Supplementary Material:
Efficient Neural Radiance Fields with Learned Depth-Guided Sampling

Haotong Lin[*] Sida Peng[*] Zhen Xu Hujun Bao Xiaowei Zhou
State Key Lab of CAD&CG, Zhejiang University

In the supplementary material, we provide network architectures, details of experimental setup, and more experimental results.

1. Network architectures

Pooling operator. Given the multi-view point features $\{f_i\}_{i=1}^N$, the pooling operator ψ aims to aggregate these features to obtain the feature f_{img}, which is used to infer the radiance field. Instead of simply concatenating these features like MVSNeRF [1], we use a weighted pooling operator proposed in IBRNet [5], which allows us to input any number of source views. Specifically, we first compute a per-element mean $\boldsymbol{\mu}$ and variance $\mathbf{v}$ of $\{f_i\}_{i=1}^N$ to capture global information. Then we concatenate each feature f_i with $\boldsymbol{\mu}$ and $\mathbf{v}$, and feed the concatenated feature into a small shared MLP to obtain a weight w_i. The feature f_{img} is blended via a soft-argmax operator using weights $\{w_i\}_{i=1}^N$ and multi-view features $\{f_i\}_{i=1}^N$.

Architectures of MLPs. The MLP ϕ is used to infer the density σ from the image feature f_{img} and the voxel feature f_{voxel}. To predict the color of the point, we use the MLP φ to yield the blending weights for image colors in the source views. We illustrate the architectures of ϕ and φ in Table 1.

MLP	Layer	Chns.	Input	Output
ϕ	LR_0	8 + 16 / 128	f_{img}, f_{voxel}	hidden feature
	LR_i	128 / 128	hidden feature	hidden feature
	LR_3	128 / 64 + 1	hidden feature	f_p, σ
φ	LR_0	64 + 16 + 4 / 128	f_p, f_i, $\Delta\mathbf{d}_i$	hidden feature
	LR_1	128 / 64	hidden feature	hidden feature
	LR_2	64 / 1	hidden feature	w_i

Table 1. **The architectures of MLPs ϕ and φ.** We denote LR to be LinearRelu layer. "Chns." shows the number of input and output channels for each layer.

2. Details of the experimental setup

Evaluation details. Our evaluation setup is taken from MVSNeRF [1] and is described as the following. To report the results on the DTU [2] dataset, we compute the metric score of foreground part in images. For metrics of SSIM and LPIPS, we set the background to black and calculate the metric score of the whole image. The segmentation mask is defined by whether there is ground-truth depth available at each pixel. Since marginal regions of images are usually invisible to input images on the Real Forward-facing [3] dataset, we only evaluate 80% area in the center of images. The image resolutions are set to 512 × 640, 640 × 960 and 800 × 800 on the DTU, Real forward-facing and NeRF Synthetic [3] datasets, respectively.

Test frame id	1	2	3	4	Mean
Per-frame training	36.73	33.07	34.46	33.24	34.38
Per-sequence training	37.68	35.55	36.33	35.30	36.22

Table 2. **Image synthesis results on 4 frames of the ZJU-MoCap [4] dataset in terms of PSNR metric.** "Per-frame training" means we simply fine-tune the pre-trained network on the test frame. "Per-sequence training" means we fine-tune the pre-trained network on the whole sequence (1200 frames).

Experimental details of PlenOctree. We convert vanilla trained NeRF models to PlenOctree models following the suggestion by [6]. Training a NeRF model on one frame of the ZJU-MoCap [4] dataset takes about 2.5 hours. It takes about 1.45 hours to convert the trained NeRF model to the PlenOctree model. After converting, we optimize the PlenOctree model using the view synthesis loss with SGD optimizer as suggested by [6]. The optimization process takes about 0.1 hours.

3. Visual results

Depth results. As shown in Figure 1, the proposed method produces reasonable depth results by supervising networks with only images. The cost volume recovers high-quality depth, which allows us to place few samples around surfaces to achieve photorealistic rendering.

