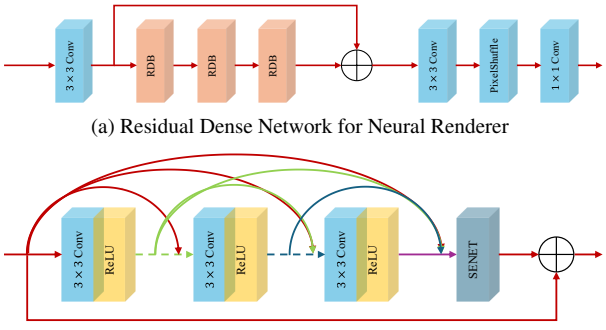
Depth-Guided Bundle Sampling for Efficient Generalizable Neural Radiance Field Reconstruction

Supplementary Material

7. Implementation and Network Details

Implementation Details. Following ENeRF [7] and MVSGaussian [8], we partition the DTU dataset [1] into 88 training scenes and 16 test scenes. The generalizable model is trained on the 88 training scenes and evaluated on the 16 test scenes using an RTX A6000 GPU. Training is performed with the Adam optimizer [5], starting with an initial learning rate of 5×10^{-4}. The batch size is set to 4, and the learning rate is halved every 50k iterations. During training, source views are selected with probabilities of 0.1, 0.8, and 0.1 for 2, 3, and 4 input views, respectively. Evaluation is conducted following the criteria outlined in prior works such as ENeRF [7] and MVSGaussian [8]. Specifically, for the DTU test set, segmentation masks are employed to evaluate performance, defined based on the availability of ground-truth depth at each pixel. For the Real Forward-facing dataset [9], where marginal regions are typically occluded in the input images, evaluations are restricted to the central 80% of the image area. The image resolutions used for evaluation are 512×640 for the DTU dataset [1], 640×960 for the Real Forward-facing dataset [9], and 800×800 for the NeRF Synthetic dataset [10].

Network Details. The mipmap query process is a well-optimized technique in the rendering community for efficient texture sampling. To leverage this, we employ the nvdiffrast library [6] to implement our bundle sampling with high efficiency. The radiance field prediction network described in Eq. (7) in Section 4.3 of the main text is adapted from ENeRF [7], with modifications to the feature dimensions of the samples. The neural renderer comprises three residual dense blocks [14] integrated with SENet [4], as illustrated in Fig. 6. Each block employs a hidden dimension of 64 and a growth rate of 32. The ray-specific representation is unfolded using a pixel shuffle operation. During training, the final target view is synthesized by fusing two intermediate views, calculated as: $\hat{\boldsymbol{I}}_{\text{tar}} = \hat{\boldsymbol{I}}_c + \hat{\boldsymbol{I}}_f$. However, for the Real Forward-facing [9] and NeRF Synthetic [10] datasets—domains that differ significantly from the DTU dataset—adjusting the weights of the intermediate views improves performance. Specifically, during evaluation on these datasets, the final view is generated as: $\hat{\boldsymbol{I}}_{\text{tar}} = 0.5\hat{\boldsymbol{I}}_c + 1.5\hat{\boldsymbol{I}}_f$. This weighted fusion strategy effectively adapts the model to cross-domain scenarios, enhancing rendering performance.



(a) Residual Dense Network for Neural Renderer

(b) Residual Dense Block

Figure 6. Neural renderer is a (a) residual dense network to decode joint bundle representation into colors, which consists of 3 (b) residual dense blocks.

8. Additional Ablation Experiments

To demonstrate the effectiveness of the proposed bundle sampling strategy, we apply a 1×1 bundle size to ENeRF, which corresponds to casting each cone through a single pixel. This setup disables bundle sampling while preserving all other components of the method. The results on the DTU test set are presented in Tab. 4, while cross-dataset results on the Real Forward-facing and NeRF Synthetic datasets are shown in Tab. 5.

As observed, the quality difference between 1×1 and 2×2 bundle sizes is quite small, demonstrating that our bundle sampling strategy effectively reduces the number of samples without compromising rendering quality. Notably, in the natural scenes of Real Forward-facing dataset, a 4×4 bundle size achieves results comparable to those of smaller bundle sizes. This further validates the efficiency of our method, particularly for scenes characterized by piece-wise smoothness.

