

Horizon-GS: Unified 3D Gaussian Splatting for Large-Scale Aerial-to-Ground Scenes

Lihan Jiang^{1,3*} Kerui Ren^{2,3*} Mulin Yu³ Lining Xu⁴
 Junting Dong³ Tao Lu⁵ Feng Zhao¹ Dahua Lin⁴ Bo Dai^{6,3✉}

¹University of Science and Technology of China, ²Shanghai Jiao Tong University,

³Shanghai Artificial Intelligence Laboratory, ⁴The Chinese University of Hong Kong,

⁵Brown University, ⁶The University of Hong Kong

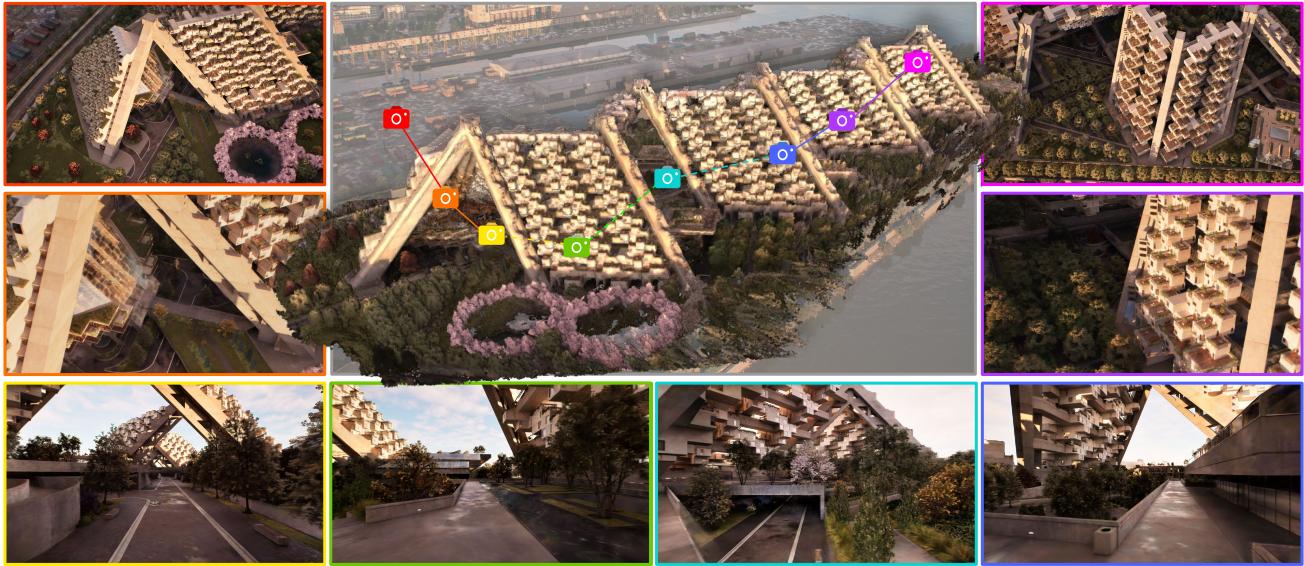


Figure 1. Horizon-GS enables high-quality rendering and reconstruction of aerial-to-ground scenes with unprecedented fidelity across scales, supporting drastic view changes. The colored camera trajectories in the center illustrate the novel viewpoints, while the reconstructed mesh is overlaid on the scene. The surrounding images show the corresponding predicted images for each viewpoint.

Abstract

Seamless integration of both aerial and street view images remains a significant challenge in neural scene reconstruction and rendering. Existing methods predominantly focus on single domain, limiting their applications in immersive environments, which demand extensive free view exploration with large view changes both horizontally and vertically. We introduce Horizon-GS, a novel approach built upon Gaussian Splatting techniques, tackles the unified reconstruction and rendering for aerial and street views. Our method addresses the key challenges of combining these perspectives with a new training strategy, overcoming viewpoint discrepancies to generate high-fidelity scenes. We also curate a high-quality aerial-to-ground views dataset encompassing both synthetic and real-world scene to ad-

vance further research. Experiments across diverse urban scene datasets confirm the effectiveness of our method. Project page: <https://city-super.github.io/horizon-gs/>.

1. Introduction

Modeling and rendering large-scale scenes has become essential across domains such as Embodied AI, digital twins, and autonomous driving. Neural radiance fields [20] have revealed the potential to realize this vision. Recently, 3D Gaussian Splatting (3D-GS) [9] has advanced this field further, achieving exceptional visual quality alongside real-time rendering speeds. These capabilities markedly accelerate the transformation of the physical world into digital spaces, supporting applications from real-to-sim simulation

and autonomous navigation to immersive VR/AR experiences and the metaverse.

Despite recent advancements in rendering aerial [3, 14, 18] and street views [10] with 3D-GS, existing methods remain limited to training on a single view type, resulting in a disjointed experience. We aim to bridge this gap by integrating both perspectives, enabling a seamless and unified experience across downstream applications. Our preliminary experiments on state-of-the-art Gaussian Splatting methods reveal two primary challenges in integrating aerial and street views: 1) conflicts in Gaussian densification arising from the substantial disparity between aerial and street views, and 2) an uneven distribution of camera data, with a bias toward street views. During the densification process, limited visibility from street views impedes consistent gradient accumulation, thereby disrupting the Gaussian instantiation process across the scene. Moreover, the variation in detail between aerial and street views makes a single static set of Gaussian primitives insufficiently flexible, restricting smooth transitions between viewpoints.

To address above challenges and achieve a seamless integration for both aerial and street views, we propose a novel method, Horizon-GS, that overcomes the limitations of existing approaches. Specifically, we propose a coarse-to-fine training strategy that explicitly divides the entire training process into two stages: the first stage benefits more from aerial images to establish a coarse geometry, and the second stage focuses on adding finer details mainly from street views. Combined with a camera distribution balance strategy, our method achieves state-of-the-art quality for both aerial and street views, delivering superior rendering quality and geometry accuracy. Moreover, our approach is easily adaptable for large-scale reconstruction. We also notice that, the lack of calibrated aerial and street datasets is a critical issue that hinders unified model training. To advance the field and evaluate our method, we construct a large-scale dataset consisting of five synthetic and two real-world scenes, encompassing both aerial and street views.

