ProDiG: *Pro*gressive *Di*ffusion-Guided *G*aussian Splatting for Aerial to Ground Reconstruction

Sirshapan Mitra Yogesh S. Rawat
CRCV, University of Central Florida
{sirshapan.mitra, yogesh}@ucf.edu
https://sirsh07.github.io/research/prodig

Abstract

Generating ground-level views and coherent 3D site models from aerial-only imagery is challenging due to extreme viewpoint changes, missing intermediate observations, and large scale variations. Existing methods[9, 10] either refine renderings post-hoc, often producing geometrically inconsistent results, or rely on multi-altitude ground-truth, which is rarely available. Gaussian Splatting and diffusion-based refinements[38] improve fidelity under small variations but fail under wide aerial-to-ground gaps. To address these limitations, we introduce ***ProDiG (Progressive Diffusion-Guided Gaussian Splatting for Aerial to Ground Reconstruction)****, a diffusion-guided framework that progressively transforms aerial 3D representations toward ground-level fidelity. ProDiG synthesizes intermediate-altitude views and refines the Gaussian representation at each stage using a* geometry-aware causal attention module *that injects epipolar structure into reference-view diffusion. A* distance-adaptive Gaussian module *dynamically adjusts Gaussian scale and opacity based on camera distance, ensuring stable reconstruction across large viewpoint gaps. Together, these components enable progressive, geometrically grounded refinement without requiring additional ground-truth viewpoints. Extensive experiments on synthetic and real-world datasets demonstrate that ProDiG produces visually realistic ground-level renderings and coherent 3D geometry, significantly outperforming existing approaches in terms of visual quality, geometric consistency, and robustness to extreme viewpoint changes. Github: https://github.com/sirsh07/ProDiG*

1. Introduction

3D site modeling is fundamental to virtual and augmented reality, digital environment construction, robotics, and autonomous navigation. Recent advances in neural scene representations-including Neural Radiance Fields (NeRFs) [25] and Gaussian Splatting [16] have led to high-fidelity reconstructions from diverse camera trajectories. Strong results have been achieved using ground-view imagery [5, 8, 42], aerial-only inputs [23, 34, 39], joint aerial-ground capture [14, 46], and even satellite-scale observations [39]. However, these systems typically operate under small viewpoint deviations, and their quality degrades sharply when asked to extrapolate across extreme view and scale differences.

A rapidly emerging application domain-driven by consumer UAVs, surveillance drones, and wide-area mapping platforms-demands *3D reconstruction from aerial-only observations*. In this setting, the goal is to generate ground-level views or full 3D site models using only aerial imagery (Figure 1). This task is exceptionally challenging: aerial and ground viewpoints vary, scene structures appear at vastly different scales, intermediate camera poses are absent, and outdoor dynamics introduce shadows, occlusions, and transient objects.

Some recent efforts attempt to bridge this gap using generative refinement or geometric reconstruction. Generative approaches such as [9] improve visual quality but operate as post-hoc enhancement, leaving the 3D structure unchanged and often inconsistent under novel views. Geometry-driven pipelines like [10] require multi-altitude ground-truth for reliable registration, which is rarely available in practice. As a result, current methods lack a way to progressively adapt the 3D representation across altitudes while enforcing strong geometric consistency.

Gaussian Splatting itself faces additional limitations in this setup. When trained solely on aerial inputs, Gaussians tend to overfit to the aerial viewing distribution, producing artifacts and unstable geometry when extrapolated to ground-level viewpoints. Diffusion-guided refinement methods [3, 4, 38, 44] can improve fidelity in near-distribution settings, but become brittle when reference and target views differ substantially. Under wide-baseline conditions, diffusion models often copy appearance from the reference image or hallucinate structures inconsistent with the scene, reflecting the absence of mechanisms for progres-

