IM360: Large-scale Indoor Mapping with 360 Cameras

Supplementary Material

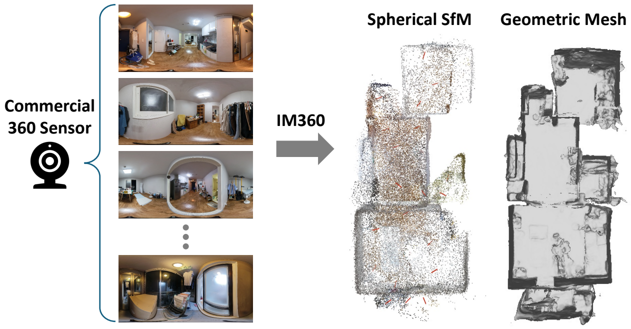

Figure 1. Qualitative results of the custom dataset. We captured 20 images of a house using the commerical 360° sensor, Insta360.

In this supplementary material, we provide additional experimental results, highlighting the superior performance of our method compared to existing approaches. Appendix 1 shows the qualitative results of our custom dataset to demonstrate the practicality of our method. Table 1 present the sensor configurations of publicly available datasets containing indoor scenes. Since Matterport3D [3] and Stanford2D3D [1] offer sparsely scanned panoramic images in large-scale environments, our experiments primarilty focus on these two datasets. Appendix 3.1 highlights the advantages of the spherical camera model and the dense matching algorithm for indoor reconstruction. Appendix 3.2 describes the details and justifies the model choices for geometry mesh reconstruction. Appendix 3.3 demonstrates the effectiveness of our novel texturing method.

1. Real World Application

To demonstrate the versatility of our proposed spherical Structure from Motion, we tested it on a custom dataset. We directly captured 20 omnidirectional images in a 56 m^2 house using an Insta360 camera. Figure 1 presents all registered cameras obtained through spherical SfM along with the 3D reconstruction results.

2. Technical Details

2.1. Spherical Structure from Motion

Spherical Dense Matching: We provide details on spherical dense matching. Dense matchers estabilsh pixel-wise correspondences and sample reliable matches using confidence scores. Ideally, if the network is well-trained, unreliable matches are filtered by confidence, and further refined through geometric filtering during SfM. Furthermore, following the detector-free setting [13, 19], we quantize each 2D match location onto a grid to enhance the consistency of closely spaced subpixel matches. Specifically, each coordinate x is rounded to the nearest grid point using the formula, $\left\lfloor \frac{x}{r} \right\rceil \cdot r$ where r denotes the grid cell size and $\lfloor \cdot \rceil$ represents the rounding operator. This quantization step combines nearby matches into the same grid cell, effectively merging redundant or noisy matches into a single representative location.

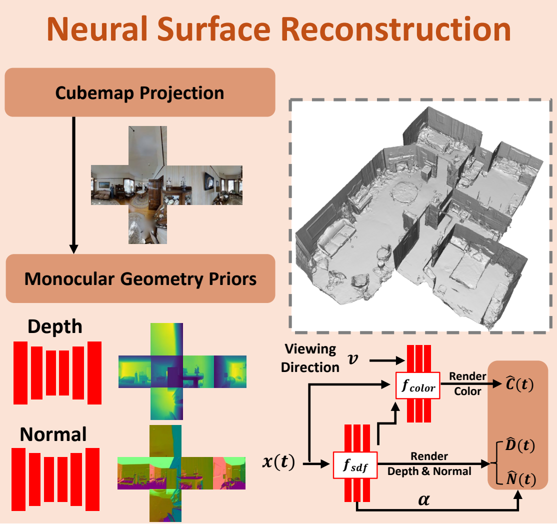

Figure 2. Overview of Neural Surface Reconstruction. In this work, we follow DebSDF [26], which estimates an implicit surface representation by utilizing volumetric rendering and monocular geometric priors. This approach refines the underlying structure and enhances surface details, leading to more accurate 3D reconstructions compared to other recent methods.

Spherical Two-view Geometry Estimation: We provide additional details on spherical two-view geometry estimation in the manuscript. Following Solarte *et al.* [21], we adopt the eight-point algorithm (8-PA) to estimate the essential matrix from corresponding points. Unlike the conventional 8-PA [12], which uses normalized image coordinates, we replace them with unit bearing vectors on a unit sphere via spherical projection. Thus, a minimum of $n \geq 8$ bearing vector correspondences is required to estimate the essential matrix. We reformulate Equation (2) from the manuscript as a least-squares problem:

$$A[E]_v = 0 \qquad (1)$$

Here, A is an $n \times 9$ matrix formed by stacking the Kronecker products of corresponding bearing vectors as $A_i = u_1^i \otimes u_2^i$, and $[E]_v$ is a vector obtained by row-wise concatenation of the essential matrix entries. The solution to Equation 1 is obtained using the Direct Linear Transformation (DLT)