In summary, the main contributions of our method are:

- We tackle the challenging and important task of unified large-scale scene reconstruction from both aerial and street views, starting with an in-depth analysis of the core conflicts that hinder integration across these perspectives.
- We propose an efficient framework that addresses the inherent discrepancies and unifies the contribution from aerial and street views. Our framework is also adaptable to both 3D Gaussians for realistic novel view synthesis and 2D Gaussians for accurate geometry reconstruction.
- We present a high-quality and diverse dataset that includes both synthetic and real-world data, specifically curated to support balanced, cross-view reconstruction.
- Our approach achieves state-of-the-art performance in small-scale and large-scale scene reconstruction, setting

a new benchmark for unified aerial and street modeling.

2. Related work

Neural Scene Representations. Neural Radiance Fields (NeRF) [20] have sparked significantly interest in novel view synthesis due to their photorealistic visual quality. Recent research [2, 6, 17, 21, 26, 34] has introduced various hybrid grid representations, reducing the number of sampling points and MLP layers, accelerating both the training and rendering processes. More recently, 3D Gaussian Splatting (3D-GS) [9] has achieved more detailed visual quality with efficient training times by using anisotropic 3D Gaussians, while enabling real-time rendering through tile-based splatting methods. It quickly revolutionizes neural rendering and is widely applied in various fields, especially in many large-scale applications, such as autonomous driving [4, 32, 38, 47], geometry reconstruction [7, 40, 41], and 3D generative tasks [28, 29, 37, 42].

Large-scale Scene Modeling. NeRF and 3D Gaussian Splatting have sparked a new trend in neural modeling for large-scale scenes, pushing the boundaries of visual fidelity and real-time performance. Extensive neural representation methods have been developed for large-scale city reconstruction and rendering. However, most methods target either aerial [3, 14, 18, 24, 30, 33, 36, 46] or street views [10, 16, 23, 27]. For aerial views, models like Vast-Gaussian [14] and CityGaussian [18] use data partitioning for parallel training, while DoGaussian [3] employs a global node to ensure consistent inference. For street scenes, Hierarchical-3DGS [10] optimizes and renders Gaussians in an LOD hierarchy. However, these methods do not naturally support a unified reconstruction from both aerial and street views. The work closest to ours is UC-GS [44], which incorporates cross-view uncertainty to enhance details in extrapolated views but lacks the scalability for large-scale applications, limiting its practicality.

Another major barrier to unified aerial and street modeling is the scarcity of datasets with calibrated images from both perspectives. Current large-scale neural rendering datasets, such as UrbanScene3D [15], Mill19 [30], Quad 6K [5], and MatrixCity [12], Waymo Block-NeRF [27], KITTI-360 [13], and Hierarchical-3DGS [10], typically cater to a single view type. MatrixCity is the only dataset with both views, but it is synthetic and lacks diversity. To fill this gap, we introduce a new dataset with a variety of synthetic and real-world scenes, supporting research on large-scale, unified reconstruction.

3. Horizon-GS

While substantial progress has been made in reconstructing urban scenes from either aerial or street images [3, 10,

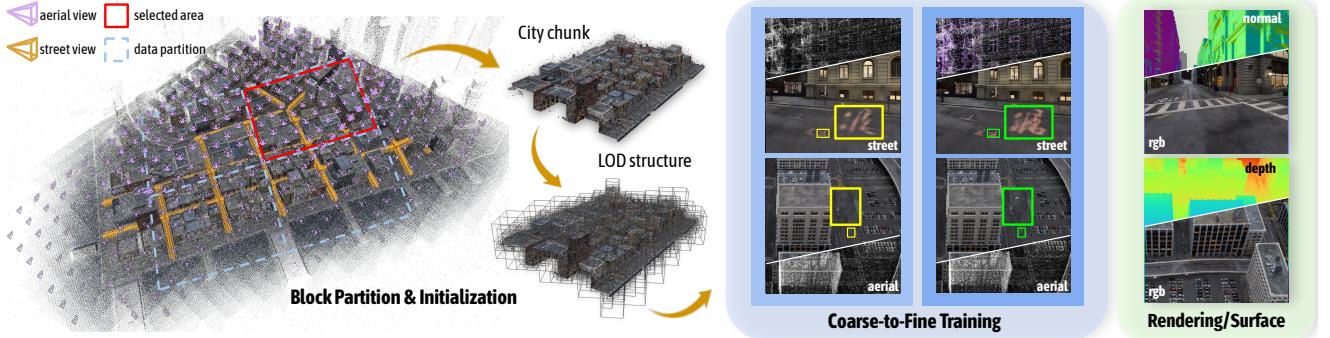


Figure 2. **Pipeline of Horizon-GS.** We divide large-scale scenes into chunks. For each chunk, we initialize LOD-structured anchors and conduct the coarse-to-fine training process. Specifically, the coarse stage reconstructs the overall scene, while the fine stage enhances street view details (highlighted in purple). We can derive RGB, depth, and normal images by utilizing different primitive attributes (2D/3D Gaussians) with a single shared underlying structure.

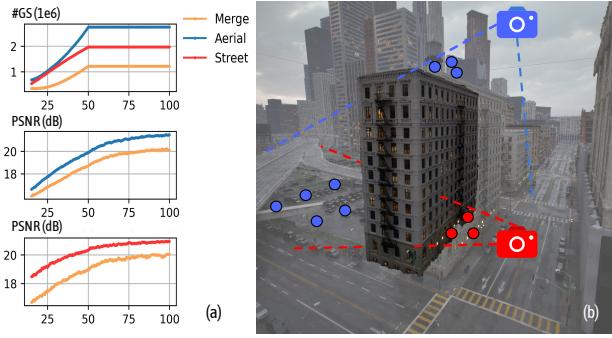


Figure 3. (a) Test curves for PSNR and the number of Gaussian primitives across aerial only, street only, and merged views from 15k to 100k iterations on our proposed Road scene. (b) Gradient conflicts restrict the optimization of Gaussian primitives because street views tend to exclude blue Gaussian primitives due to their lower contribution, while aerial views do the opposite.

