

InpaintNeRF360: Text-Guided 3D Inpainting on Unbounded Neural Radiance Fields

Dongqing Wang Tong Zhang Alaa Abboud Sabine Süsstrunk
 School of Computer and Communication Sciences, EPFL
 {dongqing.wang, tong.zhang, alaa.abboud, sabine.susstrunk}@epfl.ch



Figure 1. **Text-Guided Inpainting on NeRF.** Given a text instruction (“Remove the flowerpot and flowers”) and a pre-trained NeRF scene, InpaintNeRF360 can remove an arbitrary number of objects from the 3D scene and fill in the missing region with perceptually plausible and view-consistent content.

Abstract

Neural Radiance Fields (NeRF) can generate highly realistic novel views. However, editing 3D scenes represented by NeRF across 360-degree views, particularly removing objects while preserving geometric and photometric consistency, remains a challenging problem due to NeRF’s implicit scene representation. In this paper, we propose InpaintNeRF360, a unified framework that utilizes natural language instructions as guidance for inpainting NeRF-based 3D scenes. Our approach employs a promptable segmentation model by generating multi-modal prompts from the encoded text for multiview segmentation. We apply depth-space warping to enforce viewing consistency in the segmentations, and further refine the inpainted NeRF model using perceptual priors to ensure visual plausibility. InpaintNeRF360 is capable of simultaneously removing

multiple objects or modifying object appearance based on text instructions while synthesizing 3D viewing-consistent and photo-realistic inpainting. Through extensive experiments on both unbounded and frontal-facing scenes trained through NeRF, we demonstrate the effectiveness of our approach and showcase its potential to enhance the editability of implicit radiance fields.

1. Introduction

Recreating real-world scenarios is one of the main focuses of Virtual and Augmented Reality (VR/AR) applications. Neural Radiance Field (NeRF) and its variants [3, 22, 26] are capable of efficiently modeling unbounded real-world scenes with 2D-image input for photo-realistic novel view synthesis. Consequently, they have the potential to become widely accessible tools for representing the 3D world.

One important feature of 3D editing tools is the ability to modify the content of the created scene, including easily removing objects. However, there are several challenges associated with this seemingly straightforward task. In the NeRF framework, inpainting directly in 3D space becomes intractable due to the implicit representation of the captured scenes, which is encoded through the weights of multilayer perceptrons (MLP). This limitation hinders user control of scene contents, such as adding or removing objects, through simple user instructions. Additionally, training 3D inpainting networks is significantly more challenging than training with images. Therefore, a more feasible approach is to tackle 3D inpainting from a 2D perspective by utilizing multiview images as supervision.

Existing work on removing objects from a trained NeRF [17, 23, 37] primarily focuses on restrictive frontal-facing viewing angles, where the shape and silhouette of the object to be removed remain relatively unchanged across all viewpoints. These methods start with user-drawn segmentation masks as initial input and subsequently employ 2D image/video segmentation models [1, 6] to infer masks for the object from different viewing directions. This approach enables a relatively consistent segmentation of the removed region. However, these 2D-based segmentation techniques fail to extrapolate across unbounded scenes, where the shape of the object undergoes drastic changes across views. In such cases, the removed region needs to be filled with consistent geometry and appearance across multiple viewpoints.

To tackle these challenging issues and allow accurate object-level editing, InpaintNeRF360 utilizes semantic guidance to identify the object that needs to be removed or edited from the scene. Our approach is based on the intuition that the semantic meaning of an object’s identity remains more invariant against drastic viewpoint changes compared to its shape observed from a 2D direction. Moreover, incorporating text guidance is more intuitive from users’ perspectives as it allows them to interact without the need to manually input a segmentation mask.

InpaintNeRF360 leverages text input by encoding it as prompts to feed into a promptable segmentation model called Segment Anything Model (SAM) [13]. It enforces identity consistency for segmented objects in 3D space through inverse depth space projection warping across different views, serving as a multi-modal prompt for SAM. To generate perceptually consistent inpainting for the entire removed region, we guide the 3D inpainting process using a pre-trained image inpainter [31] in conjunction with multiview segmented images. We then fine-tune the trained NeRF model with a customized pipeline to reconstruct the inpainted 3D scene. To ensure view consistency, we exploit the 3D consistency of a converged NeRF [8]. Additionally, we utilize the perceptual loss [41] on the masked region to

encourage the creation of a visually plausible inpainted region that seamlessly blends with the surrounding environment.

