
InstaInpaint: Instant 3D-Scene Inpainting with Masked Large Reconstruction Model

Supplementary Material

A More Details Details

Camera Normalization and Selection. For SPIn-NeRF [1] and LLFF [2], we normalize all cameras in a scene into a $[-1, 1]^3$ world space. We first calculate the mean of all camera locations and make all poses relative to the mean pose. We then scale all the poses based on the maximum camera deviation. During training, we perform the same normalization procedure on each DL3DV [3] clip (i.e. 15 frames). During validation, we use an intuitive way to select the 4 input views. We choose the camera closest to the mean position for the reference viewpoint. We choose another 3 cameras from the remaining camera positions so that they span a triangle with maximum coverage.

Masking Finetuning. During training, we randomly selects one of three mask types at each iteration: instance masks, geometric masks, or random masks, with sampling probabilities of 25%, 25%, and 50% respectively. We provide ablation for the sampling rate Section B. After choosing the mask type, we randomly apply 1 to 4 masks on to the inpaint views. For random masks, we randomly sample rectangular masks with edge length in range $[size/6, size/4]$. For geometric masks, we define elliptical regions with axis lengths determined by the mask count. We sample the two axis from range $[size/8, size/6]$ if the mask number is 1 or 2, and $[size/12, size/8]$ if the mask number is 3 or 4.

B More Ablation Studies

Mask sampling probability. We provide an ablation study on different mask sampling strategies in Table 1. We can find that the introduction of geometric and random masks effectively improves inpainting performance. The optimal mask sampling distribution is allocating 50% probability to random masks and 25% each to geometric and object masks. This distribution best balance random masks’ ability to prevent inconsistent inpainting, geometric masks’ preservation of spatial relationships and object masks’ high multiview consistency,

Table 1: **Ablation study on mask sampling probability.** We underline our default settings in the method section. The **best** and **second-best** results are highlighted, respectively.

Object Mask	Geometric Mask	Random Mask	SPIn-NeRF			LLFF		
			LPIPS↓	M-LPIPS↓	FID↓	KID↓	C-FID↓	C-KID↓
33%	33%	33%	0.4289	0.3532	86.432	0.0139	199.25	0.0628
50%	25%	25%	0.4326	0.3678	86.973	0.0142	199.89	0.0623
25%	50%	25%	0.4191	0.3484	85.385	0.0132	198.94	0.0611
<u>25%</u>	<u>25%</u>	<u>50%</u>	0.4147	0.3130	84.535	0.0135	198.54	0.0613

C More Qualitative Comparisons and Visualizations

Please check our supplementary videos for a more straightforward comparison of inpainting consistency. We provide in the video: comparison on SPIn-NeRF and LLFF with previous 3D inpainting methods [4, 5, 6], comparison on object insertion ability and more text-guided 3D inpainting results.

D Boarder Impact

Our work bridges 3D reconstruction and inpainting by adapting Large Reconstruction Models for real-time 3D completion. This advancement enables two key opportunities:

First, when combined with 2D generative models, our system allows instant creation of 3D-consistent assets from text prompts - particularly valuable for VR/AR applications where users can interactively modify 3D environments.

Second, the technical approach demonstrates how reconstruction-focused models can be repurposed for generative tasks, suggesting a brand new research direction.

Admittedly, like any technology, our method could potentially be misused to generate manipulated 3D reconstructions or factually inaccurate renderings. We recognize the need for safeguards and mechanisms to attribute AI-generated assets.

References

- [1] 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*, pages 20669–20679, 2023. [1](#)
- [2] 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)*, 2019. [1](#)
- [3] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. [1](#)
- [4] Chieh Hubert Lin, Changil Kim, Jia-Bin Huang, Qinbo Li, Chih-Yao Ma, Johannes Kopf, Ming-Hsuan Yang, and Hung-Yu Tseng. Taming latent diffusion model for neural radiance field inpainting. In *ECCV*, pages 149–165, 2024. [1](#)
- [5] Yuxin Wang, Qianyi Wu, Guofeng Zhang, and Dan Xu. Learning 3d geometry and feature consistent gaussian splatting for object removal. In *ECCV*, pages 1–17, 2024. [1](#)
- [6] Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024. [1](#)