{R}^{H\times W}$, where $m \in \{0, 1, \ldots, M_l-1\}$. These refined depth hypotheses are then interpolated to the corresponding feature level $\mathbf{D}_m^l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ for adaptive cost volume aggregation and depth estimation.

## IV. Experiments

### A. Implementation Details

We implement the proposed GE-UwMVS network using PyTorch and train it on the UwMVS training set. The feature level $L$ is set to 3 to form a three-stage coarse-to-fine network. The number of input views $N$ is 5, the number of source views $N_s$ and the point distance threshold $\epsilon_l$ for depth consistency optimization are set to 8 and 0.2, respectively. Both the image and ground-truth depth map are resized to a resolution of $512 \times 640$ pixels. The depth range is configured from 425 mm to 905 mm to uniformly sample depth hypotheses. We define the number of depth hypothesis planes $M_l$ as 48, 32, and 8 for the coarse, middle, and fine levels, respectively. For depth hypothesis refinement, we use 8 neighboring pixels ($n = 8$) to form a $3\times3$ local neighborhood. The depth intervals are set to 4, 2, and 1 times the coarse-level depth interval, corresponding to the coarse, middle, and fine levels. The loss weights $\lambda_l$ are set to 1, 1, and 2 for the coarse, middle, and fine levels, respectively. We optimize the network using the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The network is trained for 30 epochs on two NVIDIA RTX 3090 Ti GPUs, with a batch size of 2 per GPU. The initial learning rate is set to 0.001, and the weight decay is 0.0001. A multi-step learning rate scheduler reduces the learning rate by a factor of 0.5 at epochs 8, 12, 16, and 20.

### B. Benchmarking Performance

Table II benchmarks the proposed GE-UwMVS method against recent learning-based MVS methods on the UwMVS test set to quantify their underwater point cloud reconstruction performance. We report reconstruction accuracy (Acc.), completeness (Comp.), and overall score (Overall) based on mean error distance ($mm$, **lower is better**) across 22 reconstruction scenarios in 4 types of underwater scenes. For all methods, we set $N = 5$ with an image resolution of $H \times W = 864 \times 1152$ for depth estimation and use the Fusible method [38] to fuse multi-view depth maps into the final point cloud. We keep original settings of each method for point cloud fusion. Extensive experiments show that our method achieves state-of-the-art underwater reconstruction performance in terms of overall score by striking an excellent trade-off between reconstruction accuracy and completeness. Figure 5 presents a qualitative comparison between our method and recent methods [10], [11], where our method produces more accurate and complete reconstructions under extreme underwater conditions. Note that the depth is directly estimated from the degraded underwater images, we add restored color [28] to the point cloud during fusion, which does not affect reconstruction performance.

### C. Ablation Experiments

**Effectiveness of Synthesis Method** We quantitatively show that our synthetic underwater images inherit visual properties

TABLE II
QUANTITATIVE RESULTS ON THE PROPOSED UWMVS DATASET FOR BENCHMARKING POINT CLOUD RECONSTRUCTION PERFORMANCE

| Methods | Mean Error Distance on 22 **Bluish** Scenarios | | | Mean Error Distance on 22 **Greenish** Scenarios | | | Mean Error Distance on 22 **Hazy** Scenarios | | | Mean Error Distance on 22 **Low-light** Scenarios | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC. ↓ $(mm)$ | Comp. ↓ $(mm)$ | **Overall*** ↓ $(mm)$ | ACC. ↓ $(mm)$ | Comp. ↓ $(mm)$ | **Overall*** ↓ $(mm)$ | ACC. ↓ $(mm)$ | Comp. ↓ $(mm)$ | **Overall*** ↓ $(mm)$ | ACC. ↓ $(mm)$ | Comp. ↓ $(mm)$ | **Overall*** ↓ $(mm)$ |
| MVSNet [8] | 0.560 | 0.475 | 0.518 | 0.588 | 0.539 | 0.564 | 0.617 | 0.587 | 0.602 | 0.667 | 0.647 | 0.657 |
| CasMVSNet [9] | 0.348 | 0.405 | 0.377 | 0.345 | 0.472 | 0.409 | 0.337 | 0.470 | 0.403 | 0.347 | 0.592 | 0.470 |
| RC-MVSNet [47] | 0.425 | 0.341 | 0.383 | 0.431 | 0.390 | 0.411 | 0.421 | 0.370 | 0.395 | 0.457 | 0.462 | 0.460 |
| UniMVSNet [40] | 0.370 | 0.306 | 0.338 | 0.387 | 0.343 | 0.365 | 0.380 | 0.363 | 0.372 | 0.338 | 0.698 | 0.518 |
| TransMVSNet [10] | 0.371 | 0.300 | 0.336 | 0.362 | 0.355 | 0.359 | 0.387 | 0.493 | 0.440 | 0.386 | 0.588 | 0.487 |
| GeoMVSNet [11] | **0.309** | 0.476 | 0.393 | **0.316** | 0.532 | 0.424 | **0.322** | 0.495 | 0.409 | **0.302** | 0.773 | 0.538 |
| DMVSNet [48] | 0.385 | 0.281 | 0.333 | 0.393 | 0.298 | 0.346 | 0.395 | 0.306 | 0.351 | 0.429 | 0.360 | 0.394 |
| Ours | 0.356 | **0.274** | **0.315** | 0.355 | **0.286** | **0.321** | 0.356 | **0.294** | **0.325** | 0.371 | **0.323** | **0.347** |

