LongSplat: Robust Unposed 3D Gaussian Splatting for Casual Long Videos

Supplementary Material

A. Implementation Details

We implement LongSplat using PyTorch. Our rendering and 3D Gaussian updates are accelerated using CUDA and cuDNN. Camera pose optimization is performed using differentiable rendering, while the PnP initialization leverages OpenCV's solver with RANSAC. All experiments run on NVIDIA 4090 GPUs.

A.1. LongSplat Algorithm: Pseudo-Code

The LongSplat pipeline incrementally reconstructs a scene from a casually captured long video, without known poses, by tightly coupling pose estimation and 3D Gaussian Splatting. The workflow can be summarized in the following pseudo-code:

Algorithm 1 LONGSPLAT: Incremental 3DGS

Input: RGB frames $\{I_t\}_{t=1}^{T}$
Output: 3DGS $\mathcal{G}$, camera poses $\{P_t\}_{t=1}^{T}$
/* **Initialization** */
$(D_t, C_t, P_t) \leftarrow \texttt{MASt3R Global Alignment}(I_{1\ldots N_{\text{init}}})$
$\texttt{OctreeAnchorFormation}(\mathcal{G}, D_t, P_t)$
/* **Incremental Joint Optimization** */
for $t \leftarrow N_{\text{init}}$ **to** T **do**
  $\texttt{GlobalOptimize}(\mathcal{G}, \{P_{1..t-1}\}, K_g)$
  $(D_t, C_t) \leftarrow \texttt{MASt3R}(I_t)$
  $P_t \leftarrow \texttt{PnP\_RANSAC}(C_t, \mathcal{G})$
  if $P_t = \textit{FAIL}$ **then**
    fallback to t
  end
  $\texttt{PoseRefine}(\mathcal{G}, P_t, I_t)$
  $\texttt{AnchorUnprojection}(\mathcal{G}, D_t, P_t)$
  $\mathcal{W} \leftarrow \texttt{VisibilityWindow}(t)$
  $\texttt{LocalOptimize}(\mathcal{G}, \{P_k\}_{k\in\mathcal{W}}, K_\ell)$
end
/* **Final Global Refinement** */
$\texttt{GlobalRefinement}(\mathcal{G}, \{P_{1..T}\}, K_r)$
return $(\mathcal{G}, \{P_t\}_{t=1}^{T})$

B. Additional Experiments

B.1. CO3Dv2 Benchmark Evaluation.

We report the results on CO3Dv2 [8] in Fig. 1 and Table 1. LongSplat surpasses CF-3DGS and HT-3DGS in all image and pose metrics, confirming the method's robustness on this more challenging benchmark.

Table 1. **Qualtitative comparison on the CO3Dv2 dataset [8]**

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	ATE↓	RPE_t↓	RPE_r↓
CO3Dv2	CF-3DGS	26.61	0.79	0.29	0.014	0.218	0.374
	HT-3DGS	28.34	0.84	0.30	0.017	0.058	0.314
	Ours	**32.59**	**0.91**	**0.17**	**0.005**	**0.023**	**0.096**

Figure 1. **Qualitative comparison on the CO3Dv2 dataset [8]**

B.2. Comparison between COLMAP and LongSplat on the Hike Dataset

We compare LongSplat with a standard COLMAP-based reconstruction pipeline on our Hike dataset. This dataset poses extreme challenges for incremental SfM due to vegetation occlusion, textureless surfaces, and long trajectories. The quantitative results in Table 5 show that LongSplat consistently outperforms COLMAP in both rendering quality and pose estimation accuracy. This highlights the advantage of our octree-anchored Gaussian formulation combined with learned 3D priors.

B.3. Pose Accuracy on Hike Dataset.

COLMAP poses are **noisy** on several Hike videos, so we use the 6 stable sequences (forest2, indoor, university1-4) as references to compute pose accuracy in Table 2. LongSplat achieves the lowest errors, beating all baselines.

