Filling the Unseen: Scene Extrapolation via 3D Gaussian Splatting

Yunlai Zhou
Case Western Reserve University
Department of Computer and Data Sciences
Cleveland, Ohio, USA
yxz3057@case.edu

Yiren Lu
Case Western Reserve University
Department of Computer and Data Sciences
Cleveland, Ohio, USA
yxl3538@case.edu

Tuo Liang
Case Western Reserve University
Department of Computer and Data Sciences
Cleveland, Ohio, USA
txl859@case.edu

Disheng Liu
Case Western Reserve University
Department of Computer and Data Sciences
Cleveland, Ohio, USA
dxl952@case.edu

Vipin Chaudhary
Case Western Reserve University
Department of Computer and Data Sciences
Cleveland, Ohio, USA
vxc204@case.edu

Yu Yin✉
Case Western Reserve University
Department of Computer and Data Sciences
Cleveland, Ohio, USA
yxy1421@case.edu

Abstract

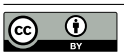
3D Gaussian Splatting achieves photorealistic reconstruction within training view distribution, yet it degrades on out-of-distribution novel views, exhibiting holes in unobserved regions and artifacts in observable areas. Recent works formulat this task as extrapolation and interpolation and tried to address it with generative models, but remain limited in extrapolation scale and quality. They repeat a generate–reconstruct–shift cycle to progressively build a scene, which introduces accumulated errors with every step conditioning on previous outcomes. In this work, we propose a holistic framework for extrapolation and interpolation. We devise an independent camera view detection mechanism to enable parallel conflict-free extrapolation, circumventing the reliance on aforementioned error-prone cycle. Building upon this, we design a hierarchical pipeline that extrapolates independent and dependent camera views separately. Additionally, previous methods overlook inconsistency between generated and original images, resulting in compromising well-reconstructed areas. We propose a plug-and-play Quality-Aware Mask (QA-Mask) module, enabling selective utilization on generated data. By calibrating learning weights with pixel-wise rendering quality, it prevents generation-induced degradations on well-constructed areas. Extensive experiments demonstrate the superior performance of our framework, with QA-Mask generalizing on multiple generative reconstruction models (project page).

CCS Concepts

• **Computing methodologies** → **Reconstruction**.

Keywords

Gaussian Splatting, Scene Extrapolation, Spherical Harmonics



MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835425

ACM Reference Format:
Yunlai Zhou, Yiren Lu, Tuo Liang, Disheng Liu, Vipin Chaudhary, and Yu Yin. 2026. Filling the Unseen: Scene Extrapolation via 3D Gaussian Splatting. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835425

1 Introduction

The emergence of Neural Radiance Fields (NeRFs) [17] and 3D Gaussian Splatting (3DGS) [12] has significantly advanced 3D scene reconstruction with their unprecedented photorealism and high fidelity. However, this ability is confined to the training view distribution. When rendering from out-of-distribution novel views, they suffer from severe degradations, including holes due to un-observation in training images, as well as artifacts and incorrect geometry in observed areas, as shown in the upper part of Figure 1. These issues undermine the viewing experience in image quality and freedom of exploration, limiting their practical applicability.

Several studies [14, 16, 28, 29, 35] employed diffusion models to repair artifacts on novel views, but they cannot handle holes due to un-observation in training views. Recently, a number of methods [11, 23, 30, 39] have been proposed, formulating and trying to tackle the aforementioned two issues as scene extrapolation and interpolation. However, they are generally limited to small-scale extrapolation, such as filling small holes within observed areas, and are unable to expand a scene at a larger scale. Moreover, their extrapolation results are often prone to blurring and inaccurate geometry. We attribute these limitations to their reliance on the generate–reconstruct–shift cycle which is first introduced in text-to-scene generation [4, 9, 19, 36, 37], where "shift" means moving the camera view to reveal new areas that need to be generated. This cycle will be iteratively repeated in a spatially-sequential manner to progressively build up a scene. However, iterative generation and reconstruction conditioned on the outcomes of previous steps tend to accumulate errors, as generation-based reconstruction is notoriously under-constrained, especially when extrapolating a scene beyond its boundaries.

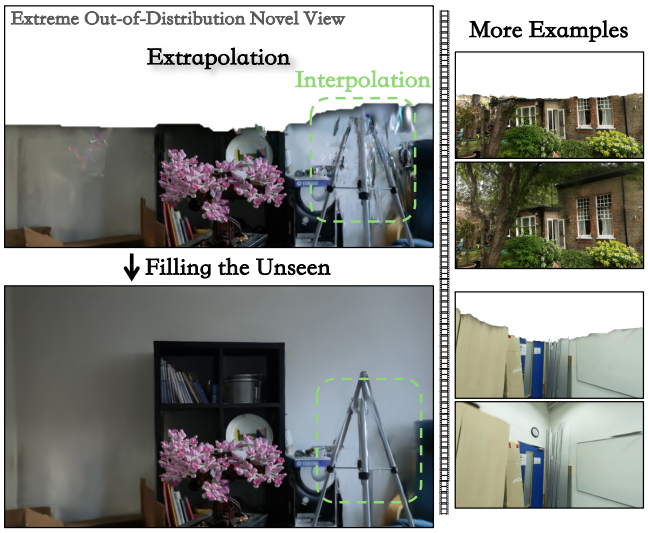


Figure 1: Our framework excels at large-scale, geometrically coherent 3D scene extrapolation and interpolation on extreme out-of-distribution novel views.

Furthermore, existing approaches [11, 14, 23, 29, 30, 39] overlook the inconsistency between generated content and the real scene. Indiscriminately applying generated content to update the scene inevitably leads to suboptimal reconstruction quality and fidelity, e.g., corrupting well-reconstructed areas in the original scene. A principled strategy is to selectively leverage generated content based on the current rendering quality, ensuring that generation only complements rather than undermines the reconstruction process.

