MVIP-NeRF: Multi-view 3D Inpainting on NeRF Scenes via Diffusion Prior

Honghua Chen Chen Change Loy Xingang Pan
S-Lab, Nanyang Technological University
honghua.chen@ntu.edu.sg ccloy@ntu.edu.sg xingang.pan@ntu.edu.sg

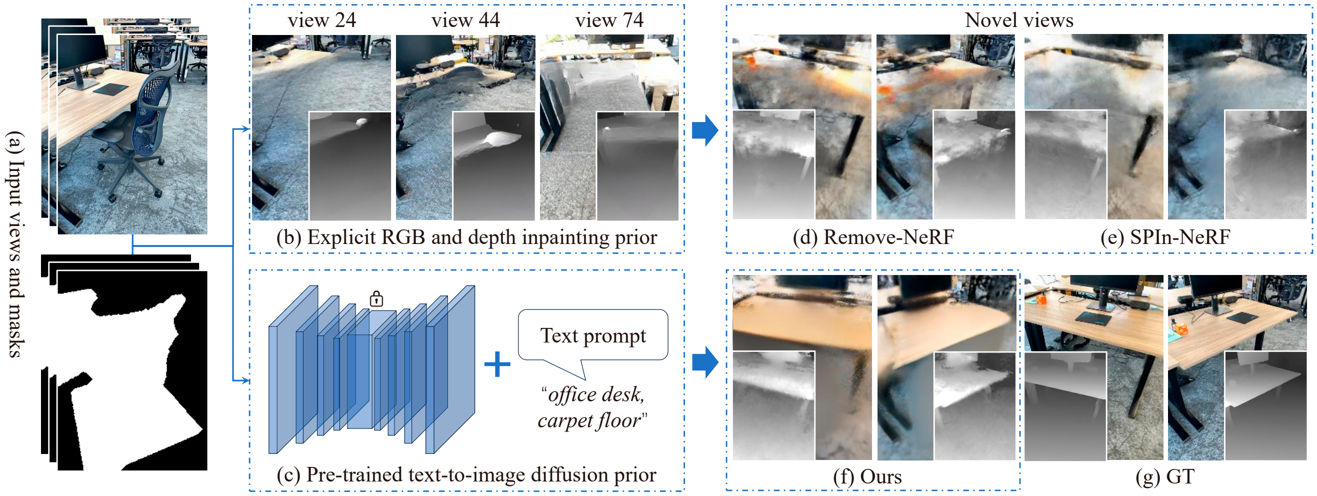
(a) Input views and masks
view 24
view 44
view 74
(b) Explicit RGB and depth inpainting prior
Text prompt
"office desk, carpet floor"
(c) Pre-trained text-to-image diffusion prior
Novel views
(d) Remove-NeRF
(e) SPIn-NeRF
(f) Ours
(g) GT

Figure 1. Comparison of our MVIP-NeRF with two state-of-the-art approaches, Remove-NeRF [35] and SPIn-NeRF [17]. Existing methods heavily depend on explicit RGB and depth inpainting results. This type of inpainting prior frequently shows inconsistency, inaccuracy, and misalignment to a certain degree (sub-figure (b)). In contrast, our approach implicitly exploits the diffusion prior (sub-figure (c)), resulting in more faithful and consistent results, in terms of both appearance and geometry.

Abstract

Despite the emergence of successful NeRF inpainting methods built upon explicit RGB and depth 2D inpainting supervisions, these methods are inherently constrained by the capabilities of their underlying 2D inpainters. This is due to two key reasons: (i) independently inpainting constituent images results in view-inconsistent imagery, and (ii) 2D inpainters struggle to ensure high-quality geometry completion and alignment with inpainted RGB images.

To overcome these limitations, we propose a novel approach called MVIP-NeRF that harnesses the potential of diffusion priors for NeRF inpainting, addressing both appearance and geometry aspects. MVIP-NeRF performs joint inpainting across multiple views to reach a consistent solution, which is achieved via an iterative optimization process based on Score Distillation Sampling (SDS). Apart from recovering the rendered RGB images, we also extract normal maps as a geometric representation and define a normal SDS loss that motivates accurate geometry inpainting and alignment with the appearance. Additionally, we formulate a multi-view SDS score function to distill generative priors simultaneously from different view images, ensuring consistent visual completion when dealing with large view variations. Our experimental results show better appearance and geometry recovery than previous NeRF inpainting methods.

1. Introduction

Neural Radiance Fields (NeRFs) [14] inpainting involves the removal of undesired regions from a 3D scene, with the objective of completing these regions in a contextually coherent, visually plausible, geometrically accurate, and consistent manner across multiple views. This form of 3D editing holds significant value for diverse applications, including 3D content creation and virtual/augmented reality.

Inpainting on NeRF scenes presents two intricate challenges: (i) how to ensure that the same region observed in multiple views is completed in a consistent way, especially when the view changes significantly; and (ii) inpainting must address not only the 2D appearance of NeRFs but also yield geometrically valid completion.

Several NeRF inpainting techniques have been developed to address specific aspects of these challenges [10, 16, 17, 35, 37, 38]. The majority of these approaches heavily rely on explicit RGB and depth inpainting priors, often employing 2D inpainters like LaMa [27] to independently inpaint all views and subsequently address the multiview inconsistency. For example, SPIn-NeRF [17] and InpaintNeRF360 [30] incorporate a perceptual loss within masked regions to account for low-level inconsistency, but the perceptual-level inconsistency still cannot be fully addressed (see from Fig. 1 (b)(e)). Another approach involves preventing inconsistent and incorrect views from being used in NeRF optimization. To achieve this, Weder et al. [35] introduce uncertainty variables to model the confidence of 2D inpainting results, facilitating automated view selection. As a simpler alternative, Mirzaei et al. [16] propose to use a single inpainted reference view to guide the entire scene inpainting process. However, this method is difficult to adapt to scenes with large view variations and requires non-trivial depth alignment. In summary, these methods remain constrained by the capabilities of underlying 2D inpainters. Besides, they share the common limitation of neglecting the correlation between inpainted RGB images and inpainted depth maps, resulting in less pleasing geometry completion.

