Single-Mask Inpainting for Voxel-based Neural Radiance Fields

Jiafu Chen, Tianyi Chu, Jiakai Sun, Wei Xing*, and Lei Zhao*

Zhejiang University
{chenjiafu, chutianyi, csjk, wxing, cszhl}@zju.edu.cn

Abstract. 3D inpainting is a challenging task in computer vision and graphics that aims to remove objects and fill in missing regions with a visually coherent and complete representation of the background. A few methods have been proposed to address this problem, yielding notable results in inpainting. However, these methods haven't perfectly solved the limitation of relying on masks for each view. Obtaining masks for each view can be time-consuming and reduces quality, especially in scenarios with a large number of views or complex scenes. To address this limitation, we propose an innovative approach that eliminates the need for per-view masks and uses a single mask from a selected view. We focus on improving the quality of forward-facing scene inpainting. By unprojecting the single 2D mask into the NeRFs space, we define the regions that require inpainting in three dimensions. We introduce a two-step optimization process. Firstly, we utilize 2D inpainters to generate color and depth priors for the selected view. This provides a rough supervision for the area to be inpainted. Secondly, we incorporate a 2D diffusion model to enhance the quality of the inpainted regions, reducing distortions and elevating the overall visual fidelity. Through extensive experiments, we demonstrate the effectiveness of our single-mask inpainting framework. The results show that our approach successfully inpaints complex geometry and produces visually plausible and realistic outcomes.

1 Introduction

Neural radiance fields (NeRFs) [28] have emerged as an outstanding technique in 3D scene representation and novel view synthesis. By simply taking hundreds or even tens of images from a scene as input, NeRFs are able to capture the intricate detail and produce photorealistic renderings from new viewpoints. However, unwanted objects may appear when capturing the scene, such as litters on the floor and tourists in scenic spots. Thus, seamlessly removing objects and filling in the background—a task known as 3D inpainting—is essential in scene editing.

Though inpainting has been well-researched in 2D image processing, the study of 3D inpainting remains to be intractable. There are three challenges

* Corresponding Authors.

Fig. 1: Results of 3D inpainting by our method. Given a set of photographs (a) with a mask annotated on a selected view (b), our model is capable of removing objects, seamlessly filling in the background and generating inpainted novel views (c), which are visually plausible.

in 3D inpainting. Firstly, when no input view captures the area obstructed by the object to be removed, it remains uncertain how that area might appear. A reasonable geometry and appearance should be generated to ensure continuity and coherence with the surrounding scene. The inpainting result should not only align with the adjacent areas in terms of texture and color, but also adhere to the overall depth and lighting conditions to achieve a natural and undistorted look. Secondly, it is complicated to manually annotate the precise mask for each view. The complexity of the scene and the number of views can significantly increase the difficulty of mask annotation. Moreover, the manual effort can be time-consuming and impractical in real-world applications where efficiency and automation are desired. Thirdly, directly adopting state-of-the-art image inpainting methods to remove objects in rendered images from NeRFs will generate inconsistent results across different views, as shown in Fig. 2 (b). On the other hand, training a NeRF with inconsistent 2D inpainted images can lead to blurry results, as shown in Fig. 2 (c).

To address the aforementioned challenges, a number of works [22, 30, 47] have explored 3D inpainting for NeRFs. They identify and mask the object targeted for removal in each input view, then employ pre-trained 2D inpainting models to produce inpainted images. Following appropriate adjustments, the refined images are integrated to NeRFs to re-model a scene without the removed object. NeRF-In [22] uses a video object segmentation method to transfer the user-drawn mask from single view to other input views. On this basis, SPIn-NeRF [30] lifts the video segmentation masks into a coherent 3D segmentation via fitting a semantic NeRF, which resolves inconsistency and improve the masks. Remove-NeRF [47] generates a 3D point cloud representation of the scene and specifies a 3D bounding box enclosing the object to be removed in the point cloud. Afterwards the empty space of the 3D bounding box is trimmed and the masks are derived by rendering this marked space from each viewpoint. However,

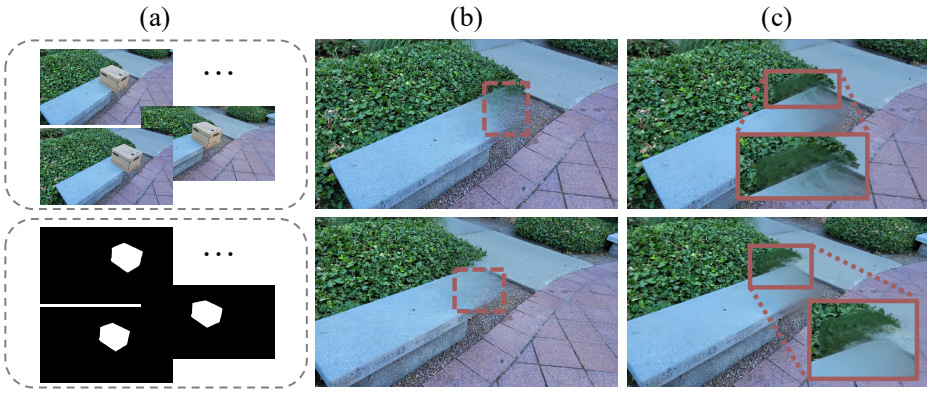

Fig. 2: Samples of challenging scenes. (a) Input views and corresponding masks. (b) Inconsistent results generated by 2D inpainting methods. (c) Blurry results when training a NeRFs model with inconsistent inpainted images.

the process of obtaining masks for all input views is cumbersome. Though users are not burdened with annotating every mask, obtaining all masks limits the scalability and efficiency of the inpainting process, making it less feasible for real-time or interactive applications. Thus, we tend to utilize the user-drawn single mask to directly remove the object and inpaint the scene.

