IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.
Cite as T-ASE paper.

Multi-View Stereo With Geometric Encoding for Large-Scale Dense Scene Reconstruction

Guidong Yang, *Member, IEEE*, Rui Cao, *Graduate Student Member, IEEE*, Junjie Wen, *Member, IEEE*, Benyun Zhao, *Graduate Student Member, IEEE*, Qingxiang Li, Xi Chen, Yun-Hui Liu, *Fellow, IEEE*, and Ben M. Chen, *Fellow, IEEE*

***Abstract*—Multi-view stereo (MVS) implicitly encodes photometric and geometric cues into the cost volume for multi-view correspondence matching, transferring insufficient geometric cues essential to depth estimation and reconstruction. This paper proposes GE-MVS, a novel multi-view stereo network with geometric encoding for more accurate and complete depth estimation and point cloud reconstruction. First, the cross-view adaptive cost volume aggregation module is proposed to strengthen the encoding of multi-view geometric cues during cost volume construction. Then, the depth consistency optimization is performed in 3D point space during learning by invoking ground-truth depth cues from adjacent views. Finally, the surface normal geometries are explicitly encoded to refine the sampled depth hypotheses to be consistent in the local neighbor regions. Extensive experiments on the standard MVS benchmarks including DTU, Tanks and Temples, and BlendedMVS demonstrate the state-of-the-art depth estimation and point cloud reconstruction performance of GE-MVS. The GE-MVS is further deployed in real-world experiments for UAV-based large-scale reconstruction, where our method outperforms the prevalent industrial reconstruction solutions in terms of reconstruction efficiency and effectiveness. Supplementary video can be found at https://youtu.be/Z4tGROatVjU**

***Note to Practitioners*—Multi-view stereo (MVS) enables dense point cloud reconstruction of target scenes from calibrated multi-view images and has been widely adopted in robotic navigation, exploration, and manipulation. Recently, learning-based MVS methods have significantly improved reconstruction accuracy and completeness compared to traditional approaches. This work aims to enhance geometric modeling during network learning by utilizing ground-truth depth cues from adjacent views and encoding surface normal geometries. Extensive experiments conducted on both datasets and real-world scenarios validate the effectiveness, scalability, and efficiency of the proposed method. The proposed method can provide accurate depth and dense point cloud representations for applications such as aerial path planning, robotic manipulation, autonomous driving, and virtual and augmented reality. Future work will focus on adapting the proposed method from terrestrial to underwater domains for real-world underwater dense scene reconstruction.**

***Index Terms*—Multi-view stereo, depth estimation, point cloud reconstruction, unmanned aerial vehicle.**

Received 27 September 2024; revised 5 March 2025 and 5 August 2025; accepted 29 September 2025. Date of publication 9 October 2025; date of current version 28 October 2025. This article was recommended for publication by Associate Editor G. Chen and Editor Z. Li upon evaluation of the reviewers' comments. This work was supported in part by the Research Grants Council of Hong Kong, SAR, under Grant 14206821, Grant 14217922, and Grant 14209623; and in part by the InnoHK Clusters of the Hong Kong SAR Government via Hong Kong Centre for Logistics Robotics. *(Corresponding author: Junjie Wen.)*

Guidong Yang, Rui Cao, Benyun Zhao, Qingxiang Li, Xi Chen, Yun-Hui Liu, and Ben M. Chen are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong (e-mail: gdyang@mae.cuhk.edu.hk; rcao@mae.cuhk.edu.hk; byzhao@mae.cuhk.edu.hk; qingxiang.li@polimi.it; xichen@mae.cuhk.edu.hk; yhliu@mae.cuhk.edu.hk; bmchen@mae.cuhk.edu.hk).

Junjie Wen is with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: wenjj@pcl.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TASE.2025.3619093, provided by the authors.

Digital Object Identifier 10.1109/TASE.2025.3619093

I. Introduction

MULTI-VIEW stereo (MVS) reconstructs a dense point cloud representation of the scene from an unordered set of multi-view calibrated images by solving multi-view correspondence matching, and it has been widely adopted in various tasks such as robotic manipulation [1], aerial path planning [2], autonomous driving [3], underwater exploration [4], structural inspection [5] and heritage preservation [6], [7]. Although traditional MVS methods [8], [9], [10], [11], [12] have achieved decent reconstruction performance based on hand-crafted matching metrics, recent learning-based MVS methods [13], [14], [15], [16], [17] significantly outperform their traditional counterparts in terms of reconstruction accuracy and completeness on standard MVS benchmarks [18], [19], [20]. Learning-based MVS methods first adopt deep networks to extract multi-view feature maps. Then, multi-view feature maps and associated camera parameters are implicitly encoded into the cost volume for enhanced multi-view correspondence matching. Afterward, the cost volume is regularized to estimate the depth map. Multi-view depth maps are then filtered and fused to the dense point cloud by applying photometric and geometric constraints as a post-processing step. Despite the promising results exhibited by learning-based MVS methods, the following improvements can be made to further improve the reconstruction performance:

1st **Motivation** MVS with varying viewpoints encounters occlusion, illumination changes, and content variations, where occlusion leads to inaccurate and incomplete depth estimation, and illumination changes make the depth estimation and subsequent point cloud reconstruction of non-Lambertian surfaces more challenging. We observe that source views spatially closer to the reference view tend to exhibit higher feature overlap and thus provide more reliable photometric and geometric cues for depth estimation and subsequent point cloud reconstruction [21], [27], [28]. Motivated by this observation, we propose a cross-view adaptive cost volume aggregation

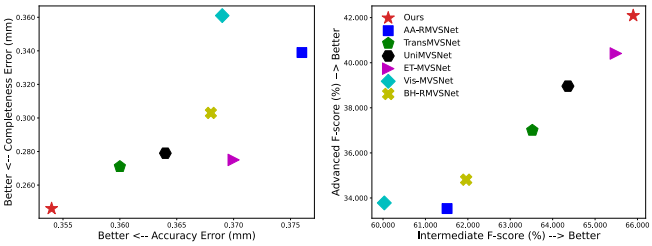
IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

module that enhances geometric cue encoding by inferring per-view, per-pixel visibility weights for pairwise matching costs between the reference and each source view. Our method retains the squared difference for pairwise matching costs due to its intrinsic ability to reflect feature differences, while enhancing computational efficiency through averaging the costs across the color channel dimension. In contrast to existing pixel-wise visibility schemes [21], [25], which typically assign a single confidence or uncertainty value per pixel, our method estimates fine-grained voxel-wise weights over the entire cost volume via a lightweight re-weighting network, enabling more accurate modeling of visibility along both spatial and depth dimensions. To obtain pixel-wise visibility, we compute the maximum weight across depth hypotheses, producing a sharp per-pixel visibility map that reflects the most confident depth response. This per-pixel visibility map is applied as a soft gating mechanism that preserves the full cost structure while suppressing unreliable matches. Experimental results show that the proposed module effectively improves both reconstruction accuracy and completeness by addressing the above challenges while maintaining competitive efficiency.

2^{nd} **Motivation** Most existing methods [22], [23], [24], [25], [26], [28], [29], [30] perform depth inconsistency checks on estimated depth maps by applying photometric and geometric constraints as a post-processing step, where inconsistent pixels are directly discarded during point cloud generation resulting in incomplete reconstruction. The differentiable homography warping exclusively performs implicit geometric modeling during network learning, which delivers insufficient geometric cues essential for depth estimation and subsequent point cloud reconstruction. Unlike existing methods, we perform explicit depth inconsistency check during learning by encoding adjacent source-view ground-truth depth cues to geometrically constrain the depth optimization process directly from the 3D point space. Experimental results show that explicit depth consistency optimization endows the network with the ability for more accurate and complete reconstruction.

3^{rd} **Motivation** Most existing methods [22], [23], [24], [25], [28], [29], [30], [31] adopt a coarse-to-fine framework [15] to gradually refine depth hypotheses of each feature level for memory efficiency and high-resolution reconstruction, where coarse-level depth hypotheses are uniformly sampled from the predefined depth range and finer-level depth hypotheses are dynamically obtained from coarser-level depth estimation. However, resulting depth hypotheses are less satisfactory as the coarser level suffers from inaccurate and incomplete depth estimation, which imposes learning ambiguity on the cross-entropy loss where ground-truth probability volume is obtained by one-hot encoding the depth hypotheses closest to the ground-truth depths. Inspired by [32], [33], [34], and [35], we propose to refine and constrain the sampled depth hypotheses to be geometrically consistent in local neighbor regions by explicitly encoding surface normal geometries. Experimental results show that the normal-assisted depth hypotheses refinement effectively enhances overall reconstruction performance.

We propose a coarse-to-fine MVS network with the above geometric encoding strategies, termed GE-MVS, to demonstrate the superiority of our modules. Extensive experiments indicate that our method achieves state-of-the-art depth estimation and point cloud reconstruction performance on standard MVS benchmarks, including DTU [18], Tanks and Temples [20], and BlendedMVS [19]. As shown in Fig. 1(a), our method achieves more accurate and complete reconstruction on the DTU dataset. Additionally, it effectively generalizes to large-scale complex scenes, obtaining higher reconstruction F-scores on both the intermediate and advanced sets of the Tanks and Temples, as demonstrated in Fig. 1(b). Real-world experiments for large-scale reconstruction based on the unmanned aerial vehicle (UAV) further validate the scalability and generalization ability of GE-MVS, significantly outperforming industrial reconstruction solutions in terms of reconstruction efficiency and effectiveness.



(a) DTU (b) Tanks and Temples

Fig. 1. Reconstruction performance comparison of our method with state-of-the-art learning-based methods [21], [22], [23], [24], [25], [26] on (a) DTU dataset and (b) Tanks and Temples benchmark. Our method obtains more accurate and complete point cloud reconstructions.

This paper is structured as follows: Section II reviews related work. Section III details the methodology of the proposed MVS method. Section IV presents comprehensive benchmark and ablation experimental results. Section V showcases the real-world experiments. Section VI discusses the limitations. Section VII concludes this paper.

II. Related Work

A. Traditional Multi-View Stereo

Traditional multi-view stereo (MVS) methods can be categorized into three main types based on their output representation: point cloud-based methods, volumetric methods, and depth map-based methods. Point cloud-based methods [9], [36] utilize a propagation strategy to sequentially densify a sparse set of key points. However, this sequential approach presents challenges for full parallelization, thereby limiting computational efficiency. In contrast, volumetric methods [37], [38] discretize the three-dimensional space into voxels. These methods employ photometric measures to assess voxel adherence to surfaces, which often leads to high memory consumption. Depth map-based methods [8], [10], [11], [12], [39] have emerged as a more flexible and efficient alternative. They decompose the MVS process into two distinct stages: depth map estimation and depth map filtering and fusion. In the first stage, per-view depth maps are estimated, while the second stage fuses these depth maps into either volumetric representations [40] or point cloud reconstructions [11]. Although traditional approaches achieve detailed reconstructions in rigid Lambertian textured scenarios, they face

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

significant challenges. Variations in illumination, low-textured regions, specular and reflective surfaces, and repeated patterns can lead to unreliable pixel correspondences. Consequently, this results in inaccuracies and incompleteness in the final reconstructions.

B. Learning-Based Multi-View Stereo

Recent advances in learning-based MVS methods have significantly outperformed traditional counterparts in MVS benchmarks [18], [19], [20]. Pioneering works such as SurfaceNet [41] and LSM [42] warp multi-view image features into voxel-based cost volumes regularized by 3D CNNs to regress surface voxels. However, these approaches suffer from common shortcomings associated with volumetric representations, making them challenging to scale for large-scale scene reconstruction. In contrast, MVSNet [13] employs differentiable homography warping to construct the cost volume from multi-view image features and predefined depth hypotheses. It then utilizes 3D CNNs to regularize the cost volume and estimate per-view depth, facilitating large-scale scene reconstruction. MVSNet is regarded as a seminal work in end-to-end learning-based MVS methods. Nevertheless, it struggles to scale to high-resolution images due to its large memory footprint and high computational cost. Additionally, it faces challenges with multi-view matching ambiguity due to its heuristic cost volume aggregation strategy, which overlooks the varying significance of different views.

To achieve high-resolution depth estimation and reconstruction, several variants of MVSNet have been proposed to enhance network scalability for high-resolution images by reducing memory footprint. These approaches often employ a coarse-to-fine framework [15] or utilize recurrent neural networks (RNNs) [14]. RNN-based methods [21], [26], [43] recurrently regularize the cost volume, effectively trading runtime for reduced memory usage. However, this can result in slower inference speeds despite their capability to handle high-resolution images. In contrast, coarse-to-fine methods [22], [23], [28], [29], [30] first infer a low-resolution (coarse) depth map using a predefined depth range with a large depth interval. They then progressively narrow the depth range and interval, which helps reduce runtime and memory consumption while achieving high-resolution (fine) depth map estimation and reconstruction. Specifically, TransMVSNet improves feature matching by introducing an attention-based transformer that aggregates long-range contextual information across multiple views, along with an adaptive receptive field to ensure a smooth transition from local to global features. GeoMVSNet leverages geometric priors at the coarse feature level by integrating coarse-level depth estimation into a dual-branch feature extraction network to enhance geometry awareness. Additionally, it incorporates coarse-level probability volumes into finer-level cost regularization, yielding more robust cost matching. In comparison, our method explicitly optimizes depth consistency by leveraging ground-truth depth cues from adjacent views and refines depth hypotheses through the encoding of surface normal geometries.

C. Cost Volume Aggregation and Geometric Encoding

Our method follows the coarse-to-fine framework to achieve a trade-off between reconstruction efficacy and efficiency. The work most related to ours can be separated into two branches: 1) cost volume aggregation and 2) geometric encoding. For cost volume aggregation, recent studies indicate that different views contribute unequally to the aggregation process. Various modules have been proposed to reduce matching ambiguity by learning view-wise weights [28], patch-wise weights [44], channel-wise weights [27], and pixel-wise weights [21]. In contrast to these existing methods, we propose learning voxel-wise weights through a re-weighting network and taking the maximum weight along the depth dimension to enhance multi-view feature similarity measurement.

Regarding geometric encoding, saddle-shaped depth geometries [45] have been introduced to achieve oscillating depth planes. Geometric structures in coarse stages [30] are embedded in feature fusion and regularization processes to promote geometry awareness. Multi-granularity geometric information [46] is encoded to address disparity ambiguities, while surface normal geometries [34] are incorporated into the regularization network to obtain geometrically consistent cost volumes. However, these methods typically perform depth inconsistency checks after network training, discarding inconsistent pixels during point cloud fusion, which can lead to incomplete reconstructions. In contrast to these approaches, our method performs depth consistency optimization directly in the 3D point space during the learning process by utilizing ground-truth depth cues from adjacent views. Additionally, we explicitly encode surface normal geometries to refine the sampled depth hypotheses, ensuring consistency within local neighboring regions.