To the best of our knowledge, InpaintNeRF360 is the first work to enable inpainting on 360-degree Neural Radiance Fields. We evaluate InpaintNeRF360 on many real-world datasets, including unbounded scenes [3] and front-facing scenes [21] to demonstrate the effectiveness of our pipeline on different types of scenes. We show that InpaintNeRF360 also allows effective editing of object appearance without prompt engineering for the natural language instructions.

2. Related Works

Image Inpainting. Recent advancements in the field of 2D image inpainting, which involves filling in masked regions to ensure visually consistent and coherent completed images, have primarily relied on generative models, particularly Generative Adversarial Networks (GAN) approaches [10, 28, 44], as well as denoising diffusion models [27, 31, 32]. These models have demonstrated impressive capabilities in generating visually plausible and photo-realistic predictions for the missing pixels. However, all of these works solely focus on image domain synthesis without explicitly considering the fidelity of the underlying 3D structure across different views. In this work, InpaintNeRF360 goes beyond traditional 2D image inpainting and addresses the task of inpainting in unbounded 3D scenes, while also preserving view consistency even with drastic changes in viewpoint.

Inpainting Neural Radiance Fields. The use of neural radiance fields, such as NeRF, to represent 3D scenes [22] has achieved high-quality, photo-realistic novel view synthesis. Subsequent works have extended the original framework for improvements such as material editing [34, 40, 43], faster training speed [7, 26, 38].

However, only a few works explore removing objects, i.e. inpainting, from a pre-trained NeRF model in 3D space. The implicit representation employed by the underlying neural networks makes NeRF challenging to edit with simple instructions, thus limiting the explicability of NeRF. Some previous works [24, 35] utilize the supervision of Contrastive Text-Image Pre-Training (CLIP) [30]. However, these methods focus on inpainting a single object and cannot be easily generalized to real-world scenes. Methods utilizing depth-based inverse approaches [17, 23, 37] or 2D image segmentation models [6] to remove a single object often require user-drawn masks or input depth sequences as priors. These approaches are typically limited to front-facing scenes where all camera poses roughly point in the same direction. In contrast, InpaintNeRF360 aims to ex-

tend the inpainting capabilities of NeRF to unbounded 360-degree scenes.

Object Segmentation with 3D consistency. Object segmentation [1, 12, 16] in the image domain is a well-studied task that involves assigning a classification label to each pixel in the image. Existing SOTA models [5, 39] utilize Grounded Language-Image Pre-training (GLIP) [15]. These models extract object-level visual representations and encode joint modalities with texts into a shared vector space to assign instance segmentation to identifiable objects. Such methods often cannot output view-consistent segmentation for desired objects due to their unawareness of 3D structure. When observed from particular angles, the semantic identity of the object may be ambiguous due to variation of the object outline across different views or occlusion, leading to detection failures.

In contrast, InpaintNeRF360 operates in 3D and uses depth information priors to encode natural language into multi-modal prompts, such as bounding boxes and points, for a promptable segmentation model, the Segment Anything Model (SAM) [13]. We creates 3D consistent segmentation for the objects to remove.

Text Instructed 3D Editing. Given the rise of text-conditioned image generative models, many works focus on generating 3D content with text instructions. Some rely on joint embeddings of CLIP to synthesize 3D meshes [20, 25] or neural radiance fields [11, 14]. Others distill a pre-trained diffusion model to optimize a 3D NeRF scene in the latent space [19, 29]. These methods all suffer from having to map the inconsistencies of 2D diffusion model outputs to a 3D consistent scene. We notice a concurrent work Instruct-NeRF2NeRF [8] that edits renderings of a pre-trained NeRF model to preserve 3D consistency. However, it lacks the ability to remove objects from the scene, and cannot perform accurate object-level editing as it operates in latent space for image editing. InpaintNeRF360 encodes semantics into image space to allow for accurate object appearance editing. Moreover, our method allows removing an arbitrary number of objects from the scene solely through text instructions.

3. Method

InpaintNeRF360 takes as input a reconstructed 3D NeRF scene with source images which the model trained on, and an instructive text prompt. It outputs the inpainted 3D scene with the desired object(s) edited or removed and filled with view-consistent background.

We initialize the process by encoding textual input, which provides high-level instructions for object detection (Sec. 3.2.1). We then leverage the depth information ob-

tained from a trained NeRF to enforce 3D consistency on the segmentation prompts (Sec. 3.2.2), followed by refining masks for inpainting (Sec. 3.2.3). At last, we employ a pre-trained 2D image inpainter [31] to edit the segmented image observations and fit the NeRF scene with edited 2D images as prior while preserving view-consistency (Sec. 3.3). Our model pipeline is shown in Fig. 2.

