This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

HAD: Hallucination-Aware Diffusion Priors for 3D Reconstruction

Xi Liu[1,2*] Weiwei Sun[1†] Zhou Ren[1] Chris Broaddus[1] Siyu Huang[2] Laurent Guigues[1]
[1]Amazon AWS [2]Clemson University

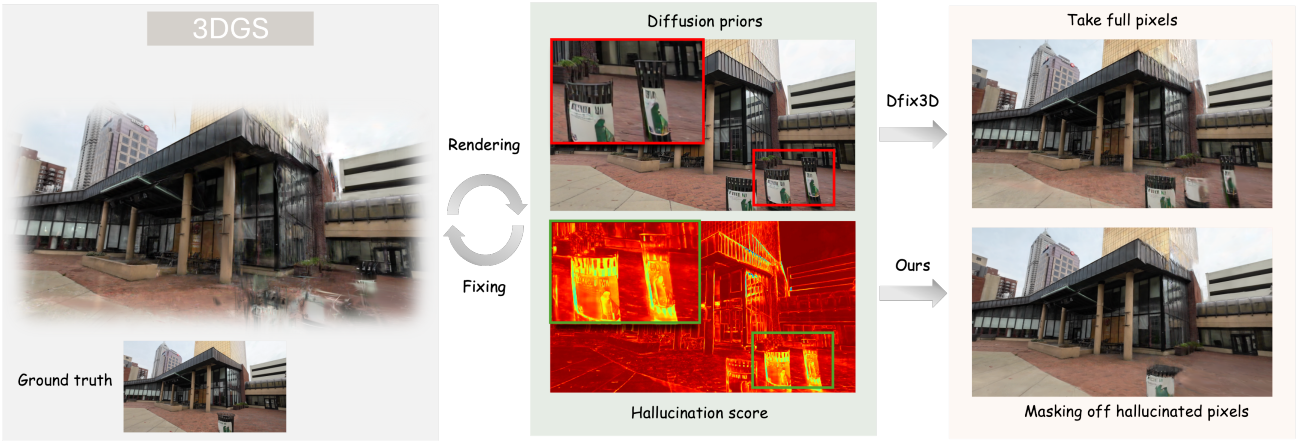


Figure 1. While diffusion priors [41] enhance the quality of 3D reconstruction, they introduce detrimental *aliens* – the hallucinated elements that do not exist in the observed regions, as highlighted in boxes. This work addresses this issue through hallucination score modeling, achieving high-quality 3D reconstruction with improved fidelity.

Abstract

Diffusion priors have recently demonstrated strong capability in enhancing the quality of sparse-view 3D reconstruction by augmenting training views at novel viewpoints, but they inevitably introduce hallucinated content – artifacts inconsistent with the input views – into the final 3D model. To address this challenge, we propose Hallucination-Aware Diffusion prior (HAD), which estimates pixel-wise hallucination score maps for augmented images by leveraging multi-view reasoning capabilities from a feedforward novel view synthesis (NVS) network pre-trained on large-scale 3D data. These hallucination scores enable selective masking of unreliable pixels during the progressive 3D reconstruction procedure, preventing the introduction of non-existent artifacts into the 3D model. To further enhance performance, we create multiple versions of augmented images at each novel view by conditioning the diffusion prior on different input views, which are then fused into a final image that leverages the broader context across all input views. We show that our method substantially reduces hallucination artifacts in diffusion-assisted 3D reconstruction, thereby achieving state-of-the-art performance across multiple benchmarks on novel view synthesis.

1. Introduction

Neural Radiance Fields (NeRF) [31] and 3D Gaussian Splatting (3DGS) [21] have emerged as groundbreaking approaches in 3D reconstruction, demonstrating remarkable capabilities in novel view synthesis (NVS) from multiple 2D views. Despite their significant achievements, these methods [21, 22, 31] require dense camera coverage and high-fidelity image inputs to maintain reliable performance. This dependency severely limits performance in data-sparse scenarios, such as sparse-view settings and extreme novel-view extrapolation tasks where the quality of rendered images degrades dramatically, as illustrated in Fig. 1. One approach to address data sparsity is to leverage generative diffusion priors to augment novel-view data by removing artifacts from rendered images through denoising conditioned on the original input views [11, 28, 41].

While these augmented views achieve photorealism, they often fail to maintain fidelity to the original input views. The root cause is that the diffusion process, by design, does not strictly preserve the content of the conditional input views, resulting in multi-view inconsistency in the generated images. Consequently, when this hallucinated content from augmented views is incorporated into the 3D

*Work done at Amazon.
†Corresponding author.

model, it leads to a *hallucination issue* in 3D reconstruction, where rendered views appear high-quality but exhibit low fidelity to the input views. While recent works [11, 28, 47] mitigate this issue by enhancing the ability of diffusion priors to respect input views – e.g., through video diffusion models [43, 51], multi-view inconsistency – as shown in Sec. 7 of `supplementary material`– still remains in generated images due to the generative nature of diffusion priors.

To address this limitation, we propose incorporating hallucination awareness into the augmented views. Rather than preventing hallucination entirely, we retain the ability to filter out hallucinated content when incorporating the augmented images into 3D models. Accordingly, we coin our method as Hallucination-Aware Diffusion priors (HAD). Specifically, we propose a hallucination score network to predict pixel-wise hallucination scores for augmented views, quantifying the inconsistency between augmented views and the original input multi-views. The network consists of a multi-view encoder that understands multiple input views and subsequently a hallucination score branch that detects hallucinated content in augmented views.

Inspired by the ability of recent feedforward novel view synthesis (NVS) networks [20] to reconstruct novel views with high fidelity, we leverage the pre-trained feature backbone of a well-trained NVS network as our multi-view encoder. The key insight is that these NVS feature backbones develop strong multi-view reasoning abilities through training on large-scale 3D data. By leveraging this pre-trained knowledge, we can train our hallucination network on a small-scale curated dataset, bypassing the challenging large-scale data curation that would otherwise be required—each training sample must include both input multi-views and augmented novel views, necessitating a complete pipeline of 3D reconstruction, novel view rendering, and diffusion prior denoising.