In this paper, we present ***Filling the Unseen***, a holistic framework for large-scale, geometrically coherent scene extrapolation and interpolation via 3DGS, as shown in Figure 1. Our key novelty is that we switch the common paradigm from many-stage cycle repetition to hierarchical two-stage extrapolation. We first devise a mechanism to identify a set of independent camera views used for extrapolation in a given 3D scene by detecting collisions between mesh-based view frustums. These independent camera views support parallel and conflict-free generation and reconstruction. Built upon this mechanism, we redesign a hierarchical pipeline that circumvents repeating the error-prone generate–reconstruct–shift cycles. Specifically, our pipeline first performs large-scale scene extrapolation on all independent camera views in a single pass, followed by a second-stage residual completion for remaining unfilled regions, thus achieving seamless scene extrapolation. In addition, to improve the rendering quality of the original scene on out-of-distribution extrapolation views, we perform scene interpolation by re-exploiting the informative generated videos produced during the extrapolation, together with an artifact removal model [29].

Furthermore, by leveraging the directional modeling capability of spherical harmonics to model "which viewing directions are supervised" for each Gaussian primitive, we design a novel method to render masks that can measure rendering quality. Spatially reweighting the reconstruction loss with this mask during scene updates enables selective utilization on generated content, as shown in Figure 4, which effectively prevents generation-induced degradations in well-reconstructed regions while also preserves desired extrapolation and interpolation. We show that this plug-and-play **Q**uality-**A**ware **Mask** (QA-Mask) module can be readily integrated into multiple generative reconstruction models [29, 30] as enhancement. In summary, our key contributions include:

- We propose *Filling the Unseen*, a hierarchical 3DGS-based framework enabled by independent camera view detection mechanism for large-scale, geometrically coherent scene extrapolation and interpolation.
- We introduce QA-Mask, a plug-and-play, rendering quality-aware mask method to enable selective utilization on generated content during scene updates.
- Extensive experiments on multiple datasets demonstrate the superior performance of our proposed framework and the generalizability of QA-Mask.

2 Related Works

2.1 Scene Extrapolation and Interpolation

ExtraNeRF [23] first explored diffusion inpainters to synthesize beyond observed regions, but was mainly limited to narrow regions around object boundaries. VEGS [11] extended this idea to urban driving scenes. For sparse inputs, Zhong et al. [39] proposed a training-free scene-grounding strategy that tames pretrained diffusion models toward generating scene-consistent image sequences. GenFusion [30] instead trains a video diffusion model on paired low-quality renderings and ground-truth images, providing a unified solution to extrapolation and interpolation, but often suffering from blurriness and color inconsistency. GaMo [10] formulates sparse-view reconstruction as multi-view outpainting by expanding the field of view to expand scene coverage.

Compared to extrapolation, scene interpolation has been more extensively studied. 3DGS-Enhancer [14] fine-tunes diffusion models on held-out degraded renderings and their ground truth. DiFix3D+ [29] improves this paradigm using single-step diffusion, diverse data hold-out strategies, and progressive 3D updates. GSFix3D [28] combines meshes and 3D Gaussians to handle diverse scenes and artifacts, while GSFixer [35] incorporates VGGT [26] into video diffusion models to condition generation on both geometry and appearance. Nevertheless, these approaches rely explicitly or implicitly on iterative generate-reconstruct-shift cycles and insufficiently address inconsistencies between generated content and real scenes, both of which require immediate and careful handling.

2.2 Rendering Quality Measurement

Rendering quality and uncertainty estimation benefit applications including next-best-view selection [20, 32], floater removal [7, 13, 27], pruning and densification [8], and scene generation [14]. For NeRFs, prior works model uncertainty through volumetric fields [7] via spatial perturbations and Bayesian Laplace approximation, local 3D diffusion priors [27] and score distillation sampling loss, or Neural Visibility Fields [32]. For Gaussian Splatting, Li et al. [13]

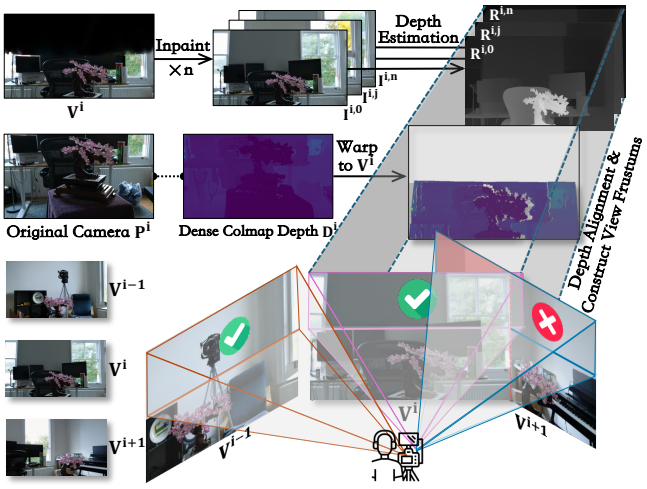


Figure 2: Independent Camera View Detection. We construct a mesh-based view frustum enclosing the geometry of the inpainted region for each candidate view $\mathbf{V^i}$. Collision detection is performed for frustums to determine whether their camera views are independent. In this example, $\mathbf{V^{i-1}}$ and $\mathbf{V^i}$ are independent, while $\mathbf{V^i}$ and $\mathbf{V^{i+1}}$ collide with each other and are therefore not independent.

infer predictive uncertainty using a variational multi-scale framework, while PUP-3DGS [8] uses second-order reconstruction-error approximations for uncertainty-aware pruning.

However, these methods are not designed for perceptual rendering quality assessment in context of scene generation. 3DGS-Enhancer [14] assumes that well-reconstructed regions are represented by small-volume Gaussians, which may fail for large, low-frequency surfaces such as plain walls. In contrast, our QA-Mask better reflects perceptual rendering quality and is specifically designed for scene generation, as validated in our experiments.

3 Methods

We present *Filling the Unseen*, a hierarchical two-stage framework for large-scale and geometrically coherent scene extrapolation and interpolation. Figure 3 provides an overview of the full framework.

Filling the Unseen is built upon a simple yet effective insight: a scene can be partitioned into mutually non-overlapping sub-regions to support parallel, conflict-free generation. These sub-regions are captured by a set of *independent camera views* $\mathbf{V}_\mathrm{I}$, identified via the mechanism described in Section 3.2. Leveraging this mechanism, we design the main pipeline in two stages. In the first stage (Section 3.3.1), we perform parallel generation and reconstruction on $\mathbf{V}_\mathrm{I}$, enabling large-scale scene expansion in a single pass and eliminating the need for repetitive generate–reconstruct–shift cycles. In the second stage (Section 3.3.2), we further process the remaining un-extrapolated gaps between views in $\mathbf{V}_\mathrm{I}$, followed by a final interpolation step to achieve seamless scene extrapolation and interpolation. In Section 3.4, we introduce a novel quality-aware mask to take into consideration the inconsistency between generated content and the original scene, with the goal of promoting reliable updates while mitigating undesirable side-effects.

