1. Data

1.1. Capture

For real-world data, the aerial view images are captured by an M350RTK DJI drone equipped with five SHARE 304S cameras, as shown in Fig. 2(a). The resolution of these images is 9552 × 6368, and each camera has a sensor size of 36 mm.

Street view images are captured by a custom designed helmet equipped with six DJI Osmo Action4 cameras, following Hierarchical-3DGS [4], as visualized in Fig. 2(b). The resolution of these images is 3840 × 2160. We use a DJI Osmo Action GPS Bluetooth remote to connect and operate all six cameras simultaneously. During the data collection process, we wear the helmet and walk to ensure image stability. The cameras are set to auto exposure, auto white balance, and timelapse mode with a 1-second interval. Each camera has a sensor size of 19.5 mm.

Following the setting of Gaussian Splatting [3], we resize the the longest side original images to 1600 pixels.

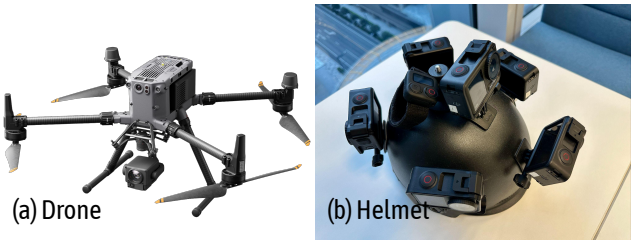

Figure 1. (a) M350RTK DJI drone for aerial images. (b) Helmet with six DJI Osmo Action4 cameras for street images

For the synthetic data in our dataset, we maintain the alignment of the cameras' roll, pitch, and yaw angles with those of the real-world scenes to ensure the uniformity of all data, as shown in Tab. 1.

Rot	Aerial			Street		
	Roll	Pitch	Yaw	Roll	Pitch	Yaw
1	0	-45	0	0	0	0
2	0	-45	90	0	25	0
3	0	-45	180	0	0	75
4	0	-45	270	0	0	145
5	0	-90	0	0	0	-145
6				0	0	-75

Table 1. Camera rotation parameters in synthetic scenes.

1.2. Discussion

In practical captures, camera density and coverage imbalance can vary significantly due to equipment differences. Our aerial-to-street dataset is a specific example; the insights and methodology can also apply to a broader range of multiscale, multi-source datasets.

2. More implementation

2.1. Global Appearance Embedding

In large-scale scenes, the data is typically captured in different environments, leading to inconsistent exposures. Inspired by Octree-GS [7] and Hierarchical-3DGS [4], we employ classical generative Latent Optimization (GLO) [1] to optimize individual appearance embedding vectors for each training image. To ensure consistent appearance codes across different chunks, we initially train the Gaussian primitives without densification for a few iterations, as the appearance codes mainly capture global and low-frequency attributes of the scene.

Scene		Aerial			Street		
Method	Metrics	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline [6]		20.18	0.539	0.549	21.22	0.626	0.394
Single Domain		22.42	0.666	0.402	21.85	0.653	0.362
Finetune		21.36	0.606	0.473	21.72	0.648	0.367
Ours		**23.23**	**0.729**	**0.322**	**22.04**	**0.669**	**0.325**

Table 2. Quantitative comparison using naive finetuning solutions.

2.2. Mesh Extraction

For mesh extraction, we adopt the 2D-GS[2] approach, rendering depth maps and fusing them into a TSDF volume, with the maximum depth range calculated based only on aerial views due to their wider coverage. The marching cube resolution is 1024.

3. More Experiments

3.1. Gradients Conflict

We visualize the maximum gradient norm of Scaffold-GS [6] for aerial-only, street-only, and combined training (calculated for each subset of data below). Separate training results in a higher gradient norm, particularly during the densification stage in early training. This observation across datasets highlights inherent gradient conflicts in the cross-domain setting, resulting in degraded performance.