[14, 18], the challenge of combining both perspectives into a unified model remains unsolved and largely overlooked. This task is non-trivial, requiring solutions to bridge the significant disparities in scale, perspective, and appearance between aerial and street views. In this section, we first analyze the core challenges in Section 3.1, briefly describe the base modules in Section 3.2, and then detail the proposed framework that leverages a two-stage training strategy and the enhanced 3D Gaussian densification operators in Section 3.3. Additionally, how to apply the framework to large-scale scenes is discussed in Section 3.4. Finally, we describe the loss functions and regularization in Section 3.5.

3.1. Challenges

Before presenting our framework, we first analyze the inherent conflicts in using 3D Gaussians for simultaneous training on aerial and street views. We apply the state-of-the-art Scaffold-GS [19] on our collected real-world scene which features both aerial and street views. Naturally, we

aim to combine these two classes of views during reconstruction, as they complement each other. But we notice that reconstructing either aerial or street views alone has achieved much better results than training them together, as shown in Fig. 2 (b) and Fig. 3. Combining these perspectives presents two major challenges: (1) conflicting gradient accumulation and (2) disparity in view information coverage.

The first challenge arises from the differences between aerial and street views, which lead to conflicting gradient updates during training. While street views cause the densification policy to remove irrelevant, occluded Gaussians, aerial views promote regrowth in these regions. This interference disrupts the Gaussian densification process, making it unreliable and ultimately degrading reconstruction quality. Fig. 3 confirms that joint training across both domains consistently underperformed separate training, even with extensive iterations, and with more observed views, the number of final Gaussian primitives are even less.

The second challenge is the imbalance in camera coverage: street views capture detailed, nearly 360-degree scenes, while aerial views are downward-focused and less frequent. This uneven distribution shifts the reconstruction focus toward street details, preventing a balanced integration of both perspectives. Together, these challenges emphasize the complexity of unified aerial-street modeling and the necessity for specialized solutions.

3.2. Gaussian Splatting Base Modules

3D Gaussian Splatting (3D-GS) [9] represents scenes using anisotropic 3D Gaussian primitives, achieving high rendering quality at significantly faster speeds. Each Gaussian G is defined by parameters: a center $\mu \in \mathbb{R}^3$, a rotation quaternion $q \in \mathbb{R}^4$, scaling vector $s \in \mathbb{R}^3$, opacity $\sigma \in [0, 1]$, and a color feature c either encoded with spherical harmonics or direct RGB feature for view-dependent/independent color. To render images, the 3D Gaussians are projected onto the

image plane as 2D Gaussians $G'(x)$, ordered front-to-back, and α -blended for pixel color calculation.

Scaffold-GS [19] further introduces anchors to reduce redundancy and enhance storage efficiency. From each anchor, k neural Gaussians are emitted, with centers determined by learnable offsets. Properties are decoded by MLPs from anchor features and viewing positions, creating a more robust and compact representation.

2D Gaussian Splatting (2D-GS) [8] collapses 3D Gaussians into 2D oriented planar Gaussian disks for surface accuracy. An adaptive rasterizer efficiently renders these 2D Gaussians, and depth distortion and normal consistency losses are added to improve reconstruction quality. Finally, a mesh is extracted using a truncated signed distance function (TSDF), which fuses multi-view depth maps from the optimized 2D-GS field.

For enhanced robustness, we adopt the anchor design from Scaffold-GS [19], incorporating a flexible neural Gaussian representation. Specifically, the properties of different neural Gaussians, both 3D and 2D, can be generated by learnable MLPs. We utilize neural 3D Gaussians to improve rendering quality and neural 2D Gaussians for high-fidelity geometry reconstruction.

3.3. Aerial-Street Joint Reconstruction

Coarse-to-Fine Training. The significant difference in aerial and street views lead to performance challenges when training them together, necessitating separate handling for each modality. Inspired by BungeeNeRF [33], we propose a two-stage coarse-to-fine training strategy where street views refine the details of aerial views. In the first stage, we focus primarily on the aerial views to establish a coarse geometry, accumulating gradients for densification from the aerial images alone and allowing the Gaussian primitives to fully develop and cover a broader feature space. Street views guide the initial placement of fine-grained 3D Gaussians, ensuring their alignment with the global scene. In the second stage, we freeze the MLP weights and attributes of the Gaussian primitives from the first stage to preserve the skeleton structure. Street views then play a key role in refining this skeleton, adding finer details and achieving a balanced integration of both views while efficiently resolving their inherent conflicts.

We modify the densification policy in the second stage to enhance the capture of fine-grained details in large-scale scenes with sparse images. The original Scaffold-GS method averages screen-space gradients of neural Gaussians within each voxel. While effective for object-centric scenes, it struggles to model local details in sparsely observed areas. As a result, subtle features remain under-optimized, leading to blurred renderings.

Here we consider neural Gaussians with higher gradients, greater opacity, and larger projected screen size

as more significant, inspired by Hierarchical-3DGS [10]. Specifically, we compute the max gradients ∇_g , average opacity σ and max projected radius r of the included neural Gaussians for each N iterations. Those neural Gaussians satisfying:

$$\nabla_g \cdot r \cdot \sigma^{\tau_\sigma} > \tau_g, \quad (1)$$

where τ_σ and τ_g are pre-defined thresholds, are treated as significant and used as new anchors.

Balanced Camera Distribution. As previously mentioned, the imbalanced distribution of cameras can lead to 3D-GS overfitting to street view details during joint training, causing blurring and loss of structural detail in aerial views. Inspired by NeRF Director [35], we emphasize the importance of view selection and introduce a straightforward yet effective mechanism for balancing camera distribution. Specifically, we randomly select a value from the interval $[0, 1]$. If the value is smaller than $\frac{R}{R+1}$, an aerial image is chosen for training; otherwise, a street image is picked. In detail, we set $R = 2$ in the first stage to ensure thorough training of the aerial views and reduce it to 1 in the second stage.

Multi-resolution LOD Construction. Considering the differing levels of detail in aerial and street view images, and to enable real-time rendering, we adopt a Level-of-Detail (LOD) strategy, inspired by Octree-GS [24]. Specifically, for content with varying details between aerial and street views, we set different LOD levels: K_{aerial} for aerial views and K for the whole views, corresponding to the two-phase training process.

$$K_{\text{aerial}} = \lfloor \log_2(D_{\text{aerial}}/d_{\text{aerial}}) \rfloor + 1, \\ K = K_{\text{aerial}} + \lfloor \log_2(d_{\text{aerial}}/d_{\text{street}}) \rfloor. \quad (2)$$

where D_{aerial} and d_{aerial} are the r_d -th largest and r_d -th smallest aerial distances, while d_{street} is the r_d -th smallest street distance. Note that, our model is set to K layers and K_{aerial} is only used in the first training stage. Specifically, in the first stage, we only activate K_{aerial} layers and add new anchors to the same LOD level as the significant source anchors. In the second stage, we open all layers and add the new anchors to the next LOD level to capture the higher-frequency information of street views.