Dataset	Environment	Floor Space (m^2)	# Perspective Images	# Panoramic Images	Annotation
Stanford2D3D [1]	6 large indoor areas	6020	70,496	1,413	Laser Scanner
Matterport3D [3]	2056 rooms (90 scenes)	46,561	194,400	10,800	Laser Scanner
ScanNet [5]	707 small rooms	34,453	2,492,518	-	RGB-D
7 Scenes [20]	7 small rooms	-	43,000	-	RGB-D
TUM Indoor [14]	building (7 floors)	16,341	48,974	-	Laser Scanner
TUM-LSI [25]	building (5 floors)	5,575	1,095	-	Laser Scanner
InLoc [23]	building (5 floors)	10,370	10328	-	Laser Scanner / RGB-D
Baidu [22]	mall	9,179	2,078	-	Laser Scanner
NAVER LABS [15]	mall and metro	53,036	136,783	-	Laser Scanner

Table 1. Among large-scale indoor datasets, Matterport3D [3] and Stanford2D3D [1] provide 360° panoramic images, unlike other datasets that primarily rely on perspective images. Compared to perspective images, 360° images significantly reduce the number of captures required to cover a scene. However, this sparse and reduced number of images introduces various challenges for visual localization and mapping pipelines. To address these challenges, we integrate spherical SfM, geometric reconstruction, and texture optimization techniques into our approach.

method [12], from which the essential matrix can be recovered.

Image Pair Selection: We do not consider image pair selection a core component of our pipeline and do not incorporate image retrieval methods such as NetVLAD, as we regard this as a separate research topic. We intentionally avoid using automatic annotation to ensure accurate evaluation of other components (Sec3.1), especially since human annotation is relatively easy in sparse view settings.

2.2. Geometric Reconstruction

Following the DebSDF [26], we jointly train two MLPs using the differentiable volumetric rendering, (i) f_{sdf}, which represents the scene geometry as a signed distance function, and (ii) f_{color}, a color network. The training process of [26] incorporates a combination of losses, including color reconstruction loss $L_{\text{rgb}} = \sum_{r\in R} ||\hat{C}(r) - C(r)||_1$, Eikonal loss [11] $L_{\text{eikonal}} = \sum_{x\in\chi}(||\bigtriangledown f_{sdf}(x)||_2 - 1)^2$, and depth and normal losses,

$$\begin{aligned} L_{\text{depth}} &= \sum_{r\in R} ||(w\hat{D}(r) + q) - D(r)||^2, \\ L_{\text{normal}} &= \sum_{r\in R} ||\hat{N}(r) - N(r)||_1 + ||1 - \hat{N}(r)^\top N(r)||_1. \end{aligned} \tag{2}$$

The depth and normal losses are derived from prior geometric cues by comparing the rendered depth $\hat{D}(t)$ and normals $\hat{N}(t)$ with the corresponding prior depth D and normals N from Omnidata [9]. Color image $\hat{C}$ is volumetrically rendered by ray marching $\hat{C} = \sum_{i\in I} \alpha_i C_i T_i$ along with $\hat{D}$ and $\hat{N}$. We then utilize the learned SDF (f_{sdf} evaluated over a uniform grid) to extract a mesh M using the Marching Cubes algorithm [16].

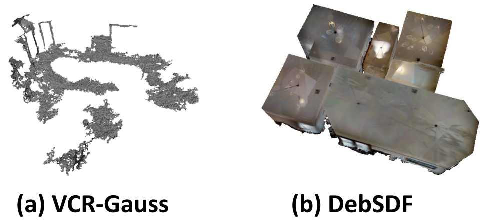
(a) VCR-Gauss **(b) DebSDF**

Figure 3. Examples of VCR-Gauss [4] and DebSDF [26]

3. More Experimental Results

3.1. Spherical Structure from Motion

Due to page limitations, we present our spherical structure from motion results in the supplementary material: 1) **OpenMVG:** An open-source SfM pipeline that supports spherical camera models [17]. 2) **SPSG COLMAP:** SuperPoint [6] and SuperGlue [18] are used with cubemap and equirectangular projection. 3) **DKM COLMAP:** This method leverages DKM [8] to establish dense correspondences, utilizing cubemap and equirectangular projection. 4) **SphereGlue COLMAP:** SuperPoint [6] with a local planar approximation [7] and SphereGlue [10] are utilized to mitigate distortion in ERP images. The experimental results discussed in the main paper for Matterport3D [3] and Stanford2D3D [1] are shown in Fig. 4 and Fig. 5, respectively.

3.2. Geometric Reconstruction

Figure 3 presents the geometry reconstruction results of VCR-Gauss [4] for comparison. VCR-Gauss is a Gaussian Splatting-based surface reconstruction method that utilizes monocular geometry priors, similar to DebSDF [26]. However, in our experiment, we completely fail to train this model, resulting in a polygonal soup.

3.3. Texture Map Optimization

We compare our method with several recent rendering approaches, including **TexRecon** [24], **SparseGS** [27], and **ZipNeRF** [2]. Our method outperforms these approaches by delivering higher frequency details and producing seamless texture maps. The results of the textured mesh and rendering are shown in Fig. 6 and Fig. 7 - 10.

