Hi-NeRF: Hybridizing 2D Inpainting with Neural Radiance Fields for 3D Scene Inpainting

Xianliang Huang[1,3][0000−0003−4368−1136], Shuhang Chen[1][0000−0002−2575−5923], Zhizhou Zhong[1][0000−0002−1211−035X], Jiajie Gou[1][0009−0004−7012−0435], Jihong Guan[2,4][0000−0003−2313−7635], and Shuigeng Zhou*[1,3][0000−0002−1949−2768]

[1] Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai, China
[2] School of Computer Science and Technology, Tongji University, Shanghai, China
[3]{huangxl21,sgzhou}@fudan.edu.cn [4]jhguan@tongji.edu.cn

Abstract. Recent developments in Neural Radiance Fields (NeRF) have showcased notable progress in the synthesis of novel views. Nevertheless, there is limited research on inpainting 3D scenes using implicit representations. Traditional approaches utilizing 3D networks for direct 3D inpainting often falter in high-resolution settings, mainly due to GPU memory constraints. This paper introduces Hi-NeRF, an innovative 3D inpainting approach designed to remove arbitrary 3D objects by hybridizing 2D inpainting strategies with NeRF techniques. Recognizing that prevailing 2D inpainting methods often fail to grasp the 3D geometric intricacies of scenes, we leverage the unique capability of NeRF in capturing these structures. Additionally, we propose a multi-view perceptual loss (MVPL) to harness multi-view data, ensuring that 2D inpainting and implicit 3D representations can mutually compensate for each other. Furthermore, we refine the output from the Segment Anything Model (SAM) using image dilation to produce accurate multi-view masks. To finalize the process, we employ Instant-NGP to efficiently retrieve 3D-consistent scenes from 3D-consistent inpainted images. As there is no multi-view 3D scene datasets with corresponding masks, we construct both real-world and synthetic scenes for the multi-view 3D scene inpainting task, which serves as a benchmark dataset. Experimental results on both indoor and outdoor scenes highlight the superiority of our approach over the existing 2D inpainting methods and NeRF-based baselines.

Keywords: Multi-view Scenes Synthesis · 2D Inpainting · Neural Radiance Fields · 3D Scene Inpainting.

1 Introduction

Scene-inpainting endeavors to eliminate undesired content, synthesize views for missing regions and maintain overall 3D consistency in both appearance and structure, which is a pivotal task in numerous applications [14,26,47,28,41,12].

* Corresponding author

Currently, there are various ways to represent 3D scenes, including voxels, point clouds, meshes, and multi-view images. Wang et al. [36] used RGB-D data to reconstruct 3D meshes, and evaluated the proposed method on both real-world and synthetic datasets. PointR [44] and PointDETR [45] leverage transformer encoder-decoder architectures to facilitate 3D completion on point clouds while struggling to capture fine-grained details.

However, 3D scene inpainting presents more serious challenges when adopting explicit representation, primarily due to the requirement for more computational memory. Additionally, the limited advancement in 3D sensors also makes 3D data collection both inefficient and costly.

Motivated by two considerations, we capture 2D images from different perspectives of 3D scenes. One reason for transitioning 3D scenes to 2D images is the comparative efficiency of 2D inpainting. This approach adeptly sidesteps the complications of directly training 3D networks. Indeed, it is more expensive to increase spatial resolution in 3D representations than in 2D images. Another reason is that 2D images are ubiquitous and large labeled datasets are available. These images can be rapidly inpainted by 2D inpainters. The prowess of advanced image inpainters [49,34], which extract generic features from comprehensive image databases like ImageNet [5], FFHQ [15], and Places2 [50], can be harnessed for this purpose.

Nevertheless, a notable drawback with existing 2D inpainting methods is that they cannot generate perceptually convincing appearances, and fail to grasp key 3D structures, such as consistent geometry and appearance across different viewpoints. This arises because 2D inpainters typically operate without leveraging multi-view information, instead, they inpaint each view in isolation. This paper tries to utilize implicit representation to remove undesirable objects from target scenes while ensuring that the substituted areas are consistent with the surrounding context and preserve a visually acceptable 3D structure.

The advent of neural radiance fields (NeRF) [23] breaks the barrier from multi-view images to high-quality realistic 3D scene synthesis. Recently, several works [21] have successfully applied NeRF to editing and manipulating 3D scenes. One of the subtasks related to editing is removing unwanted objects and inpainting 3D scenes. Compared to the explicitly discontinuous form of point clouds or meshes in 3D scenes, NeRF is swiftly becoming a mainstream 3D representation method due to the adoption of implicit function fitted by a straightforward MLP network and its ability to leverage geometric consistency. Furthermore, the self-supervised training feature of NeRF is inherently capable of learning the appearance and geometric consistency across multiple views. This inherent strength of NeRF addresses the drawback of 2D inpainters that typically treat each view of 3D scenes in isolation.

However, combining NeRF with 2D inpainting presents two significant challenges. First, there is a pressing demand for multi-view images with corresponding masks. However, suitable datasets remain still absent in the current landscapes. Second, simply inpainting multiple views independently leads to rendering blurry results. To handle these two challenges, we first construct a new

dataset and refine the latest release of the Segment Anything Model (SAM) to automatically generate accurate masks, which mitigates the labor cost of human-annotation masks. Subsequently, we iteratively update the parameters of 2D inpainters in a self-training manner. Specifically, we re-train NeRF with the initial 2D inpainted results as input. This enables NeRF to produce geometrically consistent inpainted outputs across varying viewpoints, facilitating the subsequent optimization of the 2D inpainter parameters. Finally, instead of simply combining a 2D inpainter and a 3D reconstructor, we propose a multi-view perceptual loss to guide the whole 3D inpainting process iteratively, which ensures the generation of 3D-consistent inpainted images.