3.1. Background: Neural Radiance Fields

Our work builds upon Neural Radiance Fields (NeRF), a compact representation of 3D scenes for reconstruction, through differentiable volume rendering. NeRF encodes a 3D scene as a function f_θ parametrized by an MLP with learnable parameters θ , which maps a 3D viewing position \mathbf{x} and its 2D direction \mathbf{d} to the density σ and a viewing-dependent color \mathbf{c} : $f_\theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\sigma, \mathbf{c})$. Rendering a NeRF from a posed camera is done by sampling batches of rays for the camera pose, and rendering corresponding pixel colors for each ray. For each ray $\mathbf{r} = (\mathbf{o}, \mathbf{d})$, we sample an array of 3D points (\mathbf{x}_i, t_i) , $i = 1, 2, \dots, K$, where $\mathbf{x}_i \in \mathbb{R}^2$ and t_i is the depth. We query the MLP with these points along the ray for $\{\sigma_i\}_{i=1}^K$ and $\{\mathbf{c}_i\}_{i=1}^K$.

The estimated RGB of ray $\hat{\mathbf{C}}(\mathbf{r})$ is obtained by alpha compositing [22] the densities and colors along the ray:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^K \alpha_i T_i \mathbf{c}_i, \quad (1)$$

where $T_i = 1 - \exp(-\sigma_i \|t_i - t_{i-1}\|)$ is the ray transmittance between \mathbf{x}_i and \mathbf{x}_{i+1} , and $\alpha_i = \prod_{j=1}^{i-1} T_j$ is the attenuation from ray origin to \mathbf{x}_i . The MLP is optimized through pixel loss for the distance between the estimated pixel color and ground truth color.

3.2. Multiview 3D-Consistent Segmentation

The first stage is to obtain refined 3D segmentation masks for the objects to remove/edit given by the text input that remains identity-consistent for all viewing directions. With the set of N source camera poses and a pre-trained NeRF model, we render N image observations $\{\mathbf{I}_n\}_{n=1}^N$ to be inpainted through text instructions.

3.2.1 Multi-modal encoding as Prompt

As we use natural language as input to locate the object region, relying solely on image-level visual representations in CLIP is insufficient for precise object localization based on text instructions.

To overcome this limitation, we employ a grounding model GLIP pre-trained on a large-scale dataset of image-text pairs [15]. GLIP is specifically designed to learn semantic-rich representations of objects through the task of phrase grounding. Phrase grounding involves establishing