* **Note that the overall score is the summary measure of the overall reconstruction performance**.
↓ The ↓ means that the smaller value indicates the better reconstruction performance.

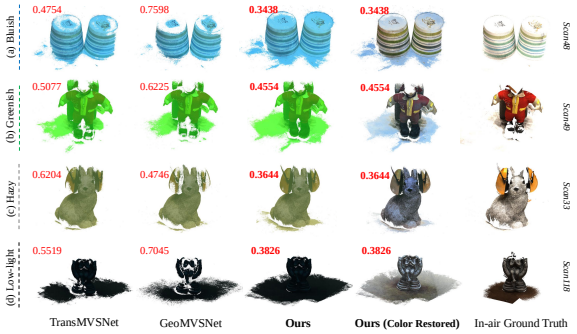


Fig. 5. Point cloud reconstruction results of recent methods [10], [11] and ours for underwater scenes: (a) bluish, (b) greenish, (c) hazy, and (d) low-light. The number denotes the overall score in $mm$, **lower the better**.

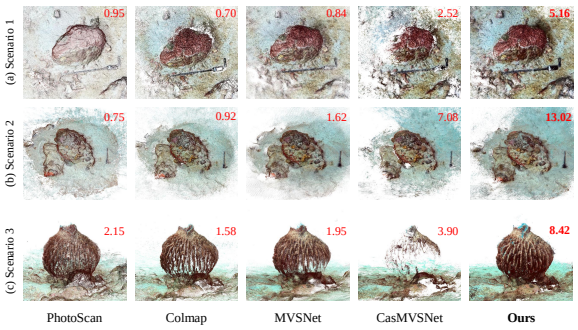


Fig. 6. Point cloud reconstruction results of industrial solutions [51], [52], recent learning-based methods [8], [9], and ours across three real-world underwater scenarios. The number is the total number of points in Million.

TABLE III
ABLATION EXPERIMENTS ON UWMVS DATASET

| Dataset Type | | Method | Mean Error Distance on 22 **Greenish** Scenes | | |
|---|---|---|---|---|---|
| In-air | Uw | | Acc. ↓ $(mm)$ | Comp. ↓ $(mm)$ | **Overall*** ↓ $(mm)$ |
| ✓ | | CasMVSNet | 0.357 | 0.495 | 0.426 |
| | ✓ | CasMVSNet | **0.345** | **0.472** | **0.409** |
| ✓ | | TransMVSNet | **0.355** | 0.527 | 0.441 |
| | ✓ | TransMVSNet | 0.362 | **0.355** | **0.359** |
| ✓ | | GeoMVSNet | **0.288** | 0.714 | 0.501 |
| | ✓ | GeoMVSNet | 0.316 | **0.532** | **0.424** |

* **The overall score is the summary measure of the overall reconstruction performance**.

of the real underwater images by computing [49] $Avg_b$, $Avg_a$, *RMS contrast*, and $Avg_L$ for **bluish**, **greenish**, **hazy**, and **low-light** scenes, respectively. The $Avg_L$, $Avg_a$, and $Avg_b$ denote the average values of the channel $L$, $a$, and $b$ in the CIELAB color space, respectively. Figure 3 shows these metrics for real and synthetic images, indicating that synthetic images closely match real underwater degradation levels. Table III verifies the effectiveness of our synthesis approach, showing that models trained on the synthetic UwMVS dataset outperform those trained on in-air images in underwater reconstruction.