In this work, we propose a novel 3D inpainting framework for forward-facing scenes that only relies on a single mask throughout the entire process, which not only simplifies the inpainting process but also avoids blurriness typically encountered in inpainted regions, a common issue when using inconsistent results from 2D inpainting models. The object targeted for removal is annotated on a randomly selected input view to create a mask. Since we focus on inpainting forward-facing scenes as previous works [22,30,47], single mask annotation does not lead to a significant loss of information compared to accurately annotating a mask for each view. An ordinary voxel-based NeRFs network is first trained to reconstruct the original scene, which contains a density voxel space, a feature voxel space and a shallow MLP mapping features to colors. Subsequently, as light rays pass through the mask into the scene from the camera, they intersect with the density and feature voxel space. This intersection enables us to unproject the mask onto the density and feature voxel space. The voxels within the mask range include both the object to be removed and the space requiring inpainting, thus pinpointing the voxels that need to be changed. By aligning the mask with the 3D scene representation in this manner, we can accurately identify, at voxel level, the areas that need inpainting. Having identified the voxels for modification, we initially focus on inpainting of the selected view, and then fine-tune from other views to enhance the realism and natural appearance of the entire inpainted scene. Specifically, we adopt a pre-trained 2D image inpainter [44] to generate reference inpainted color and depth images from the selected view, which are

used for regularizing appearance and geometry respectively. After convergence, the scene has been roughly inpainted from the perspective of the selected view. To ensure that the scene looks natural from different views, we fine-tune the inpainted area from other views. Given that large diffusion models, like *Stable Diffusion* [37], are trained on vast datasets of hundreds of millions of images and exhibit superior performance on open domain image generation tasks, we explore their potential for removing distorted areas in images. We leverage the optimized gradient from the denoising process of a pre-trained diffusion model to update the scene for better visual coherence and realism. Finally, we obtain a natural and undistorted inpainted scene without the removed object.

In summary, the main contribution of our work are as follows:

- We analyze the reasons for the blurriness in previous 3D inpainting methods and propose to address this issue by using a single reference to avoid inconsistency of 2D inpainted images.
- We propose a novel single-mask 3D inpainting approach for removing objects from 3D forward-facing scenes consistently and use a pre-trained and advanced large diffusion model to reduce distortions in inpainted regions, which tackles blurriness and makes the process efficient.
- Extensive experiments on different datasets are conducted to demonstrate the effectiveness of our method, demonstrating that our method surpasses state-of-the-art approaches in visual coherence and realism.

2 Related Work

2.1 Image Inpainting

Inpainting is a long-standing research topic in computer vision. Early works in this field focus on patch-based schemes [1, 40]. With the advent of deep learning, follow-up works turn to leverage neural networks. Pathak *et al*. [34] is a pioneering work in proposing a deep encoder-decoder architecture for image inpainting task. Since then, a series of subsequent works have been proposed to achieve better performance in many aspect, such as efficiency [38, 39], quality [10, 15, 21, 32, 53, 54], and diversity [19, 23, 57, 58]. We adopt LaMa [44] as our image inpainter, whichf introduces Fast Fourier Convolution to image inpainting for obtaining a large and effective receptive field. Yet, these image-based methods lack a mechanism for enforcing spatial consistency and do not inherently understand 3D scene structure. Consequently, they fall short in consistently inpainting multiple views of a scene, which is a critical requirement for our task.

2.2 NeRF Editing

In the past few years, rendering 3D scenes implicitly, especially NeRFs [28] has achieved incredibly high-quality results in scene reconstruction and novel view synthesis. Recent works have explored NeRFs for fast rendering [5, 12, 31, 42], improved visual quality [2–4], and sparse inputs [6, 11, 16, 18, 33, 43, 49, 52]. With

the rapid development, there are attempts [7, 8, 24, 26, 45, 48, 50, 55] aiming at editing on NeRFs, but they focus on non-inpainting tasks.

The first NeRF inpainting work is NeRF-In [22], which develops a framework to transfer a user-drawn mask to other views and model the scene with inpainted color and depth images. Later, SPIn-NeRF [30] constructs a 3D segmentation model to ensure the consistency of masks. To reduce the impact of supervising the scene with inconsistent inpainted images, SPIn-NeRF employs a perceptual loss instead of pixel loss in NeRF-In. Remove-NeRF [47] takes a different approach by marking the object to be removed within a point cloud and projecting this back to each view. It also introduces a view-selection mechanism to remove inconsistent views for optimization, thereby alleviating blurriness. RefIn-NeRF [29] uses a single inpainted 2D reference and provides controllability of inserting novel objects to into 3D scenes. Despite valuable efforts, they all necessitate extracting masks for all input views to perform inpainting, a requirement that is time-consuming and reduces inpainting quality.

