Gaussian Scenes: Pose-Free Sparse-View Scene Reconstruction using Depth-Enhanced Diffusion Priors

Soumava Paul, Prakhar Kaushik, Alan Yuille
CCVL, Johns Hopkins University
soumava2016@gmail.com, {pkaushi1, ayuille}@jhu.edu

Abstract

In this work, we introduce a generative approach for pose-free reconstruction of 360° scenes from a limited number of uncalibrated 2D images. Pose-free scene reconstruction from incomplete, unposed observations is usually regularized with depth estimation or 3D foundational priors. While recent advances have enabled sparse-view reconstruction of unbounded scenes with known camera poses using view-conditioned diffusion priors, these methods cannot be directly adapted for the pose-free setting when ground truth COLMAP poses are not available during evaluation. To address this, we propose an RGBD diffusion model designed to inpaint missing details and remove artifacts in novel view renders and depth maps of a 3D scene. We introduce context and geometry conditioning using FiLM modulation layers as a lightweight alternative to cross-attention and also propose a novel confidence measure for Gaussian representations to allow for better detection of these artifacts. By progressively integrating these novel views in a Gaussian-SLAM-inspired process, we achieve a multi-view-consistent Gaussian representation. Evaluations on the MipNeRF360 and DL3DV-10K benchmark dataset demonstrate that our method surpasses existing pose-free techniques and performs competitively with state-of-the-art posed reconstruction methods in complex 360° scenes. Our code and datasets will be open-sourced upon acceptance.

1 Introduction

Reconstructing high-quality 3D scenes from sparse images remains a fundamental challenge in computer vision. While recent methods employ various priors to stabilize NeRFs (Mildenhall et al., 2020) or Gaussian splats (Kerbl et al., 2023) in under-constrained scenarios, they typically require accurate camera parameters derived from dense observations—a restrictive assumption for real-world applications. Pose estimation from sparse views is inherently challenging; both traditional Structure from Motion and recent foundational models (Wang et al., 2024; Leroy et al., 2024) struggle with insufficient matching features. Current pose-free 3DGS approaches integrate monocular depth (Ranftl et al., 2020), semantic segmentation (Kirillov et al., 2023), or 3D priors (Wang et al., 2024), but fail on complex 360° scenes with sparse coverage, highlighting the need for additional generative regularization.

Despite numerous attempts (Deng et al., 2022; Roessle et al., 2022; Li et al., 2024; Xiong et al., 2023; Zhu et al., 2023; Chung et al., 2023; Wang et al., 2023a; Jain et al., 2021; Niemeyer et al., 2022; Wynn and Turmukhambetov, 2023), only COGS Jiang et al. (2024) and InstantSplat (Fan et al., 2024) address sparse view synthesis without precomputed cameras. While they reduce artifacts and blur, they lack generative capacity for complete 360° reconstruction. Full scene reconstruction requires robust priors from powerful diffusion models (Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022) encoding common 3D structures. Recent methods (Liu et al., 2023; Sargent et al., 2024; Wu et al., 2024; Gao* et al., 2024; Blattmann et al., 2023) incorporate view conditioning for realistic extrapolation but depend on accurate poses. Only Gaussian Object (Yang et al., 2024), iFusion (Wu et al., 2023), and UpFusion (Nagoor Kani et al., 2024) provide pose-free generative solutions, yet target object rather than scene reconstruction.

We present *GScenes*, an efficient approach using 3D foundational and RGBD diffusion priors for pose-free sparse-view reconstruction of complex 360° scenes. We first estimate a point cloud and approximate camera parameters using MASt3R Leroy et al. (2024), then jointly optimize Gaussians and cameras with 3DGS. Novel views generated from a B-spline trajectory contain artifacts that our diffusion prior refines to further optimize the scene. We condition a Stable Diffusion UNet Rombach et al. (2022) on estimated cameras, context, 3DGS renders, depth maps, and a confidence map capturing artifacts. Despite using weaker generative priors than pose-dependent methods, we demonstrate competitive performance against ReconFusion (Wu et al., 2024) and CAT3D (Gao* et al., 2024) while outperforming other techniques without requiring million-scale multi-view data or extensive compute resources.

Our contributions include:

- An image-to-image RGBD diffusion model for synthesizing plausible novel views from sparse unposed images, using lightweight FiLM modulation Perez et al. (2018) instead of cross-attention
- A confidence measure to detect artifacts in novel view renders, guiding our diffusion model toward effective rectification
- Integration of diffusion priors with MASt3R's geometry prior, enabling efficient scene reconstruction previously requiring 3D-aware video diffusion
- Superior performance compared to recent regularization and generative prior-based approaches for large 3D scene reconstruction

See Appendix for further discussion and Related Works.

2 Method

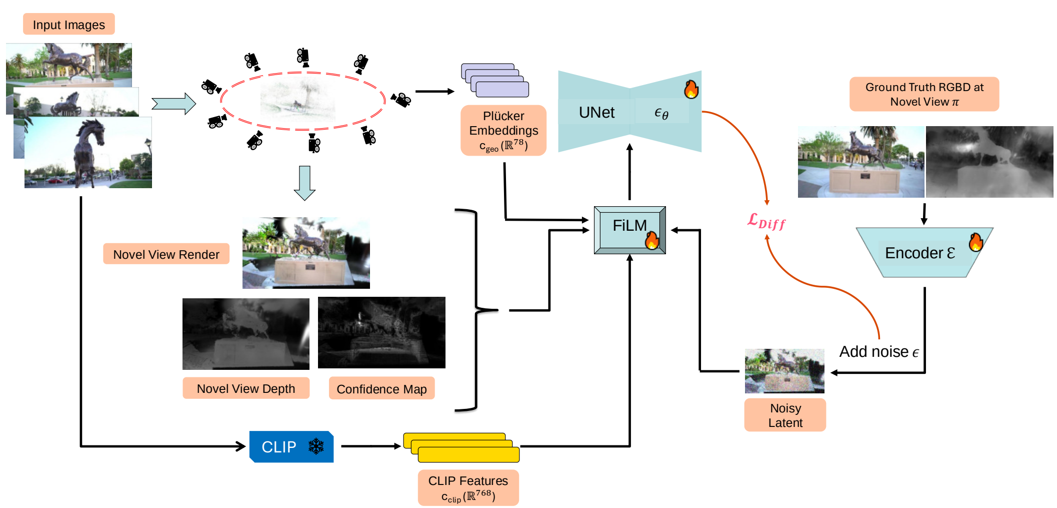


Figure 1: **Overview of *GScenes***. We render 3D Gaussians fitted to our sparse set of M views from a novel viewpoint. The resulting render and depth map have missing regions and Gaussian artifacts, which are rectified by an RGBD image-to-image diffusion model. This then acts as pseudo ground truth to spawn and update 3D Gaussians and satisfy the new view constraints. This process is repeated for several novel views spanning the 360° scene until the representation becomes multi-view consistent.

This section begins with an overview of our pipeline in Sec 2.1, detailing our approach for reconstructing a 3D scene from a sparse set of uncalibrated 2D images. In Sec 2.1, we describe how we initialize a Gaussian point cloud using MASt3R and 3DGS to provide a coarse 3D representation. Sec 2.2 introduces our RGBD image-to-image diffusion model, which refines rendered novel views by correcting artifacts and filling missing regions. In Sec 2.3, we propose a confidence measure based on cumulative transmittance and Gaussian density to guide the diffusion model in identifying unreliable regions in novel-view renders. Sec 2.4 outlines our synthetic dataset creation pipeline, which enables training the diffusion model with high-quality RGBD supervision. Sec 2.5 details our depth-augmented autoencoder finetuning process to improve latent-space encoding for RGBD data. Sec 2.5.1 explains the fine-tuning of our UNet with synthesized training data to generate photorealistic and geometrically consistent novel views. In Sec 2.5.2, we describe the inference process of our diffusion model, where novel views are synthesized using RGBD renders and confidence maps. Sec 2.6 presents our iterative 3D Gaussian optimization strategy that progressively integrates novel view constraints into the scene representation. Finally, Sec 2.7 describes a test-time pose alignment step that refines camera poses to align the rendered image with a given test view before evaluation.

2.1 Algorithm Overview

Problem Setup Given a set of M images $\mathcal{I} = \{I_1, I_2, \ldots, I_M\}$ of an underlying 3D scene with unknown intrinsics and extrinsics, our goal is to reconstruct the 3D scene, estimate the camera poses of a monocular camera at the M training views, and synthesize novel views at evaluation time given by N unseen test images $\{I_{M+1}, I_{M+2}, \ldots, I_{M+N}\}$.