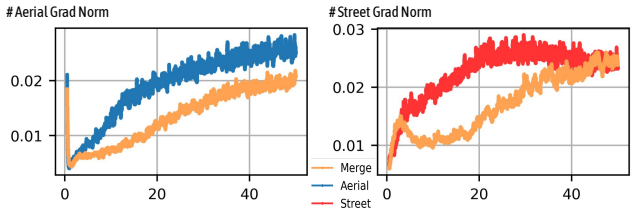

Figure 2. The maximum gradient norm of Scaffold-GS across aerial only, street only, and merged views on the Road scene.

Scene	City						Colosseum						Elevenruin					
	Aerial			Street			Aerial			Street			Aerial			Street		
Method | Metrics	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
2D-GS [2]	25.27	0.739	0.391	21.75	0.708	0.439	22.50	0.752	0.382	25.76	0.905	0.143	26.49	0.842	0.350	24.21	0.773	0.297
Our-2D-GS	32.21	0.931	0.113	23.94	0.808	0.297	25.40	0.891	0.163	26.25	0.899	0.141	33.56	0.952	0.133	26.12	0.837	0.211
3D-GS [3]	26.79	0.784	0.351	21.79	0.723	0.422	22.25	0.754	0.380	25.30	0.910	0.132	27.49	0.857	0.333	24.87	0.795	0.276
Scaffold-GS [6]	30.03	0.890	0.187	23.98	0.796	0.334	25.14	0.854	0.226	25.33	0.867	0.187	31.21	0.928	0.175	26.10	0.835	0.219
Hier-GS [4]	29.15	0.871	0.206	24.51	0.810	0.298	23.67	0.805	0.314	25.74	0.915	0.129	31.67	0.922	0.211	26.50	0.858	0.160
Ours	33.95	0.946	0.092	24.28	0.827	0.264	25.85	0.900	0.139	26.11	0.904	0.133	34.99	0.967	0.071	26.67	0.855	0.173

Table 3. Quantitative comparison on each synthetic scene of our proposed dataset.

3.2. Naive Solutions

Based on the observations discussed in Section **??**, a naive solution is to merge the results from training on individual domains. To eliminate artifacts at the seams and maintain consistency in the feature space, we conduct an experiment where we concatenate the results from training on a single domain and fine-tuned the model for an additional 10k iterations on the Road and Park scenes. As shown in Tab. 2, this fine-tuning approach inefficient, time-consuming, and fails to address the core issue.

3.3. More Results

We report quantitative results for each scene of our proposed dataset, as discussed in the main text: synthetic scenes (City, Colosseum, and Elevenruin) and real scenes (Road, Park). These results cover image quality metrics such as PSNR, SSIM [8], and LPIPS [9], as shown in Tables 3, 4, 5.

Additionally, we compare our approach with UC-GS [10] and Hier-GS [4] equipped with camera selection strategies (R=1), both of which serve as strong baselines. Despite their advanced configurations, our approach consistently outperforms these methods, particularly in textureless and high-frequency regions, as demonstrated in Fig. 3.

Scene	Road					
	Aerial			Street		
Method | Metrics	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
2D-GS [2]	19.63	0.484	0.584	19.37	0.541	0.468
Our-2D-GS	21.79	0.645	0.384	20.57	0.628	0.349
3D-GS [3]	19.95	0.509	0.562	20.17	0.573	0.435
Scaffold-GS [6]	20.36	0.532	0.532	20.08	0.580	0.422
UC-GS [10]	21.00	0.581	0.468	20.59	0.610	0.378
Hier-GS [4]	21.22	0.620	0.432	**21.30**	**0.651**	**0.312**
Hier-GS + cam bal.	21.45	0.635	0.413	20.84	0.631	0.346
Ours	**22.60**	**0.682**	**0.356**	20.94	0.637	0.341

Table 4. Quantitative comparison on Road scene.

Scene	Park					
	Aerial			Street		
Method | Metrics	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
2D-GS [2]	19.76	0.524	0.586	21.80	0.664	0.376
Our-2D-GS	23.35	0.729	0.330	22.46	0.681	0.339
3D-GS [3]	20.23	0.545	0.565	22.64	0.681	0.361
Scaffold-GS [6]	19.99	0.545	0.565	22.35	0.672	0.366
UC-GS [10]	20.62	0.586	0.511	22.91	0.688	0.341
Hier-GS [4]	21.63	0.657	0.427	**23.75**	**0.720**	**0.294**
Hier-GS + cam bal.	21.97	0.672	0.403	22.73	0.685	0.346
Ours	**23.85**	**0.776**	**0.287**	23.14	0.701	0.308

Table 5. Quantitative comparison on Park scene.