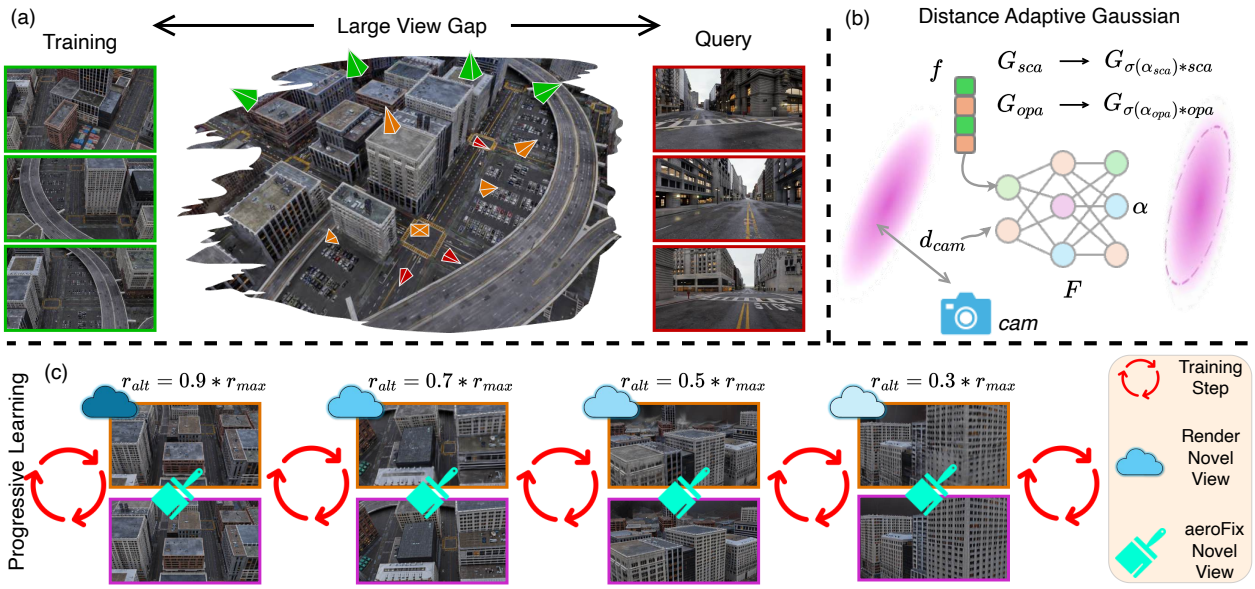

Figure 1. **Overview of ProDiG.** *(a)* Our framework reconstructs a complete 3D scene using only aerial images. A large distribution shift exists between the aerial training images and the ground-level query images. During evaluation, we render novel views at ground-level camera poses and compare them against ground-truth images. *(b)* In our Distance-Adaptive Gaussian Splatting module, each Gaussian is dynamically scaled and reweighted using a lightweight encoder that predicts adjustment factors from its learned scaling feature and its distance to the active camera. *(c)* We progressively render noisy novel views at successively lower altitudes, fix these views using our diffusion model, and iteratively retrain the Gaussian Splatting model using the fixed novel view.

sive viewpoint descent and geometry-aware conditioning.

To address these challenges, we introduce **ProDiG**, a diffusion-guided framework that progressively transforms an aerial 3D representation toward ground-level fidelity. Instead of attempting a direct aerial-to-ground leap, ProDiG synthesizes a sequence of intermediate altitudes and refines the Gaussian representation at each stage. Because camera poses encode rich geometric relationships across altitudes, we incorporate a geometry-aware conditioning module that injects epipolar structure into the diffusion model's reference-view guidance. To handle large variations in camera-to-scene distance, we further introduce a lightweight distance-adaptive module that dynamically adjusts Gaussian parameters during training, ensuring stable updates across wide-baseline transitions.

By combining progressive altitude descent, geometry-aware diffusion guidance, and distance-adaptive Gaussian refinement, ProDiG produces coherent 3D geometry and realistic ground-level renderings from aerial-only inputs without requiring ground-view or multi-altitude captures.

Contributions. The main contributions of this work are as follows:

- **Causal Attention Mixing.** We develop a pose-aware refinement module that incorporates epipolar geometry into the diffusion-based restoration process, enabling structurally consistent and geometrically grounded novel-view refinement.
- **Distance-Adaptive Gaussian Module.** We propose a lightweight camera-aware modulation mechanism that dynamically adjusts Gaussian parameters (scale, opacity) based on the camera's distance to the scene, improving rendering under extreme viewpoint changes.
- **Progressive Altitude Refinement.** We introduce a progressive learning strategy that synthesizes and integrates intermediate-altitude views, effectively bridging the large viewpoint gap between aerial and ground perspectives.

2. Related Work

Novel View Synthesis: Recent works [12, 16, 18] in novel view synthesis have primarily focused on controlled, synthetic environments where camera trajectories are well-sampled and scenes are static. Some methods have extended this to more realistic or in-the-wild settings, which introduce additional challenges such as dynamic objects, variable lighting, and scene clutter. *Wild Gaussian*[19] and *WildGS* [41] address these issues by incorporating appearance embeddings and confidence modeling to handle temporal and photometric inconsistencies.

Other approaches, such as *Mip-Splat*[45], address the problem of varying camera distances by applying mip filters to the gaussian models. More recent efforts, including *Scaffold Gaussian*[24], *City Gaussian*[23], and *Octree Gaussian*[28], focus on large-scale scene modeling using Gaussian Splatting. [24] introduces neural Gaussians to im-

prove level-of-detail representations, while [28] leverages octree structures to manage memory and rendering efficiency. However, these methods largely assume consistent viewpoints between training and testing.

In contrast, our work explicitly addresses the challenge of large viewpoint changes between aerial training data and ground-level evaluation, a setting that demands both geometric generalization and perceptual robustness.

Diffusion Models. Recent advances in diffusion models span a broad spectrum of tasks, from text and image synthesis [26, 47] to high-fidelity image restoration [22]. Closely related to our work are approaches that refine noisy Gaussian Splatting renderings using diffusion-based denoising. Prior methods such as [3, 38] progressively enhance Gaussian scene representations by cleaning noisy renderings before reintegrating them into the 3D model. [9] applies diffusion-based refinement to reduce noise in street-view renderings of large-scale 3D scenes, but does not update the underlying 3D representation. While our objective is related, our method explicitly leverages the 3D structure of the scene: we incorporate geometric constraints into the diffusion model, enabling geometry-aware, cross-view-consistent refinement that updates and improves the underlying Gaussian Splatting representation.

3. Methodology

Problem Formulation: Our goal is to reconstruct a 3D site using only aerial-view images, while enabling faithful rendering of views from ground-level camera poses. Let $\mathcal{I}_{\text{aerial}} = \{I_i\}_{i=1}^{N}$ denote a set of aerial images with known ground truth camera parameters or obtained via standard structure-from-motion[29]. We seek to learn a 3D Gaussian representation $\mathcal{G}$ such that renderings from novel ground-view poses closely match real ground-level observations, with minimal hallucination or geometric distortion.

Formally, let $\mathcal{Q}_{\text{ground}} = \{q_j\}_{j=1}^{M}$ denote a set of ground-level query poses, and let $\mathcal{I}_{\text{ground}} = \{I_j^*\}_{j=1}^{M}$ represent the corresponding ground-truth images used only for evaluation. Our objective is to optimize $\mathcal{G}$ such that:

$$R(\mathcal{G}, q_j) \approx I_j^*, \quad \forall q_j \in \mathcal{Q}_{\text{ground}},$$

where $R(\mathcal{G}, q_j)$ denotes the differentiable Gaussian Splatting renderer evaluated at pose q_j.