2.3 NeRFs with 2D Diffusion Models

2D diffusion models are first introduced by Sohl-Dickstein *et al.* [41] and have emerged as new state-of-the-art deep generative models in image synthesis. The Latent Diffusion Models (LDM) [37] carry out diffusion processes in the latent space, effectively reducing computational costs. Leveraging the 2D diffusion models' ability to generate images of high visual quality, researches begin experimenting their application in supervising 3D generation. DreamFusion [36] proposes a method to directly predicts the update direction using a 2D diffusion model for optimizing NeRFs, which provides an efficient algorithm to bridge the gap of 2D diffusion models and 3D representation NeRFs. Follow-up works [20,25] focus on improving the quality of 3D generation. Some other works [46,59] utilize 2D diffusion models to edit the scene. However, instead of generating new objects, we aim to utilize the excellent performance of diffusion models to remove distortions in the scene. To the best our knowledge, we are the first to use 2D diffusion models to remove distortions in 3D scenes.

3 Proposed Method

We now illustrate our framework for inpainting a forward-facing 3D scene with a single mask. Given a collection of images from a scene with corresponding camera parameters, our goal is to remove objects from the scene according to the given mask of a selected view and fill in the missing part of the scene in a visually coherent and plausible manner. To achieve this, we propose a framework to unproject the single mask to scene representation space and preliminarily inpaint the scene through the supervision of inpainted reference RGB and depth provided by 2D inpainters. Even though the scene has been roughly inpainted from the selected view, novel views of the scene may appear distorted and unnatural. Therefore, we propose to utilize the powerful capability of 2D diffusion models

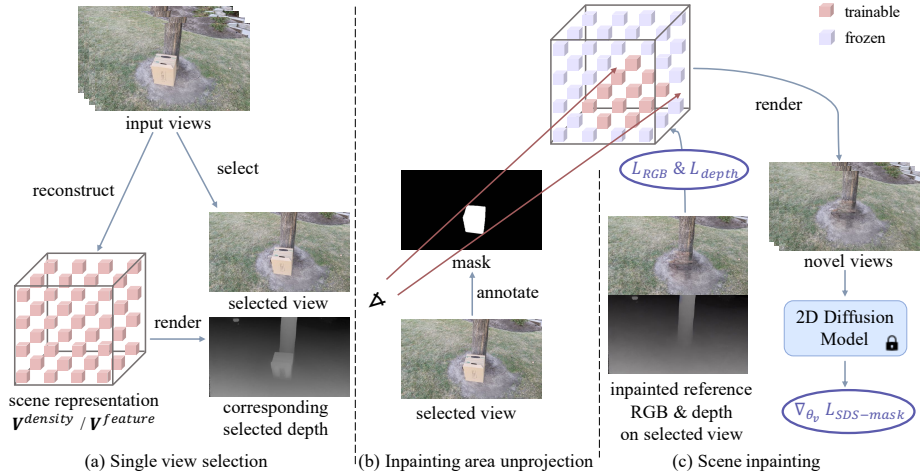

Fig. 3: An overview of our method. (a) We first use the input views to reconstruct the scene, and then randomly select one from the input views to render its depth image using the reconstructed model. (b) Then we annotate the removed object on the selected view and use 2D inpainters to obtain inpainted reference color and depth image. (c) We roughly inpaint the scene with L_{RGB} and L_{depth}. Later, we render novel views and input them into a 2D diffusion model to mitigate distortions using $\nabla_{\theta_v} L_{SDS-mask}$.

to remove distortions and fine-tune the scene for generating visually plausible and consistent results. In the next, we will first introduce some basic theories in Sec. 3.1, and then discuss the inpainting area unprojection in Sec. 3.2. Finally, we describe how the inpainting optimization process is carried out in Sec. 3.3.

3.1 Preliminary

Neural Radiance Fields. NeRFs [28] optimize a network to model a scene as continuous radiance fields, which takes 3D position $\mathbf{x}$ and viewing direction $\mathbf{d}$ as input and outputs volume density σ and color $\mathbf{c}$. To accelerate the process of training and testing, DVGO [42] uses a density voxel grid $\boldsymbol{V}^{density}$ to obtain σ and an intermediate feature voxel grid $\boldsymbol{V}^{feature}$ and a shallow MLP to obtain $\mathbf{c}$. Specifically, the process is as follows:

$$\begin{aligned} \sigma &= \log(1 + \exp(\text{interp}(\mathbf{x}, \boldsymbol{V}^{density}) + b)), \\ \mathbf{c} &= \text{MLP}(\text{interp}(\mathbf{x}, \boldsymbol{V}^{feature}), \mathbf{d}), \end{aligned} \tag{1}$$

where "interp" refers to trilinear interpolation in voxel grids, and the shift b is a hyperparameter.

To render the color of a pixel $\hat{C}(\mathbf{r})$, a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is cast from the camera center $\mathbf{o}$ along the direction $\mathbf{d}$ through the pixel. The volume rendering

process is integrating points on the ray:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t))dt, \\ \text{where } T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds), \tag{2}$$

where t_n and t_f represent the near and far bounds of the ray.