Algorithm 1 *GScenes* Algorithm

Require: Sparse input image set $\mathcal{I} = \{I_1, I_2, \ldots I_M\}$
Ensure: Set of 3D Gaussians $\mathcal{G}$ and camera poses $\pi = \{\pi_1, \pi_2, \ldots \pi_M\}$

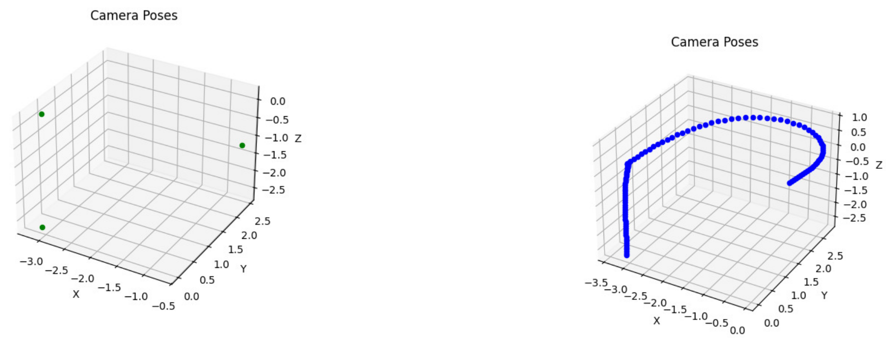
```
1: Î ← I
2: G ← Optimize 3DGS with MASt3r point cloud for 1k iterations
3: for N iterations do
4:     π ← Sample novel camera pose.
5:     I, D ← R_π(G) - Render from camera π
6:     I ← Refine(I)
7:     Î ← Î ∪ {I}
8:     G ← Optimize 3DGS for k iterations
9: end for
```

An overview of our method is given in Fig 1 and Alg 1. We initialize *GScenes* with an incomplete dense Gaussian point cloud reconstruction from sparse input images using MASt3R (Leroy et al., 2024) and *1k* iterations of 3DGS. Note that InstantSplat (Fan et al., 2024) proposes the same pipeline for sparse-view reconstruction, but with a DUSt3R initialization. We choose MASt3R to initialize scene geometry instead due to its superior performance in the sparse-view setting. We use this incomplete Gaussian scene representation as an implicit geometric prior and sample novel views along a smooth B-spline trajectory fitted to training views. An example trajectory is shown in Fig 2. We then use our RGBD diffusion prior to synthesize plausible novel views. In addition to CLIP features of source images for context and plücker embeddings of source and target cameras for geometric conditioning, we devise a novel 3DGS confidence measure to effectively guide our diffusion model towards empty regions and potential artifacts in a novel view render. We then run $10k$ iterations of 3DGS optimization, sampling a novel view before each densification step (Fu et al., 2023) to obtain our final scene representation.

Gaussian Point Cloud Initialization The MASt3r pipeline gives us a pixel-aligned dense-stereo point cloud $\mathbf{P} \in \mathbb{R}^{S\times 3}$, camera intrinsics $\{\mathbf{K_i} \in \mathbb{R}^{3\times 3}\}_{i=1}^{M}$ and extrinsics $\{\mathbf{E_i} = [\mathbf{R_i}|\mathbf{T_i}]\}_{i=1}^{M}$ for our M input images. Nevertheless, both $\mathbf{P}$ and the estimated poses demonstrate sub-optimal alignment compared to those generated by COLMAP (Schönberger and Frahm, 2016) from the dense observation dataset. Consequently, similar to the approach in InstantSplat, we initialize 3D Gaussians at each location in the globally aligned point cloud $\mathbf{P}$. We then jointly optimize both the Gaussian attributes and camera parameters over the $1k$ iterations without incorporating any form of Adaptive Density Control.

2.2 RGBD Diffusion Priors for NVS

Reconstructing complete 3D scenes from sparse observations requires inferring content in unobserved regions—a fundamental challenge that geometric regularization and 3D priors alone cannot adequately



(a) Optimized Training Poses from MASt3R + 3DGS

(b) NVS cameras with B-spline of degree 5

Figure 2: Camera Trajectory Visualization for Novel View Synthesis in pose-free sparse-view setting.

address. We introduce a diffusion-based generative approach that leverages 2D image priors to synthesize plausible content in these regions.

2.2.1 Generative Model Architecture

Despite our initial geometric reconstruction, sparse input views inevitably result in regions with no Gaussian primitives ("0-Gaussians"), causing empty areas and artifacts in novel views. Unlike regularization-based methods that merely constrain optimization without generating content, our approach directly synthesizes missing scene details.

Our model comprises:

- A variational autoencoder (encoder $\mathcal{E}$, decoder $\mathcal{D}$) operating in a compressed latent space
- A UNet denoiser ϵ_θ predicting noise in diffused latent z_t
- Multi-modal conditioning incorporating RGBD renders, confidence maps, semantic context, and geometric information

The UNet ϵ_θ receives four inputs: an artifact-laden RGBD image $\hat{I}$, a confidence map $\mathcal{C}$ identifying unreliable regions, CLIP features c_{clip} providing semantic context, and camera encodings c_{geo} establishing geometric relationships between views.

2.2.2 Multi-modal Conditioning

We initialize ϵ_θ with Stable-Diffusion-2 weights (Rombach et al., 2022) and expand the first convolutional layer to accept additional inputs by concatenating the noisy latent z_t, the encoded RGBD image $\mathcal{E}(\hat{I})$, the confidence-weighted encoded image $\mathcal{E}(\hat{I} \cdot \mathcal{C})$ and the downsampled confidence map $\hat{\mathcal{C}}$. To ensure view coherence and geometric consistency, we incorporate:

Semantic Context: CLIP features $c_{\text{clip}} \in \mathbb{R}^{M \times d}$ from source images serve as semantic anchors, ensuring generated content remains consistent with observed scene elements.

Geometric Information: For each camera with center $\mathbf{o}$ and forward axis $\mathbf{d}$, we compute its plücker coordinates $\mathbf{r} = (\mathbf{d}, \mathbf{o} \times \mathbf{d}) \in \mathbb{R}^6$ and apply frequency encoding for obtaining higher-dimensional features:

$$\mathbf{r} \mapsto [\mathbf{r}, \sin(f_1 \pi \mathbf{r}), \cos(f_1 \pi \mathbf{r}), \cdots, \sin(f_K \pi \mathbf{r}), \cos(f_K \pi \mathbf{r})] \tag{1}$$

where $K = 6$ is the number of Fourier bands, and f_k are equally spaced frequencies. This yields a 78-dimensional embedding for each camera ($c_{\text{geo}} \in \mathbb{R}^{(M+1) \times 78}$), capturing geometric relationships between viewpoints. Plücker coordinates were originally introduced by LFNs (Sitzmann et al., 2021) for per-pixel parameterization of a ray. We instead obtain a single representation per camera using extrinsics $\mathbf{E_i}$ for obtaining $\mathbf{o}$ and $\mathbf{d}$.

2.2.3 Parameter-Efficient FiLM Conditioning

We employ Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) instead of cross-attention for incorporating context and geometry information, achieving both computational efficiency and strong performance:

1. Process context and geometry through self-attention to capture inter-view relationships:
$$c^i_{\text{attn}} = \text{SelfAttention}(c_i); \quad i \in \{\text{clip}, \text{geo}\} \tag{2}$$
2. Generate scaling and shifting parameters via layer-specific networks:
$$\gamma^{(l)}, \beta^{(l)} = \text{FC}^{(l)}(c^i_{\text{attn}}) \tag{3}$$
3. Modulate feature maps through element-wise operations:
$$\mathbf{F}^{(l)}_{\text{mod}} = \gamma^{(l)} \cdot \mathbf{F}^{(l)} + \beta^{(l)} \tag{4}$$

Critically, we apply FiLM modulation only to down and mid blocks of the UNet—not up blocks—based on empirical evidence showing this selective application yields optimal results (Fig. 3).

Our FiLM-based approach requires only 8.14M parameters (7.38M for CLIP features, 758K for pose embeddings) compared to 29.8M for an equivalent cross-attention implementation—a $3\times$ reduction while maintaining comparable quality, enabling more efficient training and faster inference during iterative reconstruction.

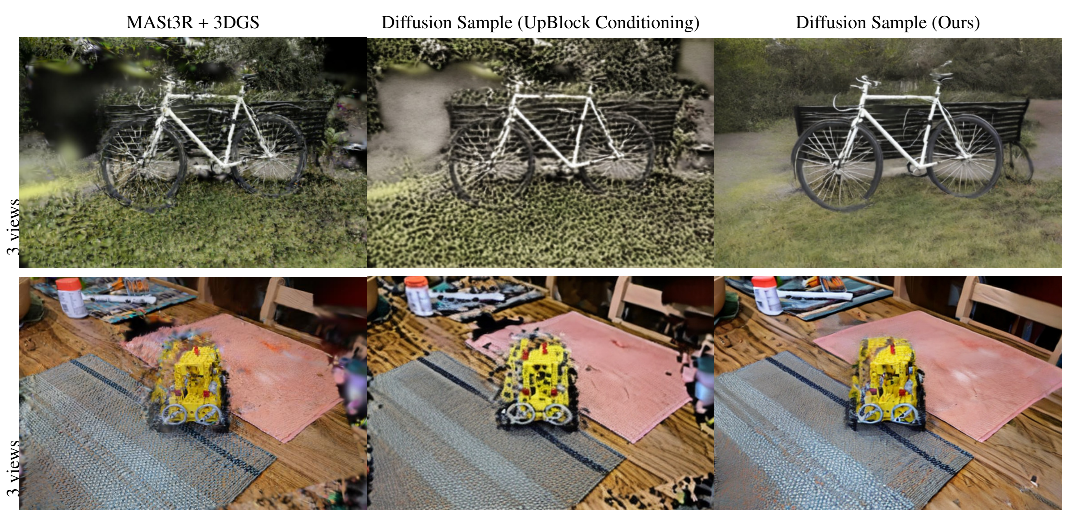


Figure 3: Incorporating context and geometry conditioning in the up blocks of the UNet negatively impacts latent and subsequent image reconstruction.