Our network identifies hallucinated content in augmented images through the hallucination score map, enabling the masking of hallucinated pixels in the augmented views during 3D model optimization. To further enhance the performance of HAD, we propose a multi-sampling strategy that generates multiple images at the same novel view by conditioning the diffusion process on different input views, thereby incorporating a broader context from the input views into HAD. We then fuse these images into an improved image for 3D reconstruction using the hallucination score maps. This strategy increases the number of conditional input views in the diffusion prior without retraining, effectively reducing the ratio of hallucinated content in diffusion-generated images.

We validate our method on two standard NVS benchmarks: DL3DV [25], on which we train HAD for in-domain evaluation, and MipNeRF 360 [3] for cross-domain evaluation. As demonstrated in Sec. 5, our method significantly outperforms state-of-the-art approaches, with PSNR improvements of 0.78dB on DL3DV and 0.69dB on MipNeRF 360. We then summarize our contributions as below:

- We identify a critical limitation where diffusion priors, while alleviating data sparsity in 3D reconstruction, introduce hallucination issues that compromise fidelity to input views despite achieving photorealistic rendering.
- We propose Hallucination-Aware Diffusion priors (HAD), the first framework to explicitly tackle the hallucination issue in diffusion-assisted 3D reconstruction through hallucination score modeling. To the best of our knowledge, this is the first work to study hallucination score modeling in this context.
- We introduce a multi-sampling strategy into HAD that generates and fuses multiple versions of a novel view into a single improved image.
- Our method achieves state-of-the-art performance in both in-domain and cross-domain evaluations. Extensive experiments show that our improvements stem from effective hallucination mitigation across diverse scenarios.

2. Related Works

Radiance fields for novel view synthesis. Novel view synthesis (NVS) aims to render photorealistic images from novel viewpoints given a set of posed input images. Among recent approaches, Neural Radiance Fields (NeRF)[31] and 3D Gaussian Splatting (3DGS)[21] have emerged as the two dominant paradigms. Subsequent research [2, 3, 5, 12, 32, 48] has focused on addressing NeRF's limitations in handling unbounded scenes and improving its rendering efficiency through hierarchical sampling, spatial factorization, and hybrid representations. 3DGS [21] leverages an efficient rasterization-based differentiable splatting pipeline [53], enabling real-time novel view synthesis with competitive visual quality. Following its success, numerous works [6, 10, 16, 40, 49] have further enhanced 3DGS by improving depth and surface estimation accuracy, reducing storage overhead, and alleviating aliasing artifacts, collectively establishing it as one of the most powerful and practical frameworks for 3D scene reconstruction. In this work, we validate the effectiveness of our hallucination-aware diffusion priors on enhancing 3DGS.

Novel view synthesis with priors. Recent advances demonstrate that incorporating priors from large foundation models can greatly enhance 3D reconstruction under sparse or imperfect observations. These priors provide strong guidance on geometry, texture, and semantics, enabling 3D reconstruction to move beyond reliance on purely original input data. Geometry-based methods [17, 23, 29, 33, 38, 45] impose additional constraints such as depth, frequency, or semantic consistency to stabilize training, but their perfor-

mance is often limited by inaccurate or inconsistent estimations. Image diffusion-based approaches [27, 34, 36, 41, 46] leverage pretrained diffusion models as powerful generative priors to reconstruct 3D scenes from text or very limited input views. More recently, video and multi-view diffusion models [4, 7, 13, 26, 28, 30, 42] further enhance view and temporal consistency through camera-conditioned generation. Nevertheless, while these diffusion priors effectively augment 3D reconstruction, they also introduce hallucinated content. We propose injecting hallucination awareness into diffusion-generated images and demonstrate that it significantly improves reconstruction performance.

Feedforward novel view synthesis. With the remarkable success of Transformer architectures [37] in both large language models and vision foundation models [9], feedforward networks [18–20, 44] have recently shown strong potential in novel view synthesis. The recent state-of-the-art LVSM [20] demonstrates that a pure transformer-based network can synthesize high-quality images from sparse inputs, while Rayzer [18] further eliminates the need for camera poses during training while still achieving competitive or superior performance. Due to their reconstruction nature, NVS networks are trained to synthesize novel view images with high fidelity to the input views and thus are prone to high-frequency information loss, whereas diffusion priors, despite their low fidelity to input data, tend to generate high-frequency information. Thus, in this work, we leverage the feature backbone of a well-trained NVS network as our multi-view encoder for understanding the context of input views to predict reliable hallucination scores for diffusion-generated images. This approach operates under the mild assumption that NVS networks possess the ability to effectively understand input context in order to render novel views with high fidelity.

Uncertainty estimation in 3D reconstruction. Recent works [1, 14, 15, 24, 35] estimate the uncertainty in 3D reconstruction, a concept related to our hallucination score. However, unlike these methods that quantify uncertainty in the final 3D representation, our method directly suppresses hallucinations in the augmented novel views for reconstruction, thereby reducing the 3D uncertainty propagation into the 3D model.

3. Preliminary

We briefly describe the preliminaries for 3D Gaussian Splatting (3DGS) – the 3D pipeline that we use to validate our HAD, feedforward novel view synthesis network and diffusion priors.

3D Gaussian Splatting. Given a set of calibrated images with known camera parameters, the goal of 3D Gaussian Splatting (3DGS) [21] is to render photorealistic image from arbitrary viewpoints in real time. 3DGS models the scene as a collection of anisotropic Gaussian primitives $\{(\boldsymbol{\mu}_i, \mathbf{S}_i, \mathbf{R}_i, \delta_i, \mathbf{c}_i)\}_{i=1}^{N}$ where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ denotes the 3D center of the i-th Gaussian, $\mathbf{S}_i = \mathrm{diag}(s_x, s_y, s_z)$ is a scaling matrix, and $\mathbf{R}_i$ is a rotation matrix derived from a quaternion parameterization. Each Gaussian also carries an learnable opacity scale δ_i and a set of spherical harmonics (SH) coefficients $\mathbf{c}_i$ for modeling view-dependent color. The opacity α_i is expressed as

$$\alpha_i = \delta_i \exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right], \tag{1}$$

where the covariance matrix is given by $\boldsymbol{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^T\mathbf{R}_i^\top$.