In this work, we are interested in addressing these challenges via a new paradigm. Instead of employing 2D inpainting independently for each view, we believe that ideally, the inpainting at different views should work jointly to reach a solution that *i)* fulfills the 2D inpainting goal at each view and *ii)* ensures 3D consistency. Fortunately, 2D diffusion models present an ideal prior for achieving this goal. While recent advances like DreamFusion [19] have demonstrated their capability in 3D generation, the adaptation of diffusion priors to tackle the NeRF inpainting problem remains an untapped area.

To this end, we present *MVIP-NeRF*, a novel approach that performs multiview-consistent inpainting in NeRF scenes via diffusion priors. Given an RGB sequence and per-frame masks specifying the region to be removed, we train a NeRF using a reconstruction loss in the observed region and an inpainting loss in the masked region. The inpainting loss is based on the Score Distillation Sampling (SDS) [19] that attempts to align each rendered view with the text-conditioned diffusion prior. This approach allows our model to progressively fill the missing regions in the shared 3D space, thus the inpainting goal at multiple views can work jointly to reach a consistent 3D inpainting solution. To further ensure a valid and coherent geometry in the inpainted region, we also adopt diffusion priors to optimize the rendered normal maps. In addition, observing that the stochasticity of SDS often leads to a sub-optimal solution under large view variations, we formulate a multi-view score distillation, which ensures that each score distillation step takes into account multiple views that share the same SDS parameters. This achieves improved consistency and sharpness within the filled regions when the view changes significantly. We summarize our contributions as follows:

(i) A diffusion prior guided approach for high-quality NeRF inpainting, achieved without the need for explicit supervision of inpainted RGB images and depth maps.

(ii) An RGB and normal map co-filling scheme with iterative SDS losses that can simultaneously complete and align the appearance and geometry of NeRF scenes.

(iii) A multi-view score function to enhance collaborative knowledge distillation from diffusion models, avoiding detail blurring when dealing with large view variations.

(iv) Extensive experiments to show the effectiveness of our method over existing NeRF inpainting techniques.

2. Related Work

2.1. NeRF Inpainting

The use of NeRFs [14] for representing 3D scenes has enabled high-quality, photorealistic novel view synthesis. Despite this, only a limited number of studies have delved into the task of object removal or inpainting from pretrained NeRF models. Early approaches such as Edit-NeRF [11], Clip-NeRF [29] and LaTeRF [15], introduced methods to modify objects represented by NeRFs. However, the efficacy of these approaches is largely limited to simple objects rather than scenes featuring significant clutter and texture. Object-NeRF [36] supports the manipulation of multiple objects, like moving, rotating, and duplicating, but does not carefully handle the inpainting scenario. More closely, Instruct-NeRF2NeRF [4] proposes to use an image-conditioned diffusion model to facilitate text-instructed NeRF scene stylization.

NeRF-In [10], Remove-NeRF [35], SPIn-NeRF [17], and InpaintNeRF360 [30] are most closely related to our method. All of these approaches use RGB and depth priors from 2D image inpainters to inpaint NeRF scenes. The main difference among them is how they resolve view inconsistencies. Remove-NeRF [35] tackles inconsistencies by adaptively selecting views based on the confidence of the 2D inpainting results. Following it, a more straightforward scheme is to only use a single inpainted reference image to guide NeRF scene inpainting [16]. However, this method necessitates tedious and exact depth alignment. SPIn-NeRF [17] and InpaintNeRF360 [30] employ a perceptual loss within inpainted regions to account for the inconsistencies between different views. In contrast to these methods, we propose to inpaint NeRF scenes through an iterative optimization process that distills appearance and geometry knowledge from the pre-trained diffusion model. Consequently, our approach attains a more consistent and realistic representation of the entire NeRF scene, without requiring explicit RGB or depth supervision. It is worth noting that a recent work, RePaint-NeRF [38], also leverages a pre-trained diffusion model for NeRF painting. However, its primary focus is on object replacement within a NeRF scene. This task differs from ours in that it does not necessitate considering the coherence with the local context or ensuring high-quality geometry filling.

2.2. Diffusion Priors

Recently, we have witnessed remarkable advancements in the field of image generation, driven by the evolution of diffusion models [3, 5, 25, 26]. These models excel by progressively removing noise from Gaussian distributions using a UNet noise predictor, enabling the generation of high-quality images that align well with the training data. By training on large-scale text-image pairs [23], diffusion models have gained unprecedented success in text-to-image generation, with Stable Diffusion [21] as a phenomenal example. Therefore, many efforts have been made to explore the use of diffusion models as priors for a range of image restoration tasks such as super-resolution, colorization, inpainting, and deburring, *etc*. [6, 7, 9, 32, 33].

Beyond their use in 2D tasks, diffusion priors have also seen successful applications in 3D generation. A pioneering work in this direction is Dreamfusion [19], which leverages multiview 2D diffusion priors for 3D generation via an SDS loss, a concept derived from the distillation process of Imagen [22]. This approach gets rid of the need for large amount of 3D training data and thus has been widely adopted in subsequent text-to-3D synthesis endeavors such as MakeIt3D [28], Magic3D [8], Fantasia3D [1], and ProlificDreamer [34]. The SDS loss can not only synthesize objects but also edit existing ones, as studied in Latent-NeRF [13], Vox-E [24], and AvatarStudio [12]. Unlike these works that aim to create or edit 3D objects, our work is targeted to inpaint undesired regions to be coherent with the context for NeRF scenes.