To achieve this, we employ a progressive altitude refinement strategy. Starting from an initial Gaussian model trained solely on aerial views, we iteratively (i) synthesize novel intermediate views from gradually lower altitudes, (ii) refine these views using a diffusion-based restoration module, and (iii) retrain the Gaussian model using the refined views. This progressive training loop enables the representation to adapt smoothly from aerial to ground viewpoints, reducing artifacts and preventing catastrophic hallucination in extreme novel-view synthesis.

We first describe our diffusion model *aeroFix* adapted for aerial scenes in Section 3.1. Next, we present our distance-adaptive modification to the Gaussian Splatting framework in Section 3.2. Finally, we introduce our progressive altitude learning strategy in Section 3.2.

3.1. aeroFix

Background: Diffusion models generate images through a gradual denoising process [11]. We build our diffusion framework on top of [38]. While Diffix+[38] is primarily trained on ground-view imagery, our work focuses on varying altitude views. Our contributions are twofold: (i) architectural enhancements to the diffusion model and (ii) revised training objectives.

Following the terminology in [38], we refer to the images rendered from Gaussian Splatting under novel camera poses before refinement as noisy novel views. The refined outputs obtained after diffusion are termed fixed novel views. The image used to condition the diffusion process is referred to as the reference view, and when this view corresponds to a ground-truth image, we denote it as the ground-truth reference view.

Causal Attention Mixing:

Diffusion-based refinement [3, 38, 44] offers strong visual generation capabilities, but when applied to aerial-to-ground synthesis it suffers from a fundamental limitation: cross-view interactions are unconstrained. Tokens in the noisy novel view freely attend to all tokens in the reference view, even when the two views are separated by large viewpoints. This unconstrained mixing causes hallucinations, weak geometric alignment, and instability across altitudes. Motivated by this observation, and drawing inspiration from prior work in 3D- and multi-view diffusion models [2, 13, 15, 31, 32], we introduce our Causal Attention Mixing module, which explicitly incorporates epipolar geometry into the reference-view conditioning mechanism.

We adapt the self-attention layer with our proposed Causal Attention Mixing module. The key idea is to exploit the underlying epipolar geometry to constrain cross-view interactions. Instead of allowing each token in the noisy novel view to attend to all tokens in the reference view, we restrict attention to only those reference tokens that lie along the corresponding epipolar line [15, 32].

As illustrated in Fig. 2, given a noisy novel view and a reference view together with their respective camera poses, we compute the epipolar relationships between the two views. For each pixel in the novel view, we first obtain its corresponding ray parametrization ℓ from the camera intrinsics and extrinsics. Using the relative pose, we reproject this ray, producing a set of projected points that trace out the corresponding epipolar line in the reference image [15, 33]. We then convert this line into a binary attention mask (after appropriate resizing), yielding an epipolar mask (E_{ref})

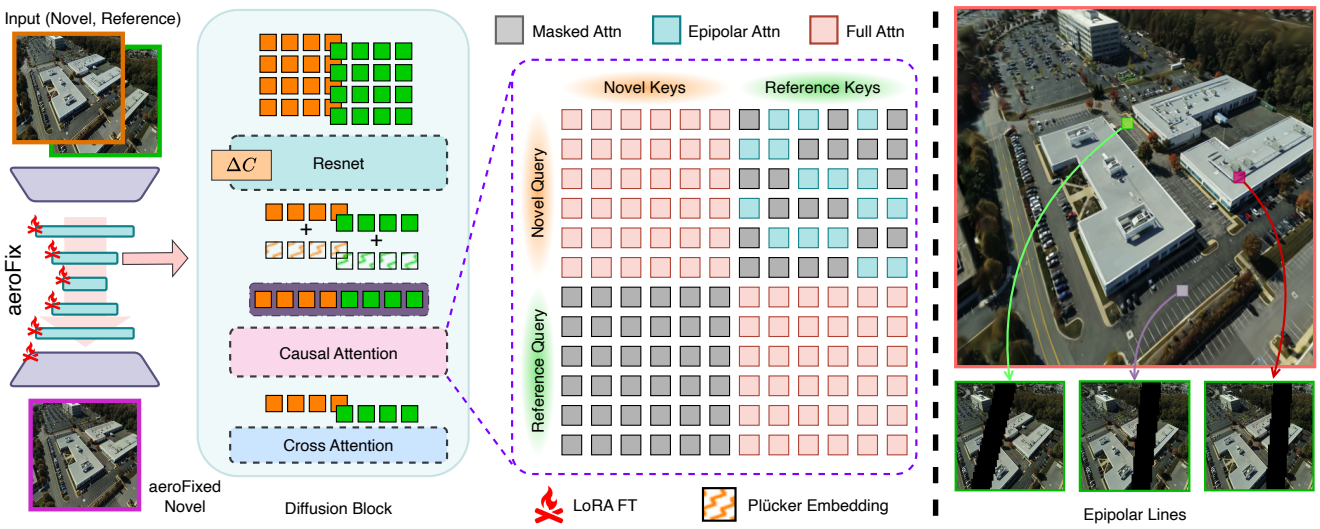

Figure 2. **Overview of aeroFix:** *(left)* Our diffusion model is fine-tuned on aerial imagery using LoRA. The noisy novel view is fixed using the reference view to fixed novel image. In the diffusion block, the relative camera pose difference is injected into the timestep embedding of the noisy image to encode geometric variation across viewpoints. We additionally include Plücker ray embeddings before the attention mixing layer to provide geometric cues. In the Causal Attention Mixing module, we enforce an epipolar constraint by masking the novel query - reference key attention map such that only tokens aligned with the corresponding epipolar lines retain attention (value 1), while all others are suppressed (value 0). The reference query - novel key block is fully masked, and the remaining attention blocks operate under standard full attention. *(right)* The figure illustrates epipolar correspondences for multiple query points on the noisy image and their corresponding lines on the reference view.

for every token in the novel view. This mask identifies the subset of reference-view tokens that lie near the epipolar line corresponding to that novel-view token. To improve robustness, masks are further dilated. This dilation increases the effective attention region and compensates for small pose inaccuracies arising during pose estimation. Because the diffusion model follows a causal generation process-corrupting only the novel view while the reference view should remain clean-we also apply a complementary constraint: reference tokens should not attend back to noisy novel tokens. This asymmetric masking mirrors the causal structure of the denoising process and prevents the stable guidance view from being contaminated by noisy features.

