Perspective-aware 3D Gaussian Inpainting with Multi-view Consistency

Yuxin Cheng, Binxiao Huang, Taiqiang Wu, Wenyong Zhou, Chenchen Ding
Zhengwu Liu, Graziano Chesi, Ngai Wong*
The University of Hong Kong, Hong Kong SAR, China
{yxcheng,huangbx7,takiwu,wenyongz,dingcc}@connect.hku.hk, {zwliu,chesi,nwong}@eee.hku.hk

Abstract

3D Gaussian inpainting, a critical technique for numerous applications in virtual reality and multimedia, has made significant progress with pretrained diffusion models. However, ensuring multi-view consistency, an essential requirement for high-quality inpainting, remains a key challenge. In this work, we present PAInpainter, a novel approach designed to advance 3D Gaussian inpainting by leveraging perspective-aware content propagation and consistency verification across multi-view inpainted images. Our method iteratively refines inpainting and optimizes the 3D Gaussian representation with multiple views adaptively sampled from a perspective graph. By propagating inpainted images as prior information and verifying consistency across neighboring views, PAInpainter substantially enhances global consistency and texture fidelity in restored 3D scenes. Extensive experiments demonstrate the superiority of PAInpainter over existing methods. Our approach achieves superior 3D inpainting quality, with PSNR scores of 26.03 dB and 29.51 dB on the SPIn-NeRF and NeRFiller datasets, respectively, highlighting its effectiveness and generalization capability. The code will be publicly available at https://pa-inpainter.github.io.

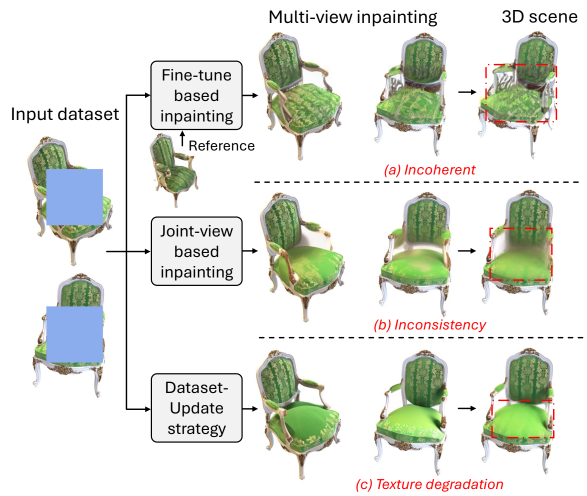

Figure 1. Current challenges in 3D Gaussian inpainting: a) the fine-tune based inpainter trained for specific tasks [7] experiences significant performance decline when applied to general inpainting scenarios; b) the joint-view inpainting method [44] struggles with inconsistency across multi-view images, resulting in noisy inpainting results; c) the DU strategy [12] leads to texture degradation in both inpainted multi-view images and the final 3D scene.

1. Introduction

As a prominent application in the realm of 3D editing, 3D inpainting plays a pivotal role in various applications and industries, including the metaverse and holographic multimedia production [5, 52]. However, traditional hand-crafted 3D completion approaches, which rely on professional designers and specialized tools, remain labor-intensive and cumbersome. With recent advancements in 3D neural representations [17] and generative models [34], 3D inpainting can be achieved by applying a two-stage paradigm: 1) using a pretrained 2D diffusion model to inpaint masked multi-view renderings of the 3D Gaussian scene with missing regions; and 2) optimizing the initial 3D Gaussian scene with the inpainted multi-view images [7, 32]. While this efficient framework shows significant potential, multi-view inconsistency remains an inherent challenge in diffusion models due to their independent view processing nature, which hinders high-quality 3D inpainting [22, 44].

Existing works have explored various approaches to improve multi-view consistency in 3D Gaussian inpainting, yet new limitations continue to emerge. Fine-tune based inpainting methods adapt diffusion models with additional control conditions (e.g., reference images [7]), but are confined to specific scenarios, as illustrated in Fig. 1(a). Without modifying pretrained diffusion models, the joint-view based inpainting approach [44] processes multi-view images in 2×2 grid tiles and achieves improved consistency,

*Corresponding author

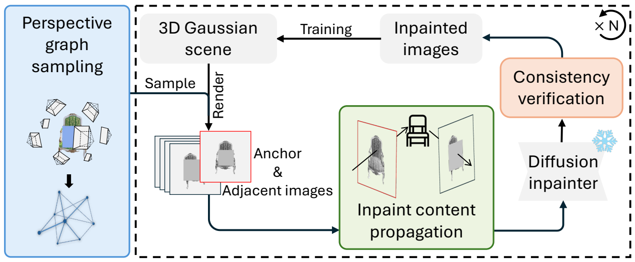

Figure 2. The overall pipeline of our proposed PAInpainter. Based on the constructed perspective graph, our approach iteratively performs multi-view image inpainting and 3D Gaussian training. The adaptive graph sampling algorithm enables efficient inpaint content propagation across adjacent viewpoints, while consistency verification ensures coherent multi-view inpainting results, thereby improving the 3D inpainting quality.

yet still exhibits artifacts in challenging regions, as shown in Fig. 1(b). Similarly, DatasetUpdate (DU) [12] alternates between 3D scene optimization and multi-view inpainting while progressively updating the dataset to improve consistency. However, it suffers from texture fading in the final results, as demonstrated in Fig. 1(c). These limitations highlight the persistent challenge of achieving high-fidelity and globally consistent inpainting across multiple views.

