Supplementary Material:
Multi-view Reconstruction via SfM-guided Monocular Depth Estimation

Haoyu Guo[1*] He Zhu[2*] Sida Peng[1] Haotong Lin[1] Yunzhi Yan[1]
Tao Xie[1] Wenguan Wang[1] Xiaowei Zhou[1] Hujun Bao[1†]
[1]Zhejiang University [2]Beijing Normal Univeristy

1. Baselines

We compare our method with the following baseline methods in four categories:

- *Monocular depth estimation*: Marigold [4], Depth-Anything [15] and Depth-Anything v2 [16] are monocular relative depth estimation methods. Due to discrepancies between their results and the true scale, we align their predictions with SfM depth before evaluation. Metric3D [18] is a monocular metric depth estimation method.
- *Depth completion:* SparseDC [6] is currently the state-of-the-art method for monocular depth completion. We input RGB images and SfM depth maps from each frame of the test scenes into SparseDC for comparison.
- *Optimization-based reconstruction:* MonoSDF [19] and StreetSurf [2] model scenes using Signed Distance Fields (SDF), optimize the SDF through differentiable rendering, and leverage monocular geometric cues to enhance reconstruction quality.
- *Learning-based MVS:* MVSNet [17], IGEV-MVS [14] and SimpleRecon [8] construct a cost volume from multi-view inputs to predict depth. NeuralRecon [10] aggregates multi-view features in world coordinates to predict TSDF volumes, thereby extracting scene geometry. Dust3R [13] uses a ViT model to reconstruct point maps from input image pairs.

For each depth estimation based method, we employ the same multi-view fusion technique as ours.

2. Comparison results

In addition to the geometric reconstruction results on DTU presented in the main paper, we provide more qualitative and quantitative comparisons and analyses in the supplementary material. These include qualitative comparisons of depth maps (in Figures 1 to 5), qualitative comparisons of geometric reconstructions (in Figures 6 to 8), and quantitative evaluations (in Tables 1 to 3). When visualizing depth maps, we normalize all methods using the same range and employ the Spectral colormap for consistent visualization. Based on these results, we draw the following conclusions:

- *Monocular relative depth estimation methods:* These methods, particularly Depth-Anything v2, exhibit visually impressive depth predictions. However, their numerical accuracy is not as strong, as evidenced by several observations. First, their quantitative evaluation results are not particularly high. Second, their reconstructed meshes exhibit some noise, often caused by inconsistencies between different views. Lastly, the color differences between their depth map visualizations and the ground truth in some areas also reflect numerical errors.
- *Monocular metric depth estimation methods:* Metric3D performs well on datasets like ScanNet and Waymo, partly because its training data includes real-world indoor and street-view data that closely resemble these scenes. However, Metric3D performs poorly on object-level and aerial datasets like DTU and UrbanScene3D.
- *Depth completion:* SparseDC is not robust to real SfM depth inputs, which often contain noise, resulting in suboptimal depth completion and final reconstruction results.
- *Optimization-based reconstruction methods:* These methods (e.g., MonoSDF, StreetSurf) achieve high-quality reconstructions in indoor scenes but suffer from very slow optimization processes. Moreover, their performance is less competitive in large-scale street-view scenes due to limited expressiveness.
- *Learning-based MVS methods:* IGEV predicts relatively accurate depth maps in areas with enough views and rich textures. However, on DTU, its performance is hindered by the limited number of views, leading to suboptimal matching. While IGEV performs well overall in indoor and outdoor scenes, it struggles in low-texture and boundary regions. NeuralRecon and SimpleRecon achieve good results on ScanNet, however, we found that they perform very poorly on other datasets.

* Equal contribution
† Corresponding authors

- *Our method:* Murre not only produces visually pleasing depth predictions but also achieves higher numerical accuracy. It is robust in low-texture regions and performs well across various datasets, demonstrating consistent and reliable results.

3. Implementation of LCM

We analyse the trade-off between speed and reconstruction quality in the main paper, where we distill our model using Latent Consistency Model (LCM) [7] to reduce the number of denoising steps. Specifically, we fix the UNet in the original model as the teacher UNet and use it to initialize the student UNet and the target UNet. During training, the student UNet is optimized using consistency objective, while the target UNet updates its parameters via exponential moving average (EMA). During inference, the trained LCM enables few-step denoising, achieving satisfactory results even with a single step.

4. Additional ablation study on sfm method

Without any retraining, we directly evaluate our performance using PixSfM with two different matchers: SuperPoint+SuperGlue and LoFTR, as shown in Figure 9.