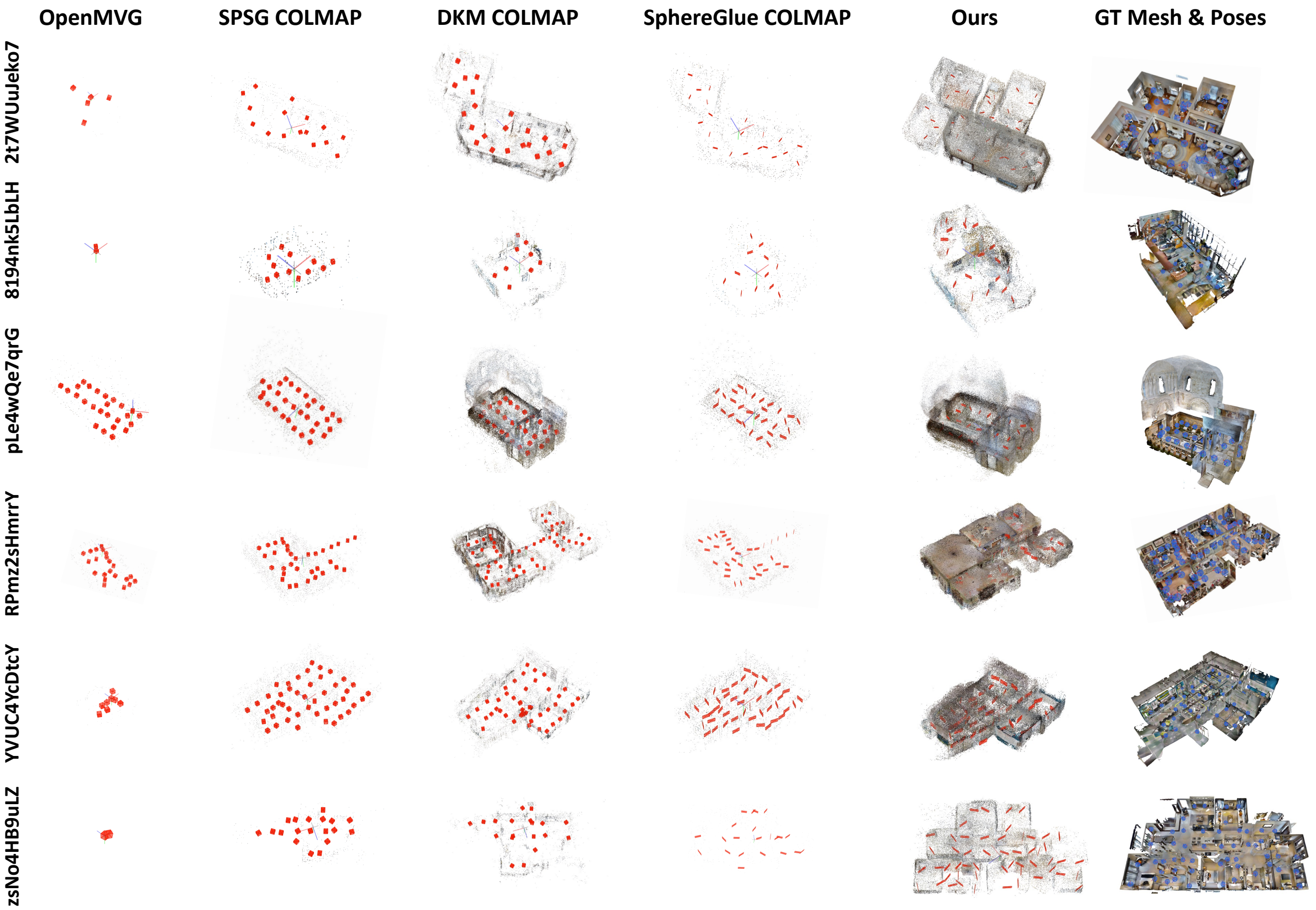

Figure 4. Qualitative Comparison of SfM results on Matterport3D. While other approaches failed to achieve pose registration, our method successfully estimates poses by leveraging the spherical camera model and dense matching.