Formally, let Q, K, V denote the queries, keys and values in causal attention mixing, respectively. Our attention is computed as:

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{Q_n K_n^\top}{\sqrt{d}} \odot \begin{bmatrix} 1 & E_{\text{ref}} \\ 0_{\frac{n}{2}\times\frac{n}{2}} & 1 \end{bmatrix}\right) V,$$

where $\odot$ denotes element-wise masking using the (dilated) epipolar constraints.

Large viewpoint differences also determine the severity of noise and the degree of ambiguity in the novel view. We inject this information explicitly by conditioning the diffusion timestep with the pose difference between the novel and reference cameras allowing the network to infer how severely the image is degraded based on the viewpoint gap. For the reference view, we use a zero-valued pose-difference embedding, indicating that it serves as a clean and stable guidance source.

Finally, to encode each pixel's 3D spatial information, we incorporate a Plücker ray embedding [15, 32] into the feature maps before the Causal Attention Mixing stage using an encoder, enhancing cross-view geometric consistency. Formally, a plücker ray embedding with camera origin $\mathbf{o} \in \mathbb{R}^3$ and direction $\mathbf{d} \in \mathbb{R}^3$ is represented as $\mathbf{P} = (\mathbf{o} \times \mathbf{d}, \mathbf{d})$, where $\times$ denotes the cross product.

Loss Functions: Refining aerial-to-ground novel views with diffusion introduces two main challenges: maintaining perceptual consistency with the reference view and preserving fine structural details that are easily lost during denoising. Thus, we add two complementary loss terms that explicitly guide both perceptual alignment and structural fidelity. First, we employ a DSSIM loss, commonly used in gaussian splatting. Since the novel view typically shares similar lighting and scene appearance with the reference view (as opposed to the ground-truth training target), DSSIM provides a perceptually meaningful constraint that encourages the refined novel view to retain appearance consistency while still denoising, which reduces the risk of color shifts or unnatural smoothing. Cross-view diffusion can blur fine-grained structures such as building edges, windows, and rooftop details, particularly in aerial imagery with large depth variations. To preserve these features, we

introduce a multi-scale Sobel loss. We compute Sobel edge magnitude maps to construct edge-aware weighting masks, which modulate the ℓ_2 loss so that high-frequency structural regions such as fine-grained details like windows and building edges in aerial images receive greater emphasis during optimization. A similar edge-guided weighting strategy was also employed in [22] for diffusion-based structural guidance. This loss is applied across multiple spatial scales by downsampling the images.

3.2. Progressive Gaussian Splatting

Background: Gaussian Splatting [16] represents a 3D scene as a set of anisotropic Gaussian primitives, each parameterized by a mean position $\mu \in \mathbb{R}^3$, a covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$ that controls its spatial extent and orientation, a color feature vector $c \in \mathbb{R}^3$, and an opacity parameter $\alpha \in [0, 1]$. Given an initial point cloud and calibrated camera poses, each point is expanded into a Gaussian defined as:

$$G(\mathbf{X}) = \alpha \cdot \exp\left(-\tfrac{1}{2}(\mathbf{X}-\mu)^\top \Sigma^{-1}(\mathbf{X}-\mu)\right).$$

During rendering, each 3D Gaussian is projected onto the image plane using the camera intrinsics, view direction, and Σ. For a pixel, the resulting color $C(p)$ is given by:

$$C(p) = \sum_i \alpha_i(p)\, c_i \prod_{j<i}(1-\alpha_j(p)),$$

where $\alpha_i(p)$ denotes the projected opacity of the i-th Gaussian at pixel p.

The Gaussian parameters $\{\mu, \Sigma, c, \alpha\}$ are optimized by minimizing the discrepancy between the rendered image I_{render} and a ground-truth image I_{gt}.

$$\mathcal{L} = \lambda_1 \cdot \text{DSSIM}(I_{\text{render}}, I_{\text{gt}}) + \lambda_2 \cdot \|I_{\text{render}} - I_{\text{gt}}\|_2^2.$$

Distance-Adaptive Gaussian Module: With cameras distributed across varying altitudes, we draw inspiration from level-of-detail (LoD) strategies [17, 24, 28] commonly used in large-scale scene modeling. We use a simple yet effective modification to regulate Gaussian growth and suppress unwanted artifacts. Inspired by [24], we adaptively scale both the size and opacity of each Gaussian based on its distance from the active camera. Specifically, we maintain a learnable feature vector (f_{sca}, f_{opa}) for each Gaussian and use a lightweight MLP (F_{sca},F_{opa}) to predict scaling (α_{sca}) and opacity (α_{opa}) adjustment factors conditioned on this feature and the camera distance (d_{gc}). This formulation allows the representation to remain stable across views captured at different altitudes. Finally, to handle illumination diversity in in-the-wild datasets, we use appearance embeddings[19] that enable consistent color adaptation under varying lighting conditions.

$$\alpha_{\text{opa}} = F_{\text{opa}}(f_{\text{opa}}, d_{\text{gc}}), \qquad \text{opa} = \sigma(\alpha_{\text{opa}}) \cdot \text{opa},$$
$$\alpha_{\text{sca}} = F_{\text{sca}}(f_{\text{sca}}, d_{\text{gc}}), \qquad \text{sca} = \sigma(\alpha_{\text{sca}}) \cdot \text{sca},$$