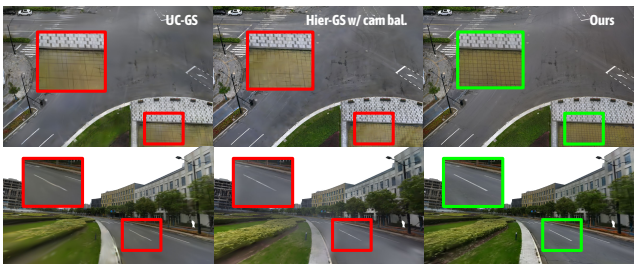

Figure 3. Qualitative comparisons of Horizon-GS against UC-GS [10] and Hier-GS [4] with Camera Balance strategy.

3.4. Ablation

We select Scaffold-GS [6] as our baseline and perform two additional ablation studies focusing on the fine stage and global appearance embedding, respectively. For quantitative results, we use the Road and Park scenes, while Block_A is used for qualitative analysis.

Scene	Aerial			Street		
Method | Metrics	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline [6]	20.18	0.539	0.549	21.22	0.626	0.394
Ours w/o fine stage	**23.32**	0.725	0.326	21.69	0.658	0.338
Ours w/o depth	23.21	0.728	0.322	**22.17**	**0.670**	0.326
Ours	23.23	**0.729**	**0.321**	22.04	0.669	**0.324**

Table 6. Ablations on our proposed real-world scenes.

Fine Stage. The second stage is used for complementing the details of aerial views. The rendering quality will decrease hugely if discarding it, as shown in Tab. 6.

Depth Supervision. Depth supervision is an optional scene-dependent parameter applied equally across all baselines. Although depth supervision does not directly improve per-view metrics, it notably enhances weakly textured regions, such as road surfaces.

4. Limitation and More Discussion

In this paper, we analyze the challenges of unified large-scale scene reconstruction from both aerial and street views, and propose a systematic solution that delivers high-quality benchmarks and results. The modules in our system are not mere incremental improvements, but essential components of a cohesive framework designed for robust aerial-to-street view reconstruction in practical, city-scale applications. While our method proves effective in reconstruction and producing high-quality results, it also has certain limitations. First, similar to most Gaussian-based methods, Horizon-GS may reach suboptimal solutions when there is insufficient input information. In future work, we plan to leverage advanced foundation models to guide the optimization process more effectively. Additionally, the divide-and-conquer approach inevitably introduces redundancy due to the required overlaps for seamless merging between chunks. Implementing more systematic approaches, such as Grendel-GS [11] or RetinaGS [5], also presents a promising solution.

References

[1] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 599–608. PMLR, 2018. 1

[2] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, page 32. ACM, 2024. 1, 2

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 1, 2

[4] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Trans. Graph.*, 43(4): 62:1–62:15, 2024. 1, 2

[5] Bingling Li, Shengyi Chen, Luchao Wang, Kaimin Liao, Sijie Yan, and Yuanjun Xiong. Retinags: Scalable training for dense scene rendering with billion-scale 3d gaussians. *CoRR*, abs/2406.11836, 2024. 3

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 20654–20664. IEEE, 2024. 1, 2

[7] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *CoRR*, abs/2403.17898, 2024. 1

[8] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 2

[9] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[10] Saining Zhang, Baijun Ye, Xiaoxue Chen, Yuantao Chen, Zongzheng Zhang, Cheng Peng, Yongliang Shi, and Hao Zhao. Drone-assisted road gaussian splatting with cross-view uncertainty. *CoRR*, abs/2408.15242, 2024. 2

[11] Hexu Zhao, Haoyang Weng, Daohan Lu, Ang Li, Jinyang Li, Aurojit Panda, and Saining Xie. On scaling up 3d gaussian splatting training. *CoRR*, abs/2406.18533, 2024. 3