D. Real-World Engineering Applications

MVS has demonstrated strong capability in reconstructing accurate and complete dense point clouds from calibrated multi-view images, leading to its widespread adoption across various engineering domains such as simultaneous localization and mapping (SLAM), aerial path planning, autonomous driving, structural inspection, and building energy analysis. Specifically, the first monocular SLAM framework [47] that couples visual odometry with learning-based unsupervised MVS has been proposed to simultaneously achieve high-precision localization and dense scene reconstruction in an efficient manner. A prediction-boosted planning framework [48] has been developed to maximize the performance of geometry-based MVS, enabling autonomous aerial path generation for accurate and complete reconstruction of unknown environments. To enhance perception in dynamic scenes, geometry-based MVS has also been integrated with a risk map-driven deep learning framework [49], which explicitly models motion and intention uncertainties to enable robust and reliable understanding of surrounding environments. In addition, a two-stage ground extraction framework [50] based on adaptive bin partitioning and grid projection has been introduced to achieve a favorable trade-off between extraction accuracy and computational efficiency. A UAV-based

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

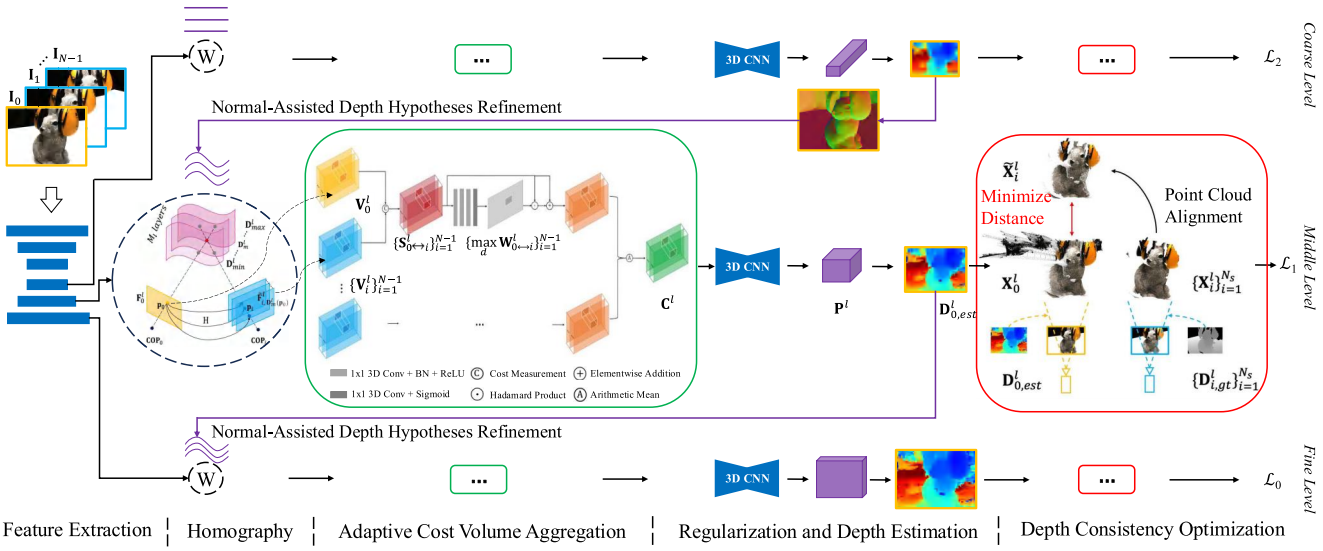


Fig. 2. Network overview of GE-MVS. L is set to 3 to form a three-stage coarse-to-fine network. The green box denotes the proposed cross-view adaptive cost volume aggregation module, the red box represents the proposed depth consistency optimization module, and the violet line indicates the proposed normal-assisted depth hypotheses refinement module. The rest is inherited from our baseline method [15].

intelligent inspection framework [51], [52] leveraging learning-based MVS has also been developed to support scalable infrastructure inspection by decomposing the process into defect detection, structural reconstruction, and defect registration, with MVS providing dense reconstruction as the foundation for downstream defect registration. Furthermore, an autonomous design framework [53], [54] integrating learning-based MVS has been proposed to support decision-making in the deployment of building-integrated photovoltaics based on reconstructed structural models. These diverse applications collectively demonstrate the effectiveness and robustness of MVS in real-world deployments, thereby underscoring the necessity of further enhancing its performance across varying scenarios. In this paper, we conduct extensive real-world experiments to evaluate the effectiveness and efficiency of the proposed method in both terrestrial and underwater environments, covering a wide range of depth scales, illumination conditions, and geometric complexities.

III. Methodology

The architecture of GE-MVS is shown in Fig. 2. Given N-view images $\{\mathbf{I}_i \in \mathbb{R}^{3\times H\times W}\}_{i=0}^{N-1}$ with their camera intrinsics $\{\mathbf{K}_i \in \mathbb{R}^{3\times 3}\}_{i=0}^{N-1}$ and extrinsics $\{[\mathbf{R}_i \in \mathbb{R}^{3\times 3}; \mathbf{t}_i \in \mathbb{R}^{3\times 1}]\}_{i=0}^{N-1}$, our goal is to estimate the depth map $\mathbf{D}_{0,\text{est}} \in \mathbb{R}^{H\times W}$ for the reference image $\mathbf{I}_0$. First, multi-scale feature pyramids of multi-view input images are extracted through a feature extractor with shared weights among multiple views. The initial 3D cost volume pyramid of the reference view is constructed via differentiable homography warping and cross-view adaptive aggregation to measure multi-view matching similarity (Subsection III). Then, the noise-contaminated cost volume pyramid is regularized to acquire the probability volume pyramid for depth map pyramid estimation (Subsection III-B). Afterward, the depth inconsistency of reference-view depth map pyramid is explicitly checked by encoding adjacent source-view ground-truth depth to geometrically constrain the depth optimization from the 3D point space (Subsection III-C). Finally, we refine the sampled depth hypotheses of the coarse-to-fine framework by explicitly encoding surface normal geometries (Subsection III-D). The depth map at the original scale is taken as the final output. The estimated depth maps are filtered and fused [55] to the final point cloud.

A. Cross-View Adaptive Cost Volume Aggregation

1) Feature Volume Construction: Given multi-view images $\{\mathbf{I}_i\}_{i=0}^{N-1}$, the Feature Pyramid Network [56] is adopted to extract L-scale feature pyramids $\{\mathbf{F}_i^l\}_{i=0}^{N-1} \in \mathbb{R}^{C_l\times \frac{H}{2^l}\times \frac{W}{2^l}}$, where $l \in \{0, 1, \ldots, L-1\}$ denotes the feature level and C_l denotes the channel number. For each feature level l, the reference-view depth range $[\mathbf{D}_{\min}^l, \mathbf{D}_{\max}^l]$ is uniformly discretized into M_l discrete depth hypotheses:

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

where $\{\mathbf{D}_{\max}^l, \mathbf{D}_{\min}^l, \mathbf{D}_{\text{ini},m}^l\} \in \mathbb{R}^{\frac{H}{2^l}\times \frac{W}{2^l}}$ denote the maximum, minimum, and sampled initial depth hypotheses at feature level l, respectively. Here, $m \in \{0, 1, \ldots, M_l - 1\}$ denotes the index of depth hypothesis plane. Note that the coarse-level depth range is predefined and the finer-level depth range is obtained from the coarser-level depth estimation. $\mathbf{D}_{\text{ini},m}^l$ is further refined by surface normal geometries to final depth hypotheses $\mathbf{D}_m^l$ for more accurate and complete depth estimation and reconstruction, as illustrated in Subsection III-D.

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

$$\mathbf{p}_i = \mathbf{K}_i\left[\mathbf{R}_{0\to i}\left(\mathbf{K}_0^{-1}\mathbf{p}_0\mathbf{D}_m^l(\mathbf{p}_0)\right) + \mathbf{t}_{0\to i}\right], \tag{2}$$

where $\mathbf{p}_0$ and $\mathbf{p}_i$ denote the reference-view and source-view pixel coordinates, respectively. $\mathbf{R}_{0\to i} = \mathbf{R}_i\mathbf{R}_0^{-1}$ and

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

$\mathbf{t}_{0\to i} = (\mathbf{t}_0 - \mathbf{R}_i\mathbf{R}_0^{-1}\mathbf{t}_i)$ are the relative rotation matrix and translation vector between the reference and source view, respectively. $\mathbf{K}_0$ and $\mathbf{K}_i$ are the scaled camera intrinsics for the reference and source view. Given $\mathbf{p}_i$ from $\mathbf{F}_i^l$, differentiable bilinear interpolation is adopted to interpolate the source-view feature map $\widetilde{\mathbf{F}}_i^l$ aligned to the reference view. The above coordinate mapping and interpolation process are performed for each depth hypothesis $\mathbf{D}_m^l(\mathbf{p}_0)$ to obtain the corresponding feature map $\widetilde{\mathbf{F}}_{i,\mathbf{D}_m^l(\mathbf{p}_0)}^l$, which is consecutively stacked along the depth dimension to construct the source-view feature volume $\left\{\mathbf{V}_i^l \in \mathbb{R}^{C_l\times M_l\times \frac{H}{2^l}\times\frac{W}{2^l}}\right\}_{i=1}^{N-1}$. The reference-view feature map $\mathbf{F}_0^l$ is repeated M_l times along the depth dimension to obtain the reference-view feature volume $\mathbf{V}_0^l \in \mathbb{R}^{C_l\times M_l\times \frac{H}{2^l}\times\frac{W}{2^l}}$.

2) Cost Volume Aggregation: Given per-view feature volumes, the next step is to aggregate multi-view feature volumes into the cost volume to measure multi-view feature matching similarity. The cross-view adaptive cost volume aggregation module is proposed to strengthen the geometric cues implicitly encoded by the homography warping and to remove multi-view matching ambiguities by considering per-view per-pixel visibility. The schematic plot of the cross-view adaptive aggregation module is shown in the green box of Fig. 2. The initial pairwise matching cost between the reference view and i_{th} source view at pixel $\mathbf{p}$ is measured as:

$$\mathbf{S}_{0\leftrightarrow i}^l(\mathbf{p}) = \frac{1}{C_l}\sum_{i=0}^{C_l-1}\left(\mathbf{V}_0^l(\mathbf{p}) - \mathbf{V}_i^l(\mathbf{p})\right)^2, \tag{3}$$

where we first compute the pairwise feature similarity and then compute the mean over the channel dimension for memory efficiency. $\left\{\mathbf{S}_{0\leftrightarrow i}^l \in \mathbb{R}^{M_l\times\frac{H}{2^l}\times\frac{W}{2^l}}\right\}_{i=1}^{N-1}$ denotes the initial pairwise cost, which is smoothed via a lightweight network to produce the per-voxel weight $\left\{\mathbf{W}_{0\leftrightarrow i}^l \in \mathbb{R}^{M_l\times\frac{H}{2^l}\times\frac{W}{2^l}}\right\}_{i=1}^{N-1}$. Multi-view feature volumes are then adaptively aggregated as:

$$\mathbf{C}^l(\mathbf{p}) = \frac{1}{N-1}\sum_{i=1}^{N-1}\underbrace{\left(1+\max_d \mathbf{W}_{0\leftrightarrow i}^l(\mathbf{p})\right)}_{\text{Adaptive Visibility}}\odot\mathbf{S}_{0\leftrightarrow i}^l(\mathbf{p}), \tag{4}$$

where $\mathbf{C}^l \in \mathbb{R}^{M_l\times\frac{H}{2^l}\times\frac{W}{2^l}}$ denotes the reference-view cost volume and $\odot$ is the Hadamard product. The per-view per-pixel adaptive visibility is obtained by taking the maximum similarity along the depth dimension and varying spatial saliency along the height and width dimension. The 1 is added to the adaptive visibility to preserve the initial pairwise cost and prevent excessive smoothing. The per-pixel adaptive visibility is masked over the initial pairwise cost and the resulting pairwise cost is accumulated over all the source views and averaged to obtain the final cost volume. In this way, the pixels with higher feature similarity will have a larger contribution, while pixels that suffer from matching ambiguities will be suppressed during cost volume aggregation.

B. Cost Volume Regularization and Depth Estimation

Following the previous work [15], [22], [23], a multi-scale 3D U-Net is utilized to regularize the initial noise-contaminated cost volume $\mathbf{C}^l \in \mathbb{R}^{M_l\times\frac{H}{2^l}\times\frac{W}{2^l}}$ and to transform multi-view matching cost into 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. The probability volume represents the probability map corresponding to M_l depth hypotheses. The depth estimation is treated as a pixel-wise depth classification problem where the depth hypothesis corresponding to the maximum probability is taken as the depth estimation result:

$$\mathbf{D}_{0,\text{est}}^l(\mathbf{p}) = \underset{d\in\{\mathbf{D}_m^l(\mathbf{p}_0)\}_{m=0}^{M_l-1}}{\arg\max}\ \mathbf{P}_0^l(\mathbf{p}), \tag{5}$$

where $\mathbf{D}_{0,\text{est}}^l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ denotes the depth estimation for the reference view at feature level l. Note that the depth estimation at the fine level is taken as the final output.

C. Depth Consistency Optimization

Most existing methods [22], [23], [24], [25], [26], [27], [28], [29] perform depth inconsistency checks after network learning and discard the inconsistent pixels during point cloud fusion, leading to incomplete reconstruction. Furthermore, the cost volume delivers insufficient geometric cues for depth estimation and subsequent reconstruction. To dynamically improve depth consistency and explicitly strengthen the geometric modeling, we perform depth consistency optimization in the 3D point space by encoding ground-truth depth cues from adjacent views to suppress the inconsistent pixels throughout the learning.

1) 2D→3D Backward Projection: At each feature level l, we first back-project the pixel coordinates from both the reference view and the 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:

$$\mathbf{X}_i^l(\mathbf{p}) = (\mathbf{K}_i\mathbf{R}_i)^{-1}\mathbf{p}\mathbf{D}_i^l(\mathbf{p}) - \mathbf{R}_i^{-1}\mathbf{t}_i, \tag{6}$$

where $\left\{\mathbf{X}_i^l \in \mathbb{R}^{3\times\frac{H}{2^l}\times\frac{W}{2^l}}\right\}_{i=0}^{N_s}$ 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.

2) Point Cloud Alignment: As shown in the red box of Fig. 2, after 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 $\mathbf{D}_{i,\text{gt}}^l$. Furthermore, there also exists misalignment between $\mathbf{X}_0^l$ and $\mathbf{X}_i^l$ due to cumulative errors in camera intrinsics and extrinsics. We hence establish pairwise coordinate mapping to align the source-view point cloud to the reference view. The coordinate mapping is formulated as:

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

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

where $\mathbf{D}^l_{0,\text{est}}(\mathbf{p}_0)$ denotes reference-view depth estimation at pixel $\mathbf{p}_0$. Given source-view pixel coordinates $\mathbf{p}_i$ from $\mathbf{X}^l_i$, we regard x, y, z point coordinates as channel features and utilize the differentiable bilinear interpolation to obtain the source-view point cloud $\widetilde{\mathbf{X}}^l_i$ aligned to the reference view.

