Pose-Free Omnidirectional Gaussian Splatting for 360-Degree Videos with Consistent Depth Priors

Supplementary Material

A. Overview

In the supplemental materials, we first present an analysis of the algorithm's runtime. We then conduct additional ablation studies on several core modules, examining the contributions of the internal depth prior and the normalized cross-correlation (NCC) similarity. To further clarify the experimental details, we report per-scene quantitative results for the comparative experiments, enabling a more comprehensive evaluation of model behavior across diverse scenarios. We also include error map visualizations for the novel view synthesis results. In addition, we present visualizations of camera pose estimation under different monocular depth priors to analyze their influence on pose accuracy. Finally, to discuss the performance boundaries of our method, we demonstrate its behavior under out-of-distribution monocular depth estimation, failure cases, and limitations.

A.1. Runtime Analysis

In this section, we analyze the runtime of each algorithmic component and compare it with ODGS[6] in Table 1 in the *classroom* scene. For a newly added frame $t+1$, SIFT matching in SCA-PE takes approximately $0.4t$ seconds, while PnP requires nearly a constant 2 seconds. The DIA-Densify stage takes $0.2t$ seconds to extract inliers and related attributes, plus 6 seconds to convert the inlier points into Gaussian seeds. Consequently, for a reasonable number of frames, components with linear complexity in frame length, such as Refinement, dominate the overall runtime.

Table 1. **Runtime analysis on OB3D NonEgocentric. Unit: mins.**

ODGS + SfM	Ours	Init	SIFT	PnP	Densify	Refinment	50 frames	100 frames
95.7	55.2	0.1	2.5	0.9	3.7	48.0	115.0	266.7

†Except for the last two columns, all results are evaluated at 25 frames.

A.2. Ablation Study on Internal Depth Priors and NCC Similarity

Due to space limitations, the main submission discusses only the roles of several key modules and stages without delving into their internal implementation details. In this supplemental material, we first analyze the depth prior used in the spherical consistency-aware pose estimation (SCA-PE) module. Specifically, we replace the internally rendered depth (Int.D) used by SCA-PE with externally predicted monocular depth (Ext.D), which is further aligned to the rendered depth to mitigate the scale mismatch between monocular predictions and the Gaussian model. As shown in the sixth row of Table 2, using an external depth prior leads to noticeable degradation in both camera pose estimation accuracy and novel view synthesis quality compared with the full model (Row 5). This decline stems from depth errors, scale inconsistencies, and the loss of structural detail in monocular depth predictions, since these issues cannot be fully resolved by scale alignment. These results validate the effectiveness of employing internally rendered depth as the depth prior. Furthermore, we observe that despite these residual errors, using an external depth prior in SCA-PE still outperforms the PnP-based pose estimation (Row 2). This demonstrates that the spherical consistency in SCA-PE provides robustness against imperfect depth priors.

Next, we examine the effect of normalized cross-correlation (NCC) similarity filtering in the depth-inlier merging (DIM) module. As shown in the last row of Table 2, when removing patch-level NCC filtering from the DIM module, the over-smoothing and structural detail loss inherent in the monocular depth prior can not be identified using geometric consistency alone. This introduces noise into the depth-inlier set, weakening the Gaussians' ability to faithfully capture the true scene structure, thereby degrading both novel view synthesis quality and camera pose estimation accuracy.

Table 2. **Ablation experiments on model components.**

Base	PnP [9]	SCA-PE	Int.D	Ext.D	DIM	NCC	GOP	PSNR	SSIM	LPIPS	RPE_t	RPE_r	ATE
✓								20.17	0.559	0.413	3.684	5.601	0.1217
✓	✓							27.50	0.815	0.164	0.229	0.156	0.0076
✓		✓	✓					28.99	0.858	0.133	0.077	0.027	0.0021
✓		✓	✓		✓	✓		*29.70*	0.872	*0.122*	0.057	0.025	0.0013
✓		✓	✓		✓	✓	✓	**30.81**	**0.886**	**0.113**	**0.040**	**0.016**	**0.0007**
✓		✓		✓	✓	✓	✓	29.15	0.852	0.140	0.163	0.049	0.0045
✓		✓	✓		✓		✓	29.64	*0.874*	*0.122*	*0.045*	*0.022*	*0.0012*

