DPGS: A Diffusion-Prior Guided Framework for Large-Scale 3D Gaussian Splatting Reconstruction

Shidong Zhang[1], Shuaixin Li[2,3⋆], Juntong Qi[4], Haoxin Zhang[5], Xiao Zhang[6], Xiaozhou Zhu[2,3], and Wen Yao[2,3⋆]

[1] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China
[2] Defense Innovation Institute, Chinese Academy of Military Science, Beijing, China
[3] Intelligent Game and Decision Laboratory, Beijing, China
[4] School of Future Technology, Shanghai University, Shanghai 200444, China
[5] School of Systems Science and Engineering, Sun Yat-Sen University, Guangzhou, China
[6] Harbin Engineering University, Harbin, China

Abstract. High-fidelity reconstruction of large-scale aerial scenes using 3DGS confronts dual challenges: restricted nadir views result in facade "blind spots" and geometric collapse in textureless regions, while the massive scale of high-resolution data induces severe computational load imbalance during parallel training. In this paper, we propose DPGS, a holistic framework that orchestrates visual foundation model priors and generative diffusion priors to achieve robust reconstruction. First, we introduce a foundation model-driven pixel-wise dense initialization to mitigate geometric incompleteness from sparse SfM points, effectively preventing floating artifacts in weak-texture areas. Second, to address computational bottlenecks, we propose an observation density balanced partitioning method. By formulating a graph partitioning problem weighted by feature track lengths and observation frequency, we dynamically balance the rendering load across sub-blocks. Finally, we devise a diffusion-prior guided geometry completion mechanism. Leveraging a 3D-aware diffusion model with spiraling side-view sampling, we recover textures for unobserved facade blind spots. DPGS achieves state-of-the-art rendering quality, geometric completeness, and efficiency across standard benchmarks and a self-collected dataset. Project page: `https://zsddd.github.io/dpgs`.

Keywords: 3DGS, Large-Scale Aerial Scenes, Diffusion Priors, Load Balancing, Dense Initialization

1 Introduction

The large-scale, high-fidelity reconstruction of 3D environments is a cornerstone of modern digital twins applications, enabling critical tasks in smart city gov-

⋆ Corresponding authors.

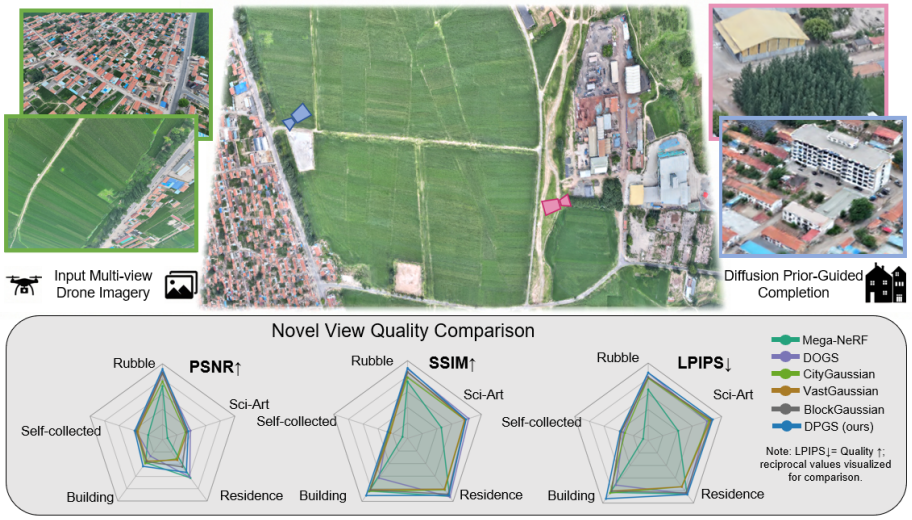

Fig. 1: Our method synthesizes high-quality scenes solely from nadir-view images, eliminating the need for additional 3D or street-level training data. Given multiple nadir inputs (top left), we leverage 3D Gaussian Splatting and a pre-trained text-to-image diffusion model within an iterative refinement framework to reconstruct photorealistic 3D scenes from limited observations (top right). Compared to existing methods, our approach achieves superior rendering fidelity across most scenarios (bottom).

ernance, urban planning and virtual reality [3, 7]. While traditional approaches, primarily based on Structure-from-Motion (SfM) and Multi-View Stereo (MVS), have long served as the industry standard for aerial mapping, they are often hindered by prohibitive computational costs and lengthy processing times.

In recent years, 3D Gaussian Splatting (3DGS) [16] has emerged as a transformative paradigm. By parameterizing the scene with anisotropic 3D Gaussian primitives and utilizing efficient differentiable rasterization, 3DGS achieves real-time rendering speeds with photorealistic quality, offering a promising alternative to implicit Neural Radiance Fields(NeRF) [29]. Its explicit point-based nature naturally scales to large environments, making it an ideal solution for next-generation aerial photogrammetry.

Nevertheless, directly applying 3DGS to large-scale aerial scenarios faces challenges stemming from the restricted observation conditions inherent to real-world flight missions. Existing "divide-and-conquer" strategies primarily focus on spatial partitioning to manage memory [23] often assuming ideal data acquisition with dense, multi-angle coverage with low-altitude (see Fig. 2 (a)), e.g., UrbanScene3D [24] and MatrixCity [21]. In contrast, practical large-area mapping relies on high-altitude nadir strip flights [31], which significantly limits parallax and provides insufficient coverage of vertical structures (e.g., building facades). As illustrated in Fig. 2 (b)-(d), these limitations lead to three critical bottlenecks.