Altitude-Based Progressive Learning: We investigate several strategies for altitude-based progressive learning, where novel viewpoints are generated at gradually reduced altitudes. Prior works [10, 20] commonly used a novel low-altitude camera trajectory. While this approach can be effective, the resulting viewpoints may sometimes differ too significantly from the reference views, making the refinement task particularly challenging for the diffusion model. To better control this variation, we study several alternative trajectory constructions that vary the degree of deviation from the original camera poses:

1. **Novel Trajectory (Baseline).** An elliptical low-altitude path is generated and uniformly sampled, with all cameras oriented toward the scene centroid.
2. **Scaled Trajectory.** Original camera poses are retained but their altitudes are scaled by a fixed factor, preserving azimuth while lowering height.
3. **Forward Trajectory.** Cameras are moved forward along their viewing direction to reduce altitude while keeping orientation unchanged.
4. **Stochastic Forward Trajectory.** The forward-translation strategy augmented with small random yaw and pitch perturbations to increase viewpoint diversity.
5. **Stochastic Scaled Forward Trajectory.** Cameras undergo altitude scaling plus a small forward shift, combined with mild pose noise for smoother variation.

The remainder of the progressive training pipeline follows [38]. We first train the Gaussian Splatting model for a fixed number of iterations. Using the strategies described above, we then generate novel views and fix them using our diffusion model. The fixed novel views are subsequently added back into the Gaussian Splatting training set, enabling iterative improvement. Finally, we apply a view-quality filtering step to downweight or discard novel views whose fixed renderings deviate excessively from their corresponding reference views, ensuring that training emphasizes reliable and consistent samples.

4. Experiment Details

We evaluate our approach in two stages. First, we assess the effectiveness of our diffusion-based refinement module *aeroFix* in denoising and correcting novel views rendered from Gaussian Splatting in aerial scenes in sec 4.1. Next, we evaluate the full *ProDiG* pipeline, where the refined novel views are iteratively incorporated back into the Gaussian Splatting optimization sec 4.2.

4.1. Effectiveness of aeroFix

We fine-tune the diffusion model introduced in [38] to operate in aerial-view settings following [27]. To construct training data, we render novel views from Gaussian Splatting models and treat the unrefined renderings as noisy inputs while using the corresponding ground-truth

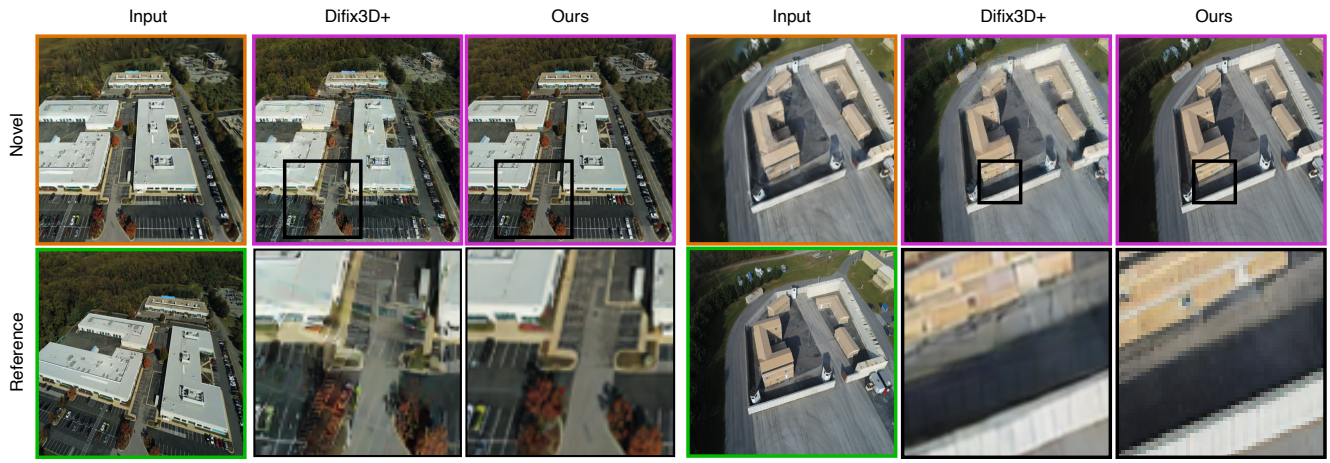

Figure 3. **Effectiveness of aeroFix:** Comparison of aerial image refinement between Difix3D+[38] and our aeroFix model. The noisy novel views are outlined in orange, the reference images in green, and the refined (fixed) novel images in pink. Difix3D+ tends to copy content from the reference view when the viewpoint difference is large, leading to inconsistencies and artifacts. In contrast, aeroFix effectively preserves structural fidelity and produces geometrically consistent, artifact-free refinements.

