DeclutterNeRF: Generative-Free 3D Scene Recovery for Occlusion Removal

Supplementary Material

A. Method Details

A.1. Architecture and Training Details

DeclutterNeRF follows the core architecture and strategy of the original NeRF [23]. Specifically, we build on NeRF−− [48] and apply DeclutterNeRF on top of this structure. Our model is implemented using PyTorch [29] and trained on a single NVIDIA GeForce RTX 4090 GPU. Since our dataset typically requires no more than 10GB of VRAM, and the image size can be adjusted flexibly to control memory usage, GPUs with significantly lower configurations can also be used to train our model. Unlike recent models that employ multiple MLPs and assign distinct names to each, we adhere to the original NeRF approach by using a single MLP for training and rendering. Our MLP consists of 8 fully connected ReLU hidden layers, each with 128 dimensions. Our further camera optimization algorithm mentioned in *Sec. 3.2* and problems encountered based on the logic of NeRF−−.

For training settings, we use a scale factor of 4 and a batch size of 4096, with 200K iterations. This aligns with the training methods of current mainstream models. Even with a scale factor of 2 and a batch size of 8192, our GPU memory usage does not exceed 15 GB. We evenly distribute the batch samples across each input image, so the number of samples per image depends on the total number of images in this scene. We train our model using the Adam optimizer [14].

A.2. Annotation Mapping Details

We directly leverage OR-NeRF's efficient multiview segmentation approach to remove obstacles and construct our dataset [55]. Its multiview segmentation process is both efficient and consistent. When given point prompts on a single view, the system projects these points into 3D space using COLMAP's sparse reconstruction, establishing correspondences between 2D points and the 3D point cloud. These 3D points are then projected back to all 2D images using camera parameters, creating consistent annotations across all views. Once annotations are propagated to all views, the SAM predicts masks for each view at approximately two frames per second, without requiring neural network training for each scene.

A.3. Evaluation Settings

Due to the irregular occlusion masks in occluded images, we rearrange valid pixels from ground truth and rendered images into rectangular formats suitable for SSIM and LPIPS patch-based evaluation. This rearrangement may introduce slight variations in metrics compared to methods that directly compare original images, as the structural changes can affect SSIM and LPIPS scores. However, these differences are typically minimal and do not impact the overall evaluation results.

Considering the unavoidable occlusions when capturing real-world scenes, we calculate the rendering accuracy only within the valid visible regions using masks. Therefore, we suggest readers interpret the quantitative evaluation metrics reasonably and place more emphasis on the qualitative results, which demonstrate the true rendering performance in scenes with occlusion removal.

B. Dataset Details

B.1. Dataset Building Process

For the DeclutterSet, we capture each scene using either a Canon R6 Mark II camera or an iPhone 12 Pro, maintaining consistent exposure and focus settings throughout the capture process. To ensure high-quality multi-view inputs, we record continuous video while moving the camera in a smooth arc trajectory around the scene. From each recording, we extract 30-35 sequential frames at regular intervals, creating a forward-facing dataset similar to the classic NeRF format. We pay attention to select scenes with varying occlusion characteristics - different depths, scales, and geometric complexity. Camera parameters are estimated using COLMAP's structure-from-motion pipeline. For occlusion annotation, we used OR-NeRF's efficient multiview segmentation approach, requiring only point prompts on a single view to generate consistent masks across all views.

B.2. Considerations

While OCC-NeRF [58] provides some occlusion datasets, community feedback (as evidenced by multiple issues raised in its repository) has identified several issues with their data. These include blurry images, missing parameters, and even mismatches in ground truth for testing. Even the authors' model and code failed to reproduce their reported results.

To address these shortcomings, we constructed *DeclutterSet*, which includes a variety of occlusion types, varying occlusion sizes and camera motions, and different occluder distances. As stated in the main text, it combines reliable data from existing references and is augmented with newly captured scenes, offering a new and robust benchmark for the community.