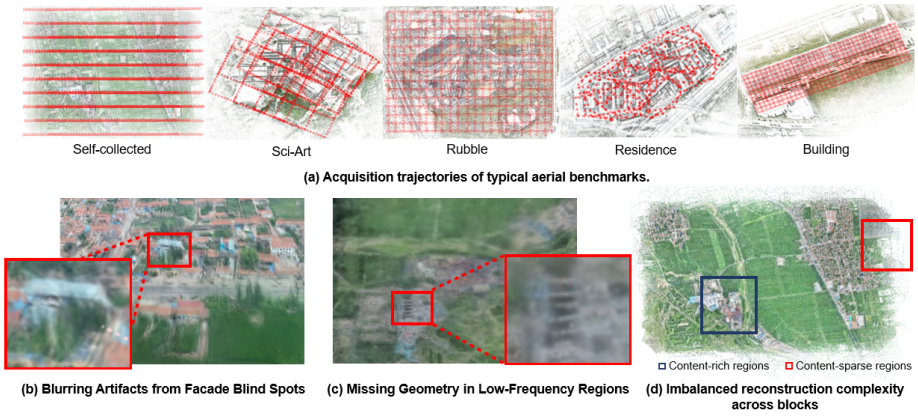

Fig. 2: Illustration of Fundamental Challenges in Practical Large-Scale Aerial 3DGS. (a) Acquisition Trajectory: Practical mapping often relies on sparse nadir strip flights (Self-collected), which contrasts sharply with the dense orbiting or grid-like camera trajectories (red markers) used in standard benchmark datasets. Scenes Sci-Art and Residence are from the UrbanScene3D dataset, whereas Rubble and Building originate from the Mill 19 dataset. This trajectory discrepancy introduces three critical bottlenecks: (b) Facade Blind Spots: The lack of supervision from oblique views leads to blurry artifacts and geometric collapse on vertical structures.(c) Missing Geometry in Low-Frequency Regions: Sparse SfM anchors result in missing geometric information in weakly textured areas.(d) Load Imbalance: Heterogeneous scene complexity leads to severe imbalances in computational load among partitioned sub-blocks.

- First, the lack of oblique views results in severe "blind spots" on vertical building facades, causing geometric collapse and blurring artifacts due to insufficient supervision.
- Second, weak textures in aerial imagery, e.g., roads and vegetation, lead to sparse triangulated point clouds, depriving 3DGS of the necessary initialization anchors and resulting in floating artifacts.
- Third, uniform spatial partitioning fails on heterogeneous aerial scenes. Since dense urban structures demand significantly more rendering resources than flat terrain, this mismatch causes severe load imbalance and GPU memory exhaustion during parallel training.

To address these limitations, we present DPGS, a diffusion-prior guided framework for large-scale 3DGS. Our core insight is that when physical observations are restricted, we must leverage external prior knowledge of the world to achieve robust 3D reconstruction.

To tackle the geometric sparsity issue, we first introduce a pixel-wise dense initialization strategy driven by dense matching foundation models, such as RoMa [8], UFM [44], and ASpanFormer [4]. This injects dense, high-fidelity anchors into the scene, ensuring robust convergence even in textureless regions. Next, to mitigate facade blind spots, we devise a diffusion-prior guided completion mechanism specifically tailored for structure-dense regions with insuf-

ficient coverage. By designing a virtual spiral descent trajectory, we employ a 3D-aware diffusion model to hallucinate and refine plausible textures for unobserved angles, effectively "growing" missing geometries under generative guidance. Finally, we move beyond rigid spatial partitioning strategies that overlook the intrinsic distribution of scene complexity. Instead, we introduce a Load-Balanced Adaptive Partitioning strategy. Recognizing that scene complexity is non-uniform, we formulate a graph partitioning problem weighted by feature track lengths—employing observation frequency as a proxy for rendering load—to dynamically balance the computational cost across sub-blocks. Experiments on challenging high-altitude aerial benchmarks demonstrate that DPGS delivers performance superior to competing methods in both rendering quality and geometric completeness.

Note that, it is crucial to address the trade-off between the generative nature of diffusion models and the metric rigor of traditional photogrammetry. While conventional mapping prioritizes pixel-wise fidelity, emerging applications like VR/AR and unmanned system perception demand structural completeness. Recent literature [2, 5, 35] corroborates that for downstream tasks such as autonomous navigation, topological continuity and geometric completeness are far more critical than photogrammetric-grade metric precision. A geometric "blind spot" causes catastrophic failures in machine perception, whereas a semantically consistent—albeit hallucinated—facade effectively closes the 3D geometry. Consequently, rather than pursuing strict photometric accuracy in observation-denied regions, our method emphasizes structural and semantic integrity. In summary, our main contributions are as follows:

1. We propose DPGS, a holistic framework tailored for large-scale aerial reconstruction that effectively overcomes the geometric and textural deficiencies caused by restricted nadir observations.
2. We integrate Visual Foundation Model and Generative Diffusion Priors into the pipeline: utilizing dense matching foundation models, e.g., RoMa, ASpanFormer, UFM for dense geometric initialization and a diffusion-guided completion module to restore missing facade details, achieving state-of-the-art performance on both standard benchmarks and challenging real-world aerial datasets.
3. We introduce an Observation Density Balanced Partitioning (ODBP) strategy, which dynamically adjusts block granularity based on a novel rendering load proxy (track length), ensuring efficient parallel training for heterogeneous scenes.

2 Related Work

2.1 Novel View Synthesis and Scene Generation

Novel View Synthesis. The emergence of NeRF has thoroughly revolutionized the field of novel view synthesis by utilizing coordinate-based Multilayer Perceptrons

(MLPs) for implicit scene representation [1, 30]. However, the heavy computational burden imposed by volumetric ray-marching prompted a paradigm shift towards explicit representations. As a milestone, 3DGS introduced anisotropic Gaussian primitives and efficient differentiable rasterization techniques, achieving real-time rendering without sacrificing visual fidelity. Building upon this foundation, recent studies have actively expanded its capabilities: Mip-Splatting [43] and AbsGS [41] mitigate aliasing artifacts; SuGaR [12] applies geometric regularization for precise mesh extraction; and LightGaussian [10] addresses storage constraints via pruning and vector quantization. Although these methods perform exceptionally well in bounded local scenes, they face severe scalability bottlenecks. When directly applied to unbounded large-scale environments, they often fail to balance rendering quality and GPU memory load.