Overall, our contributions are summarized as follows:

- We propose a novel 3D inpainting method to reconstruct 3D-consistent inpainted scenes from multi-view 2D images by hybridizing 2D inpainting with implicit representation.
- We design a multi-view perceptual loss to obtain view-consistent and perceptually acceptable outcomes.
- We construct a benchmark dataset tailored for 3D scene completion, paving the way for subsequent research in this domain.
- We conduct extensive experiments on both indoor/outdoor scenes and synthesis/real scenes, which demonstrate that our method is not only superior to 2D inpainters but also outperforms NeRF-based 3D inpainting methods.

2 Related Work

Image inpainting. Image inpainting, which is a popular task in computer vision and image processing, has received considerable attention [10,38]. Early works for filling masked regions in corrupted images can roughly be classified into two categories: diffusion-based [2] and patch-based [4]. With the development of adversarial training and transformers, various variants of 2D inpainting approaches [46,35,49,29] for improving visual fidelity have been developed. LaMa [34] applies fast Fourier convolutions to enhance the inpainting network architecture and achieves outstanding performance, even in challenging scenarios. Despite the booming research on 2D inpainting, only a few works of 3D inpainting are reported, which remains an under-explored task. Our work focuses on consistent 2D inpainting of multi-view objects or scenes to obtain reliable information for NeRF, with which eventually to synthesize 3D scenes.

Image segmentation. As a fundamental task in computer vision, Image segmentation includes interactive segmentation [39], edge detection [1], foreground segmentation [33], semantic segmentation [32], instance segmentation [19], and panoptic segmentation [17]. Traditional methods rely on pre-processing and clustering [40]. With the development of deep learning, numerous advanced approaches are proposed. The CNN-based methods have become the mainstream, which typically employ an encoder-decoder structure to keep more detailed information. Recently, transformer-based models [3] achieve state-of-the-art performance with more complex training pipeline and higher computation cost. The

latest Segmentation Anything Model (SAM) [18] is a great advance. In this paper, we employ SAM to generate multi-view masks to indicate which objects are to be removed from 3D scenes.

Novel view inpainting with NeRF. Neural radiance fields (NeRF) [23] have substantially driven the advance of computer vision in recent two years. Unlike previous voxel, point cloud, and mesh-based methods, NeRF employs a multi-layer perceptron (MLP) network to learn and represent the radiance fields in a 3D scene from a set of posed images. Instant-NGP [27] is based on NeRF and significantly improves the training speed. The success of NeRF has inspired a variety of works [25,9] on scene manipulation, including scene completion and synthesis of novel views. These methods concentrate on modifying colors or distorting shapes and lack integration with 2D methods. NeRF-In[20] is the first method that combines 2D inpainting with NeRF. However, it fails to address the problem of inconsistency and merely reduces the number of views utilized for fitting, which degrades the quality of the final results. In this paper, we incorporate 2D inpainting with NeRF to complete challenging real-world scenes that are view-consistent and photorealistic. Moreover, we leverage Instant-NGP as a rapid and powerful 3D implicit representation for the efficient reconstruction of scenarios from view-consistent inpainted images.

3 Preliminary

Neural radiance fields (NeRF) [23] introduce a neural implicit function $F_\Theta : (\mathbf{x}, \theta, \phi) \rightarrow (\mathbf{c}, \sigma)$, where the scene coordinate $\mathbf{x} = (x, y, z)$ and the azimuthal and polar viewing angles (θ, ϕ) are taken as input to output a volume density σ and an RGB color $\mathbf{c} = (r, g, b)$. This 5D function F_Θ is typically implemented by one or more Multi-Layer Perceptrons (MLPs) and Θ denotes the parameters of F. Theoretically, the rendered RGB color $C(\mathbf{r})$ can be calculated via integrating the predicted densities and colors along a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v}$ in NeRF. Due to the continuity of output value $(\mathbf{c}, \sigma)$ along the ray, the volume rendering integral equation [22] is numerically approximated by the following quadrature rule:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i, T_i = \sum_{j=1}^{i-1} \sigma_j\delta_j \tag{1}$$

where $\mathbf{r}$ denotes a ray, N is the number of samples, and $\delta_i = t_{i+1} - t_i$ is the distance between the i-th point and its following point, T_i is the accumulated transmittance. In the basic NeRF model, this is implemented by designing an MLP in two stages.

Since the process is fully differentiable, a mean square loss function is used to update the MLP parameters Θ as follows:

$$\mathcal{L}_{mse} = \sum_{\mathbf{r}\in\mathcal{R}} \|(\hat{C}^c(\mathbf{r}) - C(\mathbf{r}))\|_2^2 + \|(\hat{C}^f(\mathbf{r}) - C_i(\mathbf{r}))\|_2^2, \tag{2}$$

where $\mathcal{R}$ denotes a batch of rays. $C(\mathbf{r})$ is the ground truth, $\hat{C}^c(\mathbf{r})$ and $\hat{C}^f(\mathbf{r})$ are the coarse and fine volume predicted RGB colors of ray $\mathbf{r}$, respectively. Besides,

positional encoding is also employed before the MLP for mapping the input coordinates to a higher dimensional space, which helps represent high-frequency details.