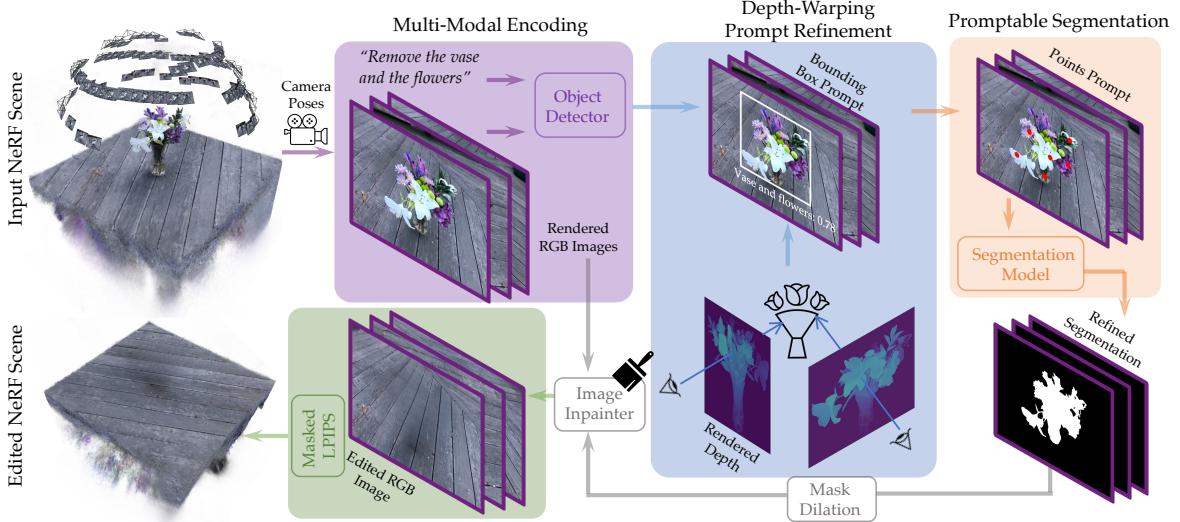


Figure 2. Overview of InpaintNeRF360 framework. InpaintNeRF360 begins with the input of a pre-trained NeRF model and its source image dataset. These images are then encoded, along with text instructions, for object detection. To improve the accuracy of the generated 2D bounding box, we incorporate depth information from the pre-trained NeRF. Using point-based prompts, we segment the rendered images to obtain detailed segmentation masks. With the generated masks and corresponding rendered RGB images, we employ a 2D image inpainter to generate inpainting priors. Finally, we train a new NeRF model using a masked LPIPS loss with the inpainting priors to remove the desired objects while preserving perceptual quality.

correspondences between natural language descriptions and specific objects or regions within an image. By leveraging this pre-trained grounding model, we are able to generate bounding boxes that accurately identify the objects mentioned in the text instructions.

The model for 2D object detection $\text{OD}(\mathbf{I}, s) = \{(B_q, p_q)\}_{q=1}^Q$ takes in an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a text s , and encodes each through respective image and language encoders. We then take these bounding boxes $\{B_q\}_{q=1}^Q$ from the model output in which $B_q = (l_q, r_q, u_q, d_q) \in \mathbb{R}^4$ with corresponding probability p_q . These boxes are typically inaccurate as some cannot enclose our desired region, see Fig. 3. Hence, we propose our depth-based prompt refinement for consistent segmentation.

3.2.2 Depth-warping prompt refinement

The inaccuracy of the bounding boxes becomes more pronounced when the orientations of the input cameras are unconstrained. Thus, utilizing initialized rough bounding boxes as prompts for object segmentation based on text instructions leads to inconsistent masks across views, as depicted in the top row images of Figure 3. Instead of relying solely on bounding box prompt engineering, our method utilizes the depth information estimated from the NeRF scene.

The depth information of a sampled ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ with origin \mathbf{o} and direction \mathbf{d} in a trained NeRF scene can be

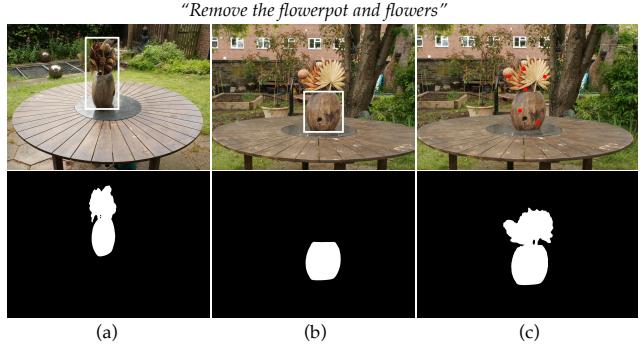


Figure 3. Inconsistent bounding box across different views (a) and (b) are from the same dataset under the same instruction. However, the generated bounding boxes are different. After applying depth warping refinement (point prompts as red dots), (c) generates accurate segmentation.

estimated via a modification of Eq. (1):

$$D(\mathbf{r}) = \sum_{i=1}^I \alpha_i T_i \cdot t_i, \quad (2)$$

After obtaining the depth of each observation, we randomly sample p points in each bounding box and estimate their depth in the 3D space along the viewing direction. We apply a "Depth Warp" on the sampled points, projecting them back into pixel space and aligning them with other 2D observations within the scene. This projection enables

us to provide point-based prompts on the object, even if the bounding box produced by the 2D object detector is incomplete or inconsistent across different views. By generating these "out-of-the-box" point-based prompts for the segmentation model, we ensure accurate capturing of the object in the segmentation masks, even in cases where the initial bounding box information may be insufficient or incomplete from certain viewpoints. This approach significantly enhances the segmentation performance and enables accurate segmentation for the desired objects in the scene.