During rendering, tile-based rasterization is used to identify, for each pixel, which Gaussians contribute to its color. And then each pixel $\mathbf{p}$ can be rendered by blending N ordered points overlapping the pixel:

$$C(\mathbf{p}) = \sum_{i=1}^{N} \mathbf{c}_i\, \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) \tag{2}$$

Feed-forward novel view synthesis network. A feedforward NVS network is a generalizable network that takes multiple views as input and outputs a 3D feature, enabling the rendering of images from novel viewpoints. We formulate it as:

$$\mathbf{i_c} = \mathrm{FFD}(\mathbf{P} \mid \mathbf{c}), \tag{3}$$

where $\mathbf{P}$ denotes the input calibrated images, $\mathbf{c}$ represents novel view pose and $\mathbf{i_c}$ is the rendered image from $\mathbf{c}$. Recent works [19, 44] adopt a *Gaussian head* to predict the parameters of a Gaussian set and then render the target image using Eq. (2). In contrast, other approaches such as LVSM [20]– the state-of-the-art NVS network – directly regresses the target image without explicitly reconstructing the intermediate 3D representation, leading to a more simplified and scalable framework than the explicit Gaussian head approach. In light of this, our hallucination score network leverages the pretrained feature backbone of LVSM [20] to extract 3D information from input views.

Diffusion priors for 3D reconstruction. Recent advances have demonstrated that diffusion-based priors are highly effective for improving 3D reconstruction and scene enhancement [8, 11, 26, 28, 41]. By learning to model the data distribution $p_{\text{data}}(x)$ through an iterative denoising process, diffusion models can transform imperfect novel-view images into samples that follow the distribution of realistic images, under the assumption that the distribution of artifacts in imperfect novel-view images rendered by 3DGS is close to the noise distribution [41]. This property allows diffusion models, with light fine-tuning, to effectively remove artifacts in the images rendered by 3DGS at novel views.

Formally, a diffusion model progressively corrupts clean data $x \sim p_{\text{data}}$ with Gaussian noise to obtain noisy samples

$$x_\tau = \alpha_\tau x + \sigma_\tau \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{4}$$

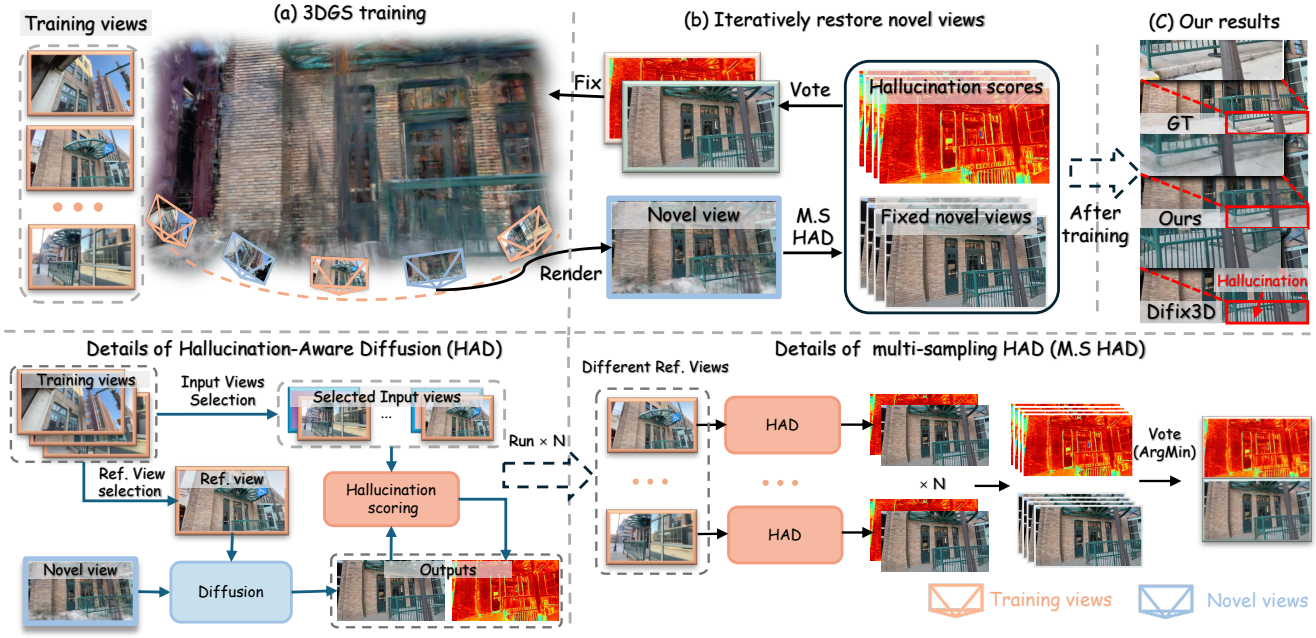


Figure 2. **Overview of framework –** We train 3DGS with input images and HAD-augmented novel views. HAD combines a pretrained diffusion prior (which generates images from 3DGS-rendered views conditioned on reference input images) with our hallucination score network (which predicts pixel-wise reliability maps). Our multi-sampling strategy fuses multiple generated versions into refined augmented views. Hallucination scores guide 3DGS optimization by masking off unreliable content, thus improving reconstruction quality.

where τ denotes the diffusion timestep, and $(\alpha_\tau, \sigma_\tau)$ control the noise schedule. The denoising network ϵ_θ learns to invert the noising process by predicting the noise term ϵ added at each step, optimized via the denoising score matching objective:

$$\mathbb{E}_{x\sim p_{\text{data}},\, \tau\sim p_\tau,\, \epsilon\sim\mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(x_\tau; c, \tau)\|_2^2\right], \quad (5)$$

where c denotes conditioning views sampled from the original input multi-views. The diffusion time τ is typically discretized into T steps (e.g., $T = 200$) with $\tau \sim \mathcal{U}(0, T)$, ensuring that the data becomes nearly Gaussian at the final step. In this work, we validate our hallucination awareness approach on Difix3D [41], a single-step diffusion prior that achieves the top performance on NVS from sparse views.