3) Depth Inconsistency Check: The pointwise distance error is then computed as the Euclidean norm between reference-view points and aligned source-view points:

$$\mathbf{E}^l_{0\leftrightarrow i}(\mathbf{p}) = \|\mathbf{X}^l_0(\mathbf{p}) - \widetilde{\mathbf{X}}^l_i(\mathbf{p})\|_2, \tag{8}$$

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

$$\mathbf{M}^l_0(\mathbf{p}) = \frac{1}{N_s}\sum_{i=1}^{N_s}[\mathbf{E}^l_{0\leftrightarrow i}(\mathbf{p}) > \epsilon_l], \tag{9}$$

where $\mathbf{M}^l_0 \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ denotes the depth inconsistency mask of the reference view at feature level l. $[\cdot]$ denotes the Iverson bracket and ϵ_l denotes the per-level point distance threshold.

4) Loss Function: As illustrated in Subsection III-B, the depth estimation is treated as a pixel-wise depth classification problem, where the depth hypothesis corresponding to the maximum probability is taken as the estimation result. Therefore, the cross-entropy loss is adopted to supervise the difference between estimated probability volume $\mathbf{P}^l(\mathbf{p})$ and ground-truth probability volume $\mathbf{P}^l_{\text{gt}}(\mathbf{p})$, where $\mathbf{P}^l_{\text{gt}}(\mathbf{p})$ is obtained by one-hot encoding the depth hypotheses closest to the ground-truth depths. The cross-entropy loss of feature level l is defined as follows:

$$\mathcal{L}^l_{CE} = \sum_{\mathbf{p}\in\{\mathbf{p}_v\}}\sum_{m=0}^{M_l-1} -\mathbf{P}^l_{\text{gt},m}(\mathbf{p})\log(\mathbf{P}^l_m(\mathbf{p})), \tag{10}$$

where $\{\mathbf{P}^l_{\text{gt},m}, \mathbf{P}^l_m\} \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ are ground-truth and estimated probability map of m_{th} depth hypothesis. $\{\mathbf{p}_v\}$ is the set of pixels with valid ground-truth depth.

The per-level cross-entropy loss $\mathcal{L}^l_{CE}$ is further weighted by the depth inconsistency mask $\mathbf{M}^l_0(\mathbf{p})$ to geometrically optimize depth consistency:

$$\mathcal{L}_l = \mathcal{L}^l_{CE} + \sum_{\mathbf{p}\in\{\mathbf{p}_v\}}\sum_{m=0}^{M_l-1} -\mathbf{M}^l_0(\mathbf{p})\left(\mathbf{P}^l_{\text{gt},m}(\mathbf{p})\log(\mathbf{P}^l_m(\mathbf{p}))\right), \tag{11}$$

where $\mathcal{L}_l$ denotes the per-level loss for depth optimization, $\mathbf{M}^l_0(\mathbf{p})$ provides the per-pixel penalty for reference-view depth inconsistent to N_s source views, the original cross-entropy loss is retained to prevent excessive depth consistency correction. The total loss for optimization is the weighted sum of the per-level loss:

$$\mathcal{L} = \sum_{l=0}^{L-1}\lambda_l\mathcal{L}_l, \tag{12}$$

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

D. Normal-Assisted Depth Hypotheses Refinement

Coarse-to-fine depth hypotheses sampling introduces learning ambiguity into the network. Furthermore, sampled depth hypotheses lead to inconsistent depth estimation in the local neighbor regions under challenging multi-view matching conditions. Inspired by [32], [34], and [35], we propose encoding surface normal geometries to refine the depth hypotheses, making them geometrically smooth and consistent in local neighbor regions. The normal map is generated by the monocular normal estimation network Omnidata [57] to help resolve multi-view matching ambiguities under challenging conditions.

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

$$\mathbf{X}(\mathbf{p}) = \mathbf{K}^{-1}\mathbf{p}\mathbf{D}^l_{\text{ini},m}(\mathbf{p}), \tag{13}$$

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

$$\mathbf{N}(\mathbf{p}_i)\cdot(\mathbf{X}(\mathbf{p}_i) - \mathbf{X}(\mathbf{p})) = 0, \tag{14}$$

where $\mathbf{N}(\mathbf{p}_i) \in \mathbb{R}^{3\times 1}$ denotes the normal vector at pixel $\mathbf{p}_i$, $i \in \{0, 1, \ldots, n-1\}$. We then refine neighboring depth hypotheses by encoding normal geometries according to Eq. 14:

$$\mathbf{D}^l_m(\mathbf{p}_i) = \frac{\mathbf{N}(\mathbf{p}_i)\cdot(\mathbf{K}^{-1}\mathbf{p})}{\mathbf{N}(\mathbf{p}_i)\cdot(\mathbf{K}^{-1}\mathbf{p}_i)}\mathbf{D}^l_{\text{ini},m}(\mathbf{p}), \tag{15}$$

where $\mathbf{D}^l_m(\mathbf{p}_i)$ denotes the refined depth hypothesis at pixel $\mathbf{p}_i$. The above refinement is performed for all M_l depth hypotheses to obtain the refined depth hypotheses $\mathbf{D}^l_m \in \mathbb{R}^{H\times W}, m \in \{0, 1, \ldots, M_l - 1\}$, which is interpolated to corresponding feature level $\mathbf{D}^l_m \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}}$ for adaptive cost volume aggregation, depth optimization and estimation.

IV. Benchmark and Ablation Experiments

In this section, we first illustrate the datasets and evaluation metrics for quantifying point cloud reconstruction and depth estimation performance. Next, we provide the implementation details and evaluate our method on three standard MVS benchmarks [18], [19], [20]. Finally, we perform systematic ablation experiments to validate the effectiveness and efficiency of each component of the proposed method.

A. Datasets

The DTU dataset [18] is an indoor MVS dataset comprising 119 object-centric scenes. Each scene contains multi-view images captured from 49 fixed camera positions under 7 different lighting conditions, with ground-truth point clouds scanned by the structured light scanner. The dataset is split

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

into 79 scenes for training, 22 scenes for validation, and 18 scenes for evaluation by following MVSNet [13]. The BlendedMVS dataset [19] is a large-scale dataset for MVS fine-tuning and validation, comprising 113 complex scenes, including cities, architectures, statues, small objects, and more. The dataset is split into 106 scenes for training and 7 scenes for validation. The Tanks and Temples [20] dataset serves as a large-scale MVS evaluation benchmark consisting of indoor and outdoor scenes. The dataset contains intermediate (8 scenes) and advanced sets (6 scenes) with variations in scene scale, exposure condition, and surface reflection.

B. Evaluation Metrics

1) DTU: The DTU dataset computes mean error distance in millimeter (*mm*) to quantify the reconstruction accuracy and completeness. For a given scene, let $\mathcal{R}$ and $\mathcal{G}$ denote the reconstructed and ground-truth point clouds, respectively. Accuracy and completeness are defined as follows:

$$e_{\mathbf{r}\to\mathcal{G}} = \min_{\mathbf{g}\in\mathcal{G}} \|\mathbf{r}-\mathbf{g}\|_2, \tag{16}$$

$$Acc = \frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}} [e_{\mathbf{r}\to\mathcal{G}} < d]\cdot e_{\mathbf{r}\to\mathcal{G}}, \tag{17}$$

$$e_{\mathbf{g}\to\mathcal{R}} = \min_{\mathbf{r}\in\mathcal{R}} \|\mathbf{g}-\mathbf{r}\|_2, \tag{18}$$

$$Comp = \frac{1}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}} [e_{\mathbf{g}\to\mathcal{R}} < d]\cdot e_{\mathbf{g}\to\mathcal{R}}, \tag{19}$$

where $\mathbf{r}$ and $\mathbf{g}$ represent every point in $\mathcal{R}$ and $\mathcal{G}$, respectively. $\|\cdot\|_2$ denotes the Euclidean distance, $|\cdot|$ represents the number of points, $[\cdot]$ is the Iverson bracket, and d denotes the outlier rejection threshold. Specifically, the accuracy measures the distance from $\mathcal{R}$ to $\mathcal{G}$, reflecting the quality of the reconstructed points, i.e., how close $\mathcal{R}$ lies to $\mathcal{G}$. The completeness is measured as the distance from $\mathcal{G}$ to $\mathcal{R}$, representing the extent to which the ground truth $\mathcal{G}$ is restored.

DTU reports mean accuracy error and mean completeness error over 22 test scenes. To avoid the imbalance between mean accuracy and completeness, the overall score takes the average of mean accuracy and mean completeness to measure the overall reconstruction performance. A lower overall score indicates better reconstruction performance:

$$Overall\ Score = \frac{1}{2N_t}\sum_{i=0}^{N_t-1}(Acc_i + Comp_i) \tag{20}$$

where N_t denotes the total number of test scenes.

2) Tanks and Temples: The Tanks and Temples dataset adopts the precision and recall in percentage (%) to quantify the reconstruction accuracy and completeness, respectively. The definitions of precision and recall are analogous to accuracy and completeness as defined for the DTU dataset. However, while the DTU dataset measures error distances in millimeters (mm), the Tanks and Temples dataset reports errors as percentages (%). For each scene, the precision and recall are defined as follows:

$$e_{\mathbf{r}\to\mathcal{G}} = \min_{\mathbf{g}\in\mathcal{G}} \|\mathbf{r}-\mathbf{g}\|_2, \tag{21}$$

$$Precision = \frac{100}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}} [e_{\mathbf{r}\to\mathcal{G}} < d], \tag{22}$$

$$e_{\mathbf{g}\to\mathcal{R}} = \min_{\mathbf{r}\in\mathcal{R}} \|\mathbf{g}-\mathbf{r}\|_2, \tag{23}$$

$$Recall = \frac{100}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}} [e_{\mathbf{g}\to\mathcal{R}} < d]. \tag{24}$$

To achieve a trade-off between precision and recall, the F-score is defined as their harmonic mean, serving as a summary measure of reconstruction performance. A high F-score indicates that the reconstruction is both accurate and complete:

$$F\text{-}score = \frac{2\cdot Precision\cdot Recall}{Precision + Recall}. \tag{25}$$

Tanks and Temples computes the mean F-score across 8 scenes in the intermediate set and 6 scenes in the advanced set as official metrics.

3) BlendedMVS: Unlike the previous two datasets focusing on evaluating point cloud reconstruction performance, the BlendedMVS dataset adopts the endpoint error (EPE), 1-threshold error (e_1), and 3-threshold error (e_3) to quantify the depth estimation performance. The EPE is defined as the mean absolute error between the predicted and ground-truth depth maps, with the absolute error scaled by the depth interval. The e_1 and e_3 denote the percentages of pixels in the predicted depth map with scaled absolute errors greater than 1 and 3, respectively.

C. Implementation Details

The proposed network is implemented by PyTorch and trained on the DTU training set. The original DTU dataset [18] only contains ground-truth point clouds scanned by the structured light scanner. Therefore, following the common practices [13], [15], [30], we generate the per-view ground-truth depth map by screened Poisson surface reconstruction and per-view depth map rendering for end-to-end training. 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 ϵ_l for depth inconsistency check are set to 8 and 0.2, respectively. The image and ground-truth depth map resolutions are resized to $H \times W = 512 \times 640$. The depth range is set to 425 *mm* to 905 *mm* to uniformly sample depth hypotheses, where the number of depth hypotheses planes M_l is defined as $48, 32, 8$ from coarse to fine level, respectively. The corresponding depth interval is set to $4, 2, 1$ times of the coarse-level depth interval and the loss weight λ_l is $1, 1, 2$ from coarse to fine level. The number of square neighboring pixels n is set to 8 to form a 3×3 local neighbor region for depth hypotheses refinement. The network is optimized by the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) for 30 epochs on two NVIDIA RTX 3090Ti GPUs with a batch size of 2 on each GPU. The initial learning rate and weight decay of the optimizer are 0.001 and 0.0001, respectively. The multi-step learning rate scheduler is adopted to decay the initial learning rate by a factor of 0.5 at epochs $8, 12, 16$, and 20, respectively.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE I

QUANTITATIVE BENCHMARKING RESULTS ON THE DTU EVALUATION SET FOR EVALUATING POINT CLOUD RECONSTRUCTION PERFORMANCE ($N = 5$, $H \times W = 864 \times 1152$, LOWER IS BETTER)

Type	Methods	Year	Mean Error Distance on 22 Scenes		
			Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)
Traditional	Camp [8]	2008	0.835	0.554	0.695
	Furu [9]	2010	0.613	0.941	0.777
	Tola [10]	2012	0.342	1.190	0.766
	Gipuma [11]	2015	**0.283**	0.873	0.578
	Colmap [12]	2016	0.400	0.664	0.532
Learning-based	SurfaceNet [41]	2017	0.450	1.040	0.745
	MVSNet [13]	2018	0.396	0.527	0.462
	R-MVSNet [14]	2019	0.385	0.459	0.422
	P-MVSNet [44]	2019	0.406	0.434	0.420
	Point-MVSNet [58]	2019	0.342	0.411	0.376
	D^2HC-RMVSNet [43]	2020	0.395	0.378	0.386
	AttMVS [27]	2020	0.383	0.329	0.356
	CVP-MVSNet [59]	2020	0.296	0.406	0.351
	UCS-Net [60]	2020	0.338	0.349	0.344
	AA-RMVSNet [21]	2021	0.376	0.339	0.357
	EPP-MVSNet [61]	2021	0.413	0.296	0.355
	PatchMatchNet [62]	2021	0.427	0.277	0.352
	IterMVS [63]	2022	0.373	0.354	0.363
	CDS-MVSNet† [29]	2022	0.365	0.281	0.323
	UniMVSNet† [23]	2022	0.364	0.279	0.321
	TransMVSNet† [22]	2022	0.360	0.271	0.316
	Vis-MVSNet [25]	2023	0.369	0.361	0.365
	IGEV-MVS† [46]	2023	0.331	0.316	0.324
	GeoMVSNet† [30]	2023	0.370	0.275	0.323
	ET-MVSNet† [24]	2023	0.359	0.265	0.312
	BH-RMVSNet [26]	2024	0.368	0.303	0.335
	LCM-MVSNet [28]	2024	0.358	0.275	0.317
	CasMVSNet† (Baseline) [15]	2020	0.359	0.339	0.349
	Ours		0.354	**0.246**	**0.300**
	Rela. Improvement (%)		1.39	27.43	14.04
	Ours (w/ Metric3D-Giant)		0.349	**0.233**	**0.291**
	Rela. Improvement (%)		2.78	31.27	16.62

* The overall score is the summary measure of the overall reconstruction performance.
↓ Smaller values indicate better performance.
† Re-evaluated by utilizing the released optimal checkpoints with the same depth map filtering & fusion method and in the same platform as ours.