In this paper, we introduce the **P**erspective-**A**ware 3D Gaussian **Inpainter** (PAInpainter) to enhance multi-view consistency. As illustrated in Fig. 2, we propose a novel perspective-aware framework with three key components: perspective graph sampling, inpaint content propagation, and consistency verification. Specifically, we construct a perspective graph that models the spatial relationships among viewpoints. Leveraging adaptive graph sampling and the inherent perspective overlap between neighboring views, we propagate inpainted content across adjacent cameras, which serves as supplementary visual priors for the diffusion model during inpainting, improving fine-grained texture preservation and consistency across multi-view inpainted images. To ensure high-quality and reliable results, we introduce a dual-feature verification mechanism that evaluates both texture and geometric coherence in latent space, effectively identifying and selecting consistent inpainting results. Combined with our framework, the versatile generation capability of the pretrained diffusion model further empowers our approach to handle various challenging 3D inpainting scenarios.

Our approach demonstrates exceptional performance in high-fidelity 3D Gaussian inpainting across diverse scenarios. Through extensive experiments on three mainstream 3D inpainting datasets, we demonstrate that PAInpainter significantly outperforms existing methods both quantitatively and qualitatively. Additionally, our PAInpainter exhibits robust generalization capability across various scenarios. Our main contributions are summarized as follows:

- We propose a novel perspective-aware framework for 3D Gaussian inpainting that systematically integrates inpainting view sampling, cross-view content propagation, and consistency verification.
- We introduce three effective components: a perspective graph to guide viewpoint sampling for inpainting, a perspective-aware projection strategy to propagate inpainting content, and a dual-feature verification mechanism to ensure multi-view consistency.
- Extensive experiments on diverse 3D scenes demonstrate that PAInpainter outperforms state-of-the-art methods in achieving superior consistency and visual fidelity.

2. Related Work

2.1. 2D Image Inpainting

2D inpainting methods aim to restore missing or obscured regions in images with coherent textures and structures [3, 30]. Early approaches relied on texture synthesis and pixel interpolation techniques by leverages information from known regions [2, 9, 10]. Learning-based approaches, especially deep learning methods [18, 19, 31, 48, 49] and recent diffusion models [34], have since emerged as powerful alternatives, demonstrating superior capabilities in high-fidelity content completion. The Latent Diffusion Models (LDMs)[34] and its variants[24, 50] achieve remarkable generation results across diverse scenarios. However, these methods process each image independently without considering 3D spatial relationships and geometric attributes among multiple viewpoints, leading to subsequent inconsistency when applied to the 3D domain.

2.2. 3D Scene Inpainting

3D inpainting extends the content completion task into 3D space. Early approaches focused on geometric completion using traditional representations like point clouds and meshes [8, 46]. Recent advances in neural representations, particularly Neural Radiance Fields (NeRF)[25] and 3D Gaussian Splatting (3DGS)[17], have revolutionized 3D scene modeling. While direct 3D diffusion models [28, 36, 42] face challenges with limited training data and computational complexity, an alternative approach combines pretrained 2D diffusion models with 3D neural representations [15, 20, 21, 23, 26, 32, 39, 44, 45]. This paradigm shows promising results by combining the powerful generation capabilities of 2D diffusion models with real-time 3D reconstruction [7, 22]. Despite the advancement in 3D inpainting efficiency, ensuring geometric and appearance consistency across different viewpoints remains challenging. In this paper, we build our method on 3DGS, which enables fast training and real-time rendering, and achieve improved multi-view consistency by propagating extra prior informa-

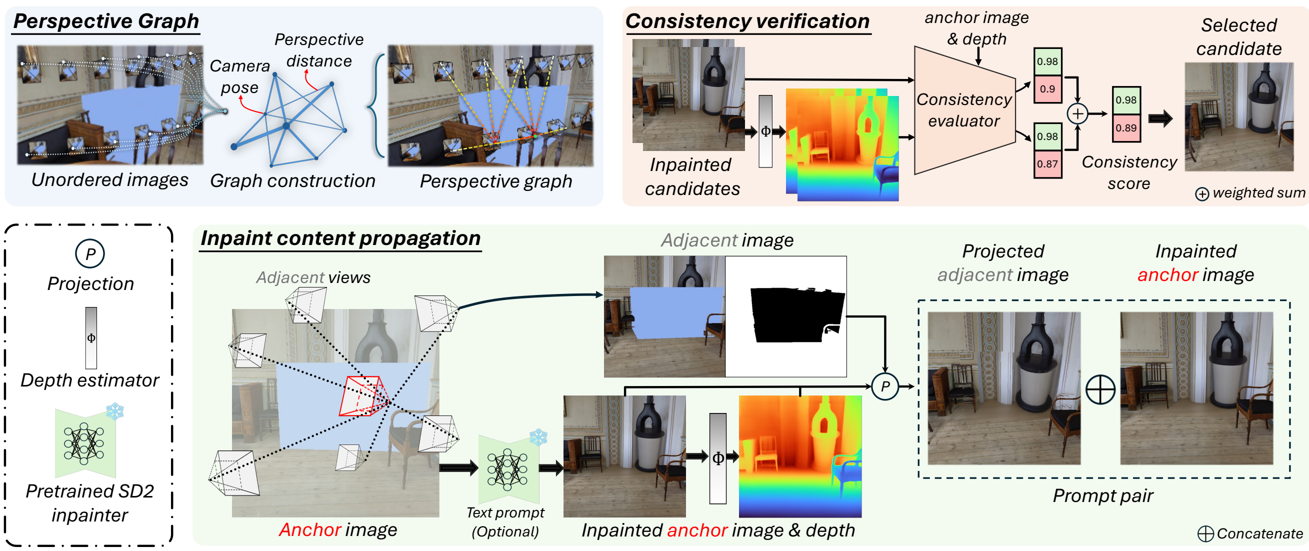

Figure 3. Overview of PAInpainter for multi-view consistent 3D Gaussian inpainting. Our method is built upon the pretrained SD2 [34] and incorporates three key components: 1) *perspective graph* models spatial relationships among cameras to guide adjacent view sampling; 2) *inpaint content propagation* transmits inpainting content across adjacent views sampled from perspective graph, providing extra visual priors for diffusion inpainting; 3) *consistency verification* evaluates inpainted results based on texture and geometric features coherence. The perspective-aware graph sampling contributes to effective content propagation and consistency verification across multiple views.

tion to guide the diffusion inpainting process.