[*] Authors contributed equally

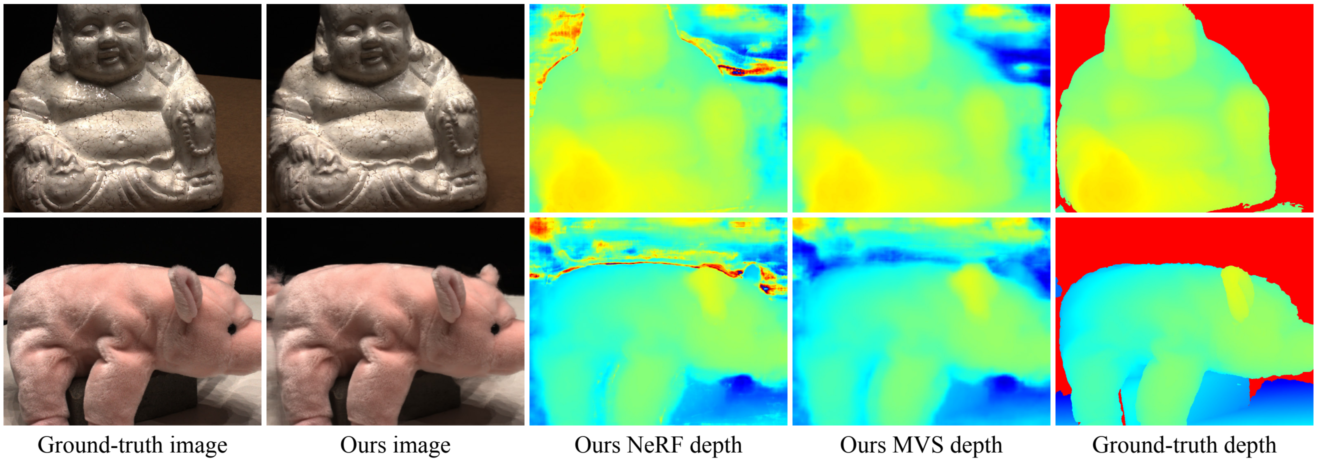

Figure 1. **Visual depth results on the DTU [2] dataset.** "Ours NeRF depth" represents the depth results recovered from volume densities. "Ours MVS depth" denotes the depth results from the fine-level cost volume.

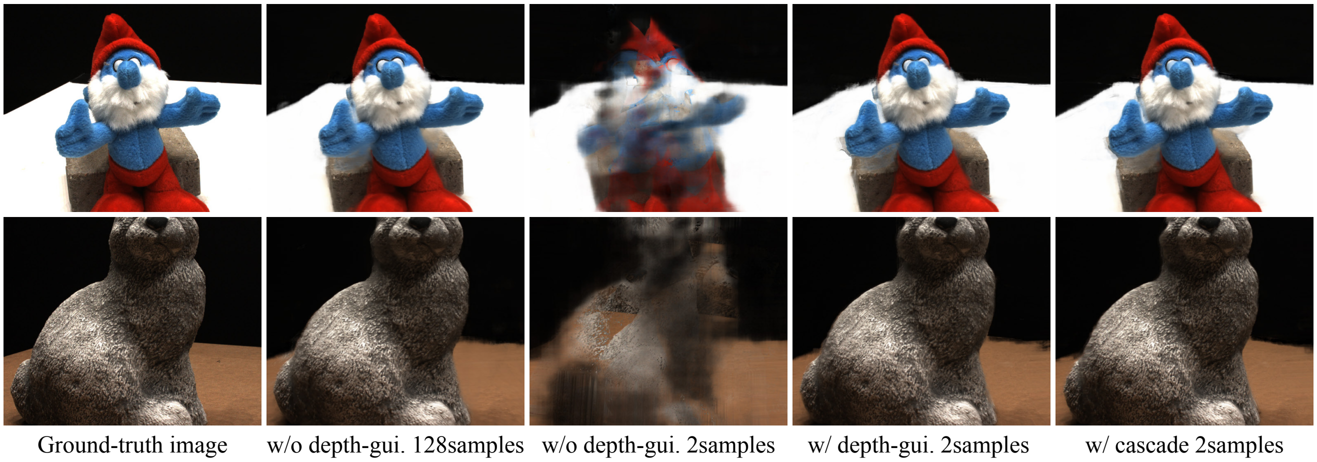

Figure 2. **Visual ablation results on the DTU [2] dataset.** "w/o depth-gui." is similar to MVSNeRF [1].

Ablation results. We provide visual ablation results in Figure 2. The results show that when we reduce the number of samples from 128 to 2, our method with depth-guided sampling almost maintains the same rendering quality. With the depth-guided sampling, the construction of a high-resolution cost volume becomes a bottleneck in the rendering speed. The cascade cost volume further speeds up the construction of the cost volume without loss of rendering quality as shown in Figure 2.