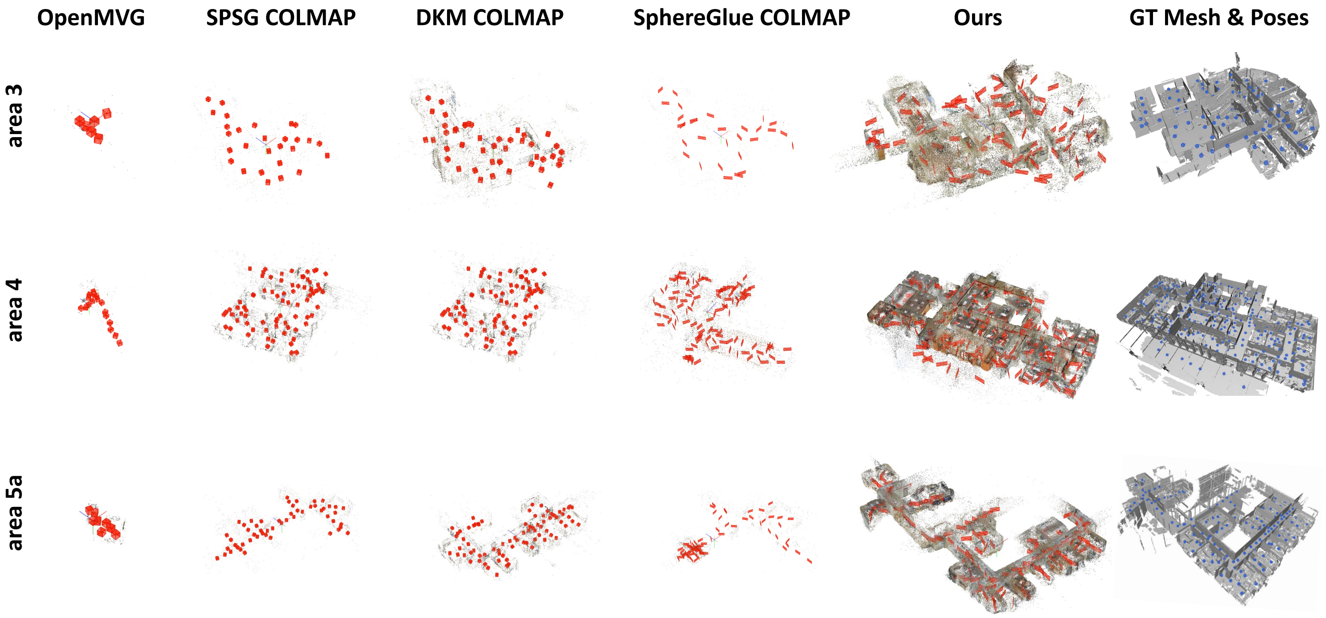

Figure 5. Qualitative Comparison of SfM results on Stanford2D3D. While other approaches failed to achieve pose registration, our method successfully estimates poses by leveraging the spherical camera model and dense matching.

