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

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

In this supplementary material, we provide more detailed information to complement the main manuscript. Specifically, we first conduct more ablation studies to analyze our method, by using Stable Diffusion (SD) inpainting results [3] as explicit prior for Remove-NeRF [5] and SPIn-NeRF [2]. Then, we formulate a depth SDS to further explain why we use the normal map as a geometry representation to distill knowledge from the pre-trained diffusion model. Next, we provide the additional controllability of our method, more qualitative results, and failure cases. Finally, we report the specific parameter configurations utilized in optimizing each NeRF scene from both datasets.

1. Comparisons with SD inpainting prior

Note that both Remove-NeRF and SPIn-NeRF leverage LaMa [4] for independent inpainting across multiple views, followed by the optimization of NeRF scenes. While LaMa has demonstrated superior quantitative performance, as reported in [1, 3], we conduct an additional comparison with SPIn-NeRF + SD and Remove-NeRF + SD. As reported in Table 1, we observe that: i) the SD-based inpainting method may not improve the LaMa-based version, and ii) our method still shows better performance than the SD-based inpainting approaches.

2. Depth SDS

As stated in our main paper, we incorporate a geometry diffusion prior to ensure a valid and coherent geometry in the inpainted region. We use normal SDS to distill geometry information from diffusion prior. So, a natural question is: *why not define a depth SDS*? Actually, in our early test, we also formulated the depth SDS as follows:

$$\nabla_\theta \mathcal{L}^d_{\text{masked}} = w(t)\big(\boldsymbol{\epsilon}^\omega_\phi(\mathbf{z}_t; m, y, t) - \boldsymbol{\epsilon}\big)\frac{\partial \mathbf{z}}{\partial \mathbf{d}}\frac{\partial \mathbf{d}}{\partial \theta}, \quad (1)$$

where $\mathbf{d}$ denotes the rendered depth map.

Fig. 1 presents the inpainting results with normal SDS and depth SDS. It is evident that the depth SDS result exhibits depth residuals in the inpainted regions, and displays a certain degree of edge distortion in its color images. This observation highlights the challenges associated with depth SDS. We argue that the less satisfactory performance of depth SDS can be attributed to the inherent limitation of depth values in conveying comprehensive geometry information, as opposed to surface normals which more clearly reveal the geometric structures.

3. Controlibility

An additional significant capability of our method is the generation of novel content within the masked region in the 3D scene, which we refer to as controllability. Examples illustrating this capability are presented in Fig.2. It is noteworthy that [1] also possesses the ability to insert novel content into the 3D scene by providing a different inpainted reference image. In our case, controllability is achieved by supplying a distinct text prompt and a large classifier-free guidance (CFG) value (set to 25 for all results).

It is essential to acknowledge that, due to our method's reliance on the SDS loss, the generated contents may not exhibit the same level of realism as [1], which employs realistic inpainted images. The inherent differences in the approaches highlight the trade-offs between controllability and photorealism in content generation within 3D scenes. How to better utilize diffusion to solve this problem would be an interesting direction.

4. Run time

As indicated in the main paper, our method, due to its reliance on the diffusion model, requires more time and memory resources. Specifically, for each scene in *Real-S* (image resolution: 1008×567), our model can be trained with 2 v100 GPUs and consumes approximately 6 hours of computation with $10,000$ iterations. In contrast, Remove-NeRF and SPIn-NeRF, employing a random batching scheme, can operate on a single GPU and complete the task in less than 1 hour. We can alleviate this problem by feeding the downscaled rendered image into a diffusion model.

Table 1. **Comparison with state-of-the-art methods with different 2D inpaiting methods.** Our method is relatively better compared to other novel-view synthesis baselines in inpainting the missing regions of the scene.

	Real-S				*Real-L*			
	PSNR↑	LPIPS↓	FID↓	Depth L_2 ↓	PSNR↑	LPIPS↓	FID↓	Depth L_2 ↓
Remove-NeRF + LaMa [5]	17.556	0.665	254.345	8.748	25.176	0.187	**88.245**	0.038
Remove-NeRF + SD [5]	17.381	0.677	245.941	9.997	24.612	0.201	110.817	0.029
SPIn-NeRF + LaMa [2]	17.466	0.574	239.990	1.534	25.403	0.215	103.573	0.090
SPIn-NeRF + SD [2]	17.497	0.604	**227.243**	1.610	25.102	0.194	108.286	0.089
Ours	**17.667**	**0.507**	255.514	**1.499**	**25.690**	**0.181**	100.452	**0.021**