NeRF achieves excellent performance in realistic view synthesis with densely captured images from calibrated cameras. For this reason, we utilize the output of NeRF to re-weight the parameters of a 2D inpainter and then obtain view-consistent inpainted results.

4 Method

Our proposed framework, named Hi-NeRF, aims to generate realistic 3D scene reconstructions through consistent inpainting of multi-view objects or scene captures. Firstly, we employ the latest SOTA large vision model SAM to generate the original masks of the object we want to remove from the multi-view images. To enable the collaboration between 2D inpainting and implicit 3D representation, our approach learns an inpainted NeRF model from the output of the original 2D inpainter. Next, the parameters of the 2D inpainter are updated by utilizing the rendered views and their corresponding masks. Meanwhile, we leverage the output of NeRF as appearance and geometry prior to enhancing the fitting of the 2D inpainter. Finally, we propose a multi-view perceptual loss to exploit global context and aggregate multi-view information. An overview of our approach can be seen in Fig. 1.

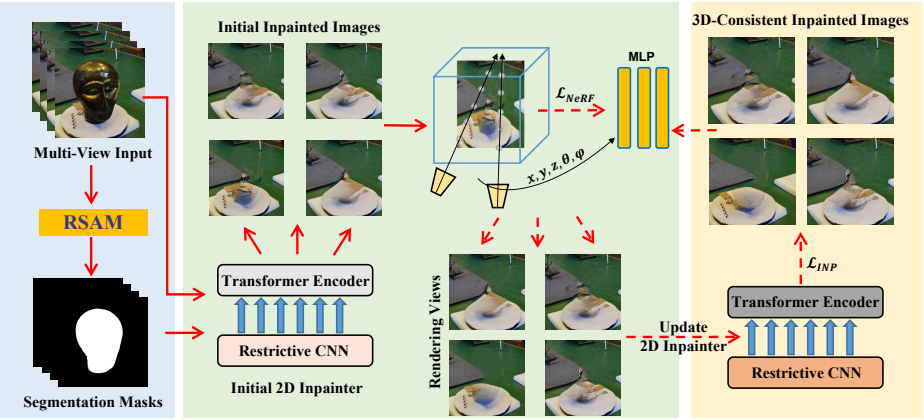

Fig. 1. The framework of Hi-NeRF. Given multi-view images and corresponding masks as input, the initial inpainted results are utilized to train NeRF. Then, the parameters of 2D inpainter are updated by using the rendering views and corresponding masks.

4.1 Problem Formulation

Our task is to inpaint scenes from multi-view images with the guidance of corresponding masks. Given multi-view input images, we need to set up viewpoints for rendering each scene. Then the camera pose for each set of images is reconstructed using a dense structure-from-motion pipeline [31]. We denote the set of original multi-view images as $\mathcal{I} = \{I_1, I_2, ..., I_k\}$. The corresponding mask set are denoted as $\mathcal{M} = \{M_1, M_2, ..., M_k\}$. Let $\mathcal{F}_\theta$ represents the initial parameter of 2D inpainter, we formulate the i-th initial inpainted multi-view image as follow:

$$\widetilde{I}_i = \mathcal{F}_\theta(I_i, M_i), (i = 1, 2, ..., k) \tag{3}$$

I_i and M_i are the i-th view and corresponding mask, respectively. We define the set of inpainted images as $\widetilde{\mathcal{I}} = \{\widetilde{I}_1, \widetilde{I}_2, ..., \widetilde{I}_k\}$. Note that these multi-view images are inpainted independently, and directly supervising a NeRF using the inpainted views may lead to blurry results due to the 3D inconsistencies between each inpainted views. Our goal is to synthesize a sequence of novel inpainted views and overcome the problems mentioned above.

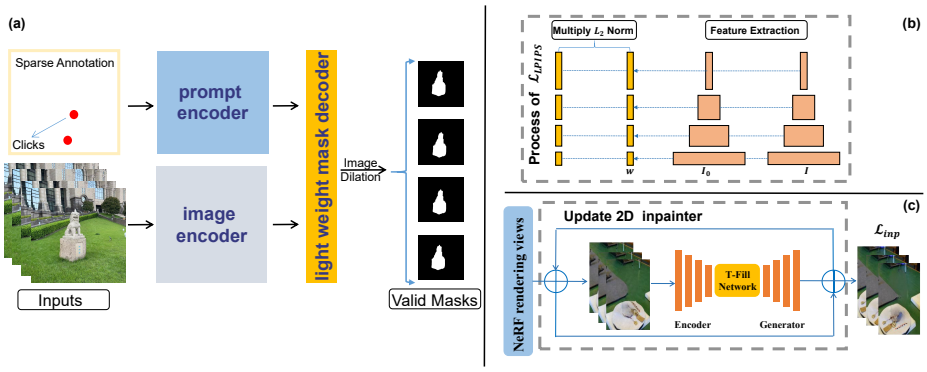

Fig. 2. Left: the process of generated multi-view masks by our Refined Segment Anything Model (RSAM). Right: the key processes of Hi-NeRF. Parameter iteratively updated with rendering views.