3. Method

In this section, we provide a brief introduction to NeRF and SDS, followed by the formulation of our problem setting.

3.1. Preliminary

Neural Radiance Fields. NeRFs [14] encodes a 3D scene, by a function g that maps a 3D coordinate $\mathbf{p}$ and a viewing direction $\mathbf{d}$ into a color value $\mathbf{c}$ and a density value σ. The function g is a neural network parameterized by θ, so that $g_\theta : (\gamma(\mathbf{p}), \gamma(\mathbf{d})) \mapsto (\mathbf{c}, \sigma)$, where γ is a positional encoding. Each expected pixel color $\hat{C}(\mathbf{r})$ is rendered by casting a ray $\mathbf{r}$ with near and far bounds t_n and t_f. Typically, we divide $[t_n, t_f]$ into N sections $(t_1, t_2, ..., t_N)$ along a ray $\mathbf{r}$ and then compute the pixel color by $\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} \mathbf{c}_i^*$. The weighted color $\mathbf{c}_i^*$ of a 3D point is computed by $\mathbf{c}_i^* = w_i \mathbf{c}_i$, where $w_i = T_i(1 - \exp(-\sigma_i \delta_i))$, $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$, and $\delta_i = t_i - t_{i-1}$. Therefore, the NeRF reconstruction loss can be formulated as

$$\mathcal{L}^a = \sum_{\mathbf{r} \in R} ||\hat{C}(\mathbf{r}) - C(\mathbf{r})||^2, \tag{1}$$

where $\hat{C}(\mathbf{r})$ represents the rendered color blended from N samples, R is a batch of rays randomly sampled from the training views, and $C(\mathbf{r})$ corresponds to the ground-truth pixel color. If equipped with the ground-truth depth information, we can add another reconstruction loss to further optimize the geometry of NeRF scenes [2]:

$$\mathcal{L}^g = \sum_{\mathbf{r} \in R} ||\hat{D}(\mathbf{r}) - D(\mathbf{r})||^2, \tag{2}$$

where $\hat{D}(\mathbf{r})$ represents the rendered depth/disparity and $D(\mathbf{r})$ corresponds to the ground-truth pixel depth.

Score distillation sampling. SDS [19] enables the optimization of any differentiable image generator, *e.g*., NeRFs or the image space itself. Formally, let $\mathbf{x} = g(\theta)$ represent an image rendered by a differentiable generator g with parameter θ, then SDS minimizes density distillation loss [18] which is essentially the KL divergence between the posterior of $\mathbf{x} = g(\theta)$ and the text-conditional density p_ϕ^ω:

$$\mathcal{L}_{\texttt{Dist}}(\theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\big[w(t)\, D_{\text{KL}}\big(q(\mathbf{x}_t|\mathbf{x}) \,\|\, p_\phi^\omega(\mathbf{x}_t; y, t)\big)\big], \tag{3}$$

where $w(t)$ is a weighting function, y is the text embedding, and t is the noise level. For an efficient implementation, SDS updates the parameter θ by randomly choosing timesteps $t \sim \mathcal{U}(t_{\texttt{min}}, t_{\texttt{max}})$ and forward $\mathbf{x} = g(\theta)$ with noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to compute the gradient as:

$$\nabla_\theta \mathcal{L}_{\texttt{SDS}}(\theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)\big(\boldsymbol{\epsilon}_\phi^\omega(\mathbf{x}_t; y, t) - \boldsymbol{\epsilon}\big)\frac{\partial \mathbf{x}}{\partial \theta}\right]. \tag{4}$$

3.2. Problem formulation and overview

Given a set of RGB images, $\mathcal{I} = \{I_i\}_{i=1}^n$, with corresponding 3D poses $\mathcal{G} = \{G_i\}_{i=1}^n$, 2D masks $\mathcal{M} = \{m_i\}_{i=1}^n$, and a text description y, our goal is to produce a NeRF model for the scene. This NeRF model should have the capability to generate an *inpainted* image from any novel viewpoint. In general, we address unmasked and masked regions separately, following this general formulation:

$$\mathcal{L} = \mathcal{L}^a_{\text{unmasked}} + \lambda_1 \mathcal{L}^g_{\text{unmasked}} + \lambda_2 \mathcal{L}^a_{\text{masked}} + \lambda_3 \mathcal{L}^g_{\text{masked}}. \tag{5}$$

The entire process is visualized in Figure 2. Specifically, for unmasked regions, we utilize pixel-wise color (Eq. 1)