Table 3. **Quantitative comparison of novel view synthsis and camera pose estimation on OB3D Egocentric dataset.**

Method	Metric	archviz-flat	barbershop	bistro	classroom	emerald-square	fisher-hut	lone-monk	pavillion	restroom	san-miguel	sponza	sun-temple	mean
CF-3DGS	PSNR	30.24	29.64	26.27	28.04	24.50	21.77	19.47	32.10	18.09	21.02	30.86	19.68	25.14
	SSIM	0.893	0.861	0.841	0.871	0.721	0.606	0.621	0.874	0.478	0.664	0.889	0.580	0.742
	LPIPS	0.107	0.224	0.098	0.190	0.154	0.336	0.268	0.182	0.641	0.308	0.113	0.409	0.253
HT-3DGS	PSNR	30.39	30.68	*26.79*	28.82	21.82	21.62	20.44	31.53	19.43	19.97	31.27	21.88	25.39
	SSIM	0.889	0.877	*0.857*	0.877	0.658	0.608	0.662	0.865	0.466	0.576	0.891	0.658	0.740
	LPIPS	0.110	0.206	*0.097*	0.169	0.192	0.322	0.235	0.205	0.594	0.365	0.110	0.300	0.242
3R-GS	PSNR	*32.85*	*31.08*	25.39	*31.70*	*28.87*	*26.27*	*31.99*	*37.84*	*29.10*	*27.45*	*36.62*	*34.60*	*31.15*
	SSIM	*0.920*	*0.887*	0.839	*0.937*	*0.930*	*0.806*	*0.955*	*0.950*	*0.752*	*0.889*	*0.963*	*0.962*	*0.899*
	LPIPS	*0.056*	*0.147*	0.118	*0.059*	0.055	**0.111**	*0.035*	*0.057*	*0.237*	*0.099*	*0.034*	*0.027*	*0.086*
Ours	PSNR	**36.62**	**35.73**	**32.88**	**37.99**	**36.38**	**27.84**	**38.74**	**41.18**	**32.50**	**30.70**	**41.37**	**37.29**	**35.77**
	SSIM	**0.956**	**0.930**	**0.979**	**0.972**	**0.978**	**0.843**	**0.988**	**0.969**	**0.846**	**0.937**	**0.984**	**0.970**	**0.946**
	LPIPS	**0.054**	**0.101**	**0.024**	**0.034**	**0.019**	*0.136*	**0.014**	**0.046**	**0.160**	**0.066**	**0.012**	**0.018**	**0.057**
CF-3DGS	RPE_t	0.278	*0.087*	0.274	0.224	0.773	8.273	2.882	3.688	10.272	2.322	0.442	8.658	3.181
	RPE_r	*0.081*	*0.081*	0.017	0.100	0.033	2.973	0.367	2.311	60.238	1.057	*0.041*	11.255	6.546
	ATE	0.0090	*0.0020*	*0.0080*	0.0060	0.0200	0.1610	0.0550	0.0810	0.1750	0.0540	0.0090	0.1680	0.0623
HT-3DGS	RPE_t	*0.270*	0.088	*0.270*	0.229	0.713	7.365	2.310	4.664	9.454	5.609	0.442	9.086	3.375
	RPE_r	0.082	*0.081*	*0.016*	*0.082*	*0.024*	1.991	0.269	5.126	70.064	5.675	*0.041*	13.047	8.041
	ATE	0.0090	*0.0020*	*0.0080*	0.0060	0.0180	0.1570	0.0480	0.0920	0.1640	0.1150	0.0090	0.1670	0.0663
3R-GS	RPE_t	0.335	0.294	1.214	*0.197*	*0.478*	*0.180*	*0.205*	*0.246*	*0.645*	*0.153*	*0.271*	*0.346*	*0.380*
	RPE_r	0.171	0.149	0.142	0.116	0.147	*0.150*	*0.128*	*0.172*	*0.155*	*0.080*	0.085	*0.068*	*0.130*
	ATE	*0.0023*	0.0028	0.0109	*0.0016*	*0.0036*	*0.0018*	*0.0025*	*0.0028*	*0.0054*	*0.0021*	*0.0026*	*0.0052*	*0.0036*
Ours	RPE_t	**0.018**	**0.018**	**0.019**	**0.018**	**0.020**	**0.019**	**0.018**	**0.020**	**0.017**	**0.016**	**0.017**	**0.017**	**0.018**
	RPE_r	**0.017**	**0.017**	**0.013**	**0.017**	**0.009**	**0.013**	**0.014**	**0.013**	**0.010**	**0.015**	**0.014**	**0.013**	**0.014**
	ATE	**0.0002**	**0.0002**	**0.0003**	**0.0001**	**0.0004**	**0.0003**	**0.0003**	**0.0004**	**0.0001**	**0.0003**	**0.0001**	**0.0001**	**0.0002**