9. Per-Scene Optimization Results

For per-scene optimization, we follow the setting of ENeRF [7], selecting 3 and 4 source views with probabilities of 0.4 and 0.6, respectively. The batch size is set to 1. For efficiency, the bundle size is fixed at 2×2 for ENeRF+Ours. The generalizable model is fine-tuned on each scene for 10 epochs (less than 1 hour). For evaluation, 4 input views are used.

The per-scene optimization results on the DTU, Real Forward-facing and NeRF Synthetic datasets are presented

Methods	3-view					2-view		
	PSNR ↑	SSIM ↑	LPIPS ↓	Avg. samples per ray	FPS ↑	PSNR ↑	SSIM ↑	LPIPS ↓
ENeRF+Ours (1×1)	<u>28.75</u>	**0.966**	**0.069**	1.63	15.3	<u>26.37</u>	**0.951**	**0.088**
ENeRF+Ours (2×2)	**28.86**	<u>0.964</u>	<u>0.073</u>	0.42	<u>28.6</u>	**26.39**	<u>0.949</u>	<u>0.089</u>
ENeRF+Ours (4×4)	28.21	0.957	0.088	0.10	**43.6**	26.09	0.942	0.105

Table 4. Ablation on bundle size on the DTU testset [1]. The best result is highlighted in bold, while the second-best is underlined.

Methods	Settings	Real Forward-facing			NeRF Synthetic		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ENeRF+Ours (1×1)	3-view	<u>24.20</u>	**0.860**	<u>0.164</u>	**26.62**	**0.948**	<u>0.076</u>
ENeRF+Ours (2×2)		**24.33**	**0.860**	**0.162**	<u>26.48</u>	**0.948**	**0.075**
ENeRF+Ours (4×4)		23.84	<u>0.850</u>	0.174	26.12	<u>0.943</u>	0.104
ENeRF+Ours (1×1)	2-view	**23.14**	<u>0.825</u>	<u>0.187</u>	<u>24.98</u>	<u>0.936</u>	<u>0.097</u>
ENeRF+Ours (2×2)		<u>23.06</u>	0.824	**0.186**	**25.01**	**0.937**	**0.087**
ENeRF+Ours (4×4)		<u>23.06</u>	**0.826**	<u>0.187</u>	24.65	0.932	0.098

Table 5. Ablation on bundle size on Real Forward-facing [9] and NeRF Synthetic [10]. The best result is highlighted in bold, while the second-best is underlined.

Methods	DTU			Real Forward-facing			NeRF Synthetic		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF$_{10.2h}$ [10]	27.01	0.902	0.263	25.97	0.870	0.236	**30.63**	**0.962**	0.093
IBRNet$_{ft-1.0h}$ [12]	**31.35**	<u>0.956</u>	0.131	24.88	0.861	0.189	25.62	0.939	0.111
MVSNeRF$_{ft-15min}$ [2]	28.51	0.933	0.179	25.45	0.877	0.192	27.07	0.931	0.168
ENeRF$_{ft-1.0h}$ [7]	28.87	0.957	0.090	24.89	0.865	0.159	27.57	0.954	0.063
MVSGaussian$_{ft-1.0h}$ [8]	-	-	-	25.92	0.891	0.135	27.87	0.956	<u>0.061</u>
ENeRF+Ours$_{ft-15min}$	29.57	**0.962**	<u>0.079</u>	<u>26.29</u>	<u>0.894</u>	<u>0.125</u>	28.07	0.957	0.062
ENeRF+Ours$_{ft-1.0h}$	<u>29.78</u>	**0.962**	**0.077**	**26.75**	**0.904**	**0.113**	<u>28.39</u>	<u>0.958</u>	**0.060**

Table 6. Per-scene optimization results on DTU [1], Real Forward-facing [9] and NeRF Synthetic [10] datasets. The best result is highlighted in bold, while the second-best is underlined.

in Tab. 6. The results demonstrate that our method outperforms other approaches on most scenes, particularly in natural scenes of the Real Forward-facing dataset. On the NeRF Synthetic dataset, our method achieves second-best performance in terms of PSNR and SSIM, trailing only NeRF [10]. However, our method significantly outperforms NeRF in LPIPS, indicating superior perceptual quality. Furthermore, our approach achieves this performance while requiring only 1/10 of NeRF's training time, highlighting its efficiency.