2.3 Pixel-Aligned Confidence Map

To guide our diffusion model in identifying problematic regions in novel views, we introduce a pixel-aligned confidence measure combining transmittance with Gaussian density:

$$\mathcal{C}_i = -\log(T_i + \epsilon) \times n_{\text{contrib}} \tag{5}$$

where $T_i = \prod_i (1 - \alpha_i)$ represents light transmission without Gaussian interaction, n_{contrib} counts contributing 3D Gaussians, and $\epsilon > 0$ prevents logarithmic singularities. This formulation captures two complementary reliability signals: (1) low transmittance indicates significant Gaussian interactions, suggesting higher rendering confidence, and (2) consensus among multiple Gaussians validates pixel reliability through primitive agreement.

Unlike 3DGS-Enhancer (Liu et al., 2024), which assumes well-reconstructed areas contain small-scale Gaussians, our measure remains effective for monotonous textures where fine-grained Gaussian representation is unnecessary. Fig. 4 demonstrates our approach accurately identifies both empty regions and reconstruction artifacts while avoiding false positives. The significance of this improved confidence measure is evident in Fig. 5, where models trained with previous confidence formulations produce implausible novel views due to misleading confidence signals.

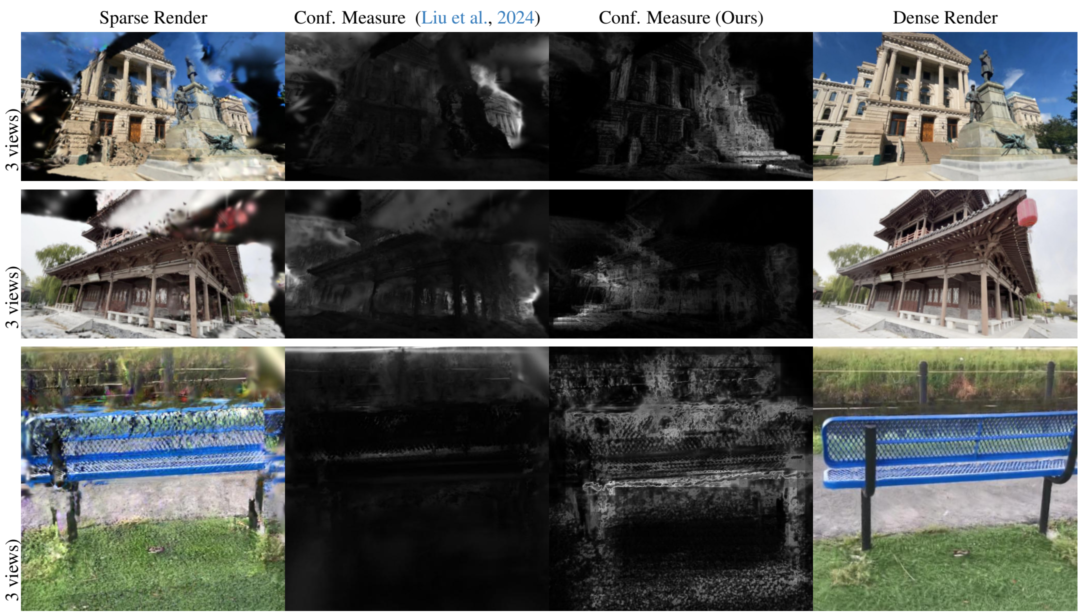


Figure 4: Confidence Measure comparison with 3DGS-Enhancer (Liu et al., 2024). Our confidence map accurately identifies artifacts and 0-Gaussian regions in the sparse-view (darker pixels) while (Liu et al., 2024) incorrectly attributes high confidence to regions with overlap of small-scale Gaussians. NVS render from a densely fitted 3DGS representation is provided for reference.

Figure 5: Conditioning the UNet with an inaccurate confidence measure((Liu et al., 2024)) leads to implausible NVS.

2.4 Synthetic RGBD Dataset Creation

For training the additional weights in ϵ_θ for RGBD image-to-image diffusion, we rely on a set $\mathcal{X} = \{(I^i, \hat{I}^i, \mathcal{C}^i, c^i_{clip}, c^i_{geo})_{i=1}^N\}$, each containing a clean RGBD image I^i, an RGBD image

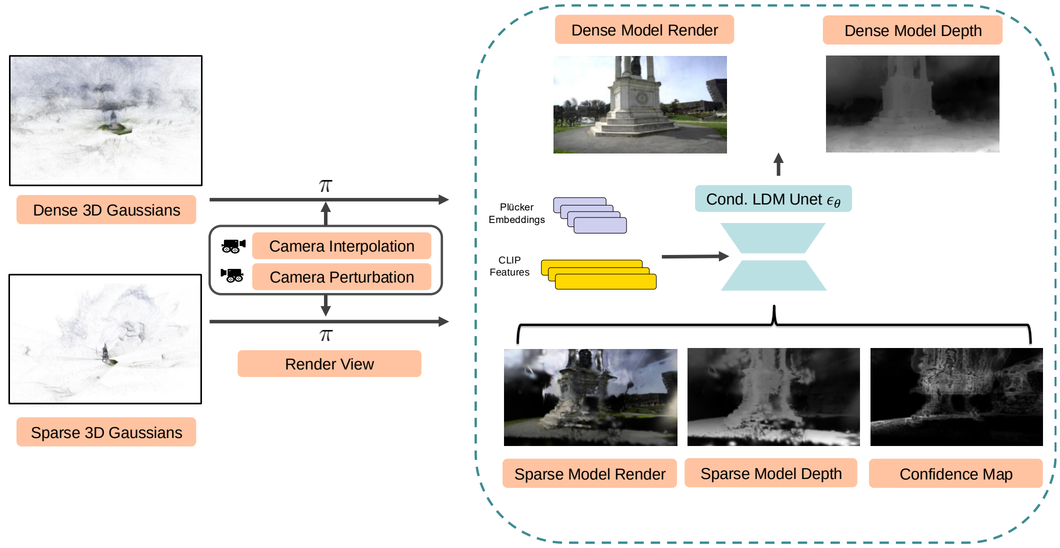


Figure 6: **Training our RGBD diffusion model.** Pairs of RGBD clean images and images with artifacts are obtained from 3DGS fitted to sparse and dense observations, respectively, across 1043 scenes. CLIP features provide semantic scene context, plücker embeddings of source and target cameras provide geometry information, and a confidence map additionally detects empty regions and artifacts in the artifact image. The Stable Diffusion UNet (Rombach et al., 2022) is then fine-tuned with a dataset of $171,461$ samples.

with artifacts $\hat{I}^i$ and the corresponding confidence map $\mathcal{C}^i$, CLIP features of source images $c^i_{clip} \in \mathcal{R}^{M\times 768}$, and plücker embeddings of source and target cameras $c^i_{geo} \in \mathcal{R}^{(M+1)\times 78}$, to "teach" the diffusion model how to inpaint missing details and detect Gaussian artifacts guided by the confidence map, context and geometry features and generate a clean version of the conditioning image. For this, we build a dataset generation pipeline comprising a high-quality 3DGS model fitted to dense views, a low-quality 3DGS model fitted to few views, and camera interpolation and perturbation modules to use supervision of the high-quality model at viewpoints beyond ground truth camera poses. The fine-tuning setup is illustrated in Fig. 6. For a given scene, we fit sparse models for $M \in \{3, 6, 9, 18\}$ number of views. We render I^i using the high-quality model and $\hat{I}^i$, $\mathcal{C}^i$ using the low-quality model. We save the CLIP features c_{clip} and plücker embeddings of the M sparse views and 1 target view per sample for conditioning $\boldsymbol{\epsilon_\theta}$.