2.3. Multi-view Consistency

Multi-view consistency ensures that the generated content in multi-view images of a 3D scene maintains geometric and texture coherence [12, 43]. Recent works have explored two main techniques to address inconsistency arising from 2D diffusion models. The first approach resorts to the diffusion model fine-tuning with additional conditions [16, 29, 43, 47], namely incorporating depth features, task-specific modules, and geometric constraints [6, 7, 22]. However, these methods typically specialize in specific tasks like object removal and struggle to generalize to broader 3D inpainting scenarios. The second technique explores solution without modifying pretrained models [12, 15, 44], but leverages depth priors or additional supervision [27, 32] to enhance cross-view consistency in generation and reconstruction process. While these approaches show potential, they often lack proactive consistency inspection of inpainted images, leading to compromised performance under challenging conditions. To address this limitation, we introduce a dual-feature verification mechanism designed to reject inconsistencies, thereby ensuring coherent inpainting across diverse scenarios for 3D scene restoration.

3. Methodology

Overview. The key components of PAInpainter are illustrated in Fig. 3. This section first introduces the overall framework (Sec. 3.1), followed by the technical details of perspective graph construction, inpaint content propagation, and consistency verification (Sec. 3.2). The adaptive graph sampling strategy and 3D Gaussian training procedure are elaborated in Sec. 3.3.

3.1. Framework

As shown in Fig. 2, PAInpainter completes unknown regions within a 3D scene by iteratively inpainting multi-view renderings and optimizing the 3D Gaussians with inpainted images. Building upon the high-fidelity image inpainting capabilities of pretrained StableDiffusion2 (SD2) [34], our framework enhances the multi-view inpainting consistency through three key techniques: perspective graph sampling, inpaint content propagation and consistency verification.

Based on the fact that the cross-attention mechanism of SD2 allows for reference-guided content generation in missing regions [7], we observe that the inpainting consistency significantly degrades with increasing perspective differences between views, as shown in Fig. 1(a). This observation motivates us to sample images with similar perspectives, thereby promoting the generation of consistent content. These adjacent views serve dual purposes: they facilitate effective content propagation by providing reliable texture and geometric priors for masked images, while enabling feature-space consistency verification among inpainting images to mitigate SD2's inherent randomness and select optimal results from multiple candidates.

Based on these findings, we develop PAInpainter based

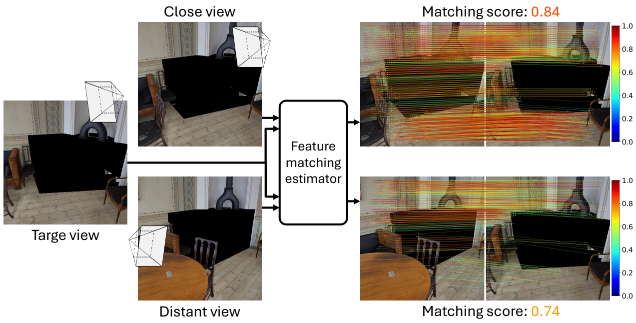

Figure 4. Perspective distance evaluation via feature matching. The color bar indicates the match's confidence. For a given target image, nearby views achieve significantly higher matching score (0.84) compared to distant views (0.74), validating the effectiveness of our perspective-aware graph construction method. This distance metric naturally captures the spatial relationships between different viewpoints.

on the following iterative framework:

1. Given a 3D Gaussian scene $\mathcal{G}_u$ with unknown regions, multi-view images $\mathbf{I} = \{I_i\}_{i=1}^n$ with corresponding camera poses $\mathbf{T} = \{T_i \in \mathrm{SE}(3)\}_{i=1}^n$ and masks $\mathbf{M} = \{M_i\}_{i=1}^n$, we construct a perspective graph $\mathbf{G}$ for $\mathbf{I}$, where edges encode the perspective distances among views.
2. For each inpainting round, we adaptively sample an anchor image I_{anchor} from the constructed graph and employ SD2 to inpaint it, obtaining I'_{anchor}. We then query its adjacent images from $\mathbf{G}$ to form a subset $\mathbf{I}_{adj}$. The inpainted content from I'_{anchor} is projected to each image in $\mathbf{I}_{adj}$, and I'_{anchor} serves as a reference image for following diffusion inpainting of these adjacent views.
3. For images in $\mathbf{I}_{adj}$, multiple inpainted candidates are generated by SD2. We then compute consistency scores between each candidate and I'_{anchor} with regard to the inpainting regions, selecting the candidate with the highest score as the final inpainting result.
4. We optimize the 3D Gaussian scene $\mathcal{G}_u$ by training on the inpainted images.

The process iteratively alternates between multi-view inpainting (steps 2-3) and 3D Gaussian optimization (step 4), progressively inpainting and refining the 3D scene.

3.2. PAInpainter

We now detail the three key modules of PAInpainter for achieving consistent 3D Gaussian inpainting.