4.2 Refined Segment Anything Model (RSAM)

We utilize the Segment Anything Model (SAM)[18] to identify specific objects in multi-view images based on sparse points, indicating the regions requiring inpainting. These sparse points are subsequently leveraged across all views to derive multi-view masks. As depicted in Fig.2, this method encompasses three principal components. Firstly, the prompt encoder integrates positional encodings with convolutions to characterize both sparse and dense prompts. Secondly, the image encoder employs a pre-trained Vision Transformer (ViT) to efficiently

process high-resolution inputs, optimizing both speed and performance. Lastly, drawing from self-attention and cross-attention concepts, the streamlined mask decoder incorporates a revised transformer decoder block, merging prompt and image embeddings.

Let $S(\cdot)$ and $I_i \in \mathbf{R}^{H\times W\times 3}$ represent the model of SAM and the input RGB images, the output segmented images can be denoted as:

$$\hat{S}_i = S\left(I_i\right)\ (i = 1, 2, \ldots, k). \tag{4}$$

To convert the segmented images into initial object masks, we define the following transformation function:

$$f(\mathbf{x}) = \begin{cases} 255, & \text{if } \mathbf{x} \text{ belong to mask region} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Then, we obtain the initial object masks:

$$M_i = f(\hat{S}_i)\ (i = 1, 2, \ldots, k) \tag{6}$$

where M_i has the same size as I_i.

We observe that the output of SAM is always close to the object silhouettes, leading to unsatisfactory inpainting results near the mask boundaries due to abrupt changes. Inspired by image dilation [7] proposed in morphology, we use a modified mask to solve the above problem. Finally, the refined mask $\hat{M}_i$ for I_i is obtained as follows:

$$\hat{M_i} = \bigvee_{(m,n)\in B} M_i(x+m, y+n),\ (i = 1, 2, \ldots, k) \tag{7}$$

(x, y) is the pixel coordinates of M_i, (m, n) is the pixel coordinate of the structure element B.

4.3 Hybridizing 2D Inpainting with NeRF

Benefiting from large-scale vision pre-training, T-Fill [49] learns a well-aligned feature and demonstrates strong power in 2D image inpainting. Our inpainting backbone model is directly adopted from T-Fill [49]. The network architecture and the updated process are illustrated in Fig. 2(b) and (c).

Specifically, we propose a self-training-like approach that utilizes the ability of NeRF to extract multi-view information and obtain inpainted images of multi-view 3D consistency. Different from directly training a 3D inpainter, our method leverages the inpainted images $\widetilde{\mathcal{I}}$ to train a NeRF model, the rendered view $\hat{I}_k$ can be obtained as follows:

$$\hat{I}_k = F_\Theta(\widetilde{I}_k, G_k, \mathbf{K}) \tag{8}$$

G_k is the corresponding camera pose of $\widetilde{I}_k$ and $\mathbf{K}$ is the camera intrinsic matrix.

Then we update the initial parameters of T-Fill $\mathcal{F}_\theta$ with the set of inpainted images $\hat{\mathcal{I}} = \{\hat{I}_1, \hat{I}_2, ..., \hat{I}_k\}$ and the corresponding masks $\hat{\mathcal{M}} = \{\hat{M}_1, \hat{M}_2, ..., \hat{M}_k\}$. We obtain the updated model parameters $\mathcal{F}'_\theta$ to generate 3D-consistent inpainted images $\bar{\mathcal{I}} = \{\bar{I}_1, \bar{I}_2, ..., \bar{I}_k\}$. Finally, we train the Instant-NGP with these 3D-consistent inpainted images $\bar{\mathcal{I}}$ and obtain the 3D inpainted synthetic result of any view.

4.4 Loss Design

Loss design is of great significance to ensure multi-view consistency and keep the texture style. To boost the synthetic performance of Hi-NeRF, our total loss function, termed Multi-View Perceptual Loss (MVPL), consists of two parts:$\mathcal{L}_{\text{INP}}$ and $\mathcal{L}_{\text{NeRF}}$ which denote 2D inpainter loss and NeRF loss, respectively. We utilize LPIPS loss [48] to increase the global-content perceptibility in the masked regions and aggregate feature information of the NeRF model from multiple perspectives. Formally, we have

$$\text{LPIPS}(\hat{I}, I) = \sum_j \frac{1}{C_j \times H_j \times W_j} \left\| \varphi_j(\hat{I}) - \varphi_j(I) \right\|^2, \tag{9}$$

$\varphi_j(I)$ and $\varphi_j(\hat{I})$ are the feature maps of generated images and target images, respectively. They have similar $C_j \times H_j \times W_j$ and are extracted by the j-th layer of the convolutional network that can be VGG, Alexnet, or Squeezenet. The NeRF regularization item with a weighting factor λ_{reg} is defined as follows:

$$\mathcal{L}_{\text{NeRF}} = \frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{\mathbf{r} \in \mathcal{R}_n} \left\| \widetilde{I}_n(\mathbf{r}) - \hat{I}_n(\mathbf{r}) \right\|^2 + \lambda_{\text{reg}} \sum_{\mathbf{r} \in \mathcal{M}_n} \text{LPIPS}(\widetilde{I}_n(\mathbf{r}), \hat{I}_n(\mathbf{r})) \Big), \tag{10}$$

where N is the number of multi-view images used for training the NeRF model in each epoch. $\mathcal{R}_n$ and $\mathcal{M}_n$ are the entire domain and masked domain of image I_n. $\widetilde{I}_n(\mathbf{r})$ is the RGB value for pixel $\mathbf{r}$ in the original inpainted images. The predicted image $\hat{I}_n(\mathbf{r})$ is obtained by volume rendering.