Table 2. **Pose Accuracy on Hike Dataset.**

Hike dataset	ATE↓	RPE_t↓	RPE_r↓
MASt3R + Scaffold-GS	0.006	0.009	0.292
MASt3R + Scaffold-GS*	0.006	0.009	0.221
LocalRF	0.004	0.011	0.211
Ours	**0.002**	**0.003**	**0.128**

B.4. Comparison between HT-3DGS and LongSplat

We report the comparison with HT-3DGS in Table 3 and Fig. 2. HT-3DGS runs only on T&T (33.53 dB), but falls to 13.75 dB on Free and runs OOM on Hike. LongSplat remains stable across all datasets. This confirms our SOTA claim for long, casually captured videos.

Table 3. **Qualitative comparison with HT-3DGS.**

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	ATE↓	RPE_t↓	RPE_r↓	Success Rate
Tanks & Temples	HT-3DGS	**33.53**	**0.96**	**0.07**	**0.00**	0.04	**0.07**	8/8
	Ours	32.83	0.94	0.08	**0.00**	**0.03**	**0.07**	8/8
Free	HT-3DGS	13.75	0.39	0.65	0.02	0.34	4.41	6/7
	Ours	**27.88**	**0.85**	**0.17**	**0.00**	**0.03**	**0.10**	7/7
Hike	HT-3DGS	OOM	OOM	OOM	OOM	OOM	OOM	0/12
	Ours	**25.39**	**0.81**	**0.19**	**0.00**	**0.01**	**0.21**	12/12

Figure 2. **Qualitative comparison with HT-3DGS**

B.5. Ablation on Using MASt3R Relative Poses

To demonstrate the importance of our proposed pose estimation pipeline, we conduct an ablation replacing LongSplat's correspondence-guided PnP with directly using MASt3R's relative pose estimates. As shown in Fig. 3, this leads to degraded novel view synthesis quality and larger pose errors, especially in long sequences. This confirms that raw MASt3R poses alone are insufficient for high-quality incremental reconstruction.

B.6. Ablation on training loss

We report the ablation study on training loss in Table 4. Removing individual losses degrades performance. Our full method achieves the best rendering quality and pose accuracy.

Table 4. **Ablation on training loss.**

Method	PSNR↑	SSIM↑	LPIPS↓	RPE_t↓	RPE_r↓	ATE↓
w/o 2d correspondence loss	26.54	0.80	0.24	0.049	0.253	0.007
w/o depth loss	26.74	0.82	0.22	0.076	0.246	0.011
Ours	**27.88**	**0.85**	**0.17**	**0.028**	**0.103**	**0.004**

C. Complete Quantitative Evaluation

C.1. Tanks and Temples

We provide full quantitative results on the Tanks and Temples benchmark in Tabs. 6 and 7. LongSplat consistently outperforms baselines in both rendering quality and pose estimation accuracy, demonstrating its effectiveness even in indoor and urban scenes with varied scales and complexities.

C.2. Free dataset

We provide full quantitative results on the Free dataset benchmark in Tab. 8. LongSplat consistently outperforms baselines in both rendering quality and pose estimation accuracy, demonstrating its effectiveness even in indoor and urban scenes with varied scales and complexities.

C.3. Hike dataset

Hike dataset benchmark in Tab. 5. LongSplat consistently outperforms baselines in both rendering quality and pose estimation accuracy, demonstrating its effectiveness even in challenging indoor and urban scenes with varied scales and complexities. Notably, in scenarios where COLMAP fails to reconstruct due to long trajectories or low-texture regions, LongSplat maintains high-quality results, preserving structural details and ensuring stable pose estimation.

D. Additional Visual Comparisons

D.1. Visual Comparison on Ablation Study

Fig. 4 shows the visual impact of removing key training components. Both trajectory estimation and novel view synthesis degrade severely when global optimization, local optimization, or final refinement is removed, emphasizing their importance.

