Supplementary Materials
MegaSurf: Learnable Sampling Guided Neural Surface Reconstruction Framework of Scalable Large Scene

Anonymous Authors

1 EXPERIMENT DETAIL

This paper experiment used a total of 71 blocks of data, including: Residence-UrbanScene3D 16 blocks, Polytech-UrbanScene3D 12 blocks, Artsci-UrbanScene3D 6 blocks, Building-Mill19 15 blocks, Rubble-Mill19 6 blocks and Songshanhu-Ours 16 blocks. The training time of MegaSurf is basically the same as Bakedangelo. Without hyperparameter optimization, it takes 8 hours and 22G GPU memory to train NSR on an A6000.

2 EVALUATION

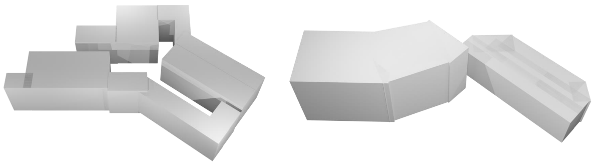


Figure 1: The proxy we used to crop the scene for evaluation.

Because the NSR results often have many floaters that affect the evaluation, we first use ray casting according to the camera pose to generate a depth map under each image pose. Then, we project the depth map to form a point cloud. Next, we use the proxy to crop the point cloud and uniformly downsample it to obtain the final point cloud for evaluation. The LiDAR GT and evaluation software are provided by UrbanScene3D [3].

3 ADDITIONAL ABLATION STUDY

To demonstrate the effectiveness of our Divide-and-Conquer training strategy and L_{nooc}, we conducted additional ablation studies. Table 1 presents the quantitative comparison results, while Figure 3 and 2 depict the visual comparisons. Among them, we used Bakedangelo [5] as the **Base** (Fig 2,3 a) method. **w/o Step1&Step2** (Fig 2,3 b) represents not using step 1 and step 2 of the Divide-and-Conquer strategy, not training the LG Guider separately, but instead training the LG Guider and radiance field simultaneously using geometry cues and images. This means that the losses involved in all three steps will simultaneously contribute. **w/o Step2** (Fig 2,3 c) means that we train the LG Guider using geometry cues but do not perform the operation of step 2 (initializing the render net using the LG Guider). **Freeze** ***Prop*** (Fig 2,3 d) implies that after training the LG Guider and initializing the radiance field, we freeze the parameters of the LG Guider. The rendering loss in step 3 can no longer modify its parameters. **w/o** L_{nocc} (Fig 2,3 e) implies not using the L_{nocc} regularization term.

The Base method always suffers into shape radiance ambiguity, although the details in its reconstruction results are very realistic. Due to the presence of irreducible errors in the loss, surfaces often appear excessively smooth in w/o Step1&Step2. In w/o Step2, not initializing the render net in step 2 leads to an immediate disruption of the accurate geometric information learned by the LG Guider and geometry net when the rendering loss is introduced, resulting in ambiguity. For Freeze *Prop*, due to the imperfect information learned by the LG Guider in step 1, which requires continuous optimization in step 3, not optimizing the LG Guider will result in significant degradation of the reconstruction results. Additionally, not using L_{nocc} often leads to ambiguity at some corners, manifested as surface protrusions. Our complete model can preserve realistic details while overcoming ambiguity.

Table 1: Quantitative evaluation of ablation study. The full MegaSurf model achieves the best surface reconstruction performance.

Method	CD ↓	Acc_{95} ↓	$Comp_{95}$ ↓	$Overall_{95}$ ↓
Artsci				
Base	1.3938	0.3319	0.5813	0.4566
w/o Step1&Step2	1.2428	0.3478	0.5244	0.4361
w/o Step2	1.1479	0.3066	0.4488	0.3777
Freeze *Prop*	1.4585	0.3736	0.6982	0.5359
w/o L_{nocc}	1.1322	0.2980	0.4649	0.3815
Ours	1.0574	0.2990	0.4138	0.3564
Polytech				
Base	1.1029	0.2989	0.3969	0.3479
w/o Step1&Step2	0.6868	0.1751	0.2338	0.2045
w/o Step2	0.6852	0.1953	0.2148	0.2051
Freeze *Prop*	0.8350	0.2218	0.3182	0.2700
w/o L_{nocc}	0.6782	0.1749	0.2120	0.1935
Ours	0.6593	0.1763	0.2086	0.1925

4 MATCHING COST

We introduce the matching cost which used as a matching quality indicator in PatchMatch operation [4].

Given N images $\{\boldsymbol{I}_i\}_{i=0}^{N-1}$ with calibrated camera parameters $\{\boldsymbol{P}_i\}_{i=0}^{N-1}$, where $\boldsymbol{P}_i = \boldsymbol{K}_i\left[\boldsymbol{R}_i \mid \boldsymbol{t}_i\right]$, we first generate a random 3D plane hypothesis in local coordinate $\boldsymbol{\theta} = \left[\boldsymbol{n}^{\top}, d\right]^{\top}$ for each pixel in reference image I_r, where $\boldsymbol{n}$ is the normal of the plane and d is distance from the origin to the plane. The normalization cross correlation (NCC) is defined as:

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{x}', \boldsymbol{H} = \boldsymbol{K}_s\left(\boldsymbol{R}_s\boldsymbol{R}_r^T - \frac{\boldsymbol{R}_s\left(\boldsymbol{R}_s^T\boldsymbol{t}_s - \boldsymbol{R}_r^T\boldsymbol{t}_r\right)\boldsymbol{n}^T}{d}\right)\boldsymbol{K}_r^{-1}, \quad (1)$$

