PDF: Point Diffusion Implicit Function for Large-scale Scene Neural Representation

Appendix

In this appendix, we provide experimental results on more benchmarks to further explore the effectiveness of our point diffusion implicit function for large-scale scenes (Sec. A and B) and conduct more ablation experiments (Sec. C).

A Experiments on the BlendMVS Dateset

Dataset and Baseline. The BlendedMVS [7] dataset is a large-scale synthetic dataset for multi-view stereo containing 113 scenes, which can be further divided into large-scale outdoor scenes part and small-scale objects part according to the scene scale. Since current large-scene NeRF methods are one model per scene, to save computational resources and time, we select the first five scenes of the large-scale outdoor scenes part and compare with Mip-NeRF 360 [2], which is the optimal baseline on the representative subset of OMMO dataset [3] as shown in our manuscript, see Tab. 4 and Fig. 6 . It can be seen that our method has a significant improvement over Mip-NeRF 360 [2] in relation to PSNR, SSIM and LPIPS, and synthesizes more realistic novel view images.

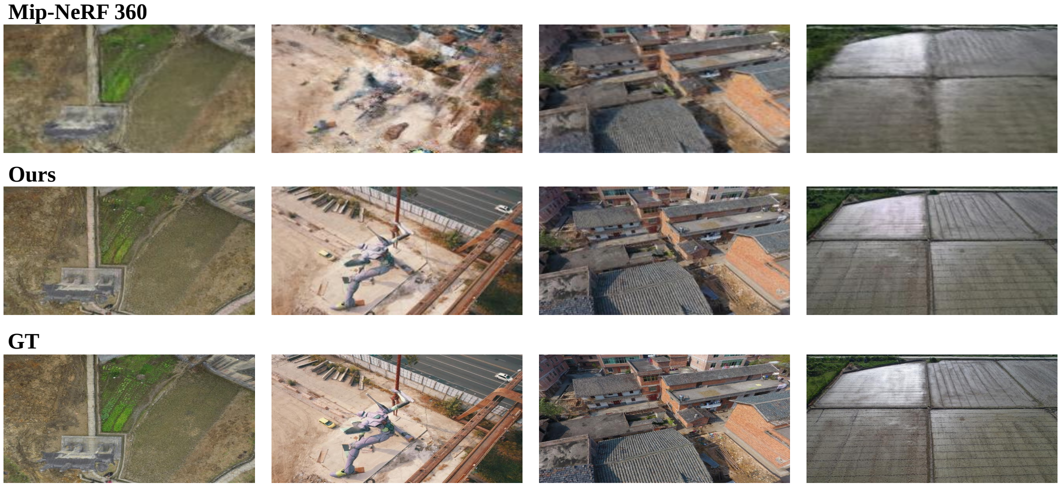


Figure 6: Qualitative results of our PDF method with the baseline Mip-NeRF 360 on the first five large-scale outdoor scenes of the BlendedMVS dataset (zoom-in for the best view).

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Scene	**Mip-NeRF 360**[2]			**Ours**		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1	15.43	0.27	0.318	**19.74**	**0.67**	**0.246**
2	13.65	0.16	0.252	**21.42**	**0.68**	**0.152**
3	19.59	0.43	0.344	**20.93**	**0.65**	**0.213**
4	18.84	0.38	0.349	**19.80**	**0.68**	**0.267**
5	17.05	0.30	**0.207**	**19.18**	**0.62**	0.271
Mean	16.91	0.31	0.294	**20.21**	**0.66**	**0.230**