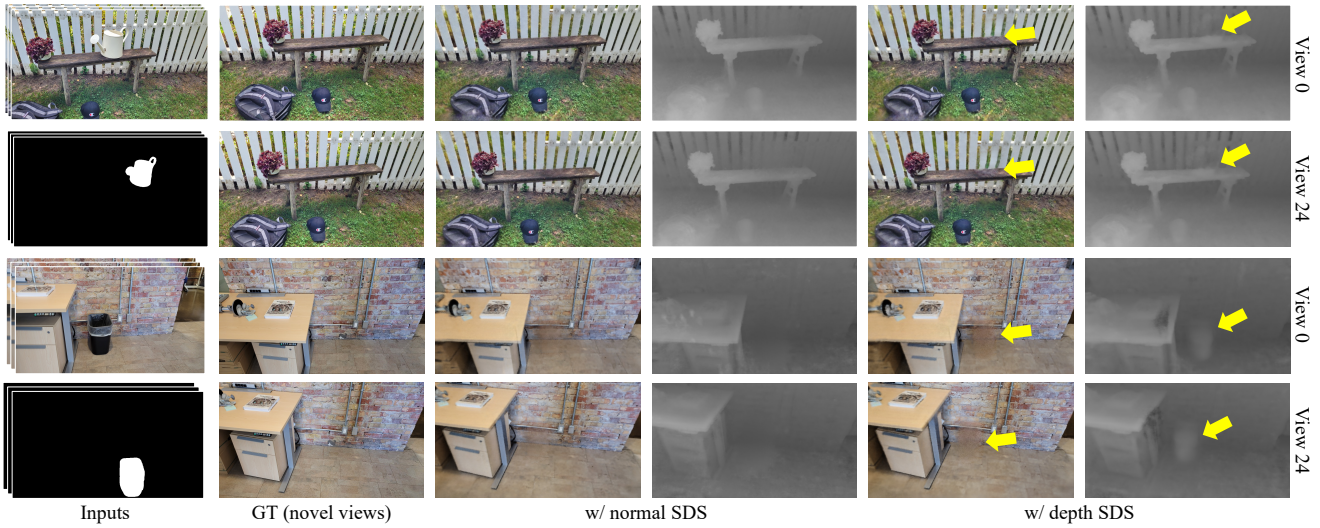

Figure 1. Comparison of depth SDS and our normal SDS. For each scene, we generate RGB images and depth maps for two novel views. Notably, the result of depth SDS reveals limitations in geometry recovery and introduces color distortions. Yellow arrows indicate the less pleasing regions.

5. Failure cases

In scenarios where a scene has very large undesired areas, and these areas are difficult to adequately describe with textual cues, our method may exhibit a tendency to generate blurred results. This limitation arises from the inherent difficulty in capturing fine details or specific features when the inpainting task involves extensive and complex regions that lack clear descriptive cues from external prompts. One potential direction for improvement is to use more accurate masks so that more information can be exploited.

6. More visual results

More detailed qualitative results for several challenging cases are demonstrated in Fig. 4. We show the inpainted results from two novel different viewpoints. We can observe that our approach not only excels in recovering large missing regions but also demonstrates proficiency in restoring intricate textures and maintaining well-aligned geometries.

7. Evaluation settings and more quantitative results

Due to the potential influence of the underlying NeRF architecture, in our evaluation settings, we replaced the unmasked region of the rendered images with their ground truth. Thus, only the masked region contributes to the final error. Observing that the masking scheme (whether the unmasked region is set to 0 or GT) and the LPIPS version (VGG or Alex) can affect the results, we report more detailed results in Table 2.

8. Detailed parameter settings

Given that our method is tailored to leverage the text-to-image diffusion model, we provide detailed parameters for each scene in Table 3, including the inputted text prompt, classifier-free guidance (CFG) for multi-view SDS, and CFG for normal SDS. This provides a comprehensive overview of the input configurations.

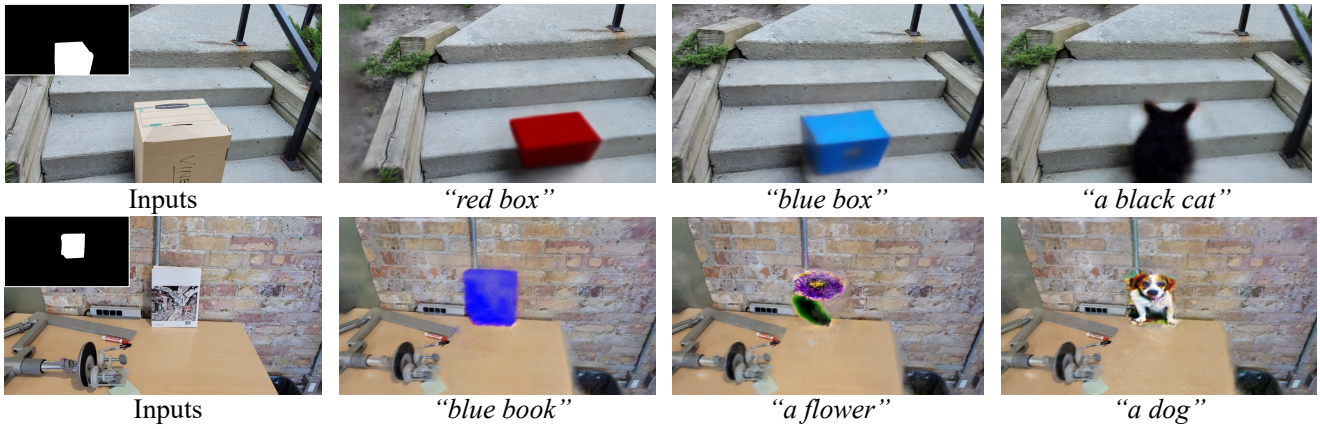

Figure 2. Controlibility of our method. Our method can yield different inpainting results by setting different text prompts.