Table 4. **Quantitative comparison of novel view synthsis and camera pose estimation on OB3D NonEgocentric dataset.**

Method	Metric	archviz-flat	barbershop	bistro	classroom	emerald-square	fisher-hut	lone-monk	pavillion	restroom	san-miguel	sponza	sun-temple	mean
CF-3DGS	PSNR	29.31	21.60	24.87	19.66	24.17	*20.59*	18.61	25.63	19.28	19.10	24.08	20.80	22.31
	SSIM	0.919	0.676	0.824	0.693	0.684	*0.601*	0.584	0.748	0.489	0.546	*0.704*	0.645	0.676
	LPIPS	0.139	0.455	0.159	0.403	0.186	*0.342*	0.315	0.317	0.597	0.401	0.316	0.330	0.330
HT-3DGS	PSNR	*30.84*	19.91	27.73	21.02	19.95	20.53	17.96	*26.12*	20.60	19.47	*24.35*	22.23	22.56
	SSIM	*0.932*	0.622	0.894	0.734	0.596	0.598	0.574	*0.753*	0.483	0.565	*0.704*	0.700	0.680
	LPIPS	0.109	0.497	0.101	0.337	0.270	0.350	0.312	*0.305*	0.546	0.374	*0.302*	0.268	0.314
3R-GS	PSNR	30.57	*24.10*	*28.95*	*26.50*	*27.50*	9.07	*26.37*	18.03	*28.29*	*21.94*	8.47	*30.85*	*23.39*
	SSIM	0.927	*0.742*	*0.947*	*0.871*	*0.838*	0.367	*0.887*	0.622	*0.747*	*0.743*	0.245	*0.920*	*0.681*
	LPIPS	*0.077*	*0.246*	*0.047*	*0.128*	*0.072*	0.849	*0.067*	0.484	**0.183**	*0.188*	0.699	*0.051*	*0.238*
Ours	PSNR	**34.95**	**30.79**	**32.28**	**32.15**	**29.33**	**23.27**	**27.90**	**36.86**	**29.62**	**24.82**	**33.32**	**34.47**	**30.81**
	SSIM	**0.965**	**0.880**	**0.966**	**0.938**	**0.879**	**0.678**	**0.903**	**0.942**	**0.795**	**0.804**	**0.929**	**0.961**	**0.887**
	LPIPS	**0.058**	**0.160**	**0.040**	**0.081**	**0.065**	**0.309**	**0.065**	**0.086**	*0.238*	**0.156**	**0.077**	**0.029**	**0.113**
CF-3DGS	RPE_t	0.134	*1.164*	0.172	1.116	0.429	5.599	1.071	3.697	5.116	1.529	*0.495*	1.232	1.813
	RPE_r	0.097	10.572	0.059	1.121	0.061	*0.797*	1.199	2.267	47.672	2.969	0.330	3.237	5.865
	ATE	0.0060	0.0800	0.0070	0.0490	0.0110	*0.1270*	0.0300	0.0700	0.1810	0.0520	*0.0130*	0.0470	0.0561
HT-3DGS	RPE_t	*0.124*	2.211	0.207	1.631	1.872	*5.164*	1.826	3.366	5.629	1.846	0.558	1.953	2.199
	RPE_r	*0.085*	7.351	*0.032*	1.804	0.281	5.990	0.770	2.258	39.243	6.932	*0.095*	5.676	5.876
	ATE	0.0060	0.1290	0.0080	0.0500	0.0450	0.1410	0.0460	*0.0680*	0.1560	0.0820	0.0160	0.0730	0.0683
3R-GS	RPE_t	0.280	1.308	*0.069*	*0.123*	*0.121*	6.523	*0.078*	*3.179*	*0.099*	*0.046*	4.985	*0.126*	*1.303*
	RPE_r	0.241	*0.269*	0.039	*0.111*	*0.056*	8.770	*0.066*	2.351	*0.068*	0.076	7.675	*0.059*	*1.522*
	ATE	*0.0041*	*0.0326*	*0.0008*	*0.0026*	*0.0024*	0.1953	0.0010	0.1291	*0.0032*	*0.0006*	0.1977	*0.0016*	*0.0439*
Ours	RPE_t	**0.009**	**0.010**	**0.009**	**0.010**	**0.025**	**0.322**	**0.042**	**0.012**	**0.012**	**0.008**	**0.011**	**0.009**	**0.040**
	RPE_r	**0.018**	**0.014**	**0.006**	**0.021**	**0.011**	**0.040**	**0.015**	**0.008**	**0.017**	**0.016**	**0.016**	**0.008**	**0.016**
	ATE	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0007**	**0.0048**	**0.0018**	**0.0002**	**0.0002**	**<0.0001**	**0.0001**	**<0.0001**	**0.0007**