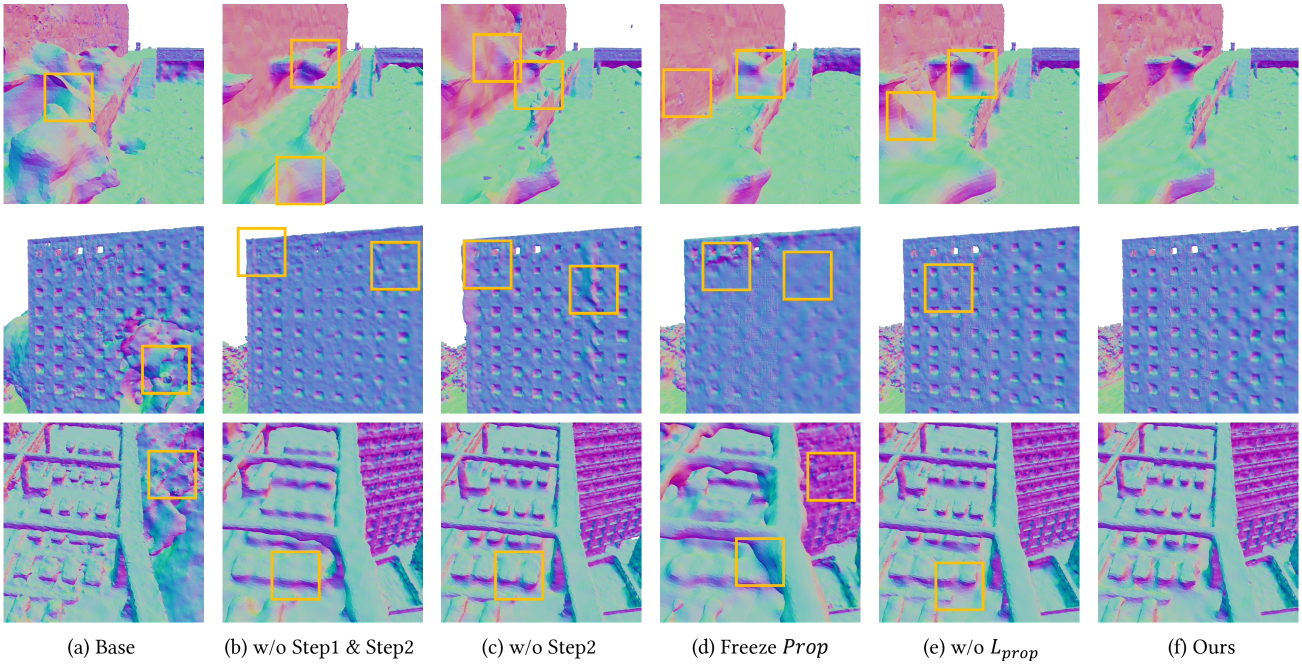

(a) Base (b) w/o Step1 & Step2 (c) w/o Step2 (d) Freeze *Prop* (e) w/o L_{prop} (f) Ours

Figure 2: The visual comparison of ablation studies conducted on the Polytech dataset. Please refer to Section 3 for the definitions of each name. The position inside the box indicates the presence of significant geometric errors or excessive smoothness.

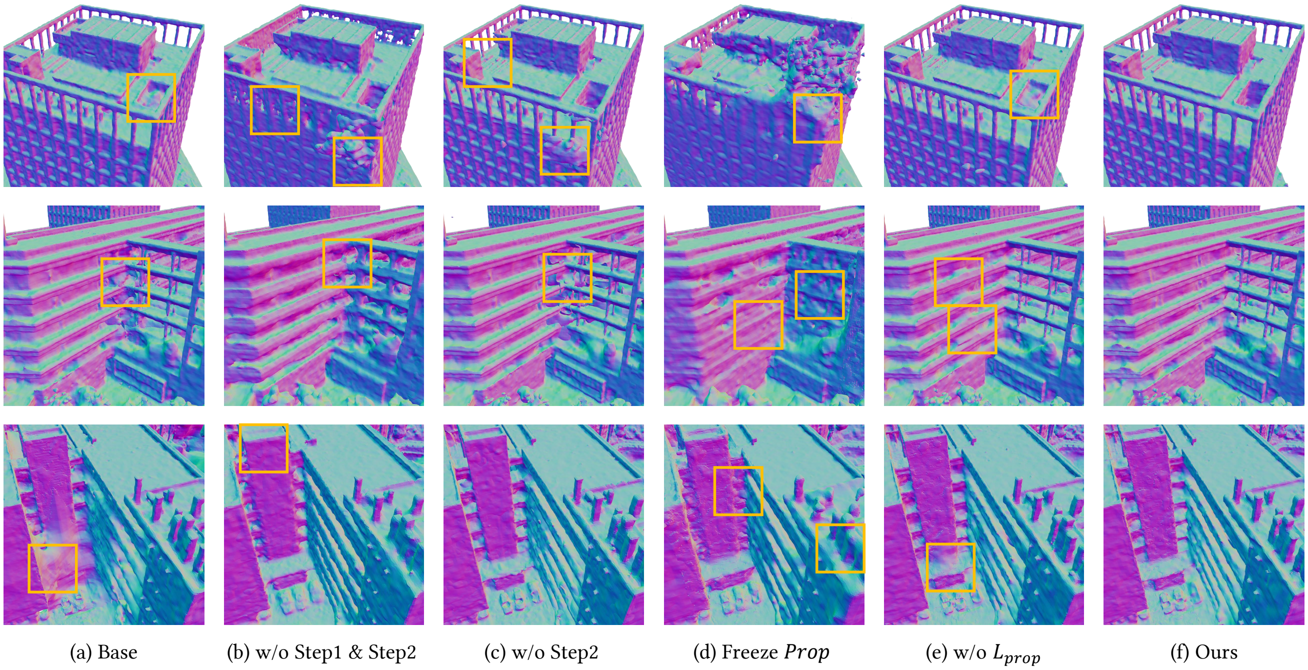

(a) Base (b) w/o Step1 & Step2 (c) w/o Step2 (d) Freeze *Prop* (e) w/o L_{prop} (f) Ours