5. Visualization with texture

To better visualize the reconstruction results of our method, we apply an off-the-shelf texture mapping method [12] to our meshes on the UrbanScene3D dataset. The results are presented in Figure 10.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 4, 7

[2] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 1

[3] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 3

[4] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *CVPR*, 2024. 1

[5] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *ECCV*, 2022. 5, 6, 8

[6] Chen Long, Wenxiao Zhang, Zhe Chen, Haiping Wang, Yuan Liu, Peiling Tong, Zhen Cao, Zhen Dong, and Bisheng Yang. Sparsedc: Depth completion from sparse and non-uniform inputs. *Information Fusion*, 2024. 1

[7] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[8] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *ECCV*, 2022. 1

[9] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3, 5, 8

[10] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *CVPR*, 2021. 1

[11] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 4, 6, 7

[12] Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *ECCV*, 2014. 2

[13] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. *CVPR*, 2024. 1

[14] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, 2023. 1

[15] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *CVPR*, 2024. 1

[16] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024. 1

[17] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 1

[18] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023. 1

[19] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 2022. 1

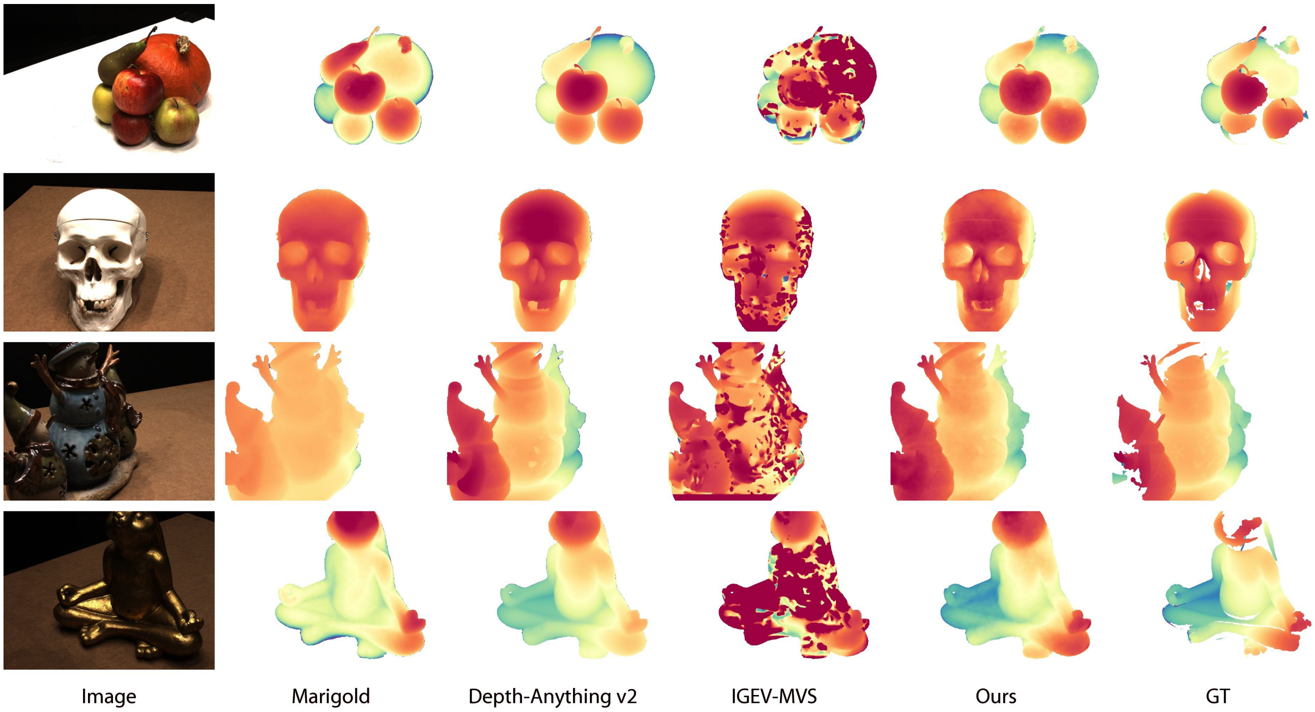

Figure 1. Qualitative comparison of depth estimation on DTU [3].

