Appendix

1 Additional Related Work on 3D Completion

Traditional 3D completion approaches such as [1,5] attempt to fill small missing regions in mesh or point clouds using smoothness or geometric priors. In recent years, 3D completion methods based on deep learning have shown promising results [2]. Wang et al. [9] proposed the use of 3D-ED-GAN to generalize geometric structures and map corrupted scans to complete shapes. Sharma et al. [6] introduced a fully convolutional autoencoder to learn volumetric representations from noisy data by estimating voxel occupancy grids. Additionally, Song et al. [8] built a dataset of 3D scenes and proposed a semantic scene completion network that produces complete 3D volumes and semantic labels for a given scene based on a single-view depth map. However, most of these works are based on 3D CNN, which consume significantly more GPU memory than methods of 2D completion, rendering these methods unfeasible to high-resolution scenarios. The advances in implicit representations [7,4] have promoted the recent developments in 3D scene reconstruction. One such method is the Scene Representation Network (SRN) [7], which uses a multi-layer perceptron (MLP) as the neural representation of a learned scene given a collection of images and associated poses. DeepSDF [4] applies deep decoders to learn implicit signed distance functions of various shape instances in the same class.

2 Additional Experiments

Effectiveness of multi-stage strategy. We examine the impact of the parameter t on the inpainting results, as depicted in Fig. 1. A noticeable enhancement in visual quality is observed when t is set to 2. When t=1, there still remains visual inconsistency. Our investigation indicates that $t = 2$ is sufficient for our framework, thus we adopt it as our default setting.

The effectiveness of refined masks. We demonstrate both qualitative and quantitative results to evaluate the effectiveness of image dilation in Fig.2 and Tab.1. Given the input images and corresponding masks, we visualize the inpainting result with both SAM masks and refined masks. Fig.2 shows additional multi-view masked results utilizing both real-world and synthetic inputs. Tab.1 reports the rendering quality of our model with the original SAM masks and the refined masks, respectively. Our multiview mask strategy benefits the further inpainting results.

2.1 Additional Multi-view Results

Here, we provide additional qualitative visualizations of our multi-view inpainting results on *Fortress* and *Combined Room* to show the effectiveness of HiNeRF. Fig. 3 demonstrates selected five views of our view-consistent inpainting

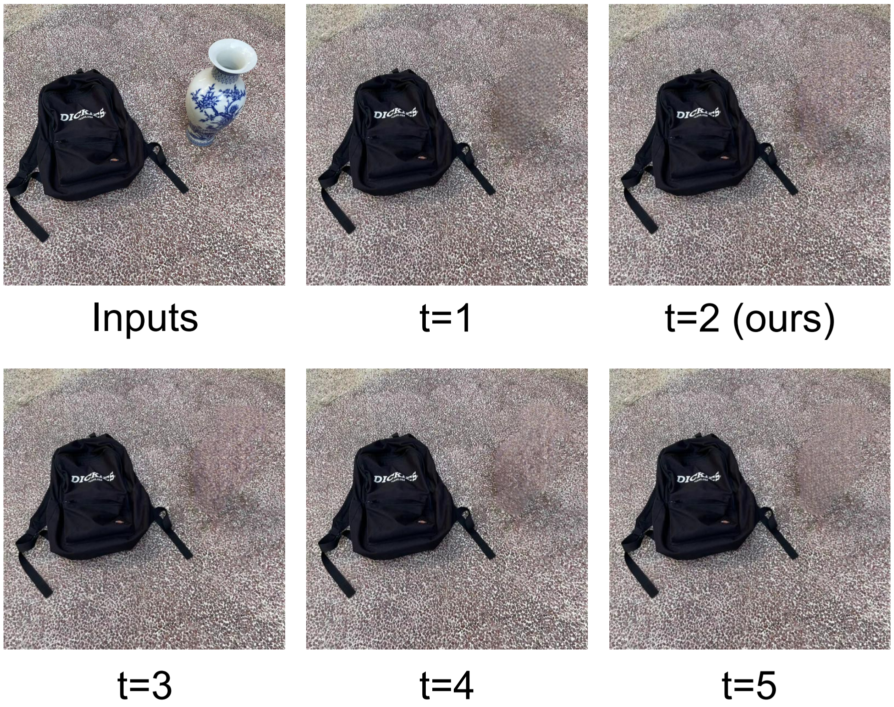

Fig. 1. Qualitative analysis of the multi-stage scheme in Hi-NeRF. The number of stages is denoted by t. We visualize the results for t=1, 2, 3, 4 and 5 in a consistent view on the *Vase* scene of our dataset.

