Supplementary Material

1 Implicit Decoder Architecture

We use the convolutional decoder network as proposed by Bemana et al. [1, 5] to output a full resolution 2D map corresponding to the camera index. The normalized scalar input index is concatenated to the coordconv [3] layer in the first layer of the decoder. This is followed by a series of convolutional and upsampling layers (bi-linear interpolation) until the desired output resolution is reached. The number of parameters is controlled using a scalar capacity factor that is multiplied to a preset of each layer, e.g., [2c, 4c, 8c] where c is the capacity factor. We use capacity factor [10, 15, 18] for the depth decoder corresponding to 2, 3 and 4 inputs, respectively. Similarly, we use [6, 10, 12] for the opacity decoder.

2 3D Inpainting

Since we constrain the movement of Gaussians, our approach, unlike existing methods, does not hallucinate details in the occluded areas. This is a unique advantage as it allows us to inpaint these regions using state-of-the-art methods. To do so, we begin by rendering a novel view with holes and alpha mask. We use stable diffusion inpainting [7] to generate the missing texture in the masked regions. Then, we estimate the monocular depth on the inpainted image. Next, we use the gradient of this depth to fill in rendered depth in the masked regions using poisson blending [6]. We then project Gaussians to the scene using the inpainted depth and image. We repeat this process sequentially for a few novel views. Once this process is done, the inpainted scene is 3D consistent and can be rendered from any novel view without holes. We show additional inpainting results in Fig. 1.

3 Comparison against DNGaussian

We compare our approach against a recent state-of-the-art sparse view synthesis approach by Li et al. [2] which also utilizes the 3D Gaussian representation. We use the author provided results for both qualitative (Fig. 2) and quantitative (Table 1) comparisons on the 3-input LLFF [4] dataset. We outperform DNGaussian both visually and numerically since DNGaussian generates blurry texture with distracting artifacts.

4 Additional Ablation Results

To improve the effectiveness of decoder, we estimate a multi-channel depth offset by utilizing a multi-channel depth-based segmentation mask. Table. 2 shows the effectiveness of this multi-channel approach.

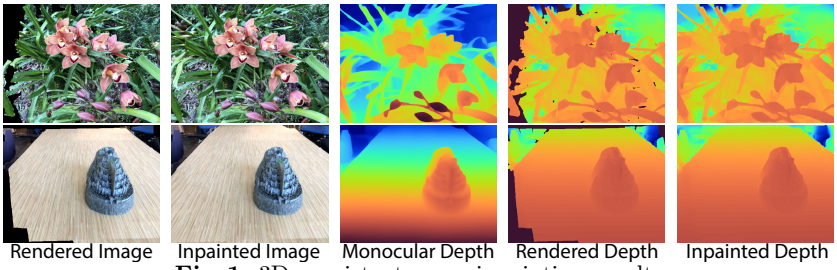

Fig. 1: 3D consistent scene inpainting results.

Fig. 2: LLFF 3 input. We show comparisons against DNGaussian [2].

5 Additional Qualitative Results

We provide additional qualitative results for 2 and 4 input configurations on the LLFF [4] and NVS-RGBD [9] datasets.

LLFF Dataset As shown in Fig. 3, our method produces significantly better texture and geometry compared to other approaches on the LLFF dataset with 2 inputs. For the Orchids scene, all the NeRF-based approaches fail to handle the complex jumble of leaves behind the flower, producing glaring artifacts. Our approach is able to capture the intricate geometry and details. The other scene, Trex, contains a lot of thin details as highlighted by the insets. NeRF-based approaches fail to capture the thin details and produce blurry texture on the trex bones. Our approach is able to reconstruct details while providing a smooth geometry. We show the 4 input results in Fig. 5. SparseNeRF [9] and FlipNeRF [8] produce noisy texture for the Fern scene, while FreeNeRF [10] produces over-blurred results. Our approach produces texture much closer to ground truth in comparison without noise or blurriness. NeRF-based approaches struggle to handle regions that contain relatively sparse supervision (e.g., regions visible in 1 or 2 views out of 4). This is highlighted in the Flower scene. As shown, NeRF-based approaches produce ghosted artifacts while our method is able to generate a coherent geometry and thus high quality details.

NVS-RGBD Dataset NVS-RGBD is a sparse input dataset with 2 and 3 views. We show the 2 input results in Fig. 4. On the Plant scene, the input views have a large angular difference in pose. NeRF-based approaches overfit to the training views and produce geometry and texture with significant artifacts. Our method is able to reconstruct a coherent geometry and high-quality texture in comparison. For the Basketball scene, SparseNeRF and FlipNeRF generate significant artifacts on the shoes while FreeNeRF is unable to capture texture details. Our approach produces details closer to ground truth without any no-

Table 1: Numerical comparisons with DNGaussian [2] on the LLFF dataset with 3 views.

Method	PSNR	SSIM	LPIPS
DNGaussian	19.55	0.647	0.264
Ours	**20.33**	**0.725**	**0.180**

Table 2: Numerical comparisons to highlight the effect of #channels in the segmentation mask.