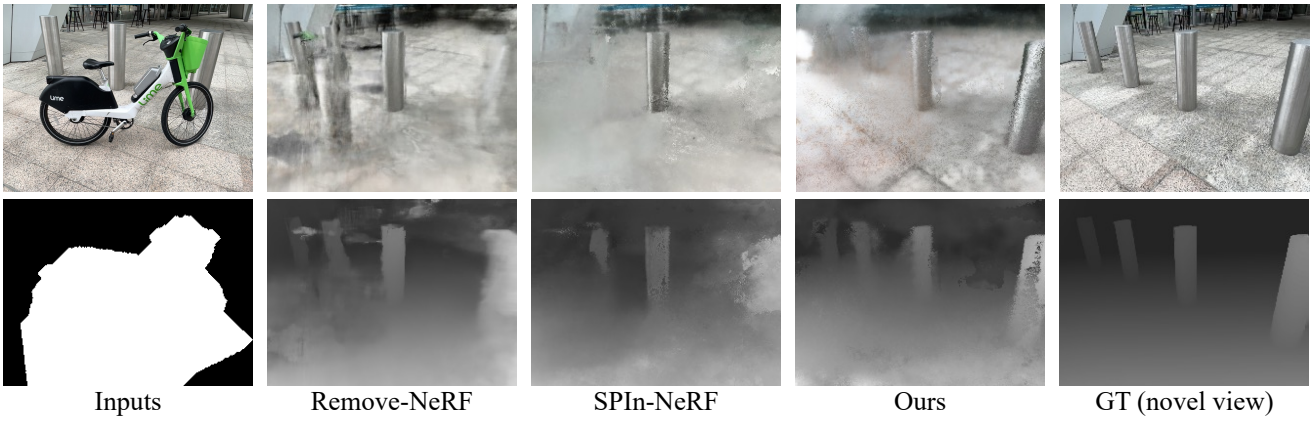

Figure 3. Failure case. The input is a scene with very large undesired areas, and these areas are difficult to adequately describe with a textual prompt (*“a group of metal poles sitting on an outdoor floor”*), our method may exhibit a tendency to generate blurred results.

References

[1] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. Reference-guided controllable inpainting of neural radiance fields. *arXiv preprint arXiv:2304.09677*, 2023. 1

[2] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. 1, 2

[3] 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. 1

[4] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1

[5] Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel J Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16528–16538, 2023. 1, 2

Figure 4. More visual results on different scenes. For each scene, we present inpainted results from two novel viewpoints. It is noteworthy that our approach not only excels in recovering large missing regions but also demonstrates proficiency in restoring intricate textures and maintaining well-aligned geometries.

Table 2. LPIPS results computed by different evaluation schemes. Left side of "/": REAL-S. Right side: REAL-L.

	VGG & umasked=0	VGG & umasked=GT	Alex & umasked=0	Alex & umasked=GT
Remove-NeRF	0.503/0.123	0.588/0.170	0.584/0.121	0.665/0.187
SPIn-NeRF	0.425/0.138	0.513/0.212	0.497/0.143	0.574/0.215
Ours	0.409/0.115	0.488/0.163	0.443/0.119	0.507/0.181

Table 3. Detailed parameter setting. We report detailed parameters for each scene, including the inputted text prompt, CFG for multi-view SDS ($\mathcal{L}^{ma}_{\text{masked}}$), and CFG for normal SDS ($\mathcal{L}^{g}_{\text{masked}}$). This provides a comprehensive overview of the input configurations

Dataset	Scene	Text prompt	CFG-ma	CFG-g
Real-S	1	a stone park bench	7.5	7.5
	2	a wooden tree trunk on dirt	7.5	2.5
	3	a red fence	7.5	7.5
	4	stone stairs	7.5	2.5
	7	a grass ground	15	7.5
	9	a corner of a brick wall and a carpeted floor	12.5	5
	10	a wooden bench in front of a white fence	7.5	7.5
	12	grass ground	15	7.5
	book	a brick wall with an iron pipe	12.5	5
	Trash	a brick wall	12.5	5
Real-L	001	a gray carpet floor	7.5	7.5
	002	office desk, carpet floor	25	7.5
	003	a sofa	7.5	7.5
	004	black door, white wall, and carpet floor	7.5	7.5
	005	an office desk	7.5	7.5
	006	a stone floor	7.5	7.5
	007	a stone bench	7.5	7.5
	008	a stone wall	7.5	7.5
	009	a wall corner	7.5	7.5
	010	a wall corner and a wooden floor	12.5	7.5
	011	a white door and a wooden floor	12.5	7.5
	012	stone staircases	7.5	7.5
	013	a wall corner	12.5	7.5
	014	a brick wall corner	12.5	7.5
	015	a brick wall	25	7.5
	016	a group of metal poles sitting on an outdoor floor	25	7.5