Figure 3: The visual comparison of ablation studies conducted on the Artsci dataset. Please refer to Section 3 for the definitions of each name. The position inside the box indicates the presence of significant geometric errors or excessive smoothness.

$$NCC\left(I_r\left(q_x\right), I_s\left(q_{x'}\right)\right) = \frac{\mathrm{Cov}\left(I_r\left(q_x\right), I_s\left(q_{x'}\right)\right)}{\sqrt{\mathrm{Var}\left(I_r\left(q_x\right)\right)\mathrm{Var}\left(I_s\left(q_{x'}\right)\right)}}, \quad (2)$$

where x' is the corresponding pixel for x in I_s, $\boldsymbol{H}$ is the plane-induced homography, q_x is the pixels in the 5x5 patch which take x as center, Cov represents the covariance and Var represents the variance. Then the mean of top k largest NCC is set as the final matching cost:

$$\boldsymbol{E}_{NCC} = \frac{1}{k}\left(\sum_k 1 - NCC_k\right). \quad (3)$$

5 GEONEUS

We also tested the GeoNeuS [1] and Geoangelo on the UrbanScene3D datasets. The results are shown in Figure 4 and Figure 5. We train each method for 200k iterations per block. The Geoangelo adds the NCC module of GeoNeuS to the Bakedangelo [5]. The Geoangelo implemented by us is basically the same as the Geoangelo implemented in SDFStudio [5] regarding reconstruction results.

The reconstruction results based on the MLP method are too smooth. Using hash tables as the scene representation is promising for large-scale scene reconstruction. In the experiment, we find that the scene often cannot be reconstructed if the NCC loss is provided. The first reason is that the NCC module requires that all sampled rays must come from pixels in one image, instead of randomly sampling from all image pixels like other NSR methods, the reconstruction quality is seriously degraded in large scenes. Another reason is that unlike depth prior, NCC cannot clearly indicate the difference between the current and target states. In a complex outdoor environment, NCC often causes optimization to fall into the local optimization, which fails to reconstruct the scene.

Because GeoNeus and Geoangelo often fail to reconstruct, we are unable to assemble all block reconstruction results together for numerical comparison with other methods. Besides, calculating NCC requires a reference image and corresponding source images. For large scenes, GPU memory cannot store all images. If we adopt a joint learning strategy to optimize NSR and MVS simultaneously, like HelixSurf [2], the data transfer between the host and the device must be carried out in each iteration, which consumes much time. (Geoangelo: 200K, 3 days)

6 MORE EXPERIMENT RESULTS

Here we present more experimental results. Figure 6,7,8,9,10,11 show the entire scene mesh of each method on four datasets. Figure 12 and 13 shows more close-up comparisons. ACMH's point cloud reconstructions often capture all details but also introduce significant noise, leading to erroneous geometry and overly smooth surfaces after triangulation. Bakedangelo suffers from shape-radiance ambiguity. Monoangelo can overcome ambiguity but sacrifices detail sharpness. Our method not only overcomes ambiguity but also preserves realistic details.

REFERENCES

[1] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. 2022. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems* 35 (2022), 3403–3416.

[2] Zhihao Liang, Zhangjin Huang, Changxing Ding, and Kui Jia. 2023. Helixsurf: A robust and efficient neural implicit surface learning of indoor scenes with iterative intertwined regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13165–13174.

[3] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. 2022. Capturing, Reconstructing, and Simulating: the UrbanScene3D Dataset. In *ECCV*. 93–109.

[4] Qingshan Xu, Weihang Kong, Wenbing Tao, and Marc Pollefeys. 2022. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4945–4963.

[5] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. 2022. SDFStudio: A Unified Framework for Surface Reconstruction. https://github.com/autonomousvision/sdfstudio