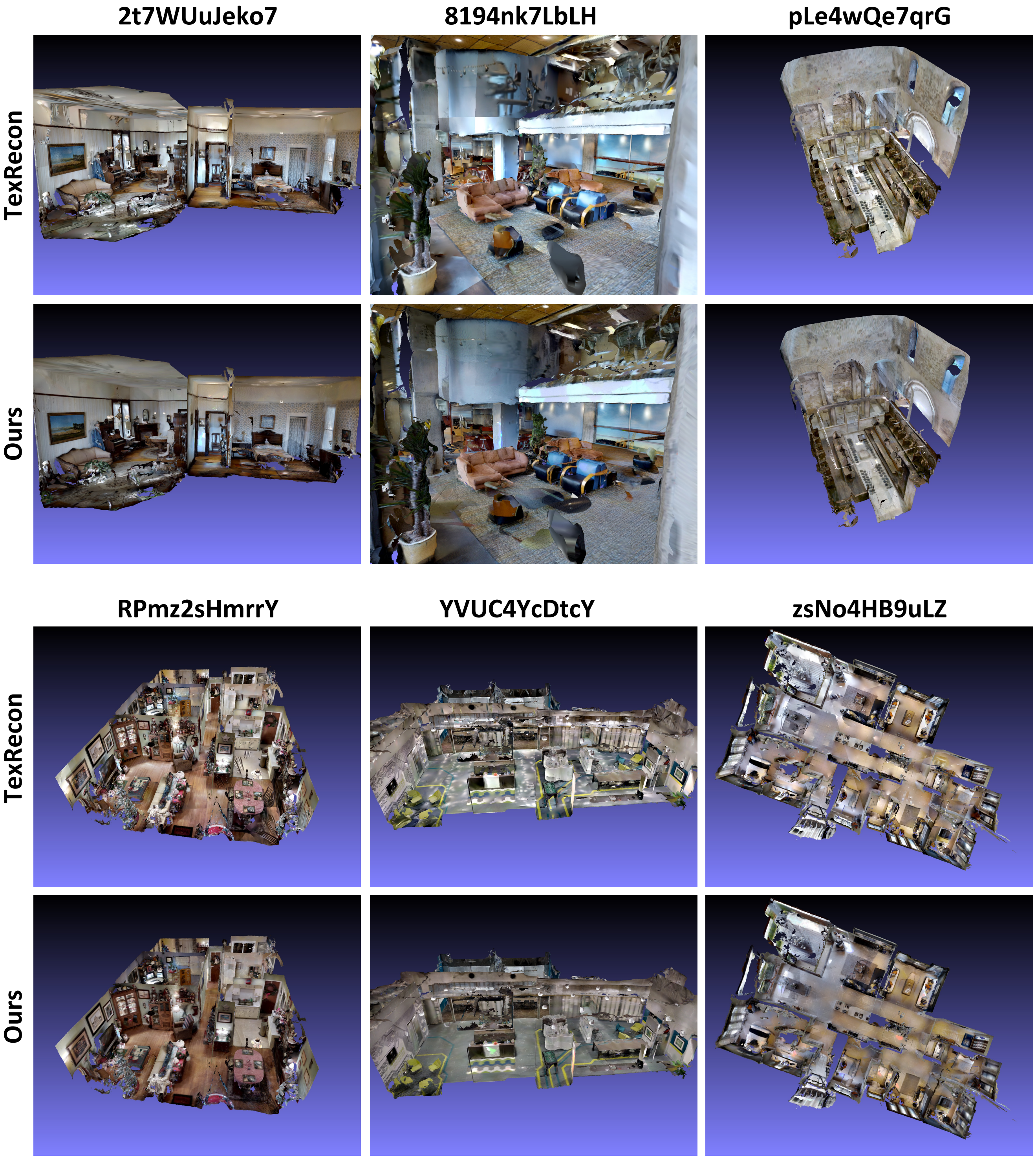

Figure 6. Qualitative Comparisons of Textured Mesh Results on Matterport3D. A comparison between TexRecon [24] and ours shows that our method effectively reduces noise in the texture maps, leading to improved visual quality and detail.

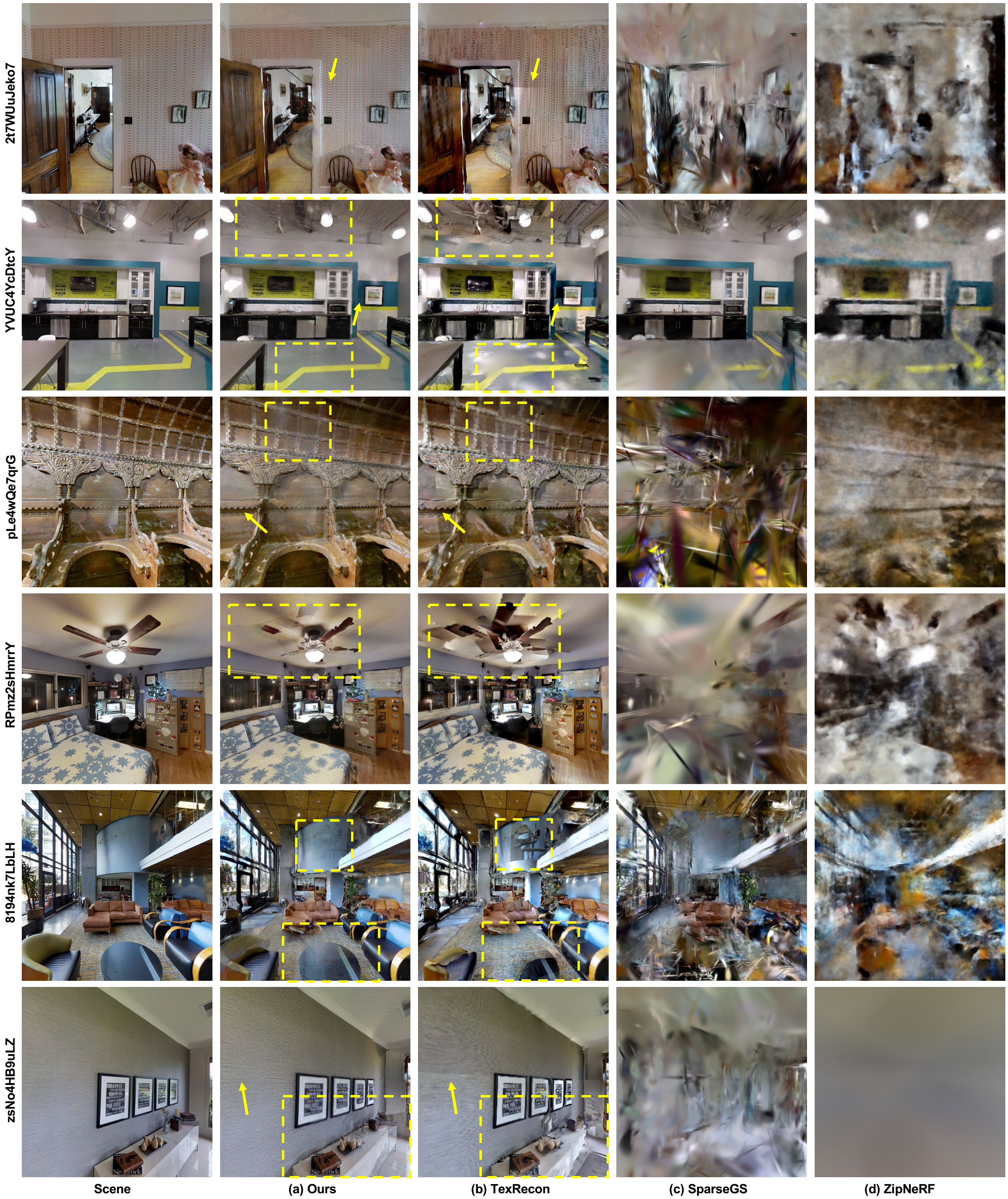

Figure 7. Qualitative Comparisons with Existing Methods. Our method can render high frequency details and results in lower noise.