10. Per-Scene Breakdown Results

The per-scene breakdown results are presented in Tab. 7 and Tab. 8 for the DTU dataset, Tab. 9 for the Real Forward-facing dataset, and Tab. 10 for the NeRF Synthetic dataset. For efficiency, the bundle size is fixed at 2×2 for ENeRF+Ours. These detailed results are consistent with the averaged results reported in the main text.

As shown in Tab. 7, Tab. 8 and Tab. 10, ENeRF+Ours achieves the best generalizable performance on most scenes. For per-scene optimization results, ENeRF+Ours ranks second only to NeRF [10] in terms of PSNR, while achieving the best performance in terms of SSIM and LPIPS. A likely explanation is that our loss function incorporates SSIM loss and perceptual loss in addition to MSE. This combination enables better preservation of high-frequency details in the images while enhancing perceptual quality.

Tab. 9 demonstrates that ENeRF+Ours consistently ranks either first or second in both generalizable and per-scene optimization results across all scenes, ranking first on most. This further validates the effectiveness of our method, particularly in natural scenes characterized by piece-wise smoothness.

11. Limitations

Our method leverages depth information to guide sampling, primarily relying on MVS depth estimation. However, its effectiveness can be compromised in scenes where MVS performance is poor. For instance, in the Ficus and Materi-

Scan	#1	#8	#21	#103	#114
Metric			PSNR		
PixelNeRF [13]	21.64	23.70	16.04	16.76	18.40
IBRNet [12]	25.97	27.45	20.94	27.91	27.91
MVSNeRF [2]	26.96	27.43	21.55	29.25	27.99
ENeRF [7]	28.85	29.05	22.53	30.51	28.86
MatchNeRF [3]	27.69	27.76	22.75	29.35	28.16
GNT [11]	27.25	28.12	21.67	28.45	28.01
MVSGaussian [8]	**29.67**	29.65	23.24	30.60	29.26
ENeRF+Ours	29.62	**30.42**	**23.40**	**32.28**	**30.11**
NeRF$_{10.2h}$ [10]	26.62	28.33	23.24	30.40	26.47
IBRNet$_{ft-1.0h}$ [12]	**31.00**	**32.46**	**27.88**	**34.40**	**31.00**
MVSNeRF$_{ft-15min}$ [2]	28.05	28.88	24.87	32.23	28.47
ENeRF$_{ft-1.0h}$ [7]	30.10	30.50	22.46	31.42	29.87
ENeRF+Ours$_{ft-15min}$	30.11	31.09	23.51	32.79	30.34
ENeRF+Ours$_{ft-1.0h}$	30.80	31.33	23.59	32.85	30.35
Metric			SSIM		
PixelNeRF [13]	0.827	0.829	0.691	0.836	0.763
IBRNet [12]	0.918	0.903	0.873	0.950	0.943
MVSNeRF [2]	0.937	0.922	0.890	0.962	0.949
ENeRF [7]	0.958	0.955	0.916	0.968	0.961
MatchNeRF [3]	0.936	0.918	0.901	0.961	0.948
GNT [11]	0.922	0.931	0.881	0.942	0.960
MVSGaussian [8]	**0.966**	**0.961**	0.930	0.970	0.963
ENeRF+Ours	0.965	0.960	**0.939**	**0.974**	**0.966**
NeRF$_{10.2h}$ [10]	0.902	0.876	0.874	0.944	0.913
IBRNet$_{ft-1.0h}$ [12]	0.955	0.945	**0.947**	0.968	0.964
MVSNeRF$_{ft-15min}$ [2]	0.934	0.900	0.922	0.964	0.945
ENeRF$_{ft-1.0h}$ [7]	0.966	0.959	0.924	0.971	0.965
ENeRF+Ours$_{ft-15min}$	0.967	0.961	0.939	**0.975**	**0.967**
ENeRF+Ours$_{ft-1.0h}$	**0.968**	**0.962**	0.940	**0.975**	**0.967**
Metric			LPIPS		
PixelNeRF [13]	0.373	0.384	0.407	0.376	0.372
IBRNet [12]	0.190	0.252	0.179	0.195	0.136
MVSNeRF [2]	0.155	0.220	0.166	0.165	0.135
ENeRF [7]	0.086	0.119	0.107	0.107	0.076
MatchNeRF [3]	0.157	0.227	0.149	0.179	0.132
GNT [11]	0.143	0.210	0.171	0.149	0.139
MVSGaussian [8]	**0.069**	**0.102**	0.088	0.098	**0.070**
ENeRF+Ours	0.070	0.105	**0.076**	**0.091**	0.071
NeRF$_{10.2h}$ [10]	0.265	0.321	0.246	0.256	0.225
IBRNet$_{ft-1.0h}$ [12]	0.129	0.170	0.104	0.156	0.099
MVSNeRF$_{ft-15min}$ [2]	0.171	0.261	0.142	0.170	0.153
ENeRF$_{ft-1.0h}$ [7]	0.071	0.106	0.097	0.102	0.074
ENeRF+Ours$_{ft-15min}$	0.066	0.097	**0.074**	0.086	0.066
ENeRF+Ours$_{ft-1.0h}$	**0.064**	**0.096**	**0.074**	**0.084**	**0.065**