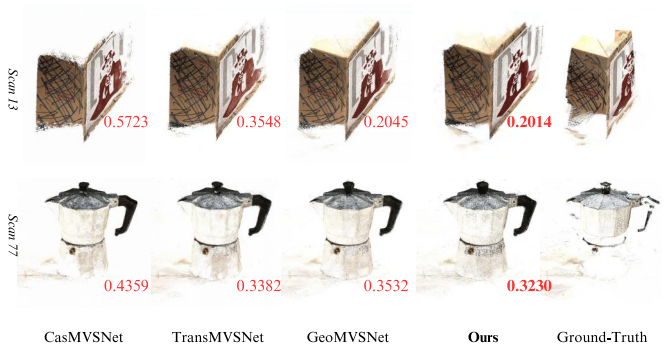


Fig. 3. Reconstruction comparison of *scan 13* and *scan 77* on the DTU evaluation set, where our method achieves more complete dense reconstruction for low-textured and non-Lambertian surfaces under bright light. The bottom-right number denotes the completeness error in mm (**lower the better**).

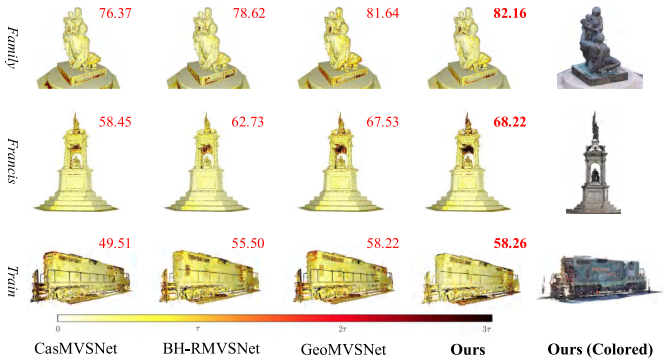


Fig. 4. Reconstruction error rendering of scene *Family*, *Francis*, and *Train* on the intermediate set of the Tanks and Temples benchmark. The darker color indicates a larger reconstruction error and τ denotes the per-scene point distance threshold. The top-right number denotes the F-score in % (**higher the better**).

D. Benchmarking Performance

1) Benchmarking on DTU: We benchmark our method on the DTU evaluation set by comparing it with several traditional and dozens of learning-based MVS methods to quantify its point cloud reconstruction performance as shown in Table I. The reconstruction accuracy (Acc.), completeness (Comp.), and overall score (Overall) in mean error distance (mm) of 22 scenes are reported (**lower the better**). N is set to 5 with the image resolution of $H \times W = 864 \times 1152$ for depth map estimation. The fusible [55] is adopted to fuse multi-view depth maps into the final point cloud. All settings follow common practices for a fair comparison. Extensive experiments demonstrate that our method outperforms pioneering works, such as P-MVSNet [44], which constructs the cost volume through patch-wise matching confidence aggregation. Our method achieves state-of-the-art reconstruction performance in terms of overall score by striking an excellent trade-off between reconstruction accuracy and completeness, verifying the effectiveness of our proposed modules for improving point cloud reconstruction. The qualitative comparison between our method and existing methods [15], [22], [30] in Fig. 3 demonstrates that our method achieves much more complete reconstruction with fine-grained details for non-Lambertian and low-textured surfaces under bright light, qualitatively demonstrating the quantitative benchmarking results.

2) Benchmarking on Tanks and Temples: We further quantify the point cloud reconstruction performance of our method on the Tanks and Temples benchmark to evaluate its generalization ability on large-scale outdoor and indoor scenes. For the intermediate set, the mean precision, recall, and F-score in mean error percentage (%) of 8 outdoor scenes are reported (**higher the better**). For the advanced set, the mean precision, recall, and F-score in mean error percentage (%) of 6 indoor and outdoor scenes are reported (**higher the better**).

The model is fine-tuned on the BlendedMVS training set for 20 epochs to improve its generalization ability on large-scale real-world scenes by setting $N = 7$ with $H \times W = 576 \times 768$. For benchmarking, N is set to 11 with $H \times W = 1080 \times 1920$ for depth estimation, and the dynamic fusion [43] is adopted for point cloud reconstruction. All settings follow common practices for a fair comparison. The benchmarking results on the intermediate set and advanced set of the Tanks and Temples are summarized in Table II and Table III, respectively. Our method achieves the highest F-score compared to traditional and dozens of learning-based methods across large-scale indoor and outdoor scenes. We render the reconstruction errors as shown in Fig. 4 and Fig. 5, where our method outperforms recent methods [15], [26], [28], [30] by a large margin.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

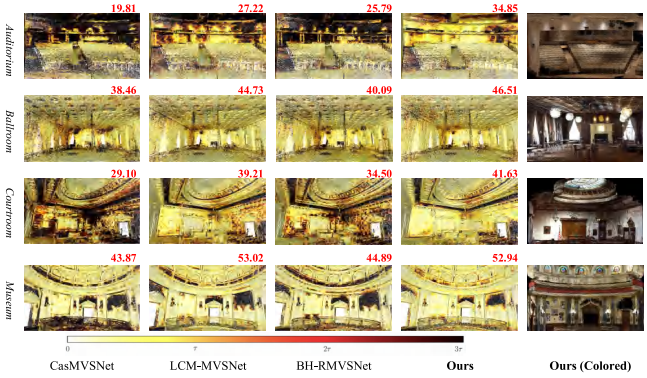


Fig. 5. Reconstruction error rendering of scene *Auditorium*, *Ballroom*, *Courtroom*, and *Museum* on the advanced set of the Tanks and Temples benchmark. The darker color indicates a larger reconstruction error and τ denotes the per-scene point distance threshold. The top-right number denotes the F-score in % (**higher the better**).

TABLE II

QUANTITATIVE BENCHMARKING RESULTS ON THE INTERMEDIATE SET OF TANKS AND TEMPLES FOR EVALUATING POINT CLOUD RECONSTRUCTION PERFORMANCE ($N = 11$, $H \times W = 1080 \times 1920$, HIGHER IS BETTER)

Type	Methods	Year	Mean Error Percentage on 8 Scenes		
			Precision ↑ (%)	Recall ↑ (%)	F-score* ↑ (%)
Traditional	MVE [64]	2015	19.67	40.16	25.37
	OpenMVG [65] + MVE [64]	2016	27.96	61.97	38.00
	Colmap [12]	2016	43.16	44.48	42.14
	Pix4D [66]	2016	46.85	41.58	43.24
	VisualSfM [67] + OpenMVS [68]	2020	21.76	30.39	24.45
	OpenMVG [66] + OpenMVS [69]	2020	35.25	55.16	41.71
Learning-based	MVSNet [13]	2018	40.23	49.70	43.48
	Point-MVSNet [58]	2019	41.27	60.13	48.27
	R-MVSNet [14]	2019	43.74	57.60	48.40
	P-MVSNet [44]	2019	49.93	63.82	55.62
	CVP-MVSNet [59]	2020	51.41	60.19	54.03
	UCSNet [60]	2020	46.66	70.34	54.83
	D^2HC-RMVSNet [43]	2020	49.88	74.08	59.20
	AttMVS [27]	2020	**61.89**	58.93	60.05
	PatchMatchNet [62]	2021	43.64	69.37	53.15
	AA-RMVSNet [21]	2021	52.68	75.69	61.51
	EPP-MVSNet [61]	2021	53.03	75.58	61.68
	IterMVS [63]	2022	46.82	73.50	56.22
	CDS-MVSNet [29]	2022	53.23	74.39	61.58
	TransMVSNet [22]	2022	55.14	76.73	63.52
	UniMVSNet [23]	2022	57.54	73.82	64.36
	Vis-MVSNet [25]	2023	54.44	70.48	60.03
	ET-MVSNet [24]	2023	58.52	75.45	65.49
	GeoMVSNet [30]	2023	59.75	74.28	65.89
	BH-RMVSNet [26]	2024	53.24	75.82	61.96
	LCM-MVSNet [28]	2024	59.25	68.56	63.33
	CasMVSNet [15] (Baseline)	2020	47.62	74.01	56.84
	Ours		57.17	**79.12**	**65.90**
	Rela. Improvement (%)		20.05	6.90	15.94
	Ours (w/ Metric3D-Giant)		58.02	**78.50**	**66.20**
	Rela. Improvement (%)		21.84	6.07	16.47

* The F-score is the summary measure of the overall reconstruction performance.
↑ Larger values indicate better performance.

3) Benchmarking on BlendedMVS: Unlike previous benchmarks, BlendedMVS reports the endpoint error (EPE), 1-threshold error (e_1), and 3-threshold error (e_3) to quantify the depth estimation performance. For fine-tuning and benchmarking, N is set to 5 with $H \times W = 576 \times 768$ by following common practices. Extensive experiments in Table IV demonstrate that our method obtains the lowest depth estimation error compared to recent learning-based MVS methods. We qualitatively compare the depth estimates of our method with those from recent learning-based MVS methods [15], [22], [30] on the BlendedMVS validation set [19] as shown in Fig. 6, where our method achieves more accurate and complete depth estimation for small-scale to large-scale scenes with varying depth ranges, geometry complexities, surface textures, and illumination conditions. This qualitative comparison corroborates the quantitative results in Table IV.

TABLE III

QUANTITATIVE BENCHMARKING RESULTS ON THE ADVANCED SET OF TANKS AND TEMPLES FOR EVALUATING POINT CLOUD RECONSTRUCTION PERFORMANCE ($N = 11$, $H \times W = 1080 \times 1920$, HIGHER IS BETTER)

Type	Methods	Year	Mean Error Percentage on 8 Scenes		
			Precision ↑ (%)	Recall ↑ (%)	F-score* ↑ (%)
Traditional	MVE [64]	2015	15.21	24.62	18.28
	OpenMVG [65] + MVE [64]	2016	19.49	31.29	22.93
	Pix4D [66]	2016	31.33	21.78	25.07
	Colmap [12]	2016	33.65	23.96	27.24
	VisualSfM [67] + OpenMVS [68]	2020	15.33	12.85	12.70
	OpenMVG [65] + OpenMVS [68]	2020	23.74	13.29	21.85
Learning-based	R-MVSNet [14]	2019	31.47	22.05	24.91
	AttMVS [27]	2020	**40.58**	27.26	31.93
	PatchMatchNet [62]	2021	27.27	41.66	32.31
	AA-RMVSNet [21]	2021	37.46	33.01	33.53
	EPP-MVSNet [61]	2021	40.09	34.63	35.72
	CDS-MVSNet [29]	2022	30.73	35.09	31.94
	IterMVS [63]	2022	28.04	42.60	33.24
	TransMVSNet [22]	2022	33.84	44.29	37.00
	UniMVSNet [23]	2022	33.76	47.22	38.96
	Vis-MVSNet [25]	2023	30.16	41.42	33.78
	ET-MVSNet [24]	2023	33.44	51.83	40.41
	GeoMVSNet [30]	2023	37.56	47.74	41.52
	BH-RMVSNet [26]	2024	30.79	43.85	34.81
	LCM-MVSNet [28]	2024	33.56	46.26	38.54
	CasMVSNet [15] (Baseline)	2020	29.68	35.24	31.12
	Ours		36.97	50.81	**42.09**
	Rela. Improvement (%)		24.56	44.18	35.25
	Ours (w/ Metric3D-Giant)		34.85	**55.21**	**42.38**
	Rela. Improvement (%)		17.42	56.67	36.18

* The F-score is the summary measure of the overall reconstruction performance.
↑ Larger values indicate better performance.

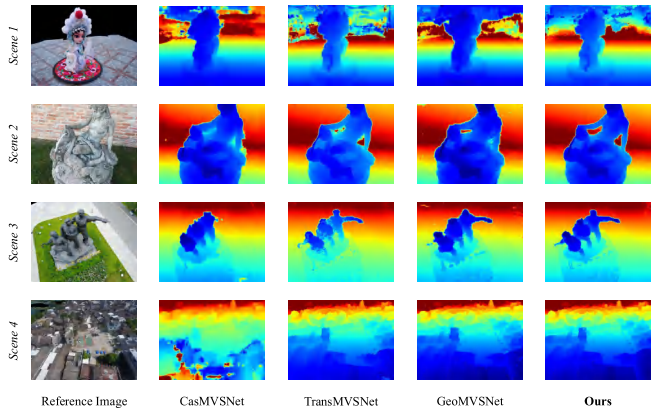


Fig. 6. Depth estimation results of recent learning-based methods [15], [22], [30] and our approach, evaluated across scenes with varying scales, depth ranges, geometry complexities, surface textures, and illumination conditions from the BlendedMVS validation set.

E. Ablation Experiments

1) Module Effectiveness and Efficiency: As shown in Table V, we perform an ablation study on the DTU

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE IV

QUANTITATIVE BENCHMARKING RESULTS ON THE BLENDEDMVS VALIDATION SET FOR EVALUATING DEPTH ESTIMATION PERFORMANCE ($N = 5$, $H \times W = 576 \times 768$, LOWER IS BETTER)

Methods	Year	Mean Depth Error on 7 Scenes		
		EPE ↓	e_1 (%) ↓	e_3 (%) ↓
MVSNet [13]	2018	1.49	21.98	8.32
CVP-MVSNet [59]	2020	1.90	19.73	10.24
EPP-MVSNet [61]	2021	1.17	12.66	6.20
CDS-MVSNet [29]	2022	1.80	22.88	9.28
TransMVSNet [22]	2022	1.05	13.74	5.47
UniMVSNet [23]	2022	1.17	11.27	4.96
IterMVS [63]	2022	0.87	12.15	4.48
ET-MVSNet [24]	2023	3.44	23.40	12.18
GeoMVSNet [30]	2023	2.37	23.20	11.76
Vis-MVSNet [25]	2023	1.56	21.68	8.36
LCM-MVSNet [28]	2024	1.02	10.15	4.54
CasMVSNet (Baseline) [15]	2020	1.43	19.73	10.24
Ours		**0.77**	**9.54**	**3.96**
Ours (w/ Metric3D-Giant)		**0.62**	**7.11**	**3.27**

↓ Smaller values indicate better performance.