Table 5. **Quantitative comparison of novel view synthsis on Ricoh360 dataset.**

Method	Metric	bricks	bridge	bridge_under	cat_tower	center	farm	flower	gallery_chair	gallery_park	gallery_pillar	garden	poster	mean
ODGS	PSNR	*24.62*	*24.37*	25.93	*25.35*	*29.39*	16.34	22.71	27.62	*26.19*	28.74	27.09	26.92	25.44
	SSIM	*0.847*	*0.815*	0.853	*0.808*	*0.894*	0.603	0.748	0.883	*0.840*	0.897	*0.838*	0.880	*0.826*
	LPIPS	**0.102**	**0.106**	0.102	**0.110**	**0.080**	0.410	**0.150**	*0.113*	**0.107**	**0.069**	**0.100**	*0.111*	*0.130*
OmniGS	PSNR	24.66	23.67	*26.60*	24.75	29.16	*22.20*	22.30	*28.79*	-	28.73	27.13	*28.33*	*26.03*
	SSIM	0.836	0.790	*0.873*	0.773	0.887	*0.726*	0.724	*0.892*	-	0.884	0.795	*0.898*	0.825
	LPIPS	0.120	0.136	**0.089**	0.159	0.100	*0.166*	0.191	**0.105**	-	*0.087*	0.155	**0.104**	**0.128**
CF-3DGS	PSNR	21.25	21.54	21.93	23.37	26.71	19.86	20.18	25.30	20.05	27.29	24.63	21.12	22.77
	SSIM	0.709	0.718	0.728	0.717	0.836	0.628	0.614	0.803	0.681	0.844	0.675	0.730	0.724
	LPIPS	0.264	0.276	0.288	0.254	0.259	0.328	0.310	0.255	0.304	0.236	0.298	0.333	0.284
HT-3DGS	PSNR	21.37	20.80	22.20	23.21	26.02	20.37	20.13	24.17	24.26	26.27	25.30	21.95	23.00
	SSIM	0.712	0.695	0.726	0.716	0.825	0.660	0.623	0.781	0.771	0.828	0.704	0.768	0.734
	LPIPS	0.262	0.316	0.276	0.254	0.297	0.324	0.316	0.297	0.250	0.290	0.318	0.323	0.294
3R-GS	PSNR	23.54	24.07	24.75	24.82	23.44	18.87	22.11	27.58	26.05	*28.92*	*27.62*	18.44	24.19
	SSIM	0.806	0.807	0.830	0.750	0.730	0.538	0.675	0.872	0.813	*0.898*	0.813	0.640	0.764
	LPIPS	0.137	0.165	0.137	0.181	0.178	0.251	0.231	0.152	0.175	0.117	0.173	0.241	0.178
Ours	PSNR	**27.33**	**25.98**	**28.15**	**26.51**	**31.31**	**23.85**	**23.91**	**30.15**	**28.15**	**30.82**	**29.60**	**30.90**	**28.05**
	SSIM	**0.897**	**0.865**	**0.900**	**0.814**	**0.920**	**0.788**	**0.768**	**0.912**	**0.846**	**0.912**	**0.861**	**0.920**	**0.867**
	LPIPS	*0.108*	*0.115*	*0.091*	*0.153*	*0.098*	**0.157**	*0.186*	0.143	*0.173*	0.107	0.137	0.139	0.134