SDS Loss. In order to utilize 2D diffusion models ϕ to supervise 3D NeRFs models, Score Distillation Sampling (SDS) loss is generally used [20,25,36,46,59]. At an arbitrary view of NeRFs model θ, an image z could be rendered. The 2D diffusion model ϕ predicts the sampled noise as $\epsilon_\phi(z_t; y, t)$ given text embedding y and the noisy image z_t by adding noise ϵ at time-step t. SDS loss is calculated as a gradient, which is a probability density distillation loss and guides the update direction of NeRFs:

$$\nabla_\theta L_{SDS}(\phi, \theta) = \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(z_t; y, t) - \epsilon)\frac{\partial z}{\partial \theta}\right], \tag{3}$$

where θ denotes parameters of NeRFs model, ϕ denotes parameters of diffusion model, $w(t)$ is a weighted function from DDPM [14]. $\nabla_\theta L_{SDS}$ directly shows the update direction, thus the backpropagation doesn't go through the diffusion model.

3.2 2D-Mask to 3D-Area Unprojection

Unprojecting a 2D mask to the 3D area, especially when the scene is represented using voxel grids, requires understanding the spatial relationships between the camera, the 2D image, and the 3D scene. For each pixel in the mask, there are multiple points in the 3D space corresponding to it. As the exact depth of the object to be removed is unknown, we take all these points into consideration, that is the entire depth range. The unprojection involves determining the rays that pass through the corresponding pixels in the 2D image and intersect the voxel grid in the 3D space. By traversing these rays, the 2D mask M can be mapped to the relevant voxels $\{v_{mask}\}$ in the scene, indicating the areas to be masked or inpainted, as shown in Fig. 3(b).

An *extra initialization* is applied to voxels marked in the density voxel grid for better convergence of the inpainted NeRFs network. Instead of optimizing on the original NeRFs network, we take a different approach by initializing the marked voxels as free space in scene. This initialization step allows us to explicitly regard the inpainted area as an empty region and gradually learn to build plausible structure. We first filter out the free space of the original NeRFs network and randomly choose a voxel from it. The chosen voxel contains a density voxel value $v_{density}$ and a corresponding hyperparameter $\widetilde{b}$, which are both used as the initial values of marked voxel in the density voxel grid. In this way, the converged geometry can better fit the inpainted reference depth image.

3.3 Scene Inpainting

Rough Inpainting on Reference View Direct training of a 3D inpainter is difficult due to a lack of prior knowledge and data on the scene. Thus, we leverage 2D single image inpainters to obtain image priors instead. Specifically, we use LaMa [44] to help with image inpainting in our method. It should be noted that LaMa is a representative image inpainting method, which may be replaced by other advanced methods.

Given a selected input image I_s and its corresponding annotated mask M, an inpainted reference color image $\hat{I}_s$ can be obtained: $\hat{I}_s = \text{LaMa}(I_s, M)$. With the inpainted color image, the NeRFs network can be optimized by minimizing the L_2 distance between the inpainted pixel $\hat{C}(\mathbf{r})$ and the rendered pixel $C(\mathbf{r})$:

$$L_{RGB} = \sum_{\mathbf{r}\in\mathcal{R}} \parallel C(\mathbf{r}) - \hat{C}(\mathbf{r}) \parallel_2, \tag{4}$$

where $\mathcal{R}$ is a ray batch from the inpainted region of $\hat{I}_s$.

With only the inpainted reference color image, only the appearance of the object is changed to fit $\hat{I}_s$ when an image is captured from the selected view. The geometry may be corrupted in the marked region. Thus, we use an inpainted depth image as an additional guidance for the NeRFs network. The original depth image D_s for inpainting is rendered from the NeRFs network for scene reconstruction under the selected view by substituting distance t for color $\mathbf{c}$ in Eq. 2:

$$D(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))t dt. \tag{5}$$

Similarly, we obtain an inpainted reference depth image $\hat{D}_s$: $\hat{D}_s = \text{LaMa}(D_s, M)$. The inpainted reference depth image is used to remove the object from depth and plausibly inpaint the scene geometry. The NeRFs network is optimized via:

$$L_{depth} = \sum_{\mathbf{r}\in\mathcal{R}} \parallel D(\mathbf{r}) - \hat{D}(\mathbf{r}) \parallel_2, \tag{6}$$

where $\mathcal{R}$ is a ray batch from the inpainted region of $\hat{D}_s$. Particularly, the gradients of the density voxels are detached in L_{RGB} to ensure that only L_{depth} interferes with the geometry.