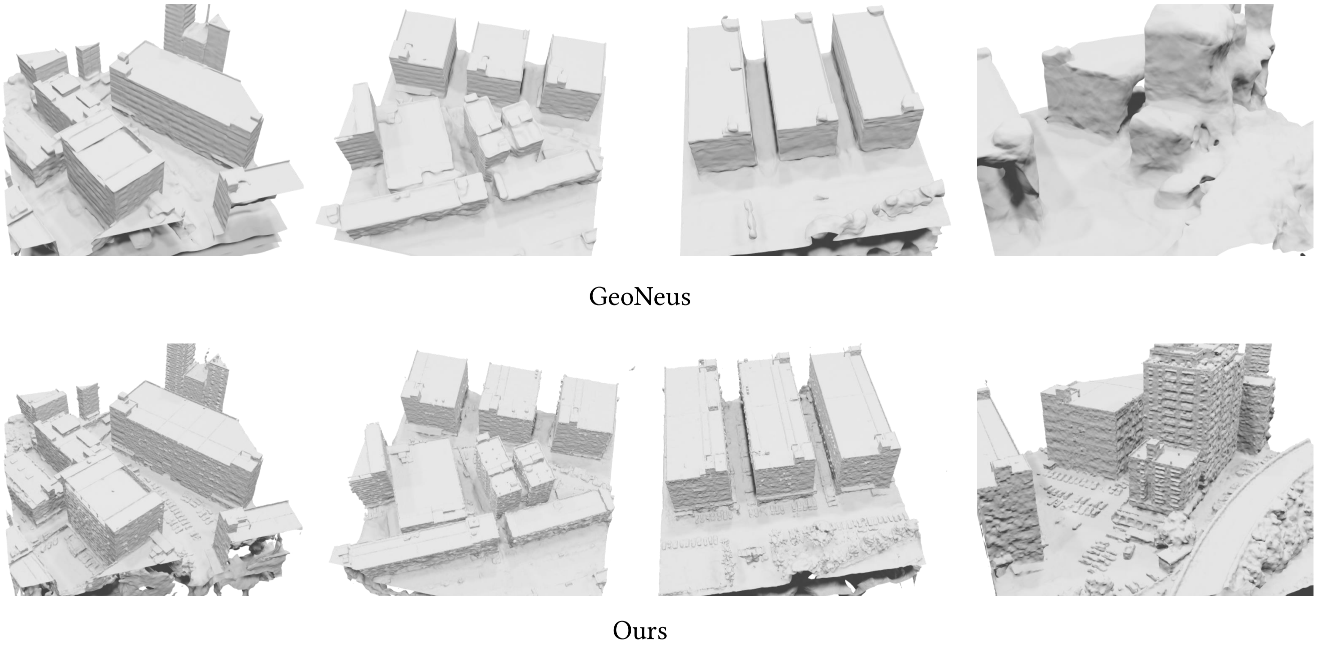


Figure 4: Reconstruction results of GeoNeuS on UrbanScene3D dataset. The reconstruction results often exhibit excessive smoothness and sometimes fail to reconstruct.

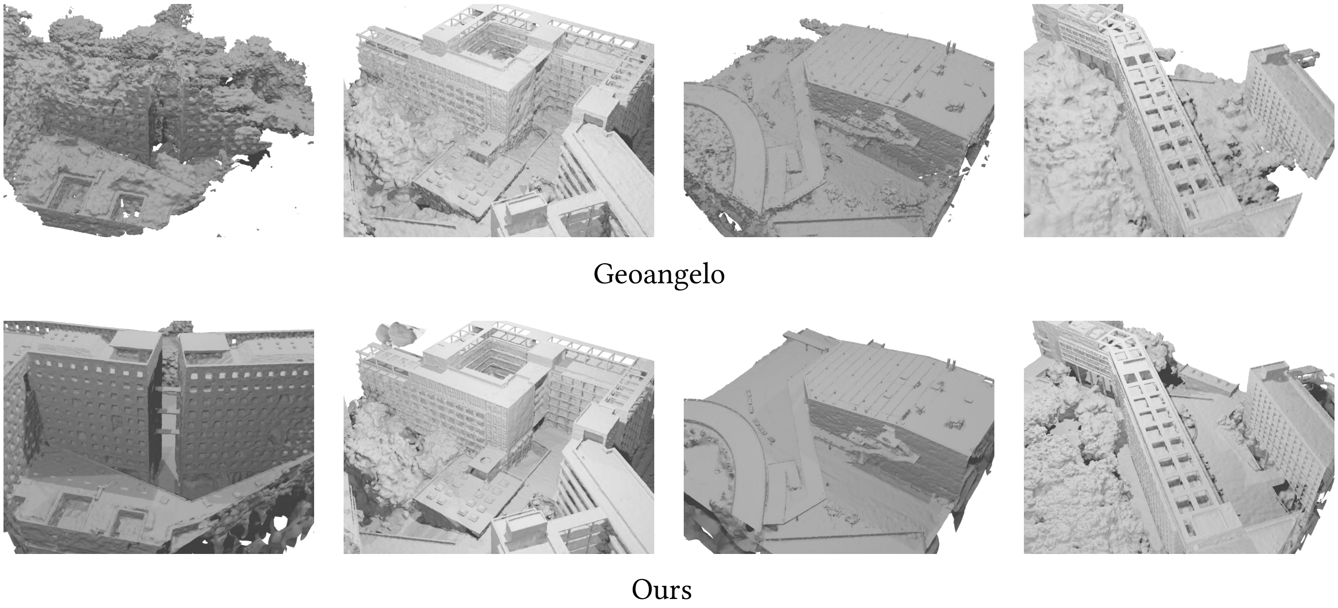
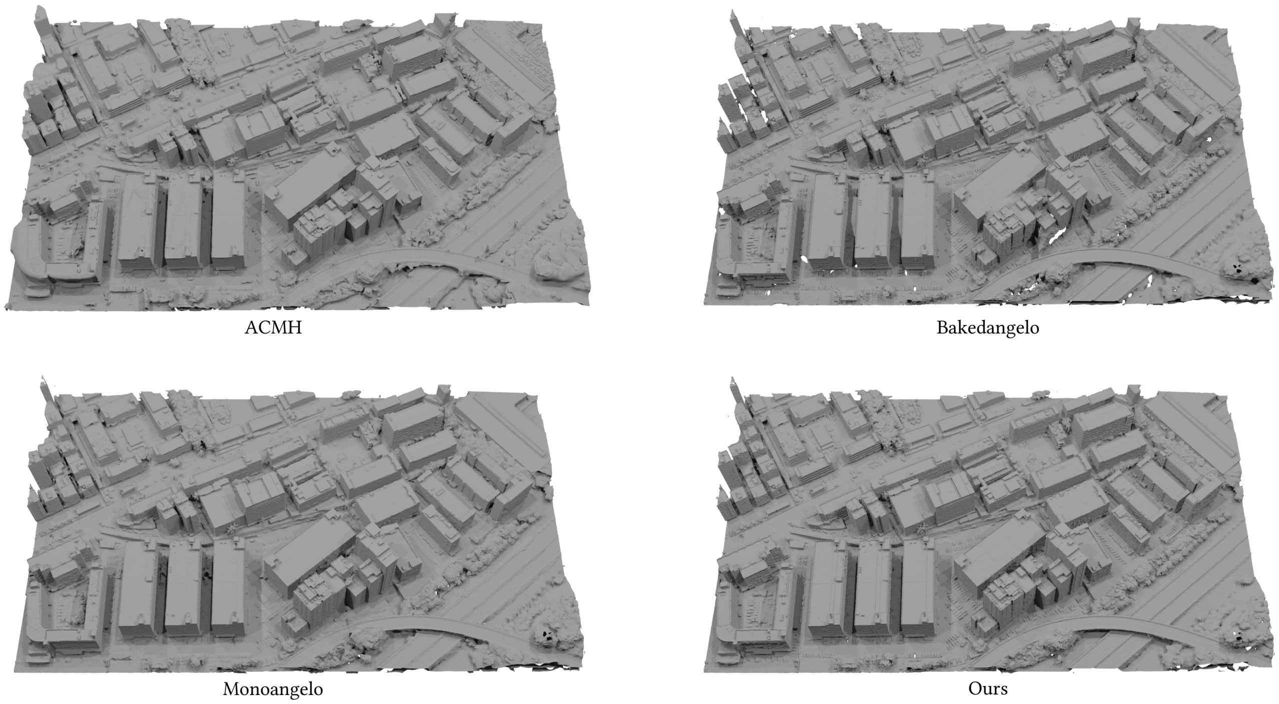