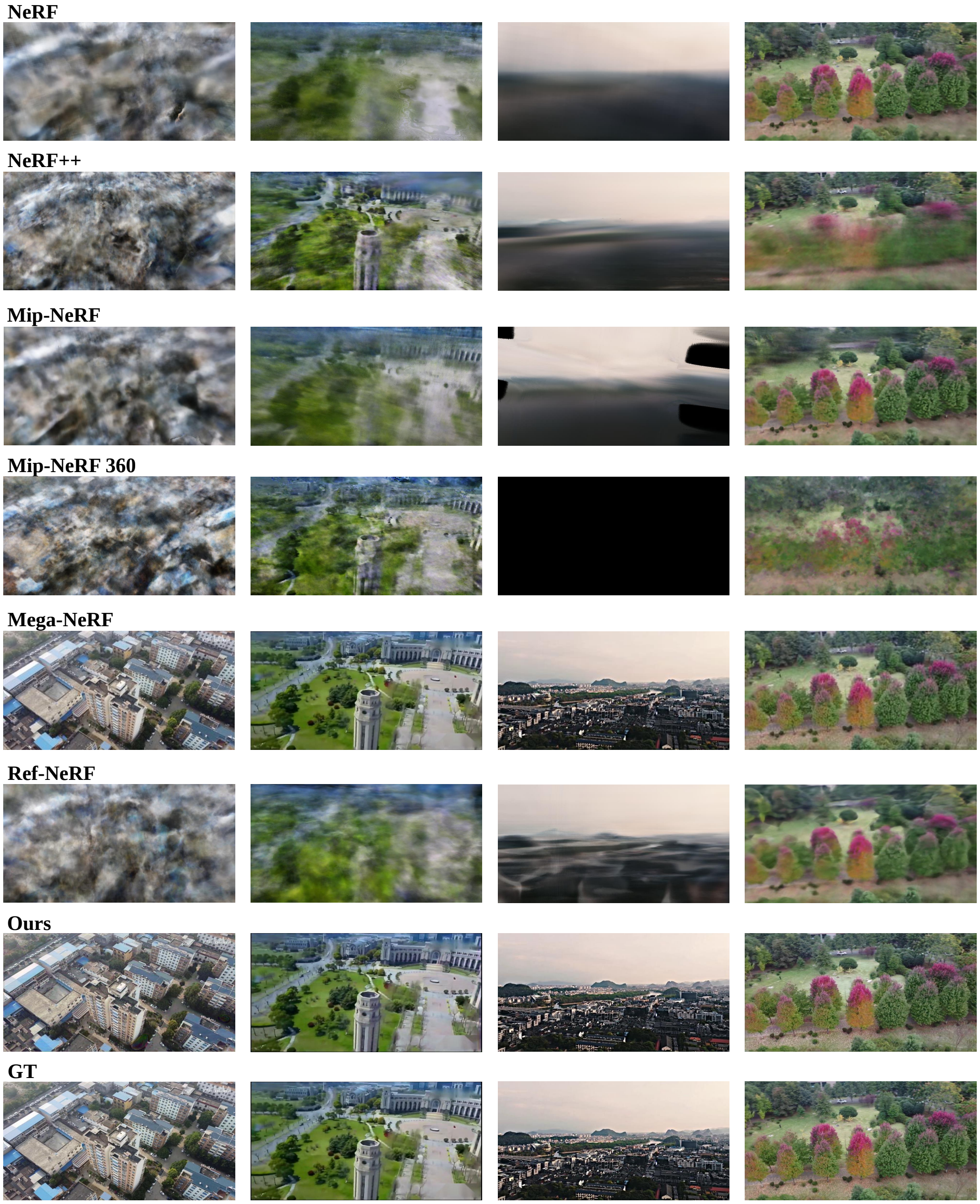
Table 4: Quantitative results of our PDF method with the baseline Mip-NeRF 360 on the first five large-scale outdoor scenes of the BlendedMVS dataset. ↑ means the higher, the better.

B Experiments on the Full OMMO Dateset

In the manuscript, we have reported performance on the representative subset of the OMMO dataset [3]. A more comprehensive evaluation on all scenes from the OMMO dataset is shown in Tab. 5 and Fig. 7. Our method still outperforms other state-of-the-art methods on average and most scenes, and synthesizes photo-realistic images, consistent with results on the representative subset.