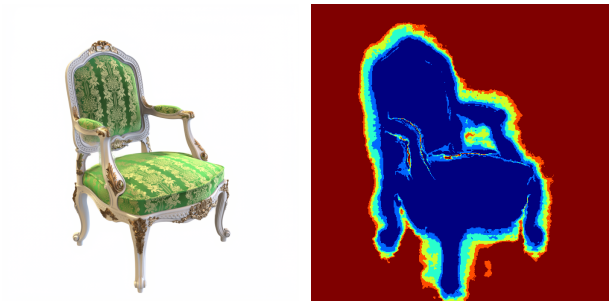
Table 7. Quantitative results of five sample scenes on the DTU [1] test set. The best result is highlighted in bold, while the second-best is underlined.

als scenes from the NeRF Synthetic dataset, challenges such as occlusion, thin structures, reflections, or glossy surfaces result in poor MVS depth estimation. Consequently, our generalizable results are not as strong as GNT [11], which

Scan	#30	#31	#34	#38	#40	#41	#45	#55	#63	#82	#110
Metric	PSNR										
ENeRF [7]	29.20	25.13	26.77	28.61	25.67	29.51	24.83	30.26	27.22	26.83	27.97
MatchNeRF [3]	29.16	24.26	25.66	27.52	25.16	28.27	23.94	26.64	**29.40**	27.65	27.15
GNT [11]	27.13	23.54	25.10	27.67	24.48	28.10	24.54	28.86	26.36	26.09	26.93
MVSGaussian [8]	30.10	25.94	26.82	29.27	26.13	30.33	24.55	31.40	28.46	27.82	28.15
ENeRF+Ours	**30.89**	**26.76**	**27.81**	**30.10**	**27.16**	**30.84**	**25.95**	**31.63**	29.39	**29.55**	**29.47**
Metric	SSIM										
ENeRF [7]	0.981	0.937	0.934	0.946	0.947	0.960	0.948	0.973	0.978	0.971	0.974
MatchNeRF [3]	0.974	0.921	0.874	0.902	0.903	0.936	0.934	0.929	0.976	0.966	0.962
GNT [11]	0.954	0.907	0.880	0.921	0.893	0.908	0.918	0.934	0.938	0.949	0.930
MVSGaussian [8]	0.983	0.946	**0.947**	0.954	**0.957**	**0.967**	0.954	**0.979**	0.980	0.974	0.976
ENeRF+Ours	**0.985**	**0.953**	0.944	**0.955**	0.954	0.966	**0.959**	**0.979**	**0.984**	**0.979**	**0.979**
Metric	LPIPS										
ENeRF [7]	0.052	0.108	0.117	0.118	0.120	0.091	0.077	0.069	0.048	0.066	0.069
MatchNeRF [3]	0.085	0.169	0.234	0.220	0.216	0.174	0.127	0.164	0.077	0.093	0.141
GNT [11]	0.110	0.172	0.201	0.231	0.116	0.168	0.134	0.155	0.127	0.138	0.127
MVSGaussian [8]	0.048	0.093	**0.097**	0.098	**0.101**	**0.075**	0.067	0.055	0.041	0.057	**0.057**
ENeRF+Ours	**0.045**	**0.083**	0.103	**0.095**	**0.101**	**0.075**	**0.057**	**0.054**	**0.035**	**0.048**	0.060

Table 8. Quantitative results of other eleven scenes on the DTU [1] test set. The best result is highlighted in bold, while the second-best is underlined.

is based on an attention mechanism, and our per-scene optimization results fall short of NeRF [10] due to fewer samples that cannot be accurately placed near surfaces.



