Supplementary Material for:
GoMVS: Geometrically Consistent Cost Aggregation for Multi-View Stereo

Jiang Wu[1] * Rui Li[1,2] * Haofei Xu[2,3] Wenxun Zhao[1] Yu Zhu[1] † Jinqiu Sun[1] Yanning Zhang[1]†
[1]Northwestern Polytechnical University [2]ETH Zürich [3]University of Tübingen, Tübingen AI Center

1. Network Efficiency

We compare the runtime, network parameters, and GPU memory with previous state-of-the-art methods. As shown in Tab. 1, our method achieves the best performances without relying on intensive network parameters. Meanwhile, our method achieves a comparable running time with previous methods. We found that the most computation lies in the GCP process, which can be further optimized for faster inference.

Methods	Overall	Params(M)	GPU(GB)	Time(s)
UniMVSNet [7]	0.315	0.934	8.8	0.271
TransMVSNet [3]	0.305	1.148	4.3	0.743
GeoMVSNet [11]	0.295	15.306	9.2	0.191
MVSFormer [2]	0.289	26.710	6.2	0.301
Ours	0.287	1.503	12.1	0.485

Table 1. **Comparison on network efficiency**.

2. Details of Monocular Normal Cues

We use an off-the-shelf surface normal network Omnidata [4] to estimate surface normal cues. Omnidata is trained using the images at the resolution of 384 × 384, which may not guarantee high-quality normal cues when the resolution of input images becomes higher. To this end, we adopt a divide-and-conquer strategy proposed by MonoSDF [10]. Specifically, 1) we divide the input image into multiple 384 × 384 patches with overlapping regions, 2) we estimate normal maps for all image patches, 3) we leverage the SVD decomposition to compute the optimal rotation matrix $\mathbf{R}$ to align the normal patches, which are then merged into high-resolution normal maps.

3. Evaluation of Different Normal Cues

In Section 4.4 of the main paper, we analyze the quantitative results of using different normal generation schemes [4, 8, 9] for geometrically consistent aggregation. In this section, we visualize the normal maps and their corresponding normal error maps for different normal generation schemes. Herein, the normal error is represented by the angle between the computed normal map and the ground truth normal map. As shown in Fig. 1, the normal computed from the intermediate depth map and cost volume present noisy predictions with artifacts, which can lead to degraded reconstruction quality. As a comparison, monocular normals provide more reasonable normal predictions with smoother surfaces, facilitating better aggregations as shown in the quantitative results. Additionally, we quantitatively evaluate the impact of normal quality on the final depth map. As shown in Tab. 2, higher normal accuracy leads to better depth prediction.

Method	Depth ACC.			Normal ACC.		
	MAE↓	<2mm↑	<4mm↑	<12.5↑	<22.5↑	<30↑
Depth2normal	15.58mm	77.30%	81.26%	31.94%	57.47%	70.43%
Cost2normal	15.44mm	78.72%	83.04%	24.36%	63.23%	**79.00**%
Omnidata	**13.04mm**	**79.83%**	**83.57**%	**32.36**%	**66.22**%	78.22%

Table 2. Normal and depth accuracy on DTU test set.

4. More Qualitative Results

Comparison of the reconstruction error. We visualize the reconstruction recall error maps on the Tank and Temple dataset as shown in Fig. 2. Our method achieves more faithful and complete reconstructions than previous methods.

Visulization of reconstructed point clouds. We present the reconstructed point clouds of DTU and Tanks and Temple datasets in Fig. 3 and 4, respectively. Our method faithfully captures rich details in both well-controlled laboratory scenes (DTU) and complex real-world scenes (TNT).

[*] indicates equal contributions and [†] indicates corresponding authors.