3.4. Large-scale Scene Training

In this section, we present a partitioning strategy to adapt our framework for large-scale urban scenes, supporting both model scalability and rendering fidelity.

Gaussian-based methods often suffer from blurring artifacts when reconstructing large-scale scenes, mainly due to challenges in densifying extensive image sets that require fine-tuned parameters. Additionally, limited GPU

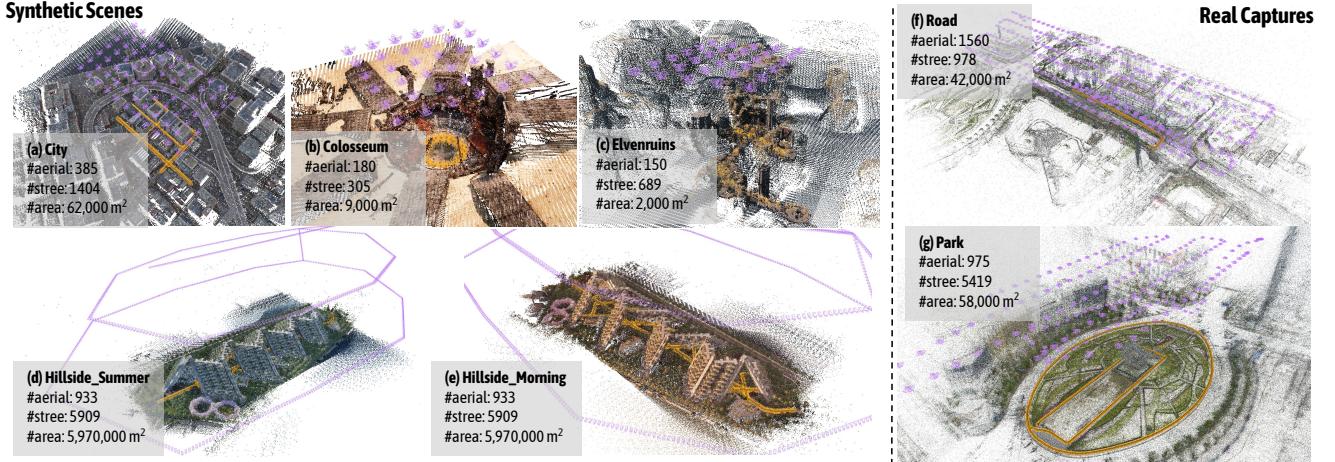


Figure 4. **Visualization of our constructed dataset.** All the 7 scenes contain calibrated aerial and street view images. We illustrate the scenes with the point clouds and the corresponding image capture poses. The trajectory of aerial views is shown in purple, while street views are represented in yellow. Our dataset contains 5 synthetic scenes (a-e) and 2 real scenes (f-g).

memory presents a hurdle when training on large-scale scenes. A common solution is the divide-and-conquer strategy, where a whole scene is split into smaller chunks. For instance, VastGaussian [14] progressive partitioning for aerial views but faces projection errors with street views during visibility-based camera selection. To better accommodate the reconstruction of large-scale scenes involving both aerial and street views, we make the following adjustments. Following VastGaussian, we first partition the scene into $m \times n$ chunks based on ground-projected camera positions, then expand the original boundaries of each chunk by a threshold to ensure sufficient overlap. Next, we apply different strategies for aerial and street views: for aerial views, we augment each chunk with additional visible cameras and accessory points, for street views, which experience significant occlusion, we generate point clouds from depth maps to ensure comprehensive training coverage. All scenes follow the Manhattan world assumption, with the z-axis perpendicular to the ground plane.

Each chunk is trained independently using the strategy discussed in Section 3.3, then merged into a single model by discarding Gaussians outside the original boundaries and concatenating Gaussian attributes across chunks. To enhance rendering speed, we convert the hybrid representation to a fully explicit one by removing view inputs from each MLP and replacing the MLPs for color with SH predictions.

3.5. Loss Function and Regularization

Photometric Loss. We use different losses for rendering and surface reconstruction. For rendering, in addition to the standard L1, SSIM Loss, we also employ volume regularization \mathcal{L}_{vol} from Scaffold-GS [19] to ensure Gaussian primitives effectively cover the entire scene.

Geometry Regularizations. For geometry awareness, we introduce depth supervision $\mathcal{L}_d = \frac{1}{hw} \sum |\hat{D} - D|$, where h and w the rendering image dimensions, D is the inverse of the rendered depth map and D_s is the pseudo ground truth disparity map generated by either rendering engines (for synthetic data) or a scale-aligned pretrained monocular depth model [39] (for real data). For surface reconstruction, we also incorporate the normal loss \mathcal{L}_n , as used in 2D-GS [8], to enforce normal consistency.

Mask Regularization. To mitigate the effects of moving pedestrians, vehicles, and the sky, which is challenging for vanilla GS methods, we employ an opacity regularization

$$\mathcal{L}_o = -\frac{1}{hw} \sum \sigma \cdot \log \sigma - \frac{1}{hw} \sum M \cdot \log(1 - \sigma), \quad (3)$$

which encourages the opacity values toward 0 or 1, where M denotes the mask region. Finally, the combined rendering and surface reconstruction losses, \mathcal{L}_R and \mathcal{L}_S , are defined as follows respectively:

$$\begin{aligned} \mathcal{L}_R &= \mathcal{L}_1 + \lambda_{\text{ssim}} \mathcal{L}_{\text{ssim}} + \lambda_{\text{vol}} \mathcal{L}_{\text{vol}} + \lambda_d \mathcal{L}_d + \lambda_o \mathcal{L}_o, \\ \mathcal{L}_S &= \mathcal{L}_R + \lambda_n \mathcal{L}_n. \end{aligned} \quad (4)$$

4. Dataset Curation

As previously mentioned, the lack of calibrated aerial and street datasets is a crucial issue that restricts the research on the unified scene reconstruction. To advance the field and evaluate our method, we construct a large-scale dataset with aerial and street views calibrated, which contains five synthetic scenes and two real scenes, as illustrated in Fig. 4. More details are listed in the supplementary materials.