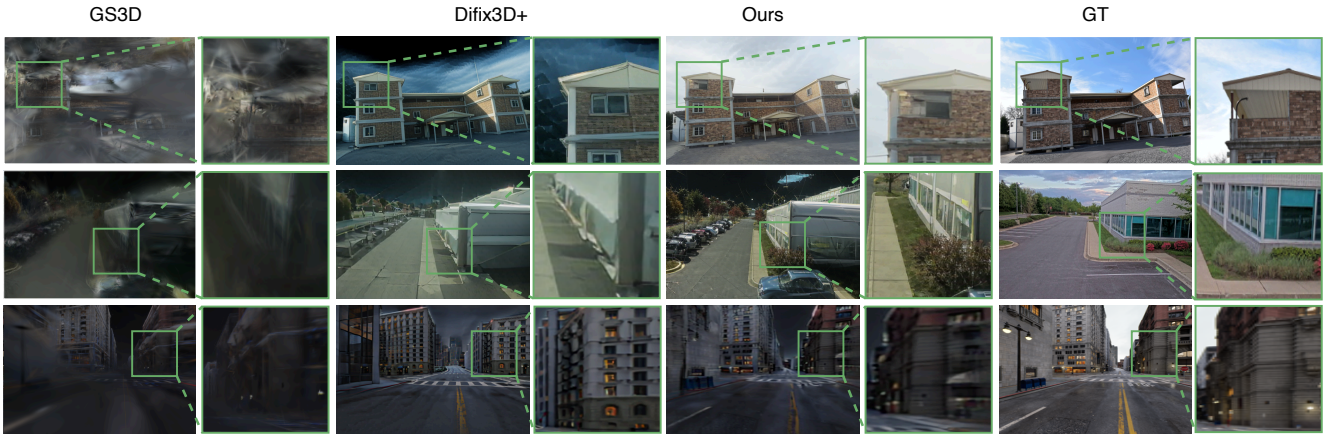

Figure 4. **Qualitative analysis of ProDiG(ours):** Comparison of our method with existing baselines on aerial-to-ground reconstruction. Gaussian Splatting [16] struggles to render complete scenes due to the absence of ground-level viewpoints, while Difix3D+[38] exhibits noisy artifacts and hallucinated structures. In contrast, ProDiG (ours) produces geometrically consistent and visually coherent reconstructions with fewer hallucinations. Notably, in the second row, the aerial inputs and reconstructed model include cars visible from above, whereas the ground-truth image - captured at a different time - does not.

views as clean targets. In total, we generate approximately 35k noisy-clean training pairs, each accompanied by camera pose information. The training data is drawn from several diverse aerial datasets, including GauUScene[40], MatrixCity[21], Mill19[35], and HorizonGS[14], where we ran the 3DGS pipeline to obtain the renderings.

For evaluation, we prepare a separate held-out set of 1k noisy-clean image pairs that is not used during training. We compare our model against the original [38] baseline, which has previously demonstrated strong performance in removing Gaussian Splatting artifacts. We report PSNR, SSIM, LPIPS, and DreamSim scores to measure reconstruction fidelity, perceptual quality, and semantic alignment. Additional implementation details and fine-tuning hyperparameters are provided in the supplementary material.

Results: Our method demonstrates consistent improvements over both the Difix3D+ and its LoRA-finetuned variant, as shown in Table 1. Figure 3 further illustrates that Difix3D+ tends to copy content from the reference view when the viewpoint difference between the reference and novel aerial image is large. In contrast, our diffusion model produces structurally consistent refinements with fewer artifacts and reduced noise. We also present a study of the key components of aeroFix in Table 1. We observe that incorporating pose embedding and causal attention yields consistent improvements over the baseline across both structural and perceptual metrics. The additional loss terms in aeroFix provide further gains in structural fidelity, while perceptual

Table 1. Comparison of aeroFix with Difix3D+[38]. Best values are in **bold**.(D: Dreamsim, P: PSNR, S: SSIM, L: LPIPS)

Method	D ↓	P ↑	S ↑	L ↓
Difix3D+ [38]	0.15	20.47	0.54	0.42
Difix3D+ (LoRA)	0.07	21.45	0.59	0.30
Pose + Plücker	0.06	22.30	0.64	0.27
Pose + Plücker + Causal	**0.03**	23.35	0.68	**0.24**
aeroFix (Ours)	**0.03**	**23.68**	**0.69**	**0.24**

quality remains mostly unchanged.

4.2. Effectiveness of ProDiG

Datasets: We conduct extensive evaluations across sites from multiple datasets. First, we evaluate our method on the WRIVA dataset [1], which consists of real-world, in-the-wild images captured at varying altitudes. Our experiments are performed using point clouds and camera poses generated by COLMAP[29, 30]. To better mimic real-world scenarios we do not use ground-truth poses for aerial images when estimating camera poses. Additionally, we exclude ground images from the point cloud generation process. Instead, the camera poses for the ground images are estimated using RANSAC[6], ensuring there is no data leakage through the point cloud. On average, each WRIVA site contains 50 aerial training images and 25 ground testing images, with large changes in viewpoint from aerial to ground. We also evaluate our approach on a synthetic dataset: Matrix City[21]. For Matrix City, we focus specifically on the small city, as it includes both aerial and street views.

Evaluation Metrics: We evaluate our method using two structural metrics, PSNR and SSIM [37] and two perceptual metrics, LPIPS [48] and DreamSim [7]. Among these, DreamSim is specifically designed to align quantitative evaluation more closely with human visual perception, making it our primary metric for assessing perceptual quality. In addition to reconstruction accuracy, we also report the total number of Gaussians used in each method to quantify memory efficiency and scalability.

Baselines: We compare our approach against several state-of-the-art baselines, including *3D Gaussian Splatting (3D-GS)*[16], *2D Gaussian Splatting (2D-GS)*[12], *Scaffold-GS*[24], *GS-MCMC*[18] and *Difix3D+*[38]. In the case of [38], since the original framework does not incorporate progressive learning from high to low altitudes, we adapt it to progressive training setup to enable altitude-aware refinement. These baselines represent a diverse set of methods for scene representations.

Implementation Details: We use point cloud to initialize the Gaussian Splatting pipeline. Subsequently, we train the corresponding Gaussian Splatting model for up to 7,000 iterations using the gsplat [43] implementation of Gaussian Splatting. To enable altitude-based refinement, we render novel views at 90% of the original altitude and apply the our model to enhance the visual quality of these views. This refinement process is then repeated at lower altitudes (70%, 50%, 30% and 10%), resulting in a total five-stage iterative improvement of the splats. While further steps can be added for more improvement, we limit our experiments to five refinement stages for efficiency. All experiments are conducted on an NVIDIA RTX A6000 GPU.