3.1 Preliminary: Spherical Harmonics in 3DGS

In 3DGS, Spherical Harmonics (SH) are employed as a set of orthonormal basis functions Y to represent view-dependent color appearance. Each 3D Gaussian stores a set of learnable SH coefficients $a_{l,m}$, where l denotes the SH degree and m indexes different directional modes. The **0**th-order term $a_{0,0}$ captures the view-independent base color, and higher-order terms represent the view-dependent color. The final color $c(\mathbf{d})$ for the viewing direction $\mathbf{d}$ is computed by adding these two parts together:

$$c(\mathbf{d}) = \underbrace{a_{0,0}Y_0^0}_{c_{\text{base-color}}} + \underbrace{\sum_{l=1}^{L}\sum_{m=-l}^{l} a_{l,m}\,Y_l^m(\mathbf{d})}_{c_{\text{view-dependent-color}}(\mathbf{d})}, \qquad Y_0^0 = \frac{1}{2\sqrt{\pi}}, \tag{1}$$

where Y_l^m denotes the SH basis function and L denotes the maximum degree of SH. This directional modeling capability is central to our QA-Mask design in Section 3.4.

3.2 Independent Camera View Detection

Given a reconstructed 3D scene, the user first specifies a set of camera views $\mathbf{V} = \{\mathbf{V}^1, \mathbf{V}^2, ..., \mathbf{V}^N\}$ that indicates the regions of interest for extrapolation. We automatically identify a subset of independent camera views $\mathbf{V}_\mathrm{I} \subseteq \mathbf{V}$, where two views $\mathbf{V}^i$ and $\mathbf{V}^j$ are considered *independent* if their inpainted regions do not overlap. Independent camera views can be inpainted and reconstructed in parallel without conflicts. Detecting independence at the 2D image level is usually insufficient. Feature-based similarity metrics such as CLIP [21] or DINO [18] are too coarse-grained for overlap detection, and feature matching methods [15, 24] fail on textureless surfaces like plain walls. We therefore operate in 3D, using mesh-based view frustum collision detection, as illustrated in Figure 2.

The key idea is to construct a mesh-based view frustum that tightly encloses the 3D geometry of the inpainted region for each candidate view $\mathbf{V}^\mathrm{i}$. If the frustums of $\mathbf{V}^\mathrm{i}$ and $\mathbf{V}^\mathrm{j}$ do not collide, their inpainted regions are guaranteed not to overlap, and the two views are independent. We proceed as follows:

(1) **Inpainting and depth estimation.** Inpaint each $\mathbf{V}^\mathrm{i}$ with inpainting mask $\mathbf{M}^\mathrm{i}$ to get $\mathbf{I}^{\mathrm{i,j}}$. Then estimate each relative depth $\mathbf{R}^{\mathrm{i,j}}$ using Depth Anything V2 [33].

(2) **Depth warping and alignment.** Find the closest camera pose $\mathbf{P}^\mathrm{i}$ to $\mathbf{V}^\mathrm{i}$ in the original training set, obtain the dense Colmap depth $\mathbf{D}^\mathrm{i}$, and use depth image-based rendering (DIBR) to warp it to $\mathbf{V}^\mathrm{i}$. Then align the warped depth with $\mathbf{R}^{\mathrm{i,j}}$ to obtain a scale $s_{i,j}$ and an offset $t_{i,j}$:

$$s_{i,j}, t_{i,j} = \arg\min_{s,t} \sum \left\| Warp^{P^i \to V^i}(D^i) - R^{i,j} \right\|^2. \tag{2}$$

(3) **View frustum construction.** Use $s_{i,j}$ and $t_{i,j}$ to adjust $\mathbf{R}^{\mathrm{i,j}}$ in the inpainted region:

$$D^{i,j} = (s_{i,j} \times R^{i,j} + t_{i,j}) \times M^i, \tag{3}$$

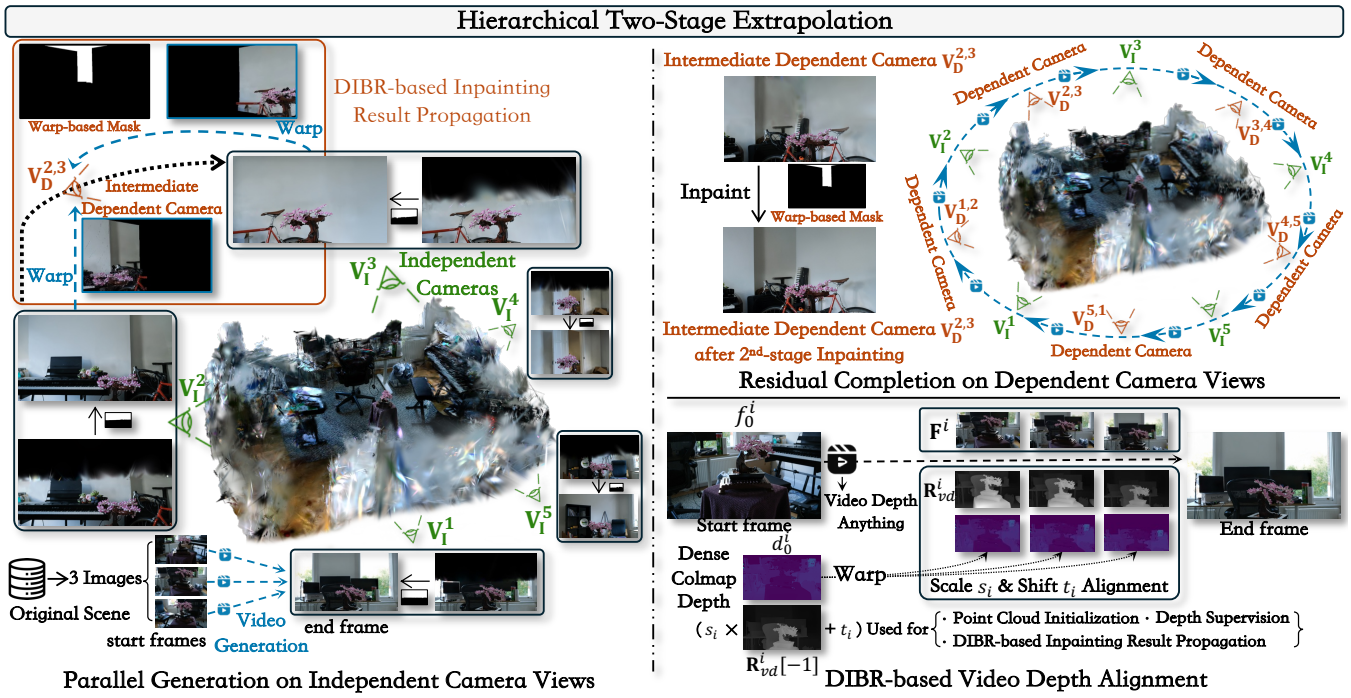


Figure 3: An overview of our main pipeline. It consists of extrapolations for independent and dependent camera views. The former achieves extrapolation on the original scene over large areas in a single pass. The latter performs a second-stage residual completion on the remaining un-extrapolated regions, enabling seamless scene extrapolation. DIBR-based video depth alignment provides robust depth priors throughout the process. Blue indicates the generated data used in each stage.

and calculate the depth range $[\mathbf{D}^{i,j}_{\min}, \mathbf{D}^{i,j}_{\max}]$ for $D^{i,j}$. Next compute an average depth range $[\mathbf{D}^{i}_{\min}, \mathbf{D}^{i}_{\max}]$ across $\mathbf{n}$ inpainting results for $\mathbf{V}^{i}$:

$$D^{i}_{min} = \frac{1}{n}\sum_{j=1}^{n} D^{i,j}_{min},\ D^{i}_{max} = \frac{1}{n}\sum_{j=1}^{n} D^{i,j}_{max}. \tag{4}$$

$[\mathbf{D}^{i}_{\min}, \mathbf{D}^{i}_{\max}]$ represents a high-probability depth range for the inpainting region of $\mathbf{V}^{i}$. We use $\mathbf{D}^{i}_{\min}$ and $\mathbf{D}^{i}_{\max}$ as the near and far planes to construct a view frustum for only the inpainting region of $\mathbf{V}^{i}$. This frustum encloses the 3D geometry of the inpainted results.

(4) **Frustum collision detection.** After constructing frustums for each $\mathbf{V}^{i}$, we identify a set of independent camera views $\mathbf{V}_{\mathrm{I}}$ via mesh-based collision detection. Next, we sort $\mathbf{V}_{\mathrm{I}}$ based on the nearest neighbor and generate a closed camera view trajectory used for the following extrapolation.

3.3 Hierarchical Two-stage Scene Extrapolation and Interpolation

3.3.1 Parallel Generation on Independent Camera Views. In the first stage, we extrapolate the original scene on $\mathbf{V}_{\mathrm{I}}$ over large areas in a single pass. We use a diffusion-based inpainting model [38] for each $\mathbf{V}^{i}_{\mathrm{I}}$ to fill up the black holes due to un-observations. Next we select 3 closest camera views to $\mathbf{V}^{i}_{\mathrm{I}}$ from the original training set and generate 3 videos using [31] in a way that one original training image being the start frame and the inpainted image of $\mathbf{V}^{i}_{\mathrm{I}}$ being the end frame. These generated videos are used for the first stage extrapolation, as shown in the left part of Figure 3.

However, seamless extrapolation cannot be achieved by the first stage alone. There is unfilled gap captured by the intermediate dependent camera view $\mathbf{V}^{i,i+1}_{\mathrm{D}}$ between $\mathbf{V}^{i}_{\mathrm{I}}$ and $\mathbf{V}^{i+1}_{\mathrm{I}}$. We will perform extrapolation on $\mathbf{V}_{\mathrm{D}}$ in the following second stage, which is the only one-time execution of the generate-reconstruct-shift cycle in our pipeline. Since image inpainting quality is highly related to the quality of the image itself, to ensure high inpainting quality for $\mathbf{V}_{\mathrm{D}}$ in the next stage, the reconstruction in this stage needs to be robust to handle view shifting. Therefore, we aim to acquire as much multi-view information as possible during the first stage reconstruction. We employ DIBR-based warping to propagate the inpainted results of $\mathbf{V}_{\mathrm{I}}$ to $\mathbf{V}_{\mathrm{D}}$. Next, we introduce how to obtain accurate depth priors and perform propagations using the depths.

DIBR-based Video Depth Alignment. Estimating relative depth by [33] and scale-aligning it with SfM points are a common way [5] to acquire depth priors. However, it is inherently unstable and error-prone because of the sparsity and noise of SfM points. So we employ Video Depth Anything [3] and DIBR to perform a robust depth alignment. As illustrated in the bottom right of Figure 3, for each $\mathbf{V}^{i}_{\mathrm{I}}$, we randomly choose one from the three generated videos and use [3] to obtain relative video depth $\mathbf{R}^{i}_{\mathrm{vd}}$. Since the start frame f^{i}_{0} of the video is one original training image, we can warp its dense Colmap depth d^{i}_{0} to all following frames $\mathbf{F}^{i}$, and then conduct a

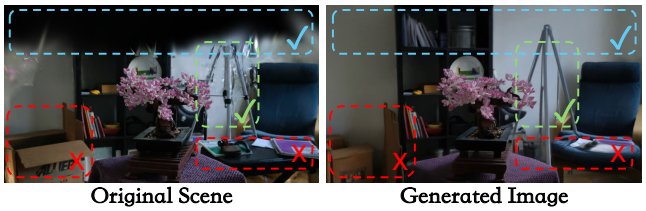


Figure 4: Under-reconstructed areas, including holes and low-quality regions, should be updated using generated images. In contrast, well-reconstructed areas within the original scene should remain unchanged to avoid compromising the overall quality due to potential generative inconsistencies.