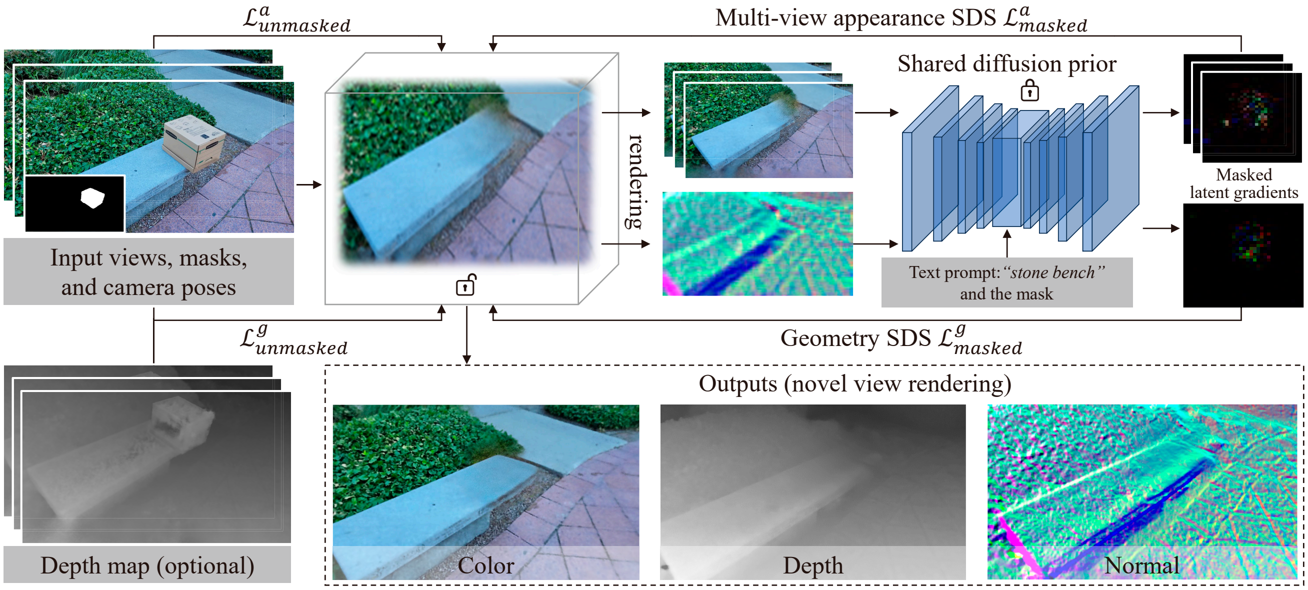
$\mathcal{L}^a_{unmasked}$
Multi-view appearance SDS $\mathcal{L}^a_{masked}$
Shared diffusion prior
rendering
Masked latent gradients
Input views, masks, and camera poses
Text prompt: *"stone bench"* and the mask
$\mathcal{L}^g_{unmasked}$
Geometry SDS $\mathcal{L}^g_{masked}$
Outputs (novel view rendering)
Depth map (optional)
Color
Depth
Normal

Figure 2. Method overview. Given posed RGB images with corresponding masks, depth maps (optional), and a text description, MVIP-NeRF can faithfully recover plausible textures and accurate surface detail. In the optimization process, for unmasked regions, we employ direct pixel-wise RGB and depth reconstruction losses. For masked regions, we introduce an RGB and normal map co-filling approach, utilizing SDS losses. This approach iteratively completes and aligns the appearance and geometry of NeRF scenes without the need for explicit supervision. Furthermore, we implement a multi-view scoring mechanism within the diffusion process to effectively handle significant variations in viewpoints. Finally, novel views can be rendered from the NeRF scene, where the object has been removed.

and depth (Eq. 2) reconstruction loss. For masked regions, we first render an RGB image and a normal map from the NeRF scene. Then, a latent diffusion model is employed as the appearance and geometry prior. Rather than directly utilizing inconsistent 2D inpainting results as supervisions and resolving these inconsistencies *post hoc*, we employ two SDS losses to compute a gradient direction iteratively for detailed and high-quality appearance and geometry completion. To further enhance consistency for large-view motion, we introduce a multi-view score function. This function ensures that multiple sampled views share the knowledge distilled from the diffusion models, thereby promoting cross-view consistency. Next, we will explain how to define $\mathcal{L}^a_{\text{masked}}$ and $\mathcal{L}^g_{\text{masked}}$, and how to extend these concepts to a multi-view version.

3.3. Appearance Diffusion Prior

We have noticed that independently inpainting individual images does not guarantee a consistent completion of the same region observed from multiple views. Sometimes, the inpainted results may even be incorrect. Therefore, instead of relying on explicit inpainting images, we incorporate a diffusion prior. We treat the inpainting task as a progressive denoising problem, which not only ensures view consistency but also enhances the visual realism of the completed scenes. To be more specific, we define the appearance SDS within the latent space of Stable Diffusion:

$$\nabla_\theta \mathcal{L}^a_{\text{masked}} = w(t)\big(\boldsymbol{\epsilon}^\omega_\phi(\mathbf{z}_t; m, y, t) - \boldsymbol{\epsilon}\big)\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \theta}, \tag{6}$$

where the noisy latent $\mathbf{z}_t$ is obtained from a novel view rendering $\mathbf{x}$ by Stable Diffusion encoder and m is the corresponding mask. It is important to note that we only backpropagate the gradient for the masked pixels. The range of timesteps $t_{\texttt{min}}$ and $t_{\texttt{max}}$ are chosen to sample from not too small or large noise levels and the text prompt y should align with the missing regions. In this work, we use the stable-diffusion-inpainting model [21] as our guidance model.

3.4. Geometry Diffusion Prior

In NeRF scenes, aside from appearance, achieving accurate geometry is a crucial component. Previous NeRF inpainting methods employ inpainted depth maps as an additional form of guidance for the NeRF model. However, we have observed that these inpainted depth maps often lead to visually unsatisfactory results and exhibit poor alignment with RGB images (see from Fig. 1 (b)). Consequently, this approach tends to be less effective in achieving high-quality geometry restoration in the masked areas.