Perspective graph construction. The graph $\mathbf{G}$ underpins our entire inpainting pipeline by modeling the proximity relationships among diverse viewpoints. Although the cameras' poses are available, the view difference, i.e., the captured content in the images, cannot be solely described by the 6-DoF distance due to variations in perspective, orientation, and scene geometry [1, 40]. To solve this problem, we propose evaluating view similarities through feature matching metrics, as shown in Fig. 4. Specifically, we employ LoFTR [37] for its transformer-based architecture that enables robust feature matching under challenging viewpoint changes. For image pair (I_i, I_j) in the dataset, we extract matches with confidence scores above threshold τ ($\tau = 0.4$ fixed in our implementation). The perspective distance is evaluated based on the average confidence score for these matches, where a higher average matching score indicates a smaller distance. In the final graph $\mathbf{G}$, nodes store images with their camera poses and masks, while edges encode the computed perspective distances. This perspective-aware graph enables effective sampling of adjacent views for consistent inpainting. As demonstrated by our experiments, this strategy provides enhanced robustness to viewpoint variations while preserving geometric interpretability.

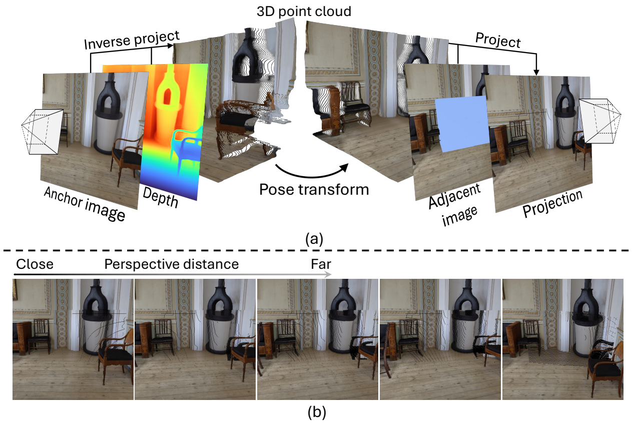

Figure 5. Inpaint content propagation mechanism. (a) Using depth information and camera poses, inpainted content from an anchor image is projected onto its adjacent masked views. (b) The projection results on neighboring views sampled from our perspective graph. Thanks to our graph-based sampling strategy, most masked regions in the selected images receive ample projected content (high coverage of effective pixels), offering rich prior information for subsequent SD2 inpainting.

Inpaint content propagation. To enhance multi-view consistency and high-fidelity inpainting, we feed supplementary priors of masked region along with the image to SD2 via inpaint content propagation. Guided by camera poses sampled from graph $\mathbf{G}$, we render the anchor image I_{anchor} and its top-k adjacent images $\mathbf{I}_{adj} = \{I_i^{adj}\}_{i=1}^k$ from 3D Gaussian scene $\mathcal{G}_u$. We first independently inpaint the anchor image I_{anchor} using SD2, obtaining I'_{anchor}, followed by propagating the I'_{anchor} to its adjacent images $\mathbf{I}_{adj}$ with masked regions through perspective projection to offer extra prior for SD2 inpainting, as shown in Fig. 5 (a).

Specifically, we use ZoeDepth [4] to estimate the depth map d_{anchor} for I'_{anchor}. With d_{anchor} and camera parameters (intrinsic K and extrinsic T_{anchor}), we inversely project the 2D image I'_{anchor} into perspective coordinates by

$$\left[x_c, y_c, z_c\right]^\top = K^{-1} \cdot (\left[u, v, 1\right]^\top \cdot d), \quad (1)$$

where $[u, v, 1]^\top$ and d represent 2D image coordinate and depth value, respectively, and $[x_c, y_c, z_c]^\top$ represents the 3D coordinate. For each adjacent image I_i^{adj} with camera pose T_i, we transform the 3D point cloud from anchor perspective to the perspective coordinates of I_i^{adj} by

$$\left[x'_c, y'_c, z'_c, 1\right]^\top = T_i \cdot T_{anchor}^{-1} \cdot \left[x_c, y_c, z_c, 1\right]^\top, \quad (2)$$

obtaining $[x'_c, y'_c, z'_c]^\top$ after homogeneous normalization. We project these coordinates onto I_i^{adj}, updating only pixels within the masked region. For regions where projection fails due to view differences or depth estimation errors, we retain the rendering RGB values from 3D Gaussian scene.

Leveraging our perspective graph sampling strategy, the projection effectively propagates the inpainted content from anchor image to adjacent frames while preserving geometric and texture consistency, as shown in Fig. 5 (b). The propagated adjacent images $\mathbf{I}_{adj}$ are then paired with I'_{anchor} as reference guidance for SD2 diffusion inpainting.

Consistency verification. Obstacles arising from perspective differences make it impossible for consistency verification to rely solely on pixel comparison. Therefore, we elevate the consistency verification process into feature space. We independently generate multiple inpainting candidates for each masked adjacent image. To handle varying 3D inpainting scenarios, we propose verifying consistency by assessing coherence in texture and geometry feature spaces. As shown in Fig. 3, for the inpainted candidates of I_i^{adj}, we use ZoeDepth to estimate the corresponding depth maps. We apply a feature extraction model as consistency evaluator (ResNet-18 [13]) to separately extract both RGB and depth features for each candidate, as well as for the inpainted anchor image and its depth map. Finally, we compute the cosine similarity between candidates and the inpainted anchor image based on fused dual features. The overall consistency score is computed as a weighted combination of RGB and depth similarities: $S = \eta S_{rgb} + (1 - \eta) S_{depth}$, where η controls the relative importance of texture and geometry consistency. The candidate with the highest consistency score is then selected as the final inpainting result. The weighting factor η is empirically set to 0.7 to balance fine texture details and structural coherence and four candidates generated for each image.