TABLE V

ABLATION STUDY OF EFFECTIVENESS OF EACH PROPOSED MODULE ON THE DTU EVALUATION SET

Model Settings				Mean Error Distance on 22 Scenes			Computational Costs	
	Ada.	Dep.	Nor.	Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)	Train / Test Memory† (MB)	Train / Test Runtime‡ (s)
(a)				0.359 (+0.0009)	0.339 (+0.0008)	0.349 (+0.0009)	13449 / 4863	0.465 / 0.222
(b)	✓			0.357 (+0.0001)	0.314 (+0.0005)	0.335 (+0.0008)	13507 / 4239	0.434 / 0.185
(c)	✓	✓		**0.343** (+0.0008)	0.272 (+0.0006)	0.308 (+0.0002)	13653 / 4239	0.456 / 0.185
(d)	✓	✓	✓	0.354 (+0.0007)	**0.246** (+0.0009)	**0.300** (+0.0008)	12709 / 10883	0.471 / 0.241

* The overall score is the summary measure of the overall reconstruction performance.
↓ Smaller values indicate better performance.
† The batch size is set to 2 and 1 to measure the memory footprint in the train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.
‡ The batch size is set to 1 to measure the runtime in the train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.

evaluation set to verify the effectiveness and efficiency of the proposed modules of our method, including 1) the cross-view adaptive cost volume aggregation (Ada.) module, 2) the depth consistency optimization (Dep.) module, and 3) the normal-assisted depth hypotheses refinement (Nor.) module. The experiments show that the Ada. module significantly enhances overall reconstruction performance by considering per-view, per-pixel visibility, outperforming the baseline method [15]. This improvement is attributed to the fact that, in contrast to the baseline method which assigns equal weight to each source view, the Ada. module dynamically infers per-view, per-pixel visibility. By leveraging the fact that source views closer to the reference view generally exhibit greater feature overlap, the module provides more reliable photometric and geometric cues. The Ada. module introduces a negligible increase in memory usage during training and reduces memory consumption during testing, while consistently accelerating runtime performance. The Dep. module further improves reconstruction accuracy and completeness, achieving state-of-the-art performance by enhancing depth consistency in the 3D point space through the incorporation of ground-truth depth cues from adjacent views. Notably, the Dep. module incurs only a minimal increase in computational cost during training. The Nor. module further enhances reconstruction performance by refining depth hypotheses to ensure they are geometrically smooth and consistent within local neighborhoods. During training, the Nor. module reduces memory consumption when combined with the classification loss-based Dep. module, resulting in more consistent and compact depth ranges. Although the Nor. module incurs additional computational costs during testing due to the increased image and normal resolution, its efficiency remains comparable to existing methods [23], [26], [28], [59].

A qualitative comparison of the point cloud reconstructions on the DTU evaluation set with different model configurations is presented in Fig. 7. Scans 33 and 49 are selected due to their intricate geometric structures, while scans 48 and 77 are chosen for their non-Lambertian, low-textured surfaces subjected to bright lighting conditions. Fig. 7(a) shows the incomplete point cloud reconstructions produced by the baseline method [15]. Replacing the heuristic variance-based cost volume aggregation with the proposed Ada. module effectively achieves more complete and dense reconstructions, as shown in Fig. 7(b). As shown in Fig. 7(c), the integration of the proposed Dep. module further improves point cloud completeness, particularly for scan 48, which contains low-textured surfaces under bright light exposure. Fig. 7(d)–(f) demonstrates the effectiveness of the proposed Nor. module in improving reconstruction completeness, where (d), (e), and (f) utilize surface normal maps predicted by the monocular normal estimation networks Omnidata [57], Metric3D-Large [69], and Metric3D-Giant [69], respectively. Notably, the angular errors of the Metric3D models are lower than those of Omnidata, with normal estimation accuracy improving from left to right. The improved accuracy of the normal maps leads to enhanced reconstruction performance. It should be noted that the focus here is on point cloud completeness, as highlighted in the qualitative comparison. The proposed modules sequentially improve the accuracy and completeness of the point cloud reconstruction compared to the baseline method, as quantitatively verified in Table V.

2) Adaptive Cost Volume Aggregation: We conduct experiments to compare the proposed cross-view adaptive cost volume aggregation module with several existing methods [13], [21], [25], [28], [70]. To ensure a fair comparison, all methods are integrated into the baseline framework, CasMVS-Net [15], a seminal work that introduced coarse-to-fine depth estimation, enabling efficient high-resolution depth estimation and point cloud reconstruction, and serving as the baseline for state-of-the-art methods.

The variance-based aggregation [13] uses the squared difference to measure pairwise matching costs between the reference-view and source-view feature volumes, treating each source-view cost equally when aggregating them into the final cost volume. However, it overlooks the varying significance of different source views, as those closer to the reference view typically provide more reliable cues for depth estimation. The view-wise learnable cost metric [28] utilizes squared difference, assigning a learnable parameter to the reference-view cost and computing a normalized matching score to weight each source-view cost, which are then adaptively aggregated into the final cost volume. The voxel-wise adaptive

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

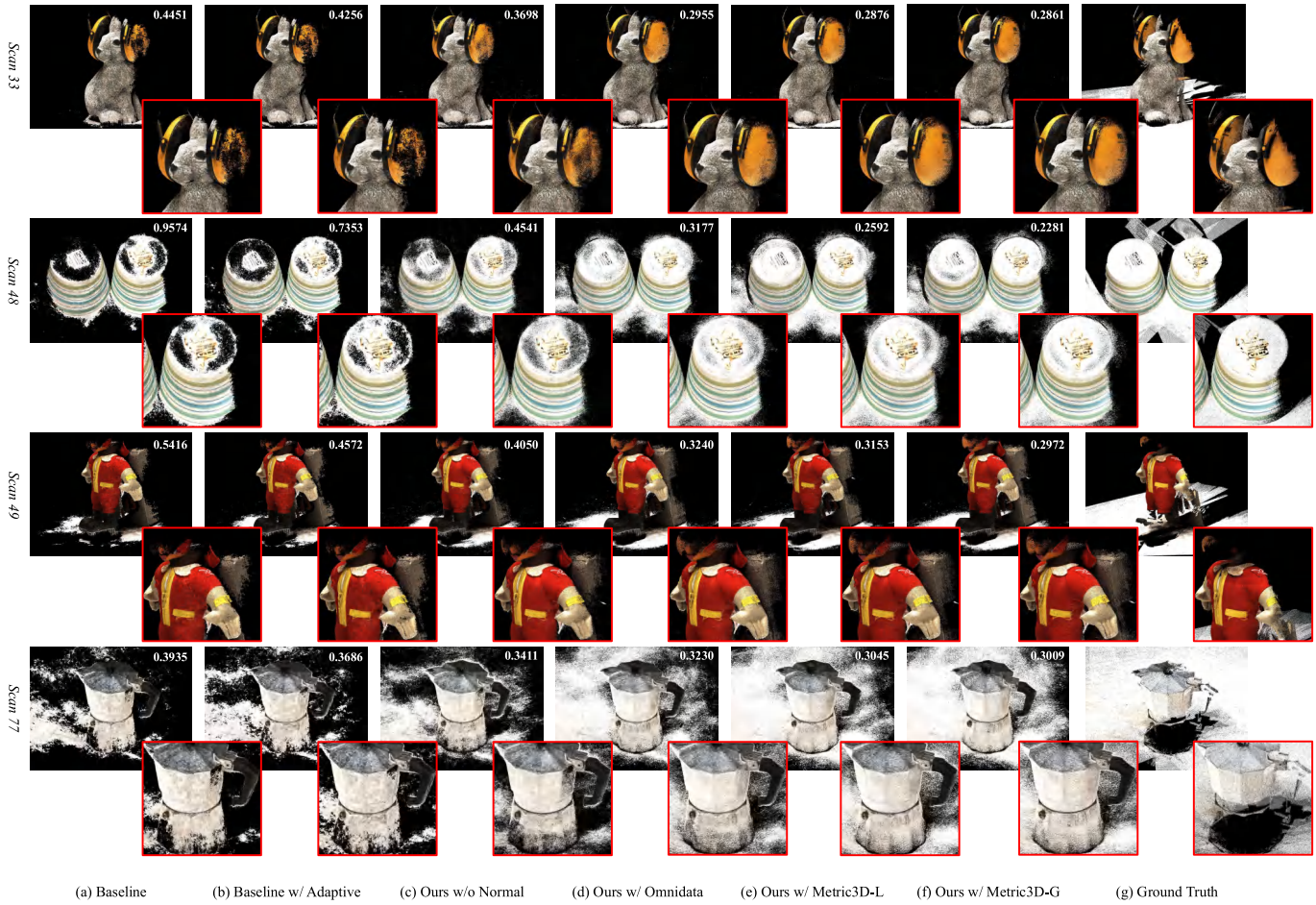


Fig. 7. Qualitative comparison of point cloud reconstructions on the DTU evaluation set. (a) Baseline method [15]. (b) Baseline method with cross-view adaptive cost volume aggregation. (c) Our method without normal-assisted depth hypotheses refinement. (d)-(f) Our method with normal-assisted depth hypotheses refinement using normal maps from Omnidata [57], Metric3D-Large [69], and Metric3D-Giant [69], respectively. (g) Ground-truth point cloud. The red box highlights the zoomed-in region, and the top-right number denotes the completeness error in *mm*, demonstrating the effectiveness of the proposed modules for point cloud reconstruction. Note that the ground-truth point clouds provided by the structured light scanner are not always complete.

aggregation [70] method computes pairwise matching costs based on the squared difference and learns a 3D mask for each source-view cost, where each voxel contains a weight for every pixel at each depth hypothesis, enabling the weighted averaging of multi-view pairwise costs. The pixel-wise inter-view adaptive aggregation [21] slices the reference-view and source-view feature volumes along the depth dimension, using the squared difference to compute pairwise matching costs, and learns pixel-wise weights for each source-view cost map to enable the weighted averaging of multi-view pairwise costs into the final cost volume. Additionally, the pixel-wise uncertainty-based aggregation [25] employs element-wise multiplication to measure pairwise matching costs and generates an uncertainty map for depth estimation, derived from the entropy of the probability distribution across all depth hypotheses, which is used as a weight map for each source-view pairwise cost. The weighted sum of multi-view pairwise matching costs is then computed to form the final cost volume.

In comparison, our method retains squared difference for pairwise matching costs, as they inherently reflect feature differences, and averages the pairwise costs along the color channel dimension to improve computational efficiency. For each source-view pairwise cost, a voxel-wise weight is first estimated using a lightweight re-weighting network, and the maximum weight along the depth dimension is then taken to obtain per-view, per-pixel visibility. The final cost volume is computed by aggregating the weighted sum of multi-view pairwise costs, where the visibility map is applied to mask each source-view pairwise cost while preserving the original source-view pairwise cost. Experimental results presented in Table VI, demonstrate that our method surpasses existing methods in reconstruction completeness and overall performance, while maintaining efficient memory usage and runtime consumption. Additionally, as illustrated in Fig. 8, the qualitative visualization of per-pixel adaptive visibility derived from adjacent source views emphasizes the maximum similarity between the reference view and the source views along the depth dimension, where the whiter regions indicate higher visibility weights assigned to the respective pixels, enabling more effective cost volume aggregation.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

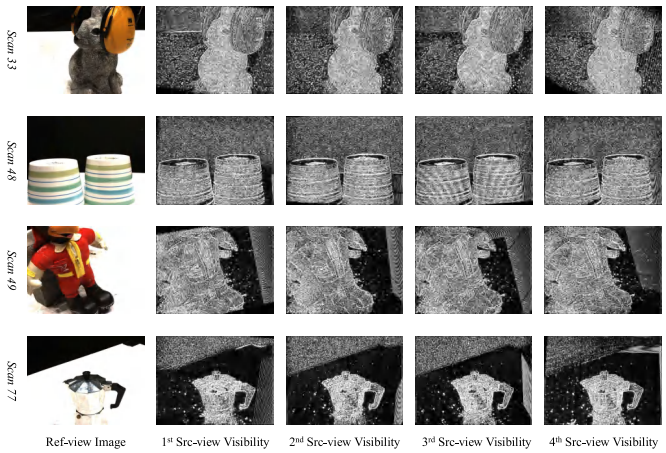


Fig. 8. Qualitative illustration of per-pixel adaptive visibility derived from adjacent source views, highlighting the maximum similarity between the reference view and source views along the depth dimension. The whiter regions indicate higher visibility weights.

TABLE VI

COMPARISON OF DIFFERENT COST VOLUME AGGREGATION MODULES FOR POINT CLOUD RECONSTRUCTION ON THE DTU EVALUATION SET

Methods	Mean Error Distance on 22 Scenes		
	Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)
Variance-Based Aggregation [15]	0.359 (+0.0009)	0.339 (+0.0008)	0.349 (+0.0009)
View-Wise Learnable Cost Metric [28]	**0.356** (+0.0000)	0.323 (+0.0001)	0.339 (+0.0005)
Voxel-Wise Adaptive Aggregation [70]	0.360 (+0.0006)	0.319 (+0.0002)	0.339 (+0.0009)
Pixel-Wise Inter-View Aggregation [21]	0.362 (+0.0004)	0.324 (+0.0006)	0.343 (+0.0005)
Pixel-Wise Uncertainty-Based Aggregation [25]	0.363 (+0.0002)	0.359 (+0.0007)	0.361 (+0.0005)
Ours	0.357 (+0.0001)	**0.314** (+0.0005)	**0.335** (+0.0008)

Methods	Computational Costs		
	Train Memory† (MB)	Test Memory† (MB)	Train / Test Runtime‡ (s)
Variance-Based Aggregation [13]	**13449**	4863	0.465 / 0.222
View-Wise Learnable Cost Metric [28]	13725	9243	0.535 / 0.231
Voxel-Wise View Aggregation [70]	16005	5887	0.604 / 0.308
Pixel-Wise Inter-View Aggregation [21]	20635	4749	0.769 / 0.925
Pixel-Wise Uncertainty-Based Aggregation [25]	14917	6987	**0.426** / 0.878
Ours	13507	**4239**	0.434 / **0.185**

↓ Smaller values indicate better performance.
† The batch size is set to 2 and 1 to measure the memory footprint in the train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.
‡ The batch size is set to 1 to measure the runtime in the train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.

3) Depth Consistency Optimization: The depth consistency optimization module is influenced by two critical hyperparameters in computing the reference-view depth inconsistency mask: the per-level point distance threshold ϵ_l, which defines the lower bound for depth inconsistency, and the number of source views N_s, across which the backward projection is performed and the reference-view depth inconsistency is accumulated. We conduct ablation experiments on ϵ_l and N_s to evaluate their impact on reconstruction performance.