In our work, we have two observations: (i) text-to-image diffusion models have a strong shape prior due to their training on diverse objects, and (ii) surface normals clearly reveal the geometric structures. Both observations encourage

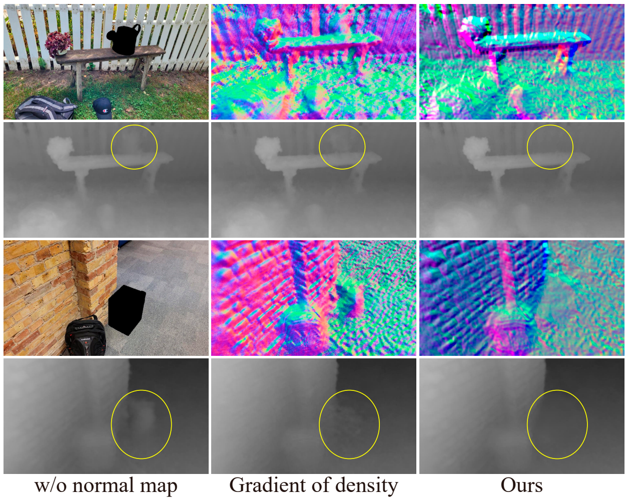

Figure 3. Effect of different normal map generation methods. In the first column, we present the input image with a mask (black region) and the depth map generated by NeRF, optimized with unmasked pixels. The second column displays the normal map derived from the density field gradient and the corresponding optimized depth map. The final column highlights the improved accuracy and reliability of geometry reconstruction achieved through the use of a smoothed normal field.

us to decouple the generation of geometry from appearance while further exploiting the potential geometry information from the diffusion prior. More specifically, considering the current NeRF function as $g(\theta)$, we generate a normal map $\mathbf{n}$ by rendering it from a randomly sampled camera pose. To update θ, we again employ the SDS loss that computes the gradient w.r.t. θ as:

$$\nabla_\theta \mathcal{L}^{g}_{\text{masked}} = w(t)\big(\boldsymbol{\epsilon}^{\omega}_{\phi}(\mathbf{z}_t; m, y, t) - \boldsymbol{\epsilon}\big)\frac{\partial \mathbf{z}}{\partial \mathbf{n}}\frac{\partial \mathbf{n}}{\partial \theta}. \tag{7}$$

Smoothed normal map generation. Surface normals are commonly derived by computing the gradient of the density field σ with respect to sampled positions. Nevertheless, the computed normals may exhibit some degree of noise, leading to an unclear geometric context. This, in turn, results in instability when generating geometry within the masked region, as demonstrated in Figure 3. Considering the readily available camera parameters and depth map, we introduce to calculate the smoothed surface normal from the depth map, treating it as a differentiable plane fitting problem.

Specifically, we denote (u_i, v_i) as the location of pixel i in the 2D image. Its corresponding location in 3D space is (x_i, y_i, z_i). We adopt the camera intrinsic matrix to compute (x_i, y_i, z_i) from its 2D coordinates (u_i, v_i), where z_i is the depth and given. Based on the assumption that pixels within a local neighborhood of pixel i lie on the same tangent plane, we then build the tangent plane to compute the surface normal of pixel i. In particular, we search the K ($K = 9$ in default) nearest neighbors in 3D space and calculate the surface normal estimate $\mathbf{n}$ based on these neighboring pixels on the tangent plane. The surface normal estimate $\mathbf{n}$ should satisfy the linear system of equations:

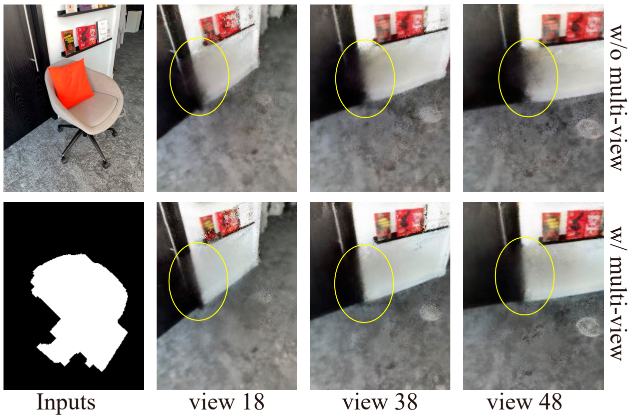

Figure 4. Effect of multi-view score distillation. The first row shows inpainting results without the multi-view score, while the second row shows the results with the multi-view score ($N = 5$).

$$\mathbf{A}\mathbf{n} = \mathbf{b}, \;\; \text{subject to} \;\; \|\mathbf{n}\|_2^2 = 1, \tag{8}$$

where $\mathbf{A} \in R^{K\times 3}$ is a matrix grouped by neighboring pixels and $\mathbf{b} \in R^{K\times 1}$ is a constant vector. Finally, we obtain the surface normal by minimizing $\|\mathbf{A}\mathbf{n} - \mathbf{b}\|^2$ whose least square solution has the closed form.

3.5. Multi-view Score Distillation

While our method consistently produces reliable results, it is worth noting that some blurring may occur in scenarios with large view variation, as shown in Fig. 4. The primary reason behind this may be that previous gradient updating, which relies on single-view information, does not adequately account for cross-view information. To this end, we define a multi-view distillation score function, which enhances the correlation in the recovery of each view. Given N viewpoints, we accordingly render N images, denoted as $\{\mathbf{x}^1, ..., \mathbf{x}^N\}$. Naturally, we can compute a multi-view score as:

$$\nabla_\theta \mathcal{L}^{ma}_{\text{masked}} = \sum_{i=1}^{N}\big(w(t)(\boldsymbol{\epsilon}^{\omega}_{\phi}(\mathbf{z}^i_t; m^i, y, t) - \boldsymbol{\epsilon}^i)\frac{\partial \mathbf{z}^i}{\partial \mathbf{x}^i}\frac{\partial \mathbf{x}^i}{\partial \theta}\big). \tag{9}$$

Note that the noise estimator and the noise level t are shared for all N images. Intuitively, this function implies that when updating θ, we take into account the interactions with other sampled views, thereby promoting view consistency.

Final loss function. Finally, we replace the $\mathcal{L}^{a}_{\text{masked}}$ with its multi-view version, and jointly train the loss as:

$$\mathcal{L} = \mathcal{L}^{a}_{\text{unmasked}} + \lambda_1 \mathcal{L}^{g}_{\text{unmasked}} + \lambda_2 \mathcal{L}^{ma}_{\text{masked}} + \lambda_3 \mathcal{L}^{g}_{\text{masked}}. \tag{10}$$

Note that we do not apply a multi-view version for the geometry diffusion prior. This is because, experimentally, we found that it contributes less while requiring more time cost.