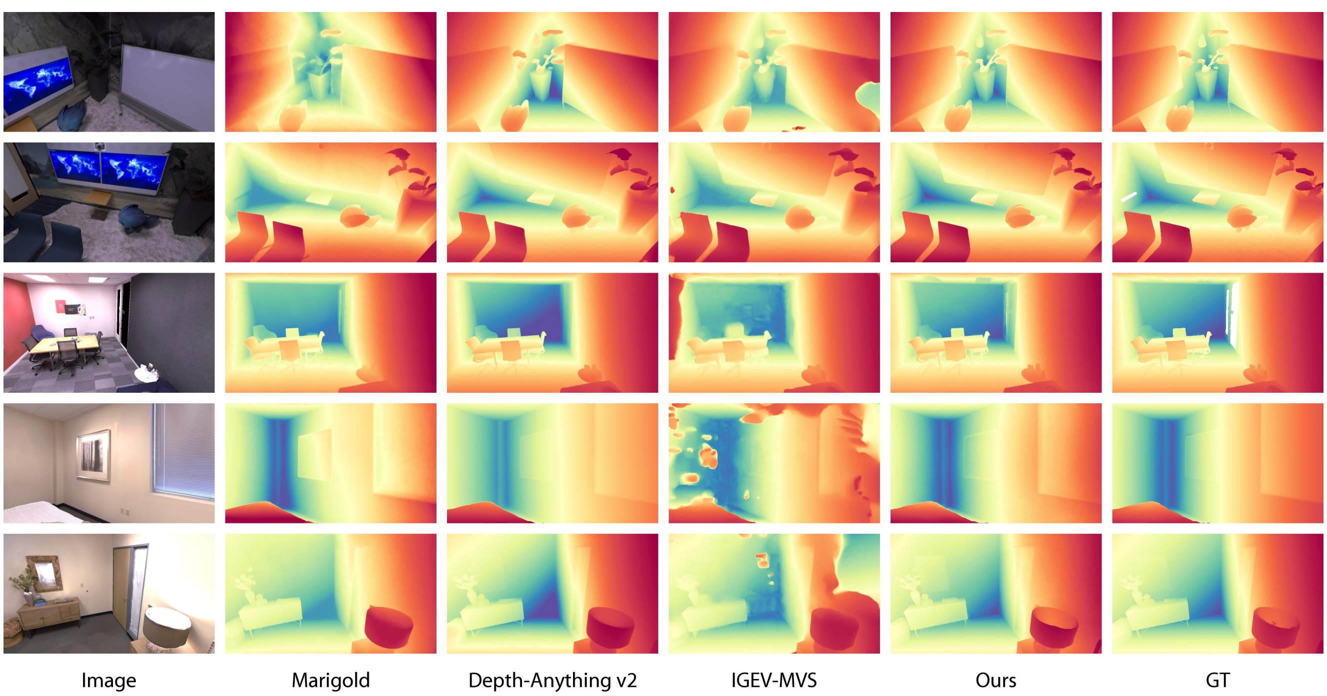

Figure 2. Qualitative comparison of depth estimation on Replica [9].

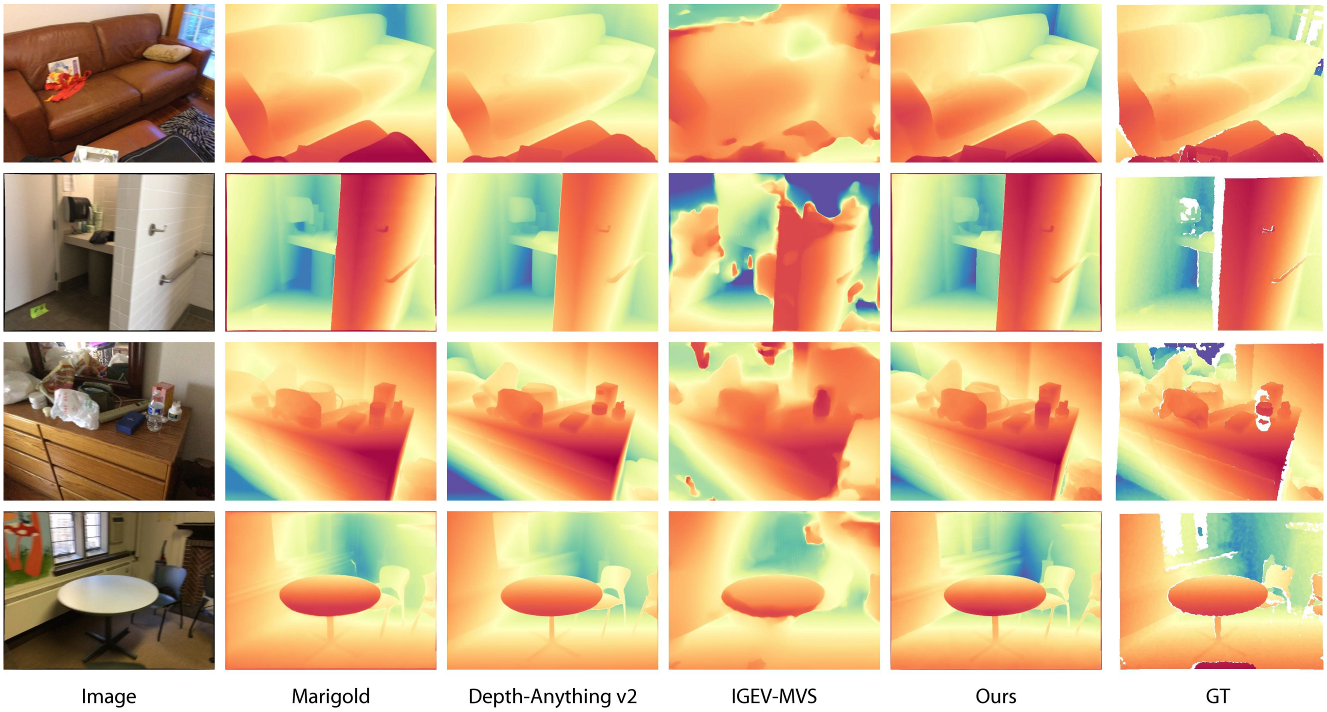

Figure 3. Qualitative comparison of depth estimation on ScanNet [1].