2.5 Depth-Augmented Autoencoder Finetuning

For encoding and decoding RGBD images, we customize a Variational AutoEncoder by introducing additional channels in the first and last convolutional layers of the Stable Diffusion VAE. A similar approach was followed by Stan et al. (2023), where their KL-autencoder was finetuned with triplets containing RGB images, depth maps, and captions to train the weights in the new channels. However, the depth maps used for fine-tuning this VAE were estimated using MiDaS(Ranftl et al., 2020), which are usually blurry monocular depth estimates. As such, reconstructing RGBD images using this VAE produces depth maps with extreme blur - not ideal for a scene reconstruction problem. Hence, we further finetune this VAE with our synthetic dataset, which contains depth maps rendered by the differentiable 3DGS rasterizer, giving accurate pixel depth with high-frequency details. Specifically, we use the following objective:

$$\mathcal{L}_{\text{autoencoder}} = \min_{\mathcal{E},D} \max_{D_\psi} \big(\mathcal{L}_{\text{rec}}(x, D(\mathcal{E}(x))) - \mathcal{L}_{\text{adv}}(D(\mathcal{E}(x))) + \log(D_\psi(x)) + \mathcal{L}_{\text{reg}}(x; \mathcal{E}, D)\big) \quad (6)$$

where $\mathcal{L}_{rec}$ is a combination of L1, perceptual losses for the RGB channels, and Pearson Correlation Coefficient (PCC), TV regularization losses for the depth channels. $\mathcal{L}_{\text{adv}}$ is the adversarial loss, D_ψ is a patch-based discriminator loss, and $\mathcal{L}_{\text{reg}}$ is the KL-regularisation loss. The incorporation of

PCC and TV terms for the depth channels leads to better retention of high-frequency details in the reconstructed depth map, as observed in Fig 7. We finetune this VAE on a subset of our dataset for 5000 training steps with batch size 16 and learning rate 1e-05.

2.5.1 UNet Finetuning

With our finetuned autoencoder, we next train the UNet with the frozen VAE on $\mathcal{X}$ with the following objective:

$$\mathcal{L} = \mathbb{E}_{i\sim\mathcal{U}(N),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon_t - \boldsymbol{\epsilon_\theta}(\mathbf{z}_t^i; t, \mathcal{E}(\hat{I}), \hat{\mathcal{C}}, \mathcal{E}(I \hat{\cdot} C), c_{attn}^{clip}, c_{attn}^{geo})\|_2^2\right] \tag{7}$$

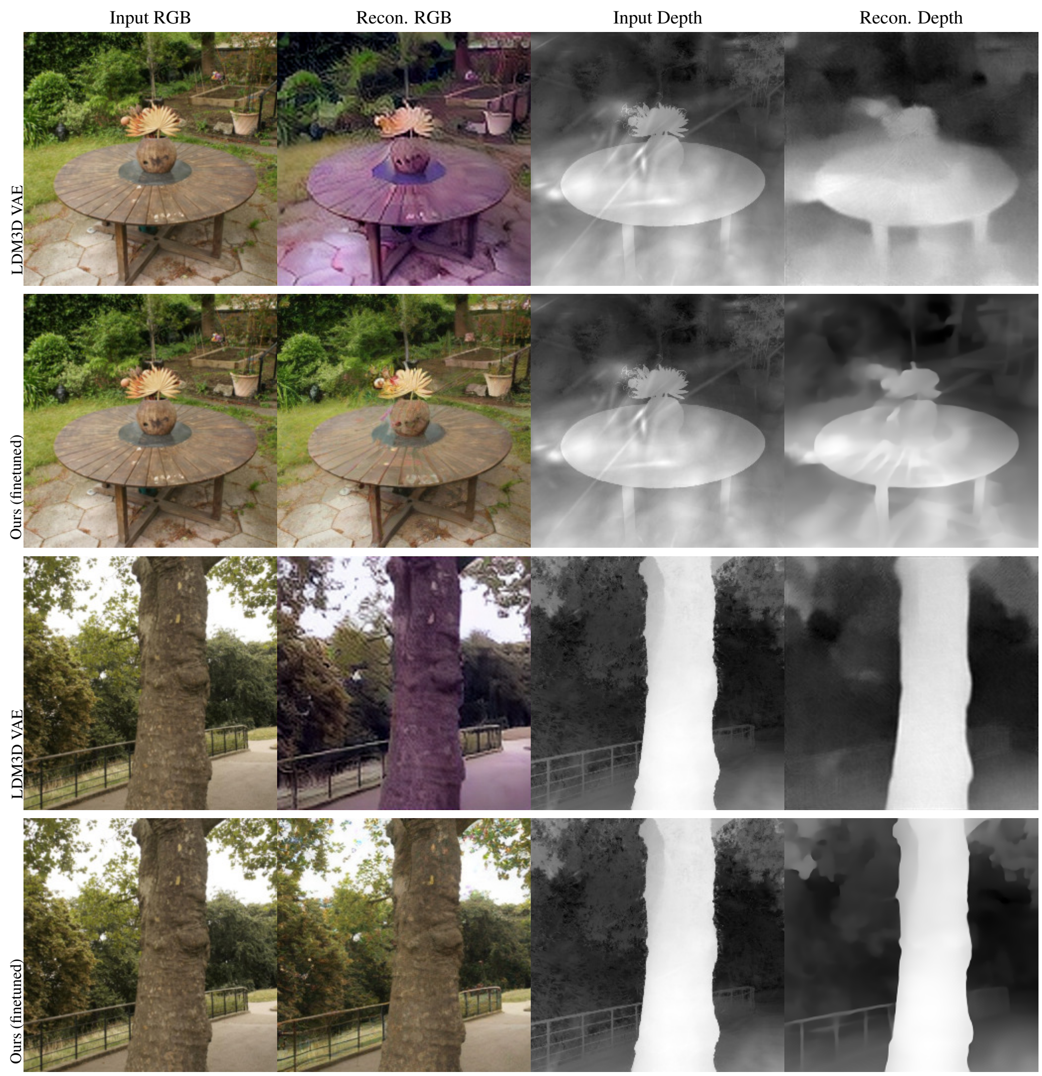


Figure 7: RGBD reconstruction comparison of our finetuned VAE with the LDM3D VAE (Stan et al., 2023). Unlike LDM3D, our VAE finetuned on a synthetic dataset preserves sharp details and edges of the input depth map while also preventing color artifacts in RGB.

2.5.2 RGBD Novel View Synthesis

At inference time, given a render and depth map with artifacts, and confidence map, CLIP features and camera embeddings for conditioning, the finetuned UNet $\boldsymbol{\epsilon_\theta}$ learns to predict the noise in latent $\mathbf{z}_t$ according to $t \sim \mathcal{U}[t_{min}, t_{max}]$ as:

$$\begin{aligned}\hat{\epsilon}_t = &\boldsymbol{\epsilon_\theta}(\mathbf{z}_t; t, \varnothing, \varnothing, \varnothing, c^{clip}_{attn}, c^{geo}_{attn}) \\ &+ s_I(\boldsymbol{\epsilon_\theta}(\mathbf{z}_t; t, \mathcal{E}(\hat{I}), \hat{\mathcal{C}}, \mathcal{E}(I\hat{\cdot}C), c^{clip}_{attn}, c^{geo}_{attn}) - \boldsymbol{\epsilon_\theta}(\mathbf{z}_t; t, \varnothing, \hat{\mathcal{C}}, \varnothing, c^{clip}_{attn}, c^{geo}_{attn})) \\ &+ s_C(\boldsymbol{\epsilon_\theta}(\mathbf{z}_t; t, \varnothing, \hat{\mathcal{C}}, \varnothing, c^{clip}_{attn}, c^{geo}_{attn}) - \boldsymbol{\epsilon_\theta}(\mathbf{z}_t; t, \varnothing, \varnothing, \varnothing, c^{clip}_{attn}, c^{geo}_{attn}))\end{aligned} \tag{8}$$

where s_I and s_C are the RGBD image and confidence map guidance scales, dictating how strongly the final multistep reconstruction agrees with the RGBD render $\hat{I}$ and the confidence map $\mathcal{C}$, respectively. After $k = 20$ DDIM (Song et al., 2021) sampling steps, we obtain our final RGBD render by decoding the denoised latent as $x_\pi = [I_\pi, D_\pi] = \mathcal{D}(z_0)$.

2.6 Scene Reconstruction with Diffusion Priors

Our diffusion priors infer plausible detail in unobserved regions. Despite view conditioning using pose embeddings, the generated images at novel poses lack complete 3D consistency. For this, we devise an iterative strategy where we first sample novel views along a B-spline trajectory fitted to the training views. We initialize the Gaussian optimization with the set of Gaussians $\mathcal{G}$ fitted to the training views (Sec 2.1). Each novel view is added to the training stack at the beginning of every densification step to encourage the optimization to adjust to the distilled scene priors. At every iteration, we sample either an observed or unobserved viewpoint from the current training stack. We bring back Adaptive Density Control to encourage densification of Gaussians in 0-Gaussian regions. We employ the 3DGS objective for the training views. For novel views, we employ the SparseFusion Zhou and Tulsiani (2023) objective in the RGB space (Wu et al., 2024) and a PCC loss for the rendered and denoised depths.

$$\mathcal{L}_{sample}(\psi) = \mathbb{E}_{\pi,t}\Big[w(t)(\|I_\pi - \hat{I}_\pi\|_1 + \mathcal{L}_p(I_\pi, \hat{I}_\pi))\Big] + w_d \cdot PCC(D_\pi, \hat{D}_\pi) \tag{9}$$

where $\mathcal{L}_p$ is the perceptual loss Zhang et al. (2018), $w(t)$ a noise-dependent weighting function, I_π, D_π are the rendered image and depth at novel viewpoint π, and $\hat{I}_\pi, \hat{D}_\pi$ are their rectified versions obtained with our diffusion prior. The PCC loss is defined as $PCC(D_\pi, \hat{D}_\pi) = 1 - \frac{\text{Cov}(D_\pi, \hat{D}_\pi)}{\sigma_{D_\pi}\sigma_{\hat{D}_\pi}}$.