Refinement on All Input Views With the prior knowledge from a 2D image inpainting model, we performed initial 3D inpainting from the selected view for the scene. However, we found that overfitting of appearance on a single view often resulted in visually unreasonable effects, *e.g.*, artifacts or distortions, from other views. In order to mitigate the distortions, we leverage the powerful capability of 2D diffusion models. Trained on hundreds of millions of data, they have learned about the distribution of realistic images, including structure and detailed textures. Unlike [20,25,36] that utilize 2D diffusion models to synthesis 3D objects and simple scenes, we take advantage of 2D diffusion models to

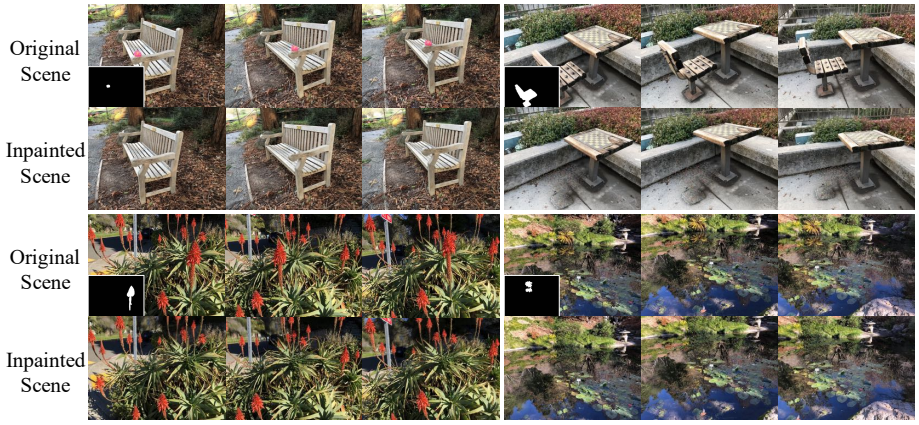

Fig. 4: Visualization of our inpainting results. Upper rows per inset show NeRFs renderings of the original scene from novel views, with left lower corner of the first image displaying the annotated mask. Lower rows show the corresponding inpainted view.

refine distorted areas in the mask regions of the scene. We use the open-source Stable Diffusion model [37] in our work, which requires a text prompt as input. To ensure that the optimization results of the diffusion model from multiple views is semantically consistent, we use the inpainted reference image $\hat{I}_s$ as a condition to assist in refining the distorted areas. To reduce the burden of getting additional text prompt input by obtaining the exact text embedding matching $\hat{I}_s$, we follow [9] to optimize the text embedding e_t in each timestamp t:

$$\min_{e_t} \parallel \bar{z}_0 - \hat{z}_0(\bar{z}_t, e_t) \parallel_2^2, \tag{7}$$

where $\bar{z}_0$ is the encoded latent of $\hat{I}_s$, and $\hat{z}_0(\bar{z}_t, e_t)$ refers to the estimated latent $\hat{z}_0$ given $\bar{z}_t$ and e_t. For every t, the optimization starts from the endpoint of the previous step $t+1$ optimization till the optimization ends at timestep 0.

From each input view, we render an image I_r and propose a mask-aware SDS loss $\nabla_{\theta_v} L_{SDS-mask}$ that restricts the loss to voxels in the mask area $\{v_{mask}\}$ and utilizies $\nabla_\theta L_{SDS}$ in Eq. 3 by replacing the text embedding y with e:

$$\nabla_{\theta_v} L_{SDS-mask} = \begin{cases} \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(z_t; e, t) - \epsilon)\frac{\partial z}{\partial \theta_v}\right], & v \in \{v_{mask}\}; \\ \texttt{STOP GRADIENT}, & v \notin \{v_{mask}\}; \end{cases} \tag{8}$$

where z_t refers to the result of adding noise ϵ at time-step t to the encoded latent of I_r. With $\nabla_{\theta_v} L_{SDS-mask}$, we refine the roughly inpainted scene from all input views and eliminate the distortions in the mask areas of the scene.

4 Experiment

4.1 Implementation Details

Our voxel-based NeRFs network is built upon DVGO [42]. We only optimize the density voxel space and feature voxel space, and keep the shallow MLP frozen. Following [42], we use the Adam optimizer with a learning rate of 0.1 for voxels marked by inpainted area unprojection. We carry out rough inpainting on reference view for 500 epochs and refine the scene for 100 more epochs. All experiments are performed on a single NVIDIA RTX A6000 (48G) GPU. Specially, the annotated mask is slightly dilated using two iterations with a 5×5 kernel to ensure the complete coverage of the removed object.

Datasets. Following SPIn-NeRF [30], we focus on forward-facing scenes. We utilize scenes provided by LLFF [27] and SPIn-NeRF [30]. All of them are captured using handheld cameras in real-scenes.

Baselines. We compare our approach with four models:

- Object-NeRF [51]: a NeRFs-based method for object manipulation that directly removes points masked in 3D without background filling with inpainters.
- Masked NeRFs: a NeRFs model trained exclusively on unmasked pixels, while masked pixels are disregarded, relying on the NeRFs model itself to interpolate plausible reconstructions for the masked regions.
- LaMa [44] + NeRFs: a NeRFs model trained on images inpainted by LaMa.
- SPIn-NeRF [30]: a state-of-the-art method designed especially for 3D inpainting tasks, which implements both multi-view consistent segmentation for the object to be removed and multi-view consistent inpainting. We use their results by running their open-source code in default setting.

4.2 Quantitative Results

We conduct quantitative comparisons on SPIn-NeRF dataset [30], which contains ground-truth captures of scenes without the removed object. Considering the complex and ambiguous nature of the task, we follow both 2D [44] and 3D [30] inpainting researches to assess the perceptual quality and realism of our inpainted scenes.