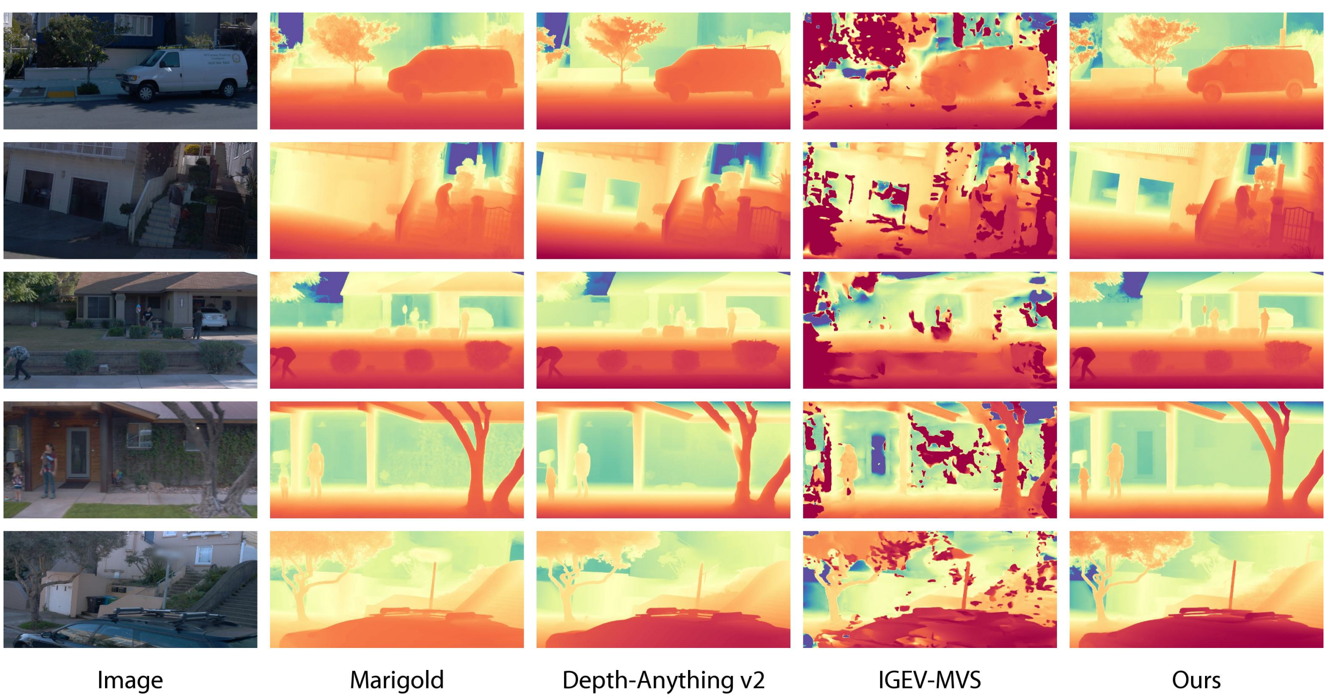

Figure 4. Qualitative comparison of depth estimation on Waymo [11].