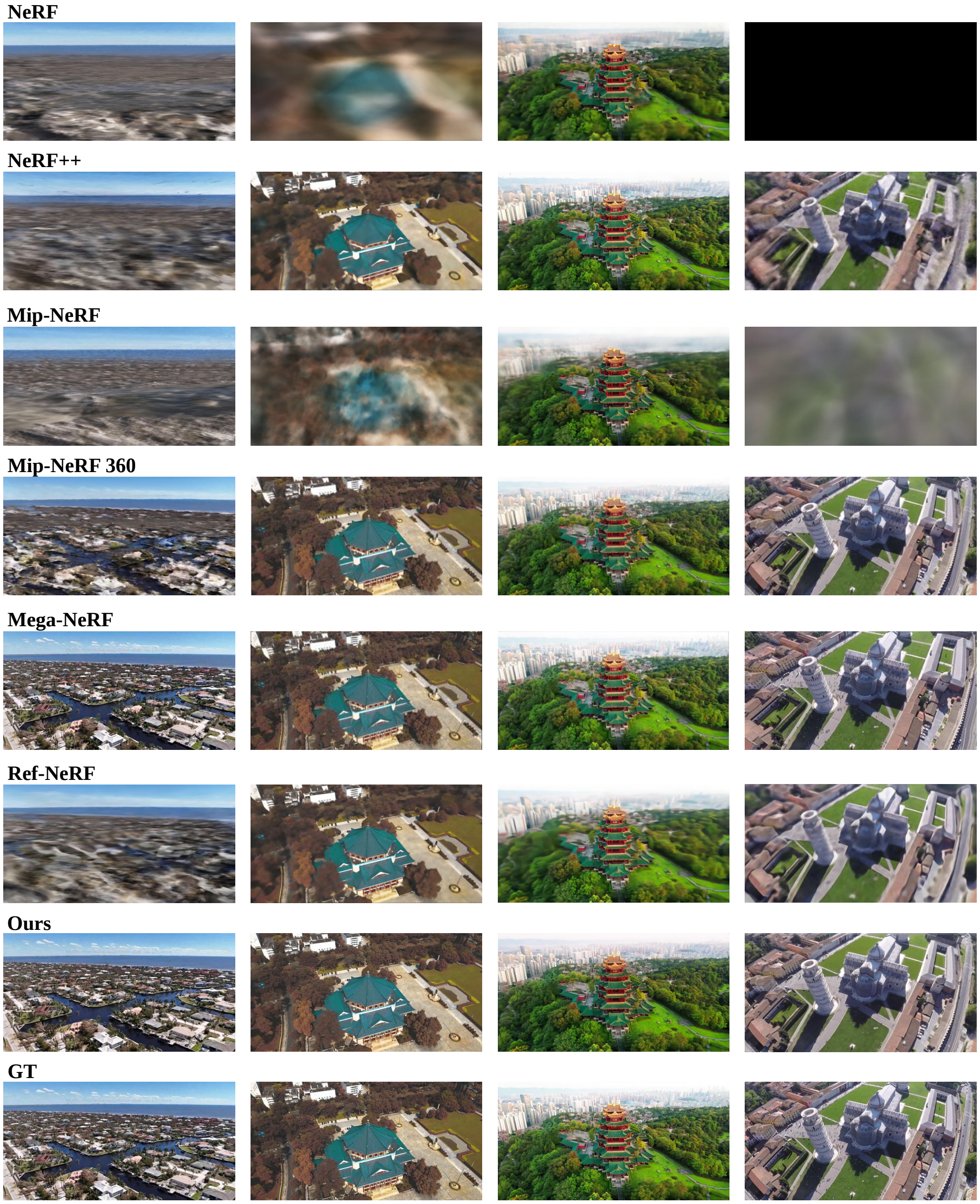
Scene ID	**NeRF**[4]			**NeRF++**[8]			**Mip-NeRF**[1]			**Mip-NeRF 360**[2]			**Mega-NeRF**[5]			**Ref-NeRF**[6]			**Ours**		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1	**16.93**	**0.37**	**0.744**	16.86	0.36	0.780	16.84	**0.37**	0.793	13.91	0.31	0.771	16.12	0.34	0.782	15.10	0.34	0.755	14.80	0.32	0.755
2	15.31	0.44	0.694	14.89	0.47	0.653	15.16	0.40	0.731	15.06	0.44	0.646	15.64	0.47	0.679	15.90	0.49	0.632	**19.63**	**0.62**	**0.374**
3	14.38	0.28	0.556	14.64	0.29	0.547	14.56	0.29	0.533	14.25	0.31	0.526	15.21	0.33	0.517	**15.44**	**0.37**	0.526	14.74	0.34	**0.515**
4	25.39	0.86	0.431	27.47	0.90	0.380	21.78	0.76	0.469	27.68	**0.94**	0.292	23.36	0.86	0.419	27.86	0.91	0.404	**31.74**	**0.94**	**0.202**
5	22.26	0.67	0.531	24.32	0.73	0.450	14.98	0.54	0.633	25.76	0.80	0.317	25.78	0.76	0.436	23.54	0.71	0.491	**27.58**	**0.90**	**0.162**
6	24.09	0.68	0.504	25.59	0.75	0.396	23.18	0.66	0.529	**28.86**	**0.90**	**0.211**	24.92	0.77	0.393	26.07	0.72	0.459	23.69	0.87	0.212
7	5.36	0.17	0.747	21.93	0.71	0.542	15.57	0.64	0.624	23.05	0.73	0.523	22.33	0.69	0.552	**25.79**	0.73	0.511	21.46	**0.81**	**0.193**
8	21.14	0.50	0.594	22.91	0.57	0.509	19.82	0.46	0.638	25.07	0.71	0.354	16.65	0.48	0.431	21.21	0.49	0.606	**27.62**	**0.92**	**0.101**
9	14.92	0.34	0.744	14.57	0.34	0.732	14.58	0.34	0.746	15.40	0.30	0.706	17.32	**0.49**	0.673	**20.34**	0.43	0.649	15.77	**0.49**	**0.381**
10	22.26	0.55	0.626	24.37	0.60	0.578	19.80	0.53	0.643	**26.68**	0.72	0.420	21.78	0.62	0.558	24.23	0.58	0.597	25.74	**0.83**	**0.136**
11	22.36	0.82	0.420	24.61	0.85	0.342	22.81	0.82	0.423	27.06	0.93	0.217	24.37	0.84	0.392	23.81	0.84	0.355	**30.29**	**0.95**	**0.188**
12	22.41	0.59	0.533	24.29	0.68	0.447	22.13	0.60	0.526	28.12	0.83	0.274	21.60	0.62	0.493	23.06	0.60	0.524	**27.92**	**0.86**	**0.063**
13	22.27	0.59	0.608	23.52	0.62	0.581	18.90	0.54	0.673	**26.63**	**0.77**	0.403	25.50	0.72	0.517	23.29	0.61	0.594	25.94	0.74	**0.205**
14	19.85	0.55	0.569	23.89	0.74	0.417	17.06	0.48	0.655	28.06	0.89	0.224	24.42	0.75	0.411	21.76	0.63	0.508	**28.11**	**0.94**	**0.127**
15	20.35	0.53	0.552	21.71	0.61	0.490	19.44	0.49	0.594	28.63	**0.89**	0.179	22.69	0.67	0.445	20.33	0.50	0.576	**27.22**	**0.89**	**0.136**
16	17.86	0.40	0.631	18.75	0.41	0.597	18.49	0.40	0.610	10.01	0.34	0.850	**20.26**	**0.53**	0.509	19.64	0.43	0.572	18.70	0.47	**0.392**