Scene	Block_Small						Synthetic						Real						
	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 [8]		24.52	0.743	0.365	22.46	0.763	0.360	24.75	0.778	0.374	23.91	0.795	0.293	19.69	0.504	0.585	20.59	0.603	0.422
Our-2D-GS		29.60	0.899	0.121	23.60	0.837	0.216	30.39	0.925	0.136	25.44	0.848	0.216	22.57	0.687	0.357	21.52	0.655	0.344
3D-GS [9]		25.44	0.781	0.325	21.81	0.744	0.371	25.51	0.798	0.355	23.99	0.809	0.277	20.09	0.527	0.564	21.41	0.627	0.398
Scaffold-GS [19]		28.44	0.863	0.191	23.84	0.819	0.271	28.79	0.891	0.196	25.14	0.833	0.247	20.18	0.539	0.549	21.22	0.626	0.394
Hier-GS [10]		28.31	0.861	0.189	23.75	0.823	0.220	28.16	0.866	0.244	25.58	0.861	0.196	21.43	0.639	0.430	22.53	0.686	0.303
Ours		30.59	0.913	0.094	23.80	0.839	0.209	31.60	0.938	0.101	25.69	0.862	0.190	23.23	0.729	0.321	22.04	0.669	0.324

Table 1. Quantitative comparison on small-scale datasets. Horizon-GS consistently achieves superior rendering quality compared to baselines in both aerial and street views.

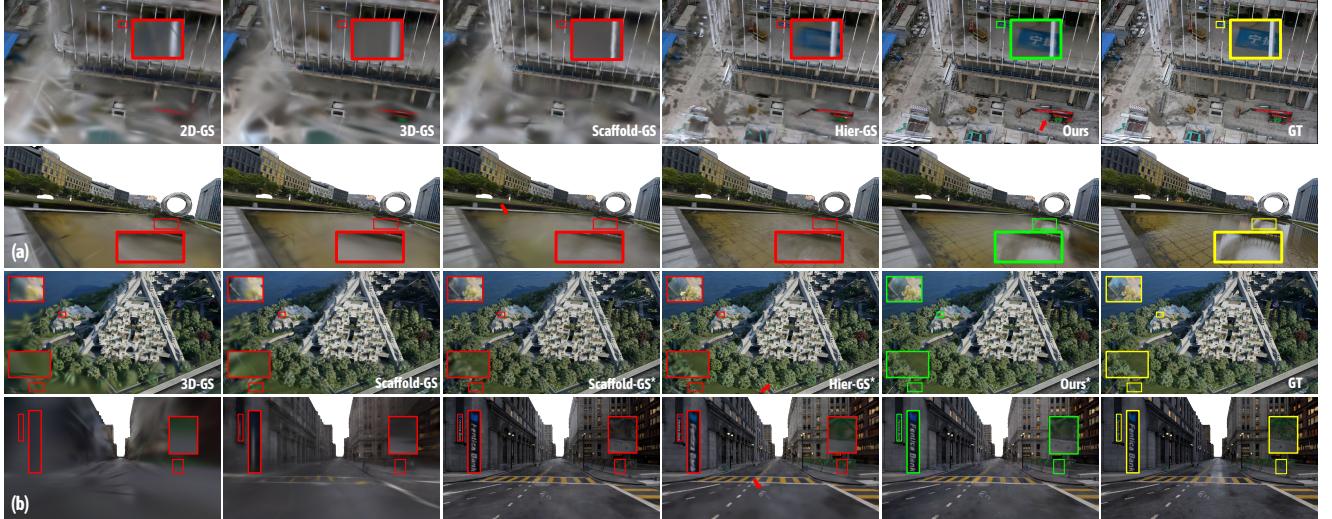


Figure 5. Qualitative comparisons of Horizon-GS against baselines [8–10, 19] across (a) small-scale and (b) large-scale scenes.

Synthetic Data. Following MatrixCity [12] curation process, we create 5 additional synthetic scenes to simulate real-world data using custom data collection plugins. To ensure diversity, these synthetic datasets encompass various scales, seasons, and environments, ranging from realistic cityscapes to imaginative gaming scenarios. In addition to RGB images, we render depth maps for geometry supervision and initial point clouds for 3D-GS training.

Real Data. To evaluate Horizon-GS in real-world data, we collect 2 additional scenes. We use a DJI drone equipped with five cameras to capture aerial images and a custom-designed helmet equipped with six Action4 cameras to capture street data. To enhance the performance of the structure-from-motion (SfM) technique, we include transitional views to increase image overlap and utilize a powerful commercial software, ContextCapture¹. Depth maps are generated using Depth-Anything-V2 [39], and scale/offset are estimated by aligning inverse depth of SfM points. Moving objects, such as humans and cars, are removed by Grounded-SAM [25], while sky masking is handled by the

pretrained SkyRemoval model [22]. More details about the dataset are provided in the supplementary materials.

5. Experiments

5.1. Experimental Setup

Datasets and Metrics. We conduct comprehensive evaluations across 11 scenes containing both aerial and street views, sourced from the MatrixCity dataset [12], the UC-GS dataset [44], and our captured dataset. Specifically, in the MatrixCity dataset, we evaluate two synthetic scenes: Block_Small and Block_A. For the UC-GS dataset, we select two synthetic scenes, New York City (NYC) and San Francisco (SF). Consistent with the original UC-GS settings, evaluations are performed solely on street views. For our newly captured dataset, we select one out of every 32 images for evaluation. Among them, we divide Block_A into 4×2 chunks and divide Hillside_Morning and Hillside_Summer into 6×2 chunks to improve GPU compatibility and achieve satisfactory visual results. We use image resolution 1920×1080 for synthetic data; 1600×1066 and 1600×900 for real-world aerial and street view images.