(a) Target View (b) Sample Distribution

Figure 7. Visualization of sample allocation in depth-guided adaptive sampling.

Additionally, the efficiency of our method faces challenges on object-centric datasets, such as the NeRF Synthetic dataset or certain human datasets. In these cases, the lack of background hinders depth estimation, leading our method to allocate the maximum number of samples in these regions. This misallocation reduces the overall efficiency of our approach, as illustrated in Fig. 7. However, this issue can be mitigated by employing a mask to exclude these areas from sampling.

For MVSGaussian+Ours, our method is applied to the NeRF branch of MVSGaussian. Since the NeRF branch in MVSGaussian uses only a single sample per ray, the speed improvement achieved by our approach is relatively small. One reason our bundle sampling enhances rendering quality is its ability to exploit correlations between adjacent rays. However, in MVSGaussian, a post-processing network is employed to leverage the context between adjacent 3D-GS elements, which partially diminishes the effectiveness of our method. Exploring more effective ways to integrate our approach with 3D-GS will be the focus of our future research.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 11, 12, 13, 14

[2] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 12, 13, 15, 16

[3] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023. 13, 14, 15, 16

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 11

[5] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 11

[6] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for

Scene	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	Trex
Metric	PSNR							
PixelNeRF [13]	12.40	10.00	14.07	11.07	9.85	9.62	11.75	10.55
IBRNet [12]	20.83	22.38	27.67	22.06	18.75	15.29	27.26	20.06
MVSNeRF [2]	21.15	24.74	26.03	23.57	17.51	17.85	26.95	23.20
ENeRF [7]	21.92	24.28	30.43	24.49	19.01	17.94	29.75	21.21
MatchNeRF [3]	20.98	23.97	27.44	23.14	18.62	18.07	26.77	20.47
GNT [11]	22.21	23.56	29.16	22.80	19.18	17.43	29.35	20.15
MVSGaussian [8]	22.45	**25.66**	30.46	24.70	19.81	17.86	29.86	21.75
ENeRF+Ours	**22.55**	25.44	**31.30**	**26.45**	**19.88**	**18.96**	**30.16**	**23.80**
$\text{NeRF}_{10.2h}$ [10]	**23.87**	26.84	31.37	25.96	21.21	19.81	**33.54**	25.19
$\text{IBRNet}_{ft-1.0h}$ [12]	22.64	26.55	30.34	25.01	22.07	19.01	31.05	22.34