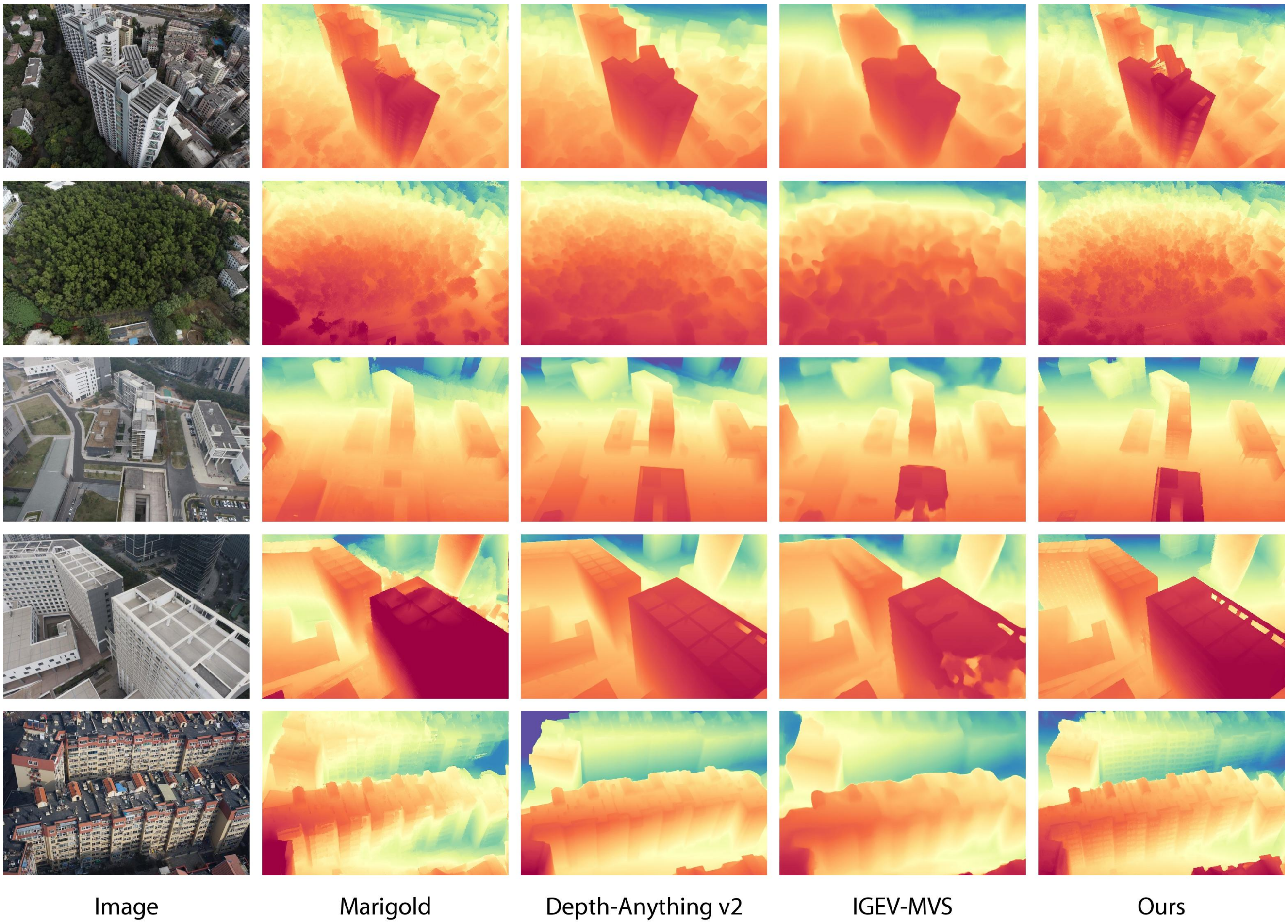

Figure 5. Qualitative comparison of depth estimation on UrbanScene3D [5].

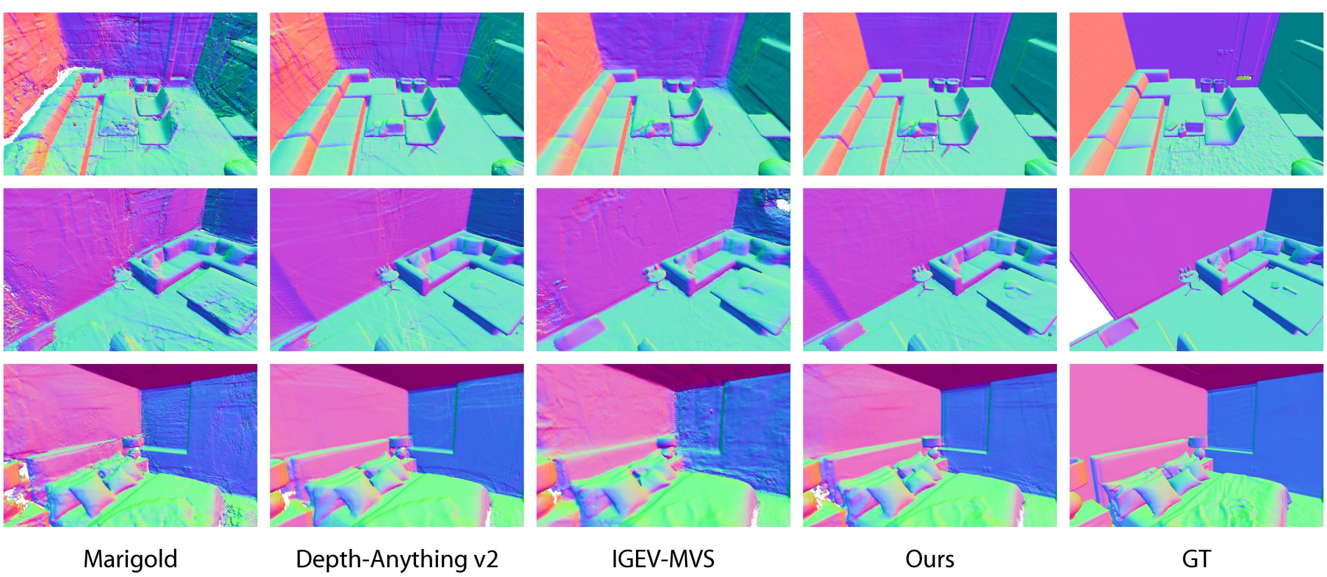

Figure 6. Qualitative comparison of geometric reconstruction on Replica [9].