Results are evaluated using standard visual quality met-

¹<https://www.daspatial.com/cn/gcluster>

Scene	Block_A						Hillside_Morning						Hillside_Summer						
	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 [8]		20.63	0.585	0.595	19.57	0.635	0.477	22.02	0.629	0.513	19.20	0.637	0.470	20.87	0.560	0.524	18.70	0.614	0.481
Our-2D-GS*		28.52	0.886	0.166	23.00	0.789	0.301	29.98	0.900	0.135	20.44	0.686	0.404	28.88	0.889	0.126	20.23	0.674	0.404
3D-GS [9]		21.02	0.590	0.595	19.06	0.628	0.477	22.87	0.662	0.489	19.35	0.643	0.466	21.50	0.592	0.503	18.63	0.617	0.480
Scaffold-GS [19]		22.98	0.654	0.523	20.89	0.681	0.436	23.05	0.688	0.433	20.08	0.656	0.461	22.02	0.634	0.436	19.36	0.629	0.473
Scaffold-GS*		27.62	0.860	0.206	23.10	0.808	0.277	25.46	0.796	0.271	20.66	0.685	0.416	23.91	0.752	0.279	20.30	0.667	0.423
Hier-GS* [10]		26.42	0.792	0.340	23.59	0.811	0.275	25.41	0.783	0.269	21.13	0.692	0.398	23.53	0.735	0.273	20.47	0.673	0.402
Ours*		28.89	0.890	0.151	23.66	0.816	0.255	31.09	0.918	0.107	20.62	0.689	0.399	29.34	0.900	0.110	20.34	0.681	0.393

Table 2. Quantitative comparison on large-scale datasets. Horizon-GS achieves superior rendering quality compared to baselines.

rics: PSNR, SSIM [31], and LPIPS [43]. Aerial and street view images are evaluated separately. To minimize the impact of pedestrians, cars, and the sky, we mask out those moving entities out when comparing with ground truths.

Baselines. We compare our method against 3D-GS [9], Scaffold-GS [19], UC-GS [44] and Hierarchical-3DGS [10] for rendering, as well as 2D-GS [8] for surface reconstruction. To ensure full convergence, we extend the total training iterations from 30k to 100k, with densification continuing until the 50k iteration. For all baselines, we run their official code on all datasets except for the UC-GS dataset, where we adopt the metrics from their original paper to maintain consistency with their settings. For Hierarchical-3DGS, we report the best visual quality results after optimizing the hierarchy (leaves). For each baseline, we also add depth supervision same as our method. In the large-scale setting, we partition the scenes and train each chunk in parallel, using * to denote the strategy discussed in Sec. 3.4.

Implementation Details. For a fair comparison, we train the first stage for 60k iterations and the second stage for 40k iterations, with densification stopping at 30k and 20k iterations, respectively. In the first stage, we follow the default settings of Scaffold-GS [19] for learning rate and densification. In the second stage, we reduce the learning rate of the offset by a factor of 10 and set τ_σ , τ_g and N to 0.2, 0.15 and 100, respectively. In LOD construction stage, r_d is set to 0.999. The loss function parameters are $\lambda_{\text{ssim}} = 0.2$, $\lambda_{\text{vol}} = 0.01$, $\lambda_o = 0.05$, and $\lambda_n = 0.05$. The weight for \mathcal{L}_d is exponentially decayed from 1 to 0.01 over both stages. Depth supervision is activated after 500 iterations, and normal supervision after 7k iterations. We denote neural Gaussians that MLP decode into 2D as ‘Our-2D-GS’ and denote ‘Ours’ if the primitives are generated into 3D. All experiments are conducted on NVIDIA A100 80G GPUs.

5.2. Results Analysis

Our evaluation covers diverse scenes, including aerial and street views, intricate and large-scale scenes, both synthetic

Scene	Held-out			View(+1m)			View(+1m 5°down)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3D-GS [9]	23.47	0.668	0.406	20.83	0.605	0.440	21.25	0.643	0.402
Scaffold-GS [19]	25.40	0.744	0.320	22.62	0.671	0.375	23.28	0.711	0.334
UC-GS [19]	25.95	0.763	0.291	23.52	0.702	0.340	24.15	0.741	0.298
Ours	25.35	0.757	0.280	25.46	0.760	0.280	25.37	0.761	0.278

Table 3. Quantitative comparison on UC-GS dataset [44]. The metrics are extracted from the original paper, in which all methods are trained for 900k iterations.

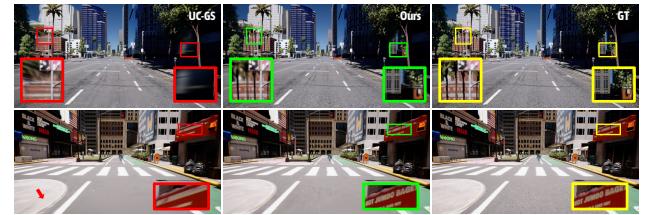


Figure 6. Qualitative comparisons of Horizon-GS against UC-GS [44]. The first row shows view shifting in the SF scene, while the second row shows view shifting and rotation in the NYC scene.

and real-world. We demonstrate that Horizon-GS preserves fine details and significantly improves the quality of surface reconstruction , as shown in Tab. 1, 3, 2 and Fig. 5, 6, 7.

Rendering Performance Analysis. Our method reconstructs a unified Gaussian model, delivering excellent visual quality in both aerial and street views. Compared to 2D-GS [8], Our-2D-GS consistently outperforms by 4 dB in aerial PSNR and 1.5 dB in street PSNR. Furthermore, compared to all the baselines, Horizon-GS exceeds the performance of all baselines, except Hierarchical-3DGS, which is competitive in street views due to its street-focused design. As shown in the highlighted patches and arrows above of Fig. 5 (a), our method excels in fine details (1st row) and reflective regions (2nd row). In the UC-GS setting (Tab. 3), under the held-out viewpoints condition where test views share the same heights as training views, Horizon-GS, trained for 100k iterations, shows slightly lower PSNR and SSIM values compared to UC-GS trained for 900k iterations. However, our approach produces richer details, as indicated by better LPIPS metrics. Moreover, when eval-

