Appendix

TABLE I
COMPARISON BETWEEN THE EVALUATION METRICS ON DIFFERENT BENCHMARKS

Benchmark	Evaluation Metrics	Indication
DTU	Accuracy ↓, Completeness ↓, Overall Score ↓	Reconstruction Performance
Tanks and Temples	Precision ↑, Recall ↑, F-score ↑	Reconstruction Performance
BlendedMVS	EPE ↓, e_1 ↓, e_3 ↓	Depth Estimation Performance

A. Datasets

DTU [4] is an indoor multi-view stereo (MVS) dataset comprising 119 object-centric scenes. Each scene includes multi-view images captured from 49 fixed camera positions under 7 different lighting conditions, with ground-truth point clouds acquired using a structured light scanner. The dataset is partitioned into 79 scenes for training, 22 scenes for validation, and 18 scenes for evaluation.

BlendedMVS [5] is a large-scale dataset designed for MVS fine-tuning and validation. It includes 113 complex scenes such as cities, architectures, statues, and small objects. This dataset is divided into 106 scenes for training and 7 scenes for validation.

Tanks and Temples [6] serves as a large-scale MVS evaluation benchmark featuring both indoor and outdoor scenes. The dataset is categorized into intermediate (8 scenes) and advanced (6 scenes) sets, each exhibiting variation in scene scale, exposure conditions, and surface reflection.

B. Evaluation Metrics

We adhere to the standard evaluation procedures to assess our approach across various benchmark datasets. A systematic comparison of the evaluation metrics is presented in Table I.

1) DTU Dataset: The DTU dataset provides an official evaluation protocol to quantify the reconstruction error of MVS methods in terms of 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:

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 $\mathcal{G}$. Specifically, for every point $\mathbf{r}$ in $\mathcal{R}$, we compute its Euclidean distance to the closest point in $\mathcal{G}$, and then iterate this computation for all points in $\mathcal{R}$ to obtain the distance distribution. The mean of this distribution is defined as the accuracy:

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

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

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

Completeness is measured as the distance from $\mathcal{G}$ to $\mathcal{R}$, representing the extent to which the ground truth $\mathcal{G}$ is restored. Specifically, for every point $\mathbf{g}$ in $\mathcal{G}$, we compute its Euclidean distance to the closest point in $\mathcal{R}$ and iterate this computation over all points in $\mathcal{G}$ to get the distance distribution. The mean of this distribution is defined as completeness:

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

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

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

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

For quantitative benchmarking on the DTU evaluation set (see Table I of the manuscript), accuracy and completeness are averaged over 22 scenes. To balance accuracy and completeness, the overall score is computed as the arithmetic mean of the mean accuracy and mean completeness. A lower overall score indicates better reconstruction performance.

2) Tanks and Temples Dataset: The Tanks and Temples dataset uses precision and recall to quantify 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 (%). The precision and recall are defined as follows:

$$e_{\mathbf{r}\to\mathcal{G}} = \min_{\mathbf{g}\in\mathcal{G}} \|\mathbf{r}-\mathbf{g}\|_2, \tag{5}$$

$$\textit{Precision} = \frac{100}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}} [e_{\mathbf{r}\to\mathcal{G}} < d], \tag{6}$$

$$e_{\mathbf{g}\to\mathcal{R}} = \min_{\mathbf{r}\in\mathcal{R}} \|\mathbf{g}-\mathbf{r}\|_2, \tag{7}$$

$$\textit{Recall} = \frac{100}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}} [e_{\mathbf{g}\to\mathcal{R}} < d]. \tag{8}$$

To achieve a trade-off between precision and recall, the Tanks and Temples dataset uses the **F-score** to summarize