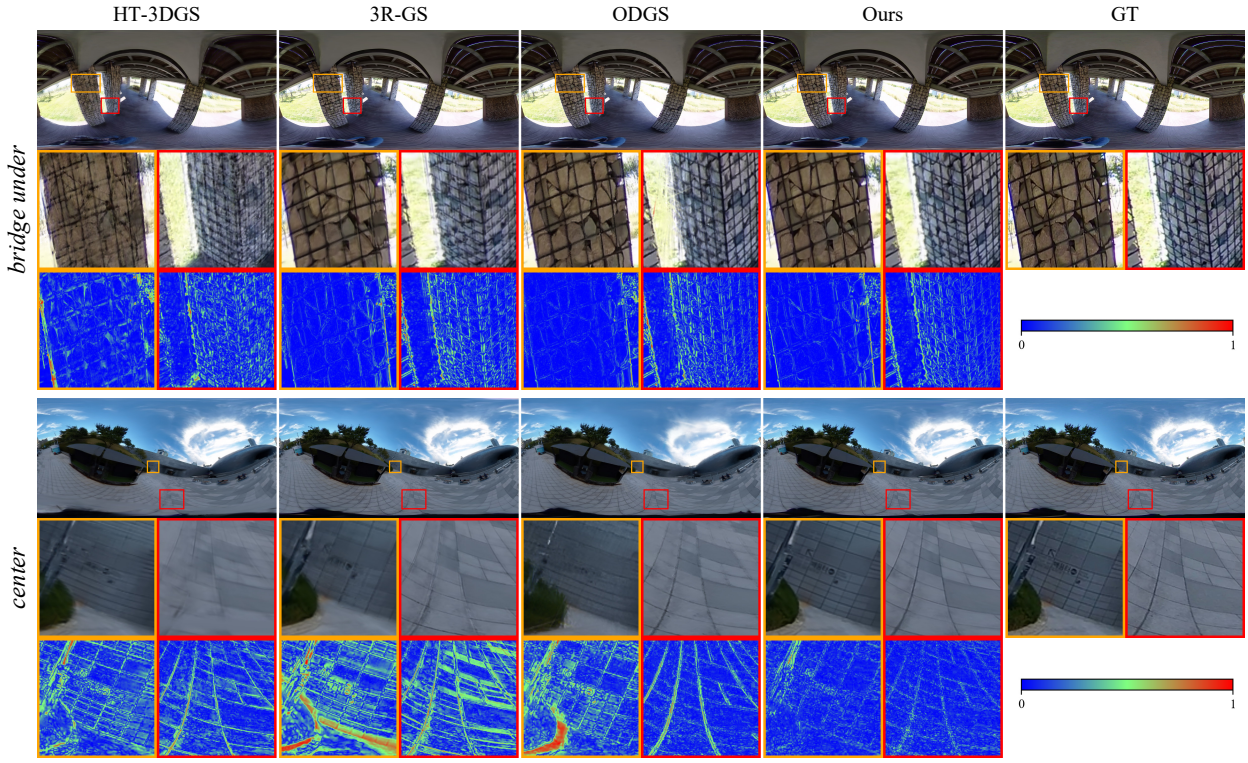

Figure 1. **Visualization results for novel view synthesis on Ricoh360.**

A.3. Detail Results for Quantitative Comparisons

We first report detailed results for all methods on the *OB3D* dataset, as summarized in Tables 3 and 4. In the *OB3D Egocentric* subset (Table 3), 3R-GS[3] substantially outperforms CF-3DGS[1] and HT-3DGS[4] in both novel view synthesis and camera pose estimation across nearly all scenes. In contrast, within the *OB3D NonEgocentric* subset (Table 4), the much larger camera motion prevents 3R-GS from recovering valid poses in scenes such as *fisher-hut* and *sponza*, leading