Table 2. Comparison of our method with existing methods on the WRIVA dataset across two sites. Best values are in **bold**.(D: Dreamsim, P: PSNR, S: SSIM, L: LPIPS)

Method	Wriva S06				Wriva S01			
	D ↓	P ↑	S ↑	L ↓	D ↓	P ↑	S ↑	L ↓
3DGS [16]	0.79	7.74	0.27	0.94	0.66	9.74	0.40	0.85
3DGS-MCMC [18]	0.71	8.71	0.31	0.90	0.57	9.73	0.39	0.80
2DGS [12]	0.78	7.08	0.21	0.95	0.68	9.68	0.36	0.85
Scaffold-GS [24]	0.72	7.79	0.22	0.88	0.67	10.35	0.37	0.82
Difix3D+ [38]	0.66	8.91	0.28	0.80	0.38	11.83	0.37	0.67
Ours	**0.50**	**11.26**	**0.33**	**0.67**	**0.29**	**13.10**	**0.45**	**0.58**

Table 3. Comparison of our method with existing methods on the MatrixCity dataset. Best values are in **bold**. (D: Dreamsim, P: PSNR, S: SSIM, L: LPIPS)

Method	D ↓	P ↑	S ↑	L ↓
3DGS [16]	0.51	10.71	0.40	0.77
2DGS [12]	0.62	9.29	0.28	0.81
3DGS-MCMC [18]	0.49	10.84	0.41	0.77
Scaffold-GS [24]	0.54	10.19	0.35	0.75
Difix3D+ [38]	0.48	11.38	0.38	0.63
Ours	**0.39**	**12.39**	**0.41**	**0.50**

4.3. Results on Wriva Dataset

The WRIVA dataset presents a challenging setup due to the large viewpoint gap between aerial and ground views. As shown in Table 2, existing Gaussian Splatting baselines yield significantly lower performance across all metrics. Among them, 3DGS-MCMC shows the most competitive results, but still underperforms compared to our method.

Our method yields an average improvement of 0.2 in DreamSim, indicating better alignment with human visual perception. While the structural performance varies across sites, we observe that on the S06 site, our method shows 2 improvement PSNR and 0.02 in SSIM compared to the baselines. Also, on the **S01** site, our method outperforms the baseline by 1.3 in PSNR and 0.05 in SSIM.

4.4. Results on Matrix City

The MatrixCity dataset provides a synthetic yet well-controlled environment, enabling detailed evaluation of both structural and perceptual reconstruction quality. As shown in Table 3, our method consistently outperforms all baselines.

Specifically, our approach improves the DreamSim score by 0.09 and LPIPS by 0.13 compared to the best-performing baselines, comparable performance in PSNR and SSIM. These results are consistent with our findings on the WRIVA dataset: our method enhances the perceptual quality of the synthesized views-closely aligned with human visual preferences-while only marginally affecting structural fidelity.

4.5. Ablation Studies

Effectiveness of Progressive Strategies: We observe that different progressive strategies perform better across different sites; however, on average, the Stoch. strategy yields the most stable and consistent results, as reported in Fig. 6. We attribute this performance to its balanced viewpoint variation-it introduces moderate positional changes compared to the purely scaled trajectory while avoiding the large geometric deviations observed in the novel trajectory, resulting in smoother adaptation during progressive refinement.

Effectiveness of Distance-Adaptive Gaussian Module: Although the Distance-Adaptive Gaussian Module yields comparable DreamSim and LPIPS scores, it consistently improves PSNR and SSIM metrics Fig 6. This enhancement is particularly beneficial in scenarios where camera distances vary significantly.

4.6. Analysis and Discussions

Qualitative Analysis: In Figure 4, we present qualitative comparisons of our method against prior approaches such as 3DGS and 2DGS. Our method consistently produces sharper, more accurate reconstructions, while baseline methods often generate noisy or overly blurred outputs, particularly in regions with large viewpoint shifts.

Generalizability to Varying Altitude, Sparse Views: We demonstrate the generalizability of our approach across varying-altitude scenarios by evaluating on multiple sites from the Aerial MegaDepth[36] dataset. As shown in Fig. 5, our method consistently outperforms both vanilla Gaussian Splatting and Difix3D+ in visual quality, producing sharper and more coherent reconstructions across diverse altitude ranges and real-synthetic mixed data conditions.

5. Conclusion

We present **ProDiG**, a progressive altitude Gaussian splatting framework for aerial-to-ground 3D reconstruction. By combining intermediate-altitude synthesis, geometry-aware causal attention, and distance-adaptive Gaussian refinement, ProDiG produces stable, geometrically consistent 3D representations and realistic ground-level renderings from aerial-only inputs. Unlike prior post-hoc or geometry-dependent methods, our approach requires no additional ground-truth views and effectively handles extreme viewpoint and scale variations. Extensive experiments on synthetic and real-world datasets validate the effectiveness of proDiG, demonstrating significant improvements in visual fidelity, structural consistency, and robustness under wide-baseline aerial-to-ground scenarios. This work highlights the potential of progressive, geometry-guided diffusion for challenging 3D site modeling applications and lays the groundwork for further exploration of aerial-to-ground synthesis in unconstrained environments.