Figure 5: Reconstruction results of Geoangelo on UrbanScene3D dataset. Sometimes NCC can help overcome ambiguity, while other times it may fail to reconstruct.



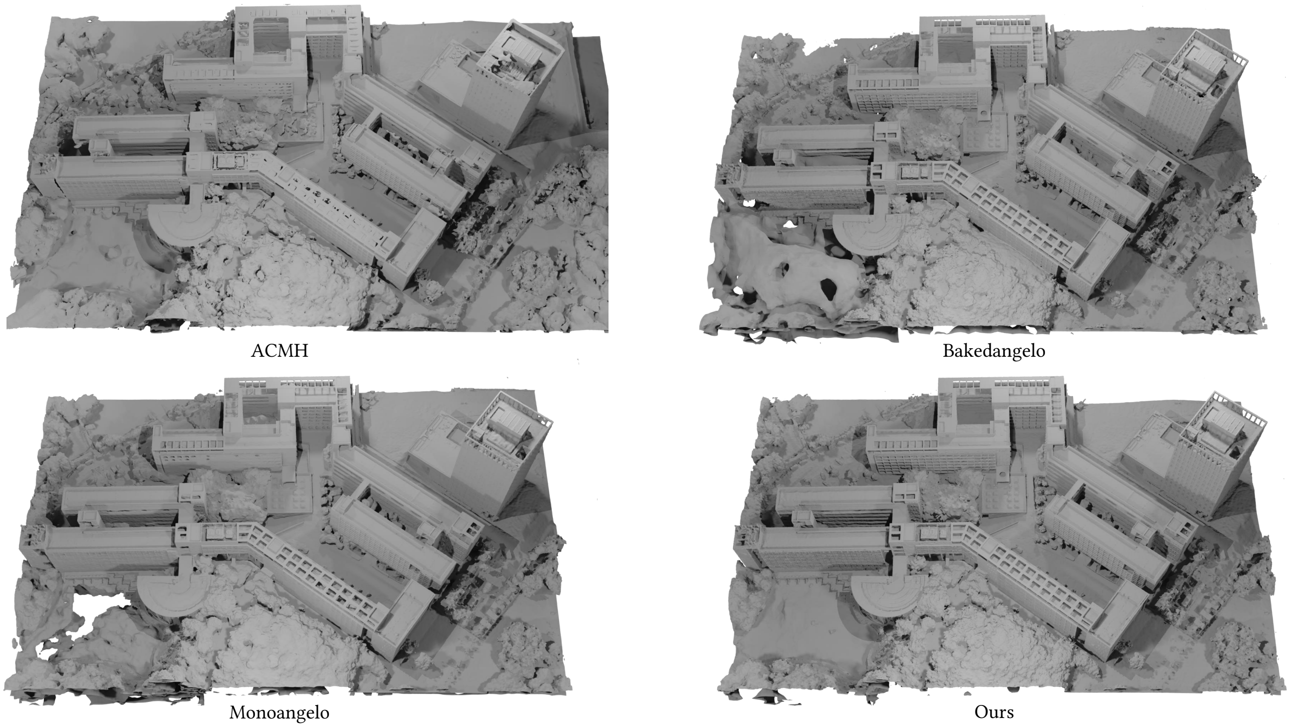
ACMH

Bakedangelo

Monoangelo

Ours

Figure 6: The reconstruction results of the entire Residence. Zoom in for observation.