4. Experiments

4.1. Experimental Setup

Datasets. We collect two real-world datasets for the experiments, called *Real-S* and *Real-L*. Next, we provide a detailed description of both datasets.

Real-S. This dataset comprises all 10 real-world scenes with slight viewpoint variations from [17]. In each scene, there are 60 training RGB images that include the unwanted object and the corresponding human-annotated masks. Additionally, 40 test images without the object are provided for quantitative evaluations. When assessing depth quality, we adhere to a standard scheme [31] of training a robust NeRF model exclusively on test views to produce pseudo-ground-truth depth maps. It is also important to note that both RGB and depth images have been resized to 1008×567, aligning with the dimensions used in [17].

Real-L. This dataset, sourced from [35], comprises 16 scenes with large viewpoint variations. They offer a wide range of diversity, encompassing differences in background texture, object size, and scene geometry. For each scene, two RGB-D sequences are available. One sequence includes the object, while the other does not, facilitating comprehensive evaluation and analysis. Since the depth maps are provided, we do not need to generate pseudo-ground-truth depth maps. Also, note that the RGB-D images have been resized to 192×256.

Metrics. To assess inpainting quality, we compare the output novel-view images generated by different approaches to the corresponding ground-truth images for each test image. Specifically, for *Real-S*, all metrics in the paper are exclusively calculated within the bounding boxes of masked regions, while for *Real-L*, due to the limited input resolution, we directly evaluate the full image. For appearance evaluation, we employ three standard metrics: PSNR, LPIPS, and FID. To assess geometric recovery, we compute the L_2 errors between the depth maps rendered by our system and the (pseudo) ground-truth depth maps. Observing that the masking scheme (whether the unmasked region is set to 0 or GT) and the LPIPS version (VGG or Alex) can affect the results, we provided more results in the supplementary file.

Parameters. We implemented our NeRF inpainting model built upon SPIn-NeRF [17] and trained it on 4 NVIDIA V100 GPUs for $10,000$ iterations using the Adam optimizer with a learning rate of 10^{-4}. As for the diffusion prior, the size of all latent inputs is set to 256×256. We set the range of timesteps as $t_{\texttt{min}} = 0.02$ and $t_{\texttt{max}} = 0.98$. In addition, we implement an annealing timestep scheduling strategy [39], which allocates more training steps to lower values of t. For the classifier-free guidance (CFG), we choose values within the range of $[7.5, 25]$ for $\mathcal{L}^{ma}_{\text{masked}}$ and $[2.5, 7.5]$ for $\mathcal{L}^{g}_{\text{masked}}$. The number of sampled views N is set to 5 in all cases. When rendering a single view, we select its four nearby views to calculate the multi-view score. Lastly, for the balance weights in Eq. 10, we empirically set $\lambda_1 = 0.1$ and $\lambda_2 = \lambda_3 = 0.0001$.

Training with high-resolution images. To enable training on high-resolution images, such as those with dimensions like 1008×567, we employ a separate batching scheme. In each iteration, we randomly select 1024 rays from the unmasked regions across all training views to reconstruct those areas. Given the context-sensitive nature of the text-to-image diffusion model, for the masked region, we select all rays within a single image that correspond to the masked regions. Following this, we combine the rendered colors from these rays with the unmasked pixels to create a complete image, which is subsequently fed into the diffusion model for further processing.

4.2. Results

Baselines. Considering the superior performance demonstrated by recent NeRF inpainting methods [17, 35] compared to traditional video and image-based inpainting pipelines, our focus is primarily on approaches that leverage the foundational NeRF representation. In total, we compare two NeRF inpainting approaches: SPIn-NeRF [17] with LaMa [27], and Remove-NeRF [35] with LaMa [27]. Although the two baselines have provided several evaluation results on their datasets, since they both require LaMa inpainting results, we re-executed their released code and reported the results accordingly to ensure a fair comparison. Also, it is noteworthy that we employ LaMa [27] as a 2D inpainter instead of Stable Diffusion. This choice is based on the former's superior quantitative performance [16, 21].

Quantitative inpainting results. We conducted a quantitative evaluation to assess the effectiveness of our method compared to the two baselines in terms of both appearance and geometry aspects. The results are reported in Table 1.

In detail, on the *Real-S* dataset where the view range is limited, all methods yield similar PSNR values. However, when considering the metric of LPIPS, which measures the perceptual quality and realism of the inpainted image, our method excels and demonstrates superiority over existing methods. When evaluating the *Real-L* dataset, which features significant view variations, we observed that Remove-NeRF performs better than SPIn-NeRF on the LPIPS metric. This difference in performance can be attributed to the presence of incorrect and inconsistent 2D inpainting results. In such cases, the view selection mechanism of