Given the small perspective differences between adjacent images and our dual-feature coherence approach, our consistency verification mechanism effectively identifies and excludes inconsistent inpainting results. Specifically, by leveraging hierarchical feature extraction capability of ResNet-18 at multiple scales and the complementary nature of RGB-depth feature pairs, this method significantly enhances the multi-view consistency of images fed into the 3D Gaussian optimization process, thereby improving the overall quality of 3D Gaussian inpainting.

3.3. Adaptive Sampling & 3D Gaussian Training

Adaptive sampling. During the iterative process, we strategically sample an anchor image and its k nearest neighbors on our proposed perspective graph $\mathbf{G}$ for inpainting and 3D Gaussian training. In the first iteration, the anchor image is selected from the entire dataset to initialize the process. For each subsequent iteration, we adopt a distance-aware sampling strategy: the anchor image is sampled from the pool of previously inpainted images, excluding both previously selected anchor images and their $k/2$ nearest neighbors from future anchor image selection. Here, k is scene-dependent, determining both the number of adjacent views for inpainting and the spatial separation between anchor images. This spatial constraint ensures well-distributed scene coverage and mitigates the risk of local region trapping. We further maintain a priority queue based on consistency scores from previous iterations, prioritizing images with worse coherence for refinement (the algorithm flowchart in Appendix B.3). This adaptive mechanism progressively improves both global consistency and local detail quality.

3D Gaussian Training. Given the dataset consisting of inpainted and masked multi-view images, we optimize 3D Gaussians following the vanilla 3DGS framework [17]. The optimization objective combines L1 and D-SSIM losses:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_1(I'_i - I_i) + \lambda\mathcal{L}_{SSIM}(I'_i - I_i),\ \lambda = 0.2, \quad (3)$$

where I'_i and I_i denote the rendering and inpainted images respectively. For masked images, we exclude the missing regions from loss computation during optimization.

Our method achieves high-quality multi-view consistent inpainting without fine-tuning pretrained diffusion models. As demonstrated in Fig. 7, PAInpainter effectively handles diverse 3D inpainting scenarios.

4. Experiments

In our experimental evaluation, we conduct comprehensive comparisons between PAInpainter and state-of-the-art approaches across 28 scenes, spanning 4 distinct inpainting tasks. Our framework employs the pretrained StableDiffusion2 (SD2) [34] as the backbone for image inpainting and ZoeDepth [4] for monocular depth estimation. Detailed implementation specifics and explanations of hyperparameters are provided in Appendix B.1 and Appendix B.2.

Datasets. To rigorously evaluate the effectiveness and generalization capability of PAInpainter, we utilize three mainstream datasets. The first dataset consists of 8 object-centric scenes derived from the *NeRF Blender* dataset [25], with

	NeRF Blender [25]				SPIn-NeRF [27]				NeRFiller [44]			
	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	FID ↓
Masked 3DGS	11.57	0.83	0.19	-	13.46	0.41	0.40	-	12.95	0.76	0.28	-
SD2 [34]	20.42	0.90	0.09	102.2	23.48	0.73	0.23	140.3	20.36	0.84	0.17	105.2
MVInpainter [7]	19.42	0.81	0.17	148.6	24.80	0.74	0.21	152.2	21.13	0.80	0.18	117.7
GridPrior + DU * [44]	22.77	0.92	0.08	104.2	25.19	0.79	0.20	151.2	26.97	0.92	0.13	121.9
NeRFiller * [44]	23.27	0.92	0.09	153.7	25.20	0.79	0.17	146.1	22.35	0.88	0.15	110.4
PAInpainter (ours)	24.19	0.92	0.08	101.8	26.03	0.81	0.15	121.7	29.51	0.94	0.08	96.1

Table 1. Quantitative comparison on the multiple datasets. We compare our method against advanced approaches on three datasets. Higher PSNR and SSIM, as well as lower LPIPS and FID indicate better performance. Cells are highlighted as follows: best, second best, third best. Our method surpasses *all baselines* across these metrics, demonstrating its efficacy and robust generalization. * represents replacing the original NeRF backbone with 3DGS for fair comparison. Detailed performance of each scene are provided in the Appendix C.

multi-view images at a resolution of 512×512. Missing regions are generated by masking the central 192×192 pixels in each image (as exemplified by the "chair" scene in Fig. 7). Additionally, we use the *SPIn-NeRF* dataset [27] as the second dataset for 3D unbounded scene inpainting. Since the SPIn-NeRF dataset only covers a single 3D inpainting task (foreground object removal), we also incorporate the dataset introduced by *NeRFiller* [44], which includes 10 real-world complex 3D inpainting scenes. Collectively, our experimental corpus of 28 scenes (details in Appendix C) encompasses multiple 3D inpainting scenarios: 1) large indoor missing region, 2) object-centric large missing region, 3) object-centric removal, and 4) multiple disjoint missing regions (illustrated in Fig. 7).

Baselines. We establish comprehensive comparisons with four representative state-of-the-art approaches, each embodying distinct technical paradigms:

- *SD2* [34]. A fundamental baseline that performs independent simultaneous inpainting across all multi-view images;
- *GridPrior+DU* [44]. An extension of the Dataset Update (DU) framework that processes random image batches in 2×2 grid patterns during iterative updates;
- *NeRFiller* [44]. A progressive joint-view inpainting strategy built upon SD2, emphasizing view-consistent content generation;
- *MVInpainter* [7]. A reference-guided approach that fine-tunes SD2 and incorporates single-view inpainting results as reference information.

We configure *GridPrior+DU* and *NeRFiller* to process twelve images per batch to meet computational constraints and provide *MVInpainter* with one inpainted reference image per scene. All above methods are built upon 3DGS and are evaluated on 3D Gaussian scenes with identical masked regions and camera poses to ensure fair comparison.