Table VII presents the ablation results of the per-level point distance threshold for point cloud reconstruction on the DTU evaluation set. We first evaluate ϵ_l by setting it to 0.1, 0.2, and 0.3 *mm*, maintaining a uniform threshold across coarse to fine feature levels. The initial choice of around 0.2 is based on a conservative estimate of the ground-truth point cloud accuracy produced by the structured light scanner [18]. The results demonstrate that our method, without normal-assisted depth hypotheses refinement, achieves the best point cloud reconstruction accuracy and completeness when a threshold of 0.2 is applied uniformly across all levels. We further experiment by setting ϵ_l to 0.3, 0.2, and 0.1 for the coarse, middle, and fine feature levels, respectively. Specifically, we increase the coarse-level threshold from 0.2 to 0.3, as coarse-level depth estimation is less accurate, and reduce the fine-level threshold from 0.2 to 0.1, as fine-level depth estimation is more accurate. The overall reconstruction performance with varying thresholds outperforms configurations where 0.1 and 0.3 are applied uniformly across all levels but remains inferior to the setting with a uniform threshold of 0.2.

TABLE VII

ABLATION STUDY OF POINT DISTANCE THRESHOLD FOR POINT CLOUD RECONSTRUCTION ON THE DTU EVALUATION SET

Point Distance Threshold	Mean Error Distance on 22 Scenes		
$(\epsilon_0, \epsilon_1, \epsilon_2)^\dagger$ (mm)	Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)
(0.1, 0.1, 0.1)	0.351 (+0.0008)	0.274 (+0.0003)	0.313 (+0.0001)
(0.2, 0.2, 0.2)	**0.343** (+0.0008)	**0.272** (+0.0006)	**0.308** (+0.0002)
(0.3, 0.3, 0.3)	0.344 (+0.0003)	0.286 (+0.0003)	0.315 (+0.0003)
(0.3, 0.2, 0.1)	0.344 (+0.0009)	0.278 (+0.0003)	0.311 (+0.0006)

† The number of source views N_s for depth consistency optimization module is set to 8.
* The overall score is the summary measure of the overall reconstruction performance.
↓ Smaller values indicate better performance.

TABLE VIII

ABLATION STUDY OF NUMBER OF SOURCE VIEWS FOR POINT CLOUD RECONSTRUCTION ON THE DTU EVALUATION SET

Number of Source Views	Mean Error Distance on 22 Scenes		
$N_s^\dagger$	Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)
4	0.351 (+0.0002)	0.269 (+0.0006)	0.310 (+0.0004)
5	0.342 (+0.0001)	0.278 (+0.0008)	0.310 (+0.0005)
6	0.347 (+0.0008)	0.268 (+0.0009)	0.308 (+0.0004)
7	0.349 (+0.0002)	**0.268** (+0.0002)	0.308 (+0.0007)
8	0.343 (+0.0008)	0.272 (+0.0006)	**0.308** (+0.0002)
9	**0.327** (+0.0007)	0.298 (+0.0007)	0.313 (+0.0002)

† The point distance threshold ϵ_l for depth consistency optimization module is set to 0.2.
* The overall score is the summary measure of the overall reconstruction performance.
↓ Smaller values indicate better performance.

Table VIII summarizes the ablation results of the number of source views N_s for point cloud reconstruction on the DTU evaluation set. The number of source views N_s is incrementally increased from 4 to 9. We start with $N_s = 4$ because the number of input images N is fixed at 5 during training and benchmarking, following common practices in state-of-the-art methods. With $N_s = 4$, only the remaining 4 adjacent source views, relative to the reference view, are utilized. The upper limit of $N_s = 9$ is set, as each reference view is paired with a maximum of 9 adjacent source views after view selection. The results show that the overall reconstruction performance improves as N_s increases up to 8, but deteriorates at $N_s = 9$. This decline is likely due to occlusion and scene content variations in the excessive views with lower feature similarity, causing a significant drop in reconstruction completeness. Based on these findings, we set $\epsilon_l = 0.2$ and $N_s = 8$ for the proposed depth consistency optimization module.

As shown in Fig. 9(a)-(b), we visualize the pointwise distance error $\mathbf{E}^l_{0\leftrightarrow i}$ between the back-projected reference-view points $\mathbf{X}^l_0$ and the aligned i-th source-view points $\mathbf{X}^l_i$, with darker regions indicating smaller errors. The distribution evolution of this error is presented in Fig. 9(c), where the depth

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

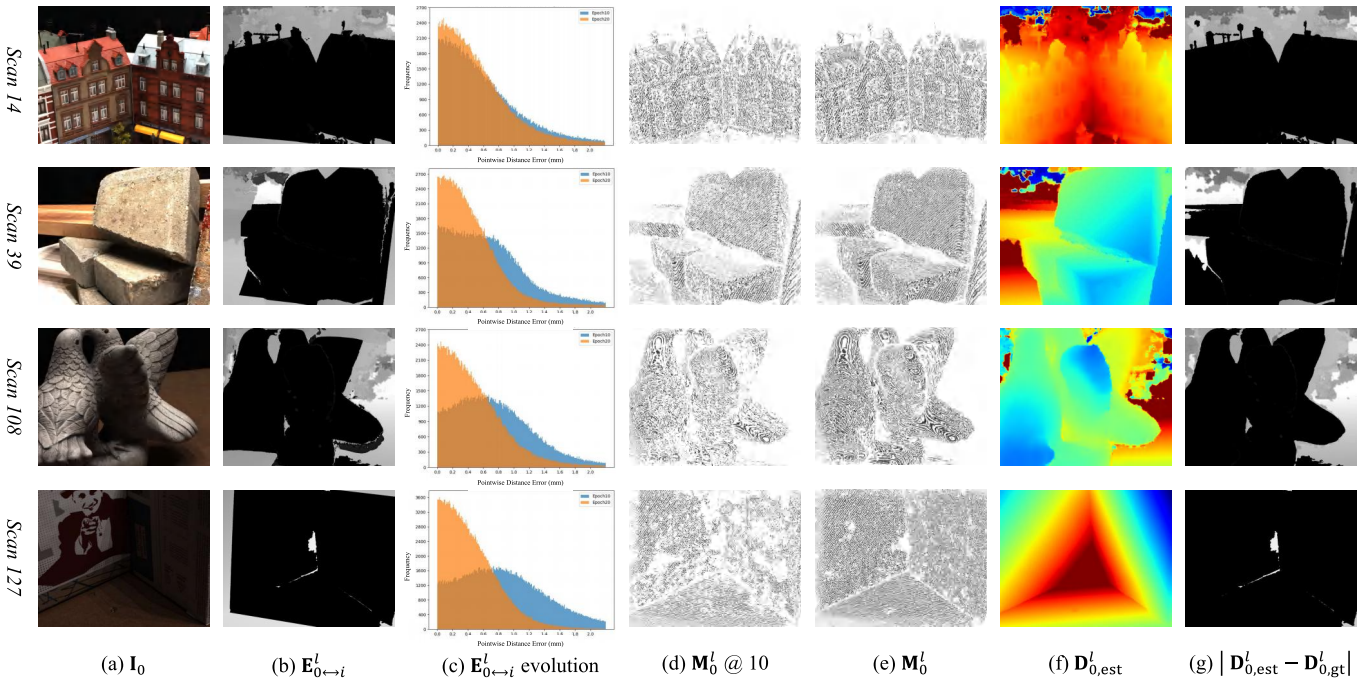


Fig. 9. Visualizations on the DTU training set to validate the effectiveness of the depth consistency optimization module. (a) Reference image $\mathbf{I}_0$. (b) Pointwise distance error $\mathbf{E}^l_{0\leftrightarrow i}$ between the back-projected reference-view points $\mathbf{X}^l_0$ and the aligned i-th source-view points $\mathbf{X}^l_i$, where darker regions indicate smaller errors. (c) Evolution of the distribution of $\mathbf{E}^l_{0\leftrightarrow i}$, showing that the depth consistency optimization module drives the point error distribution towards the threshold $\epsilon_l = 0.2$ during training. (d)-(e) Depth inconsistency masks $\mathbf{M}^l_0$ at epochs 10 and 20, where dark regions correspond to pixels with point errors within the threshold ϵ_l. (f) Estimated depth map $\mathbf{D}^l_{0,\text{est}}$. (g) Absolute depth error map between $\mathbf{D}^l_{0,\text{est}}$ and $\mathbf{D}^l_{0,\text{gt}}$, where darker colors indicate smaller errors. Note that the ground-truth depth map $\mathbf{D}^l_{0,\text{gt}}$ is incomplete in certain regions, and figures (b), (e), (f), and (g) are obtained at the convergence epoch 20.

consistency optimization module drives the error distribution toward the threshold $\epsilon_l = 0.2$ during training. Moreover, Fig. 9(d)-(e) shows depth inconsistency masks $\mathbf{M}^l_0$ at intermediate and converged epochs, where dark regions represent pixels with errors within the threshold, demonstrating the effectiveness of the depth consistency optimization module in suppressing inconsistent depth values. Fig. 9(f)-(g) shows that our method, even without encoding normal cues, achieves accurate depth estimations across various scenes with varying surface textures, geometric shapes, and lighting conditions.

4) Normal-Assisted Depth Hypotheses Refinement: We conduct ablation experiments to systematically assess the impact of normal accuracy on depth estimation and point cloud reconstruction performance. We employ three pre-trained monocular normal estimation networks: Omnidata [57], Metric3D-Large [69], and Metric3D-Giant [69] to generate normal maps for the normal-assisted depth hypothesis refinement module. These networks take the reference-view image as input and produce progressively more accurate normal maps, as indicated by the benchmark results in Tables IX and X. The generated normal maps are interpolated to the original resolution of the input image and normalized to the range $[-1, 1]$. For visualization, we use a color-encoded representation of the normal maps, where the RGB channels correspond to the X, Y, and Z components of the normal vector, providing an intuitive interpretation of surface orientations. A qualitative comparison of the normal maps is presented in Fig. 10(b)-(d). The Metric3D variants generate more accurate normal vectors with fewer angular errors and effectively handle low-light backgrounds, whereas Omnidata overemphasizes surface details, leading to higher angular errors.

TABLE IX
ABLATION STUDY OF NORMAL ESTIMATION ACCURACY FOR DEPTH ESTIMATION ON THE DTU VALIDATION SET

Methods	Mean Depth Error on 18 Validation Scenes with 7 Lighting Conditions			
	MAE ↓ (mm)	MAE_2 ↓ (%)	MAE_4 ↓ (%)	MAE_8 ↓ (%)
Ours w/o Normal	**5.312**	14.050	9.479	6.711
Ours w/ Ominidata [57]	5.839	13.713	9.517	7.078
Ours w/ Metric3D-Large [69]	5.533	12.540	8.692	6.570
Ours w/ Metric3D-Giant [69]	5.392	**11.847**	**8.232**	**6.242**
Methods	**Mean Normal Accuracy on ibims-1 Normal Benchmark**			
	11.25° ↑ (0-1)	22.5° ↑ (0-1)	30° ↑ (0-1)	Mean ↓ (°)
Ominidata [57]	0.647	0.734	0.768	20.8
Metric3D-Large [69]	0.694	0.758	0.785	**19.4**
Metric3D-Giant [69]	**0.697**	**0.762**	**0.788**	19.6

↓ Smaller values indicate better performance.
↑ Larger values indicate better performance.

We evaluate the impact of normal accuracy on depth estimation performance using the DTU validation set, which provides ground-truth depth maps for quantifying depth errors. The results, summarized in Table IX, demonstrate that our method with normal-assisted depth hypotheses refinement, when compared to ours without normal-assisted refinement, leads to a degradation in depth estimation quality, as evidenced by an increase in the mean absolute error (MAE) between predicted and ground-truth depth maps. This degradation is attributed

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

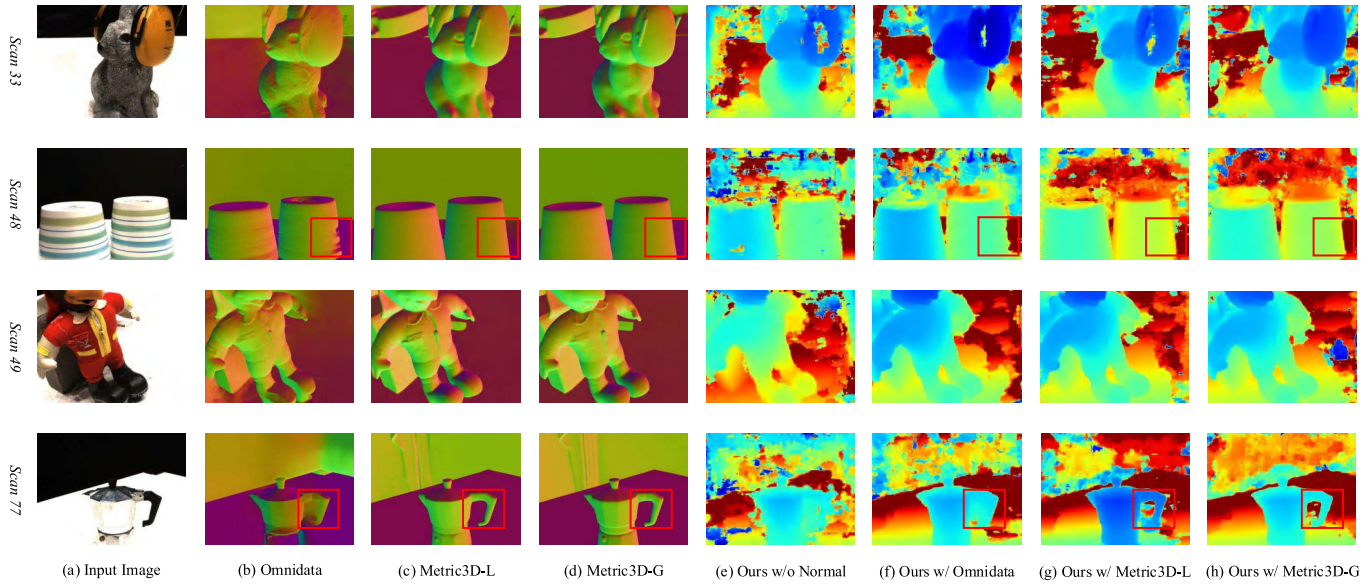


Fig. 10. Qualitative comparison of surface normal and depth maps on the DTU evaluation set. (a) Input image. (b)-(d) Surface normal maps estimated by Omnidata [57] and the Large and Giant variants of Metric3D [69]. The Metric3D variants exhibit more accurate normal vectors with reduced angular errors. (e) Depth maps predicted by our method without normal-assisted depth hypotheses refinement. (f)-(h) Depth maps with normal-assisted depth hypotheses refinement, using normal maps from Omnidata, Metric3D-Large, and Metric3D-Giant, respectively.