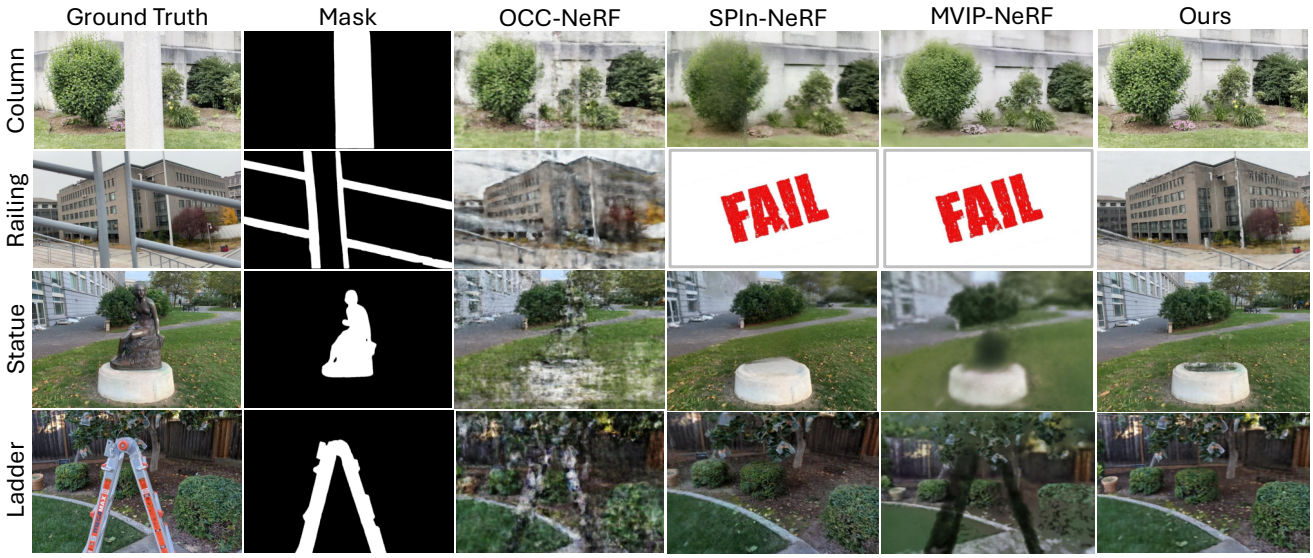

Figure 8. **Additional Qualitative Comparisons With Baselines.** Our method consistently produces desirable results, while generative models still suffer from artifacts and floaters during rendering. Notably, DeclutterNeRF maintains geometric fidelity and cross-view consistency in challenging occlusion scenarios with complex depth relationships. A detailed analysis of failure cases is provided in Sec. C.

B.3. Samples Exhibition

Figure 9 and Fig. 10 show more samples from our DeclutterSet. We select image frames that are evenly distributed to characterize our dataset: (i) wider distance distribution, (ii) larger occluded regions, (iii) greater relative motion between viewpoints and occluders, and (iv) more uncertain occluder shapes and mask layouts.

C. Additional Qualitative Results

Figure 8 shows additional visual results on our collected dataset. Beyond normal results, our method demonstrates remarkable robustness by producing high-quality renderings even when faces with incorrect camera parameters from OCC-NeRF data. This issue originates from the OCC-NeRF dataset itself. Specifically, while incorporating existing scenes to complement DeclutterSet, we observed that the *Railing* scene in OCC-NeRF suffers from camera calibration inconsistencies. Although we attempted to re-estimate the camera poses using COLMAP, the anomalies persisted. Nonetheless, we retained this scene in our dataset to reflect the realistic challenges posed by imperfect calibration—an inherent difficulty in occlusion removal tasks. As shown, baseline methods without camera parameter optimization fail to generate converged results and coherent reconstructions. OCC-NeRF produces only blurred representations, while our method successfully recovers a clear scene despite the adverse calibration conditions.

Failure Cases. The label "FAIL" in qualitative results is used to denote two distinct failure cases. (i) For SPIn-NeRF, it indicates that reconstruction was not accessible even before rendering, due to the lack of reliable depth information provided by COLMAP. (ii) For MVIP-NeRF, it refers to a failure that occurred during rendering, where the training process did not converge, resulting in extremely blurred and semantically meaningless images.