**Effectiveness of GE-UwMVS Modules** Table IV presents the impact of different modules in our GE-UwMVS method on the UwMVS test set: adaptive cost volume aggregation (Ada.), depth consistency optimization (Dep.), and normal-assisted depth hypotheses refinement (Nor.). The Ada. module improves overall reconstruction compared to the baseline [9] by inferring per-view per-pixel visibility. The Dep. module enhances reconstruction accuracy and completeness to the state of the art by reducing depth inconsistency. The Nor. module further boosts the overall reconstruction performance by promoting local depth consistency, reducing memory footprint during training when combined with the Dep. module

TABLE IV
ABLATION EXPERIMENTS ON DIFFERENT MODULES OF GE-UWMVS

| Model Settings | | | | Mean Error Distance on 22 **Greenish** Scenes | | | Computational Costs | |
|---|---|---|---|---|---|---|---|---|
| | Ada. | Dep. | Nor. | Acc. ↓ $(mm)$ | Comp. ↓ $(mm)$ | **Overall*** ↓ $(mm)$ | Train / Test Memory† | Train / Test Runtime‡ $(s)$ |
| (a) | | | | 0.382 | 0.388 | 0.385 | 13449 / 4863 | 0.465 / 0.222 |
| (b) | ✓ | | | 0.362 | 0.347 | 0.355 | 13507 / 4239 | 0.434 / 0.185 |
| (c) | ✓ | ✓ | | 0.359 | 0.318 | 0.339 | 13653 / 4239 | 0.456 / 0.185 |
| (d) | ✓ | ✓ | ✓ | **0.355** | **0.286** | **0.321** | 12709 / 10883 | 0.471 / 0.241 |

* **Note that the overall score is the summary measure of the overall reconstruction performance**.
† The batch size is 2 and 1 to measure the memory footprint for train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.
‡ The batch size is 1 to measure the runtime for train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.

by achieving more consistent and compact depth ranges, and retaining comparable efficiency to existing methods [8], [40], [47], [50] despite the increased computational costs caused by higher image and normal resolution during testing.

### D. Real-world Underwater Experiments

We conducted real-world experiments in three distinct underwater scenarios, each at a depth of 10 meters, with varying color casts, visibility conditions, and illumination levels, in Puerto Galera, Philippines. Fig. 6 shows that our GE-UwMVS trained on our dataset demonstrates a high generalization ability to real-world underwater scenarios by reconstructing more complete and denser point clouds with finer details, with color restored [30] during point cloud fusion.

## V. CONCLUSION

We have introduced a physically-guided approach for synthesizing multi-view underwater images and have constructed the first large-scale UwMVS dataset for learning-based methods. Furthermore, we have proposed a novel UwMVS network that enhances performance in underwater point cloud reconstruction. Extensive experiments on dataset and real-world underwater scenes validate the effectiveness of our synthesis approach and demonstrate the state-of-the-art reconstruction performance of our method.

## REFERENCES


[1] A. Pydyn, M. Popek, M. Kubacka, and Ł. Janowski, "Exploration and reconstruction of a medieval harbour using hydroacoustics, 3-d shallow seismic and underwater photogrammetry: A case study from puck, southern baltic sea," *Archaeological Prospection*, vol. 28, no. 4, pp. 527–542, 2021.

[2] B. Joshi, M. Xanthidis, S. Rahman, and I. Rekleitis, "High definition, inexpensive, underwater mapping," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1113–1121.

[3] T. Hitchcox and J. R. Forbes, "Improving self-consistency in underwater mapping through laser-based loop closure," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1873–1892, 2023.

[4] Y. Wang, Y. Ji, H. Tsuchiya, H. Asama, and A. Yamashita, "Learning pseudo front depth for 2d forward-looking sonar-based multi-view stereo," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8730–8737.

[5] M. Qadri, M. Kaess, and I. Gkioulekas, "Neural implicit surface reconstruction using imaging sonar," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1040–1047.

[6] N. Jaber, B. Wehbe, and F. Kirchner, "Sonar2depth: Acoustic-based 3d reconstruction using cgans," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 5828–5835.

[7] S. Arnold and B. Wehbe, "Spatial acoustic projection for 3d imaging sonar reconstruction," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3054–3060.

[8] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[9] 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.

[10] 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.

[11] 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.

[12] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[13] D. Cernea, "OpenMVS: Multi-view stereo reconstruction library," 2020. [Online]. Available: https://cdcseacave.github.io/openMVS

[14] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Q. Xu, W. Kong, W. Tao, and M. Pollefeys, "Multi-scale geometric consistency guided and planar prior assisted multi-view stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] 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.

[17] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1790–1799.