4. Methodology

Given N calibrated images of a scene $\mathbf{P} = \{(\mathbf{I}_n, \mathbf{C}_n) | n \in \{0, \dots N\}\}$, where $\mathbf{I}$ and $\mathbf{C}$ denote the RGB images and camera poses respectively, our goal is to reconstruct a high-fidelity 3DGS model capable of producing high-quality renderings for both training views and ill-constrained novel views with insufficient observations. Specifically, we train the 3DGS model under supervision from both the input views and augmented views generated by our hallucination-aware diffusion prior (HAD).

4.1. 3DGS training

We formulate the 3DGS training as

$$\arg\min_{\mathbf{\Phi}} \lambda_{\text{input}}\mathcal{L}_{\text{input}} + \lambda_{\text{novel}}\mathcal{L}_{\text{novel}} \quad (6)$$

where $\mathcal{L}_{\text{input}}$ and $\mathcal{L}_{\text{novel}}$ are the rendering losses for input views and augmented novel views, respectively, λ_{input} and λ_{novel} are the coefficients for the two loss terms, and $\mathbf{\Phi}$ denotes the parameters of the 3DGS model described in Sec. 3.

We follow 3DGS [21] to compute the rendering loss as in at input views by combining $\mathcal{L}_1$ and $\mathcal{L}_{\text{D-SSIM}}$:

$$\mathcal{L}_{\text{input}} = 0.8\mathcal{L}_1\left(\mathcal{R}_{\mathbf{\Phi}}\left(\mathbf{c}\right), \mathbf{i}\right) + 0.2\mathcal{L}_{\text{D-SSIM}}\left(\mathcal{R}_{\mathbf{\Phi}}\left(\mathbf{c}\right), \mathbf{i}\right) \quad (7)$$

where $\mathbf{i}$ and $\mathbf{c}$ are the image and camera pose sampled from the input views, and $\mathcal{R}_{\mathbf{\Phi}}$ is the rendering function parameterized by $\mathbf{\Phi}$ that takes camera pose $\mathbf{c}$ as input.

Our hallucination-aware diffusion prior (HAD) augments novel views to form $\mathcal{L}_{\text{novel}}$. Specifically, we follow Difix3D [41] to sample novel view poses by progressively departing from the input poses toward the target poses. HAD then generates the images and their corresponding hallucination score maps. We denote the resultant hallucination masks, poses, and images of T novel views as $\tilde{\mathbf{P}} = \{(\mathbf{M}_t, \tilde{\mathbf{I}}_t, \tilde{\mathbf{C}}_t) \mid t \in \{0, \dots, T\}\}$. We then calculate $\mathcal{L}_{\text{novel}}$ as:

$$\begin{aligned}\mathcal{L}_{\text{novel}} =& \mathcal{L}_1(\neg\mathbf{m} \odot \mathcal{R}_{\mathbf{\Phi}}\left(\tilde{\mathbf{c}}\right), \neg\mathbf{m} \odot \tilde{\mathbf{i}}) \\ &+ \mathcal{L}_{\text{D-SSIM}}(\neg\mathbf{m} \odot \mathcal{R}_{\mathbf{\Phi}}\left(\tilde{\mathbf{c}}\right), \neg\mathbf{m} \odot \tilde{\mathbf{i}})\end{aligned} \quad (8)$$

where $\odot$ denotes the Hadamard product, $\neg$ inverts the binary mask, and $\mathbf{m}$, $\tilde{\mathbf{c}}$, and $\tilde{\mathbf{i}}$ are the hallucination mask, pose, and RGB image, respectively, of a novel view generated through HAD, which we detail in Sec. 4.2.

We illustrate the training process in Fig. 2. Specifically, unlike Difix3D [41], which employs a two-phase training strategy that first fully trains a 3DGS model and then progressively updates it with diffusion priors using a small learning rate, we adopt a simplified single-phase training approach. Our method trains the 3DGS model with both input views and diffusion-augmented views simultaneously from the start. The key insight is that our hallucination score maps enable masking of unreliable pixels, allowing us to directly incorporate diffusion-augmented views from the beginning of training without requiring a pre-trained 3DGS initialization.

Nevertheless, we follow Difix3D [41] in alternating between view augmentation and training steps. In the view augmentation step, we sample novel view camera poses by interpolating between input and target views. HAD then generates augmented images $\tilde{\mathbf{i}}$ along with their corresponding pixel-wise hallucination score maps $\mathbf{s}$, which are converted into binary masks $\mathbf{m}$ via simple thresholding.

4.2. Hallucination-Aware Diffusion Prior

To enhance novel view synthesis quality, we propose the hallucination-aware diffusion prior (HAD) to augment images rendered at novel views and optimize the 3DGS model with these augmented views. Specifically, HAD consists of the diffusion prior (Sec. 4.2.1) from Difix3D [41] and our novel hallucination score estimator (Sec. 4.2.2). As shown in Fig. 2, the diffusion prior enhances novel views by removing artifacts through a denoising step conditioned on a sampled input view close to the novel view, while the hallucination score network predicts hallucination masks for the augmented views. Furthermore, to improve performance, we propose a multi-sampling strategy (Sec. 4.2.3) that creates multiple versions of augmented views by conditioning the diffusion prior on different input views, then fuses them into a higher-quality image.