To balance reconstruction quality and memory usage when using SPIn-NeRF with COLMAP, we uniformly apply a downsampling factor of 4.

D. Statement

D.1. Ethics Statement

Due to concerns about the misuse of generative models and image processing techniques, both 2D and 3D generation have to face these issues. Our DeclutterNeRF, which does not employ any generative priors, mitigates these concerns to a certain extent. This approach helps to avoid potential ethical issues associated with generative models while still achieving effective results in our specific domain.

D.2. Open Source Statement

Through extensive experimentation with numerous baseline methods, we have identified some opportunities for improvement in the field. Many technical repositories lack proper maintenance and guidance. We recognize that to achieve occlusion removal in NeRF, 3DGS and similar fields, it is first necessary to remove the barriers that exist in the dissemination and communication of these technologies. To this end, all code and data will be open-sourced under the MIT license for community use, fostering transparency and collaborative advancement in the field.

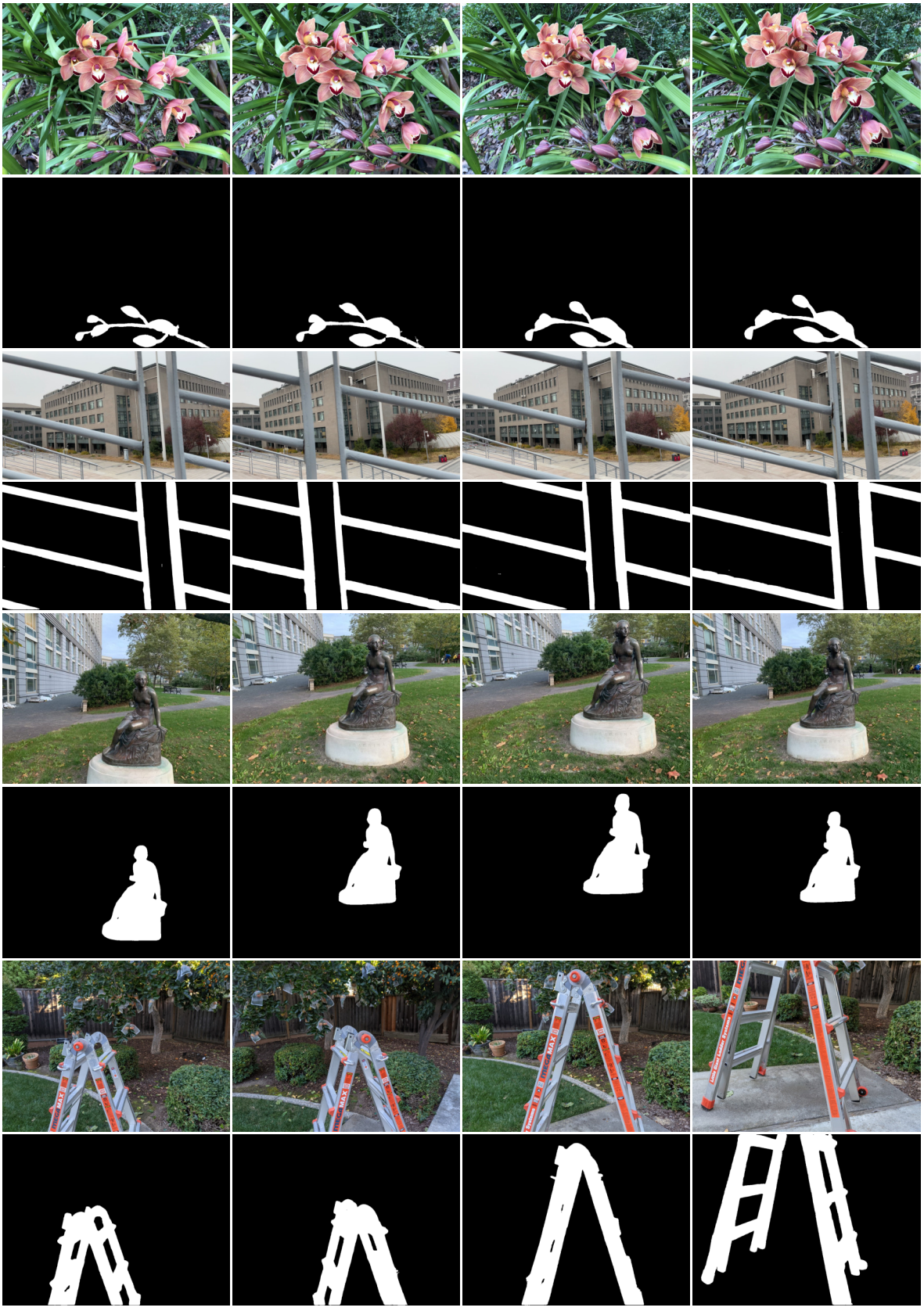

Figure 9. **DeclutterSet Illustration (Part I).** From the top to the bottom: (a) Orchids, (b) Railing, (c) Statue, (d) Ladder.

Figure 10. **DeclutterSet Illustration (Part II).** (e) Stone Column, (f) Lamp Post, (g) Chain Fence, (h) Chair Back.

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001. 2

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, page 417–424, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2

[3] Chenjie Cao, Chaohui Yu, Yanwei Fu, Fan Wang, and Xiangyang Xue. Mvinpainter: Learning multi-view consistent inpainting to bridge 2d and 3d editing. *ArXiv*, abs/2408.08000, 2024. 2

[4] Honghua Chen, Chen Change Loy, and Xingang Pan. Mvip-nerf: Multi-view 3d inpainting on nerf scenes via diffusion prior. In *CVPR*, 2024. 3, 6

[5] Jiafu Chen, Tianyi Chu, Jiakai Sun, Wei Xing, and Lei Zhao. Single-mask inpainting for voxel-based neural radiance fields. In *European Conference on Computer Vision*, 2024. 2, 3

[6] Ming-Ming Cheng, Fang-Lue Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Repfinder: finding approximately repeated scene elements for image editing. *ACM transactions on graphics (TOG)*, 29(4):1–8, 2010. 1, 3

[7] Chen Du, Byeongkeun Kang, Zheng Xu, Ji Dai, and Truong Nguyen. Accurate and efficient video de-fencing using convolutional neural networks and temporal information. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 2