On the other hand, we employ a combination of $\mathcal{L}_{\text{LPIPS}}$ and the perceptual loss [13] $\mathcal{L}_{\text{Per}}$ to supervise the fitting of T-Fill for aggregating the multi-view inpainted images $\hat{\mathcal{I}}$. Then we finetune the parameters of T-Fill using the following inpainting loss:

$$\mathcal{L}_{\text{INP}} = \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{Per}} \mathcal{L}_{\text{Per}}, \tag{11}$$

Finally, the total multi-view perceptual loss of our framework is as follows:

$$\mathcal{L}_{\text{MVPL}} = \lambda_1 \mathcal{L}_{\text{INP}} + \lambda_2 \mathcal{L}_{\text{NeRF}}, \tag{12}$$

5 Experiments

We perform extensive experiments on both realistic and synthetic scenes in our proposed dataset to evaluate our method. We compare the results of Hi-NeRF

with state-of-the-art 2D inpainting methods and several NeRF-based baselines. Furthermore, we also conduct ablation study to evaluate the key components and demonstrate the robustness of Hi-NeRF with various 2D inpainters.

5.1 Implementation Details

NeRF setting. We implement our framework with Python 3.9 and train NeRF in PyTorch [42]. The coarse network and fine network use 64 samples each. The backbone of NeRF is trained for 50,000 iterations with a batch size of 2048 rays, requiring approximately 7 hours on a single GeForce RTX 3090 GPU. The MLPs consist of 4 fully-connected hidden layers, each with 64 channels and ReLU activation. The Adam optimizer [16] is used with default parameters. In each iteration, we compute the Multi-view perceptual loss with a batch number N set to 20. Each view is divided into perceptual patches of size 50×50, and we randomly sample these patches to calculate LPIPS by emitting 2500 rays. To maintain consistency in input image size between T-Fill and NeRF, we resize the captured images to 512×512 during the 2D inpainting process and scale up the inpainted result to align with the NeRF training process.
Training details. We rely on the pre-trained generator to synthesize high-quality inpainting images and refine them during training. Our backbone T-Fill follows the implementation in [49]. T-Fill setting: we optimize the encoders, mappers, and recurrent module with learning rate 1×10^{-5}, 1×10^{-3}, and 1×10^{-3}, respectively. We update the parameters of T-Fill with 1000 iterations and a batch size of 20. It takes less than an hour to fine-tune a T-Fill Network that can handle different views. Loss weights $(\lambda_{\text{reg}}, \lambda_{\text{LPIPS}}, \lambda_{\text{Per}}, \lambda_1, \lambda_2)$ are set to $(0.1, 0.5, 0.5, 1.0, 1.0)$.

5.2 Datasets, Baselines and Evaluation Metrics

Datasets. Since there are few multi-view 3D scene datasets and corresponding masks available for evaluating NeRF-based inpainting methods, we construct a multi-view 3D scene dataset with corresponding labeled masks, which contains both synthetic scenes and realistic scenes. **Real-word scenes** consists of fifteen indoor and fifteen outdoor scenes and each scene contains $20 \sim 30$ image triplets. We provide the input GT views, camera poses, object masks, and GT without the target object. **Synthetic scenes** is generated from the Real Forward-Facing dataset [23], DTU [11], and Combined Room dataset[3]. All the realistic images are captured by an Apple iPhone 12 with a resolution of 1276×1276 pixels in size. We utilize COLMAP [31] to generate the camera parameters of all synthetic and realistic scenes. A visualization of synthetic and real-world scenes of our dataset is presented in Fig.3. This dataset contains different challenging 3D scenarios and can be used for further research on quantitative evaluations.
Baselines. We compare Hi-NeRF with several existing methods that are widely utilized in 2D inpainting tasks and self-defined NeRF-based inpainting methods.

[3] https://www.3dzx.net/

Fig. 3. Overview of our proposed dataset. In this dataset, real scenes consist of indoor scenes and outdoor scenes, including ground truth, target object masks of different viewpoints, and ground truth without target objects.

Additionally, related strong baselines, Masked-NeRF, NeRF-In [20] and SPIn-NeRF [24] are also compared to demonstrate the effectiveness of our method. **Inpainting-NeRF** is proposed by infilling the masked region with latent diffusion models [30] and training inpainted images on vanilla NeRF. **Masked-NeRF** trains a NeRF without calculating the masked regions in the loss function. **NeRF-In** [20] is the first proposed method for inpainting objects based on NeRF. It utilizes the unmasked regions to optimize the parameters of NeRF. **SPIn-NeRF**[24] is a recently proposed state-of-the-art method for image generation with a pre-trained NeRF. It learns a NeRF by leveraging a learned 2D image inpainter.

Metrics. We adopt a variety of metrics to quantitatively evaluate our results from the perspectives of both visual and geometric quality. Specifically, we employ peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [37], the Frechet Inception Distance (FID) [8] and learned perceptual image patch similarity (LPIPS) [48] between the ground-truth and the output results of Hi-NeRF.