4.2.1. Refining Novel Views w/ Diffusion Prior

Formally, given the novel view camera $\tilde{\mathbf{c}}$, the pretrained diffusion prior generates the novel view image $\tilde{\mathbf{i}}_{\mathcal{G}}$ as

$$\tilde{\mathbf{i}}_{\mathcal{G}} = \mathcal{G}\left(\mathcal{R}_{\Phi}\left(\tilde{\mathbf{c}}\right) | \, \mathbf{i}_{\text{ref}}\right) \tag{9}$$

where $\mathbf{i}_{\text{ref}}$ is a reference view sampled from the original input multi-views serving as conditioning 3D context for the view enhancement. The diffusion generated $\tilde{\mathbf{i}}_{\mathcal{G}}$, while photorealistic, tend to introduce aliens that do not exist in the input multi-views.

4.2.2. Hallucination Score Estimation

Thus, we propose a hallucination score network to estimate a pixel-wise hallucination score map $\mathbf{s}$ for the generated image $\tilde{\mathbf{i}}_{\mathcal{G}}$, enabling hallucination-aware diffusion priors for 3DGS enhancement. Specifically, the hallucination score network consists of two components: a multi-view feature encoder $\mathcal{V}$ that processes multiple input views, and a score estimation branch $\mathcal{S}$ that predicts hallucination scores for novel view images using both the multi-view features and the novel view images. For further details, please refer to Sec. 8 of `supplementary material`.

To leverage the multi-view reasoning capability of existing novel view synthesis networks, we base our hallucination score network on the pre-trained LVSM [20], a state-of-the-art feedforward NVS model with a feature backbone that processes multiple input views and an NVS head for image generation. We set the pretrained feature backbone of LVSM [20] as $\mathcal{V}$ and realize the $\mathcal{S}$ as a simple three layer UNet. We train the framework with the $\mathcal{V}$ frozen on a small curated dataset of novel view and original multi-view pairs.

Formally, we calculate the hallucination score as

$$\mathbf{s} = \mathcal{S}_{\theta}(\tilde{\mathbf{i}}_{\mathcal{G}} | \mathcal{R}_{\mathbf{\Phi}}(\tilde{\mathbf{c}}), \mathbf{F}_{\tilde{\mathbf{c}}}) \tag{10}$$

where θ denotes the parameters of $\mathcal{S}$, $\tilde{\mathbf{c}}$ is the novel view pose, and $\mathbf{F}_{\tilde{\mathbf{c}}}$ represents the multi-view features, which we formulate as:

$$\mathbf{F}_{\tilde{\mathbf{c}}} = \mathcal{V}(\mathbf{P} \mid \tilde{\mathbf{c}}) \tag{11}$$

where $\mathbf{P}$ denotes the original multiple input views. Thus, the multi-view encoder $\mathcal{V}$ outputs features aggregated at the novel view pose $\tilde{\mathbf{c}}$ from the input views.

We then concatenate $\mathbf{F}_{\tilde{\mathbf{c}}}$, $\tilde{\mathbf{i}}_{\mathcal{G}}$, and optionally the 3DGS-rendered image $\mathcal{R}_{\mathbf{\Phi}}(\tilde{\mathbf{c}})$, from which $\mathcal{S}$ predicts a pixel-wise hallucination score map $\mathbf{s}$ that reflects the hallucination probability of each pixel in the diffusion-augmented image $\tilde{\mathbf{i}}_{\mathcal{G}}$. To train $\mathcal{S}$, we define the ground-truth hallucination score as the mean absolute error (MAE) between the diffusion-generated image and the ground-truth image. We supervise $\mathcal{S}$ with an L2 loss between the predicted score and the ground-truth score.

4.2.3. Multi-Sampling Strategy

To further enhance HAD, we propose a multi-sampling strategy that creates multiple versions of augmented views and fuses them to produce higher-quality novel views for 3DGS optimization. Specifically, the diffusion model augments novel views with different sampled conditioning input views, generating multiple augmented views along with their corresponding hallucination score maps.

In Difix3D [41], each diffusion refinement relies solely on the nearest reference image $\mathbf{i}_{\text{ref}}$, which limits the exploitation of complementary cues from other viewpoints. Our multi-sampling strategy creates multiple versions of

augmented images—$\{(\tilde{\mathbf{i}}_{\mathcal{G}}^{k}, \mathbf{s}^{k}) \mid k \in \{1, \ldots, K\}\}$—by conditioning on K sampled references from the input views. We then fuse the images by selecting pixels with the lowest hallucination score across all candidate versions:

$$\tilde{\mathbf{i}}[i] = \tilde{\mathbf{i}}_{\mathcal{G}}^{k^*}[i], \quad k^* = \arg\min_{k} \mathbf{s}^{k}[i],$$

where $[i]$ denotes pixel indexing, $\tilde{\mathbf{i}}_{\mathcal{G}}^{k}$ and $\mathbf{s}^{k}$ denote the diffusion-generated image and its predicted hallucination score from the k-th reference view. This pixel-wise selection effectively fuses reliable regions from multiple refinements, yielding augmented novel views with improved fidelity. More importantly, this strategy allows the diffusion prior to exploit broader multi-view information, effectively reducing hallucination issues in the final 3DGS model.

5. Experiments

5.1. Experimental Setup

Dataset and metrics. We primarily use Peak Signal-to-Noise Ratio (PSNR), structural (SSIM [39]) and perceptual (LPIPS [50]) similarities as metrics to quantify the performance of novel view synthesis. And we conduct experiments on two well-known datasets.

- `DL3DV` [25] is a large-scale scene dataset containing calibrated multi-view images. We first curate training dataset of randomly selected 116 scenes from benchmark dataset for hallucination score network training. To validate the performance of HAD on novel view synthesis, we follow Difix3D [41] to select other 24 scenes for testing. For each testing scene, we partition the available views into three subsets: training, target, and test views. Specifically, we select nine sparse views as input for 3D reconstruction. The remaining views are then divided into test and target subsets. We follow [42] sample test views randomly from the unused views at a ratio of 25%, and compute the final performance metrics, providing a representative assessment of novel-view reconstruction quality. Target views are uniformly sampled from the last few viewpoints of each scene at a ratio of 50%. The target views are used to sample novel views for generative prior augmentation in 3DGS training. This design ensures that no leaked information from the target subset is used for model adaptation, maintaining fair and unbiased comparisons across methods.
- `MipNeRF360` [3] is a dataset comprising 9 scenes commonly used for evaluating novel view synthesis. We adopt the training and testing view splits provided by GenFusion [43], which consists of 9 training input views per scene. For the target views, similarly, we uniformly sample 50% of the remaining unused views.