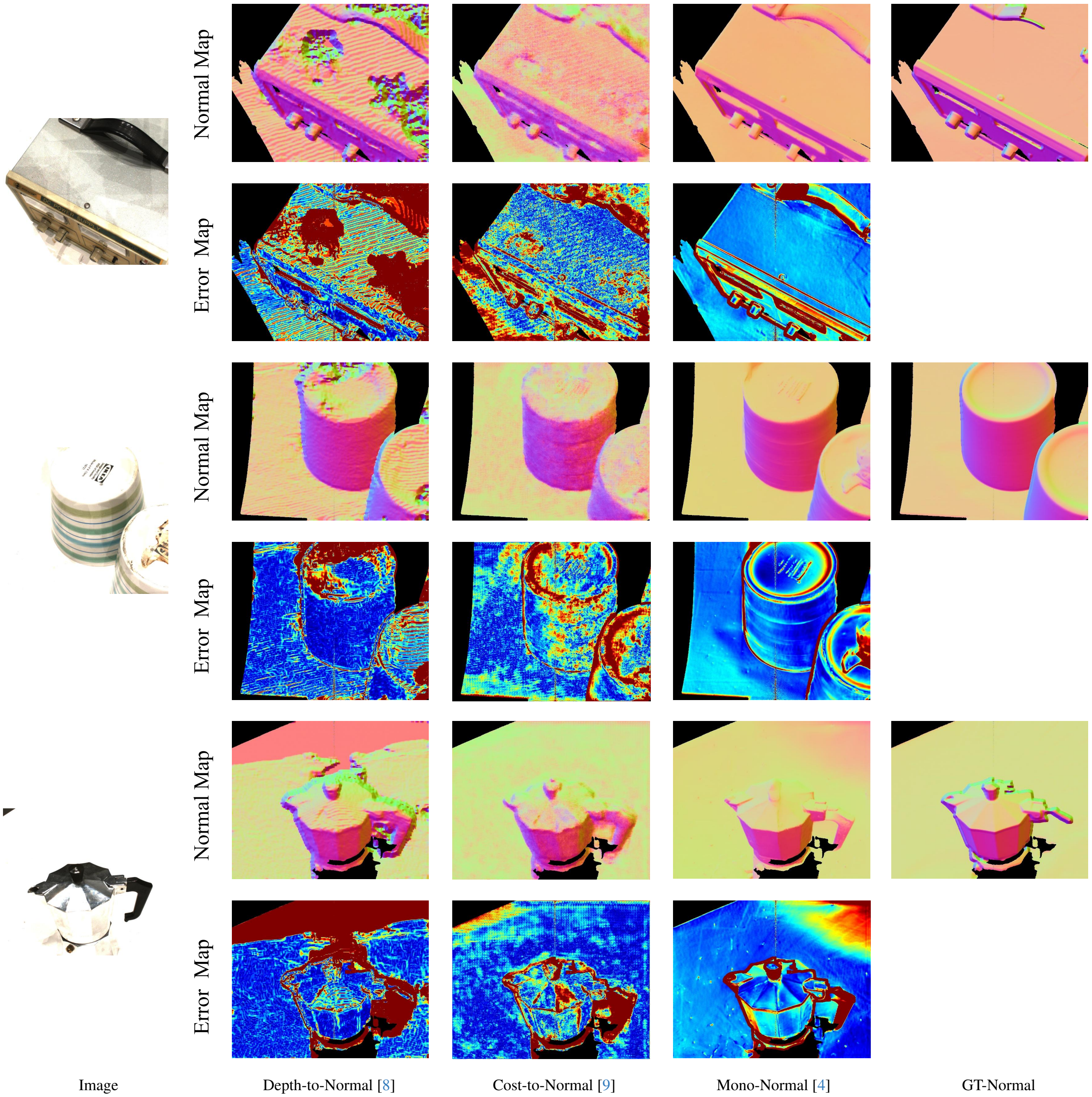

Figure 1. **Visualization of different normal maps and the corresponding error maps.** We visualize the normal maps and normal error maps of different methods, where the error maps are represented by the angle between the generated normals and the ground truth normals (red indicates larger errors, while blue indicates lower errors). Monocular normals demonstrate smoother and more reasonable predictions than other schemes.