$\text{MVSNeRF}_{ft-15min}$ [2]	23.10	27.23	30.43	26.35	21.54	20.51	30.12	24.32
$\text{ENeRF}_{ft-1.0h}$ [7]	22.08	27.74	29.58	25.50	21.26	19.50	30.07	23.39
$\text{ENeRF+Ours}_{ft-15min}$	23.65	27.63	31.68	27.62	22.46	19.02	32.85	25.37
$\text{ENeRF+Ours}_{ft-1.0h}$	23.81	**28.07**	**31.69**	**27.83**	**22.72**	**20.56**	33.51	**25.82**
Metric	SSIM							
PixelNeRF [13]	0.531	0.433	0.674	0.516	0.268	0.317	0.691	0.458
IBRNet [12]	0.710	0.854	0.894	0.840	0.705	0.571	0.950	0.768
MVSNeRF [2]	0.638	0.888	0.872	0.868	0.667	0.657	0.951	0.868
ENeRF [7]	0.774	0.893	0.948	0.905	0.744	0.681	0.971	0.826
MatchNeRF [3]	0.726	0.861	0.906	0.870	0.690	0.675	0.949	0.767
GNT [11]	0.736	0.791	0.867	0.820	0.650	0.538	0.945	0.744
MVSGaussian [8]	0.792	0.908	0.948	0.913	**0.784**	0.701	0.973	0.841
ENeRF+Ours	**0.802**	**0.911**	**0.960**	**0.934**	0.766	**0.738**	**0.977**	**0.887**
$\text{NeRF}_{10.2h}$ [10]	**0.828**	0.897	0.945	0.900	0.792	0.721	0.978	0.899
$\text{IBRNet}_{ft-1.0h}$ [12]	0.774	0.909	0.937	0.904	0.843	0.705	0.972	0.842
$\text{MVSNeRF}_{ft-15min}$ [2]	0.795	0.912	0.943	0.917	0.826	0.732	0.966	0.895
$\text{ENeRF}_{ft-1.0h}$ [7]	0.770	0.923	0.940	0.904	0.827	0.725	0.965	0.869
$\text{ENeRF+Ours}_{ft-15min}$	0.823	0.928	**0.964**	0.947	0.868	0.727	0.984	0.912
$\text{ENeRF+Ours}_{ft-1.0h}$	0.826	**0.932**	0.963	**0.949**	**0.874**	**0.782**	**0.985**	**0.919**
Metric	LPIPS							
PixelNeRF [13]	0.650	0.708	0.608	0.705	0.695	0.721	0.611	0.667
IBRNet [12]	0.349	0.224	0.196	0.285	0.292	0.413	0.161	0.314
MVSNeRF [2]	0.238	0.196	0.208	0.237	0.313	0.274	0.172	0.184
ENeRF [7]	0.224	0.164	0.092	0.161	0.216	0.289	0.120	0.192
MatchNeRF [3]	0.285	0.202	0.169	0.234	0.277	0.325	0.167	0.294
GNT [11]	0.223	0.203	0.157	0.208	0.255	0.341	0.103	0.275
MVSGaussian [8]	**0.193**	**0.133**	0.096	0.148	**0.189**	0.275	0.104	0.177
ENeRF+Ours	0.195	0.136	**0.086**	**0.125**	0.225	**0.254**	**0.095**	**0.148**
$\text{NeRF}_{10.2h}$ [10]	0.291	0.176	0.147	0.247	0.301	0.321	0.157	0.245
$\text{IBRNet}_{ft-1.0h}$ [12]	0.266	0.146	0.133	0.190	0.180	0.286	0.089	0.222
$\text{MVSNeRF}_{ft-15min}$ [2]	0.253	0.143	0.134	0.188	0.222	0.258	0.149	0.187
$\text{ENeRF}_{ft-1.0h}$ [7]	0.197	0.121	0.101	0.155	0.168	0.247	0.113	0.169
$\text{ENeRF+Ours}_{ft-15min}$	0.164	0.099	0.069	0.097	0.135	0.258	0.059	0.115
$\text{ENeRF+Ours}_{ft-1.0h}$	**0.156**	**0.093**	**0.067**	**0.093**	**0.126**	**0.201**	**0.055**	**0.109**