Scene Generation and Diffusion Priors. Diffusion models [33], initially the cornerstone of 2D generation, have profoundly reshaped the landscape of 3D content creation. Pioneering works such as DreamFusion [32] and Magic3D [22] established the "text-to-3D" generative paradigm via Score Distillation Sampling (SDS). In novel view synthesis tasks based on sparse observations, leveraging generative priors has become a mainstream trend. Strategies include employing diffusion models as 3D-aware scorers [26,37] or generating "pseudo-observations" for data augmentation [27]. Recent Multi-View Diffusion (MVD) models (e.g., EscherNet [18], Cat3D [11], ViewCrafter [42]) demonstrate robust capabilities in synthesizing plausible novel views. Furthermore, the rise of generative world models in autonomous driving [14,15,38] shows that agents can effectively navigate within hallucinated but semantically consistent environments. Nevertheless, applying these generative priors to large-scale geometrically rigorous mapping pipelines remains a challenge. Domain-specific works such as Skyfall-GS [20] successfully recover missing facades for satellite imagery, but the generated scenes lack realism. Difix3D [39] is limited to small-scale geometric refinement. Integrating diffusion priors into a unified framework for geometrically consistent, large-scale reconstruction remains an unresolved challenge.

2.2 Large-Scale Reconstruction

Traditionally, large-scale 3D environment reconstruction has relied heavily on SfM and MVS. These classical photogrammetry pipelines extract and match sparse features to estimate camera poses, followed by dense depth estimation to fuse point clouds and textured meshes [13, 36]. While they have long served as the industry standard, it is well known that they are hindered by prohibitive computational costs and lengthy processing times. Moreover, they heavily rely on high-contrast textures, often failing to effectively reconstruct textureless regions or highly specular surfaces.

Recent large-scale reconstruction frameworks universally adopt a "divide-and-conquer" paradigm. Specifically, VastGaussian [23] and DOGS [6] partition scenes based on camera distribution; Switch-NeRF [45] learns decomposition via sparse NeRFs but suffers from slow rendering; CityGaussian [28] employs

a global coarse model for guidance; and BlockGaussian [40] utilizes sparse SfM point density for partitioning.

However, these existing partition-based methods fail to concurrently address the load imbalance and observation constraints (e.g., facade blind spots and sparse initial anchors) inherent to high-altitude nadir flights, which motivates our proposed DPGS framework.

3 Method

The overall pipeline of our proposed DPGS is illustrated in Fig. 3. To achieve high-fidelity reconstruction under restricted observations, we integrate visual and generative priors, supported by an efficient partitioning strategy: First, we implement Dense Co-visibility Geometric Initialization (Sec. 3.1). Addressing geometric incompleteness in weak-texture regions, we utilize the visual foundation model, e.g., RoMa, UFM, ASpanFormer, to propagate dense geometric anchors, effectively preventing floating artifacts. Second, we introduce Diffusion Prior-Guided Geometry Completion (Sec. 3.2). To resolve perspective deficiencies on vertical facades, we employ a 3D-aware diffusion model along a spiral trajectory to hallucinate plausible textures, providing pseudo-supervision for blind spots. Finally, to mitigate the computational burden of these dense priors on large-scale scenes, we propose ODBP (Sec. 3.3). Unlike rigid spatial grids, we formulate a graph partitioning problem weighted by feature track lengths to dynamically balance the rendering load, ensuring efficient parallel training.

3.1 Dense Co-visibility Geometric Initialization

Standard 3DGS relies on sparse SfM points, which often fail in textureless aerial scenes (e.g., roads, rivers), causing floating artifacts. To mitigate this, we introduce a dense initialization framework utilizing the RoMa foundation model.

Dense Geometric Hypothesis. For co-visible image pairs (I_A, I_B), we employ RoMa to predict a dense correspondence field $\mathcal{W}_{A\to B}$ and a certainty map $\mathcal{C}_{A\to B}$. A pixel $\mathbf{u}_A \in I_A$ matches $\mathbf{u}_B = \mathbf{u}_A + \mathcal{W}_{A\to B}(\mathbf{u}_A)$. After removing low-confidence matches ($\mathcal{C}_{A\to B}(\mathbf{u}_A) < \tau_c$), we triangulate the remaining 2D correspondences into an initial dense 3D point cloud $\mathbf{P}_{\text{dense}}$.

Adaptive Geometric Filtering. Raw dense matching inevitably introduces noisy outliers in occluded regions. To distill high-fidelity anchors, we define a joint reliability score $S(\mathbf{p})$ for each triangulated candidate $\mathbf{p} \in \mathbf{P}_{\text{dense}}$. This fuses the network-predicted certainty with the multi-view reprojection error:

$$S(\mathbf{p}) = \mathcal{C}_{A\to B}(\mathbf{u}_A) \cdot \exp\left(-\frac{\|\pi_A(\mathbf{p}) - \mathbf{u}_A\|_2^2 + \|\pi_B(\mathbf{p}) - \mathbf{u}_B\|_2^2}{2\sigma^2}\right) \tag{1}$$

where $\pi(\cdot)$ denotes the perspective projection and σ controls geometric sensitivity. We prune points with $S(\mathbf{p}) < \tau_s$ and initialize the 3D Gaussians at the surviving locations ($\mu_i = \mathbf{p}_i$). This provides reliable geometric anchors, effectively preventing overfitting in texture-poor areas.

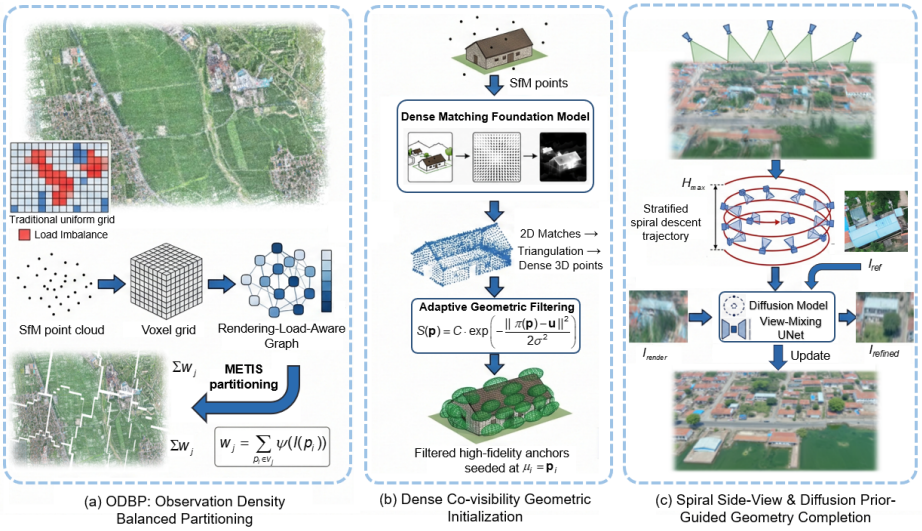

Fig. 3: Overview of the DPGS framework. (a) ODBP: The scene is segmented into computationally balanced sub-blocks via rendering-load-aware graph partitioning. (b) Dense Initialization: A foundation model and adaptive filtering are used to generate robust geometric anchors, mitigating SfM sparsity. (c) Diffusion-Guided Completion: A spiral trajectory exposes facade blind spots, which are rendered and enhanced by a diffusion model to provide pseudo-supervision ($\mathcal{L}_{side}$) for recovering missing geometry.