3.2.3 Promptable Segmentation

After obtaining the refined point-based prompts, we employ the Segment-Anything Model (SAM) [13] to identify and segment all the rendered observations $\{\mathbf{I}_n\}_{n=1}^N$ for refined masks $\{\mathbf{M}_n\}_{n=1}^N$. Specifically, SAM takes as input an image \mathbf{I} and a prompt in the form of j points pinpointing the object p_q , and produces an accurate segmentation mask \mathbf{M}_q of the same size as \mathbf{I} :

$$\mathbf{M}_q = \text{SAM}(\mathbf{I}, p_q).$$

For each image $\mathbf{I}_i \in \{\mathbf{I}_n\}_{n=1}^N$, we can get an intersected binary mask for all the objects to be removed or edited:

$$\mathbf{M}_i = \bigcap_{q=1}^Q \mathbf{M}_q.$$

As a final step, we apply dilation on each segmentation mask by m pixels to include contextual background information.

3.3. Inpainting NeRF

Once we obtain the refined 2D masks using cues from multi-view images, we proceed to the process of performing inpainting on the NeRF model.

3.3.1 Unbounded NeRF

We build upon the Nerfacto model from NeRFStudio [33] as our underlying NeRF architecture. This architecture is specifically designed to optimize the performance of NeRF on real-world image captures. It incorporates various recent advances in NeRF research, including hash encoding [26], proposal sampling and scene contraction [2], per-image appearance encoding [18], and camera pose refinements.

3.3.2 Optimizing with Inpainting Priors

With the rendered observations $\{\mathbf{I}_n\}_{n=1}^N$ and corresponding masks $\{\mathbf{M}_n\}_{n=1}^N$, we adopt a 2D zero-shot image inpainter [31] to synthesize edited observations through independent inpainting as priors for the optimization. However,

each inpainted image has slightly varying pixels across different views despite being perceptually plausible as a standalone image. Therefore, relying solely on RGB supervision to fine-tune our NeRF model leads to blurring artifacts. To address this challenge, we introduce a patch-based loss [41] to enhance the model's robustness and alleviate blurring effects.

Specially, we sample all the pixels from the input inpainted images to get W image patches $\{P_w\}_{w=1}^W$ with the size of $m \times m$. These patches, $\{P_w\}_{w=1}^W$, can be divided into two non-overlapping groups of patches P_{w_i} and P_{w_o} depending on whether a patch contains pixels within the inpainted region. For patches in P_{w_o} , we apply a pixel-wise L1 loss. This loss is obtained by comparing the RGB values of each pixel p in the inpainted patch, denoted as $\tilde{\mathbf{C}}_p$, with the corresponding rendered RGB value, denoted as $\hat{\mathbf{C}}_p$,

$$\mathcal{L}_{\text{pix}} = \frac{1}{m^2 |P_{w_o}|} \sum_{p \in P_{w_o}} \left\| \hat{\mathbf{C}}_p - \tilde{\mathbf{C}}_p \right\|_1. \quad (3)$$

On the other hand, for patches in P_{w_i} containing the inpainted region, we compute the perceptual similarity using LPIPS between the inpainted image patch $\tilde{\mathbf{I}}$ and the corresponding patch P on the rendered image. We denote \mathbf{C}_P as the set of pixel colors in patch P .

$$\mathcal{L}_{\text{inpaint}} = \frac{1}{|P_{w_i}|} \sum_{P \in P_{w_i}} \text{LPIPS}(\hat{\mathbf{C}}_P, \tilde{\mathbf{C}}_{\tilde{\mathbf{I}}}). \quad (4)$$

The inpainting loss measures the perceptual difference between the inpainted patch and the target inpainted image, while the pixel loss quantifies the pixel-level discrepancy between the inpainted and rendered RGB values. Together, these losses provide a comprehensive assessment of the reconstruction quality, accounting for both perceptual similarity and pixel-wise accuracy. Our optimization is defined as the weighted sum of each loss:

$$\mathcal{L} = \mathcal{L}_{\text{pix}} + \lambda_{\text{inpaint}} \cdot \mathcal{L}_{\text{inpaint}}. \quad (5)$$

4. Experiments

4.1. Experimental Setup

In this section, we evaluate various real-world captured datasets for text-guided inpainting on InpaintNeRF360.

Baseline To the best of our knowledge, no other work has worked on the exact same problem, i.e. inpainting on 360-degree Neural Radiance Fields. We therefore identify the closest works to InpaintNeRF360 as SPIn-NeRF (SPN) [23] which only allows inpainting on front-facing scenes. The concurrent work Instruct-NeRF2NeRF (In2n) [8] stylizes 3D NeRF scenes but does not allow removing objects. We provide a comparative analysis with SPN and In2n to

highlight the unique features and advantages of InpaintNeRF360, showcasing its effectiveness in accurate object manipulation and capacity of being extended to stylization for 360-degree scenes.