robust multi-image depth alignment using RANSAC and obtain the scale s_i and offset t_i. We acquire depth $\mathbf{D}_{\mathrm{I}}^{i}$ for the inpainted image of $\mathbf{V}_{\mathrm{I}}^{i}$ by adjusting the last frame of $\mathbf{R}_{\mathrm{vd}}^{i}$ with s_i and t_i:

$$s_i, t_i = \arg\min_{s,t} \sum \left\| Warp^{f_0^i \to \mathbf{F}^i}(d_0^i) - \mathbf{R}_{\mathrm{vd}}^{i} \right\|^2, \tag{5}$$

$$\mathbf{D}_{\mathrm{I}}^{i} = s_i \times \mathbf{R}_{\mathrm{vd}}^{i}[-1] + t_i. \tag{6}$$

$\mathbf{D}_{\mathrm{I}}^{i}$ can be used in extrapolation point cloud initialization, DIBR-based inpainting result propagation, and depth supervision.

DIBR-based Inpainting Result Propagation. We utilize depth priors $\mathbf{D}_{\mathrm{I}}^{i}$ and $\mathbf{D}_{\mathrm{I}}^{i+1}$ to warp the inpainted results on $\mathbf{V}_{\mathrm{I}}^{i}$ and $\mathbf{V}_{\mathrm{I}}^{i+1}$ to the intermediate dependent camera view $\mathbf{V}_{\mathrm{D}}^{i,i+1}$ between $\mathbf{V}_{\mathrm{I}}^{i}$ and $\mathbf{V}_{\mathrm{I}}^{i+1}$, as shown in the upper left of Figure 3. This provides additional supervision for improving the rendering quality of $\mathbf{V}_{\mathrm{D}}^{i,i+1}$ to facilitate inpainting the remaining un-extrapolated gap in $\mathbf{V}_{\mathrm{D}}^{i,i+1}$ at the second stage. Additionally, a warp-based mask indicating un-extrapolated gap is also acquired during DIBR-based warping to guide the later inpainting in the next stage.

3.3.2 Residual Completion on Dependent Camera Views.
In the second stage, we first inpaint each intermediate dependent camera view $\mathbf{V}_{\mathrm{D}}^{i,i+1}$. Then every independent or dependent camera view in the closed extrapolation trajectory corresponds to a complete image without un-extrapolated regions. Next, as illustrated in the upper right of Figure 3, we generate a video for every two neighboring views in the trajectory with one being the start frame and the other being the end frame. These videos are used for the second stage extrapolation.

With these two stages, the original scene has been extrapolated at a large scale along the extrapolation trajectory.

3.3.3 Scene Interpolation via Video Reuse.
Scene interpolation can improve overall rendering quality in the original observed regions, which suffer from degradations on views from the extrapolation trajectory, as shown in Figure 1. We exploit two sources of interpolation data. First, the videos generated during extrapolation inherently contain intermediate frames between original training images and inpainted views, providing multi-view supervision for the original scene. Second, after stage 2 extrapolation, we render images along the closed extrapolation trajectory and pass them through the artifact removal model [29], whose outputs are added to the training set on-the-fly. Both strategies are applied jointly during this final interpolation stage.

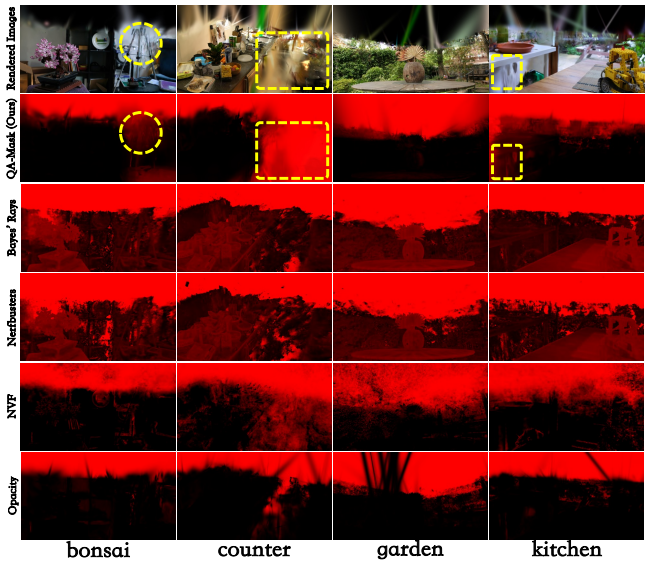


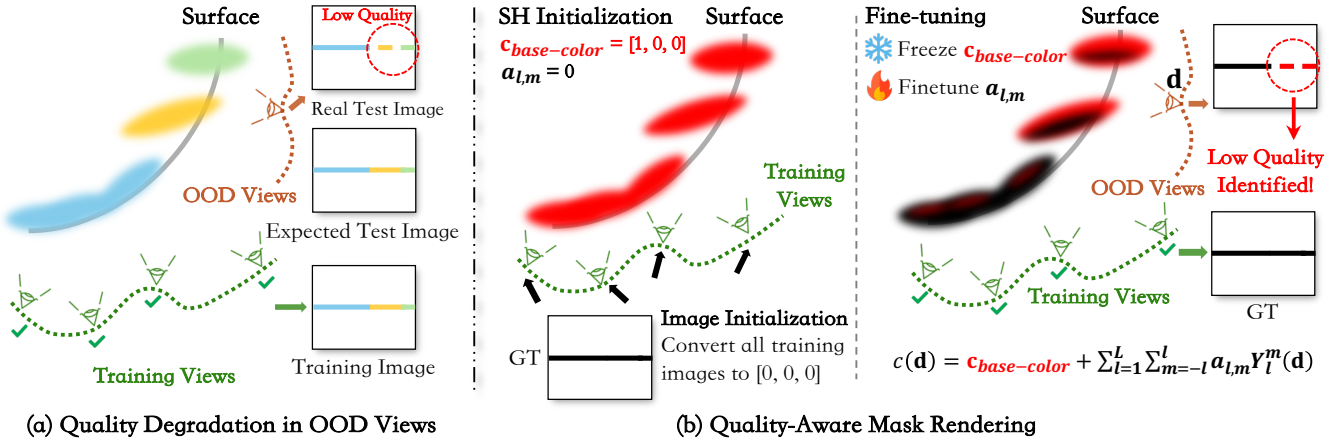
Figure 5: Visualizations of rendered RGB images and QA-Mask from extrapolation views on four scenes in Mip-NeRF 360. QA-Mask identified holes and low rendering quality regions (dashed yellow) that align well with human perception. However, Bayes' Rays and Nerfbusters fail to clearly distinguish between low-quality and well-reconstructed areas, while opacity maps and NVF are unable to accurately identify the low-quality areas within the original scene or maintain high (uncertainty) values across the hole areas.