3.2 Diffusion Prior-Guided Geometry Completion

In high-altitude nadir mapping, the complete absence of oblique observations leads to a fundamental "blind spot" for vertical structures. Standard 3DGS optimization, lacking gradient supervision on facades, inevitably degenerates into geometric collapse or "floating" artifacts. To address this, we propose a Diffusion Completion strategy driven by generative priors. Unlike traditional inpainting which operates in 2D, we leverage a 3D-consistent diffusion framework to "hallucinate" plausible facade geometry and texture, prioritizing structural integrity for downstream machine perception tasks over strict pixel-wise ground truth.

Stratified Spiral Side-View Sampling. To expose the blind spots, we first construct a virtual camera trajectory $\mathcal{T}_{side}$. Simply placing lateral cameras is insufficient; the transition from observed roofs to unobserved facades must be smooth to ensure stable Gaussian optimization. We design a stratified spiral descent trajectory centered on the scene origin c. As illustrated in Fig. 3(c), we define L distinct spiral layers extending from the flight altitude H_{max} down to the ground. For the k-th view in the l-th layer, the camera pose $P_{l,k}$ is computed to ensure comprehensive coverage of the facade while maintaining overlap with the known roof texture:

$$\mathbf{P}_{l,k} = \text{LookAt}(\mathbf{c} + \Delta h_l \cdot \mathbf{u}_z, \mathbf{o}_{l,k}, \mathbf{u}_{up}) \tag{2}$$

where $\mathbf{o}_{l,k}$ denotes the orbital position and Δh_l is a height-dependent look-at offset.

Reference-Conditioned Hallucination. Images rendered from these virtual viewpoints, denoted as I_{render}, initially exhibit severe artifacts (holes and blur). To restore them, we employ the SD-Turbo [34] diffusion model ($\mathcal{M}_\theta$) for its efficiency in single-step inference. To better adapt the model to the specific distribution of high-altitude perspectives, we fine-tuned SD-Turbo using high-resolution images from the real aerial dataset as ground truth, paired with their corresponding degraded initial 3DGS renderings as input. Crucially, to ensure semantic and stylistic consistency—preventing the generation of textures that contradict the real scene (e.g., generating a glass wall for a brick building)—we condition the generation on the nearest available real training view, I_{ref}. We incorporate a View-Mixing mechanism within the UNet architecture to propagate texture cues from the reference (roof) to the target (facade). The refined image $I_{refined}$ is obtained as:

$$I_{refined} = \mathcal{M}_\theta(I_{render}, I_{ref}; \text{prompt}) \tag{3}$$

Here, $I_{refined}$ serves as a high-quality pseudo-label.

Side-View Supervision. We incorporate these hallucinated views into the 3DGS optimization via a dedicated Side-View Loss ($\mathcal{L}_{side}$). This loss is applied periodically to guide the Gaussians to "grow" into the blind spots:

$$\mathcal{L}_{side} = \lambda_{L1}\|I_{render} - I_{refined}\|_1 + \lambda_{SSIM}(1 - \text{SSIM}(I_{render}, I_{refined})) \tag{4}$$

By minimizing $\mathcal{L}_{side}$, the geometry is forced to conform to the plausible structure predicted by the diffusion model, effectively closing the mesh surfaces and eliminating the geometric collapse typical of nadir-only reconstruction.

3.3 Observation Density Balanced Partitioning

The misalignment between scene geometry and camera distribution causes severe computational load imbalance in large-scale reconstruction. In differentiable rasterization pipelines, existing spatial-partitioning methods relying on point density neglect visibility variance: a point observed by hundreds of cameras imposes a much higher rendering load than an occluded one. To address this, we propose ODBP (Fig. 4), formulating scene segmentation as a rendering-load-aware graph partitioning problem.

Rendering-Load-Aware Graph Formulation. Let $\mathcal{P}$ denote the global SfM point cloud, where the visibility complexity $l(p_i)$ of point p_i is defined as its number of visible views. We voxelize the scene and construct a 6-connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The rendering load of voxel v_j is defined as

$$w_j = \sum_{p_i \in \text{Points}(v_j)} l(p_i). \tag{5}$$

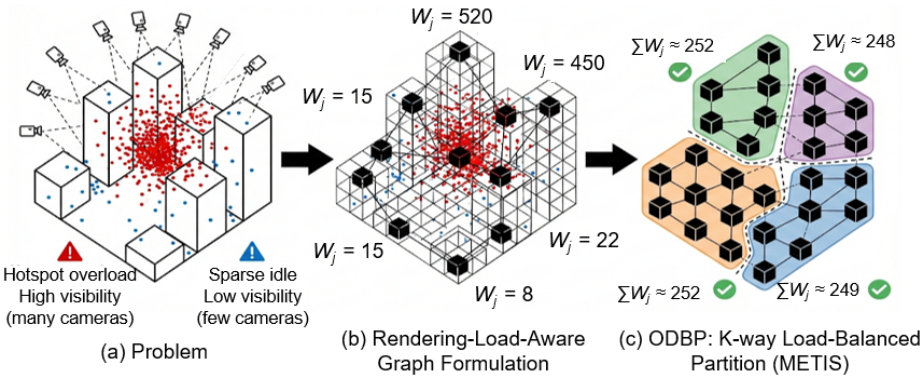

Fig. 4: Overview of ODBP. (a) Traditional spatial partitioning suffers from severe load imbalance due to highly visible "hotspots" (red) and occluded sparse regions (blue). **(b)** We model the scene as a Rendering-Load-Aware graph, where each voxel's weight W_j reflects its cumulative visibility complexity rather than mere point density. **(c)** By applying METIS partitioning constrained by these weights, we achieve a balanced computational load ($\sum W_j$) across all sub-blocks, dynamically adapting spatial boundaries to rendering demands.