Table 1. Quantitative comparison between the effectiveness of the original SAM and refined SAM on our dataset. The best results are in **bold**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
SAM masks	22.53	0.86	0.27	9.73
Ours refined masks	**24.03**	**0.88**	**0.25**	**8.60**

approach. Our method demonstrates great perceptual qualities when inpainting novel views on both real-world datasets and synthetic dataset.

Effectiveness of multi-stage strategy. We train a vanilla NeRF with the inpainted images $\widetilde{I}$ and re-train the 2D inpainter with the rendering images $\hat{I}_n(\mathbf{r})$ from NeRF. This process is done iteratively t times. We check the impact of times t on the inpainting results. Tab. 2 presents the changes of metrics when t increases from 1 to 5. We train NeRF 50000 epochs and the 2D inpainter 1000 epochs in each stage. When t=1, there is still visual inconsistency (Qualitative

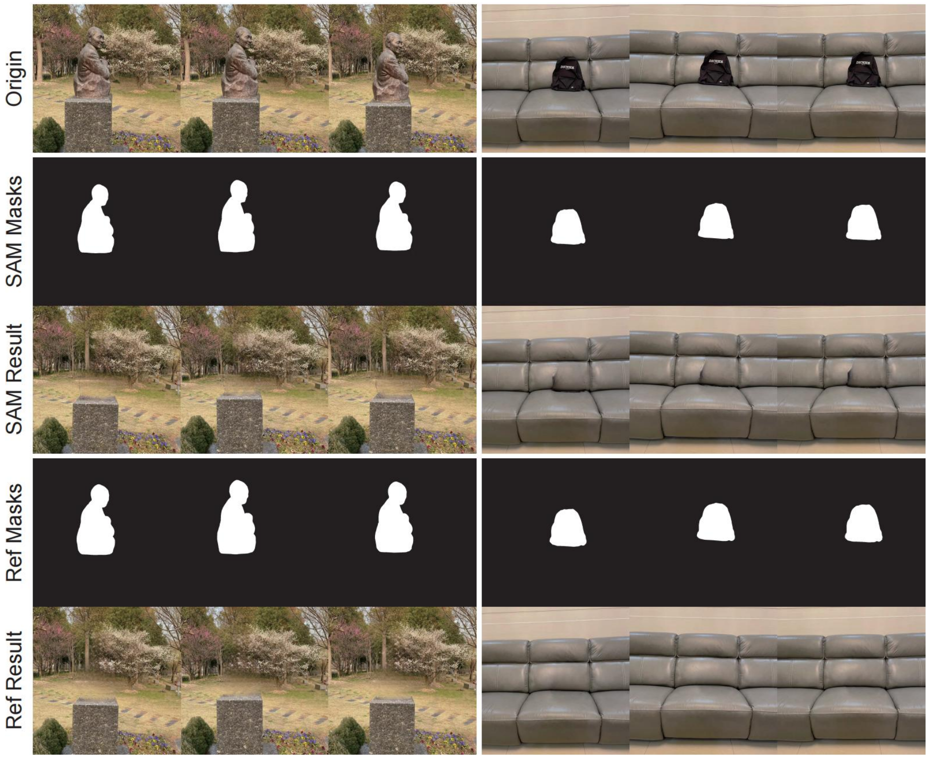

Fig. 2. Qualitative comparison of SAM masks and refined masks on *Statue* scene and *Bag* scene of our dataset.

Table 2. Effect of the multi-stage scheme. We find that $t = 2$ is enough for our framework and set to 2 as our default setting.

t-Stage	1	2	3	4	5
LPIPS	0.4716	**0.4712**	0.4713	0.4720	0.4721
FID	26.95	**26.56**	26.59	26.64	26.86

results are available in supplementary material). When t is set to 2, the metric of LPIPS and FID demonstrate minimal changes.