17	22.02	0.57	0.610	24.20	0.67	0.461	17.01	0.53	0.696	**29.53**	0.83	0.247	17.23	0.57	0.529	23.17	0.59	0.529	26.59	**0.88**	**0.111**
18	26.06	0.75	0.428	25.57	0.73	0.461	24.61	0.73	0.469	**28.55**	0.86	0.265	24.76	0.73	0.448	22.79	0.67	0.569	28.07	**0.91**	**0.152**
19	14.20	0.40	0.726	13.86	0.37	0.703	13.84	0.39	0.738	14.72	0.37	0.676	23.81	0.68	0.465	14.34	0.39	0.691	**27.55**	**0.84**	**0.170**
20	22.84	0.61	0.499	23.28	0.64	0.475	22.41	0.60	0.519	**28.33**	**0.86**	0.228	21.11	0.63	0.490	21.54	0.55	0.574	26.88	0.81	**0.197**
21	22.59	0.51	0.532	21.84	0.47	0.593	22.31	0.51	0.537	25.64	0.75	0.344	21.92	0.51	0.578	21.07	0.44	0.672	**28.62**	**0.94**	**0.141**
22	16.53	0.47	0.733	20.66	0.56	0.575	13.37	0.42	0.776	24.79	0.77	0.362	20.84	0.60	0.527	20.31	0.53	0.615	**26.33**	**0.85**	**0.074**
23	18.99	0.41	0.669	19.51	0.42	0.597	18.09	0.39	0.671	21.25	0.51	0.539	20.13	0.44	0.585	19.94	0.41	0.622	**21.64**	**0.65**	**0.206**
24	19.32	0.39	0.696	23.14	0.52	0.535	16.89	0.37	0.715	25.86	0.71	0.373	23.87	0.56	0.518	22.17	0.45	0.616	**30.90**	**0.87**	**0.097**
25	24.72	0.55	0.528	22.42	0.51	0.613	24.24	0.54	0.542	28.91	0.79	0.306	25.98	0.63	0.457	23.62	0.50	0.598	**30.85**	**0.94**	**0.083**
26	8.56	0.24	0.564	19.94	0.59	0.513	13.43	0.35	0.688	14.59	0.46	0.626	19.23	0.67	0.467	21.00	0.62	0.489	**23.88**	**0.83**	**0.311**
27	4.54	0.01	0.705	21.25	0.55	0.546	14.82	0.45	0.674	21.26	0.60	0.235	20.59	0.61	0.543	20.82	0.52	0.590	**21.77**	**0.66**	**0.164**
28	24.48	0.66	0.479	23.28	0.64	0.475	24.76	0.66	0.406	**29.62**	0.87	0.240	25.87	0.72	0.442	22.17	0.45	0.616	29.22	**0.91**	**0.153**
29	22.98	0.61	0.540	23.17	0.62	0.529	23.01	0.61	0.539	25.51	0.74	0.400	21.57	0.61	0.557	21.11	0.54	0.631	**25.86**	**0.84**	**0.174**
30	20.23	0.52	0.605	23.27	0.64	0.476	18.63	0.46	0.675	**26.54**	0.84	0.296	24.04	0.69	0.459	21.62	0.54	0.586	26.10	**0.93**	**0.096**
31	18.97	0.37	0.645	19.05	0.37	0.643	18.91	0.36	0.659	13.08	0.23	0.708	20.93	0.60	0.545	19.18	0.37	0.645	**26.68**	**0.90**	**0.208**
32	17.99	0.58	0.621	18.99	0.61	0.540	11.28	0.42	0.687	17.16	0.57	0.601	21.29	**0.70**	0.475	18.98	0.60	0.565	**23.43**	0.69	**0.142**
33	5.79	0.01	0.745	20.19	0.50	0.597	14.31	0.42	0.755	22.76	0.63	0.457	22.89	0.64	0.478	21.23	0.52	0.578	**22.91**	**0.75**	**0.134**
Mean	18.72	0.48	0.600	21.45	0.58	0.538	18.39	0.50	0.623	23.10	0.67	0.419	21.63	0.62	0.508	21.28	0.55	0.574	**25.10**	**0.79**	**0.205**