We partition $\mathcal{G}$ into K subgraphs by minimizing the edge-cut cost while balancing the rendering load:

$$\min_{\Pi} \sum_{(u,v)\in\mathcal{E}} \mathbb{I}\big[\pi(u) \neq \pi(v)\big], \quad \text{s.t.} \quad (1-\epsilon)\bar{W} \leq \sum_{v_j \in V_k} w_j \leq (1+\epsilon)\bar{W}, \tag{6}$$

where $\bar{W}$ is the average rendering load and $\epsilon = 0.05$. We employ METIS to solve this graph partitioning problem efficiently, thereby merging sparse regions and splitting dense rendering hotspots for balanced parallel rendering.

Context-Aware Auxiliary Supervision. Rigid partitioning introduces gradient noise at boundaries due to Gaussian truncation. Following [40], we employ a Context-Aware Auxiliary Supervision strategy. For a sub-block k with core bounding box $\mathcal{B}^k_{core} = [\mathbf{c}_{min}, \mathbf{c}_{max}]$, we geometrically extend its boundaries by a ratio $\alpha = 0.2$ to define an expanded auxiliary volume $\mathcal{B}^k_{aux}$:

$$\mathcal{B}^k_{aux} = \left[\mathbf{c}_{min} - \frac{\alpha}{2}(\mathbf{c}_{max} - \mathbf{c}_{min}), \mathbf{c}_{max} + \frac{\alpha}{2}(\mathbf{c}_{max} - \mathbf{c}_{min})\right] \tag{7}$$

Points in the difference set $\mathcal{S}_{aux} = \text{Points}(\mathcal{B}^k_{aux}) \setminus \text{Points}(\mathcal{B}^k_{core})$ act as Auxiliary Anchors. Initialized as "Ghost Gaussians," they participate in the forward rasterization to provide correct occlusion cues and ensure boundary photometric consistency for the composited color C:

$$C = \sum_{i\in\mathcal{N}_{core}} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) + \sum_{k\in\mathcal{N}_{aux}} c_k\alpha_k \prod_{j=1}^{k-1}(1-\alpha_j) \tag{8}$$

To prevent memory inflation, a Dual-Mode Update strategy is applied: the structural parameters of auxiliary Gaussians are frozen and their densification is disabled, ensuring contextual supervision without creating redundant primitives.

4 Evaluation

Implementation Details. Our framework is implemented in PyTorch using an NVIDIA RTX 4090 GPU. We use ODBP to discretize the scene into K sub-blocks, each initialized with RoMa dense priors. Each block is optimized independently for 40k iterations. Following a coarse-to-fine paradigm, the diffusion-guided Side Optimization is introduced at the 20k-th iteration, while adaptive densification aligns with standard 3DGS.

Datasets. We evaluate DPGS on four standard benchmarks (Rubble and Building from Mill 19; Sci-Art and Residence from UrbanScene3D) and a challenging self-collected dataset. The self-collected scene comprises 2,282 high resolution (4032×3024) images captured via high-altitude nadir strip flights. It features severe challenges like extensive vegetation, lack of facade observations, and weak textures, serving as a rigorous testbed for blind-spot completion.

Evaluation Metrics. The experimental setup follows the data partition specifications of Mega-NeRF, with images downsampled by a factor of 4 to balance the computational load, consistent with mainstream approaches. Quantitative evaluation covers four aspects: PSNR based on pixel differences measures signal fidelity; SSIM focuses on structural and luminance correlation; LPIPS, based on deep features, aims to assess the degree to which texture details fit human visual perception; and FID is employed to evaluate the overall generative quality and realism of the synthesized images.

4.1 Comparison with State-of-the-Arts Methods.

To evaluate the effectiveness of our proposed DPGS, we compare it against leading large-scale novel view synthesis methods, including the NeRF-based method (Mega-NeRF) and state-of-the-art 3DGS-based frameworks (DOGS, CityGaussian, VastGaussian, Momentum-GS [9], MixGS [25], and BlockGaussian). To ensure rigorous comparison, all baseline methods were trained and evaluated on our local hardware environment using identical configurations. The comparative experiments are conducted across diverse scene typologies, ranging from dense urban buildings to textureless rubble, as well as our self-collected aerial dataset.

Quantitative Results. As summarized in Tab. 1, our DPGS achieves SOTA performance across the majority of evaluated scenes. Specifically, in complex and challenging scenarios (such as Rubble, Residence, and the Self-collected dataset), DPGS consistently outperforms all baseline methods in PSNR, SSIM, and LPIPS metrics. Although some baseline methods (e.g., Mega-NeRF and DOGS) show competitive performance in specific constrained scenes (e.g., Sci-Art), DPGS overall demonstrates superior robustness and generalization capabilities. Notably, the significant improvement in the LPIPS metric indicates that our method synthesizes details with higher perceptual fidelity, which validates the effectiveness of integrating visual and generative priors into the reconstruction pipeline. To further validate the capability of our method under extreme conditions, as shown in Tab. 2, we conducted a targeted evaluation of DPGS against recently proposed baseline methods (Momentum-GS and MixGS) on