2.7 Test-time Pose Alignment

GScenes reconstructs a plausible 3D scene from unposed source images. However, reconstruction with few views is inherently ambiguous as several solutions can satisfy the train view constraints. Hence, the reconstructed scene would most likely be quite different from the actual scene from which M views were sampled. Hence, for a given set of test views, following prior work (Fan et al., 2024; Jiang et al., 2024), we freeze the Gaussian attributes and optimize the camera pose for each target view by minimizing a photometric loss between the rendered image and test view. Following this alignment step performed for 500 iterations per test image, we evaluate the NVS quality.

3 Experiments

We compare *GScenes* with state-of-the-art pose-free and pose-required sparse-view reconstruction methods in Fig 8, 9 and Table 1, 2. We also ablate the different components and design choices of our diffusion model.

3.1 Experimental Setup

Evaluation Dataset We evaluate GScenes on the 9 scenes of the MipNeRF360 dataset (Barron et al., 2022), and 15 scenes (out of 140) of the DL3DV-10K benchmark dataset Ling et al. (2024). For MipNeRF360, We pick the M-view splits as proposed by ReconFusion Wu et al. (2024) and CAT3D Gao* et al. (2024) and evaluate all baselines on the official test views where every 8^{th} image is held out for testing. For DL3DV-10K scenes, we create M-view splits using a greedy view-selection heuristic for maximizing scene coverage given a set of dense training views, similar to the heuristic proposed in Wu et al. (2024). For test views, we hold out every 8^{th} image as in

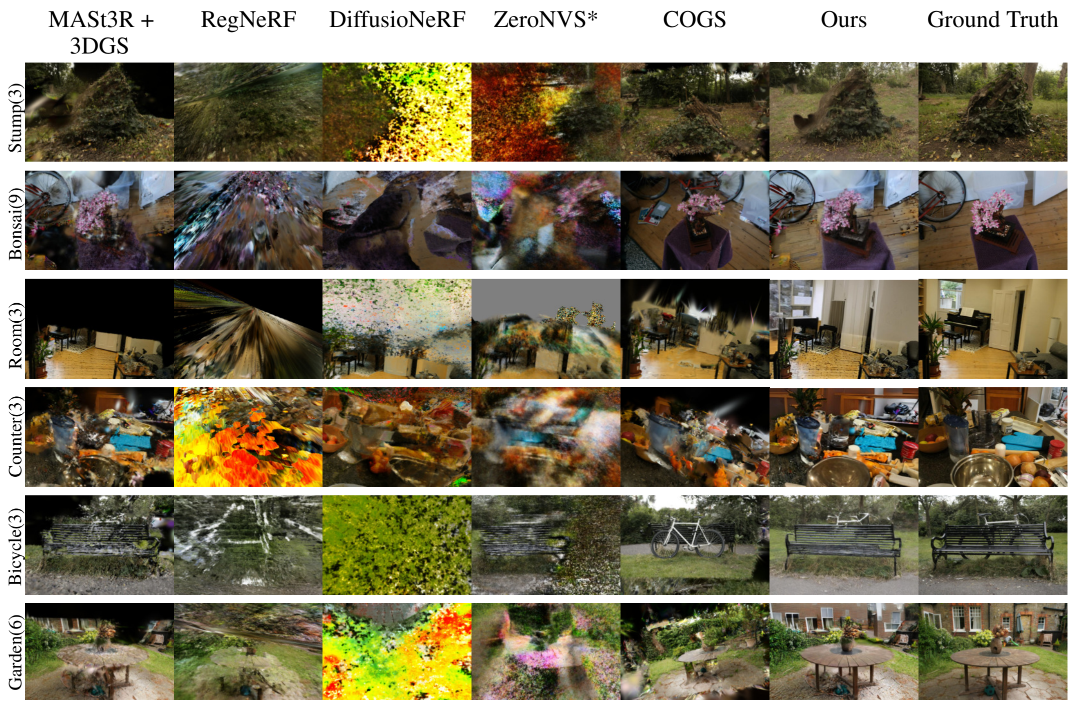


Figure 8: **Qualitative comparison** of GScenes with few-view methods. Our approach consistently fairs better in recovering image structure from foggy geometry, where baselines typically struggle with "floaters" and color artifacts.

MipNeRF360. Additionally, we pick the *plant* scene of CO3D for qualitative comparison with ReconFusion and CAT3D.

Fine-tuning Dataset We fine-tune our diffusion model on a mix of 1043 scenes encompassing Tanks and Temples (Knapitsch et al., 2017), CO3D (Reizenstein et al., 2021), Deep Blending (Hedman et al., 2018), and the $1k$ subset of DL3DV-10K Ling et al. (2024) to obtain a total of $171,461$ data samples. We first train 3DGS on sparse and dense subsets of each scene for $M \in \{3, 6, 9, 18\}$. For $M > 18$, novel view renders and depth maps mostly show Gaussian blur as artifacts. Finetuning this model takes about 4-days on a single A6000 GPU.

Metrics Our quantitative metrics are used to evaluate two tasks - quality of novel views post reconstruction and camera pose estimation. For the former, we compute 3 groups of metrics - FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018) due to the generative nature of our approach, perceptual metrics LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) to measure similarity in image structure and texture in the feature space, and pixel-aligned metrics PSNR and SSIM. However, PSNR and SSIM are not suitable evaluators of generative techniques (Chan et al., 2023; Sargent et al., 2024) as they favor pixel-aligned blurry estimates over high-frequency details.

Baselines We compare our approach against 8 baselines. FreeNeRF (Yang et al., 2023), RegNeRF (Niemeyer et al., 2022), DiffusioNeRF (Wynn and Turmukhambetov, 2023) are pose-required few-view regularization methods based on NeRFs. ZeroNVS (Sargent et al., 2024) reconstructs a complete 3D scene from a single image using a novel camera normalization scheme and anchored SDS loss. We use the ZeroNVS* baseline introduced in ReconFusion (Wu et al., 2024), designed to adapt ZeroNVS to multi-view inputs. ReconFusion and CAT3D are state-of-the-art multi-view conditioned diffusion models for sparse-view reconstruction. We use the reported average performance on the 9 scenes of MipNeRF360 with classical metrics - PSNR, SSIM, and LPIPS for quantitative comparison and pick the relevant scenes and test views from the two papers for qualitative comparison. For pose-free methods, we pick our 3D reconstruction engine (MASt3R + 3DGS) and COGS (Jiang et al., 2024) and show that our approach outperforms them in NVS both qualitatively and quantitatively.

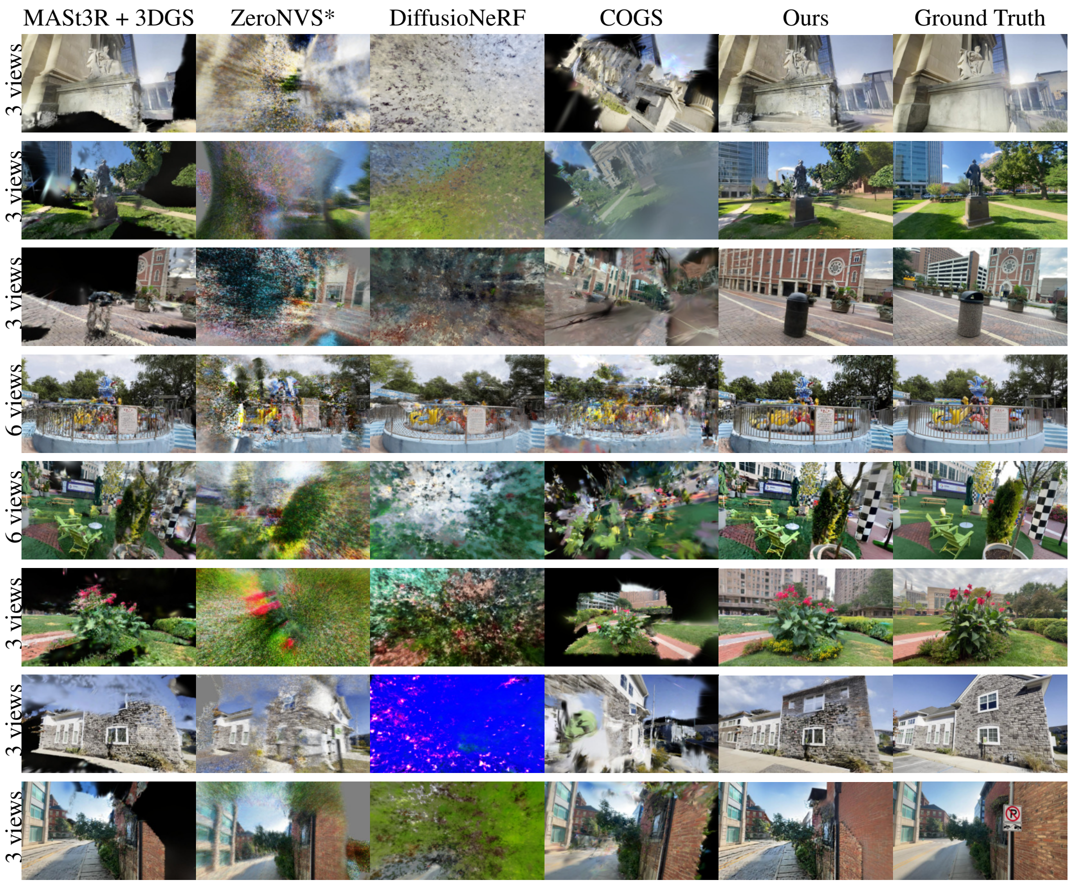


Figure 9: **Qualitative comparison** of *GScenes* with few-view methods on the DL3DV benchmark. Our method achieves plausible reconstruction in unobserved areas of complex scenes where even posed reconstruction techniques struggle.