Metric. We assess the performance through quantitative analysis of rendered image quality from inpainted 3D Gaussian scenes. Our evaluation protocol employs four well-established metrics: PSNR [11] for pixel-wise accuracy, SSIM [41] for structural similarity, LPIPS [51] for perceptual quality, and FID [14] for distribution alignment between the generated and ground truth images. For progressive methods that iteratively refine the inpainting results, we evaluate by comparing the rendered images from inpainted 3D scene against the inpainted image of each view after the last iteration. For single-round methods, we adopt the train-test split strategy on inpainted images: 80% of images in the training set are used for optimizing the masked 3D Gaussian scene, while the remaining images as test set serve for evaluating the 3D inpainting results.

4.1. Experimental Results and Analysis

Quantitative. The experiment results of PAInpainter and state-of-the-art methods are presented in Tab. 1. The *Masked 3DGS* reports the rendering quality on the initial reconstructed 3D Gaussian scene with missing regions. Our proposed PAInpainter outperforms all other methods across the evaluated metrics. Specifically, on the NeRFiller dataset, PAInpainter achieves significant improvements with PSNR of 29.51 dB, surpassing the strongest baseline (GridPrior+DU) by 2.54 dB. This demonstrates its superiority in generating high-quality 3D Gaussian inpainting results and highlights its strong generalization capability across diverse inpainting scenarios.

In contrast, *SD2* inpaints images independently without considering multi-view consistency, resulting in lower-quality renderings of the inpainted 3D scene. However, due to its pretraining on large-scale image datasets, *SD2* achieves competitive FID scores, suggesting its ability to generate plausible visual content without fine-tuning. *MVInpainter*, originally designed for object-level and forward-facing task, struggles with general 3D inpainting scenarios, leading to unsatisfactory performance. *NeRFiller* and *GridPrior+DU* perform better on structural metrics due to their joint-view mechanism that incorporate cross-view priors. Nevertheless, these two approaches show limitations in perceptual quality, as reflected by higher FID scores.

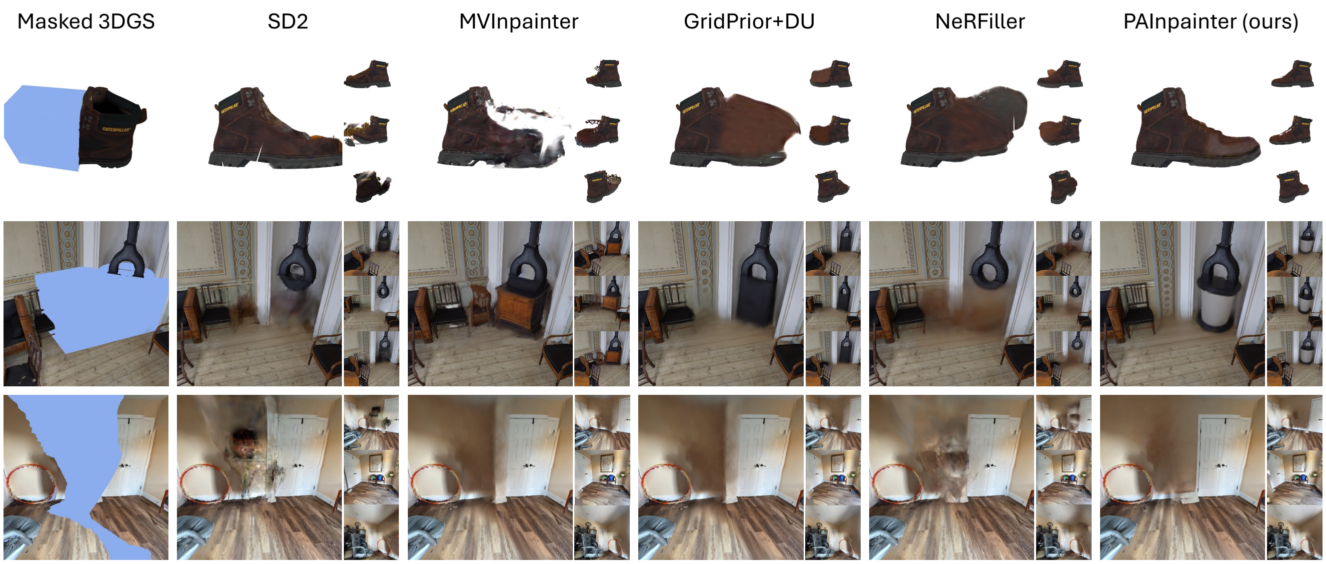

Figure 6. Qualitative comparison of 3D Gaussian inpainting results. Three scenes are shown (rows): Boot, Norway, and Office. For each scene, we present the initial masked 3D Gaussian scene (blue regions indicate missing content) and inpainting results from different methods. Four viewpoints are rendered to demonstrate multi-view consistency.

Graph sampling	Inpaint content propagation	Consistency verification	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	FID ↓
-	-	-	27.62	0.928	0.100	113.7
✓	-	-	27.94	0.929	0.091	109.4
✓	✓	-	<u>28.52</u>	0.932	<u>0.083</u>	<u>101.7</u>
✓	-	✓	28.47	<u>0.933</u>	0.085	106.0
✓	✓	✓	**29.51**	**0.935**	**0.081**	**96.1**

Table 2. Ablation study on NeRFiller dataset. Each row represents an ablated setting of our key components. The baseline uses basic iterative framework based on 3DGS without our proposed modules. Check marks (✓) indicate the presence of corresponding module. Results demonstrate PAInpainter (all modules present) achieves optimal performance.