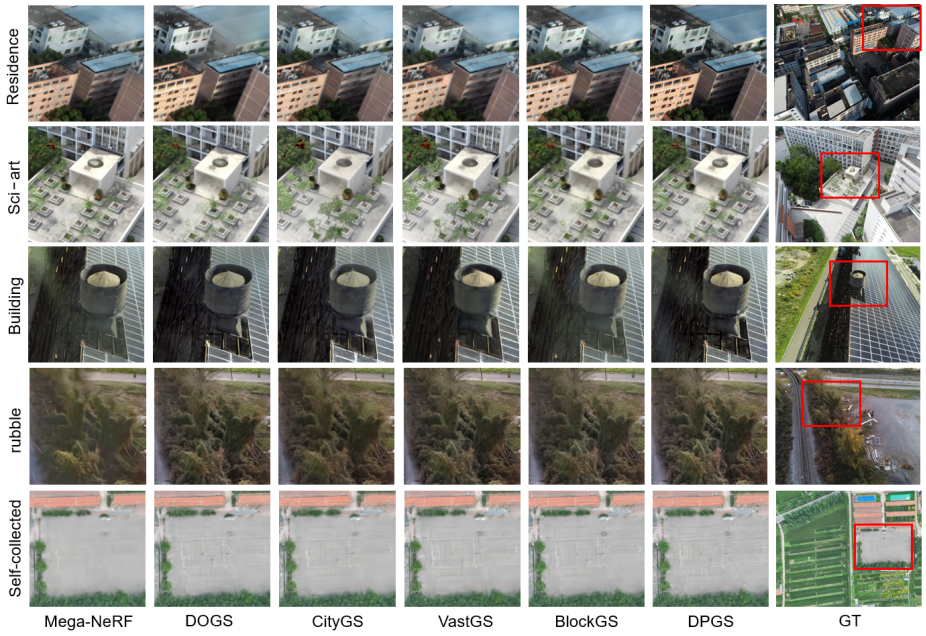

Fig. 5: Qualitative comparisons on standard benchmarks and our self-collected dataset. The rightmost column shows the GT. Red boxes mark the zoomed-in regions. DPGS gives sharper textures and fewer floating artifacts than other methods.

the ultra-large-scale MatrixCity [21] dataset and our self-collected dataset. The results demonstrate that DPGS achieves the best reconstruction quality.

Qualitative Results. As shown in Fig. 5 and Fig. 6, DPGS visually outperforms the baselines. Relying solely on sparse SfM points, methods like VastGaussian and CityGaussian suffer from floating artifacts and geometric collapse in textureless regions (e.g., the Self-collected dataset). Furthermore, insufficient oblique observations cause them to fail on vertical facades. Our method resolves these issues: the Dense Co-visibility Geometric Initialization explicitly anchors weak-texture areas, while the Diffusion Prior-Guided Geometry Completion recovers missing structures. Consequently, DPGS renders sharp, continuous facades that closely match the ground truth.

Table 1: Quantitative comparison across various scene typologies. We evaluate DPGS on four standard benchmarks and one self-collected dataset. ↑ denotes higher values are better, while ↓ denotes lower values are better.

Method	Rubble			Building			Sci-Art			Residence			Self-collected		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Mega-NeRF	23.71	0.707	0.321	19.07	0.564	0.421	23.12	0.807	0.242	21.31	0.762	0.252	20.01	0.361	0.556
DOGS	25.12	0.792	0.234	21.38	0.758	0.263	23.02	0.812	0.243	20.53	0.654	0.299	21.20	0.462	0.449
CityGaussian	24.18	0.735	0.262	20.89	0.731	0.252	20.85	0.752	0.289	21.39	0.765	0.248	21.14	0.454	0.457
VastGaussian	24.96	0.772	0.257	21.08	0.730	0.244	20.99	0.747	0.288	21.11	0.752	0.261	21.06	0.447	0.436
BlockGaussian	24.97	0.771	0.257	21.00	0.739	0.240	21.81	0.789	0.250	21.17	0.753	0.261	21.25	0.463	0.438
DPGS (Ours)	25.36	0.798	0.234	21.17	0.740	0.233	22.46	0.798	0.239	21.74	0.800	0.212	21.36	0.467	0.431

Fig. 6: Further qualitative comparison on the self-collected dataset. The rightmost column displays the GT images. Our method outperforms all baselines in both geometric accuracy and texture quality from low-altitude novel views. Notably, our approach correctly preserves distinctive side-view features—such as buildings and trees—that other methods fail to capture.

Table 2: Quantitative comparison against recent SOTA baselines. We evaluate DPGS against recently proposed baseline methods (Momentum-GS and MixGS) on the ultra-large-scale MatrixCity dataset and our self-collected dataset.

Method	MatrixCity			Self-collected		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Momentum-GS	29.11	0.881	0.180	21.05	0.455	0.476
MixGS	-	-	-	20.13	0.389	0.587
DPGS (Ours)	29.80	0.895	0.115	21.36	0.467	0.431

4.2 Ablation Studies

To verify the effectiveness of the proposed DPGS framework, we conduct extensive ablation experiments to evaluate the individual contribution of each core component, including the ODBP, dense co-visibility geometric initialization and diffusion prior-guided geometry completion.

Observation Density Balanced Partitioning. Fig. 7 and Tab. 3 show the partitioning results and corresponding statistics of ODBP. The method adjusts partition boundaries according to local observation distributions, subdividing densely observed regions while merging sparse ones. Although the point counts N_{pts} re-

Table 3: Analysis of ODBP. N_{blk} denotes the total number of partitioned subscenes. N_{pts} and N_{views} denote points (10^6) and cameras per block. L_{track} is the mean track length. W_{blk} represents the balanced Rendering Load Proxy weight (10^6) achieved by ODBP.

Dataset	$N_{\mathbf{blk}}$	$N_{\mathbf{pts}}$		$N_{\mathbf{views}}$		$L_{\mathbf{track}}$	$W_{\mathbf{blk}}$ **(Target)**
		mean	max	mean	max		consistency
Rubble	8	0.44	0.65	327	569	5.6	1.10 ± 0.08
Building	7	0.33	0.42	326	451	6.8	2.15 ± 0.05
Sci-Art	6	0.82	0.96	836	1032	7.2	2.17 ± 0.06
Residence	4	0.65	0.72	410	520	3.8	1.25 ± 0.04
Self-collected	8	0.35	0.43	466	587	3.4	0.60 ± 0.01

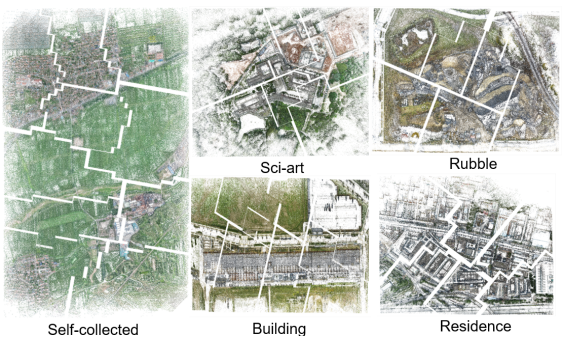

Fig. 7: Visualization of partitioning results across various datasets. White boundaries adaptively adjust to the rendering load: subdividing complex regions and aggregating sparse ones.