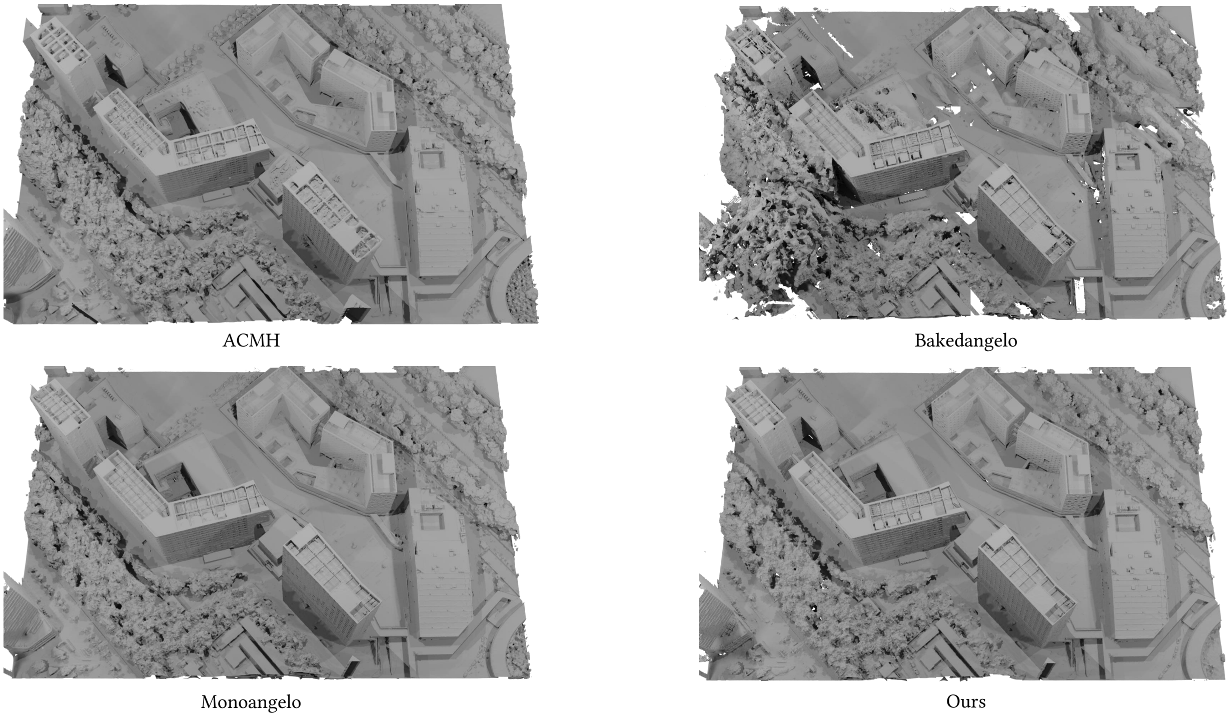
ACMH

Bakedangelo

Monoangelo

Ours

Figure 7: The reconstruction results of the entire Artsci. Zoom in for observation.

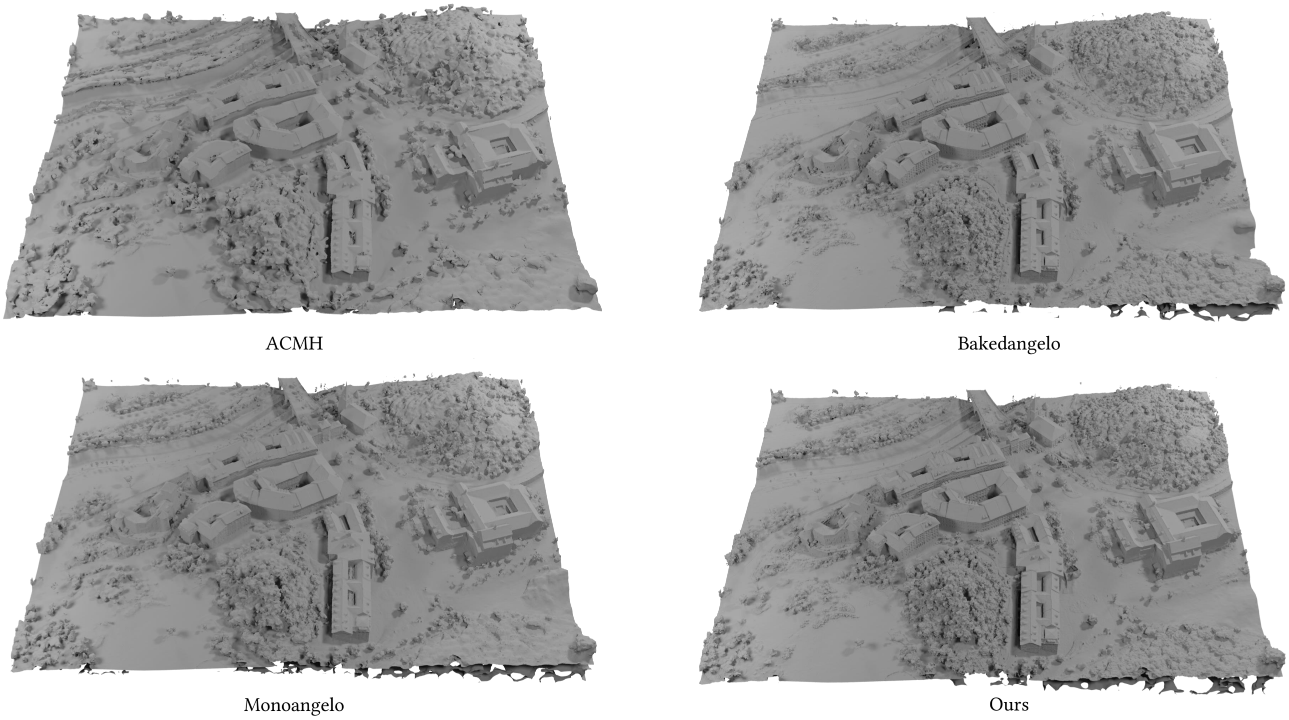

ACMH

Bakedangelo

Monoangelo

Ours

Figure 8: The reconstruction results of the entire Polytech. Zoom in for observation.