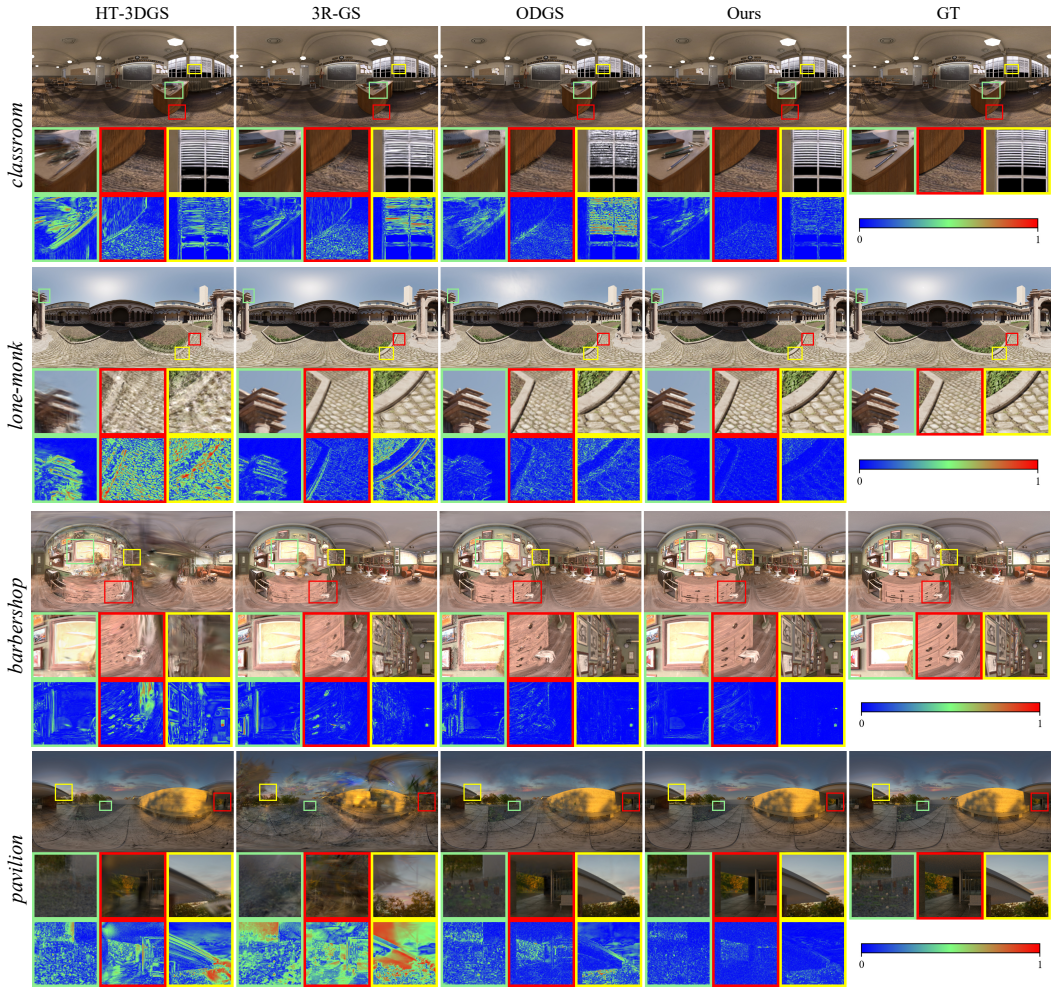

Figure 2. **Visualization results for novel view synthesis on OB3D.**

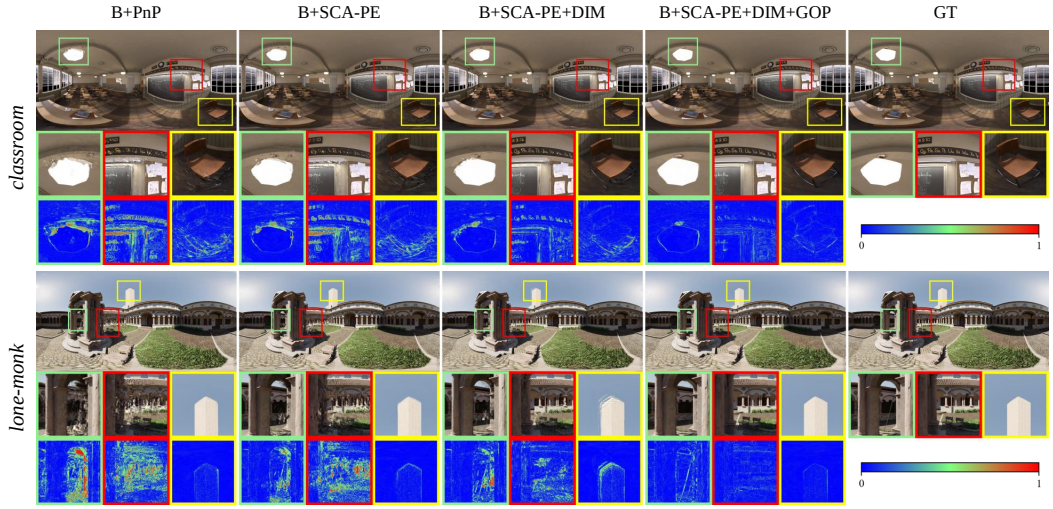

Figure 3. **Visual ablation results on model components.**

to failures in novel view synthesis. By comparison, our approach reliably recovers camera poses in all scenes under both trajectory settings. Moreover, it consistently produces the most photorealistic panoramic novel views, demonstrating substantially stronger robustness.

We additionally provide detailed novel view synthesis results on the Ricoh360 dataset. As shown in Table 5, our method consistently outperforms both pose-free [1, 3, 4] and pose-aware [6, 8] baselines across all scenes.

A.4. Visualization of Error Maps

To further highlight the visual differences across methods and model components, we provide supplementary error maps for novel view synthesis. As shown in Figures 1, 2, and 3, our method achieves the lowest overall reconstruction error. Moreover, the upper bound of our error distribution is significantly lower than that of both pose-free and pose-aware baselines, as reflected by the near absence of high-error (red) regions in the maps.

A.5. Camera Poses with Different Depth Priors

In the main submission, we analyzed several monocular depth priors, and we observed that the adjacent-frame scale inconsistency in DepthAnywhere [10] and DA^2 [7] negatively impacts both camera pose estimation accuracy and novel view synthesis quality. As shown in