3.4 Quality-Aware Mask Rendering

When updating a scene with generated content, inconsistency between the generated data and the original scene often occurs. This necessitates selective utilization on the generated data. A reasonable strategy is to spatially re-weight loss functions based on current rendering quality. As exemplified in Figure 4, in areas without Gaussians or with low rendering quality, generated data should be applied for extrapolation and interpolation. While in well-reconstructed areas, they should not be used, to prevent data inconsistency from corrupting the original scene. Based on this idea, in this section, we propose our novel method to render Quality-Aware Mask (QA-Mask).

The core of 3DGS is to find one feasible solution for a set of 3D Gaussian primitives under multi-view constraints. This solution overfits on training views but underfits on out-of-distribution novel views. In other words, rendering quality is highly related to the viewing directions (Figure 6 (a)). Since Spherical Harmonics (SH) possesses directional modeling ability, we exploit SH to model rendering quality.

Next, we first outline the steps to obtain QA-Mask from a 3DGS model pretrained on the original scene, as illustrated in Figure 6 (b), and then explain how it works:



(a) Quality Degradation in OOD Views (b) Quality-Aware Mask Rendering

Figure 6: (a) Rendering degradations on out-of-distribution views. 3D Gaussians overfit on training views but underfit on out-of-distribution (OOD) views which results in degraded rendering. (b) Quality-Aware Mask rendering method. We first prepare a pretrained 3DGS model by setting the SH base color to red, view-dependent color coefficients $a_{l,m}$ to 0 and converting all original training images to black. We fine-tune this model with only $a_{l,m}$ receiving gradients. After fine-tuning, within the training view distribution with high rendering quality, the base color is canceled out by $a_{l,m}$ with the final color being black. While in OOD views with low rendering quality, the lack of black supervisions causes Gaussians to remain the base color.

(1) **SH coefficient initialization.** We set $c_{\text{base-color}}$ to one arbitrary color from RGB, e.g., red in our example, and the view-dependent coefficients $a_{l,m}$ to 0.
(2) **Training image initialization.** We convert original training images to all black with RGB of $(0, 0, 0)$.
(3) **Fine-tuning.** We fine-tune this 3DGS model on converted black training images with only $a_{l,m}$ receiving gradients.

During fine-tuning, the Gaussians with the fixed red base color rely on view-dependent $a_{l,m}$ to shift the final color from red to black. However, on out-of-distribution novel views that are far from training supervisions, since without black signals, the final SH color remains largely red. At its core, we use SH to model "which directions are supervised." SH colors such as black and red in this example, can then serve as quantitative indicators for how well a pixel is supervised from this direction, i.e., the rendering quality. We show some visualization examples in Figure 5, where the background color is set to red during rendering. It is worth noting the choice of red/black is arbitrary and any two distinguishable colors suffice. We verified it with alternative colors in the supplementary.

After normalizing the R channel of QA-Mask $\mathbf{M}$, they are employed as per-pixel learning weights applied to our reconstruction loss, emphasizing under-reconstructed areas while suppressing updates in well-reconstructed areas:

$$\mathcal{L} = \lambda_1 \|\mathbf{M}\odot(\mathbf{I}_{\text{p}}-\mathbf{I}_{\text{gt}})\|_1 + \lambda_2 \text{SSIM}(\mathbf{M}\odot\mathbf{I}_{\text{p}}, \mathbf{M}\odot\mathbf{I}_{\text{gt}}) + \lambda_3 \|\mathbf{M}\odot(\mathbf{D}_{\text{p}}-\mathbf{D}_{\text{gt}})\|_1. \quad (7)$$

where $\mathbf{I}_{\text{p}}$, $\mathbf{I}_{\text{gt}}$ denotes rendered and ground truth images, $\mathbf{D}_{\text{p}}$, $\mathbf{D}_{\text{gt}}$ denotes rendered and ground truth depths.

Table 1: Non-reference metrics and runtime on Mip-NeRF 360. C-IQA denotes CLIP-IQA. QUALI-C denotes QUALI-CLIP.

Methods	C-IQA↑	QUALI-C↑	MVC↑	SemSim↑	Time
Few-Shot GS	0.278	0.352	1792	0.707	24min
DiFix3D+	0.354	0.451	1755	0.687	53min
GenFusion	0.340	0.378	1157	0.712	44min
Guidedvd-3dgs	0.230	0.298	1845	0.691	306min
Ours	**0.404**	**0.498**	**1972**	**0.737**	77min

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on two datasets Mip-NeRF 360 [2] and ScanNet++ [34]: 4 scenes from Mip-NeRF 360 (bonsai, counter, garden, kitchen) and 10 scenes from ScanNet++.
Evaluation Settings. Since there are no ground-truth images for extrapolation viewpoints in Mip-NeRF 360 by construction, reference metrics (PSNR/SSIM/LPIPS) are inapplicable. We therefore adopt non-reference metrics including **CLIP-IQA** [25], **QualiCLIP** [1], matching-based Multi-View Consistency (**MVC**) [6, 22], and CLIP-based Semantic Similarity (**SemSim**) [21] as proxies for perceptual quality, multi-view consistency, and semantic fidelity.

For ScanNet++, we design a controlled evaluation protocol to enable quantitative assessment using reference metrics, including **PSNR**, **SSIM**, and **LPIPS**. To implement this, we reconstruct only half of each scene, and treat the other half as the unseen extrapolation target, where original training images serve as ground truth for reference-metric evaluation. Details about metrics and settings can be found in the supplementary.