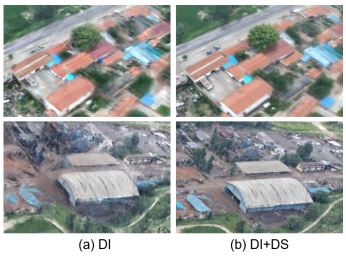

Fig. 8: Visual ablation of the diffusion prior optimization. By progressively incorporating challenging viewpoints, our method recovers fine-grained geometry in occluded areas.

main within a comparable range across blocks, the mean and maximum values of N_{views} still differ noticeably, especially in the Rubble scene. This indicates that point count alone cannot fully reflect rendering cost. We therefore introduce the mean track length L_{track} to characterize visibility complexity and refine the partition boundaries accordingly. The resulting rendering load weights W_{blk} exhibit only small variations across blocks.

Comparison of Dense Matching Models. We evaluate various dense matching methods $\mathcal{M}$ for initializing our splats. Throughout this paper, we use RoMa as our primary matching algorithm, but we also experiment with ASpanFormer [4] and UFM [44]. While RoMa yields superior performance (PSNR 21.43), ASpanFormer struggles due to its adaptive attention mechanism prone to drifting in repetitive aerial textures, where ambiguity is high. In addition to specialist matchers, we evaluate the generalist foundation model UFM. However, it remains less effective than RoMa in capturing fine-grained structural details. See Tab. 4 for a detailed comparison of performance.

Comparison of Diffusion Models. We evaluate various diffusion-based methods for image optimization, as illustrated in Tab. 6. FID evaluation images were obtained by extracting low-altitude rendering frames (2300 images in total). In this paper, we primarily utilize SD-Turbo as our core diffusion model, while also experimenting with DiffusionSat and FlowEdit. Although SD-Turbo and FlowEdit achieve comparable quantitative performance, FlowEdit fails to incorporate adjacent reference views, which occasionally leads to partial hallucinations when generating details. Furthermore, DiffusionSat struggles in our task, primarily because its architecture and priors are specifically designed for satellite imagery, making it less suitable for low-altitude or close-range novel view reconstruction. Fig. 9 showcases the images directly output by SD-Turbo.

Dense Initialization and Diffusion Supervision. As ablated in Tab. 7, DI anchors structures to improve PSNR (21.43) and SSIM (0.474). As visually compared in Fig. 8, DS further recovers fine-grained facade textures, achieving the best LPIPS (0.415) and FID (43.53). Although hallucinated details from DS

Fig. 9: Visual comparison of unobserved facade generation. The top row shows the top-down aerial images used for training. The middle row illustrates the novel side-view rendering results without diffusion guidance, while the bottom row demonstrates the results directly generated by SD-Turbo.

Table 4: Quantitative Comparison of Matching Models. We evaluate the impact of different dense initialization priors on the final 3D reconstruction quality. Our approach utilizing the RoMa foundation model provides robust geometric anchors, achieving the best fidelity across all metrics.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
ASpanFormer	21.38	0.458	0.441
UFM	21.30	0.452	0.459
Ours (RoMa)	21.43	0.474	0.436

Table 5: 3D Geometry Evaluation. Comparison of point clouds extracted from unobserved facade blind spots against the complete COLMAP ground truth. DPGS effectively recovers missing geometries, achieving the best Accuracy and Completeness.

Method	Accuracy ↓	Completeness ↑
BlockGaussian	32.34	32.85
Momentum-GS	30.67	32.31
MixGS	35.35	25.53
DPGS (Ours)	24.69	49.41

cause minor pixel-wise drops (e.g., PSNR: 21.34), their joint use successfully balances structural completeness with generative realism.

Effectiveness of the Proposed Framework on Geometric Completeness. We convert the Gaussian model trained only on Sci-Art top-down views into point clouds to compare with the COLMAP ground truth. As shown in Tab. 5 and Fig. 10, DPGS achieves the best Completeness score (49.41) and the lowest Accuracy error, outperforming all baselines and proving its effectiveness in reconstructing missing geometry in blind spots. Furthermore, Fig. 9 isolates the image refinement effect of the diffusion module on the self-collected and Sci-Art scenes. We directly refine degraded side-view 3DGS renderings, obtained from

Table 6: Quantitative Comparison of Diffusion Models. We evaluate the performance of the diffusion prior (SD-Turbo) against other generative methods. It consistently achieves the best performance.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DiffusionSat [17]	20.83	0.406	0.481	82.36
FlowEdit [19]	21.29	0.462	0.423	58.20
Ours (SD-Turbo)	21.34	0.465	0.415	43.53

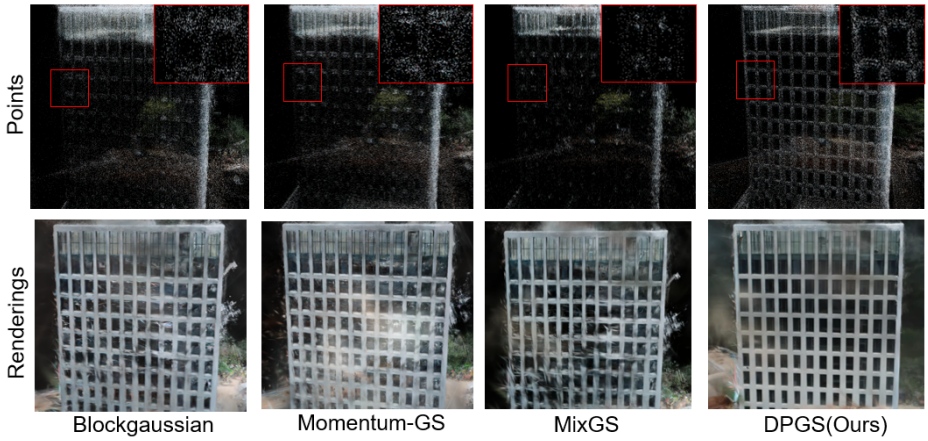

Fig. 10: 3D geometry comparison. Top (point clouds): Baselines leave severe holes in blind spots (red boxes). Bottom (renderings): DPGS successfully reconstructs coherent 3D structures and high-fidelity facades.