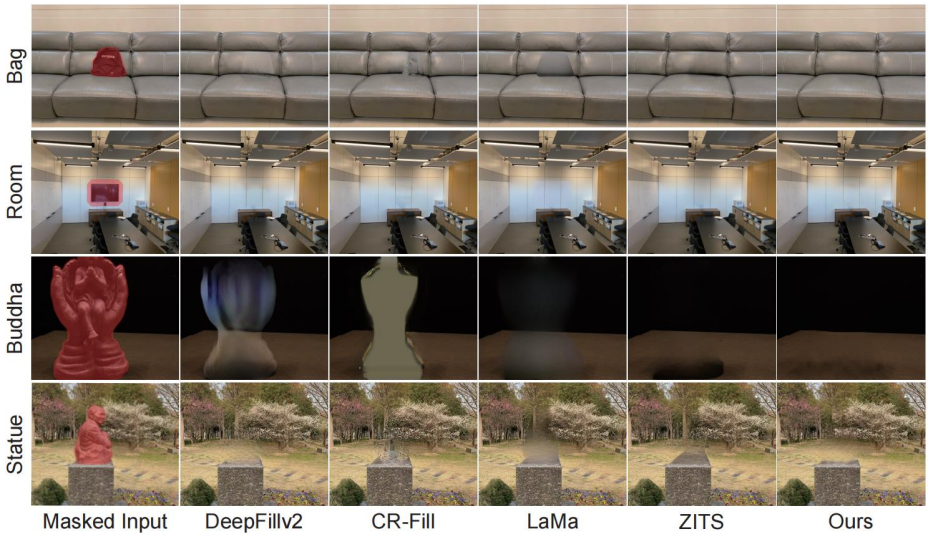

Fig. 4. Qualitative comparison on *realistic* and *synthetic* scenes of our dataset. Given the input images and corresponding masks, we visualize the output results of DeepFillv2 [43], LaMa [34], ZITS [6] and LDMs [30] in a consistent view. The above scenes are sampled from our realistic dataset.

5.3 Comparison with 2D Inpainters

We compare our method with 2D inpainting methods and inpainting-NeRF. For these methods, we follow the parameter settings in their proposed paper.

Qualitative results. Qualitative results are shown in Fig. 4, which indicate that our method generates visual-consistent inpainted results and outperforms the existing single-view independent inpainted methods on all the presented datasets. Furthermore, our results are almost invisible for the bag that is initially placed on the sofa in the *Bag* scene. The same observation can be seen in the *Buddha* and *Statue* scenes. In contrast, the inpainted views of LaMa demonstrate blurry output and inconsistent inpainted results in the masked regions. DeepFillv2 [43] demonstrates obviously severe ghost artifacts in the inpainted regions on the *Buddha* scene. The overall textural color is changed in LDMs [30]. The boundary of the target object in the inpainted results of ZITS [6] is still clearly visible on the *Room* scene and *Bag* scene.

Quantitative results. The ground truth of our real scene data is used to evaluate the quantitative metrics. The comparison results are presented in Table 1. Our method outperforms the second-best model LaMa 8.1% in the metric of PSNR. Similar improvements could be observed in terms of the SSIM, LPIPS and FID metrics.

Table 1. Quantitative comparison with existing 2D inpainting methods and the effectiveness between the SAM masks(Ours-SAM) and refined SAM(Ours-RSAM) masks as input on our proposed dataset. The best results are in **bold**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DeepFillv2	20.41	0.86	0.29	22.90
LDMs	21.30	0.86	0.28	18.77
LaMa	22.62	0.87	0.26	18.63
ZITS	22.24	0.86	0.27	19.14
Ours-SAM	22.53	0.87	0.27	10.73
Ours-RSAM	**24.43**	**0.90**	**0.25**	**8.60**

Table 2. Quantitative comparison with NeRF-based baselines on our realistic scenes. The best results are in **bold**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Masked-NeRF	17.42	0.84	0.39	13.29
Inpainting-NeRF	21.71	0.86	0.27	9.47
NeRF-In	21.64	0.86	0.31	11.92
SPIn-NeRF	22.05	0.87	0.27	10.35
Ours	**24.43**	**0.90**	**0.25**	**8.60**

5.4 Comparison with NeRF-based Baselines

In Fig. 5, we compare our method with NeRF-based inpainting baselines. From the results, we could see that our synthetic regions are with better 3D consistency and visual plausibility than other NeRF-based baselines. Both Masked-NeRF and SPIn-NeRF generate artificial textures and are unable to keep the geometry consistent through all the views in the *Fortress* and *Orchids* scenes. Additionally, NeRF-In results in the loss of fine structures in the inpainting area because there is no constraint on the multi-view inpainted images. In contrast, our method produces view-consistent outcomes that are comparable in quality and show remarkable photo-realistic inpainting effects for unseen views. Quantitative results are present in Tab. 2, which demonstrate the superior performance of Hi-NeRF in 3D inpainting. Our approach exhibits significant improvements in PSNR metric, achieving 10.8%, 12.9%, 12.3% and 40.2% superiority over SPIn-NeRF, NeRF-In, Inpainting-NeRF and Masked NeRF in our realistic scenes. Similar improvements are observed in SSIM, LPIPS and FID.