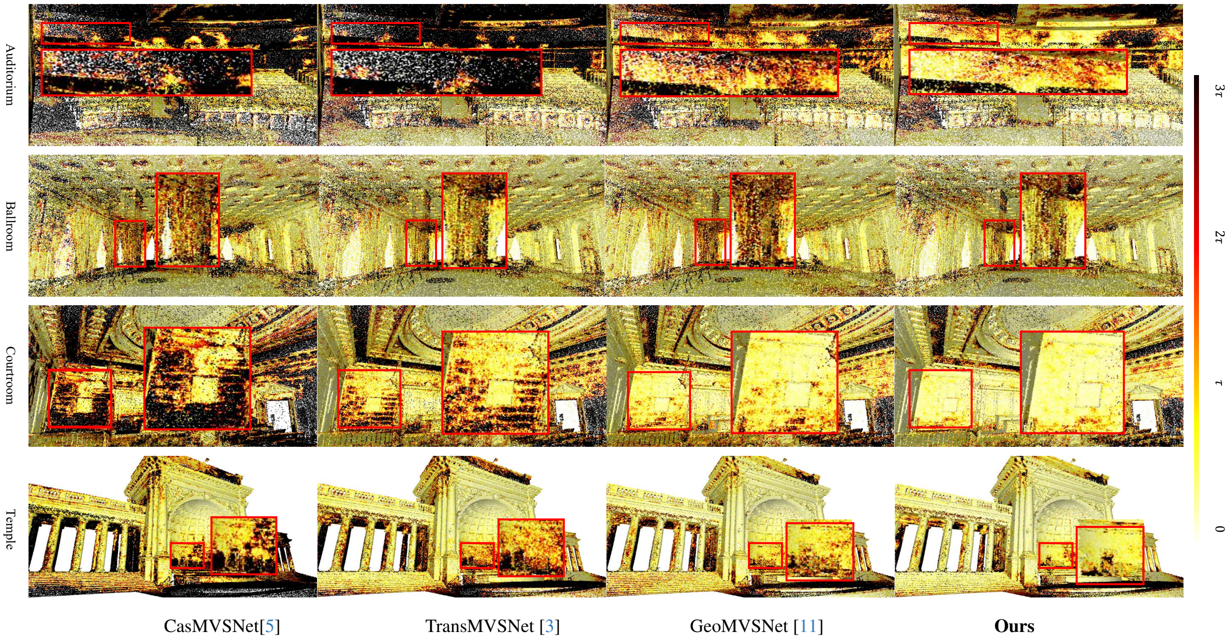

Figure 2. **Comparison of reconstruction recall error maps on the Tank and Temple [6] dataset.** Our method reconstructs more complete results than previous methods.

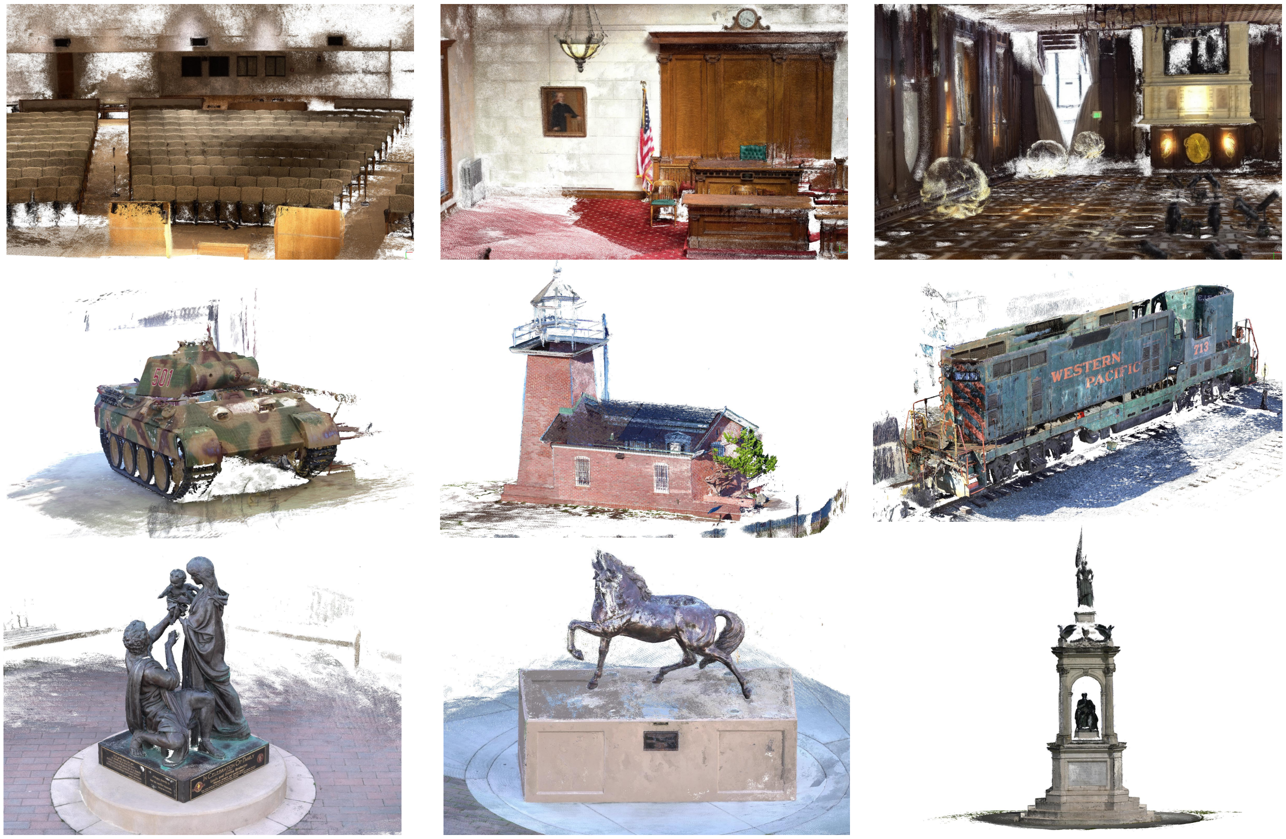

Figure 3. **Point cloud reconstructions of our method on the Tanks and Temples [6] dataset.**