4. Integrating information across video frames

We observe that training on a sequence produces higher rendering quality on the test frame compared to training on one frame of the sequence as shown in Table 2. This indicates that our method is able to integrate observations across video frames to produce higher quality images.

5. Per-scene breakdown

Tables 3, 4 and 5 present the per-scene comparisons. These results are consistent with the averaged results shown in the paper and show that our method achieves comparable performance to baselines.

References

[1] 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 *ICCV*, 2021. 1, 2

[2] Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 1, 2

[3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

Scan	#1	#8	#21	#103	#114
PSNR↑					
PixelNeRF	21.64	23.70	16.04	16.76	18.40
IBRNet	25.97	27.45	20.94	27.91	27.91
MVSNeRF	26.96	27.43	21.55	29.25	27.99
Ours	**28.86**	**28.98**	**22.69**	**30.64**	**29.00**
NeRF$_{10.2h}$	26.62	28.33	23.24	30.40	26.47
IBRNet$_{ft-1h}$	31.00	32.46	**27.88**	34.40	**31.00**
MVSNeRF$_{ft-15min}$	28.05	28.88	24.87	32.23	28.47
Ours$_{ft-20min}$	**31.93**	**32.69**	27.21	**34.66**	30.66
SSIM↑					
PixelNeRF	0.827	0.829	0.691	0.836	0.763
IBRNet	0.918	0.903	0.873	0.950	0.943
MVSNeRF	**0.937**	**0.922**	**0.890**	**0.962**	**0.949**
Ours	0.916	0.895	0.880	0.924	0.935
NeRF$_{10.2h}$	0.902	0.876	0.874	0.944	0.913
IBRNet$_{ft-1h}$	0.955	0.945	0.947	0.968	0.964
MVSNeRF$_{ft-15min}$	0.934	0.900	0.922	0.964	0.945
Ours$_{ft-20min}$	**0.966**	**0.956**	**0.948**	**0.971**	**0.965**
LPIPS ↓					
PixelNeRF	0.373	0.384	0.407	0.376	0.372
IBRNet	0.190	0.252	0.179	0.195	0.136
MVSNeRF	0.155	0.220	0.166	0.165	0.135
Ours	**0.105**	**0.149**	**0.121**	**0.128**	**0.094**
NeRF$_{10.2h}$	0.265	0.321	0.246	0.256	0.225
IBRNet$_{ft-1h}$	0.129	0.170	0.104	0.156	0.099
MVSNeRF$_{ft-15min}$	0.171	0.261	0.142	0.170	0.153
Ours$_{ft-20min}$	**0.088**	**0.133**	**0.092**	**0.119**	**0.086**

Table 3. **Quantitative comparison on the DTU dataset.**

Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[4] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1

[5] 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. 1

[6] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021. 1

	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
PSNR↑								
PixelNeRF	7.18	8.15	6.61	6.80	7.74	7.61	7.71	7.30
IBRNet	24.20	18.63	21.59	27.70	22.01	20.91	22.10	22.36
MVSNeRF	23.35	20.71	21.98	28.44	23.18	20.05	22.62	23.35
Ours	**27.01**	**22.67**	**23.22**	**31.66**	**24.24**	**23.05**	**25.00**	**25.20**
NeRF	**31.07**	**25.46**	**29.73**	**34.63**	**32.66**	**30.22**	**31.81**	**29.49**
IBRNet$_{ft-1h}$	28.18	21.93	25.01	31.48	25.34	24.27	27.29	21.48
MVSNeRF$_{ft-15min}$	26.80	22.48	26.24	32.65	26.62	25.28	29.78	26.73
Ours$_{ft-20min}$	27.81	24.01	24.25	33.15	25.16	24.79	27.38	25.81
SSIM↑								
PixelNeRF	0.624	0.670	0.669	0.669	0.671	0.644	0.729	0.584
IBRNet	0.888	0.836	0.881	0.923	0.874	0.872	0.927	0.794
MVSNeRF	0.876	0.886	0.898	0.962	0.902	0.893	0.923	**0.886**
Ours	**0.942**	**0.912**	**0.899**	**0.963**	**0.908**	**0.901**	**0.950**	0.791
NeRF	**0.971**	**0.943**	**0.969**	0.980	**0.975**	**0.968**	**0.981**	**0.908**
IBRNet$_{ft-1h}$	0.955	0.913	0.940	0.978	0.940	0.937	0.974	0.877
MVSNeRF$_{ft-15min}$	0.934	0.898	0.944	0.971	0.924	0.927	0.970	0.879
Ours$_{ft-20min}$	0.961	0.932	0.927	**0.981**	0.933	0.941	0.975	0.889
LPIPS ↓								
PixelNeRF	0.386	0.421	0.335	0.433	0.427	0.432	0.329	0.526
IBRNet	0.144	0.241	0.159	0.175	0.202	0.164	0.103	0.369
MVSNeRF	0.282	0.187	0.211	0.173	0.204	0.216	0.177	0.244
Ours	**0.060**	**0.097**	**0.101**	**0.066**	**0.108**	**0.102**	**0.058**	**0.205**
NeRF	**0.055**	0.101	**0.047**	0.089	**0.054**	0.105	**0.033**	0.263
IBRNet$_{ft-1h}$	0.079	0.133	0.082	0.093	0.105	0.093	0.040	0.257
MVSNeRF$_{ft-15min}$	0.129	0.197	0.171	0.094	0.176	0.167	0.117	0.294
Ours$_{ft-20min}$	0.056	**0.092**	0.085	**0.048**	0.095	**0.081**	0.038	**0.207**