5.5 Ablation Study

In this section, we study the robustness of our loss design and explore the effectiveness of multi-stage strategy, respectively. Furthermore, we replaced the

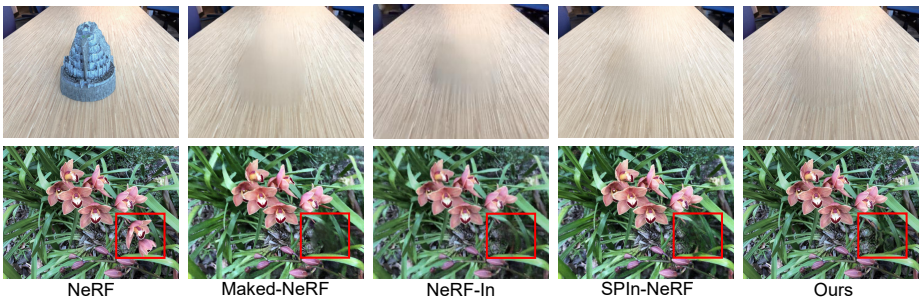

Fig. 5. Qualitative comparison with NeRF-based baselines. Since the code of NeRF-In is unavailable, we compare the experimental results using the same perspectives of LLFF data presented in the original paper of NeRF-In. Our results show more vivid results and better temporal consistency than the baseline.

Table 3. Ablation result of our proposed loss function and quantitative evaluation of different backbones on realistic scenes. The best results are in **bold**. † indicates the backbone of Hi-NeRF is replaced by the 2D inpainter.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DeepFillv2	18.84	0.87	0.26	255.05
DeepFillv2†	20.01	0.90	0.24	217.51
CR-Fill	19.45	0.89	0.25	259.62
CR-Fill†	21.57	0.91	0.22	195.80
A:T-Fill	21.72	0.89	0.21	101.04
B: A + $\mathcal{L}_{\text{INP}}$	22.42	0.90	0.20	94.14
C: B+ $\mathcal{L}_{\text{NeRF}}$	**23.15**	**0.92**	**0.17**	**90.74**

inpainting backbone of Hi-NeRF with different 2D inpainters to verify the plug-and-play capability.

Effectiveness of loss design. We conducted ablation experiments on our proposed dataset. As discussed in Section 4.4, the purpose of $\mathcal{L}_{\text{MVPL}}$ is to optimize the inpainting result of the 2D inpainter. The quantitative results are presented in Tab. 3, from which we can see that the best result is achieved when $\mathcal{L}_{\text{NeRF}}$ and $\mathcal{L}_{\text{INP}}$ are jointly used. Our method integrates these two losses and eliminates undesirable objects while retaining 3D consistency in the masked regions by learning multi-view information as well as global content. This experiment shows that our $\mathcal{L}_{\text{MVPL}}$ significantly improves all metrics.

Effectiveness of different inpainters. To demonstrate the effectiveness of Hi-NeRF, we replace our inpainting backbone with different inpainters, DeepFillv2 and CR-Fill, denoted as DeepFillv2† and CR-Fill† respectively. The results on the *Ipad* scene from our dataset are depicted in Fig. 6. Notably, the original inpainting result of DeepFillv2 is exceptionally terrible, while DeepFillv2† improves the visual consistency. CR-Fill displays poor performance in inpainting

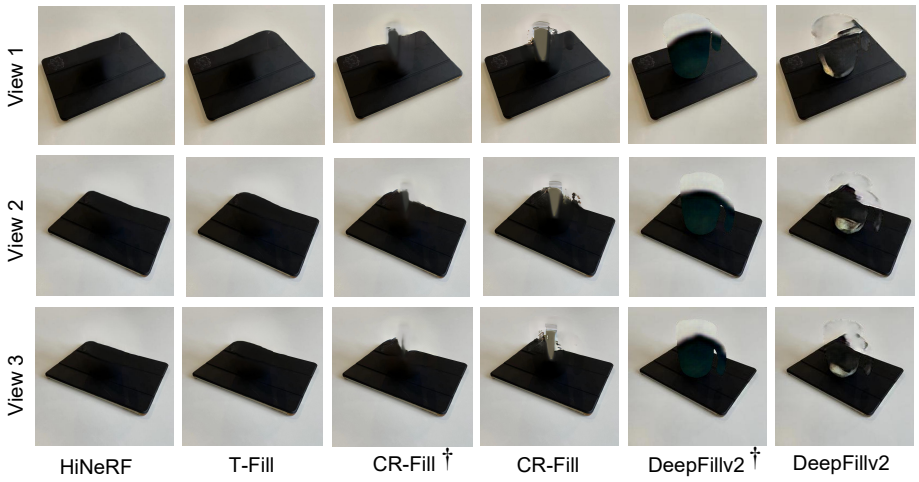

Fig. 6. Qualitative comparison with different 2D inpainters. We demonstrate three views of generated results. Methods marked with † indicate that the backbone of Hi-NeRF has been replaced with this inpainter.

region. Nevertheless, CR-Fill† surpasses CR-Fill by a significant margin due to its detailed learning of multi-view information and high-quality fitting of geometric structures. Consequently, the appearance and 3D structure of masked regions are highly consistent with the ground truth. The quantitative results presented in Tab. 3 also indicate that a significant improvement with the introduction of Hi-NeRF compared to the previous version of 2D inpainting.