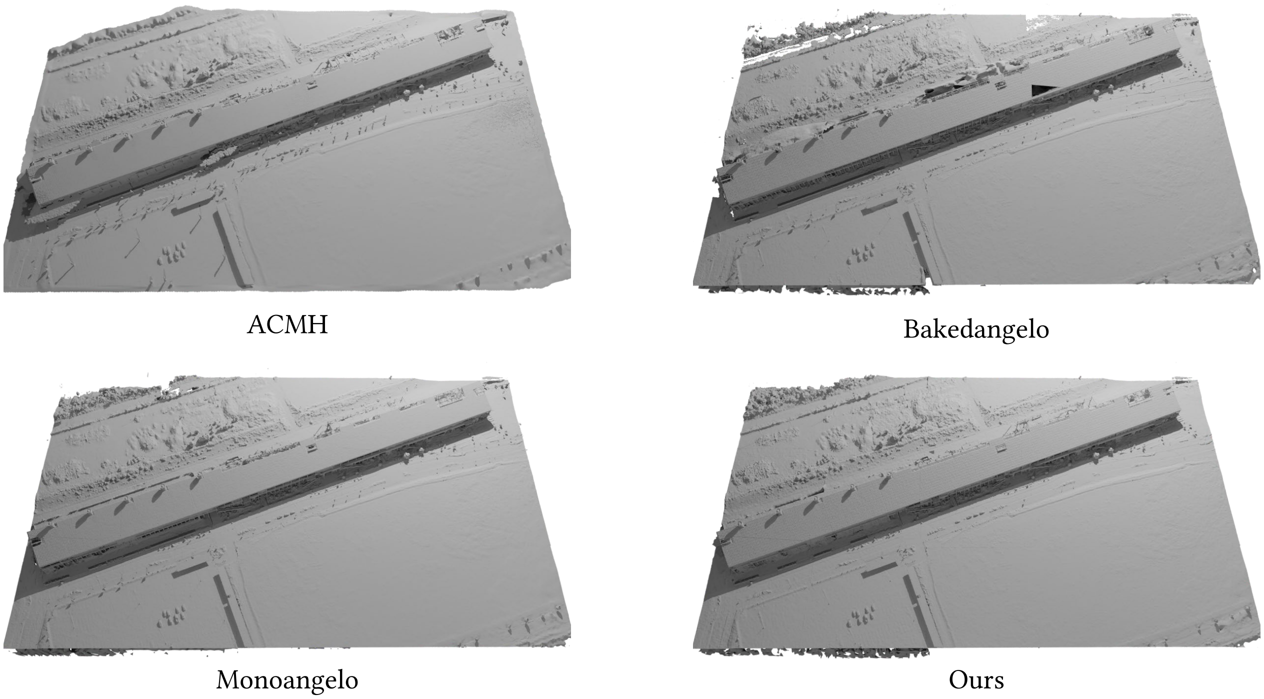

ACMH

Bakedangelo

Monoangelo

Ours

Figure 9: The reconstruction results of the entire Songshanhu. Zoom in for observation.



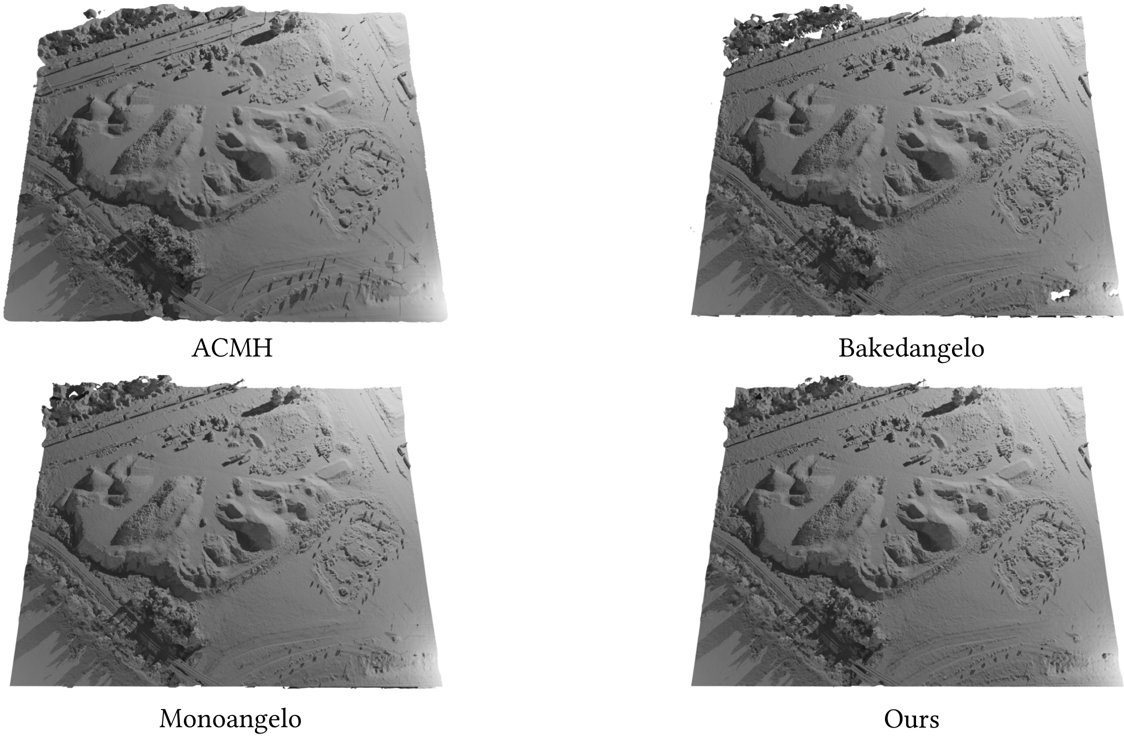
ACMH

Bakedangelo

Monoangelo

Ours

Figure 10: The reconstruction results of the entire Building. Zoom in for observation.



ACMH

Bakedangelo

Monoangelo

Ours

Figure 11: The reconstruction results of the entire Rubble. Zoom in for observation.