TABLE X
ABLATION STUDY OF NORMAL ESTIMATION ACCURACY FOR POINT CLOUD RECONSTRUCTION ON THE DTU EVALUATION SET

Methods	Mean Error Distance on 22 Scenes		
	Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)
Ours	**0.343** (+0.0008)	0.272 (+0.0006)	0.308 (+0.0002)
Ours w/. Ominidata [57]	0.354 (+0.0007)	0.246 (+0.0009)	0.300 (+0.0008)
Ours w/. Metric3D-Large [69]	0.350 (+0.0002)	0.237 (+0.0001)	0.293 (+0.0007)
Ours w/. Metric3D-Giant [69]	0.349 (+0.0004)	**0.233** (+0.0008)	**0.291** (+0.0006)

Methods	Mean Normal Accuracy on ScanNet Normal Benchmark			
	11.25° ↑ (0-1)	22.5° ↑ (0-1)	30° ↑ (0-1)	Mean ↓ (°)
Ominidata [57]	0.629	0.806	0.847	15.1
Metric3D-Large [69]	0.760	0.885	0.923	9.9
Metric3D-Giant [69]	**0.778**	**0.901**	**0.935**	**9.2**

* The overall score is the summary measure of the overall reconstruction performance.
↓ Smaller values indicate better performance.
↑ Larger values indicate better performance.

to angular errors introduced by the normal maps predicted by the monocular normal estimation network. Notably, higher normal accuracy correlates with lower MAE, highlighting the sensitivity of depth estimation quality to normal accuracy. Despite the MAE degradation, normal-assisted refinement enhances depth consistency in local regions, reflected in the reduction of MAE_2, quantifying the percentage of absolute errors greater than 2 mm. As normal accuracy improves, depth estimation errors, including MAE, MAE_2, MAE_4, and MAE_8, are consistently reduced. Among these metrics, MAE and MAE_2 are particularly indicative of the relationship between depth estimation accuracy and point cloud reconstruction quality. Specifically, lower MAE leads to improved reconstruction accuracy, while lower MAE_2 correlates with higher reconstruction completeness, as shown in the comparison between Table IX and Table X. Ultimately, our method achieves the best overall reconstruction performance by leveraging normal maps with the lowest angular errors.

The depth map comparison in Fig. 10(e)-(h) qualitatively demonstrates the effectiveness of the proposed normal-assisted depth hypotheses refinement module in improving depth estimation quality and verifies that higher normal accuracy results in more accurate depth estimation. Furthermore, the point cloud comparison presented in Fig. 7(c)-(f) qualitatively validates the proposed module's ability to enhance point cloud completeness, with the results showing that increased normal accuracy leads to a more complete point cloud reconstruction.

5) Number of Input Views and Image Resolution: We conduct ablation experiments to evaluate the impact of the number of input views, N, and image resolution, $H \times W$, on the quality of point cloud reconstruction, using our method with normal maps generated by Metric3D-Giant. As shown in Table XI, increasing the number of input views from 3 to 10 consistently improves reconstruction accuracy, with the highest accuracy achieved at $N = 10$. However, reconstruction completeness exhibits a distinct trend: it initially improves, reaching a peak at $N = 7$, before deteriorating as the number of views increases further. This decline in completeness is primarily due to redundant viewpoints, which cause occlusions and excessive overlap, leading to ambiguities and inconsistencies in the reconstructed geometry, ultimately reducing completeness. Our method achieves the best overall performance by maintaining an optimal balance between reconstruction accuracy and completeness. For quantitative benchmarking on the DTU evaluation set, we set the number of input views to $N = 5$, following standard practices for fair comparison. Furthermore, when varying the image resolution from low to high, our method attains the highest reconstruction accuracy, completeness, and overall score at $H \times W = 864 \times 1152$, which is the image resolution commonly adopted by state-of-the-art methods in DTU benchmarking.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE XI

ABLATION STUDY OF NUMBER OF INPUT VIEWS N, PROBABILITY THRESHOLD τ, NUMBER OF CONSISTENT VIEWS N_c, AND IMAGE RESOLUTION $H \times W$ FOR POINT CLOUD RECONSTRUCTION ON THE DTU TEST SET

N	τ	N_c	$H \times W$	Mean Error Distance on 22 Scenes		
				Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)
3	0.0	2	864 × 1152	0.366 (+0.0002)	0.252 (+0.0009)	0.309 (+0.0006)
4				0.355 (+0.0003)	0.238 (+0.0007)	0.297 (+0.0000)
5				0.349 (+0.0004)	0.233 (+0.0008)	0.291 (+0.0006)
6				0.350 (+0.0000)	0.231 (+0.0002)	0.290 (+0.0006)
7				0.347 (+0.0009)	**0.229** (+0.0001)	0.288 (+0.0005)
8				0.345 (+0.0005)	0.235 (+0.0002)	0.290 (+0.0004)
9				0.343 (+0.0008)	0.232 (+0.0000)	**0.287** (+0.0009)
10				**0.342** (+0.0008)	0.233 (+0.0007)	0.288 (+0.0003)
5	0.0	2	576 × 768	0.349 (+0.0005)	0.257 (+0.0000)	0.303 (+0.0003)
			864 × 1152	**0.349** (+0.0004)	**0.233** (+0.0008)	**0.291** (+0.0006)
			1152 × 1536	0.350 (+0.0008)	0.234 (+0.0003)	0.292 (+0.0006)
5	0.0	2	864 × 1152	0.349 (+0.0004)	**0.233** (+0.0008)	**0.291** (+0.0006)
	0.1			0.339 (+0.0005)	0.244 (+0.0003)	0.291 (+0.0009)
	0.2			0.326 (+0.0001)	0.270 (+0.0003)	0.298 (+0.0002)
	0.3			0.310 (+0.0004)	0.307 (+0.0008)	0.309 (+0.0001)
	0.4			0.292 (+0.0000)	0.381 (+0.0006)	0.336 (+0.0008)
	0.5			0.272 (+0.0002)	0.500 (+0.0003)	0.386 (+0.0003)
	0.6			**0.250** (+0.0009)	0.679 (+0.0009)	0.465 (+0.0004)
5	0.0	2	864 × 1152	0.349 (+0.0004)	**0.233** (+0.0008)	**0.291** (+0.0006)
		3		0.300 (+0.0004)	0.302 (+0.0005)	0.301 (+0.0005)
		4		0.268 (+0.0004)	0.404 (+0.0000)	0.336 (+0.0002)
		5		0.244 (+0.0007)	0.561 (+0.0002)	0.403 (+0.0000)
		6		**0.226** (+0.0005)	0.784 (+0.0001)	0.505 (+0.0003)

* The overall score is the summary measure of the overall reconstruction performance.
↓ Smaller values indicate better performance.

6) Point Cloud Fusion: Recall that we adopt the fusible method, in line with most state-of-the-art methods, to fuse multi-view depth maps into dense point cloud reconstruction, where the probability threshold τ and the number of consistent views N_c are used to eliminate depth outliers and reduce multi-view geometric inconsistencies, respectively. As τ increases incrementally from 0.0 to 0.6, our method achieves higher reconstruction accuracy but lower reconstruction completeness. Similarly, as N_c increases from 2 to 6, reconstruction accuracy improves while completeness decreases. This behavior is expected, as higher values of τ and N_c result in the exclusion of more potentially valid depth estimations. Our method achieves the best overall reconstruction quality when $\tau = 0.0$ and $N_c = 2$, indicating that our method is less reliant on depth map filtering after multi-view depth estimation, as our depth consistency optimization module checks for depth inconsistencies during learning by explicitly encoding adjacent source-view ground-truth depth cues and geometrically constraining the depth optimization process directly in the 3D point space.

F. Memory and Runtime Consumption

As shown in Table XII, we conduct a systematic comparison of the proposed GE-MVS method with several recent state-of-the-art learning-based MVS methods, including AA-RMVSNet [21], TransMVSNet [22], UniMVSNet [23], Vis-MVSNet [25], CDS-MVSNet [29], and GeoMVSNet [30]. The comparison encompasses key performance metrics such as model parameters, memory footprint, and inference runtime during both the training and testing phases. The results demonstrate that our method exhibits a competitive memory footprint and achieves the highest test runtime relative to other methods, while preserving superior depth estimation and reconstruction accuracy and completeness, as validated in Tables I to IV.

TABLE XII

COMPARISON OF MEMORY AND RUNTIME CONSUMPTION

Methods	#Params	Computational Costs		
		Train Memory† (MB)	Test Memory† (MB)	Train / Test Runtime‡ (s)
AA-RMVSNet [21]	187203	20349	8387	1.463 / 23.526
TransMVSNet [22]	1148924	20492	**5012**	0.769 / 0.558
UniMVSNet [23]	934375	16810	9428	0.424 / 0.324
Vis-MVSNet [25]	1162696	14917	6987	0.426 / 0.878
CDS-MVSNet [29]	981622	**8800**	9024	**0.297** / 0.284
GeoMVSNet [30]	15306124	17512	9550	0.666 / 0.303
Ours	1164107	12709	10883	0.471 / **0.241**

↓ Smaller values indicate better performance.
† The batch size is set to 2 and 1 to measure the memory footprint in the train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.
‡ The batch size is set to 1 to measure the runtime in the train ($H \times W = 512 \times 640$) and test ($H \times W = 864 \times 1152$) mode.

TABLE XIII

TRADE-OFF EXPERIMENTS BETWEEN PERFORMANCE AND EFFICIENCY ON THE DTU EVALUATION SET

$H \times W$	Mean Error Distance on 22 Scenes			Computational Costs	
	Acc. ↓ (mm)	Comp. ↓ (mm)	Overall* ↓ (mm)	Memory† (MB)	Runtime† (s)
576 × 768	0.349 (+0.0005)	0.257 (+0.0000)	0.303 (+0.0003)	6159	0.107
864 × 1152	**0.349** (+0.0004)	**0.233** (+0.0008)	**0.291** (+0.0006)	10883	0.241
1152 × 1536	0.350 (+0.0008)	0.234 (+0.0003)	0.292 (+0.0006)	15201	0.423

* The overall score is the summary measure of the overall reconstruction performance.
↓ Smaller values indicate better performance.
† The batch size is set to 1 to measure the memory footprint and runtime in the test mode.

As presented in Table XIII, we conduct trade-off experiments to analyze the balance between point cloud reconstruction quality and computational efficiency, focusing on the impact of varying input image resolution. We evaluate three resolutions: low (576 × 768), medium (864 × 1152), and high (1152 × 1536). The results demonstrate that our method consistently maintains robust reconstruction accuracy and completeness across these variations. Notably, optimal reconstruction quality is achieved at the medium resolution, striking a balance between reconstruction fidelity and computational efficiency, with the highest performance while maintaining a reasonable memory footprint and competitive runtime consumption.

V. REAL-WORLD EXPERIMENTS

We compare the proposed GE-MVS method with prevalent industrial reconstruction solutions [12], [66], [71] and recent learning-based MVS methods [22], [30] using real-world underwater, indoor, and outdoor scenes, which vary in scattering media (air vs. water), illumination conditions, texture richness, and geometric complexity. Specifically, 42 images ($H \times W = 720 \times 1280$) of an underwater scene were captured with a GoPro Hero12 Black at a depth of 10 meters in Puerto Galera Island, Philippines. Additionally, 301 images ($H \times W = 1080 \times 1920$) of an indoor scene were collected from the advanced set of the Tanks and Temples dataset [20], and 189 images ($H \times W = 1200 \times 1600$) of a low-textured outdoor scene were acquired using a self-developed UAV in Shangshui, Hong Kong SAR, China. As shown in Fig. 11, our

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

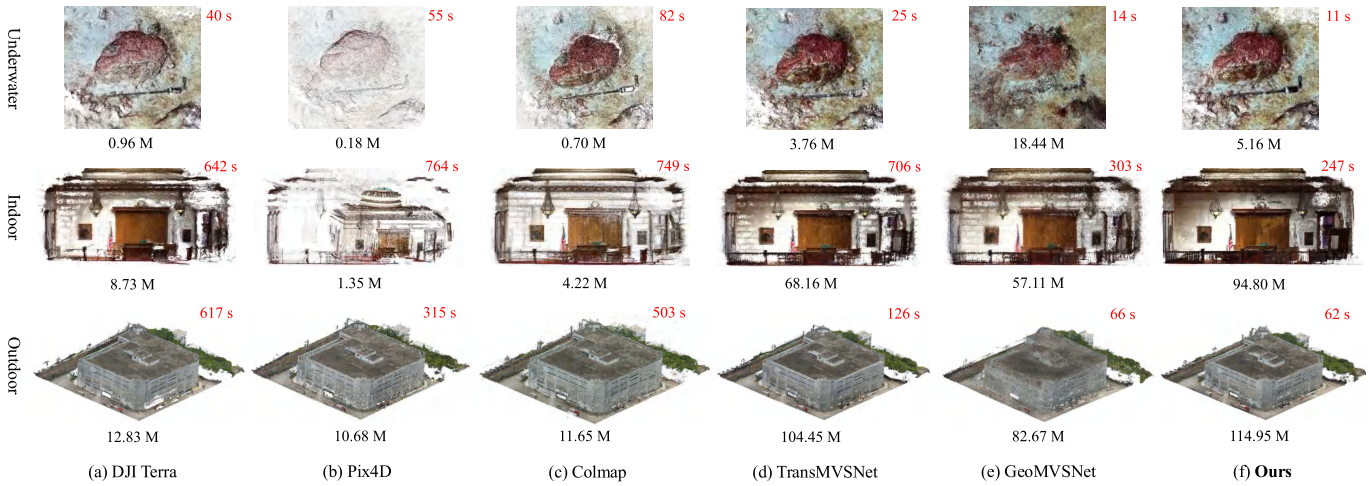


Fig. 11. Comparison of point cloud reconstruction results for real-world underwater, indoor, and outdoor scenes with varying scattering media, illumination conditions, and surface textures. In comparison to industrial reconstruction solutions such as DJI Terra [71], Pix4D [66], and COLMAP [12], as well as recent learning-based methods like TransMVSNet [22] and GeoMVSNet [30], our method produces more complete and denser reconstructions efficiently. The top-right and bottom numbers show the MVS processing time (in seconds) and the number of points in the entire scene (in millions), respectively.

method outperforms industrial solutions such as DJI Terra [71] (V4.4.6), Pix4D [66] (V4.5.6), and COLMAP [12] (V3.12.0), as well as recent learning-based methods like TransMVSNet [22] and GeoMVSNet [30], producing more complete and denser reconstructions with superior efficiency. Notably, we exclude the structure-from-motion processing time from all methods, focusing solely on MVS processing time. All experiments were conducted on a system equipped with an NVIDIA RTX 4090 GPU and an AMD Ryzen 9 7950X CPU. The superiority of GE-MVS over industrial solutions and recent learning-based methods is quantitatively verified in Table II and Table III.

VI. Limitations

While our method achieves competitive performance across a variety of benchmarks, it still presents certain limitations. In particular, similar to other learning-based MVS approaches, its effectiveness is sensitive to several hyperparameters, including the number of input views, the input image resolution, the probability threshold for depth selection, and the minimum number of consistent views required for reliable fusion. Improper configuration of these parameters can lead to suboptimal reconstruction, particularly in challenging scenarios with severe occlusions or complex surface geometries.