Table 5: Quantitative results of our PDF method with the baselines on the full OMMO dataset. ↑ means the higher, the better.



Part 1 / 2

Figure 7: More qualitative visualization results for novel view synthesis (zoom-in for the best view) on the full OMMO dataset.



Part 2 / 2

Figure 7. More qualitative visualization results for novel view synthesis (zoom-in for the best view) on the full OMMO dataset.

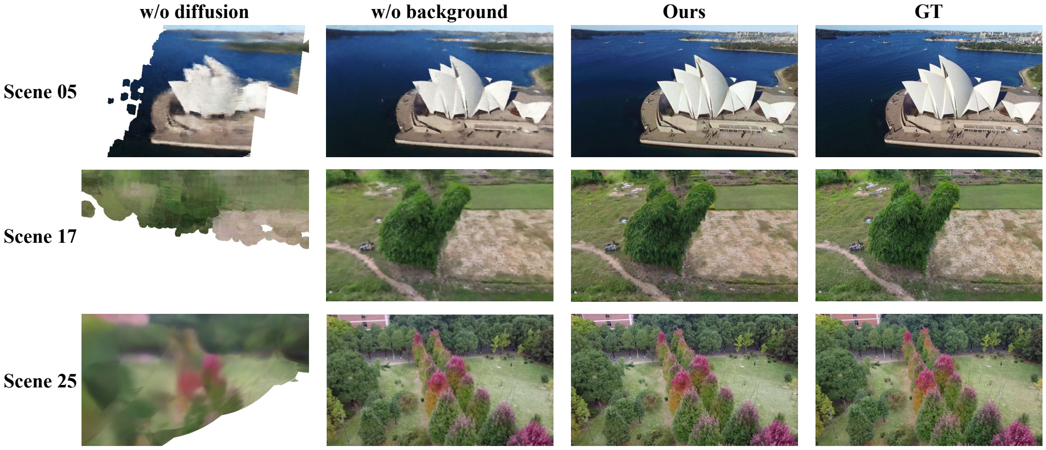


Figure 8: Qualitative ablation performance for more scenes on the OMMO dataset. From left to right: removing the diffusion-based point cloud up-sampling module, removing the background fusion module, our PDF method, and the groundtruth.

C Ablation Study

In order to further demonstrate the effectiveness of each module, we provide ablation experiment results of more scenes on whether to use the point cloud up-sampling diffusion module and whether to use the background fusion module(cf. Fig. 8 and Tab. 6). It can be seen that both the diffusion-based point cloud up-sampling module and the background fusion module play an important role in our method. The former generates dense point cloud surfaces from sparse surface priors and reduces the sampling space; the latter complements background features that point clouds cannot provide.

Table 6: Quantitative ablation performance for more scenes (scene id 05, 17, and 25) on the OMMO dataset, including removing the diffusion-based point cloud up-sampling module, removing the background fusion module, our PDF method.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o diffusion	9.48	0.47	0.325
w/o background	24.77	0.81	0.187
Ours	**28.34**	**0.91**	**0.119**

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[3] Chongshan Lu, Fukun Yin, Xin Chen, Tao Chen, Gang Yu, and Jiayuan Fan. A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. *arXiv preprint arXiv:2301.06782*, 2023.

[4] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[5] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs, 2022.

[6] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.

[7] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

[8] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.