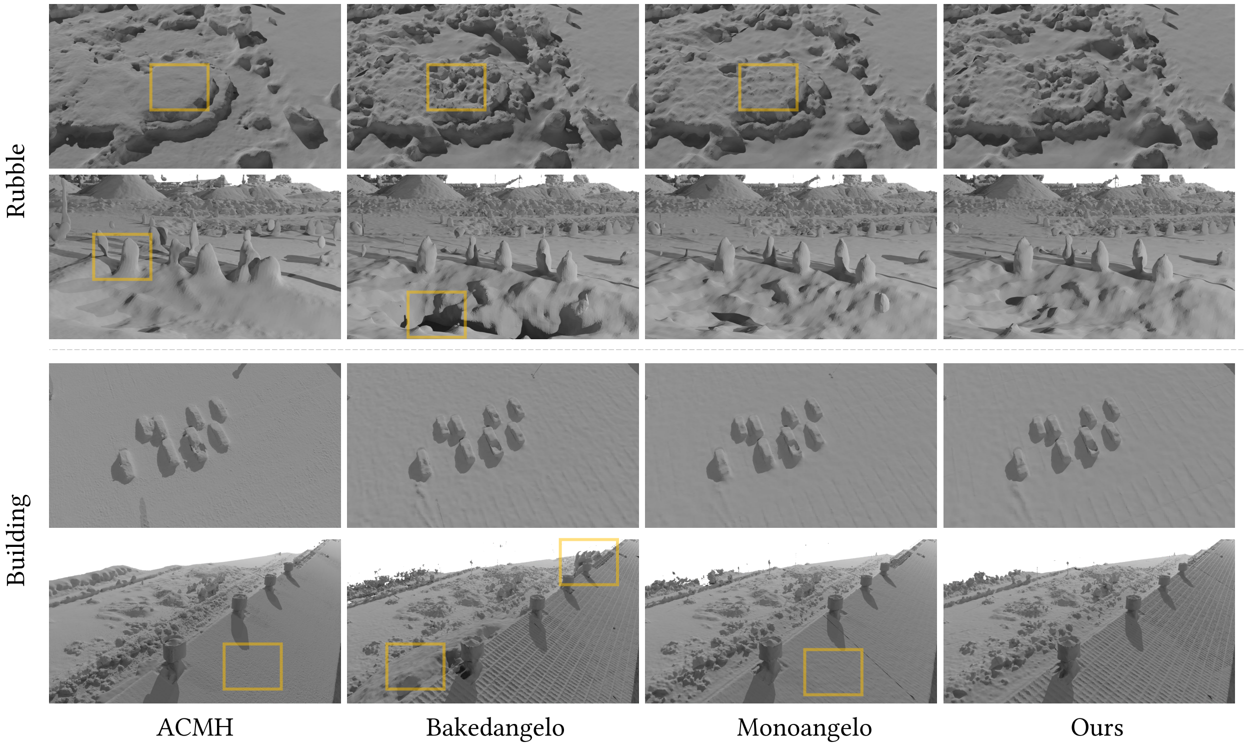


Figure 12: More qualitative results on Mill19 dataset. Zoom in for observation.

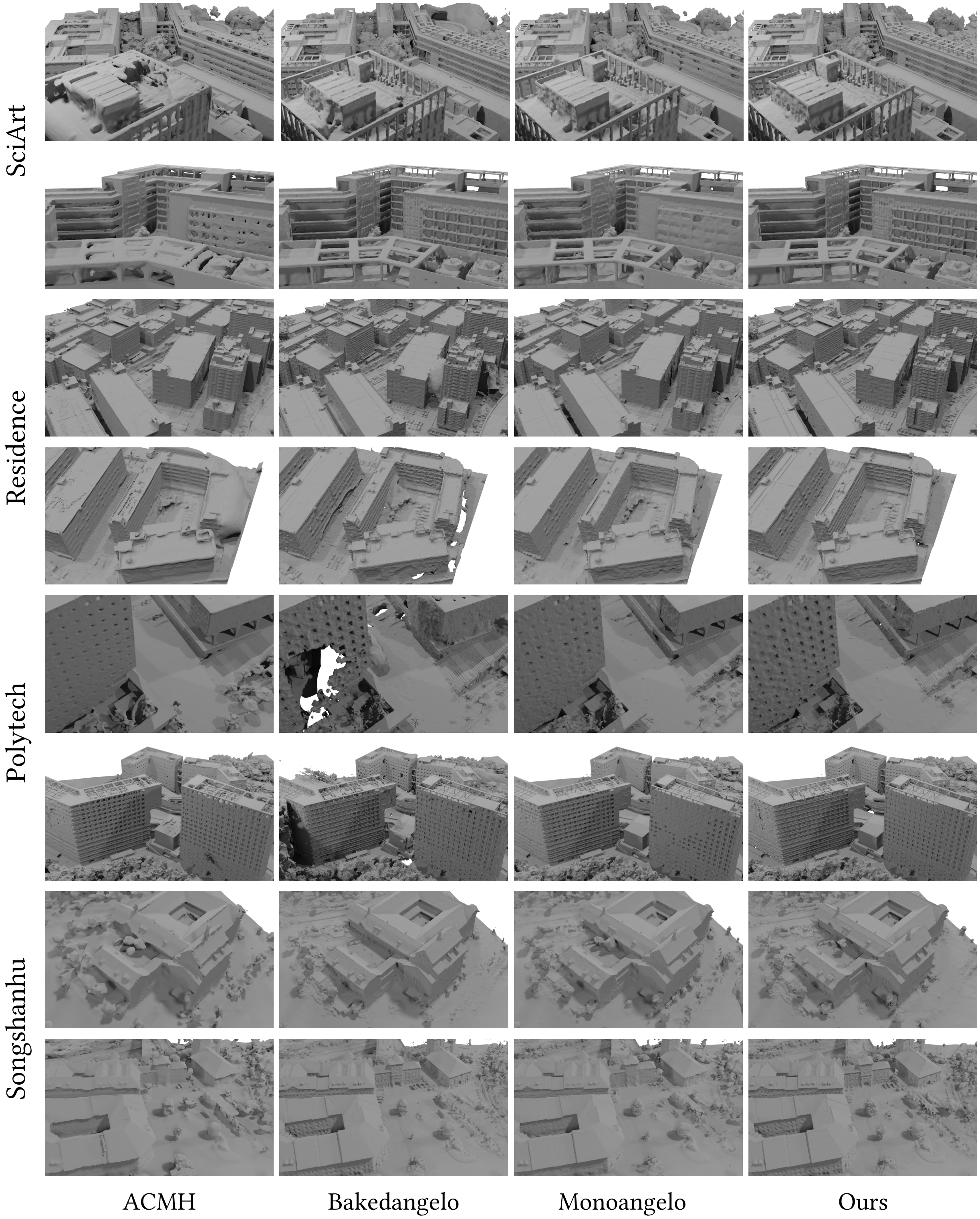


Figure 13: More qualitative results on UrbanScene3D dataset. Zoom in for observation.