D.2. Additional Trajectory Results

We include additional visualizations of camera trajectories estimated by LongSplat. As shown in Fig. 5, our method reconstructs stable, drift-free trajectories even in long and complex sequences.

D.3. Additional Tanks and Temples Results

We provide additional qualitative comparisons on the Tanks and Temples benchmark. LongSplat produces sharper and more visually consistent results across diverse scenes, demonstrating strong generalization across both indoor and outdoor environments.

D.4. Additional Free Dataset Results

Additional qualitative comparisons on the Free dataset are shown in Fig. 7. Our method preserves more fine details, produces fewer artifacts, and achieves sharper novel view synthesis than all baselines.

D.5. Additional Hike Dataset Results

Finally, we present more qualitative results on the Hike dataset in Fig. 8, Fig. 9. LongSplat reconstructs complex natural scenes with higher visual quality, capturing vegetation, terrain, and large-scale geometry with remarkable accuracy.

References

[1] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *CVPR*, 2023. 3, 4, 6

[2] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *CVPR*, 2024. 3, 4, 5, 6

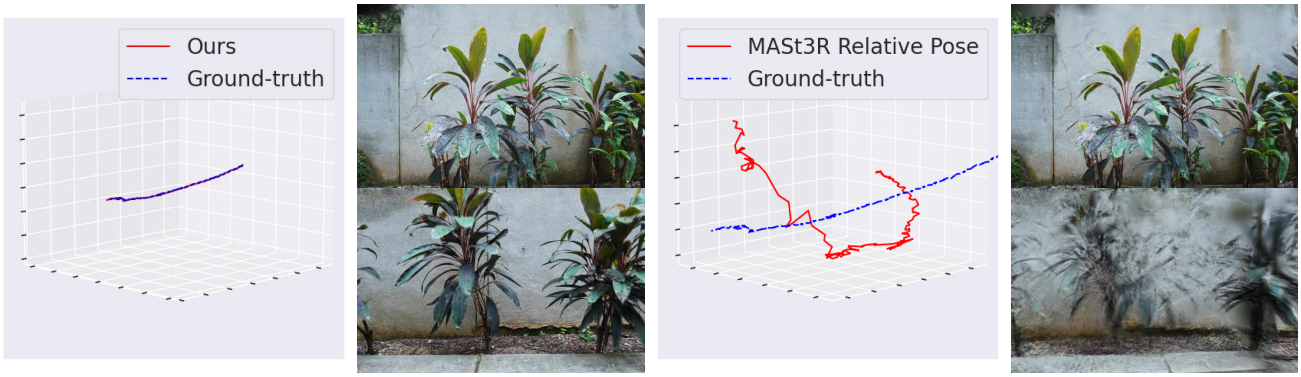

Figure 3. **Visual comparisons on ablation MASt3R relative pose.**

Table 5. **Quantitative evaluation on the Hike dataset [7].** Our method consistently outperforms baselines across diverse scenes with complex trajectories and extended sequences, highlighting LongSplat's robustness and superior scene representation capability. CF-3DGS [2] encounters OOM in all scenes and is thus omitted.