In addition, the refinement stage relies on accurate surface normals that satisfy the geometric constraint in Eq. 14. Inaccurate normal vectors can generate erroneous depth candidates, ultimately compromising depth estimation and the quality of the reconstructed point cloud. As illustrated in Fig. 10, Scans 48 and 77 exhibit noticeable depth errors near object boundaries, caused by inaccurate normal predictions (highlighted in red boxes) from the monocular estimation network Omnidata [57]. This network tends to overemphasize fine surface textures, resulting in increased normal angular errors. A potential mitigation strategy involves computing ground-truth normal maps from ground-truth depth maps using least squares plane fitting, followed by bilateral filtering to suppress noise while preserving edges. However, this approach is constrained by the incompleteness of ground-truth depth maps, which can lead to missing or sparse normal supervision.

To address these limitations, future work will focus on enhancing the robustness of our approach through multi-view depth scale consistency optimization [72] and the integration of geometry-consistent priors via multi-view calibration [73]. Moreover, we plan to explore improved normal supervision strategies, including multi-view normal fusion and weakly supervised constraints derived from photometric or geometric consistency, as well as joint optimization of depth and normal estimation within an end-to-end framework. These directions hold promise for mitigating the impact of noisy or inconsistent surface normals, especially in complex real-world scenarios.

VII. Conclusion

In this paper, we have presented the GE-MVS to strengthen the geometric cues encoding during network learning for more accurate and complete depth estimation and point cloud reconstruction. Extensive experiments on three standard MVS benchmarks, including DTU, Tanks and Temples, and BlendedMVS demonstrate the state-of-the-art depth estimation and reconstruction performance of GE-MVS. Systematic ablation experiments validate the effectiveness and efficiency of each component of the proposed method. Real-world experiments for UAV-based large-scale reconstruction witness the generalization ability and superiority of GE-MVS over prevalent industrial reconstruction solutions. Our method enables accurate, complete, and scalable dense point cloud reconstruction for scenes ranging from small-scale to large-scale.

References

[1] M. Van der Merwe, Q. Lu, B. Sundaralingam, M. Matak, and T. Hermans, "Learning continuous 3D reconstructions for geometrically aware grasping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 11516–11522.

[2] J. Lim, N. Lawrance, F. Achermann, T. Stastny, R. Bähnemann, and R. Siegwart, "Fisher information based active planning for aerial photogrammetry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 1249–1255.

[3] M. Lv, D. Tu, X. Tang, Y. Liu, and S. Shen, "Semantically guided multi-view stereo for dense 3D road mapping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 11189–11195.

[4] G. Yang et al., "End-to-end underwater multi-view stereo for dense scene reconstruction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2025, pp. 7616–7623.

[5] G. Yang et al., "Det-Recon-reg: An intelligent framework towards automated large-scale infrastructure inspection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 12742–12749.

[6] Q. Li et al., "Single drone-based 3D reconstruction approach to improve public engagement in conservation of heritage buildings: A case of hakka tulou," *J. Building Eng.*, vol. 87, Jun. 2024, Art. no. 108954.

[7] L. Long, Z. Gan, G. Yang, and Q. Li, "A 3D reconstruction pipeline for generating textured models of large-scale architectural heritage," *J. Building Eng.*, vol. 113, Nov. 2025, Art. no. 114064.

[8] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 766–779.

[9] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[10] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 903–920, Sep. 2012.

[11] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.

[12] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 501–518.

[13] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 785–801.

[14] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5520–5529.

[15] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2492–2501.

[16] X. Guan, W. Tong, S. Jiang, P. Z. H. Sun, E. Q. Wu, and G. Chen, "Multistage pixel-visibility learning with cost regularization for multi-view stereo," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 2, pp. 751–762, Apr. 2023.

[17] W. Tong et al., "Edge-assisted epipolar transformer for industrial scene reconstruction," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 701–711, 2025.

[18] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, Nov. 2016.

[19] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1787–1796.

[20] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, Aug. 2017.

[21] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6167–6176.

[22] Y. Ding et al., "TransMVSNet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8585–8594.

[23] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8635–8644.

[24] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18042–18051.

[25] J. Zhang, S. Li, Z. Luo, T. Fang, and Y. Yao, "Vis-MVSNet: Visibility-aware multi-view stereo network," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 199–214, Jan. 2023.

[26] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "Bidirectional hybrid LSTM based recurrent neural network for multi-view stereo," *IEEE Trans. Vis. Comput. Graphics*, vol. 30, no. 7, pp. 3062–3073, Jul. 2024.

[27] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1587–1596.

[28] G. Yang, X. Zhou, C. Gao, X. Chen, and B. M. Chen, "Learnable cost metric-based multi-view stereo for point cloud reconstruction," *IEEE Trans. Ind. Electron.*, vol. 71, no. 9, pp. 11519–11528, Sep. 2024.

[29] K. T. Giang, S. Song, and S. Jo, "Curvature-guided dynamic scale networks for multi-view stereo," 2021, *arXiv:2112.05999*.

[30] Z. Zhang, R. Peng, Y. Hu, and R. Wang, "GeoMVSNet: Learning multi-view stereo with geometry perception," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21508–21518.

[31] G. Yang et al., "Multi-view stereo with learnable cost metric," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023, pp. 3017–3024.

[32] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2186–2196.

[33] W. Tong et al., "Normal assisted pixel-visibility learning with cost aggregation for multiview stereo," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24686–24697, Dec. 2022.

[34] J. Wu et al., "GoMVS: Geometrically consistent cost aggregation for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 20207–20216.

[35] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "GeoNet: Geometric neural network for joint depth and surface normal estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 283–291.

[36] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 418–433, Mar. 2005.

[37] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *Int. J. Comput. Vis.*, vol. 38, no. 3, pp. 199–218, Jul. 2000.

[38] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 1067–1073.

[39] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5478–5487.

[40] R. A. Newcombe et al., "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.

[41] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2326–2334.

[42] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 365–376.

[43] J. Yan et al., "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 674–689.

[44] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10451–10460.

[45] X. Ye, W. Zhao, T. Liu, Z. Huang, Z. Cao, and X. Li, "Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17615–17624.

[46] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21919–21928.

[47] W. Tong, Y. Cai, Y.-W. Jie, Y. Duan, Y. Hou, and E. Q. Wu, "Neural rendering and flow-assisted unsupervised multi-view stereo for real-time monocular tracking and scene perception," *IEEE Trans. Autom. Sci. Eng.*, early access, Feb. 28, 2025, doi: 10.1109/TASE.2025.3546713.

[48] C. Feng, H. Li, F. Gao, B. Zhou, and S. Shen, "PredRecon: A prediction-boosted planning framework for fast and high-quality autonomous aerial reconstruction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 1207–1213.

[49] X. Liu, Y. Wang, K. Jiang, Z. Zhou, K. Nam, and C. Yin, "Interactive trajectory prediction using a driving risk map-integrated deep learning method for surrounding vehicles on highways," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19076–19087, Oct. 2022.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

[50] R. Li, Y. Wang, S. Sun, Y. Zhang, F. Ding, and H. Gao, "UE-extractor: A grid-to-point ground extraction framework for unstructured environments using adaptive grid projection," *IEEE Robot. Autom. Lett.*, vol. 10, no. 6, pp. 5991–5998, Jun. 2025.

[51] G. Yang et al., "Det-Recon-reg: An intelligent framework toward automated UAV-based large-scale infrastructure inspection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025.

[52] G. Yang et al., "Datasets and processing methods for boosting visual inspection of civil infrastructure: A comprehensive review and algorithm comparison for crack classification, segmentation, and detection," *Construct. Building Mater.*, vol. 356, Nov. 2022, Art. no. 129226.

[53] Q. Li et al., "Autonomous design framework for deploying building integrated photovoltaics," *Appl. Energy*, vol. 377, Jan. 2025, Art. no. 124760.

[54] Q. Li, "Life cycle cost analysis of circular photovoltaic façade in dense urban environment using 3D modeling," *Renew. Energy*, vol. 238, Jan. 2025, Art. no. 121914.

[55] S. Galliani, K. Lasinger, and K. Schindler. (2015). *Fusible*. [Online]. Available: https://github.com/kysucix/fusibile

[56] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[57] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Mali, Oct. 2021, pp. 10766–10776.

[58] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1538–1547.

[59] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4877–4886.

[60] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2521–2531.

[61] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5732–5740.

[62] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14194–14203.

[63] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "IterMVS: Iterative probability estimation for efficient multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8606–8615.

[64] S. Fuhrmann, F. Langguth, N. Moehrle, M. Waechter, and M. Goesele, "MVE—An image-based reconstruction environment," *Comput. Graph.*, vol. 53, pp. 44–53, Dec. 2015.

[65] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open multiple view geometry," in *Proc. Int. Workshop Reproducible Res. Pattern Recognit.*, 2016, pp. 60–74.

[66] EPFL.(2024). *Pix4dmapper: The Leading Photogrammetry Software for Professional Drone Mapping*. [Online]. Available: https://www.pix4d.com/

[67] C. Wu. (2011). *VisualSFM: A Visual Structure From Motion System*. [Online]. Available: http://www.cs.washington.edu/homes/ccwu/vsfm

[68] D. Cernea. (2020). *OpenMVS: Multi-view Stereo Reconstruction Library*. [Online]. Available: https://cdcseacave.github.io/openMVS

[69] M. Hu et al., "Metric3Dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," 2024, *arXiv:2404.15506*.

[70] H. Yi et al., "Pyramid multi-view stereo net with self-adaptive view aggregation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 766–782.

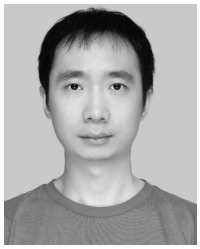
[71] DJI.(2024). *Dji Terra: Make the World Your Digital Asset*. [Online]. Available: https://enterprise.dji.com/zh-tw/dji-terra

[72] Z. Min, Y. Luo, J. Sun, and Y. Yang, "Epipolar-free 3D Gaussian splatting for generalizable novel view synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 39573–39596.

[73] Z. Min, Y. Luo, W. Yang, Y. Wang, and Y. Yang, "Entangled view-epipolar information aggregation for generalizable neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 4906–4916.

Guidong Yang (Member, IEEE) received the B.Eng. degree in mechanical and automation engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2018, the M.Eng. degree in vehicle engineering from SJTU, and the M.Sc. degree in mechanical engineering from the Politecnico di Milano, Milan, Italy, in 2021. He is currently pursuing the Ph.D. degree in mechanical and automation engineering with The Chinese University of Hong Kong, Hong Kong, China. His current research interests include multi-view stereo and object detection.

Rui Cao (Graduate Student Member, IEEE) received the B.Eng. degree in electrical engineering and its automation from Beijing Jiaotong University, Beijing, China, in 2017. He is currently pursuing the Ph.D. degree in mechanical and automation engineering with The Chinese University of Hong Kong, Hong Kong, China. His research interests include object perception, grasp synthesis, and object manipulation.

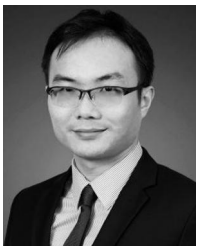


Junjie Wen (Member, IEEE) received the B.Sc. degree in automotive engineering from Dalian University of Technology, Dalian, China, in 2013, and the M.Sc. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree in mechanical and automation engineering with The Chinese University of Hong Kong, Hong Kong, China. His research interests include image restoration and novel view synthesis.

Benyun Zhao (Graduate Student Member, IEEE) received the M.Sc. degree in mechanical and automation engineering from The Chinese University of Hong Kong (CUHK), Hong Kong, China, in 2021, where he is currently pursuing the Ph.D. degree in mechanical and automation engineering. He was a Research Assistant with CUHK and Hong Kong Center of Logistics Robotics (HKCLR) from 2021 to 2022. His current research interests include object detection, semantic segmentation, and 3D scene understanding.

Qingxiang Li received the B.Sc. degree in civil engineering and the M.Sc. degree in architecture from Tianjin University (TJU), Tianjin, China, in 2016 and 2019, respectively, and the Ph.D. degree in architectural engineering from the Politecnico di Milano, Milan, Italy, in 2023. He is currently a Post-Doctoral Fellow of mechanical and automation engineering with The Chinese University of Hong Kong, Hong Kong, China. His research interests include multi-view stereo and building automation.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.



Xi Chen is currently a Research Assistant Professor of mechanical and automation engineering with The Chinese University of Hong Kong (CUHK), Hong Kong. He has over ten-year experience in sustainable building technology related to the urban energy systems, renewable application in buildings, and built environment modeling. He has led or managed multiple research projects, including ARC, MOST, RGC and consultancy projects with the local government and industry. He has published over 40 articles in peer-reviewed international journals and co-authored a book in green building and renewable application areas.

Dr. Chen has been awarded as the DECRA Fellow with Australian Research Council and the Fulbright Scholar with the Lawrence Berkeley National Laboratory. In addition, he services as an Editorial Board Member for *Buildings*, *Energies*, and *Advances in Applied Energy*.

Yun-Hui Liu (Fellow, IEEE) is currently a Choh-Ming Li Professor of mechanical and automation engineering with The Chinese University of Hong Kong (CUHK) and the Director of the CUHK T Stone Robotics Institute. He is also an Adjunct Professor with the State Key Laboratory of Robotics Technology and System, Harbin Institute of Technology, Harbin, China. He has authored or co-authored more than 300 articles in refereed journals and refereed conference proceedings. He was listed in the Highly Cited Authors (Engineering) by Thomson Reuters in 2013. His research interests include visual servoing, medical robotics, multifingered grasping, mobile robots, and machine intelligence.

Dr. Liu was a recipient of several research awards from international journals and international conferences in robotics and automation and government agencies. He was the General Chair for 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. He was the Editor-in-Chief of *Robotics* and *Biomimetics*. He served as an Associate Editor for IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION.

Ben M. Chen (Fellow, IEEE) is currently a Professor of mechanical and automation engineering with The Chinese University of Hong Kong (CUHK), Hong Kong. He was a Provost's Chair Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), before joining CUHK, in 2018. He was an Assistant Professor with the Department of Electrical Engineering, State University of New York at Stony Brook, NY, USA, from 1992 to 1993. He has authored/co-authored over 100 journals and conference papers, and a dozen research monographs in control theory and applications, unmanned systems, and financial market modeling. His current research interests are in unmanned systems and their applications.

Dr. Chen is a fellow of the Academy of Engineering, Singapore. He has served on the editorial boards of a dozen international journals, including *Automatica* and IEEE TRANSACTIONS ON AUTOMATIC CONTROL. He is currently serving as the Editor-in-Chief for *Unmanned Systems* and an Editor for *International Journal of Robust and Nonlinear Control*.