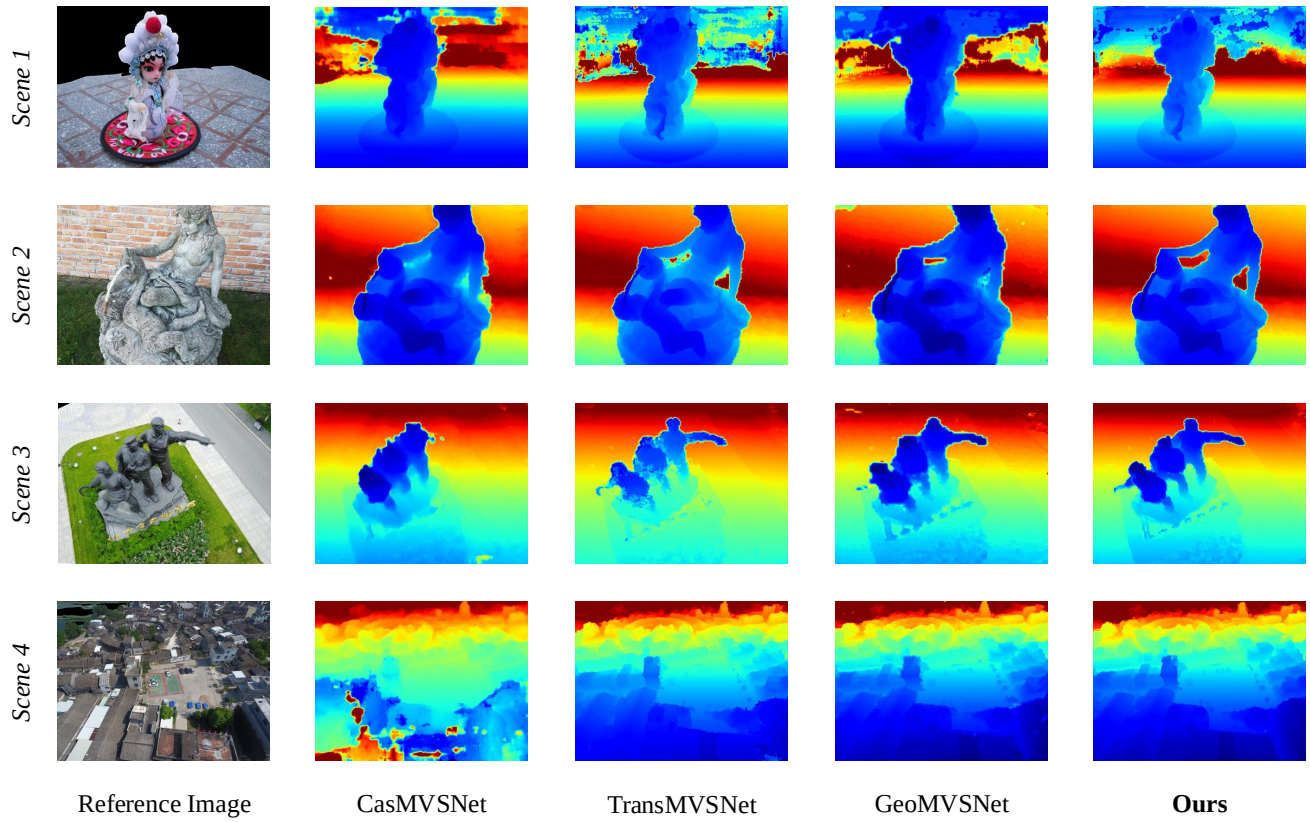

Fig. 1. Depth estimation results of recent learning-based methods [1]–[3] and our approach, evaluated across scenes with varying scales, depth ranges, geometry complexities, surface textures, and illumination conditions from the BlendedMVS validation set. This qualitative comparison corroborates the quantitative results presented in Table III of the manuscript.

overall reconstruction performance. A high F-score indicates that the reconstruction is both accurate and complete:

$$F\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{9}$$

For benchmarking on the Tanks and Temples dataset (see Table II of the manuscript), the mean F-score across 8 scenes from the intermediate set is reported as a summary measure of reconstruction performance.

3) BlendedMVS Dataset: Unlike the previous two datasets, the BlendedMVS dataset focuses on measuring depth estimation quality (see Table III of the manuscript). The endpoint error (**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. Benchmark Visualizations

We qualitatively compare the depth estimates of our method with those from recent learning-based MVS methods [1]–[3] on the BlendedMVS validation set [5] as shown in Fig. 1, 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 III of our manuscript, where our method obtains the lowest depth estimation errors.

The point reconstruction results for the complete DTU evaluation set [4] and the intermediate set of the Tanks and Temples [6] are shown in Fig. 2 and Fig. 3, respectively. Our method achieves accurate and complete point cloud reconstructions for both objects and large-scale outdoor scenes.

REFERENCES

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

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

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

[4] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 153–168, 2016.

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

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

Fig. 2. Point cloud reconstruction results for the DTU evaluation set.

Family

Francis

Horse

Lighthouse

M60

Panther

Playground

Train

Fig. 3. Point cloud reconstruction results for the intermediate set of the Tanks and Temples benchmark.