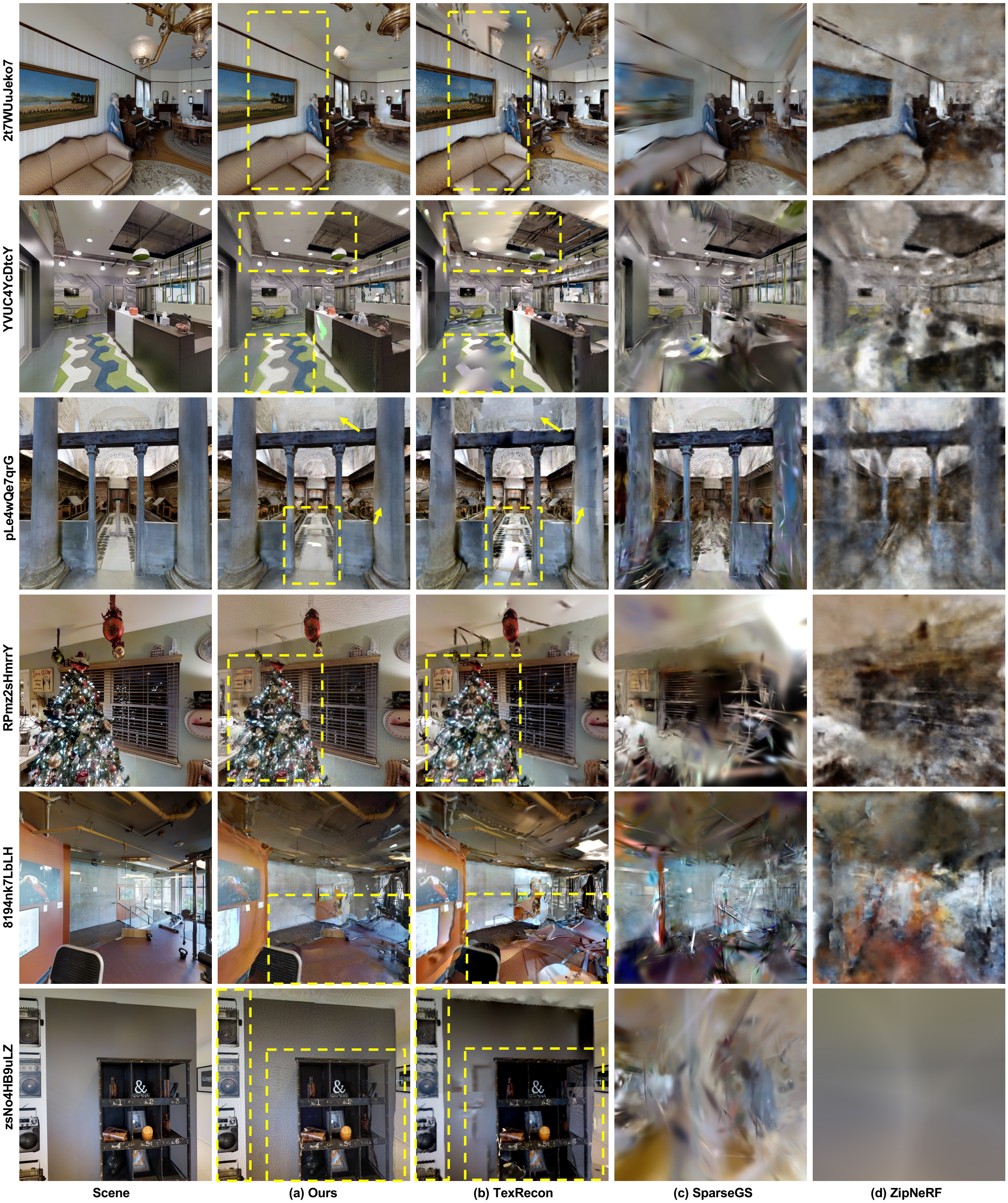

Figure 8. Qualitative Comparisons with Existing Methods. Our method can render high frequency details and results in lower noise.