6 Conclusion

In this paper, we propose a novel method named Hi-NeRF that exploits the strength of 2D painters and the synthesis of NeRF to complete the missing regions in 3D scenes. Since each view is inpainted independently, directly supervising a NeRF using the inpainted views will lead to blurry results due to 3D in consistency. So we take advantage of NeRF's ability to aggregate multi-view information and hybridize 2D inpainting methods iteratively via the MVPL loss to improve the final 3D results. Extensive experiments on our multi-view scene (mask) dataset show that Hi-NeRF is not only superior to its counterparts but also effective in improving 2D inpainters on quality and stability.

Acknowledgments. Jihong Guan was supported by National Key R&D Program of China under grant No. 2021YFC3300304.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence **33**(5), 898–916 (2010)
2. Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. IEEE transactions on image processing **12**(8), 882–889 (2003)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III. pp. 205–218. Springer (2023)
4. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on image processing **13**(9), 1200–1212 (2004)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dong, Q., Cao, C., Fu, Y.: Incremental transformer structure enhanced image inpainting with masking positional encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
7. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. IEEE transactions on pattern analysis and machine intelligence (4), 532–550 (1987)
8. 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)
9. Huang, X., Gou, J., Chen, S., Zhong, Z., Guan, J., Zhou, S.: Iddr-ngp: Incorporating detectors for distractors removal with instant neural radiance field. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1343–1351 (2023)
10. Jam, J., Kendrick, C., Walker, K., Drouard, V., Hsu, J.G.S., Yap, M.H.: A comprehensive review of past and present image inpainting methods. Computer vision and image understanding **203**, 103147 (2021)
11. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 406–413 (2014)
12. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. arXiv preprint arXiv:2406.03877 (2024)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
14. Kang, S.K., Shin, S.A., Seo, S., Byun, M.S., Lee, D.Y., Kim, Y.K., Lee, D.S., Lee, J.S.: Deep learning-based 3d inpainting of brain mr images. Scientific reports **11**(1), 1–11 (2021)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

17. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9404–9413 (2019)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
20. 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)
21. 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)
22. Max, N.: Optical models for direct volume rendering. TVCG **1**(2), 99–108 (1995)
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
24. 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. arXiv preprint arXiv:2211.12254 (2022)
25. Mirzaei, A., Kant, Y., Kelly, J., Gilitschenski, I.: Laterf: Label and text driven object radiance fields. In: European Conference on Computer Vision. pp. 20–36. Springer (2022)
26. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: Lens: Localization enhanced by nerf synthesis. In: Conference on Robot Learning. pp. 1347–1356. PMLR (2022)
27. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv preprint arXiv:2201.05989 (2022)
28. Niu, Y., Pu, Y., Yang, Z., Li, X., Zhou, T., Ren, J., Hu, S., Li, H., Liu, Y.: Lightzero: A unified benchmark for monte carlo tree search in general sequential decision scenarios. Advances in Neural Information Processing Systems **36** (2024)
29. Ren, Y., Wu, J., Lu, Y., Kuang, H., Xia, X., Wang, X., Wang, Q., Zhu, Y., Xie, P., Wang, S., et al.: Byteedit: Boost, comply and accelerate generative image editing. arXiv preprint arXiv:2404.04860 (2024)
30. 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)
31. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
32. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation. In: European conference on computer vision (ECCV) (2006)
33. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149). vol. 2, pp. 246–252. IEEE (1999)

34. 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)
35. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4692–4701 (2021)
36. Wang, W., Huang, Q., You, S., Yang, C., Neumann, U.: Shape inpainting using 3d generative adversarial network and recurrent convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2298–2306 (2017)
37. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
38. Xiang, H., Zou, Q., Nawaz, M.A., Huang, X., Zhang, F., Yu, H.: Deep learning for image inpainting: A survey. Pattern Recognition **134**, 109046 (2023)
39. Xu, Z., Chen, Z., Zhang, Y., Song, Y., Wan, X., Li, G.: Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17503–17512 (2023)
40. Yang, Y., Huang, S.: Image segmentation by fuzzy c-means clustering algorithm with a novel penalty term. Computing and informatics **26**(1), 17–31 (2007)
41. Yang, Z., Jia, X., Li, H., Yan, J.: Llm4drive: A survey of large language models for autonomous driving. arXiv e-prints pp. arXiv–2311 (2023)
42. Yen-Chen, L.: Nerf-pytorch. `https://github.com/yenchenlin/nerf-pytorch/` (2020)
43. 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 International Conference on Computer Vision. pp. 4471–4480 (2019)
44. Yu, X., Rao, Y., Wang, Z., Liu, Z., Lu, J., Zhou, J.: Pointr: Diverse point cloud completion with geometry-aware transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12498–12507 (2021)
45. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19313–19322 (2022)
46. Zeng, Y., Lin, Z., Lu, H., Patel, V.M.: Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14164–14173 (2021)
47. Zhang, M., Zhang, S., Yang, Z., Chen, L., Zheng, J., Yang, C., Li, C., Zhou, H., Niu, Y., Liu, Y.: Gobigger: A scalable platform for cooperative-competitive multi-agent interactive simulation. In: The Eleventh International Conference on Learning Representations (2023)
48. 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)
49. Zheng, C., Cham, T.J., Cai, J., Phung, D.: Bridging global context interactions for high-fidelity image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11512–11522 (2022)
50. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)