Implementation details. To construct the hallucination score network, we utilize the pretrained multiview encoder from the LVSM model [20] and train an additional score prediction branch on a small curated dataset. We prepare the dataset using the Difix3D [41] pipeline. For each scene, we first train a 3DGS model using 9 input training views, then generate 100 augmented novel views via diffusion priors [41] at a resolution of 960×540. The poses for these augmented views are sampled from views excluded from the 3DGS training set, providing ground-truth images that enable computation of the ground-truth hallucination score map for each augmented view. We fine-tune the network for 10k iterations with a batch size of 2 per GPU, requiring approximately 28 hours on eight NVIDIA V100 32GB GPUs. For 3DGS training, we set the learning rates to $8e^{-5}$ for Gaussian means, $5e^{-2}$ for opacity, $1e^{-3}$ for rotation, $5e^{-4}$ for the 0-th order spherical harmonics (SH_0), and $2.5e^{-5}$ for higher-order spherical harmonics (SH_N), and train for 30k iterations. And we set λ_{input} and λ_{novel} as 1. We empirically set the threshold as 0.9 to calculate the $\mathbf{m}$.

Table 1. **Quantitative comparison of different methods on DL3DV [25].** Best, second, and third results are highlighted in 1st, 2nd, and 3rd, respectively. (↑: higher is better, ↓: lower is better). Note that ours* denotes a variant following the two-phase 3DGS optimization strategy of Difix3D, enabling a fair comparison between diffusion priors with and without hallucination awareness. We denote "+" as post-rendering, which evaluates the rendering quality from 3DGS augmented by a diffusion model, rather than direct rendering from 3DGS.

Method	Type	PSNR↑	SSIM↑	LPIPS↓
Depthsplat [44]	FFD	18.324	0.640	0.378
LVSM [20]	FFD	19.855	0.636	0.252
Gsplat-3DGS [21]	Optimization	19.004	0.679	0.281
Gsplat-mcmc [22]	Optimization	20.532	0.721	0.225
Difix3D [41]	Optimization	21.355	0.734	0.199
Ours*	Optimization	21.983	0.755	0.195
Ours	Optimization	22.134	0.757	0.190
Difix3D+ [41]	Post-rendering	20.673	0.709	0.181
Ours+	Post-rendering	21.213	0.719	0.177

5.2. In-domain evaluation

We evaluate HAD on DL3DV, where we train our hallucination network, thereby providing an in-domain evaluation setting. We compare against feedforward NVS networks (DepthSplat [44], LVSM [20]), two variants of 3DGS, and state-of-the-art diffusion prior-assisted 3DGS pipelines (Difix3D [41] and Difix3D+ [41]). Note Difix3D+, an enhanced version of Difix3D, further improves image photorealism in the post-rendering stage.

Quantitative results – Tab. 1. Our method outperforms the baselines by a large margin across all metrics. To enable fair comparison with our closest pipeline, we adopt Difix3D's two-phase optimization strategy for 3DGS. In this approach, we first train an initial 3DGS model using only

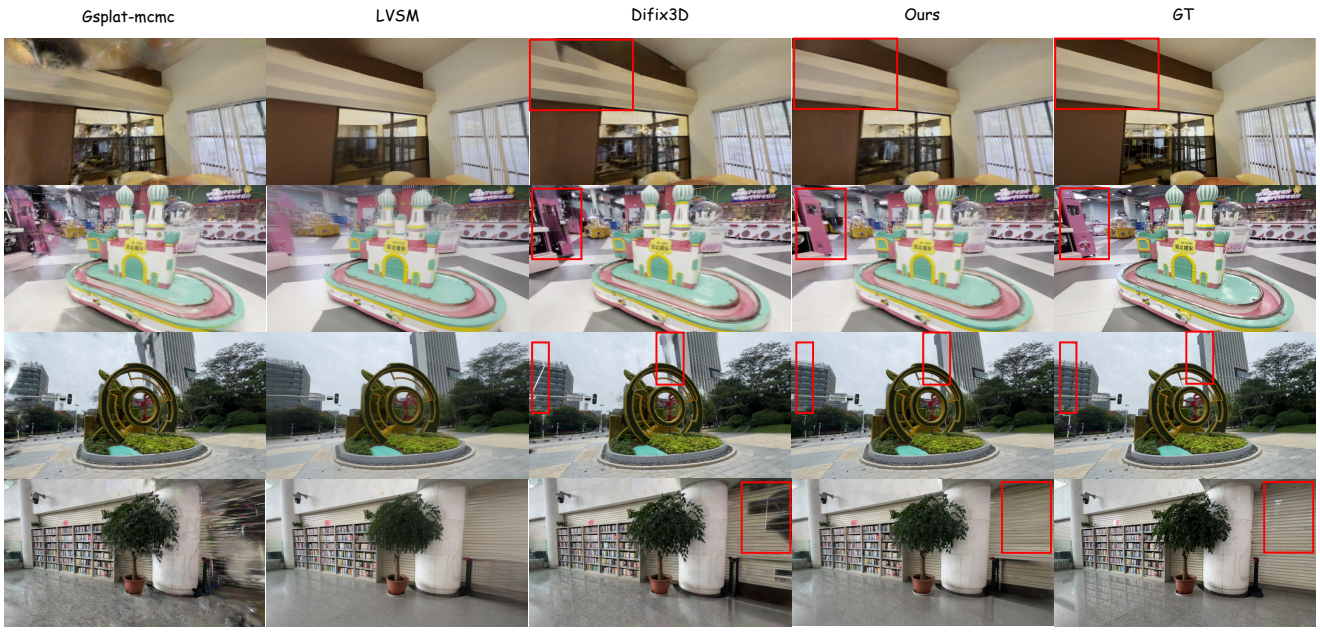


Figure 3. **Examples on DL3DV [25] –** We show novel-view rendering obtained by ours, Gspat-mcmc [22], LVSM [20] and Difix3D [41]. Our approach achieves the sharper rendering as well as the better fidelity to the ground-truth.