We report the average learned perceptual image patch similarity (LPIPS) [56] and the average Frechet inception distance (FID) [13] as evaluation metrics between the distribution of the ground-truth test views and the model outputs. Given our specific focus on inpainting, we calculate the LPIPS and FID metrics only within the bounding box of the object mask. We expand each side of the bounding box containing the mask in every direction by 10% following [30]. LPIPS provides a quantitative assessment of the visual similarity between the ground-truth test views and the inpainted regions, and FID captures the similarity between two distributions of images based on features extracted from a pretrained Inception network. The second and third column of Tab. 1 show that

Method	LPIPS ↓	FID ↓	MUSIQ ↑	Sharpness ↑
Object-NeRF	0.326	304.21	22.72	233.47
Masked NeRF	0.278	321.97	25.31	257.83
LaMa + NeRF	0.221	253.25	30.09	<u>316.28</u>
SPIn-NeRF	<u>0.187</u>	<u>204.26</u>	<u>56.95</u>	294.92
Ours	**0.186**	**168.70**	**64.20**	**585.05**

Table 1: Quantitative comparisons on LPIPS and FID with ground truth images provided and MUSIQ and sharpness on the visual quality. The reported results are average values on SPIn-NeRF dataset. **Best** and <u>second best</u> results are marked.

methods designed especially for 3D inpainting tasks (*i.e.*, SPIn-NeRF and ours) outperforms others. These specialized methods are tailored to address the unique challenges and requirements of inpainting unknown content in 3D scenes, leading to superior performance and results. While SPIn-NeRF trains an additional 3D segmentation model to obtain view-consistent masks for all input views, our method directly uses the single mask for inpainting, yet still achieves comparable results. Our method provides a more efficient and streamlined approach to 3D inpainting compared to SPIn-NeRF. Please refer to supplementary material for details on computational complexity.

In addition, to further assess the quality of inpainted scenes, we adopt two additional quantitative metrics, MUSIQ [17] and sharpness to evaluate the rendered image from a video computed by using a camera in a spiraling pattern. A classical measure of sharpness is the variance of the image Laplacian [35]. MUSIQ is meant to reproduce human perceptual judgments. As shown in the forth and fifth column of Tab. 1, our method is superior in both sharpness and MUSIQ, demonstrating our results are more realistic.

4.3 Qualitative Results

In Fig. 4, we show our inpainting results on different scenes. It can be seen that the annotated object is seamlessly removed and the background is plausibly filled in. The inpainted regions align well with the scene's geometry and maintain the overall visual appearance of the scene. It is demonstrated that our method can achieve scene inpainting that is visually coherent and contextually consistent, showcasing the effectiveness of our approach.

In addition, we compare our inpainting results with state-of-the-art 3D inpainting method SPIn-NeRF [30] in Fig. 5. SPIn-NeRF obtains masks for each input view and uses 2D image inpainter to generate the inpainted results. To eliminate the blurriness brought by training with inconsistent 2D inpainted images, SPIn-NeRF proposes to use a perceptual loss rather than mean square error to optimize the masked area. However, the effect is limited. From the zoom-in part in Fig. 5, we observe that the inpainted regions still exhibit some blurriness and lack fine details. The limited effectiveness of the perceptual loss in addressing blurriness and lack of fine details is primarily due to the inherent challenges

Fig. 5: Qualitative comparisons with state-of-the-art baseline SPIn-NeRF. SPIn-NeRF struggles to capture fine details due to optimizing based on inconsistent inpainted images, while our method use a single reference inpainted image.

in reconstructing high-frequency information from incomplete or inpainted regions. The perceptual loss, which leverages pre-trained deep neural networks to capture high-level visual features, can help maintain global structure and texture consistency, but it may struggle to capture fine details at a pixel level.

In comparison, our method uses a single mask and roughly inpaints the masked region from single view, which avoids blurriness caused by using inconsistent 2D inpainted images from different views. As our method focuses on capturing the visual appearance and context specific to a particular viewpoint, we can generally maintain the sharpness of the reference inpainted image. Furthermore, with the help of 2D diffusion model, we can not only remove distorted area in other views, but also add more reasonable details to the scene. Meanwhile, the use of a single mask also provides a simplified and more efficient inpainting process. Instead of obtaining and dealing with multiple masks and integrating inpainted regions from different viewpoints, our method streamlines the workflow by inpainting the masked region at once. This reduces the complexity that arises when combining inpainted regions from different views.

4.4 Ablation Study

The impact of L_{depth}. As discussed in Sec. 3.3, we introduce a depth loss L_{depth} to remove the object geometrically and fill in plausible geometry. In Fig. 6, we show inpainting results of our method with and without L_{depth}. We find that our full model exhibits a more accurate estimation of the underlying scene geometry, successfully removing the object's geometry and generating a smooth

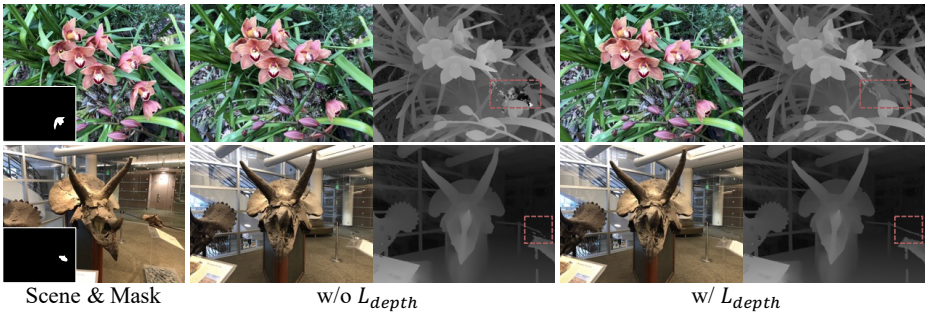

Fig. 6: Ablation study on the impact of L_{depth}. Using depth priors helps align the inpainted area seamlessly with the surrounding scene in geometry.