Method	PSNR	SSIM	LPIPS
Ours w/o mask	19.70	0.673	0.214
Ours w/ 3ch mask	19.86	0.704	0.198
Ours w/ 5ch mask	**20.33**	**0.725**	**0.180**
Ours w/ 7ch mask	20.13	0.713	0.187

Fig. 3: LLFF 2 input. We show comparisons against other sparse-view NeRF-based approaches, SparseNeRF [9], FlipNeRF [8] and FreeNeRF [10].

ticeable artifacts. The differences are more prominent in the video comparisons provided in supplementary video.

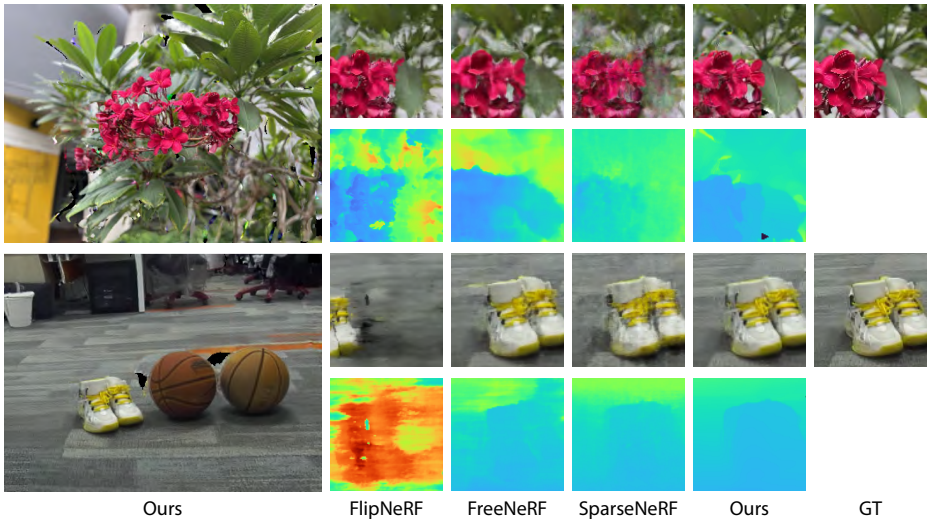

Fig. 4: NVS-RGBD 2 input (iPhone on the top and ZED 2 on the bottom). We show comparisons against other sparse-view NeRF-based approaches.

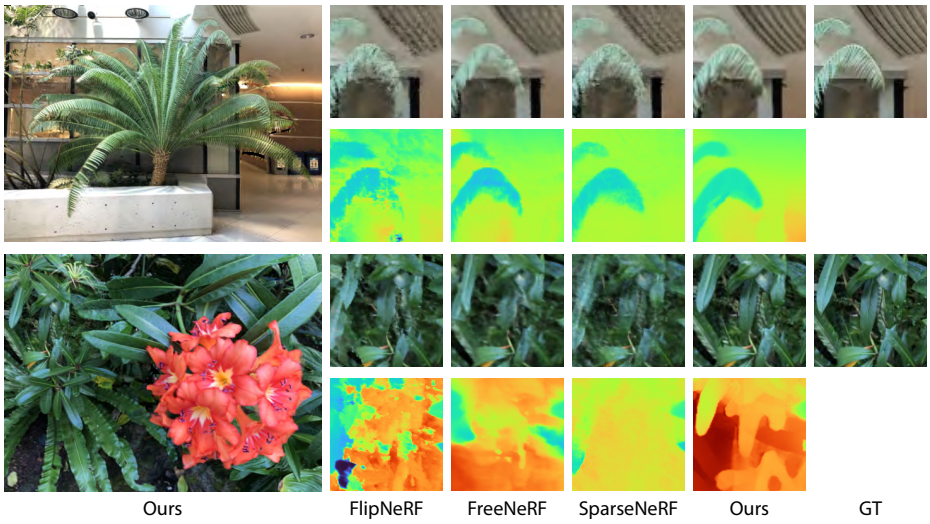

Fig. 5: LLFF 4 input. We show comparisons against other sparse-view NeRF-based approaches, SparseNeRF [9], FlipNeRF [8] and FreeNeRF [10].

References

1. Bemana, M., Myszkowski, K., Seidel, H.P., Ritschel, T.: X-fields: Implicit neural view-, light- and time-image interpolation. ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2020) **39**(6) (2020). `https://doi.org/10.1145/3414685.3417827`

2. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20775–20785 (2024)
3. Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. Advances in neural information processing systems **31** (2018)
4. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
5. Paliwal, A., Tsarov, A., Kalantari, N.K.: Implicit view-time interpolation of stereo videos using multi-plane disparities and non-uniform coordinates. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023)
6. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers. p. 313–318. SIGGRAPH '03, Association for Computing Machinery, New York, NY, USA (2003). https://doi.org/10.1145/1201775.882269, https://doi.org/10.1145/1201775.882269
7. 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 (CVPR). pp. 10684–10695 (June 2022)
8. Seo, S., Chang, Y., Kwak, N.: Flipnerf: Flipped reflection rays for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22883–22893 (2023)
9. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
10. Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8254–8263 (2023)