Failure cases and directions for improvement. We demonstrate the handling of shadows and reflections in Fig. 4. When confronted with the removal of multiple small objects, SAM faces challenges in precisely labeling masks across multi-view images, hindering Hi-NeRF from effectively filling these objects. Another improvement of Hi-NeRF is its runtime, which is a common issue in NeRF-based methods. However, recent advancements in expediting and parallelizing radiance field computations, exemplified by techniques like Instant-NGP [3], offer

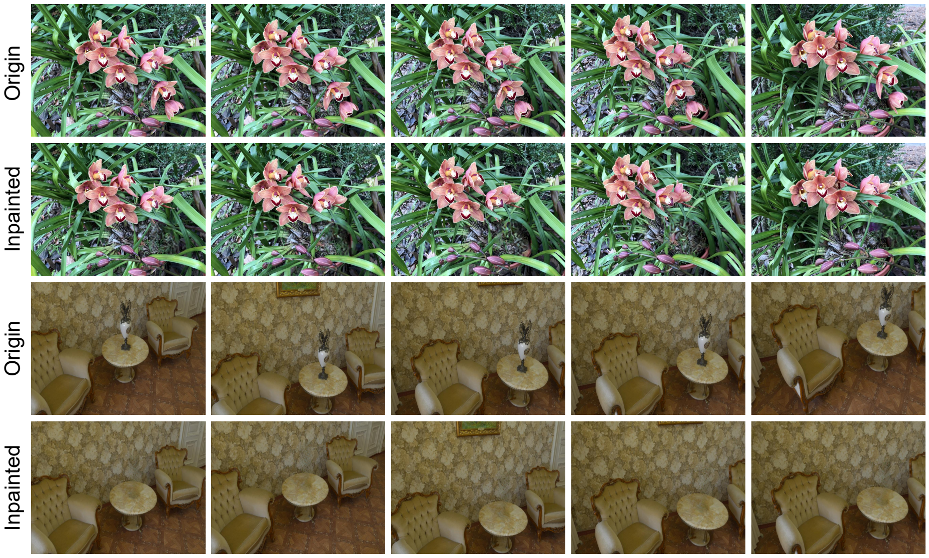

Fig. 3. Qualitative visualizations of our multi-view results.

encouraging outcomes. Consequently, ongoing efforts to mitigate computational costs remain imperative.

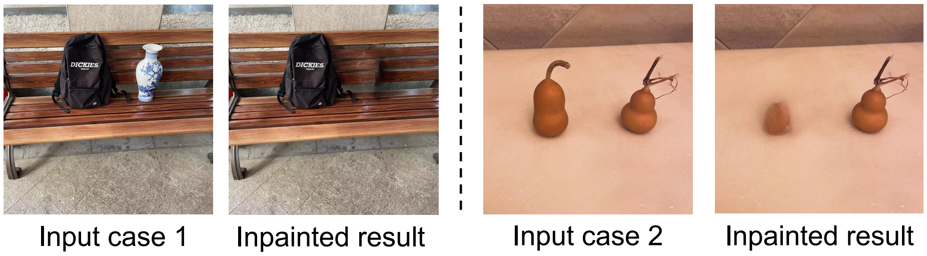

Fig. 4. Failure cases and limitations. Our method can not recover the scenes with shadows and reflections. Specifically, our method keeps the shadows of the removed objects, if they are not included in the objects masks.

References

1. Davis, J., Marschner, S.R., Garr, M., Levoy, M.: Filling holes in complex surfaces using volumetric diffusion. In: Proceedings. First international symposium on 3d data processing visualization and transmission. pp. 428–441. IEEE (2002)

2. 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)
3. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv preprint arXiv:2201.05989 (2022)
4. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
5. Sahay, P., Rajagopalan, A.: Geometric inpainting of 3d structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–7 (2015)
6. Sharma, A., Grau, O., Fritz, M.: Vconv-dae: Deep volumetric shape learning without object labels. In: European conference on computer vision. pp. 236–250. Springer (2016)
7. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019)
8. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1746–1754 (2017)
9. 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)