the input views, then optimize the model using both input and augmented views. Under this setting, we replace Difix3D's diffusion prior with our HAD. We demonstrate that even under this suboptimal configuration, our method surpasses the state of the art, highlighting the critical importance of hallucination awareness in diffusion priors. We further show that our method outperforms Difix3D in the post-rendering setting, where the diffusion prior refines the rendered images of the final 3DGS model. Notably, while the post-rendering improves the photorealism metric (LPIPS), it degrades fidelity metrics (PSNR and SSIM) for both Difix3D and our method.

Qualitative results – Fig. 3. Qualitatively, we demonstrate that HAD generates augmented views with both photorealism and high consistency with the input views. Notably, LVSM [20], despite producing blurry images, achieves better fidelity, validating our approach of leveraging LVSM's feature backbone as multi-view encoder for hallucination score prediction.

5.3. Cross-domain evaluation

We also examine whether our HAD can generalize to a different dataset – `MipNeRF360` [3]. We compare against baselines including FSGS [52], GenFusion [43] and Difix3D [41]. As shown in Tab. 2, similar to the in-domain evaluation, our method achieves state-of-the-art performance. Note that GenFusion [43] leverages video diffusion priors that exhibit better multi-view consistency, thereby outperforming baselines using image diffusion priors such as Difix3D. Nevertheless, our method demonstrates the better performance due to our hallucination-aware strategy. As shown in Sec. 9, we further validate that HAD generalizes across different diffusion models (e.g., video and multi-view diffusion [43, 51]).

Table 2. **Quantitative comparison of different methods on MipNerf360 [3].** Note the results of Genfusion and FSGS are from Genfusion [43].

Method	Type	PSNR↑	SSIM↑	LPIPS↓
Gsplat-3DGS [21]	Optimization	15.748	0.424	0.431
Gsplat-mcmc [22]	Optimization	17.102	0.454	0.385
FSGS [52]	Optimization	17.940	0.492	0.468
GenFusion [43]	Optimization	18.360	0.496	0.465
Difix3D [41]	Optimization	18.001	0.475	0.350
Ours	Optimization	18.689	0.5094	0.334
Difix3D+ [41]	Post-rendering	17.571	0.433	0.326
Ours+	Post-rendering	18.137	0.456	0.309

5.4. Ablation studies

We conduct ablation studies on different components, the number of versions and fusion strategy in the multi-sampling strategy (M.S.), the pretrained multiview encoder, and the performance of our method under dense view settings. We report the performance on `DL3DV` [25]. Except for the dense view setting where we use 24 views, all ablation studies employ the 9-views setting.

Impact of different components – Tab. 3. As shown in Tab. 3, HAD notably improves performance over Difix3D, confirming the importance of hallucination awareness. The multi-sampling strategy further enhances reconstruction consistency and perceptual quality. In addition, thanks to the high fidelity of augmented views from HAD,

Table 3. **Impact of different components**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Gsplat-mcmc	20.532	0.721	0.225
Difix3D	21.355	0.734	0.199
Difix3D + HAD	21.779	0.749	0.195
Difix3D + HAD + M.S.	21.983	0.755	0.195
Ours (full model)	**22.134**	**0.757**	**0.190**

Table 4. **Number of versions in multi-sampling strategy**.

Num. of versions	PSNR ↑	SSIM ↑	LPIPS ↓
1 (No M.S.)	21.779	0.749	0.195
2	21.938	0.754	0.195
3	**21.983**	**0.755**	**0.195**

Table 5. **Fusion methods in multi-sampling strategy**. We compare two approaches: (1) `ArgMin`: selecting pixels with the lowest hallucination score; (2) `Weighted Average`: computing the weighted mean.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
`Weighted Average`	21.856	0.747	0.190
`ArgMin` (Ours)	**22.134**	**0.757**	**0.190**

Table 6. **Different hallucination score estimators**. We use Mean Absolute Error (MAE) of the predicted hallucination score maps as our evaluation metric. We demonstrate that our hallucination score network, *with the pretrained multiview encoder*, achieves the best performance.

Method	Retrained Difix3D	Ours (w/o pretrained)	Ours (full)
MAE ↓	0.058	0.054	**0.043**

we simplify 3DGS training to a single-phase approach, differing from Difix3D which employs a two-phase strategy.

Impact of the number of versions in M.S. –Tab. 4. We can create multiple versions of augmented views with diffusion priors by conditioning on different sampled input images and then fusing them into a better final view. As shown in Tab. 4, the more versions we create, the better the performance. We select 3 views to achieve a trade-off between marginal improvement and computational overhead.

Fusion strategy in M.S. – Tab. 5. We study how to fuse the multi-sampling results. We evaluate two strategies: weighted averaging based on hallucination scores maps and ArgMin fusion, which selects the pixel with the lowest hallucination score across samples. As shown in Tab. 5, ArgMin fusion achieves the best overall performance, producing sharper and more consistent reconstructions.

Different hallucination score estimators – Tab. 6. We study the performance of three hallucination score estimators: retrained Difix3D, ours without the pretrained multi-view encoder, and our full method. For the retrained Difix3D, we follow Difix3D [41] to finetune the diffusion prior to augment novel views while predicting hallucination scores. We evaluate the performance of hallucination estimation on our curated dataset of 114 training scenes and 26 testing scenes. We calculate the Mean Absolute Error (MAE) of the predicted hallucination score maps on the test scenes. As shown in Tab. 6, the finetuned diffusion prior, due to its lack of multi-view reasoning ability, predicts the worst hallucination scores. Similarly, our method without the pretrained multi-view encoder performs worse. This comparison demonstrates the importance of leveraging multi-view reasoning ability of a pretrained NVS network.