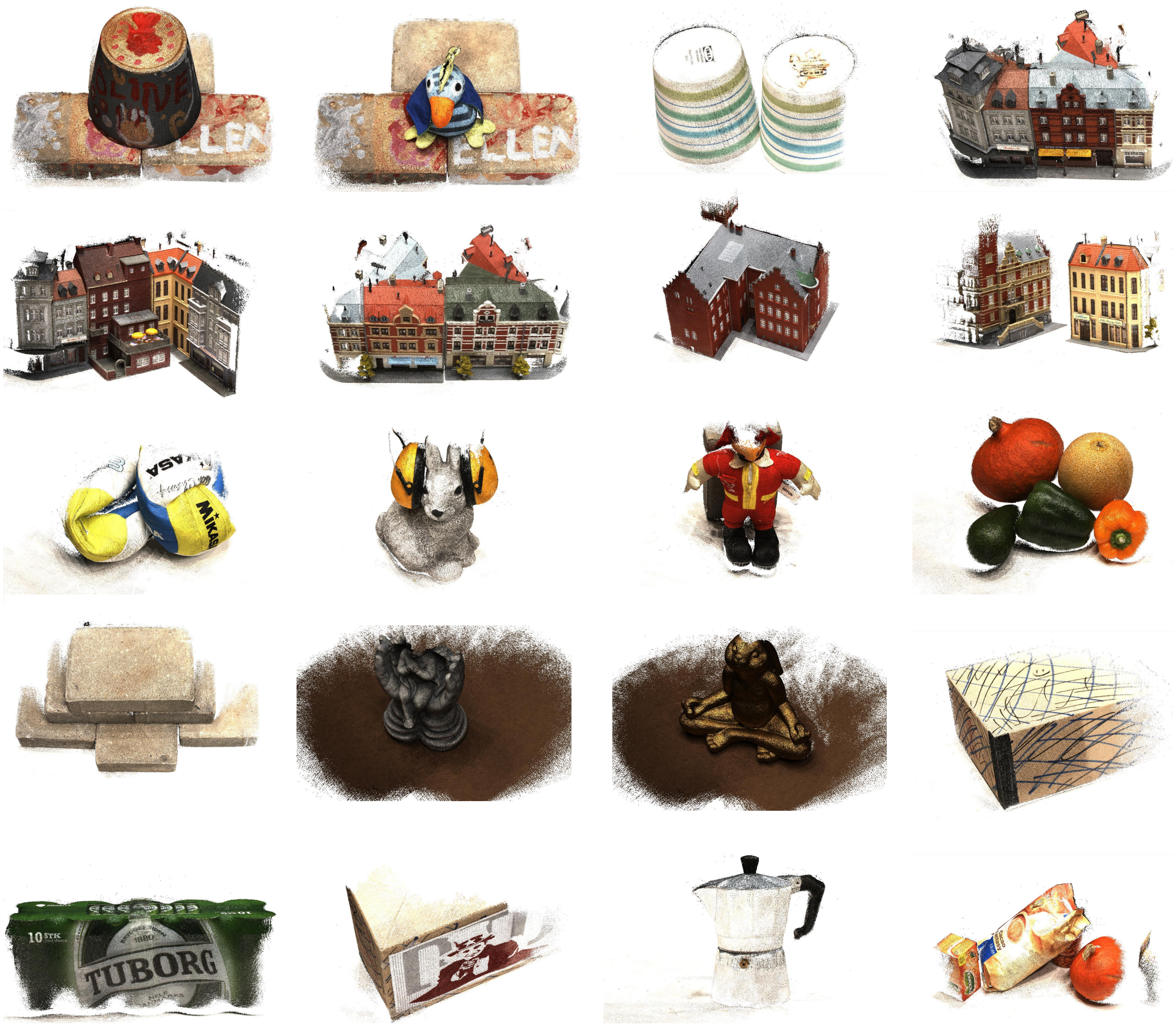

Figure 4. **Point clouds reconstructions of our method on the DTU [1] dataset.**

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 4

[2] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo. *arXiv preprint arXiv:2208.02541*, 2022. 1

[3] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 1, 3

[4] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 1, 2

[5] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping 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*, pages 2495–2504, 2020. 3

[6] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 3

[7] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022. 1

[8] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 1, 2

[9] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8626–8634, 2022. 1, 2

[10] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 1

[11] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21508–21518, 2023. 1, 3