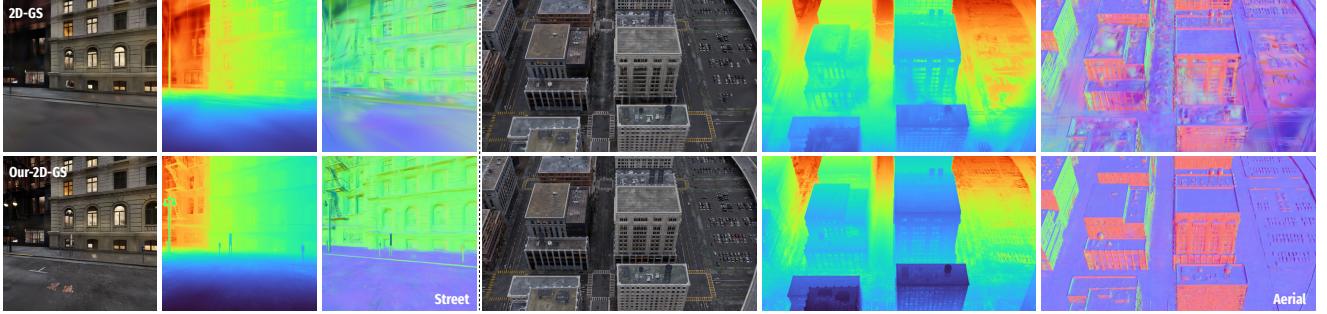


Figure 7. Qualitative comparisons of our method with 2D-GS [8]. We visualize the rendered images and geometry results (depth and world-space normal) in the Block_Small scene from both aerial and street views.

uating shifted and rotated views, our method consistently outperforms the baselines, as shown in Fig. 6.

Large-scale Rendering Performance. For the challenging large-scale scenes, we conduct two types of experiments: direct training without chunking and chunked training. As shown in Table 2, partitioning and then merging proves effective, especially when comparing Scaffold-GS and Scaffold-GS*. Horizon-GS consistently delivers excellent rendering quality in large-scale scenes compared to other baselines. In contrast, Hierarchical-3DGS [10] tends to overfit the training views, producing artifacts in novel views, while other baselines fail to capture fine-grained details (Figure 5 (b)). Each scene is divided into chunks and trained in parallel, with a total training time of about four hours per scene. Due to the LOD design, our method achieves real-time rendering speeds of 51.5 FPS in the Hillside_Morning and Hillside_Summer scenes, which are the largest scenes among our captured dataset.

Surface Reconstruction Performance. As shown in Fig. 7, our method reconstructs more complete and detailed geometry compared to 2D-GS [8]. 2D-GS produces artifacts, resulting in incomplete and lackluster geometry. In contrast, our method, using the two-stage training approach and enhanced densification, delivers detailed, geometrically accurate, and artifact-free reconstruction.

5.3. Ablation Studies

In this section, we ablate each individual module to validate their effectiveness. We use the scenes from our proposed real dataset, Road and Park. Quantitative results can be found in Tab. 4.

Balanced Camera Distribution. To assess the effectiveness of the balanced camera distribution (Sec. 3.3), we conduct an ablation study with randomly selected training views. The results in Tab. 4 reveal a notable decline in the visual quality of aerial views, while the quality of street views remains similar. This suggests that a well-distributed

Scene Method Metrics	Aerial			Street		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours w/o camera bal.	22.49	0.684	0.375	22.00	<u>0.671</u>	0.323
Ours w/o densify bal.	23.02	<u>0.721</u>	<u>0.332</u>	21.98	0.670	0.325
Ours w/o multi LOD	<u>23.05</u>	0.717	0.337	<u>22.03</u>	0.672	0.321
Ours w/o densify poli.	22.83	0.690	0.369	21.68	0.655	0.346
Ours	23.23	0.729	0.321	22.04	0.669	0.324

Table 4. Ablations of our method on our proposed real datasets.

set of observation perspectives enhances the model performance, rather than converging to a single-level view.

Balanced Densification. As outlined in Sec. 3.3, we meticulously design the densification strategy in the two training stages. Specifically, gradients of the Gaussian primitives are accumulated from aerial views, while in the second stage, they are derived solely from street views. To assess the effectiveness of this approach, we compute the gradients of the Gaussian primitives across all the images in both stages. The results reveal conflicts during the densification process, highlighting the need to explicitly separate the densification steps during optimization.

Multi-Resolution LOD. To assess the impact of LOD design, we conduct an experiment without LOD. The quality of aerial views significantly decreases, indicating that the strategy enhances the capture of fine-scale details.

Densification Policy. We perform an ablation of the densification policy (Eq. 1) by replacing it in the second stage with the one from Scaffold-GS. This leads to a significant reduction in rendering details for both aerial and street views, indicating that relying solely on average gradients for spawning Gaussian primitives is insufficient to capture local fine-grained details.

6. Conclusion

We introduce Horizon-GS, a novel framework for large-scale aerial-to-ground scene reconstruction. We first explore the challenges of unified scene reconstruction from

both aerial and street views, and propose a end-to-end coarse-to-fine training framework that mitigates inherent conflicts and is adaptable for large-scale reconstruction. Additionally, we present a comprehensive dataset with two cross views. Our experiments show that Horizon-GS outperforms existing methods in visual quality and geometry accuracy. In the future, we aim to minimize input dependency by leveraging advanced techniques such as pre-trained 3D scene foundation models and enable more systematic solutions, such as distributional training.