Table 7. **The performance of our method in dense view setting (24 views).**

Method (GSplat-MCMC)	PSNR ↑	SSIM ↑	LPIPS ↓
Gsplat-MCMC	26.280	0.869	0.101
Difix3D	26.770	0.874	0.0926
Ours	**26.969**	**0.876**	**0.0921**

The performance of our method in dense view setting – Tab. 7. We also examine whether our method remains effective when the input coverage becomes denser. On `DL3DV` [25]with 24 input views, the Difix3D already achieves a high PSNR of 26.77 dB, where hallucination artifacts are largely reduced. Even in this setting, our method still improves PSNR by 0.20 dB over Difix3D and by 0.69 dB over the original 3DGS model, indicating that the hallucination-aware diffusion prior continues to provide benefits beyond the sparse-view regime.

6. Conclusion and Future Work

In this work, we identify and address a critical limitation in diffusion-assisted 3D reconstruction: while diffusion priors effectively alleviate data sparsity, they introduce hallucinated content that compromises fidelity to input views despite achieving photorealistic rendering. We propose Hallucination-Aware Diffusion priors (HAD), the first framework to explicitly tackle the hallucination issue in diffusion-assisted 3D reconstruction through hallucination score modeling. The experimental results on both in-domain and cross-domain benchmarks demonstrate that HAD achieves state-of-the-art performance, with consistent improvements resulting from effective hallucination mitigation. Our work establishes hallucination awareness as a crucial component for diffusion-assisted 3D reconstruction.

Limitations and future works. Our method is bounded by the pretrained multi-view encoder. An interesting direction for future work is to scale up the training of our model by removing the need for complex data requirements– for instance, using uncalibrated multi-view images as in Rayzer [18]. Additionally, we validate our method only on reconstruction of seen regions, and the extension of hallucination-aware modeling to improve generation in unseen regions represents an important direction for future research.

References

[1] Niki Amini-Naieni, Tomas Jakab, Andrea Vedaldi, and Ronald Clark. Instant uncertainty calibration of nerfs using a meta-calibrator. In *European Conference on Computer Vision*, pages 309–324. Springer, 2024. 3

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 2

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2, 6, 7, 9

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 2

[6] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[7] 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. 3

[8] 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. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, Zhangyang Wang, et al. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *Advances in neural information processing systems*, 37: 140138–140158, 2024. 2

[11] Tobias Fischer, Samuel Rota Bulò, Yung-Hsu Yang, Nikhil Keetha, Lorenzo Porzi, Norman Müller, Katja Schwarz, Jonathon Luiten, Marc Pollefeys, and Peter Kontschieder. Flowr: Flowing from sparse to dense 3d reconstructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27702–27712, 2025. 1, 2, 3

[12] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2

[13] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 3

[14] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20061–20070, 2024. 3

[15] Jiaxin Guo, Jiangliu Wang, Ruofeng Wei, Di Kang, Qi Dou, and Yun-Hui Liu. Uc-nerf: Uncertainty-aware conditional neural radiance fields from endoscopic sparse views. *IEEE Transactions on Medical Imaging*, 44(3):1284–1296, 2024. 3

[16] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 2

[17] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, pages 5885–5894, 2021. 2

[18] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4929, 2025. 3, 8

[19] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025. 3

[20] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 5, 6, 7

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2, 3, 4, 6, 7

[22] 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. 1, 6, 7

[23] 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 *Proceedings of the IEEE/CVF conference on com-*

puter vision and pattern recognition, pages 20775–20785, 2024. 2

[24] Ruiqi Li and Yiu-ming Cheung. Variational multi-scale representation for estimating uncertainty in 3d gaussian splatting. In *Advances in Neural Information Processing Systems*, pages 87934–87958, 2024. 3

[25] 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. 2, 6, 7, 8

[26] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *IEEE Transactions on Image Processing*, 2026. 3

[27] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[28] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024. 1, 2, 3

[29] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 2

[30] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *International Conference on Machine Learning*, 2024. 3

[31] 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, 2

[32] Muhammad Husnain Mubarik, Ramakrishna Kanungo, Tobias Zirr, and Rakesh Kumar. Hardware acceleration of neural graphics. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–12, 2023. 2

[33] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5480–5490, 2022. 2

[34] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[35] Jianxiong Shen, Ruijie Ren, Adria Ruiz, and Francesc Moreno-Noguer. Estimating 3d uncertainty field: Quantifying uncertainty for neural radiance fields. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2375–2381. IEEE, 2024. 3

[36] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9065–9076, 2023. 2

[39] 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. 6

[40] Jiang Wu, Rui Li, Yu Zhu, Rong Guo, Jinqiu Sun, and Yanning Zhang. Sparse2dgs: Geometry-prioritized gaussian splatting for surface reconstruction from sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11307–11316, 2025. 2

[41] 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, 4, 5, 6, 7, 8

[42] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21551–21561, 2024. 3, 6

[43] Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6078–6088, 2025. 2, 6, 7, 1

[44] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16453–16463, 2025. 3, 6

[45] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8254–8263, 2023. 2

[46] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024. 3

[47] Xingyilang Yin, Qi Zhang, Jiahao Chang, Ying Feng, Qingnan Fan, Xi Yang, Chi-Man Pun, Huaqi Zhang, and Xiaodong Cun. Gsfixer: Improving 3d gaussian splatting

with reference-guided video diffusion priors. *arXiv preprint arXiv:2508.09667*, 2025. 2

[48] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5752–5761, 2021. 2

[49] 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

[50] 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. 6

[51] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12405–12414, 2025. 2, 7, 1, 3

[52] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European conference on computer vision*, pages 145–163. Springer, 2024. 7

[53] Matthias Zwicker, Hanspeter Pfister, Jeroen van Baar, and Markus Gross. Surface splatting. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, page 371–378, New York, NY, USA, 2001. Association for Computing Machinery. 2