Datasets We take real-world datasets from MipNeRF, MipNeRF-360, and NeRFStudio [2, 3, 33] to show the inpainting ability of InpaintNeRF360 on 360-degree scenes that our baseline methods cannot handle. In addition, due to the absence of ground truth data for 360-degree scene inpainting, we capture new datasets with and without the object removed for quantitative evaluation. Details of the datasets are provided in supplementary materials. We also evaluate front-facing datasets from SPIn-NeRF [23] and IBRNet [36] to display our improved inpainting quality over baseline methods.

4.2. Inpainting Evaluation

4.2.1 Qualitative Results

360-degree scenes InpaintNeRF360 can handle a wide range of scenarios for object removal and inpainting in 3D scenes, as depicted in the first two scenes of Fig. 5 and in the supplementary materials. For example, when instructed to “remove the vase and the flowers,” our method generates view-consistent content that seamlessly fills the missing region. In the second scene, when tasked with removing disconnected objects like “slippers” and “piano sheet” simultaneously, InpaintNeRF360 edits the scene while maintaining view consistency. This showcases one of the key strengths of InpaintNeRF360, namely its ability to inpaint multiple objects located anywhere in the 3D scenes without introducing blurry artifacts in the resulting NeRF scene. Consequently, the inpainted regions seamlessly blend with the surrounding context, yielding visually coherent and high-quality inpainting results. We encourage the readers to check our supplementary videos to inspect the view consistency achieved by our method. Note the consistent wooden patterns we synthesize in the garden scene.

Front-facing scenes InpaintNeRF360 also works for front-facing scenes with text instructions, and it only requires text input without hand-drawn masks as SPN. A few examples are shown in the last two scenes in Fig. 5, as InpaintNeRF360 successfully removes the object and fills in perceptually consistent content. We show baseline comparisons with SPN on how our method produces inpaintings that are more consistent with the ground truth regions in the supplementary materials.

Ablation We show the effectiveness of our depth-warping prompt refinement in Fig. 3. Only relying on text for the detection of all data images can be problematic due to the

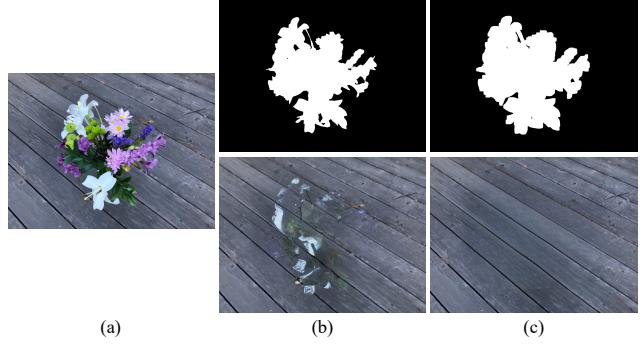


Figure 4. **Contextualized segmentation.** (a) Original image; (b) top: segmentation masks *without* dilation; bottom: resulting inpainted image; (c) top: segmentation masks *with* dilation; bottom: resulting inpainted image. Note that dilating the mask to provide contextual pixel information can improve inpainting quality.

Datasets	InpaintNeRF360	Per-frame
Garden	86%	14%
Room	69%	31%
Vasedeck	78%	22%

Table 1. **User study comparing with per-frame inpainting on visual consistency between consecutive frames.** In each of the scenes, our inpainted NeRF renders higher view consistency than per-frame inpainting. Per-frame editing lacks a 3D understanding of each scene and inpaints each image independently.

ambiguity of the object’s semantic identity. After depth projection on point-based prompts, we produce refined point-based prompts for the segmentation model. Fig. 4 demonstrates that allowing contextual information for the 2D image inpainter is very important.

Fig. 7 shows the results of the garden scene on (a) InpaintNeRF360 compared to (b) a vanilla NeRF trained directly on pixel-wise L1 loss. Without masked-perceptual loss during training, the inconsistent inpainted region on 2D images will produce blurry artifacts across the whole 3D scene, not only the space containing removed objects. In comparison, our trained renderings have much higher quality.

4.2.2 Quantitative Evaluation

Inpainting consistency compared to per-frame inpainting A naive approach for 3D scene inpainting is to independently inpaint every rendered image of the scene with a 2D image inpainter.