Scenes	COLMAP + Scaffold-GS [6]			MASt3R [5] + Scaffold-GS [6]			MASt3R [5] + Scaffold-GS [6]			LocalRF [7]			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
forest1	20.12	0.55	0.44	17.68	0.30	0.64	17.54	0.34	0.55	19.12	0.45	0.41	23.86	0.79	0.21
forest2	28.35	0.89	0.14	20.91	0.53	0.36	21.11	0.54	0.35	27.23	0.84	0.15	27.87	0.88	0.11
forest3	-	-	-	9.54	0.15	0.70	9.62	0.15	0.70	17.05	0.38	0.59	19.59	0.62	0.31
garden1	20.77	0.67	0.28	13.09	0.23	0.75	14.84	0.27	0.72	22.11	0.66	0.28	24.12	0.80	0.19
garden2	-	-	-	13.21	0.19	0.75	15.67	0.26	0.74	23.34	0.61	0.33	24.35	0.74	0.25
garden3	23.46	0.73	0.23	11.82	0.13	0.64	11.89	0.13	0.64	23.33	0.67	0.27	24.01	0.75	0.23
indoor	28.85	0.90	0.19	23.64	0.81	0.33	24.64	0.83	0.31	30.17	0.91	0.17	30.62	0.92	0.17
playground	-	-	-	19.31	0.49	0.40	19.73	0.52	0.38	22.29	0.63	0.28	24.30	0.78	0.18
university1	25.36	0.78	0.27	19.38	0.47	0.53	19.62	0.48	0.52	25.22	0.71	0.32	25.50	0.79	0.24
university2	27.25	0.87	0.13	20.27	0.58	0.36	20.72	0.60	0.35	24.56	0.75	0.23	26.82	0.85	0.15
university3	26.98	0.89	0.13	18.59	0.51	0.39	19.31	0.57	0.35	23.23	0.73	0.23	25.57	0.86	0.13
university4	25.03	0.82	0.17	20.23	0.61	0.39	20.13	0.61	0.39	25.08	0.79	0.22	27.00	0.88	0.12
Avg	25.13	0.79	0.22	17.30	0.42	0.52	17.90	0.44	0.50	23.56	0.68	0.29	**25.39**	**0.81**	**0.19**

Table 6. **Quantitative evaluation of novel view synthesis quality on the Tanks and Temples dataset [4].** Our proposed LongSplat consistently surpasses existing methods across multiple challenging scenes.

Scenes	COLMAP+3DGS [3]			NoPe-NeRF [1]			CF-3DGS [2]			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Church	29.93	0.93	0.09	25.17	0.73	0.39	30.23	0.93	0.11	30.96	0.93	0.10
Barn	31.08	0.95	0.07	26.35	0.69	0.44	31.23	0.90	0.10	32.57	0.92	0.09
Museum	34.47	0.96	0.05	26.77	0.76	0.35	29.91	0.91	0.11	33.78	0.95	0.06
Family	27.93	0.92	0.11	26.01	0.74	0.41	31.27	0.94	0.07	33.67	0.96	0.06
Horse	20.91	0.77	0.23	27.64	0.84	0.26	33.94	0.96	0.05	33.42	0.96	0.06
Ballroom	34.48	0.96	0.04	25.33	0.72	0.38	32.47	0.96	0.07	32.80	0.95	0.06
Francis	32.64	0.92	0.15	29.48	0.80	0.38	32.72	0.91	0.14	33.80	0.92	0.15
Ignatius	30.20	0.93	0.08	23.96	0.61	0.47	28.43	0.90	0.09	31.61	0.94	0.07
Avg.	30.21	0.92	0.10	26.34	0.74	0.39	31.28	0.93	0.09	**32.83**	**0.94**	**0.08**

Table 7. **Quantitative evaluation of camera pose estimation accuracy on the Tanks and Temples dataset [4].** Our method achieves consistently low errors across diverse scenes, outperforming CF-3DGS and NoPe-NeRF, especially in terms of global trajectory accuracy (ATE) and local translation consistency (RPE_t).

Scenes	CF-3DGS			NoPe-NeRF			Ours		
	ATE↓	RPE_r↓	RPE_t↓	ATE↓	RPE_r↓	RPE_t↓	ATE↓	RPE_r↓	RPE_t↓
Church	0.002	0.018	0.008	0.008	0.008	0.034	0.001	0.048	0.011
Barn	0.003	0.034	0.034	0.004	0.032	0.046	0.004	0.061	0.025
Museum	0.005	0.215	0.052	0.020	0.202	0.207	0.001	0.046	0.025
Family	0.002	0.024	0.022	0.001	0.015	0.047	0.002	0.043	0.021
Horse	0.003	0.057	0.112	0.003	0.017	0.179	0.001	0.046	0.086
Ballroom	0.003	0.024	0.037	0.002	0.018	0.041	0.002	0.053	0.021
Francis	0.006	0.154	0.029	0.005	0.009	0.057	0.009	0.213	0.036
Ignatius	0.005	0.032	0.033	0.002	0.005	0.026	0.002	0.034	0.032
Avg.	0.004	0.069	0.041	0.006	**0.038**	0.080	**0.003**	0.068	**0.032**