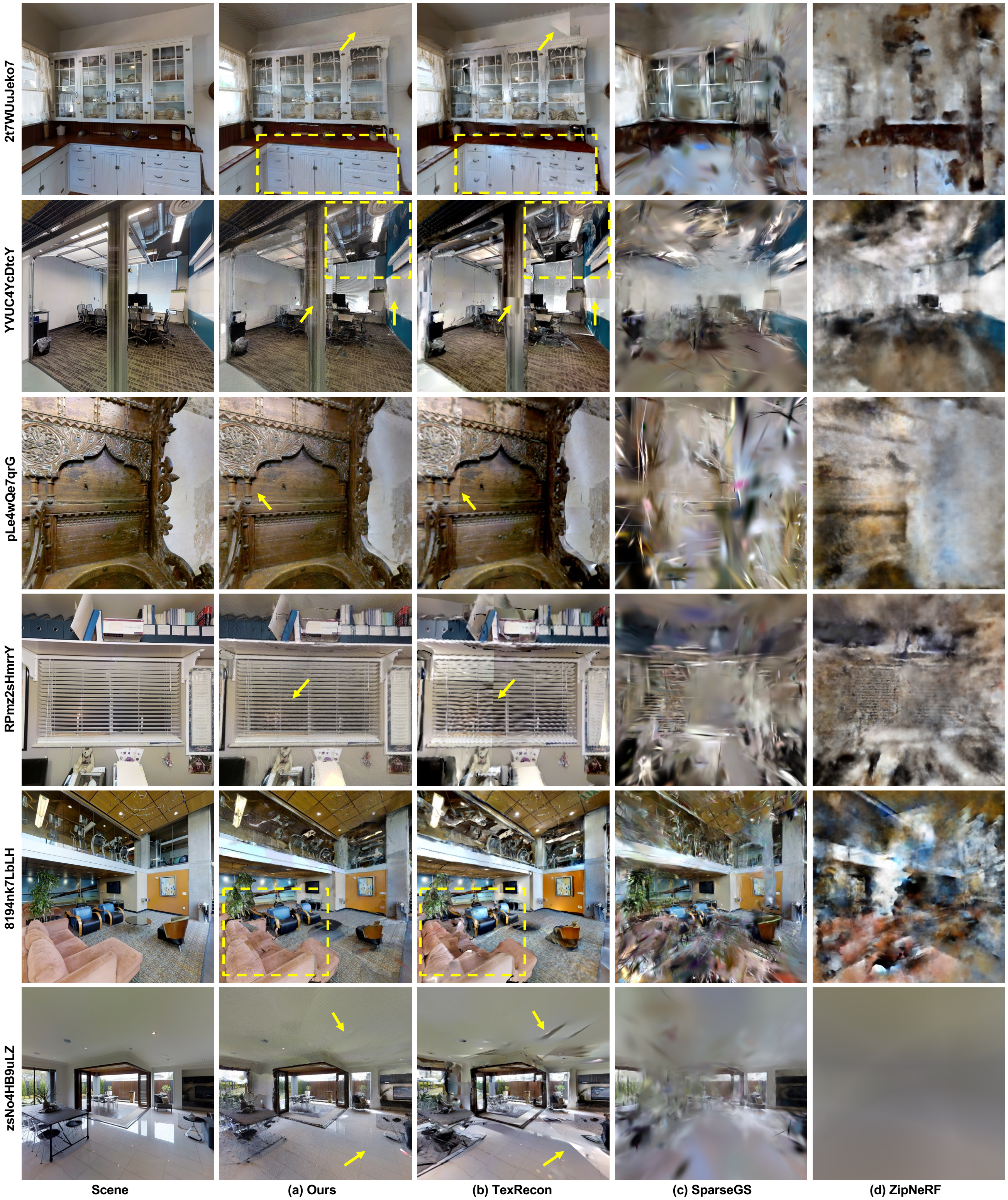

Figure 9. Qualitative Comparisons with Existing Methods. Our method can render high frequency details and results in lower noise.