Figure 7: Qualitative comparisons of extrapolation on Mip-NeRF 360. The first row shows the inpainted images on independent camera views $\mathbf{V_I}$. The following rows show the images rendered on the camera views between each pair of $\mathbf{V_I^i}$, $\mathbf{V_I^{i+1}}$. Our framework achieves high rendering quality on extrapolated areas while also leaves well-reconstructed original areas untouched.

Table 2: Comparisons of reference metrics on ScanNet++. Our method achieves the best performance averaged on 10 scenes.

Methods	PSNR↑	SSIM↑	LPIPS↓
Few-Shot GS	18.92	0.80	0.3438
DiFix3D+	20.93	0.84	0.2554
GenFusion	21.93	0.85	0.2502
Guidedvd-3dgs	22.10	0.84	0.2662
Ours	**23.79**	**0.88**	**0.2300**

4.2 Baselines and Implementation Details

Baselines. We compare with recent SOTA methods: Few-Shot Gaussian Splatting (FSGS) [40], DiFix3D+ [29], GenFusion [30], and Guidedvd-3dgs [39]. For experiments on Mip-NeRF 360, we supply FSGS and DiFix3D+ with the same inpainted images on independent camera views produced by our framework. As these methods are not explicitly designed for scene extrapolation, we adapt their settings accordingly to enable a meaningful comparison.

Implementation Details. We use $\lambda_1 = 0.7$, $\lambda_2 = 0.2$, $\lambda_3 = 0.1$ in our experiments. When performing reconstructions, our method is trained for 30,000 iterations for each stage. For QA-Mask, we use 5-degree SH and fine-tune for 20,000 iterations. All experiments are conducted on a NVIDIA RTX A6000 GPU.

4.3 Comparisons on Non-reference Setting

The non-reference metrics and runtime evaluated on Mip-NeRF 360 are reported in Table 1. Our method outperforms these baselines over all metrics at an acceptable time cost, demonstrating the superiority of our method in terms of perceptual image quality, multi-view consistency, and high-level semantic consistency. We show qualitative comparisons in Figure 7. Our method enjoys high rendering quality and multi-view consistency in extrapolated regions compared to baselines that are plagued by blurring and floaters. This stems from our hierarchical two-stage extrapolation design that effectively prevents the compounding of errors observed in prior approaches. In the same time, no sign of rendering deterioration can be observed within original areas for our method thanks to the proposed QA-Mask. While other models such as GenFusion and Guidedvd-3dgs suffer from quality drops in these areas.

Table 3: Generalizability of QA-Mask on ScanNet++. Exst. denotes existing and Extrapltd. denotes extrapolated.

Methods	PSNR↑		LPIPS↓	
	Exst.	Extrapltd.	Exst.	Extrapltd.
DiFix3D+	28.11	**10.23**	0.1187	0.4654
w/ Bayes' Rays	27.87	10.20	0.1235	0.4701
w/ Nerfbusters	27.73	10.09	0.1193	0.4659
w/ NVF	28.19	10.02	0.1093	0.4700
w/ Opacity Maps	28.15	10.17	0.1109	0.4655
w/ QA-Mask (Ours)	**28.58**	10.20	**0.1075**	**0.4653**
GenFusion	26.49	15.04	0.1367	**0.4156**
w/ Bayes' Rays	27.79	14.99	0.1146	0.4161
w/ Nerfbusters	28.01	15.00	0.1203	0.4176
w/ NVF	28.07	14.77	0.1121	0.4203
w/ Opacity Maps	28.03	15.04	0.1296	0.4186
w/ QA-Mask (Ours)	**28.33**	**15.10**	**0.1056**	0.4193

The runtime for our method in Table 1 includes the two-stage extrapolation pipeline, interpolation, and the training of QA-Mask. More images and videos are provided in the supplementary.

4.4 Comparisons on Reference Setting

Quantitative results for reference metrics on ScanNet++ are reported in Table 2. Qualitative comparisons are shown in the supplementary, where our method delivers noticeably better visual quality in both extrapolated and existing regions. Other methods exhibit artifacts and geometric distortions to varying extents, despite the availability of partial ground-truth images in this setting. More images and videos on ScanNet++ are provided in the supplementary.

4.5 Generalizability of QA-Mask

In this section, we compare our QA-Mask with a couple of baselines including opacity maps, Bayes' Rays [7], Nerfbusters [27], and Neural Visibility Field (NVF) [32] by plugging them into DiFix3D+ and GenFusion, and carry out a comparative study on ScanNet++. We report PSNR and LPIPS for the extrapolated and existing regions separately in Table 3. After introducing QA-Mask, the metrics of the existing regions have been improved for both DiFix3D+ and GenFusion, which demonstrates that QA-Mask can prevent generated data from compromising the original scene. Meanwhile, extrapolated regions maintain comparable rendering quality than the counterpart without QA-Mask, proving that QA-Mask does not hinder desired scene updates using generated data. This appealing feature of our QA-Mask, which we did not observe in other baselines as shown in Figure 5, makes it serve as a plug-and-play enhancement module for a series of generation-based reconstruction models. Qualitative comparative examples are shown in Figure 8.

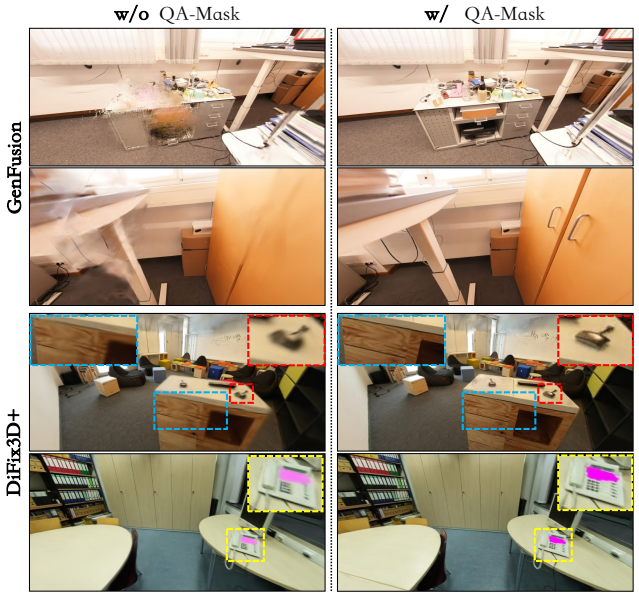


Figure 8: Comparison between w/ and w/o QA-Mask on GenFusion and DiFix3D+. QA-Mask prevents generated data from disrupting well-reconstructed areas in the original scene. Noticeable improvements can be observed for both models.