Table 7: Ablation study of Dense Initialization (DI) and Diffusion Supervision (DS) on the self-collected dataset.

DI	DS	PSNR↑	SSIM↑	LPIPS↓	FID↓
–	–	21.15	0.458	0.441	72.37
–	✓	21.27	0.453	0.428	48.00
✓	–	21.43	0.474	0.436	63.34
✓	✓	21.34	0.465	0.415	43.53

models trained with sparse top-down views, using SD-Turbo. Compared with the raw renderings, SD-Turbo suppresses blur and floating artifacts while improving edge sharpness and facade appearance consistency.

5 Conclusion

We present DPGS, a novel framework designed to overcome geometric and textural degradation in large-scale high-altitude 3DGS reconstruction. By integrating an adaptive ODBP strategy, RoMa-based dense initialization, and diffusion-guided facade completion, DPGS effectively resolves the challenges of load imbalance, weakly textured regions, and sparse nadir observations. Extensive evaluations demonstrate its SOTA performance in both geometric integrity and perceptual quality across public benchmarks and a real-world dataset. This work effectively bridges generative priors with rigorous 3D mapping, providing a strong foundation for future exploration into dynamic large-scale scene reconstruction.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021)
2. Beyer, L.L., Karaman, S.: Joint localization and planning using diffusion. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 15603–15609. IEEE (2025)
3. Bozcan, I., Kayacan, E.: Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 8504–8510. IEEE (2020)
4. Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., McKinnon, D., Tsin, Y., Quan, L.: Aspanformer: Detector-free image matching with adaptive span transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 20–36. Springer Nature Switzerland, Cham (2022)
5. Chen, K., Chen, J.K., Chuang, J., Vázquez, M., Savarese, S.: Topological planning with transformers for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11276–11286 (2021)
6. Chen, Y., Lee, G.H.: Dogs: Distributed-oriented gaussian splatting for large-scale 3d reconstruction via gaussian consensus. Advances in Neural Information Processing Systems **37**, 34487–34512 (2024)
7. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 370–386 (2018)
8. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19790–19800 (2024)
9. Fan, J., Li, W., Han, Y., Tang, Y.: Momentum-gs: Momentum gaussian self-distillation for high-quality large scene reconstruction. arXiv preprint arXiv:2412.04887 (2024)
10. Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., Wang, Z., et al.: Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. Advances in neural information processing systems **37**, 140138–140158 (2024)
11. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
12. Guédon, A., Lepetit, V.: 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. pp. 5354–5363 (2024)
13. Han, L., Gu, S., Zhong, D., Quan, S., Fang, L.: Real-time globally consistent dense 3d reconstruction with online texturing. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1519–1533 (2020)
14. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)

15. Jia, F., Mao, W., Liu, Y., Zhao, Y., Wen, Y., Zhang, C., Zhang, X., Wang, T.: Adriver-i: A general world model for autonomous driving. arXiv preprint arXiv:2311.13549 (2023)
16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
17. Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D., Ermon, S.: Diffusionsat: A generative foundation model for satellite imagery. arXiv preprint arXiv:2312.03606 (2023)
18. Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J.: Eschernet: A generative model for scalable view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9503–9513 (2024)
19. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
20. Lee, J.Y., Liu, Y.R., Tsai, S.R., Chang, W.C., Wu, C.H., Chan, J., Zhao, Z., Lin, C.H., Liu, Y.L.: Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. arXiv preprint arXiv:2510.15869 (2025)
21. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3205–3215 (2023)
22. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023)
23. Lin, J., Li, Z., Tang, X., Liu, J., Liu, S., Liu, J., Lu, Y., Wu, X., Xu, S., Yan, Y., et al.: Vastgaussian: Vast 3d gaussians for large scene reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5166–5175 (2024)
24. Lin, L., Liu, Y., Hu, Y., Yan, X., Xie, K., Huang, H.: Capturing, reconstructing, and simulating: the urbanscene3d dataset. In: European Conference on Computer Vision. pp. 93–109. Springer (2022)
25. Liu, C., Wang, H., Yu, L., Xia, G.S.: Holistic large-scale scene reconstruction via mixed gaussian splatting. arXiv preprint arXiv:2505.23280 (2025)
26. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9298–9309 (2023)
27. Liu, X., Chen, J., Kao, S.h., Tai, Y.W., Tang, C.K.: Deceptive-nerf/3dgs: Diffusion-generated pseudo-observations for high-quality sparse-view reconstruction. In: European Conference on Computer Vision. pp. 337–355. Springer (2024)
28. Liu, Y., Luo, C., Fan, L., Wang, N., Peng, J., Zhang, Z.: Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In: European Conference on Computer Vision. pp. 265–282. Springer (2024)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
30. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)

31. Nex, F., Remondino, F.: Uav for 3d mapping applications: a review. Applied geomatics **6**(1), 1–15 (2014)
32. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
35. Schmid, L., Pantic, M., Khanna, R., Ott, L., Siegwart, R., Nieto, J.: An efficient sampling-based method for online informative path planning in unknown environments. IEEE Robotics and Automation Letters **5**(2), 1500–1507 (2020)
36. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
37. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
38. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
39. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26024–26035 (2025)
40. Wu, Y., Qi, Z., Shi, Z., Zou, Z.: Blockgaussian: Efficient large-scale scene novel view synthesis via adaptive block-based gaussian splatting. arXiv preprint arXiv:2504.09048 (2025)
41. Ye, Z., Li, W., Liu, S., Qiao, P., Dou, Y.: Absgs: Recovering fine details in 3d gaussian splatting. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1053–1061 (2024)
42. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
43. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19447–19456 (2024)
44. Zhang, Y., Keetha, N., Lyu, C., Jhamb, B., Chen, Y., Qiu, Y., Karhade, J., Jha, S., Hu, Y., Ramanan, D., et al.: Ufm: A simple path towards unified dense correspondence with flow. arXiv preprint arXiv:2506.09278 (2025)
45. Zhenxing, M., Xu, D.: Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In: The Eleventh International Conference on Learning Representations (2022)