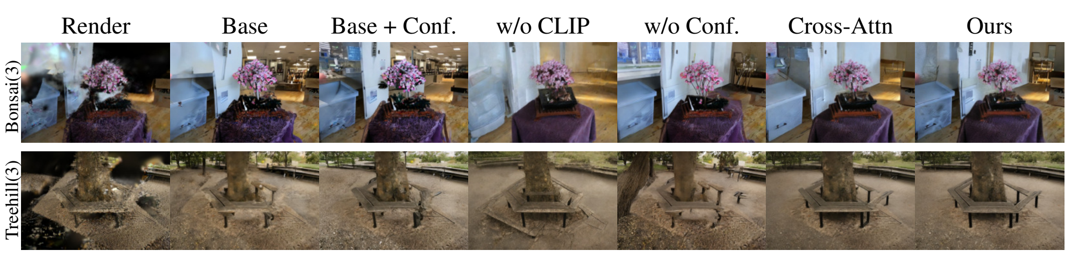


Figure 10: Ablation Study with the 3-view splits of *Bonsai* and *Treehill* scenes from MipNeRF360. Images above are samples from different variants of our diffusion model depending on the conditioning.

3.2 Implementation Details

Our framework is implemented in PyTorch 2.3.1 on single A5000/A6000 GPUs. Images and depth maps are rendered at 400-600 pixels to align with Stable Diffusion's resolution. The diffusion model is finetuned for $100k$ iterations (batch size 16, learning rate $1e-4$) with conditioning element dropout probability of 0.05 for classifier-free guidance.

Following Fan et al. (2024), we fit 3D Gaussians to sparse inputs and MASt3r point clouds for $1k$ iterations to obtain $\mathcal{G}$. We use classifier-free guidance scales $s_I = s_C = 3.0$ and sample with $k = 20$ DDIM steps. We linearly decay w_d from 1 to 0.01 and $\mathcal{L}_{sample}$ weight from 1 to 0.1 over $10k$ iterations. *GScenes* completes full 3D reconstruction in approximately 5 minutes on a single A6000 GPU.

3.3 Comparison Results

Table 1: **Quantitative comparison** with state-of-the-art sparse-view reconstruction techniques on classical metrics.

Method	PSNR ↑ 3-view	PSNR ↑ 6-view	PSNR ↑ 9-view	SSIM ↑ 3-view	SSIM ↑ 6-view	SSIM ↑ 9-view	LPIPS ↓ 3-view	LPIPS ↓ 6-view	LPIPS ↓ 9-view
FreeNeRF	11.888	12.877	13.680	0.146	0.180	0.197	0.675	0.654	0.638
RegNeRF	12.297	13.209	13.802	0.147	0.170	0.180	0.668	0.656	0.625
DiffusioNeRF	13.134	16.191	16.732	0.167	0.283	0.337	0.680	0.543	0.530
ZeroNVS*	11.902	11.789	11.729	0.142	0.133	0.128	0.710	0.702	0.694
ReconFusion	15.50	16.93	18.19	0.358	0.401	0.432	0.585	0.544	0.511
CAT3D	16.62	17.72	18.67	0.377	0.425	0.460	0.515	0.482	0.460
MASt3R + 3DGS	13.657	14.733	14.926	0.222	0.265	0.246	0.603	0.540	0.590
COGS	12.267	12.428	13.030	0.173	0.186	0.201	0.616	0.632	0.596
GScenes	14.976	15.972	16.563	0.284	0.295	0.314	0.593	0.517	0.577

Table 2: **Quantitative comparison** with few-view reconstruction techniques on metrics suited for generative reconstruction.

Method	FID ↓ 3-view	FID ↓ 6-view	FID ↓ 9-view	KID ↓ 3-view	KID ↓ 6-view	KID ↓ 9-view	DISTS ↓ 3-view	DISTS ↓ 6-view	DISTS ↓ 9-view
FreeNeRF	342.484	343.378	334.805	0.252	0.262	0.268	0.376	0.363	0.357
RegNeRF	343.593	335.606	325.228	0.260	0.276	0.257	0.387	0.383	0.371
DiffusioNeRF	323.047	269.995	246.265	0.230	0.166	0.141	0.395	0.325	0.328
ZeroNVS*	356.395	350.362	343.930	0.283	0.294	0.299	0.433	0.427	0.413
MASt3R + 3DGS	268.526	226.110	242.853	0.191	0.131	0.140	0.294	0.298	0.310
COGS	231.274	274.827	248.220	0.136	0.183	0.152	0.288	0.310	0.286
GScenes	253.659	220.292	210.255	0.176	0.098	0.106	0.284	0.245	0.241

We report qualitative and quantitative comparisons of *GScenes* against all related baselines in Fig 8 and 9 and Tables 1 and 2. Out of the 8 baselines, only *MASt3R + 3DGS* and *COGS* are pose-free techniques, while the remaining require ground truth poses for both training and evaluation. Note that *COGS* relies on ground truth camera intrinsics while *GScenes* and *MASt3R + 3DGS* do not. On the classical metrics (Tab 1), we are third behind the 2 SOTA posed reconstruction methods - *CAT3D* and *ReconFusion*. Except for FID and KID on the 3-view split, we outperform all related baselines in Table 2. Due to the unavailability of open-source code, evaluating *ReconFusion* and *CAT3D* on these measures is unfortunately not possible.

3.4 Ablation Studies

In Fig 10, we thoroughly ablate different components of our diffusion model. We pick the *bonsai* and *treehill* scenes and their 3-view splits for this experiment. The leftmost column shows the initial novel-view render obtained from the 3D reconstruction pipeline (MASt3r + 3DGS). The *Base* variant only performs image-to-image diffusion with no other conditioning, and this already provides a strong baseline for inpainting and artifact elimination in novel-view renders. Without CLIP context guidance, the model fails to adhere to the semantics of the input images when inpainting missing details. Without our confidence measure, the model typically fails to differentiate between image structures and artifacts, often producing implausible images despite context and geometry conditioning. Additionally, we train a *Cross-Attn* variant where context and geometry conditionings are incorporated using cross-attention instead of FiLM modulation layers in the down and middle blocks of the UNet. Our method achieves similar performance with more than 2x fewer additional trainable parameters.

3.5 Qualitative Comparison with ReconFusion / CAT3D

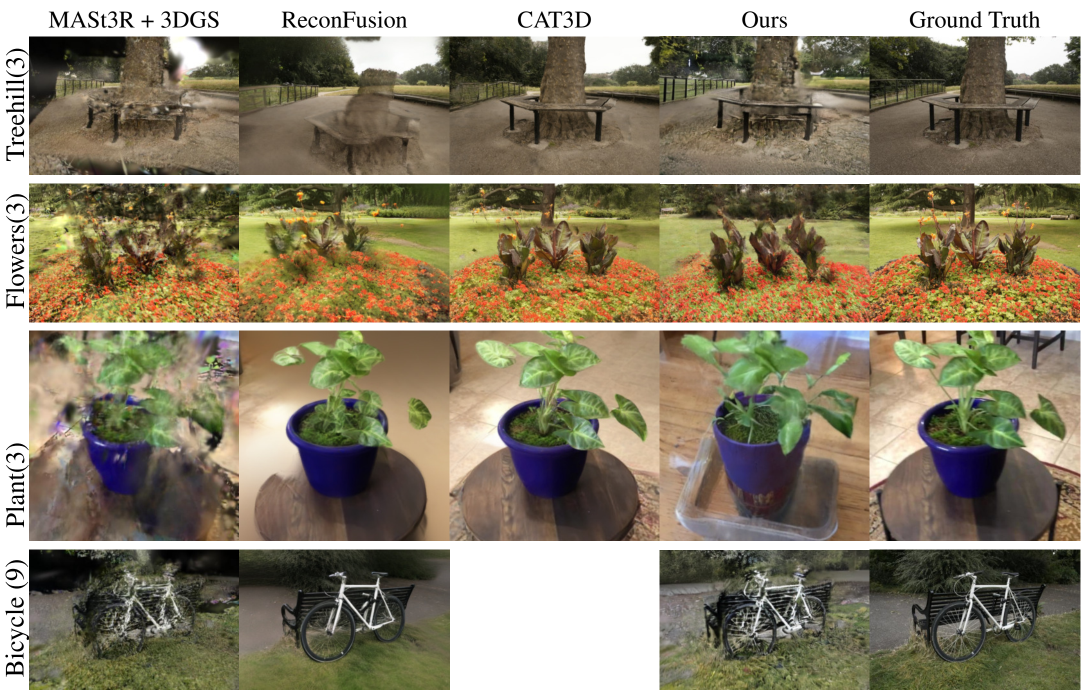


Figure 11: **Qualitative comparison** of *GScenes* with ReconFusion and CAT3D (posed techniques). Despite being a pose-free pipeline built with weaker diffusion priors, our method achieves competitive NVS quality with SOTA sparse-view reconstruction techniques. No image available for CAT3D in the last row, hence kept blank.