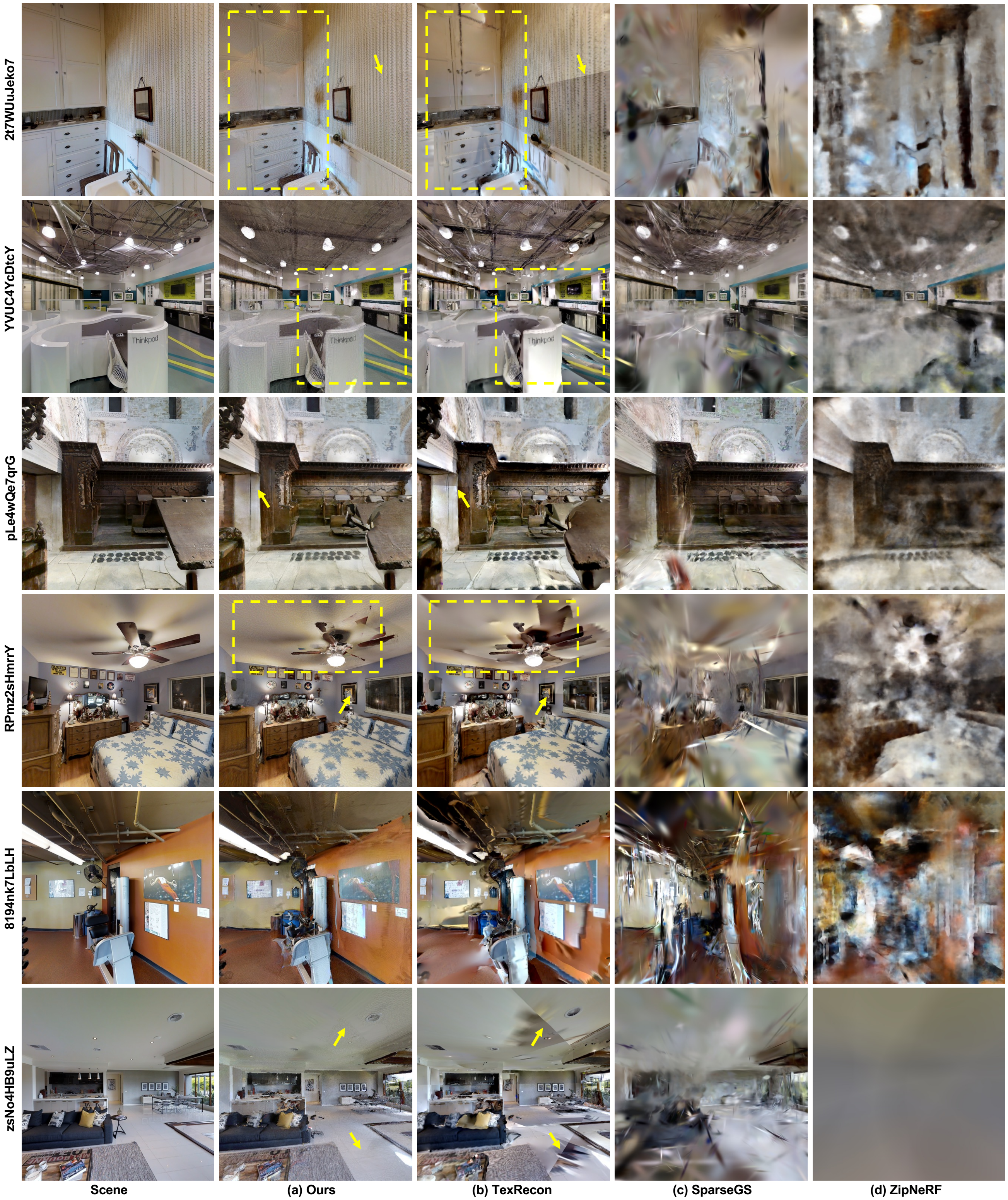

Figure 10. Qualitative Comparisons with Existing Methods. Our method can render high frequency details and results in lower noise.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1, 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 3

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2

[4] Hanlin Chen, Fangyin Wei, Chen Li, Tianxin Huang, Yunsong Wang, and Gim Hee Lee. Vcr-gaus: View consistent depth-normal regularizer for gaussian surface reconstruction. *Advances in Neural Information Processing Systems*, 37: 139725–139750, 2025. 2

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[7] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2020. 2

[8] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2

[9] 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. 2

[10] Christiano Gava, Vishal Mukunda, Tewodros Habtegebrial, Federico Raue, Sebastian Palacio, and Andreas Dengel. Sphereglue: Learning keypoint matching on high resolution spherical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6134–6144, 2023. 2

[11] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 2

[12] Richard Hartley. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2

[13] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21594–21603, 2024. 1

[14] Robert Huitl, Georg Schroth, Sebastian Hilsenbeck, Florian Schweiger, and Eckehard Steinbach. Tumindoor: An extensive image and point cloud dataset for visual indoor localization and mapping. In *2012 19th IEEE International Conference on Image Processing*, pages 1773–1776, 2012. 2

[15] Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guérin, Gabriela Csurka, et al. Large-scale localization datasets in crowded indoor spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3236, 2021. 2

[16] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2

[17] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *Reproducible Research in Pattern Recognition: First International Workshop, RRPR 2016, Cancún, Mexico, December 4, 2016, Revised Selected Papers 1*, pages 60–74. Springer, 2017. 2

[18] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[19] Zehong Shen, Jiaming Sun, Yuang Wang, Xingyi He, Hujun Bao, and Xiaowei Zhou. Semi-dense feature matching with transformers and its applications in multiple-view geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7726–7738, 2022. 1

[20] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013. 2

[21] Bolivar Solarte, Chin-Hsuan Wu, Kuan-Wei Lu, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Robust 360-8pa: Redesigning the normalized 8-point algorithm for 360-fov images. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11032–11038. IEEE, 2021. 1

[22] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7436–7444, 2017. 2

[23] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 2

[24] Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions.

In *European conference on computer vision*, pages 836–850. Springer, 2014. 3, 6

[25] Florian Walch, Caner Hazirbas, and Laura Leal-Taixe. Imagebased localization using lstms for structured feature correlation. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2

[26] Yuting Xiao, Jingwei Xu, Zehao Yu, and Shenghua Gao. Debsdf: Delving into the details and bias of neural indoor scene reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[27] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Realtime 360° sparse view synthesis using gaussian splatting. *Arxiv*, 2023. 3