Compared to previous methods, PAInpainter achieves improvement in both multi-view consistency and perceptual quality. This is particularly evident in its superior FID scores while maintaining leading performance in structural metrics, which indicates that the inpainted content generated by PAInpainter is not only coherent across views but also aligned with the original scene distribution, demonstrating its effectiveness and reliability in 3D inpainting.

Qualitative. Visualization comparisons are shown in Fig. 6, where PAInpainter exhibits remarkable performance in both detail preservation and multi-view consistency. In the "boot" scene (first row), our method accurately reconstructs the intricate leather textures while maintaining geometric continuity across different viewpoints. This advantage is further evidenced in the "Norway" scene (second row), where PAInpainter faithfully recovers the fine details of the mural paintings (beside the chair) with consistent artistic style and structural integrity. Similarly, in the "office" scene (third row), our method precisely reconstructs the architectural elements of the door frame while preserving the surrounding context. Additionally, as shown in the indoor scene in Fig. 7, PAInpainter showcases creativity ability by generating diverse content while maintaining scene consistency. In contrast, existing methods such as *GridPrior+DU* and *MVInpainter* , while achieving basic view consistency, often produce over-smoothed or distorted results, particularly in regions requiring high-fidelity detail. This comparison highlights PAInpainter's superior capability in enhancing both global structure coherence and local detail fidelity.

4.2. Ablation Study

We conduct ablation experiments to validate the effectiveness of each key component in PAInpainter, as shown in Tab. 2. The baseline method adopts a basic iterative framework for multi-view image inpainting without our proposed modules. When incorporating only the graph sampling strategy, the PSNR is improved to 27.94 dB, demonstrating that reference-guided inpainting effectively promotes view consistency. Adding either inpaint content propagation or consistency verification further improves performance (PSNR: 28.52 dB and 28.47 dB, respectively), indicating both modules contribute to high-quality inpainting.

The PAInpainter equipped with all components achieves the best performance (PSNR: 29.51 dB), showing a significant improvement of 1.89 dB over the baseline. This validates our design: graph sampling provides adjacent views, content propagation enhances inpainting consistency across views, and consistency verification nominates coherent re-

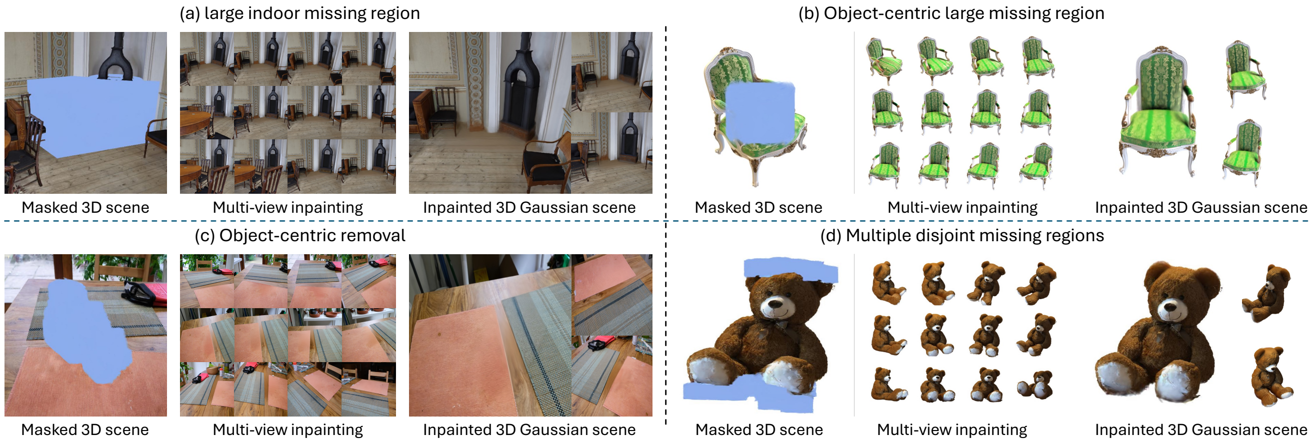

Figure 7. Visualization of PAInpainter on four representative inpainting scenarios. Each scenario shows the input, multi-view inpainting results, and the reconstructed 3D Gaussian scene, demonstrating consistent completion across varying viewpoints.

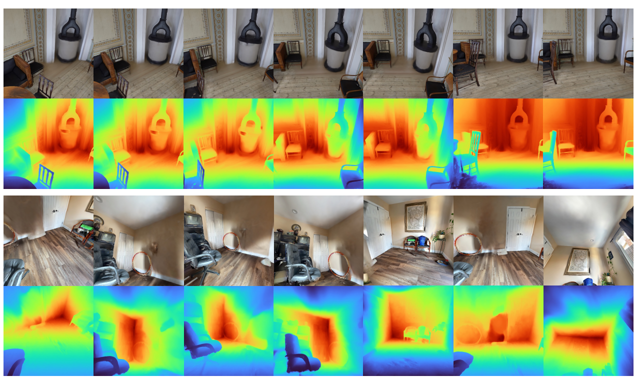

Figure 8. RGB and depth renderings of inpainted 3D Gaussian scenes from multiple viewpoints. The depth maps reveal consistent geometric reconstruction along with texture restoration.

sults. Notably, removing either propagation or verification module leads to similar performance degradation, suggesting these components are complementary in maintaining multi-view consistency while preserving texture details.

These ablation results corroborate our framework and previous experiments, confirming that the combination of all three proposed components is crucial for high-quality 3D Gaussian inpainting.

4.3. Versatility & Geometric Consistency

We showcase diverse 3D Gaussian inpainting scenarios of PAInpainter in Fig. 7, demonstrating its effectiveness across four distinct inpainting tasks. From object removal to large-area completion, our method consistently generates visually coherent results while preserving scene-specific geometric and textural details. The reconstructed 3D Gaussian scenes exhibit high fidelity across multiple viewpoints, validating the robustness of our approach in handling varying inpainting requirements.