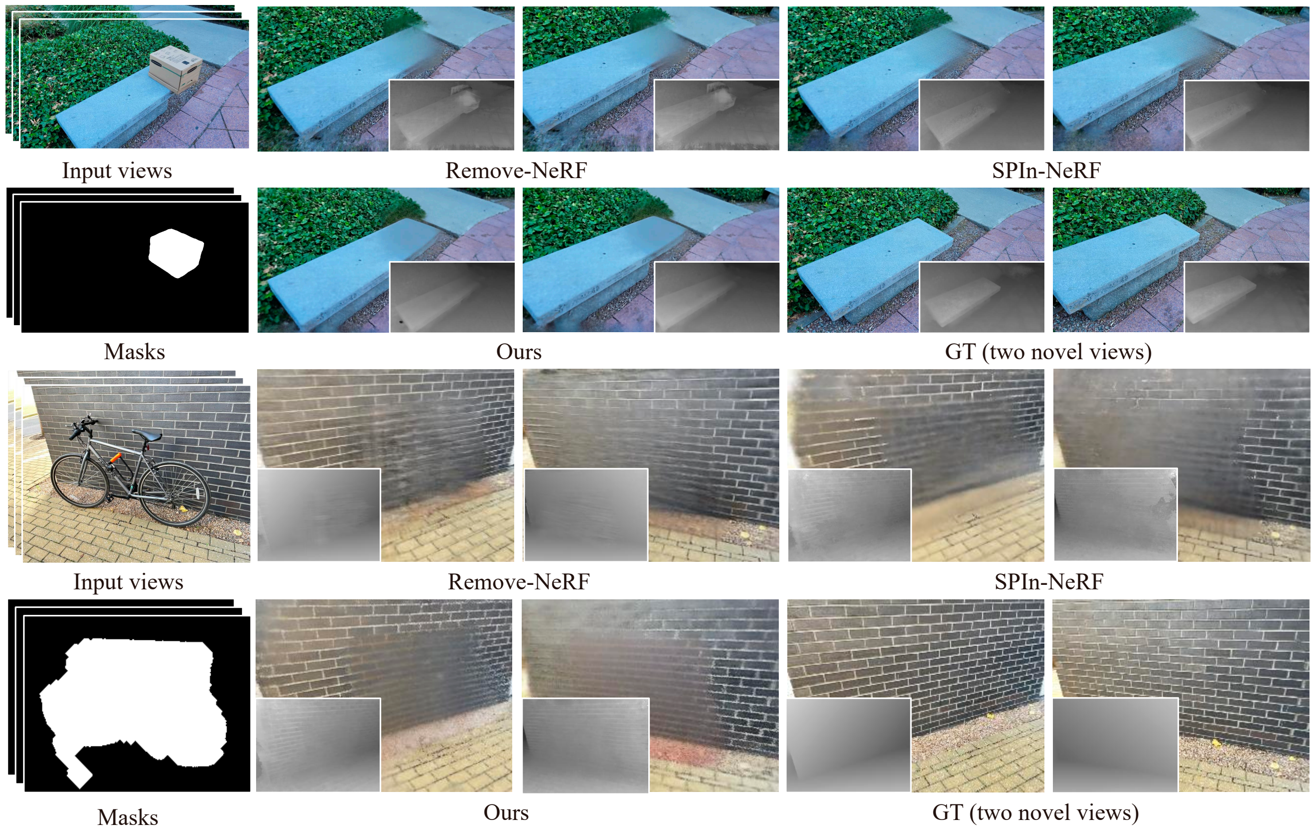

Figure 5. Visual comparison with two representative approaches [17, 35] on two scenes. The first scene is from the *Real-S* dataset with accurate masks, while the latter is from the *Real-L* dataset with large, roughly-covered masks. In the first scene, the input text prompt is *"A stone bench"* and for the second scene, it is *"A brick wall"*. Our method effectively handles both types of scenes, successfully generating view-consistent scenes with valid geometries (see the bench shape) and detailed textures (see the brick seam).

Table 1. **Comparison with state-of-the-art methods on two real-world datasets.** Our method is **best** compared to other novel-view synthesis baselines in inpainting the missing regions of the scene. Columns show the deviation from known ground-truth RGB images or depth maps of the scene (without the target object), based on the peak signal-to-noise ratio (PSNR), perceptual metric (LPIPS), feature-based statistical distance (FID), and pixel-wise L_2 depth errors.

	Real-S				*Real-L*			
	PSNR↑	LPIPS↓	FID↓	Depth L_2 ↓	PSNR↑	LPIPS↓	FID↓	Depth L_2 ↓
Remove-NeRF + LaMa [35]	17.556	0.665	254.345	8.748	25.176	0.187	**88.245**	0.038
SPIn-NeRF + LaMa [17]	17.466	0.574	**239.990**	1.534	25.403	0.215	103.573	0.090
Ours	**17.667**	**0.507**	255.514	**1.499**	**25.690**	**0.181**	100.452	**0.021**

Remove-NeRF can help avoid the incorporation of many incorrect views, resulting in improved performance. Our MVIP-NeRF achieves superior performance compared to other methods for several reasons. It not only avoids direct dependence on 2D inpainting but also leverages rich knowledge distilled from diffusion prior and shares multiple sampled views. This approach, in turn, enhances performance and contributes to its exceptional results.

Visual inpainting results. Apart from the quantitative comparisons, we also present visual comparisons. For each scene, we select two views with relatively large viewing angle deviations to showcase the respective inpainting results. Figure 5 illustrates the inpainting results for a scene from the *Real-S* dataset, characterized by highly accurate masks, in the first two rows. In contrast, the latter two rows depict results for a scene from the *Real-L* dataset, featuring large masks that roughly cover unwanted areas. Visual results demonstrate that our method effectively handles both

Table 2. **Ablation analysis.** Our method is **best** compared to different variants of our method in inpainting the missing regions of the scene. Columns show the deviation from known ground-truth RGB images or depth maps of the scene (without the target object). By ablating each component of our approach, we can observe a clear overview of their individual contributions.