In contrast, InpaintNeRF360 produces inpaintings with higher view consistency across all viewpoints. To demonstrate this, we conducted a user study where participants were presented with two video clips of each inpainted

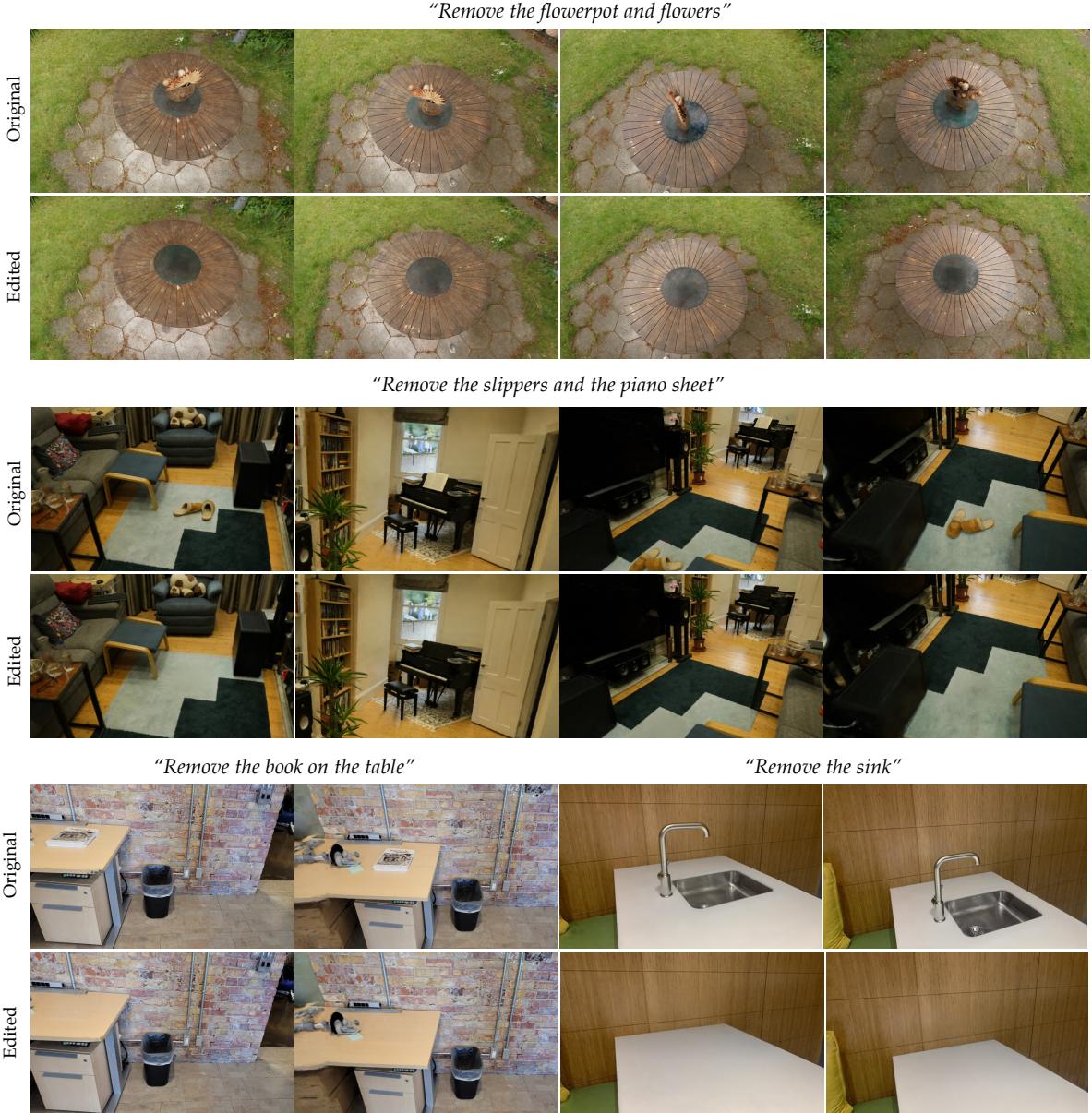


Figure 5. **Qualitative inpainting results on 360 scenes and front-facing only scenes.** Our method works with a variety type of NeRF scenes, and we also can remove arbitrary numbers of objects given the text input.

scene, rendered with sequential camera trajectories. They were then asked to identify which clip appeared more consistent. Additional details about the user study can be found in Section 2 of the supplementary material. The results, presented in Table 1, clearly show that our rendered inpaintings exhibit superior temporal consistency compared to per-frame edits.