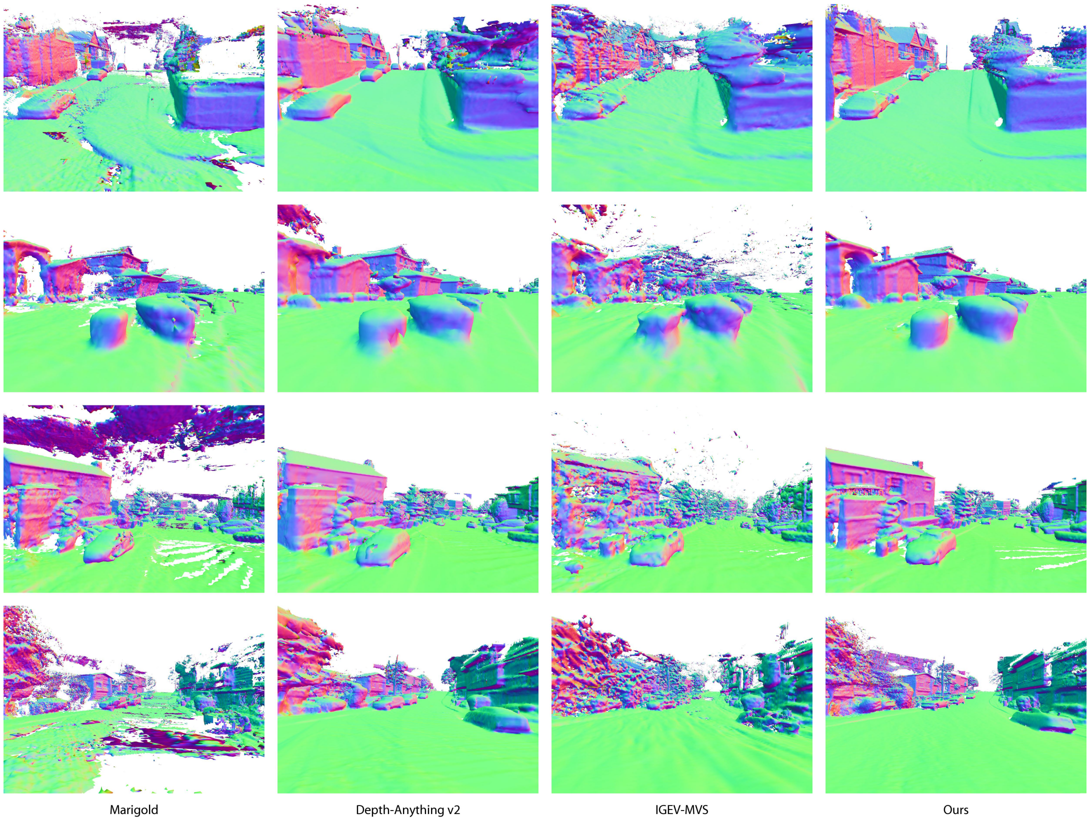

Figure 7. Qualitative comparison of geometric reconstruction on Waymo [11].

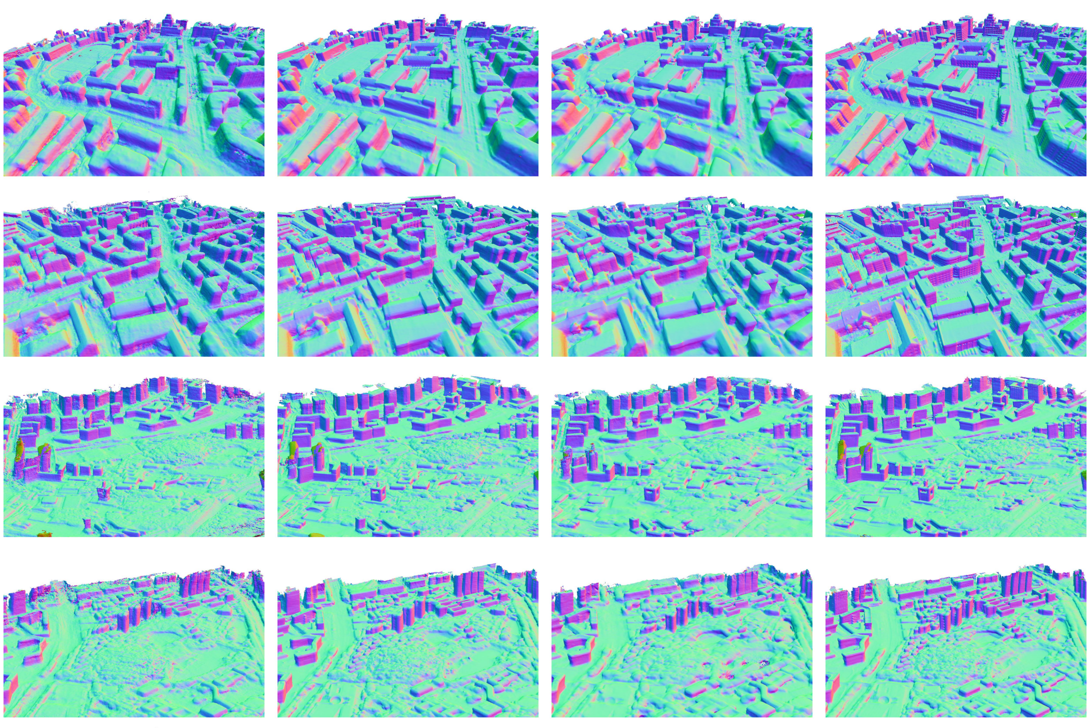

Figure 8. Qualitative comparison of geometric reconstruction on UrbanScene3D [5].

Table 1. **Quantitative results on Waymo [11].** The metrics for COLMAP, F2NeRF, and StreetSurf are sourced from the StreetSurf paper. Note that their evaluations are conducted in LiDAR space, whereas ours and other baselines are in image space. While the assessment results from both approaches should be closely aligned, they may not be identical. We report their metrics for reference.