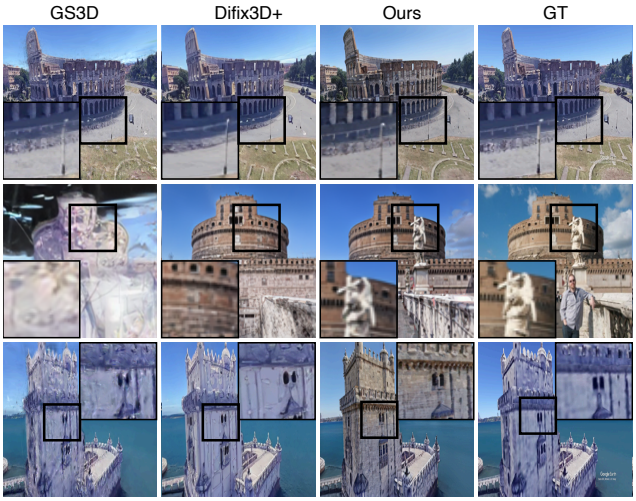

Figure 5. **Generalization across Varying Altitudes.** We evaluate our method on the Aerial MegaDepth[36] dataset, which contains sites captured at diverse altitude ranges.

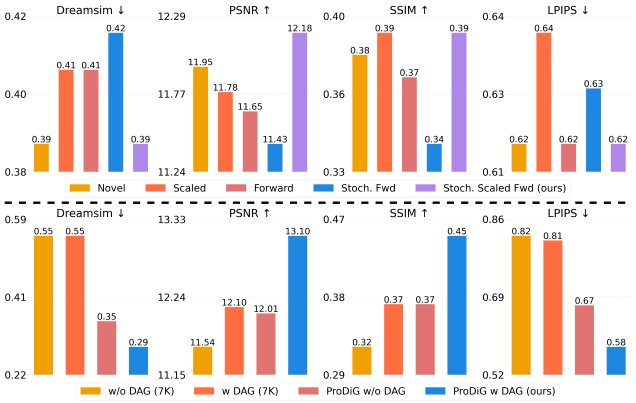

Figure 6. **Ablations.** *(top)* Comparison of different progressive methods. *(bottom)* Effectiveness of Distance Adaptive Gaussian Module. 7k represents, evalution at 7k iteration, after initial training.

Acknowledgments This work was supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes, notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Myron Brown, Michael Chan, and Michael Twardowski. Wriva public data, 2024. 7

[2] Eric Ming Chen, Sidhanth Holalkere, Ruyu Yan, Kai Zhang, and Abe Davis. Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23242–23251, 2023. 3

[3] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 1, 3

[4] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024. 1

[5] Tobias Fischer, Jonas Kulhanek, Samuel Rota Bulo, Lorenzo Porzi, Marc Pollefeys, and Peter Kontschieder. Dynamic 3d gaussian fields for urban areas. *arXiv preprint arXiv:2406.03175*, 2024. 1

[6] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 7

[7] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 7

[8] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024. 1

[9] Zhiyuan Gao, Wenbin Teng, Gonglin Chen, Jinsen Wu, Ningli Xu, Rongjun Qin, Andrew Feng, and Yajie Zhao. Skyeyes: Ground roaming using aerial view images. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3045–3054. IEEE, 2025. 1, 3

[10] Yujin Ham, Mateusz Michalkiewicz, and Guha Balakrishnan. Dragon: Drone and ground gaussian splatting for 3d building reconstruction. In *2024 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2024. 1, 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[12] 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. 2, 7

[13] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 3

[14] Lihan Jiang, Kerui Ren, Mulin Yu, Linning Xu, Junting Dong, Tao Lu, Feng Zhao, Dahua Lin, and Bo Dai. Horizon-gs: Unified 3d gaussian splatting for large-scale aerial-to-ground scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26789–26799, 2025. 1, 6

[15] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10026–10038, 2024. 3, 4

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 5, 6, 7

[17] 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. 5

[18] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *Advances in Neural Information Processing Systems*, 37:80965–80986, 2024. 2, 7

[19] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 2024. 2, 5

[20] Jie-Ying Lee, Yi-Ruei Liu, Shr-Ruei Tsai, Wei-Cheng Chang, Chung-Ho Wu, Jiewen Chan, Zhenjun Zhao, Chieh Hubert Lin, and Yu-Lun Liu. Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. *arXiv preprint arXiv:2510.15869*, 2025. 5

[21] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi 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. 6, 7

[22] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion

prior. In *European Conference on Computer Vision*, pages 430–448. Springer, 2024. 3, 5

[23] Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *European Conference on Computer Vision*, pages 265–282. Springer, 2024. 1, 2

[24] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 2, 5, 7

[25] 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

[26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 3

[27] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 5

[28] 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, 3, 5

[29] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 7

[30] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 7

[31] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34: 19313–19325, 2021. 3

[32] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pages 156–174. Springer, 2022. 3, 4

[33] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8269–8279, 2022. 3

[34] Jiadong Tang, Yu Gao, Dianyi Yang, Liqi Yan, Yufeng Yue, and Yi Yang. Dronesplat: 3d gaussian splatting for robust 3d reconstruction from in-the-wild drone imagery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 833–843, 2025. 1

[35] 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. 6

[36] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21674–21684, 2025. 8

[37] 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

[38] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025. 1, 3, 5, 6, 7

[39] Yuanbo Xiangli, Linning 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. 1

[40] Butian Xiong, Nanjun Zheng, Junhua Liu, and Zhen Li. Gauu-scene v2: Assessing the reliability of image-based metrics with expansive lidar image dataset using 3dgs and nerf. *arXiv preprint arXiv:2404.04880*, 2024. 6

[41] Jiacong Xu, Yiqun Mei, and Vishal Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. *Advances in Neural Information Processing Systems*, 37:103334–103355, 2024. 2

[42] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024. 1

[43] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. 7

[44] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1, 3

[45] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19447–19456, 2024. 2

[46] Chenhao Zhang, Yuanping Cao, and Lei Zhang. Crossview-gs: Cross-view gaussian splatting for large-scale scene reconstruction. *arXiv preprint arXiv:2501.01695*, 2025. 1

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3

[48] 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