Inpainting quality Due to the lack of baseline and ground truth datasets on inpainting unbounded NeRF scenes, we compare inpainted renderings with ground truth

Methods	Cup		Starbucks	
	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓
Per-Frame Inpaint	0.6375	181.57	0.5962	172.79
InpaintNeRF360	0.6201	168.39	0.5189	184.02

Table 2. Quantitative comparison between InpaintNeRF360 and per-frame inpainting

on our captured datasets. We report the LPIPS [42] and Frechet Inception Distance (FID) [9] results in Tab. 2

"Turn the slippers into red slippers"



Figure 6. **Editing comparison with In2n.** InpaintNeRF360, combined with appropriate mask-conditioned image editing models, can generate accurate editing on desired objects. In contrast, In2n gives the wrong texture to unwanted regions.

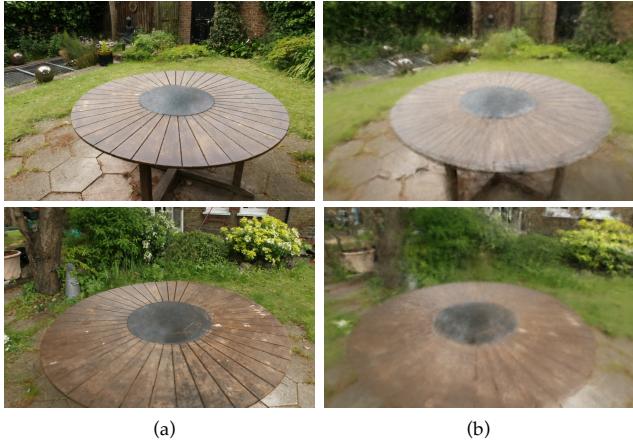


Figure 7. **Qualitative Ablation on masked LPIPS loss.** (a) NeRF trained with Masked LPIPS; (b) NeRF trained with pixel-wise L1. The latter contains blurry artifacts throughout the 3D scene.

4.3. Editing Accuracy

With a simple modification of our method by replacing the image inpainter with a mask-conditioned image editor [27], our method can produce view-consistent editing on specific objects instructed by text. Our baseline method In2n is also capable of editing, but it lacks the ability to pinpoint a particular object. In2n relies on [4] and operates solely in latent space for image editing. It applies stylization to a large undesired area of the NeRF scene, as illustrated in Fig. 6. In comparison, the modified version of InpaintNeRF360 works with object-level modification, and

delivers accurate editing results that accurately address the requested object.

5. Conclusion and Discussion

Limitations InpaintNeRF360 uses inpainted 2D images as priors for training the new NeRF scene. When the surrounding of the object to remove has irregular patterns such as a forest environment, our method cannot converge to a perceptually consistent model, as the outputs of the 2D inpainter are almost completely different. We show an example of such cases in the supplementary materials.

Conclusion We present InpaintNeRF360, a unified framework that employs natural language instructions as guidance for inpainting 3D NeRF scenes. Our method synthesizes perceptually-plausible and view-consistent content for the inpainted 3D region while being able to perform accurate object stylization, advancing the controllability of implicit radiance fields.

Acknowledgement. This work was supported in part by the Swiss National Science Foundation via the Sinergia grant CRSII5-180359.

References

- [1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017. 2, 3
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan.

- Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 5, 6
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1, 2, 6
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 8
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3
- [6] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021. 2
- [7] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2022. 2
- [8] 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, 3, 5
- [9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7
- [10] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2
- [11] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3
- [12] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, Jan. 2023. 3
- [13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3, 5
- [14] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. 3
- [15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3
- [16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3
- [17] 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
- [18] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 5
- [19] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 3
- [20] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 3
- [21] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2
- [22] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3
- [23] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, Jonathan Kelly, Marcus A. Brubaker, Igor Gilitschenski, and Alex Levinshtein. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *CVPR*, 2023. 2, 5, 6
- [24] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 20–36. Springer, 2022. 2
- [25] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 3
- [26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 2, 5
- [27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 8
- [28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature

- learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2
- [29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3
- [30] 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. 2
- [31] 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. 2, 3, 5
- [32] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2
- [33] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, et al. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023. 5, 6
- [34] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [35] 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 (CVPR)*, pages 3835–3844, 2022. 2
- [36] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6
- [37] Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. *arXiv preprint arXiv:2212.11966*, 2022. 2
- [38] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2
- [39] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3
- [40] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5453–5462, 2021. 2
- [41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 5
- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7
- [43] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 40(6):1–18, 2021. 2
- [44] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2