References

- [1] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017. 1
- [2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2
- [3] Yu Chen and Gim Hee Lee. Dogaussian: Distributed-oriented gaussian splatting for large-scale 3d reconstruction via gaussian consensus. *arXiv preprint arXiv:2405.13943*, 2024. 2
- [4] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv:2311.18561*, 2023. 2
- [5] David J Crandall, Andrew Owens, Noah Snavely, and Daniel P Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2841–2853, 2012. 2
- [6] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2
- [7] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 2
- [8] 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*, pages 1–11, 2024. 4, 5, 6, 7, 8, 1, 2
- [9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1, 2, 3, 6, 7
- [10] 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 Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 2, 4, 6, 7, 8, 1
- [11] Bingling Li, Shengyi Chen, Luchao Wang, Kaimin Liao, Si-jie Yan, and Yuanjun Xiong. Retinags: Scalable training for dense scene rendering with billion-scale 3d gaussians. *arXiv preprint arXiv:2406.11836*, 2024. 2
- [12] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhen-zhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023. 2, 6
- [13] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 2
- [14] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, You-liang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5166–5175, 2024. 2, 3, 5
- [15] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022. 2
- [16] Jeffrey Yunfan Liu, Yun Chen, Ze Yang, Jingkang Wang, Sivabalan Manivasagam, and Raquel Urtasun. Real-time neural rasterization for large scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8416–8427, 2023. 2
- [17] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2
- [18] Yang Liu, He Guan, Chuanchen Luo, Lue Fan, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. *arXiv preprint arXiv:2404.01133*, 2024. 2, 3
- [19] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. *arXiv preprint arXiv:2312.00109*, 2023. 3, 4, 5, 6, 7, 1, 2
- [20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2
- [21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2
- [22] OpenDroneMap. Skyremoval. <https://github.com/OpenDroneMap/SkyRemoval/>, 2022. 6
- [23] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022. 2
- [24] 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. *arXiv preprint arXiv:2403.17898*, 2024. 2, 4, 1
- [25] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kun-chang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 6
- [26] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 2
- [27] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2
- [28] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2
- [29] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 2
- [30] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12922–12931, 2022. 2
- [31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7, 1
- [32] Xuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15077–15087, 2024. 2
- [33] Yuanbo Xiangli, Lining Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *European conference on computer vision*, pages 106–122. Springer, 2022. 2, 4
- [34] Yuanbo Xiangli, Lining Xu, Xingang Pan, Nanxuan Zhao, Bo Dai, and Dahua Lin. Assetfield: Assets mining and re-configuration in ground feature plane representation. *arXiv preprint arXiv:2303.13953*, 2023. 2
- [35] Wenhui Xiao, Rodrigo Santa Cruz, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Leo Lebrat. Nerf director: Revisiting view selection in neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20742–20751, 2024. 4
- [36] Lining Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8306, 2023. 2
- [37] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 2
- [38] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 2
- [39] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jia Shi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 5, 6
- [40] Mulin Yu, Tao Lu, Lining Xu, Lihan Jiang, Yuanbo Xiangli, and Bo Dai. Gsdf: 3dgs meets sdf for improved rendering and reconstruction. *arXiv preprint arXiv:2403.16964*, 2024. 2
- [41] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv preprint arXiv:2404.10772*, 2024. 2
- [42] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025. 2
- [43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7, 1
- [44] 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. *arXiv preprint arXiv:2408.15242*, 2024. 2, 6, 7
- [45] Hexu Zhao, Haoyang Weng, Daohan Lu, Ang Li, Jinyang Li, Aurojit Panda, and Saining Xie. On scaling up 3d gaussian splatting training. *arXiv preprint arXiv:2406.18533*, 2024. 2
- [46] MI Zhenxing and Dan Xu. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *The Eleventh International Conference on Learning Representations*, 2022. 2
- [47] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21634–21643, 2024. 2

7. Data

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. 8(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 [10], as visualized in Fig. 8(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 [9], we resize the longest side original images to 1600 pixels.



Figure 8. (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. 5.

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 5. Camera rotation parameters in synthetic scenes.

8. More implementation

8.1. Global Appearance Embedding

In large-scale scenes, the data is typically captured in different environments, leading to inconsistent exposures. Inspired by Octree-GS [24] and Hierarchical-3DGS [10], 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.

Method	Metrics	Aerial			Street		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline [19]		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 6. Quantitative comparison using naive finetuning solutions.

8.2. Mesh Extraction

For mesh extraction, we adopt the 2D-GS[8] 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.

9. More Experiments

9.1. Naive Solutions

Based on the observations discussed in Section 3.1, 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. 6, this fine-tuning approach inefficient, time-consuming, and fails to address the core issue.

9.2. More Quantitative 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 [31], and LPIPS [43], as shown in Tables 7, 8, 9.

9.3. Ablation

We select Scaffold-GS [19] 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	City				Colosseum				Elevenruin			
	Aerial		Street		Aerial		Street		Aerial		Street	
Method	PSNR↑ SSIM↑ LPIPS↓											
2D-GS [8]	25.27	0.739	0.391	21.75	0.708	0.439	22.50	0.752	0.382	25.76	0.905	0.143
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
3D-GS [9]	26.79	0.784	0.351	21.79	0.723	0.422	22.25	0.754	0.380	25.30	0.910	0.132
Scaffold-GS [19]	30.03	0.890	0.187	23.98	0.796	0.334	25.14	0.854	0.226	25.33	0.867	0.187
Hier-GS [10]	29.15	0.871	0.206	24.51	0.810	0.298	23.67	0.805	0.314	25.74	0.915	0.129
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

Table 7. Quantitative comparison on each synthetic scene of our proposed dataset.

Scene	Road					
	Aerial		Street			
Method	PSNR↑ SSIM↑ LPIPS↓	PSNR↑ SSIM↑ LPIPS↓	PSNR↑ SSIM↑ LPIPS↓	PSNR↑ SSIM↑ LPIPS↓		
2D-GS [8]	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 [9]	19.95	0.509	0.562	20.17	0.573	0.435
Scaffold-GS [19]	20.36	0.532	0.532	20.08	0.580	0.422
Hier-GS [10]	21.22	0.620	0.432	21.30	0.651	0.312
Ours	22.60	0.682	0.356	20.94	0.637	0.341

Table 8. Quantitative comparison on Road scene.

Scene	Park					
	Aerial		Street			
Method	PSNR↑ SSIM↑ LPIPS↓	PSNR↑ SSIM↑ LPIPS↓	PSNR↑ SSIM↑ LPIPS↓	PSNR↑ SSIM↑ LPIPS↓		
2D-GS [8]	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 [9]	20.23	0.545	0.565	22.64	0.681	0.361
Scaffold-GS [19]	19.99	0.545	0.565	22.35	0.672	0.366
Hier-GS [10]	21.63	0.657	0.427	23.75	0.720	0.294
Ours	23.85	0.776	0.287	23.14	0.701	0.308

Table 9. Quantitative comparison on Park scene.

Scene	Aerial			Street		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline [19]	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	23.23	0.729	0.321	22.04	0.669	0.324

Table 10. 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. 10.

10. Limitation and More Discussion

While our method proves effective in reconstructing large-scale aerial-to-ground scenes and producing high-quality results, it also has certain limitations. First, similar to most Gaussian-based methods, Horizon-GS may reach sub-

optimal 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 [45] or RetinaGS [11], also presents a promising solution.