In Fig 11, we provide an additional qualitative comparison with ReconFusion and CAT3D. Their code is not available, and hence, we could not perform an evaluation across all test scenes in their paper. From the figures in the 2 papers, we pick the relevant test views for the *treehill*, *flowers*, *bicycle* scenes in MipNeRF360, and the *plant* scene from CO3Dv2 to show how *GScenes* compares with their reconstruction. We use the same training views as open-sourced in their data splits. Despite being a pose-free pipeline using weaker generative priors, we observe that *GScenes* compares competitively with both methods.

4 Conclusion

In this work, we present *GScenes* where we integrate an image-to-image RGBD diffusion model with a pose-free reconstruction pipeline in MASt3R to reconstruct a 360° 3D scene from a few uncalibrated 2D images. We introduce context and geometry conditioning through FiLM layers achieving similar performance as a cross-attention variant. We also introduce a pixel-aligned confidence measure to further guide the diffusion model in uncertain regions with missing details and artifacts. Our experiments show that *GScenes* outperforms existing pose-free reconstruction methods in scene reconstruction and performs competitively with state-of-the-art posed sparse-view reconstruction methods.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 9

Wenjing Bian, Zirui Wang, Kejie Li, Jiawang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. 2023. 20

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 10

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 1, 18

Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *ICCV*, 2023. 10, 19

David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024. 19

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, 2021. 19

Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023a. 19

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023b. 19

Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24–34, 2023. 20

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. 2024. 19

Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *arXiv*, 2023. 1, 18, 19

Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 2023. 19

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 1, 18, 19

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. In *IEEE TPAMI*, 2020. 10

Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. In *arXiv*, 2024. 1, 3, 9, 12, 18, 20

Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. 2023. 3, 20

Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv*, 2024. 1, 2, 9, 18, 19

Pengsheng Guo, Miguel Angel Bautista, Alex Colburn, Liang Yang, Daniel Ulbricht, Joshua M. Susskind, and Qi Shan. Fast and explicit neural view synthesis. In *WACV*, 2022. 19

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. In *ACM TOG*, 2018. 10

Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *CVPR*, 2021. 19

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 10

Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In *ICCV*, 2023. 19

Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. 1, 18, 19

Kaiwen Jiang, Yang Fu, Mukund Varma T, Yash Belhe, Xiaolong Wang, Hao Su, and Ravi Ramamoorthi. A construct-optimize approach to sparse view synthesis without camera pose. *SIGGRAPH*, 2024. 1, 9, 10, 18, 20

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM TOG*, 2023. 1, 18, 19

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 18

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. In *ACM TOG*, 2017. 10

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 1, 2, 3, 18, 20

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *CVPR*, 2024. 1, 18, 19

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 20

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023a. 19

Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023b. 19

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 9, 10

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 1, 18, 20

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *arXiv preprint arXiv:2410.16266*, 2024. 6

Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *CVPR*, 2023. 19

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. In *ECCV*, 2020. 1, 18

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, 2023. 19

Bharath Raj Nagoor Kani, Hsin-Ying Lee, Sergey Tulyakov, and Shubham Tulsiani. Upfusion: Novel view diffusion from unposed sparse view observations. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 18

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 1, 18

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 1, 10, 18, 19

Soumava Paul, Christopher Wewer, Bernt Schiele, and Jan Eric Lenssen. Sp2360: Sparse-view 360○ scene reconstruction using cascaded 2d diffusion priors. In *ECCV 2024 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild*, 2024. 19, 20

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 2, 5, 20

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 19

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 19

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. In *arXiv*, 2022. 1, 18

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *IEEE TPAMI*, 2020. 1, 7, 18

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 10

Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, 2022. 1, 18, 19

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4, 7, 18

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 18

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*, 2024. 1, 10, 18, 20

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3

Philipp Schröppel, Christopher Wewer, Jan Eric Lenssen, Eddy Ilg, and Thomas Brox. Neural point cloud diffusion for disentangled 3d shape and appearance generation. In *CVPR*, 2024. 19

J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, 2023. 19

Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. NeurIPS*, 2021. 4

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 9

Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, and Vasudev Lal. Ldm3d: Latent diffusion model for 3d, 2023. 7, 8

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 19

Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. In *ICCV*, 2021. 19

Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*, 2023a. 1, 18, 19

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023b. 19

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 18

Christopher Wewer, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. SimNP: Learning self-similarity priors between neural points. In *ICCV*, 2023. 19

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *arXiv*, 2024. 19

Chin-Hsuan Wu, Yen-Chun Chen, Bolivar Solarte, Lu Yuan, and Min Sun. ifusion: Inverting diffusion for pose-free reconstruction from sparse views. *arXiv preprint arXiv:2312.17250*, 2023. 1, 18

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024. 1, 2, 9, 10, 18, 19, 20

Jamie Wynn and Daniyar Turmukhambetov. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *CVPR*, 2023. 1, 10, 18, 19

Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360° sparse view synthesis using gaussian splatting. In *arXiv*, 2023. 1, 18, 19

Chen Yang, Sikuang Li, Jiemin Fang, Ruofan Liang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. *ACM Transactions on Graphics*, 2024. 1, 18

Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *CVPR*, 2023. 10, 19

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 19, 20

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 9, 10

Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, 2021. 19

Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 9, 20

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *arXiv*, 2023. 1, 18, 19

A Problem Introduction and Discussion

Obtaining high-quality 3D reconstructions or novel views from a sparse set of images has been a long-standing goal in computer vision. Recent methods for sparse-view reconstruction often employ generative, geometric, or semantic priors to stabilize the optimization of NeRFs (Mildenhall et al., 2020) or Gaussian splats (Kerbl et al., 2023) in highly under-constrained scenarios. However, they typically assume access to accurate intrinsic and extrinsic parameters, often derived from dense observations. This reliance on ground-truth poses is a restrictive assumption, making these methods impractical for real-world applications. Moreover, pose estimation from sparse views is error-prone; both traditional Structure from Motion approaches and recent 3D foundational models (Wang et al., 2024; Leroy et al., 2024) struggle with insufficient matching features between image pairs. In response, recent pose-free approaches using 3D Gaussian splatting (3DGS) integrate monocular depth estimation (Ranftl et al., 2020), 2D semantic segmentation (Kirillov et al., 2023), or 3D foundational priors (Wang et al., 2024), optimizing 3D Gaussians and camera poses together during training. However, these methods are typically designed for scenes with high view overlap, and they often fail to reconstruct complex, large-scale 360° scenes with sparse coverage. Additionally, despite extensive regularization to prevent overfitting, the limited observations impede coherent synthesis of unobserved regions. This challenge underlines the need for additional generative regularization to enable accurate extrapolation and complete scene reconstruction.

Despite several previous attempts (Deng et al., 2022; Roessle et al., 2022; Li et al., 2024; Xiong et al., 2023; Zhu et al., 2023; Chung et al., 2023; Wang et al., 2023a; Jain et al., 2021; Niemeyer et al., 2022; Wynn and Turmukhambetov, 2023) in the field of sparse novel view synthesis, only a few - COGS Jiang et al. (2024), InstantSplat (Fan et al., 2024) have been successful in dealing with the problem without precomputed camera parameters. These methods jointly optimize Gaussians and camera parameters during training and use different 2D priors for regularization. Although they find success in artifact and blur reduction, they are less effective in 360° scene reconstruction due to a lack of generative priors.

Reconstructing full 360° scenes from limited views requires robust priors, such as those from powerful 2D image or video diffusion models (Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022) that encode common 3D structures. Recent methods (Liu et al., 2023; Sargent et al., 2024; Wu et al., 2024; Gao* et al., 2024; Blattmann et al., 2023) incorporate view and context conditioning into these models, fine-tuning them on extensive real-world and synthetic multi-view datasets to support radiance field optimization. This enables realistic extrapolation into unobserved areas of complex 360° scenes. However, these models depend on accurate pose information, making them unsuitable for scenarios where pose estimates deviate significantly from ground truth, leading to misaligned view generation. To our knowledge, only Gaussian Object (Yang et al., 2024), iFusion (Wu et al., 2023), and UpFusion (Nagoor Kani et al., 2024) provide pose-free generative solutions, but they are designed specifically for 3D object reconstruction, not large-scale scenes.