Table 8. **Quantitative evaluation of camera pose estimation accuracy on the Free dataset [9].** "-" indicates methods that encountered out-of-memory issues. Our method consistently achieves superior performance across most scenes, significantly reducing pose errors compared to state-of-the-art approaches. "*": Initialized with MASt3R poses, then jointly optimized.

Scenes	MASt3R [5] + Scaffold-GS [6]			MASt3R [5] + Scaffold-GS [6]*			CF-3DGS [2]			NoPe-NeRF [1]			LocalRF [7]			Ours		
	ATE ↓	RPE_r ↓	RPE_t ↓	ATE ↓	RPE_r ↓	RPE_t ↓	ATE ↓	RPE_r ↓	RPE_t ↓	ATE ↓	RPE_r ↓	RPE_t ↓	ATE ↓	RPE_r ↓	RPE_t ↓	ATE ↓	RPE_r ↓	RPE_t ↓
Grass	0.038	0.554	0.559	0.002	0.152	0.016	-	-	-	0.431	9.333	3.044	0.056	6.026	0.612	0.000	0.058	0.002
Hydrant	0.013	0.168	0.145	0.013	0.165	0.144	-	-	-	0.480	4.068	5.844	0.060	8.487	1.068	0.013	0.111	0.069
Lab	0.009	0.294	0.175	0.009	0.265	0.178	-	-	-	0.533	2.623	5.774	0.041	4.405	1.072	0.004	0.217	0.067
Pillar	0.003	0.225	0.024	0.003	0.199	0.016	0.023	4.744	0.328	0.576	4.176	2.013	0.025	3.553	0.526	0.001	0.066	0.003
Road	0.013	0.153	0.088	0.013	0.159	0.088	-	-	-	0.584	4.087	6.045	0.023	9.798	0.699	0.005	0.080	0.036
Sky	0.010	0.203	0.091	0.010	0.197	0.090	-	-	-	0.807	6.661	9.775	0.031	11.075	0.894	0.002	0.114	0.017
Stair	0.006	0.260	0.050	0.006	0.247	0.050	0.021	2.139	0.140	0.624	2.809	11.120	0.008	6.257	0.563	0.000	0.078	0.001
Avg.	0.013	0.265	0.162	0.008	0.198	0.083	0.019	4.365	0.290	0.576	4.822	6.231	0.035	7.086	0.776	**0.004**	**0.103**	**0.028**