Table 9. Quantitative results of other eleven scenes on the Real Forward-facing dataset [9]. The best result is highlighted in bold, while the second-best is underlined.

high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020. 11

[7] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 11, 12, 13,

Scene	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
Metric	PSNR							
PixelNeRF [13]	7.18	8.15	6.61	6.80	7.74	7.61	7.71	7.30
IBRNet [12]	24.20	18.63	21.59	27.70	22.01	20.91	22.10	22.36
MVSNeRF [2]	23.35	20.71	21.98	28.44	23.18	20.05	22.62	23.35
ENeRF [7]	28.29	21.71	23.83	34.20	24.97	24.01	26.62	25.73
MatchNeRF [3]	25.23	19.97	22.72	24.19	23.77	23.12	24.46	22.11
GNT [11]	27.98	20.27	**26.86**	29.34	23.17	**30.75**	23.19	24.86
MVSGaussian [8]	28.93	**22.20**	23.55	**35.01**	24.97	24.49	26.80	25.75
ENeRF+Ours	**29.56**	21.47	23.86	34.42	**25.55**	24.25	**27.15**	**26.33**
NeRF$_{10.2h}$ [10]	**31.07**	**25.46**	**29.73**	34.63	**32.66**	**30.22**	**31.81**	**29.49**
IBRNet$_{ft-1.0h}$ [12]	28.18	21.93	25.01	31.48	25.34	24.27	27.29	21.48
MVSNeRF$_{ft-15min}$ [2]	26.80	22.48	26.24	32.65	26.62	25.28	29.78	26.73
ENeRF$_{ft-1.0h}$ [7]	28.94	25.33	24.71	35.63	25.39	24.98	29.25	26.36
ENeRF+Ours$_{ft-15min}$	29.93	23.07	25.38	36.53	27.01	26.56	28.81	27.26
ENeRF+Ours$_{ft-1.0h}$	30.32	23.27	25.50	**36.75**	27.42	27.08	29.42	27.37
Metric	SSIM							
PixelNeRF [13]	0.624	0.670	0.669	0.669	0.671	0.644	0.729	0.584
IBRNet [12]	0.888	0.836	0.881	0.923	0.874	0.872	0.927	0.794
MVSNeRF [2]	0.876	0.886	0.898	0.962	0.902	0.893	0.923	0.886
ENeRF [7]	0.965	0.918	0.932	0.981	0.948	0.937	0.969	0.891
MatchNeRF [3]	0.908	0.868	0.897	0.943	0.903	0.908	0.947	0.806
GNT [11]	0.935	0.891	**0.941**	0.940	0.897	**0.974**	0.791	0.874
MVSGaussian [8]	0.969	**0.927**	0.935	**0.984**	**0.953**	0.946	**0.974**	0.895
ENeRF+Ours	**0.972**	0.922	0.937	0.983	**0.953**	0.942	0.972	**0.901**
NeRF$_{10.2h}$ [10]	0.971	0.943	**0.969**	0.980	**0.975**	**0.968**	0.981	**0.908**
IBRNet$_{ft-1.0h}$ [12]	0.955	0.913	0.940	0.978	0.940	0.937	0.974	0.877
MVSNeRF$_{ft-15min}$ [2]	0.934	0.898	0.944	0.971	0.924	0.927	0.970	0.879
ENeRF$_{ft-1.0h}$ [7]	0.971	**0.960**	0.939	0.985	0.949	0.947	**0.985**	0.893
ENeRF+Ours$_{ft-15min}$	0.974	0.936	0.948	0.987	0.961	0.956	0.983	0.907
ENeRF+Ours$_{ft-1.0h}$	**0.976**	0.938	0.950	**0.988**	0.963	0.959	**0.985**	**0.908**
Metric	LPIPS							
PixelNeRF [13]	0.386	0.421	0.335	0.433	0.427	0.432	0.329	0.526
IBRNet [12]	0.144	0.241	0.159	0.175	0.202	0.164	0.103	0.369
MVSNeRF [2]	0.282	0.187	0.211	0.173	0.204	0.216	0.177	0.244
ENeRF [7]	0.055	0.110	0.076	0.059	0.075	0.084	0.039	0.183
MatchNeRF [3]	0.107	0.185	0.117	0.162	0.160	0.119	0.060	0.398
GNT [11]	0.065	0.116	**0.063**	0.095	0.112	**0.025**	0.243	**0.115**
MVSGaussian [8]	0.036	**0.091**	0.069	**0.040**	**0.066**	0.063	**0.027**	0.179
ENeRF+Ours	**0.034**	0.104	0.073	0.051	0.072	0.075	0.029	0.184
NeRF$_{10.2h}$ [10]	0.055	0.101	**0.047**	0.089	0.054	0.105	0.033	0.263
IBRNet$_{ft-1.0h}$ [12]	0.079	0.133	0.082	0.093	0.105	0.093	0.040	0.257
MVSNeRF$_{ft-15min}$ [2]	0.129	0.197	0.171	0.094	0.176	0.167	0.117	0.294
ENeRF$_{ft-1.0h}$ [7]	0.030	**0.045**	0.071	0.028	0.070	0.059	**0.017**	0.183
ENeRF+Ours$_{ft-15min}$	0.028	0.082	0.061	0.027	0.053	0.048	0.019	**0.175**
ENeRF+Ours$_{ft-1.0h}$	**0.026**	0.077	0.059	**0.025**	**0.048**	**0.044**	0.018	0.179

Table 10. Quantitative results of other eleven scenes on the NeRF Synthetic [10] test set. The best result is highlighted in bold, while the second-best is underlined.

14, 15, 16

[8] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference on Computer Vision*, pages 37–53. Springer, 2024. 11, 12, 13, 14, 15, 16

[9] 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)*, 38(4):1–14, 2019. 11, 12, 15

[10] 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. 11, 12, 13, 14, 15, 16

[11] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all that nerf needs? *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 13, 14, 15, 16

[12] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 12, 13, 15, 16

[13] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 13, 15, 16

[14] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 11