Table 4: Ablation study for core components and hyperparameter selection on both Mip-NeRF 360 and ScanNet++. Idpdt Cam Detection denotes Independent Camera View Detection.

Mip-NeRF 360	CLIP-IQA↑	MVC↑	SemSim↑
w/o QA-Mask	0.343	1673	0.719
w/o DIBR Propagation	0.316	1468	0.697
Ours	**0.404**	**1972**	**0.737**
ScanNet++	**PSNR↑**	**SSIM↑**	**LPIPS↓**
w/o Idpdt Cam Detection	20.27	0.80	0.3161
w/o Depth Supervision	23.66	**0.88**	0.2302
w/ Degree-3 SH QA-Mask	22.45	0.84	0.2432
w/ Degree-4 SH QA-Mask	23.28	0.86	0.2331
Ours	**23.79**	**0.88**	**0.2300**

4.6 Ablation Study

In this section, we perform ablation studies for core components and hyperparameter selection on both Mip-NeRF 360 and ScanNet++, and present the results in Table 4. Experiments on Mip-NeRF 360 showed that removing QA-Mask or DIBR-based propagation both yields worse performance. Since QA-Mask is designed to suppress unintended updates without impeding desired extrapolation and interpolation. DIBR-based propagation leverages depth priors to provide additional multi-view cues for the first stage reconstruction so that facilitates the transition to the next stage.

We also conducted experiments on ScanNet++. First by removing the independent camera view detection mechanism and evenly choosing the same number of camera views in a conflict-agnostic manner, the rendered results are severely plagued by content conflicts reflecting as considerable performance drop. Second, the w/o depth supervision variant achieved slightly lower performance, indicating our video depth alignment module provides accurate depth information. Additionally, we quantitatively justified our choice for the 5-degree SH in QA-Mask over 3 or 4-degree, where high degree enhances SH's ability to capture high-frequency angular variations, enabling more precise directional modeling.

5 Conclusion

We introduce ***Filling the Unseen***, a holistic framework for large-scale scene extrapolation and interpolation based on 3D Gaussian Splatting. By first detecting independent camera views that support parallel generation and reconstruction, we redesign a hierarchical two-stage extrapolation pipeline and achieved high-quality scene extrapolation. We hope this hierarchical pipeline can provide new design perspective on scene generation tasks for the community. Furthermore, the proposed QA-Mask effectively prevents generated content from negatively impacting the well-reconstructed areas of original scenes, which can serve as a plug-and-play enhancement module for a series of generation-based reconstruction models.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. 2024. Quality-Aware Image-Text Alignment for Opinion-Unaware Image Quality Assessment. *arXiv preprint arXiv:2403.11176* (2024).

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

[3] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2025. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 22831–22840.

[4] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. 2023. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384* (2023).

[5] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. 2024. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 811–820.

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 224–236.

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

[8] Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. 2025. Pup 3d-gs: Principled uncertainty pruning for 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5949–5958.

[9] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7909–7920.

[10] Yi-Chuan Huang, Hao-Jen Chien, Chin-Yang Lin, Ying-Huan Chen, and Yu-Lun Liu. 2025. GaMO: Geometry-aware Multi-view Diffusion Outpainting for Sparse-View 3D Reconstruction. *arXiv preprint arXiv:2512.25073* (2025).

[11] Sungwon Hwang, Min-Jung Kim, Taewoong Kang, Jayeon Kang, and Jaegul Choo. 2024. Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. In *European Conference on Computer Vision*. Springer, 1–18.

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

[13] Ruiqi Li and Yiu-ming Cheung. 2024. Variational multi-scale representation for estimating uncertainty in 3d gaussian splatting. *Advances in Neural Information Processing Systems* 37 (2024), 87934–87958.

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

[15] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[16] Yiren Lu, Jing Ma, and Yu Yin. 2024. View-consistent object removal in radiance fields. In *Proceedings of the 32nd ACM international conference on multimedia*. 3597–3606.

[17] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[19] Hao Ouyang, Kathryn Heal, Stephen Lombardi, and Tiancheng Sun. 2023. Text2immersion: Generative immersive scene with 3d gaussians. *arXiv preprint arXiv:2312.09242* (2023).

[20] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. 2022. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*. Springer, 230–246.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[22] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4938–4947.

[23] Meng-Li Shih, Wei-Chiu Ma, Lorenzo Boyice, Aleksander Holynski, Forrester Cole, Brian Curless, and Janne Kontkanen. 2024. Extranerf: Visibility-aware view extrapolation of neural radiance fields with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20385–20395.

[24] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8922–8931.

[25] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*.

[26] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5294–5306.

[27] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. 2023. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18120–18130.

[28] Jiaxin Wei, Stefan Leutenegger, and Simon Schaefer. 2025. GSFix3D: Diffusion-Guided Repair of Novel Views in Gaussian Splatting. *arXiv preprint arXiv:2508.14717* (2025).

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

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

[31] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2024. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*. Springer, 399–417.

[32] Shangjie Xue, Jesse Dill, Pranay Mathur, Frank Dellaert, Panagiotis Tsiotra, and Danfei Xu. 2024. Neural visibility field for uncertainty-driven active mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18122–18132.

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *Advances in Neural Information Processing Systems* 37 (2024), 21875–21911.

[34] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. 2023. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[35] Xingyilang Yin, Qi Zhang, Jiahao Chang, Ying Feng, Qingnan Fan, Xi Yang, Chi-Man Pun, Huaqi Zhang, and Xiaodong Cun. 2025. Gsfixer: Improving 3d

gaussian splatting with reference-guided video diffusion priors. *arXiv preprint arXiv:2508.09667* (2025).

[36] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. 2025. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5916–5926.

[37] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2024. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics* 30, 12 (2024), 7749–7762.

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

[39] Yingji Zhong, Zhihao Li, Dave Zhenyu Chen, Lanqing Hong, and Dan Xu. 2025. Taming Video Diffusion Prior with Scene-Grounding Guidance for 3D Gaussian Splatting from Sparse Inputs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6133–6143.

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