The geometric consistency of our method is further validated through depth visualization, as shown in Fig. 8. Despite the absence of explicit depth supervision during 3D Gaussian optimization, the inpainted regions demonstrate naturalistic depth transitions and structural coherence with surrounding areas. The continuous depth maps demonstrate that strong multi-view consistency of PAInpainter inherently leads to accurate 3D geometry reconstruction, validating the capability in preserving both appearance and structural fidelity across different viewpoints.

5. Conclusion

In this paper, we present PAInpainter, an effective 3D Gaussian inpainter that substantially enhances multi-view consistency in 3D scene completion. Our technical contributions center on the novel perspective-aware inpainting framework, which integrates a perspective graph for adaptive view sampling, guided content propagation, and consistency verification mechanisms. Through this systematic design, PAInpainter preserves fine-grained scene details while ensuring multi-view consistency. Extensive experiments demonstrate that our method achieves superior performance across diverse inpainting scenarios.

The current PAInpainter implementation offers promising results while suggesting directions for future enhancements. The key modules in the proposed framework could be integrated into LDM in an end-to-end manner for further improved performance and deployment efficiency. Meanwhile, further exploration of sparse-view scenarios and unbounded outdoor scenes remains valuable for future work. Despite these considerations, PAInpainter demonstrates robust performance and practical utility for applications from stereo vision production to AR/VR development.

Acknowledgements

This research was supported by the Theme-based Research Scheme (TRS) project T45-701/22-R of the Research Grants Council (RGC), Hong Kong SAR. We thank all anonymous reviewers for their constructive feedback to improve our paper.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 4

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch. *ACM Transactions on Graphics*, page 1–11, 2009. 2

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH '00*, 2000. 2

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4, 5

[5] Alberto Cannavò, Francesco De Pace, Federico Salaroglio, and Fabrizio Lamberti. A visual editing tool supporting the production of 3d interactive graphics assets for public exhibitions. *International journal of human-computer studies*, 141:102450, 2020. 1

[6] Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu. Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7705–7715, 2024. 3

[7] Chenjie Cao, Chaohui Yu, Yanwei Fu, Fan Wang, and Xiangyang Xue. Mvinpainter: Learning multi-view consistent inpainting to bridge 2d and 3d editing. *arXiv preprint arXiv:2408.08000*, 2024. 1, 2, 3, 6

[8] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5868–5877, 2017. 2

[9] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B. Goldman, and Pradeep Sen. Image melding. *ACM Transactions on Graphics*, page 1–10, 2012. 2

[10] A.A. Efros and T.K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999. 2

[11] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. 6

[12] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 1, 2, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[15] Han Jiang, Haosen Sun, Ruoxuan Li, Chi-Keung Tang, and Yu-Wing Tai. Inpaint4dnerf: Promptable spatio-temporal nerf inpainting with generative diffusion models. *arXiv preprint arXiv:2401.00208*, 2023. 2, 3

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

[17] 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, 5

[18] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13169–13178, 2023. 2

[19] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 2

[20] Chieh Hubert Lin, Changil Kim, Jia-Bin Huang, Qinbo Li, Chih-Yao Ma, Johannes Kopf, Ming-Hsuan Yang, and Hung-Yu Tseng. Taming latent diffusion model for neural radiance field inpainting. *arXiv preprint arXiv:2404.09995*, 2024. 2

[21] Hao-Kang Liu, I Shen, Bing-Yu Chen, et al. Nerf-in: Free-form nerf inpainting with rgb-d priors. *arXiv preprint arXiv:2206.04901*, 2022. 2

[22] Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024. 1, 2, 3

[23] Yiren Lu, Jing Ma, and Yu Yin. View-consistent object removal in radiance fields. *arXiv preprint arXiv:2408.02100*, 2024. 2

[24] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 5, 6

[26] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. 2

[27] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, Jonathan Kelly, Marcus A. Brubaker, Igor Gilitschenski, and Alex Levinshtein. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *CVPR*, 2023. 3, 6

[28] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023. 2

[29] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10258–10268, 2024. 3

[30] Nikos Paragios, Yunmei Chen, and Olivier D Faugeras. *Handbook of mathematical models in computer vision*. Springer Science & Business Media, 2006. 2

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[32] Kira Prabhu, Jane Wu, Lynn Tsai, Peter Hedman, Dan B Goldman, Ben Poole, and Michael Broxton. Inpaint3d: 3d scene content generation using 2d inpainting diffusion. *arXiv preprint arXiv:2312.03869*, 2023. 1, 2, 3

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 5, 6

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[36] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 2

[37] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 4

[38] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[39] Dongqing Wang, Tong Zhang, Alaa Abboud, and Sabine Süsstrunk. Innerf360: Text-guided 3d-consistent object inpainting on 360-degree neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12677–12686, 2024. 2

[40] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 4

[41] 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

[42] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18120–18130, 2023. 2

[43] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[44] Ethan Weber, Aleksander Holynski, Varun Jampani, Saurabh Saxena, Noah Snavely, Abhishek Kar, and Angjoo Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20731–20741, 2024. 1, 2, 3, 6

[45] Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel J Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16528–16538, 2023. 2

[46] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning 3D Shape Priors for Shape Completion and Reconstruction. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[47] 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

[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contex-

tual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[49] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 2

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[52] Yuheng Zhao, Jinjing Jiang, Yi Chen, Richen Liu, Yalong Yang, Xiangyang Xue, and Siming Chen. Metaverse: Perspectives from graphics, interactions and visualization. *Visual Informatics*, 6(1):56–67, 2022. 1