Figure 4, in most scenes, such as *bistro*, the scale alignment between monocular depth and the Gaussians effectively compensates for adjacent-frame scale drift, producing consistent camera poses. However, in the challenging *fisher-hut* scene, heavy vegetation causes alignment failures, leading to large pose estimation errors for both DepthAnywhere and DA^2. Consequently, their overall pose estimation performance degrades substantially on the *OB3D NonEgocentric* subset.

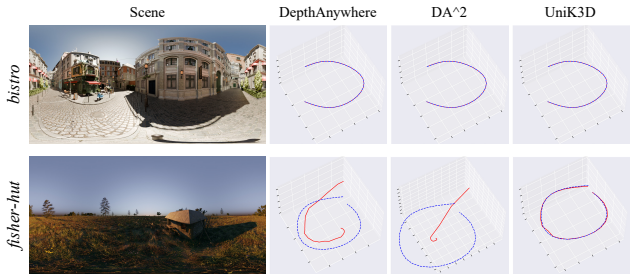

Figure 4. **Camera pose estimation with different depth priors.**

A.6. Results with Poor Monocular Depth Maps

As described in the submission, our algorithm leverages monocular depth estimation to provide geometric cues. Although monocular depth predictions may become unreliable in out-of-distribution scenarios, our method still reconstructs photorealistic Gaussians through consistency-based filtering. As shown in Figure 5, the monocular depth estimation model (UniK3D) produces a poor depth map in the *bridge_under* scene, where structural details around the pillars and the grass are missing. In contrast, the reconstruction result of PFGS360 generates a depth map that is consistent with the underlying scene geometry. This result validates the robustness of our method to the reliability of depth priors. Even when the monocular depth estimates are unreliable, the algorithm can still extract dependable geometric cues in most scenes through depth-consistency checks.