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 3

[4] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen

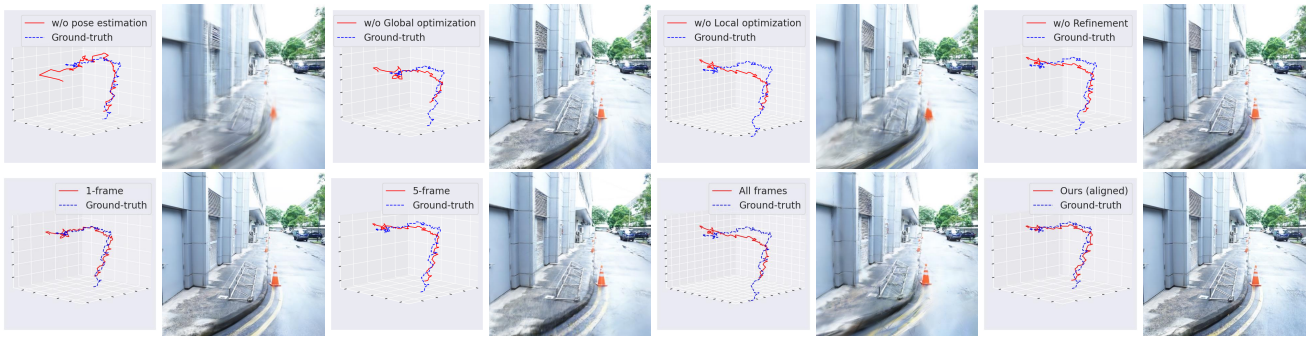

Figure 4. **Visual comparisons on ablation studies.** The top row shows the camera trajectory estimation and novel view synthesis results when different training components are removed, demonstrating the importance of each proposed module. Removing global optimization, local optimization, or final refinement significantly degrades pose accuracy and reconstruction quality. The bottom row evaluates different settings for the visibility-adapted local window size. Too small a window leads to unstable geometry and pose drift, while too large a window dilutes local visibility priors, slowing convergence. LongSplat achieves the best balance using the proposed adaptive window.

Figure 5. **Visualization of camera trajectories on Free dataset [9].** CF-3DGS [2] encounters OOM and fails for long sequences, whereas our method reliably estimates accurate, stable trajectories, demonstrating superior robustness.

Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 2017. 3, 4, 6

[5] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 3, 4, 6, 7, 8

[6] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *CVPR*, 2024. 3, 4, 6, 7, 8