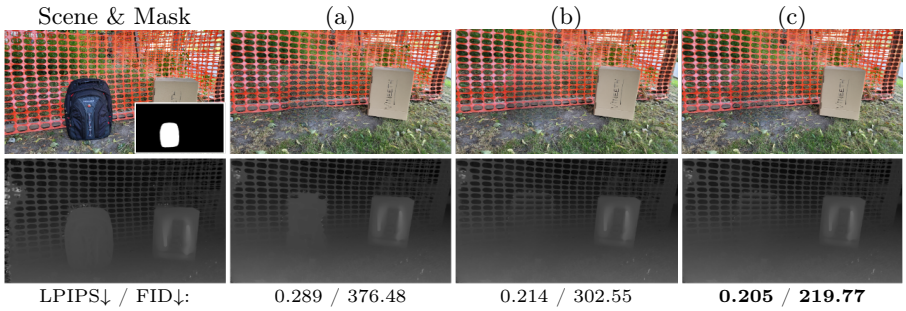

Fig. 7: Ablations study on the impact of depth initialization and $\nabla_{\theta_v} L_{SDS-mask}$. (a) Using the original scene as initialization. The results may optimize to a suboptimal scene's geometry. (b) The results w/o $\nabla_{\theta_v} L_{SDS-mask}$. The FID score slightly drops. (c) Our full model with both initializing the marked regions as empty space and $\nabla_{\theta_v} L_{SDS-mask}$.

transition between the inpainted area and the surrounding scene. This results in visually plausible and coherent inpainting outcomes that blend seamlessly with the rest of the scene. Conversely, the absence of L_{depth} prevents the model from predicting convincing geometry within the masked region, since it is difficult to learn geometry from the single reference inpainted color image.

The impact of depth initialization. The initialization of depth has a significant impact on the inpainting process. Here we verify the effectiveness of initializing the marked geometry as free space before optimization rather than optimizing based on the original geometry. As shown in Fig. 7 (a), we can observe that optimizing based on the original geometry struggles to convergence and can result in suboptimal inpainting results. However, by considering the marked regions as empty space, the NeRFs model is not constrained by the limitations of the initial geometry and can adaptively convergence to the reference geometry provided by inpainted depth image. It is important to note that the effectiveness of treating the marked geometry as free space may depend on the specific char-

acteristics of the scene and the complexity of the masked regions. In some cases, optimizing based on the original geometry may still yield satisfactory results, especially when the masked regions have relatively simple or regular geometry.

The impact of $\nabla_{\theta_v} L_{SDS-mask}$. To mitigate distortions of the scene that roughly inpainted with the supervision of the selected view, we incorporate the capability of 2D diffusion model, which helps refine the inpainting results by considering the plausibility of the rendered image. In Fig. 7 (b) and (c), we present a comparison of the inpainting results obtained with and without $\nabla_{\theta_v} L_{SDS-mask}$. We also present the quantitative scores, which validates that $\nabla_{\theta_v} L_{SDS-mask}$ helps to improve visual quality of inpainted scenes and results in visually pleasing and more natural-looking inpainted scenes.

5 Limitation

Due to the lack of depth supervision during scene reconstruction, the accuracy of the rendered depth image may be compromised. Therefore, using this depth image as input for the image inpainting model and then using the output as depth guidance for 3D inpainting could lead to errors. The inaccurate depth information may lead to unsatisfactory inpainting results in the following process. Also, if the randomly selected view contains only a part of the object to be removed, that is, the object is not fully captured within the selected perspective's range, annotating a mask on this view would evidently lead to an incomplete removal of the object.

Currently, our method can only use in forward-facing scenes. We will explore to expand our work to 360° scene in the future. As LaMa output deterministic image inpainting results, we are not able to achieve controllable scene inpainting.

6 Conclusion

In this work, we present an efficient and streamlined framework for 3D inpainting using merely a single mask. Our framework unprojects the 2D mask to voxel-based NeRFs space and only carry out inpainting within the masked regions. We leverage both image and geometry priors to roughly inpaint scenes from the selected view. Refinement is achieved by using 2D diffusion models to implicitly remove the unnatural and distorted areas when observed from other views. Extensive experimental results demonstrate the effectiveness of our method on forward-facing scenes and shows the strength of our approach against state-of-the-art in terms of visual quality and sharpness.