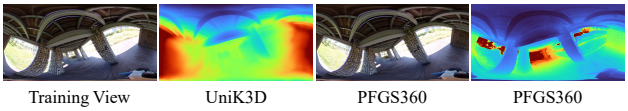

Figure 5. **Failed monocular depth estimation and corrected rendering results.**

A.7. Failure Cases.

Our method leverages 2D–3D correspondences and consistency checks to recover camera poses for panoramas, significantly outperforming approaches based on appearance loss. However, when inter-frame motion or scene changes are excessively large, such as in videos with abrupt discontinuous jumps, reliable 2D-3D correspondences between new and previous frames cannot be established, leading to the failure cases shown in 6.

Figure 6. **Results of 360Roam[2] with 5x motion speed.**

A.8. Limitations

Our method relies on 3DGS under the static-scene assumption, thus it cannot handle fully dynamic scenes. Despite this, we can reconstruct static objects across dynamic video frames. As shown in Figure 7 with a moving camera operator (red box), the consistency checks can extract 2D–3D correspondences and inlier Gaussians from static objects (green box), to recover camera poses and reconstruct their 3DGS representations.

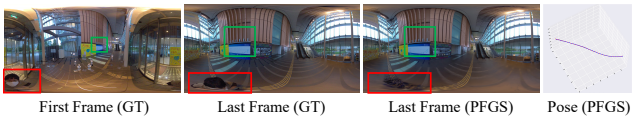

Figure 7. **Results of Miraikan 360° Video[5].**

References

[1] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20796–20805, 2024. 3, 5

[2] Huajian Huang, Yingshu Chen, Tianjia Zhang, and Sai-Kit Yeung. Real-time omnidirectional roaming in large scale indoor scenes. In *SIGGRAPH Asia 2022 Technical Communications*. Association for Computing Machinery, 2022. 5

[3] Zhisheng Huang, Peng Wang, Jingdong Zhang, Yuan Liu, Xin Li, and Wenping Wang. 3r-gs: Best practice in optimizing camera poses along with 3dgs. *arXiv preprint arXiv:2504.04294*, 2025. 3, 5

[4] Bo Ji and Angela Yao. Sfm-free 3d gaussian splatting via hierarchical training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21654–21663, 2025. 3, 5

[5] Seita Kayukawa, Keita Higuchi, Shigeo Morishima, and Ken Sakurada. 3dmoviemap: An interactive route viewer for multi-level buildings. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2023. 5

[6] Suyoung Lee, Jaeyoung Chung, Jaeyoo Huh, and Kyoung Mu Lee. Odgs: 3d scene reconstruction from omnidirectional images with 3d gaussian splattings. *Advances in Neural Information Processing Systems*, 37:57050–57075, 2024. 1, 5

[7] Haodong Li, Wangguangdong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. Da^2: Depth anything in any direction. *arXiv preprint arXiv:2509.26618*, 2025. 5

[8] Longwei Li, Huajian Huang, Sai-Kit Yeung, and Hui Cheng. Omnigs: Fast radiance field reconstruction using omnidirectional gaussian splatting. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2260–2268. IEEE, 2025. 5

[9] Chin-Yang Lin, Cheng Sun, Fu-En Yang, Min-Hung Chen, Yen-Yu Lin, and Yu-Lun Liu. Longsplat: Robust unposed 3d gaussian splatting for casual long videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27412–27422, 2025. 1

[10] Ning-Hsu Albert Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *Advances in Neural Information Processing Systems*, 37:127739–127764, 2024. 5