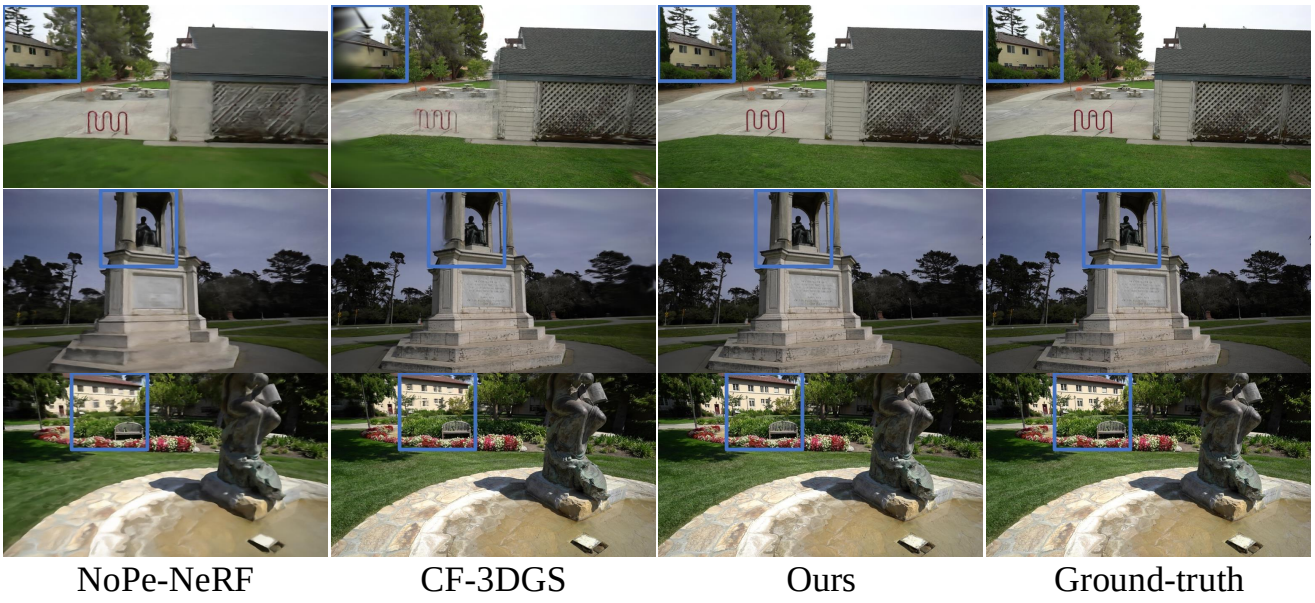

Figure 6. **More Qualitative comparison on the Tanks and Temples dataset [4].** NoPe-NeRF [1] produces visibly blurred results with inaccurate geometries, while CF-3DGS [2], despite better sharpness, fails to reconstruct fine details accurately. In contrast, our LongSplat method achieves superior rendering quality, closely matching the ground truth with sharper textures, more accurate geometry, and consistent lighting.

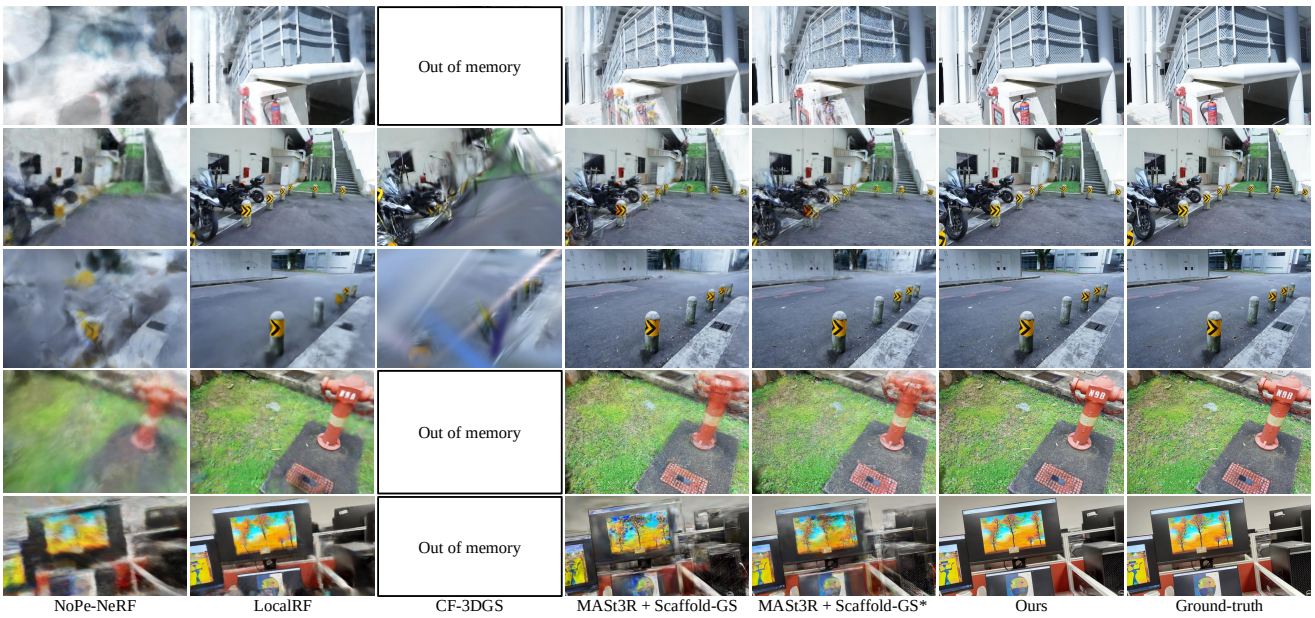

Figure 7. **More Qualitative comparison on the Free dataset [9].** We compare our method with state-of-the-art approaches including NoPe-NeRF [1], LocalRF [7], CF-3DGS [2], and MASt3R [5] combined with Scaffold-GS [6]. CF-3DGS fails due to memory constraints (OOM), and other baseline methods exhibit artifacts or blurry reconstructions. In contrast, our method produces results closest to the ground truth, demonstrating clearer details, accurate geometry, and visually consistent rendering, particularly under challenging scene structures and complex camera trajectories. "*": Initialized with MASt3R poses, then jointly optimized.