[18] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[19] C. Beall, B. J. Lawrence, V. Ila, and F. Dellaert, "3d reconstruction of underwater structures," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4418–4423.

[20] X. Xu, R. Che, R. Nian, B. He, M. Chen, and A. Lendasse, "Underwater 3d object reconstruction with multiple views in video stream via structure from motion," in *OCEANS 2016 - Shanghai*, 2016, pp. 1–5.

[21] E. Iscar, K. A. Skinner, and M. Johnson-Roberson, "Multi-view 3d reconstruction in underwater environments: Evaluation and benchmark," in *OCEANS 2017 - Anchorage*, 2017, pp. 1–8.

[22] K. A. Skinner, E. Iscar, and M. Johnson-Roberson, "Automatic color correction for 3d reconstruction of underwater scenes," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 5140–5147.

[23] W. Wang, B. Joshi, N. Burgdorfer, K. Batsosc, A. Q. Lid, P. Mordohaia, and I. Rekleitisb, "Real-time dense 3d mapping of underwater environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5184–5191.

[24] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognition*, vol. 98, p. 107038, 2020.

[25] D. H. Foster, "Color constancy," *Vision research*, vol. 51, no. 7, pp. 674–700, 2011.

[26] W. Zhang, Y. Wang, and C. Li, "Underwater image enhancement by attenuated color channel correction and detail preserved contrast enhancement," *IEEE Journal of Oceanic Engineering*, vol. 47, no. 3, pp. 718–735, 2022.

[27] J. Wen, J. Cui, Z. Zhao, R. Yan, Z. Gao, L. Dou, and B. M. Chen, "Syreanet: A physically guided underwater image enhancement framework integrating synthetic and real images," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5177–5183.

[28] Z. Wang, L. Shen, M. Xu, M. Yu, K. Wang, and Y. Lin, "Domain adaptation for underwater image enhancement," *IEEE Transactions on Image Processing*, vol. 32, pp. 1442–1457, 2023.

[29] Y. Wang, J. Guo, W. He, H. Gao, H. Yue, Z. Zhang, and C. Li, "Is underwater image enhancement all object detectors need?" *IEEE Journal of Oceanic Engineering*, 2023.

[30] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li, "Contrastive semi-supervised learning for underwater image restoration via reliable bank," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 145–18 155.

[31] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE Transactions on Image Processing*, 2023.

[32] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6723–6732.

[33] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *IEEE Journal of Oceanic Engineering*, vol. 15, no. 2, pp. 101–111, 1990.

[34] B. McGlamery, "A computer model for underwater camera systems," in *Ocean Optics VI*, vol. 208. SPIE, 1980, pp. 221–231.

[35] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1682–1691.

[36] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[37] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin institute*, vol. 310, no. 1, pp. 1–26, 1980.

[38] S. Galliani, K. Lasinger, and K. Schindler, "Fusible," *https://github.com/kysucix/fusibile*, 2015.

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[40] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654.

[41] G. Yang, X. Zhou, C. Gao, X. Chen, and B. M. Chen, "Learnable cost metric-based multi-view stereo for point cloud reconstruction," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 9, pp. 11 519–11 528, 2024.

[42] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2189–2199.

[43] W. Tong, X. Guan, J. Kang, P. Z. Sun, R. Law, P. Ghamisi, and E. Q. Wu, "Normal assisted pixel-visibility learning with cost aggregation for multiview stereo," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 686–24 697, 2022.

[44] J. Wu, R. Li, H. Xu, W. Zhao, Y. Zhu, J. Sun, and Y. Zhang, "Gomvs: Geometrically consistent cost aggregation for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 207–20 216.

[45] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 283–291.

[46] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 786–10 796.

[47] D. Chang, A. Božič, T. Zhang, Q. Yan, Y. Chen, S. Süsstrunk, and M. Nießner, "Rc-mvsnet: Unsupervised multi-view stereo with neural rendering," in *European conference on computer vision*. Springer, 2022, pp. 665–680.
[48] T. Liu, X. Ye, W. Zhao, Z. Pan, M. Shi, and Z. Cao, "When epipolar constraint meets non-local operators in multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 088–18 097.
[49] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4861–4875, 2020.
[50] G. Yang, X. Zhou, C. Gao, X. Chen, and B. M. Chen, "Learnable cost metric-based multi-view stereo for point cloud reconstruction," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 9, pp. 11 519–11 528, 2024.
[51] "Agisoft metashape: Discover intelligent photogrammetry," *https://www.agisoft.com/*, 2024.
[52] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.