[8] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, page 341–346, New York, NY, USA, 2001. Association for Computing Machinery. 1, 3

[9] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2023. 2

[10] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 1, 3

[11] Sheng-Yu Huang, Zi-Ting Chou, and Yu-Chiang Frank Wang. 3d gaussian inpainting with depth-guided cross-view consistency. *ArXiv*, abs/2502.11801, 2025. 2, 3

[12] Sankaraganesh Jonna, Sukla Satapathy, and Rajiv R Sahay. Stereo image de-fencing using smartphones. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1792–1796. IEEE, 2017. 2

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 1

[15] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[16] Axel Levy, Mark J. Matthews, Matan Sela, Gordon Wetzstein, and Dmitry Lagun. Melon: Nerf with unposed images using equivalence class estimation. *ArXiv*, abs/2303.08096, 2023. 2, 4

[17] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 5

[18] 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. *ArXiv*, abs/2404.09995, 2024. 2, 3

[19] 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, 3

[20] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14215–14224, 2020. 2

[21] Zhiheng Liu, Ouyang Hao, 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*, abs/2404.11613, 2024. 3

[22] 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. 6

[23] 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, 5

[24] 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. 1, 2, 3, 6

[25] Jingcheng Ni, Weiguang Zhao, Daniel Wang, Ziyao Zeng, Chenyu You, Alex Wong, and Kaizhu Huang. Efficient interactive 3d multi-object removal. *ArXiv*, abs/2501.17636, 2025. 2, 3

[26] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[27] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12404–12411. IEEE, 2024. 1

[28] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 5

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703, 2019. 1

[30] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019. 5

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[32] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J. Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20626–20636, 2023. 2

[33] Sara Sabour, Lily Goli, George Kopanas, Mark J. Matthews, Dmitry Lagun, Leonidas J. Guibas, Alec Jacobson, David J. Fleet, and Andrea Tagliasacchi. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *ArXiv*, abs/2406.20055, 2024. 2