Acknowledgements

This work was supported in part by Zhejiang Province Program (2023C03199, 2022C01222, 2023C03201), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), Ningbo Science and Technology Plan Project

(022Z167, 2023Z137), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. ICCV (2023)
5. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision. pp. 333–350. Springer (2022)
6. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
7. Chen, J., Xing, W., Sun, J., Chu, T., Huang, Y., Ji, B., Zhao, L., Lin, H., Chen, H., Wang, Z.: Pnesm: Arbitrary 3d scene stylization via prompt-based neural style mapping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1091–1099 (2024)
8. Chen, Y., Chen, Z., Zhang, C., Wang, F., Yang, X., Wang, Y., Cai, Z., Yang, L., Liu, H., Lin, G.: Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21476–21485 (2024)
9. Cheng, B., Liu, Z., Peng, Y., Lin, Y.: General image-to-image translation with one-shot image guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22736–22746 (2023)
10. Chu, T., Chen, J., Sun, J., Lian, S., Wang, Z., Zuo, Z., Zhao, L., Xing, W., Lu, D.: Rethinking fast fourier convolution in image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23195–23205 (2023)
11. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
12. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

15. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36**(4), 1–14 (2017)
16. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5885–5894 (2021)
17. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
18. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12912–12921 (2022)
19. Li, A., Zhao, L., Zuo, Z., Wang, Z., Xing, W., Lu, D.: Migt: Multi-modal image inpainting guided with text. Neurocomputing **520**, 376–385 (2023)
20. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
21. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European conference on computer vision (ECCV). pp. 85–100 (2018)
22. Liu, H.K., Shen, I., Chen, B.Y., et al.: Nerf-in: Free-form nerf inpainting with rgb-d priors. arXiv preprint arXiv:2206.04901 (2022)
23. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9371–9381 (2021)
24. Liu, S., Zhang, X., Zhang, Z., Zhang, R., Zhu, J.Y., Russell, B.: Editing conditional radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5773–5783 (2021)
25. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023)
26. Mikaeili, A., Perel, O., Safaee, M., Cohen-Or, D., Mahdavi-Amiri, A.: Sked: Sketch-guided text-based 3d editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14607–14619 (2023)
27. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
29. Mirzaei, A., Aumentado-Armstrong, T., Brubaker, M.A., Kelly, J., Levinshtein, A., Derpanis, K.G., Gilitschenski, I.: Reference-guided controllable inpainting of neural radiance fields. In: ICCV (2023)
30. Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K.G., Kelly, J., Brubaker, M.A., Gilitschenski, I., Levinshtein, A.: Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20669–20679 (2023)

31. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
32. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
33. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
34. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
35. Pertuz, S., Puig, D., Garcia, M.A.: Analysis of focus measure operators for shape-from-focus. Pattern Recognition **46**(5), 1415–1432 (2013)
36. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2022)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
38. Sagong, M.c., Shin, Y.g., Kim, S.w., Park, S., Ko, S.j.: Pepsi: Fast image inpainting with parallel decoding network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11360–11368 (2019)
39. Shin, Y.G., Sagong, M.C., Yeo, Y.J., Kim, S.W., Ko, S.J.: Pepsi++: Fast and lightweight network for image inpainting. IEEE transactions on neural networks and learning systems **32**(1), 252–265 (2020)
40. Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: 2008 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2008)
41. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
42. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2022)
43. Sun, J., Zhang, Z., Chen, J., Li, G., Ji, B., Zhao, L., Xing, W., Lin, H.: Vgos: Voxel grid optimization for view synthesis from sparse inputs. arXiv preprint arXiv:2304.13386 (2023)
44. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2149–2159 (2022)
45. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3835–3844 (2022)
46. Wang, D., Zhang, T., Abboud, A., Süsstrunk, S.: Inpaintnerf360: Text-guided 3d inpainting on unbounded neural radiance fields. arXiv preprint arXiv:2305.15094 (2023)

47. Weder, S., Garcia-Hernando, G., Monszpart, A., Pollefeys, M., Brostow, G.J., Firman, M., Vicente, S.: Removing objects from neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16528–16538 (2023)
48. Wu, Q., Liu, X., Chen, Y., Li, K., Zheng, C., Cai, J., Zheng, J.: Object-compositional neural implicit surfaces. In: European Conference on Computer Vision. pp. 197–213. Springer (2022)
49. Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H., Wang, Z.: Sinnerf: Training neural radiance fields on complex scenes from a single image. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
50. Yang, B., Bao, C., Zeng, J., Bao, H., Zhang, Y., Cui, Z., Zhang, G.: Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In: European Conference on Computer Vision. pp. 597–614. Springer (2022)
51. Yang, B., Zhang, Y., Xu, Y., Li, Y., Zhou, H., Bao, H., Zhang, G., Cui, Z.: Learning object-compositional neural radiance field for editable scene rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13779–13788 (2021)
52. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
53. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)
54. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4471–4480 (2019)
55. Yuan, Y.J., Sun, Y.T., Lai, Y.K., Ma, Y., Jia, R., Gao, L.: Nerf-editing: geometry editing of neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18353–18364 (2022)
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
57. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D.: Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5741–5750 (2020)
58. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1438–1447 (2019)
59. Zhou, X., He, Y., Yu, F.R., Li, J., Li, Y.: RePaint-NeRF: Nerf editting via semantic masks and diffusion models. In: IJCAI (2023)