Sequence	COLMAP	F2-NeRF	StreetSurf	Marigold	Depth-Anything	Depth-Anything v2	MVSNet	IGEV-MVS	Metric3D	SparseDC	Ours
seg1006130..	7.10	8.87	2.99	2.92	2.42	2.76	15.53	4.13	2.29	7.93	2.11
seg1027514..	7.47	16.52	2.91	2.95	2.70	2.77	13.98	6.06	2.67	10.54	2.47
seg1067626..	9.06	35.59	4.34	5.74	5.29	5.50	13.95	9.55	4.57	17.36	5.00
seg1137922..	12.39	20.10	5.70	7.41	6.08	6.57	13.35	10.77	5.01	16.85	5.61
seg1172406..	13.62	9.00	2.57	2.31	1.88	2.16	16.73	4.15	1.49	7.40	1.57
seg1287964..	10.34	6.73	3.19	3.59	3.27	3.34	10.14	5.99	3.39	10.52	3.05
seg1308545..	8.64	15.50	4.12	3.81	3.52	3.66	14.15	6.12	2.91	9.94	3.18
seg1314219..	6.75	19.30	3.48	4.27	3.82	3.81	12.61	7.18	3.60	12.19	3.28
seg1319679..	7.63	23.50	4.76	4.73	4.31	4.45	14.46	5.58	3.67	11.11	3.99
seg1323841..	7.32	20.19	3.13	3.57	3.47	3.44	12.88	6.74	3.33	12.20	2.95
seg1347637..	5.93	21.72	1.84	2.74	2.75	2.66	17.72	2.54	2.09	5.19	1.85
seg1400454..	8.08	39.85	3.29	2.89	2.63	2.72	11.66	6.07	2.58	11.18	2.43
seg1434813..	8.48	35.96	4.74	5.93	6.19	6.20	17.82	6.05	4.50	10.79	4.30
seg1442480..	7.85	36.35	2.97	3.70	3.40	3.40	12.92	7.06	2.98	12.80	2.96
seg1486973..	5.52	3.53	2.82	2.25	1.72	2.10	18.86	3.15	1.70	6.71	1.48
seg1506235..	7.84	27.61	2.40	2.36	2.19	2.12	13.32	6.02	2.22	11.00	1.83
seg1522170..	11.28	16.66	4.87	5.49	5.30	5.58	17.53	6.75	4.61	11.72	4.35
seg1527063..	2.62	7.82	1.98	1.80	1.56	1.81	20.80	3.83	1.38	7.38	1.32
seg1534950..	4.31	7.80	2.56	2.94	2.67	2.80	14.00	4.48	2.29	7.53	1.92
seg1536582..	6.57	10.41	2.47	1.94	1.54	1.76	20.86	3.15	1.48	7.44	1.46
seg1586862..	5.94	18.78	2.60	3.16	2.98	3.14	14.71	5.45	2.53	8.64	2.47
seg1634531..	5.31	11.85	2.23	2.33	1.97	2.16	15.59	3.59	1.53	7.54	1.79
seg1647019..	10.36	12.25	4.31	4.64	4.20	4.27	14.20	7.28	3.88	12.05	3.74
seg1660852..	5.11	4.72	3.91	3.50	2.92	2.93	17.28	4.23	2.62	7.95	2.68
seg1664636..	6.54	13.86	2.26	2.53	2.54	2.61	17.42	4.04	1.94	7.45	1.66
seg1776195..	14.52	25.24	3.90	4.22	3.72	3.76	12.04	7.24	3.58	12.24	3.56
seg3224923..	5.42	7.16	3.53	3.00	2.43	2.72	14.79	4.49	2.07	8.57	2.21
seg3425716..	18.81	30.68	3.00	3.67	3.20	3.03	18.46	7.55	3.23	9.94	2.95
seg3988957..	6.07	5.66	3.30	3.36	2.95	2.98	12.66	5.78	3.07	10.91	2.90
seg4058410..	5.46	7.02	2.62	3.05	3.00	2.92	12.62	4.62	2.37	8.24	2.48
seg8811210..	7.16	27.30	3.83	3.28	2.94	3.04	16.42	6.40	2.75	10.75	2.70
seg9385013..	9.10	49.34	4.52	5.03	4.34	4.42	17.68	9.89	4.08	14.63	4.33
Average	8.08	18.65	3.35	3.60	3.25	3.36	15.22	5.81	2.89	10.21	2.83

Table 2. Quantitative results on ScanNet [1].

	COLMAP	Manhattan-SDF	MonoSDF	Marigold	Depth-Anything	Depth-Anything v2	Metric3D	SparseDC	NeuralRecon	SimpleRecon	MVSNet	IGEV-MVS	Ours
0050_00	0.563	0.673	-	0.669	0.674	0.669	0.507	0.250	0.661	0.718	0.075	0.391	0.750
0084_00	0.631	0.630	-	0.733	0.692	0.863	0.516	0.204	0.805	0.881	0.066	0.600	0.732
0580_00	0.590	0.632	-	0.627	0.688	0.644	0.441	0.245	0.484	0.542	0.137	0.512	0.720
0616_00	0.365	0.472	-	0.543	0.597	0.578	0.437	0.179	0.518	0.590	0.077	0.430	0.596
Average	0.537	0.602	0.733	0.643	0.663	0.689	0.475	0.220	0.617	0.683	0.089	0.483	0.700

Table 3. Quantitative results on Replica [9].