Table 4. **Quantitative comparison on the NeRF Synthetic dataset.**

	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	Trex
PSNR↑								
PixelNeRF	12.40	10.00	14.07	11.07	9.85	9.62	11.75	10.55
IBRNet	20.83	22.38	27.67	22.06	**18.75**	15.29	27.26	20.06
MVSNeRF	**21.15**	24.74	26.03	23.57	17.51	**17.85**	26.95	**23.20**
Ours	20.84	**24.84**	**28.81**	**23.58**	18.20	17.50	**28.63**	20.70
NeRF$_{10.2h}$	**23.87**	26.84	**31.37**	25.96	21.21	19.81	**33.54**	**25.19**
IBRNet$_{ft-1h}$	22.64	26.55	30.34	25.01	**22.07**	19.01	31.05	22.34
MVSNeRF$_{ft-15min}$	23.10	27.23	30.43	**26.35**	21.54	**20.51**	30.12	24.32
Ours$_{ft-20min}$	21.92	**27.42**	29.88	25.49	21.28	19.01	30.82	23.42
SSIM↑								
PixelNeRF	0.531	0.433	0.674	0.516	0.268	0.317	0.691	0.458
IBRNet	**0.710**	0.854	**0.894**	0.840	**0.705**	0.571	0.950	0.768
MVSNeRF	0.638	**0.888**	0.872	**0.868**	0.667	**0.657**	**0.951**	**0.868**
Ours	0.628	0.830	0.864	0.795	0.633	0.541	0.921	0.701
NeRF$_{10.2h}$	**0.828**	0.897	**0.945**	0.900	0.792	0.721	**0.978**	**0.899**
IBRNet$_{ft-1h}$	0.774	0.909	0.937	0.904	**0.843**	0.705	0.972	0.842
MVSNeRF$_{ft-15min}$	0.795	**0.912**	0.943	**0.917**	0.826	**0.732**	0.966	0.895
Ours$_{ft-20min}$	0.751	0.911	0.933	0.902	0.818	0.706	0.966	0.861
LPIPS ↓								
	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	Trex
PixelNeRF	0.650	0.708	0.608	0.705	0.695	0.721	0.611	0.667
IBRNet	0.349	0.224	0.196	0.285	0.292	0.413	**0.161**	0.314
MVSNeRF	**0.238**	0.196	0.208	0.237	0.313	**0.274**	0.172	**0.184**
Ours	0.257	**0.181**	**0.137**	**0.218**	**0.256**	0.325	0.179	0.285
NeRF$_{10.2h}$	0.291	0.176	0.147	0.247	0.301	0.321	0.157	0.245
IBRNet$_{ft-1h}$	0.266	0.146	0.133	0.190	**0.180**	0.286	**0.089**	0.222
MVSNeRF$_{ft-15min}$	**0.253**	**0.143**	0.134	**0.188**	0.222	**0.258**	0.149	**0.187**
Ours$_{ft-20min}$	0.256	0.166	**0.126**	0.193	0.196	0.286	0.162	0.225

Table 5. **Quantitative comparison on the Real Forward-facing dataset.**