[7] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023. 3, 4, 6, 7, 8

[8] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common

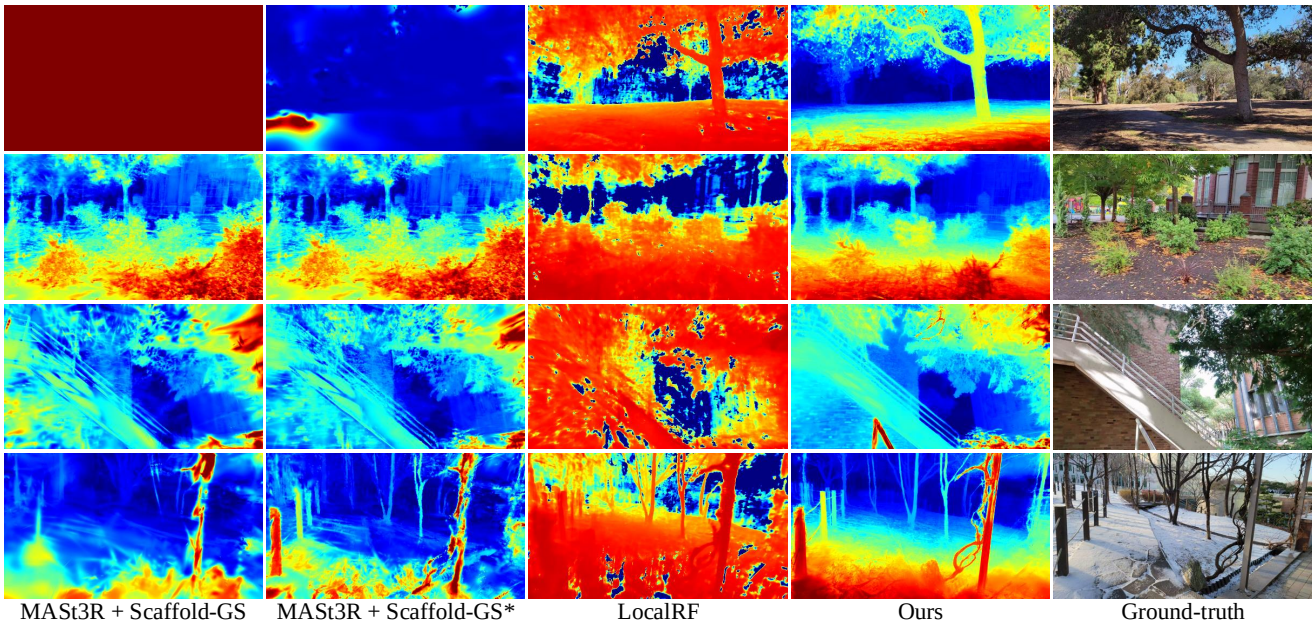

Figure 8. **Qualitative results on the Hike dataset [7].** Compared to existing methods such as LocalRF [7] and MASt3R [5]+Scaffold-GS [6], our approach significantly improves visual clarity and reconstruction fidelity, accurately capturing complex details and textures in challenging scenes captured during long, casual outdoor trajectories. Notably, our method better preserves structural details and reduces artifacts, demonstrating enhanced robustness and visual quality. "*": Initialized with MASt3R poses, then jointly optimized.

objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 1

[9] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *CVPR*, 2023. 4, 5, 6

Figure 9. **More Qualitative results on the Hike dataset [7].** Compared to existing methods such as LocalRF [7] and MASt3R [5]+Scaffold-GS [6], our approach significantly improves visual clarity and reconstruction fidelity, accurately capturing complex details and textures in challenging scenes captured during long, casual outdoor trajectories. Notably, our method better preserves structural details and reduces artifacts, demonstrating enhanced robustness and visual quality. "*": Initialized with MASt3R poses, then jointly optimized.