	MonoSDF	Marigold	Depth-Anything	Depth-Anything v2	Metric3D	SparseDC	MVSNet	IGEV-MVS	Ours
room_1	-	0.61	0.79	0.86	0.77	0.25	0.54	0.84	0.84
office_0	-	0.52	0.69	0.71	0.47	0.29	0.73	0.85	0.90
office_2	-	0.58	0.52	0.62	0.57	0.23	0.56	0.78	0.82
Average	0.86	0.57	0.67	0.73	0.61	0.25	0.61	0.82	0.85

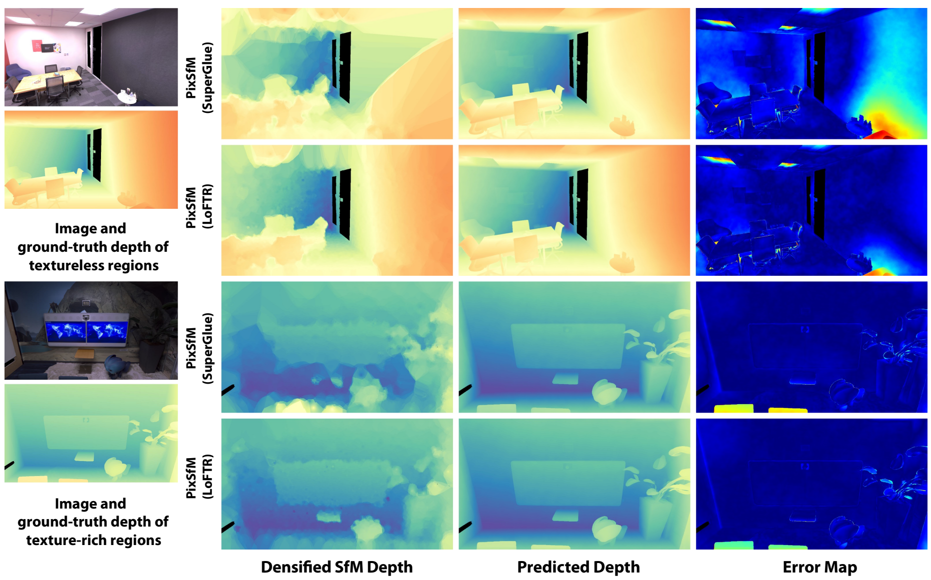

Figure 9. Results of our method based on PixSfM.

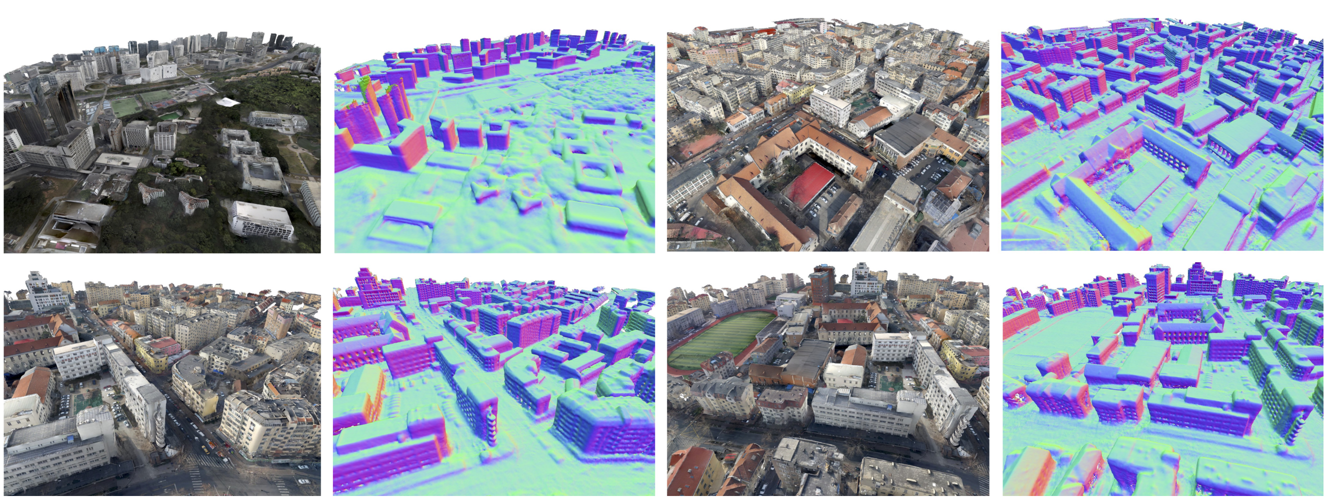

Figure 10. Visualization of our results with texture from texture mapping on UrbanScene3D [5].