	Real-S				*Real-L*			
	PSNR↑	LPIPS↓	FID↓	Depth L_2 ↓	PSNR↑	LPIPS↓	FID↓	Depth L_2 ↓
(i) (Masked NeRF)	17.121	0.761	353.766	3.089	24.189	0.222	137.960	0.040
(ii) (+appearance diffusion)	17.020	0.587	349.066	2.574	24.870	0.191	134.520	0.029
(iii) (+inpainted depth map)	17.028	0.565	355.002	1.869	24.206	0.254	138.256	0.044
(iv) (+geometry diffusion)	17.246	0.556	346.558	1.542	25.303	0.195	129.919	0.031
(v) (+multi-view)	**17.667**	**0.507**	**255.514**	**1.499**	**25.690**	**0.181**	**100.452**	**0.021**

types of scenes, successfully reconstructing view-consistent scenes with detailed textures (as evidenced by the visible brick seams in the second scene) and reasonable geometries (see the completed shape of the bench in the first scene). For more visual results, please refer to the supplemental file.

Ablation study. We first demonstrate the impact of ablating various components of our approach. We begin with training a Masked-NeRF (i) as a baseline and progressively incorporate the core modules: (i) Masked-NeRF, namely training a NeRF only using the unmasked pixels; (ii) introducing the appearance diffusion prior; (iii) using explicit depth inpainting results; (iv) replace the explicit depth inpainting prior with our geometry diffusion prior; and (v) employing a multi-view diffusion score function.

Table 2 provides a clear overview of the contribution of each module on the two datasets. Comparing (i) and (ii), the notable improvement in the LPIPS metric underscores the effectiveness of the appearance diffusion prior. However, it is worth noting that the PSNR metric may occasionally show less significant improvements due to the blurring effect, which can result in higher PSNR values. In addition, through a comparison of (ii), (iii), and (iv), we observed that when employing explicit depth inpainting results for NeRF reconstruction, the occasional inaccuracies in the inpainted depth lead to suboptimal geometry recovery. However, by leveraging the geometry diffusion prior, the reconstructed scenes exhibit enhanced geometry quality. Finally, by comparing (iv) and (v), the observed improvement further affirms the effectiveness of the multi-view score.

Results with CLIP prior. Interestingly, we found that many previous approaches use the CLIP loss [20] to supervise the alignment between synthesized views and the input text cues. To further exploit the potential prior and validate the effectiveness of our work, we replace the diffusion prior with the CLIP loss, which computes a feature loss between the inpainted image and the given text prompt. As illustrated in Fig. 6, we believe that the CLIP loss is relatively weak, making it challenging to recover the underlying appearance and geometry.

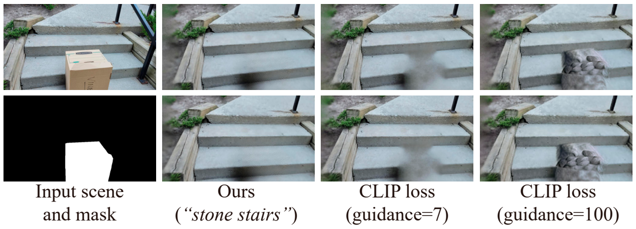

Figure 6. Comparison with CLIP guidances. The input text prompt is *"Stone Stairs"*. For each method, we show two novel view renderings. Our method can faithfully remove the unwanted object and recover the underlying structure.

5. Conclusion

In this work, we introduce *MVIP-NeRF*, a novel paradigm that harnesses the expressive power of diffusion models for multiview-consistent inpainting on NeRF scenes. Technically, to ensure a valid and coherent recovery of both appearance and geometry, we employ diffusion priors to co-optimize the rendered RGB images and normal maps. To handle scenes with large view variations, we propose a multi-view SDS score function, distilling generative priors from multiple views for consistent visual completion. We demonstrate the effectiveness of our approach over existing 3D inpainting methods and validate our key ideas by carefully crafting model variants. However, our work has several limitations: (i) the use of diffusion priors for iterative detail recovery affects efficiency, (ii) our method requires effort to tune hyper-parameters of diffusion priors, such as the CFGs, and (iii) as previous work [17, 35], our method cannot remove shadows.

Acknowledgements

This study is supported by NTU SUG-NAP and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20221-0011).

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 3

[2] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 3

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[4] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. 2

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[6] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 3

[7] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023. 3

[8] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[9] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 3

[10] 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

[11] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5773–5783, 2021. 2

[12] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, TEWARI A R MB, V GOLYANIK, A KORTYLEWSKI, and C THEOBALT. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM ToG (SIGGRAPH Asia)*, 1:16–18, 2023. 3

[13] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 3

[14] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 1, 2, 3

[15] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 2

[16] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. Reference-guided controllable inpainting of neural radiance fields. *arXiv preprint arXiv:2304.09677*, 2023. 2, 6

[17] 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. 1, 2, 6, 7, 8

[18] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR, 2018. 3

[19] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3

[20] 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. 8

[21] 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*, pages 10684–10695, 2022. 3, 4, 6

[22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 3

[24] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 430–440, 2023. 3

[25] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, page 2256–2265, 2015. 3

[26] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[27] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2, 6

[28] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 3

[29] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 2

[30] Dongqing Wang, Tong Zhang, Alaa Abboud, and Sabine Süsstrunk. Inpaintnerf360: Text-guided 3d inpainting on unbounded neural radiance fields. *arXiv preprint arXiv:2305.15094*, 2023. 2

[31] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv preprint arXiv:2303.16196*, 2023. 6

[32] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. In *arXiv preprint arXiv:2305.07015*, 2023. 3

[33] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023. 3

[34] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[35] 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. 1, 2, 6, 7, 8

[36] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021. 2

[37] Youtan Yin, Zhoujie Fu, Fan Yang, and Guosheng Lin. Or-nerf: Object removing from 3d scenes guided by multiview segmentation with neural radiance fields. *arXiv preprint arXiv:2305.10503*, 2023. 2

[38] Xingchen Zhou, Ying He, F Richard Yu, Jianqiang Li, and You Li. Repaint-nerf: Nerf editting via semantic masks and diffusion models. *arXiv preprint arXiv:2306.05668*, 2023. 2, 3

[39] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 6