In the absence of poses estimated from dense observations, we instead rely on recent 3D foundational priors Leroy et al. (2024) for scene initialization from sparse views and encode its estimated cameras for conditioning a Stable Diffusion UNet Rombach et al. (2022) during both training and evaluation. We also augment the UNet with additional channels for context, 3DGS render, depth maps with Gaussian artifacts, and a pixel-aligned confidence map capturing missing regions and reconstruction artifacts in the RGBD image. During inference, the model predicts a clean, inpainted version of the conditioning artifact image and an aligned depth map. This formulation prevents the requirement of accurate ground truth poses for pose conditioning of a multiview diffusion model. This also alleviates dependency on large-scale 3D datasets which are usually synthetic and low quality compared to real-world scenes.

We present *GScenes*, an efficient method that uses 3D foundational (Leroy et al., 2024) and RGBD diffusion priors for pose-free sparse-view reconstruction of complex 360° scenes. We first estimate a point cloud and approximate scene intrinsics and extrinsics using MASt3R and then jointly optimize both Gaussians and camera parameters with just a few iterations of 3DGS. We then generate novel views by sampling from a B-spline trajectory fitted to the training views. Resulting renders and depth maps contain missing details and Gaussian artifacts that are refined by our diffusion prior and then used to further optimize the 3D Gaussians and camera poses.

Despite using weaker generative priors compared to state-of-the-art pose-dependent methods, we show competitive qualitative and quantitative performance with ReconFusion (Wu et al., 2024) and CAT3D (Gao* et al., 2024) and comprehensively outperform all other pose-dependent and pose-free techniques on reconstruction quality. Our method leverages stronger priors than simple regularizers while not relying on million-scale multi-view data or huge compute resources to train a 3D-aware diffusion model. We also ablate all conditioning elements in our diffusion model and identify which conditioning features contribute the most towards coherent NVS from sparse views.

B Related Work

Reconstructing 3D scenes from limited observations requires generative priors or, more specifically, inpainting missing details in unseen regions in 3D and removing artifacts introduced through observing scene areas from a few observations. Our work builds on recent developments in 2D diffusion priors for 3D reconstruction (Paul et al., 2024), where knowledge learned from abundant 2D datasets is lifted to 3D for rectifying novel views rendered by sparse 3D models. Next, we discuss how our work is related to the current line of research.

Regularization Techniques Both NeRF and 3DGS rely on hundreds of scene captures for photorealistic novel view synthesis. When the input set becomes sparse, the problem becomes ill-posed, as several simultaneous 3D representations can agree with the training set. Regularization techniques are amongst the earliest techniques to address this limitation. Typical methods leverage depth from Structure-from-Motion (SfM)Deng et al. (2022); Roessle et al. (2022), monocular estimationLi et al. (2024); Xiong et al. (2023); Zhu et al. (2023); Chung et al. (2023), or RGB-D sensors Wang et al. (2023a). DietNeRF Jain et al. (2021) uses a semantic consistency loss based on CLIP Radford et al. (2021) features, while FreeNeRF Yang et al. (2023) regularizes the frequency range of NeRF inputs. Approaching generative priors, RegNeRF Niemeyer et al. (2022) and DiffusioNeRF Wynn and Turmukhambetov (2023) maximize the likelihoods of rendered patches using normalizing flows or diffusion models, respectively. However, such techniques usually fail under extreme sparsity like 3, 6, or 9 input images for a 360° scene due to weaker priors. Generative priors can be viewed as a stronger form of regularization as they provide extrapolation capabilities for inferring details in unknown parts of a scene.

Generalizable Reconstruction When only a few or a single view is available, regularization techniques are often insufficient to resolve reconstruction ambiguities. To address this, recent research focuses on training priors for novel view synthesis across multiple scenes. pixelNeRF Yu et al. (2021) uses pixel-aligned CNN features as conditioning for a shared NeRF MLP, while other approaches Trevithick and Yang (2021); Chen et al. (2021); Henzler et al. (2021); Lin et al. (2023b) condition NeRF on 2D or fused 3D features. Further priors have been learned on triplanes Irshad et al. (2023), voxel grids Guo et al. (2022), and neural points Wewer et al. (2023). Leveraging 3D Gaussian Splatting Kerbl et al. (2023), methods like pixelSplat Charatan et al. (2024) and MVSplat Chen et al. (2024) achieve state-of-the-art performance in stereo view interpolation and real-time rendering. However, these regression-based models often produce blurry outputs under high uncertainty. In contrast, generative methods like GeNVS Chan et al. (2023) and latentSplat Wewer et al. (2024) aim to sample from multi-modal distributions, offering better handling of ambiguous novel views.

Generative Priors for NVS For ambiguous novel views, predicting expectations over all reconstructions may be unreliable. Consequently, regression approaches fall short, whereas generative methods attempt to sample from a multi-modal distribution.

While diffusion models have been applied directly on 3D representations like triplanes Shue et al. (2023); Chen et al. (2023a), voxel grids Müller et al. (2023), or (neural) point clouds Zhou et al. (2021); Melas-Kyriazi et al. (2023); Schröppel et al. (2024), 3D data is scarce. Given the success of large-scale diffusion models for image synthesis, there is a great research interest in leveraging them as priors for 3D reconstruction and generation. DreamFusion Poole et al. (2023) and follow-ups Wang et al. (2023b); Lin et al. (2023a); Chen et al. (2023b); Deng et al. (2023); Tang et al. (2024) employ score distillation sampling (SDS) to iteratively maximize the likelihood of radiance field renderings under a conditional 2D diffusion prior. For sparse-view reconstruction, existing approaches incorporate view-conditioning via epipolar feature transform Zhou and Tulsiani (2023), cross-attention to encoded relative poses Liu et al. (2023); Sargent et al. (2024), or pixelNeRF Yu et al. (2021) feature renderings Wu et al. (2024). However, this fine-tuning is expensive and requires large-scale multi-view data, which we circumvent with *GScenes*.

Pose-Free 3D Reconstruction For building a generalizable sparse-view reconstruction method, the assumption of camera poses during inference limits applications to real-world scenarios where usually only an uncalibrated set of 2D images are available with no known camera extrinsics or intrinsics. Several recent works (Jiang et al., 2024; Fan et al., 2024) have attempted to solve reconstruction in a pose-free setting by jointly optimizing poses and NeRF or 3D Gaussian parameters during scene optimization. These methods typically outperform previous techniques (Chen and Lee, 2023; Lin et al., 2021; Bian et al., 2023; Fu et al., 2023) where reconstruction is done in two stages - first estimating poses and then optimizing the 3D representation. However, errors in the initial pose estimation harm subsequent scene optimization, resulting in inferior NVS quality. In our work, we use the MASt3R Leroy et al. (2024) pipeline with 3DGS for predicting 3D Gaussians and camera parameters in a global coordinate system from a set of unposed 2D images.

Our work is most closely related to Sp^2360 (Paul et al., 2024) where an instruction-following RGB diffusion model is finetuned for the task of rectifying novel views rendered by 3DGS fitted to sparse observations. We extend the problem setting to the more challenging pose-free scenario and jointly model RGB and depth to aid optimization of 3D Gaussians during the distillation phase. Additionally, we introduce CLIP context and pose conditioning through FiLM layers (Perez et al., 2018) and a pixel-aligned confidence measure for more accurate novel view synthesis.

C LIMITATIONS & FUTURE WORK

Table 3: **Pose estimation accuracy with *GScenes***. SfM-poses for training views estimated from the full observation set are used as ground truth. We report errors in camera rotation and translation using Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) as in Bian et al. (2023). Despite the DUSt3r initialization of camera extrinsics and subsequent pose optimization, there are large errors w.r.t COLMAP poses of the dense observation set, harming test-pose alignment and subsequent NVS quality.

	$\text{RPE}_t \downarrow$			$\text{RPE}_r \downarrow$			ATE ↓		
	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view
GScenes	37.684	27.893	30.916	124.501	58.915	40.424	0.261	0.294	0.266

GScenes is a first step towards a generative solution for pose-free sparse-view reconstruction of large complex scenes. However, it is not free of its fair share of limitations. The quality of the final reconstruction depends heavily on the initial relative pose estimation by the MASt3r pipeline, and even though the poses are further optimized jointly with Gaussian attributes during training, there are still large differences with the ground truth COLMAP poses estimated from dense views as we show in Tab 3. This limits fair comparison with posed reconstruction methods, as even the test-time pose alignment step (Sec 3.5) cannot compensate for the initial errors in the pipeline. Our diffusion model, much like related methods, is not agnostic to the 3D representation from which novel view renders, depth, and confidence maps are obtained for fine-tuning the diffusion model for 3D-aware sparse-view NVS. Moreover, our diffusion model trained with 3DGS renders with MASt3R initialization would not be able to rectify novel views from a 3DGS representation with SfM initialization due to the slight difference in the distribution of rendered images. To ensure multiview consistency across all synthesized views, we employ view conditioning in the form of plucker embeddings and use a fixed noise latent across all novel views for multistep reconstruction. However, this does not alleviate the multiview consistency issue completely as novel views are synthesized in an autoregressive manner and not simultaneously synthesized like in video diffusion models. We aim to address these limitations of our diffusion model in future work.