[34] Ahmad Salimi, Tristan Aumentado-Armstrong, Marcus A. Brubaker, and Konstantinos G. Derpanis. Geometry-aware diffusion models for multiview scene inpainting. *ArXiv*, abs/2502.13335, 2025. 2, 3

[35] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[36] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[37] Zhihao Shi, Dong Huo, Yuhongze Zhou, Kejia Yin, Yan Min, Juwei Lu, and Xinxin Zuo. Imfine: 3d inpainting via geometry-guided multi-view refinement. 2025. 2, 3

[38] 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. 2

[39] Richard Szeliski and Philip HS Torr. Geometrically constrained structure from motion: Points on planes. In *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pages 171–186. Springer, 1998. 4

[40] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14895–14904, 2024. 1

[41] Vaibhav Vaish, Marc Levoy, Richard Szeliski, C Lawrence Zitnick, and Sing Bing Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 2331–2338. IEEE, 2006. 2

[42] Dongqing Wang, Tong Zhang, Alaa Abboud, and Sabine Süsstrunk. Innerf360: Text-guided 3d-consistent object inpainting on 360-degree neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12677–12686, 2024. 2, 3

[43] Jiaping Wang, Shuang Zhao, Xin Tong, John Snyder, and Baining Guo. Modeling anisotropic surface reflectance with example-based microfacet synthesis. In *ACM SIGGRAPH 2008 papers*, pages 1–9. Association for Computing Machinery, 2008. 1, 3

[44] John YA Wang and Edward H Adelson. Representing moving images with layers. *IEEE transactions on image processing*, 3(5):625–638, 1994. 4

[45] 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 *CVPR*, 2021. 6

[46] Yuxin Wang, Qianyi Wu, Guofeng Zhang, and Dan Xu. Gscream: Learning 3d geometry and feature consistent gaussian splatting for object removal. In *European Conference on Computer Vision*, 2024. 2, 3

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 6

[48] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2, 4, 1

[49] Ethan Weber, Aleksander Holynski, Varun Jampani, Saurabh Saxena, Noah Snavely, Abhishek Kar, and Angjoo Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *CVPR*, 2024. 2, 3

[50] Silvan Weder, Guillermo Garcia-Hernando, Áron Monszpart, Marc Pollefeys, Gabriel Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *CVPR*, 2023. 3

[51] Chung-Ho Wu, Yang-Jung Chen, Ying-Huan Chen, Jie-Ying Lee, Bo-Hsu Ke, Chun-Wei Tuan Mu, Yi-Chuan Huang, Chin-Yang Lin, Min-Hung Chen, Yen-Yu Lin, and Yu-Lun Liu. Aurafusion360: Augmented unseen region alignment for reference-based 360° unbounded scene inpainting. *ArXiv*, abs/2502.05176, 2025. 2, 3

[52] Zeke Xie, Xindi Yang, Yujie Yang, Qi Sun, Yixiang Jiang, Haoran Wang, Yunfeng Cai, and Mingming Sun. S3im: Stochastic structural similarity and its unreasonable effectiveness for neural fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18024–18034, 2023. 2, 4, 6

[53] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG)*, 34(4): 1–11, 2015. 2

[54] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8254–8263, 2023. 5

[55] Youtan Yin, Zhoujie Fu, Fan Yang, and Guosheng Lin. Ornerf: Object removing from 3d scenes guided by multiview segmentation with neural radiance fields, 2023. 2, 3, 1

[56] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*, 2023. 1

[57] Xiao Zhao, Bo Chen, Mingyang Sun, Dingkang Yang, Youxing Wang, Xukun Zhang, Mingcheng Li, Dongliang Kou, Xiaoyi Wei, and Lihua Zhang. Hybridocc: Nerf enhanced transformer-based multi-camera 3d occupancy prediction. *IEEE Robotics and Automation Letters*, 2024. 1

[58] Chengxuan Zhu, Renjie Wan, Yunkai Tang, and Boxin Shi. Occlusion-free scene recovery via neural radiance fields. 2023. 1, 2, 4, 6

[59] C Lawrence Zitnick and Sing Bing Kang. Stereo for image-based rendering using image over-segmentation. *International Journal of Computer Vision*, 75(1):49–65, 2007. 1, 3