WaveNeRF: Wavelet-based Generalizable Neural Radiance Fields
Supplementary Material

1. Additional Technical Details

Apart from the implementation details in the main paper, we show more details of our network in this section.

As described in the main paper, our WMVS module utilizes varying granularities for the spatial sweep planes, specifically D_s values of 8, 32, and 48, while using a D_w value of 32 for the frequency sweep plane. Additionally, we employ zero padding in the wavelet transform and maintain a consistent number of channels (C = 8) for all feature volumes. To accommodate input images whose sizes are not divisible by 8, we apply padding to make them multiples of 8 before computation, and then restore their original sizes after passing through the operators. In the HNR module, we employ occlusion masks, angle-based weight calculation, and token generation as in GeoNeRF [5]. We kept these techniques consistent for the spatial domain and extended the same idea to the frequency domain.

We train our WaveNeRF model for 225k iterations using one RTX 3090 GPU. Each iteration randomly samples one scene and 512 rays are randomly selected as a training batch. We use an Adam optimizer with an initial learning rate of $5e-4$ and a cosine learning rate scheduler without restart.

We also mentioned in the main paper that we adopted a depth supervision technique and depth losses. DS-NeRF [3] shows that depth supervision can help NeRF train faster with fewer input views. Also, some generalizable NeRF studies [2, 5] display the effectiveness of depth supervision when the samples are from the dataset with ground truth depths. Therefore, we follow their depth supervision loss by comparing our predicted depth $\hat{d}$ with the ground truth depth d_{gt}.

$$\mathcal{L}_d = \frac{1}{|R_d|} \sum_{r \in R_d} ||\hat{d}(r) - d_{gt}(r)||_{s1}, \tag{1}$$

where R_d is the set of rays from samples with ground truth depths and $||\cdot||_{s1}$ is the smooth L1 loss.

In addition, although we design a new WMVS module to create cascade feature volumes as well as a wavelet feature volume, we can still predict the depth maps for each level of the cascade feature volumes. Therefore, we can follow the self-supervised depth loss function in GeoNeRF [5] for the dataset without the ground truth depth. We take the rendered ray depths as pseudo-ground truth $\hat{D}_v^{(l)}(r_v)$ and warp their corresponding colors and estimated depths from all source views using camera transformation matrices.

$$\mathcal{L}_{d_{us}}^{(l)} = \frac{2^{-l}}{|V||R|} \sum_{v=1}^{V} \sum_{r \in R} ||\hat{D}_v^{(l)}(r_v) - \hat{d}(r_v)||_{s1}, \tag{2}$$

$$r_v = T_{\to v}(r, \hat{d}(r)). \tag{3}$$

Given a ray r at a novel pose with rendered depth $\hat{d}(r)$, $T_{\to v}(r, \hat{d}(r))$ transforms the ray to its correspondent ray from source view v using camera matrices. And $\hat{d}(r_v)$ denotes the rendered depth of the correspondent ray with respect to source view v. Finally, the whole depth loss function of our model is represented as:

$$\mathcal{L}_D = \mathcal{L}_d + \sum_{l=0}^{2} \mathcal{L}_{d_{us}}^{(l)}. \tag{4}$$

2. Additional Experiment

2.1. Different Wavelet Transform Levels

The Wavelet Transform is capable of breaking down an image into various frequency feature maps with different scales. The number of scales employed in the transformation determines the number of levels of feature volumes in the WMVS module. Generally, the greater the number of levels of volumes, the more refined the depth hypotheses planes become, which theoretically leads to better performance. However, using more scales comes with a higher memory cost. Furthermore, when employing a scale J wavelet transform, the resulting decomposition yields feature maps of size $\frac{H}{2^J} \times \frac{W}{2^J}$, which presents a challenge when J is large due to the difficult padding operation. As such, we trained our WaveNeRF model using different scales ranging from 1 to 3 and evaluated them under the same settings and the same memory cost as the previous evaluation experiments. The quantitative results are shown in Tab. 1 and we can see from the table that the performance of the

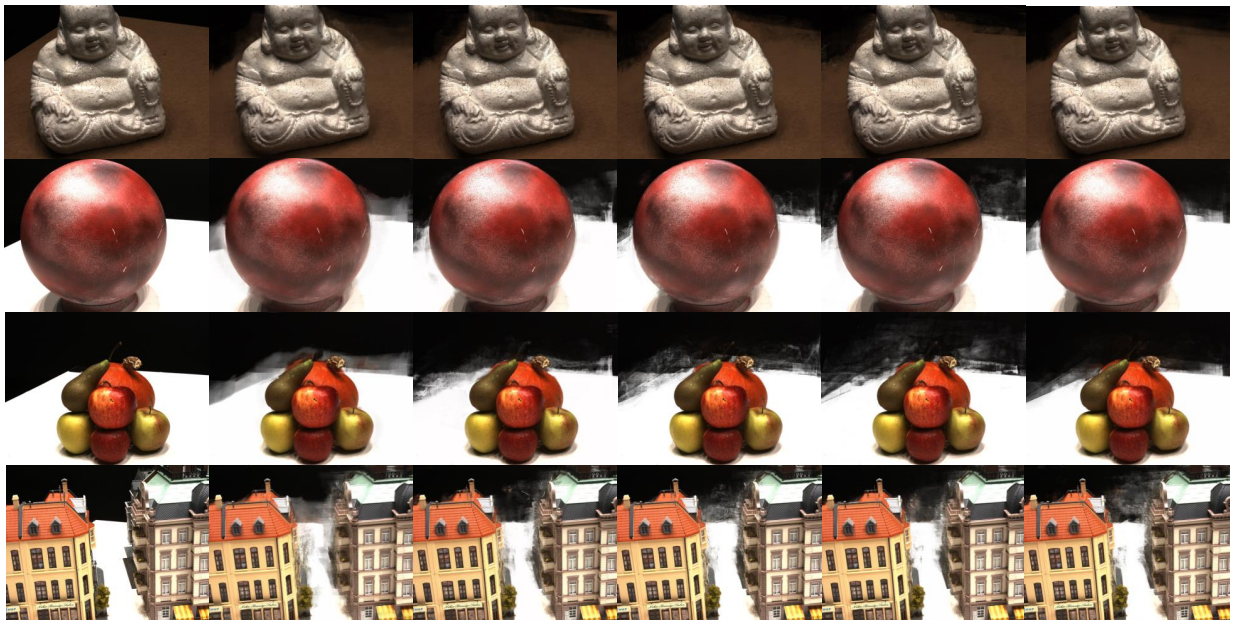


Figure 1: Qualitative comparisons of (a) the baseline model without any of our novel modules, (b) the baseline model + our WMVS module, (c) the baseline model + our WMVS module + our FSS sampling strategy, (d) the baseline model + all three of our proposed modules but without the WFL loss, and (e) our complete model, on the DTU datset [4].

scale 3 Wavelet Transform even drops severely. There may be two reasons for this result. One is that the complicated padding operations may disturb the learning process and the other potential reason is that the number of depth hypotheses planes for each volume is low due to memory limitation. If we further increase the number of scales, the padding becomes extremely hard and may even lead to the failure of training.

Scale	DTU [4]			LLFF [6]		
Level	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1	28.67	0.933	0.101	23.92	0.781	0.221
2	29.55	0.948	0.0749	24.28	0.794	0.212
3	27.31	0.899	0.134	23.54	0.761	0.251

Table 1: The results of different wavelet scales in terms of PSNR ↑, SSIM↑, and LPIPS↓ metrics.

2.2. Per-scene Breakdown

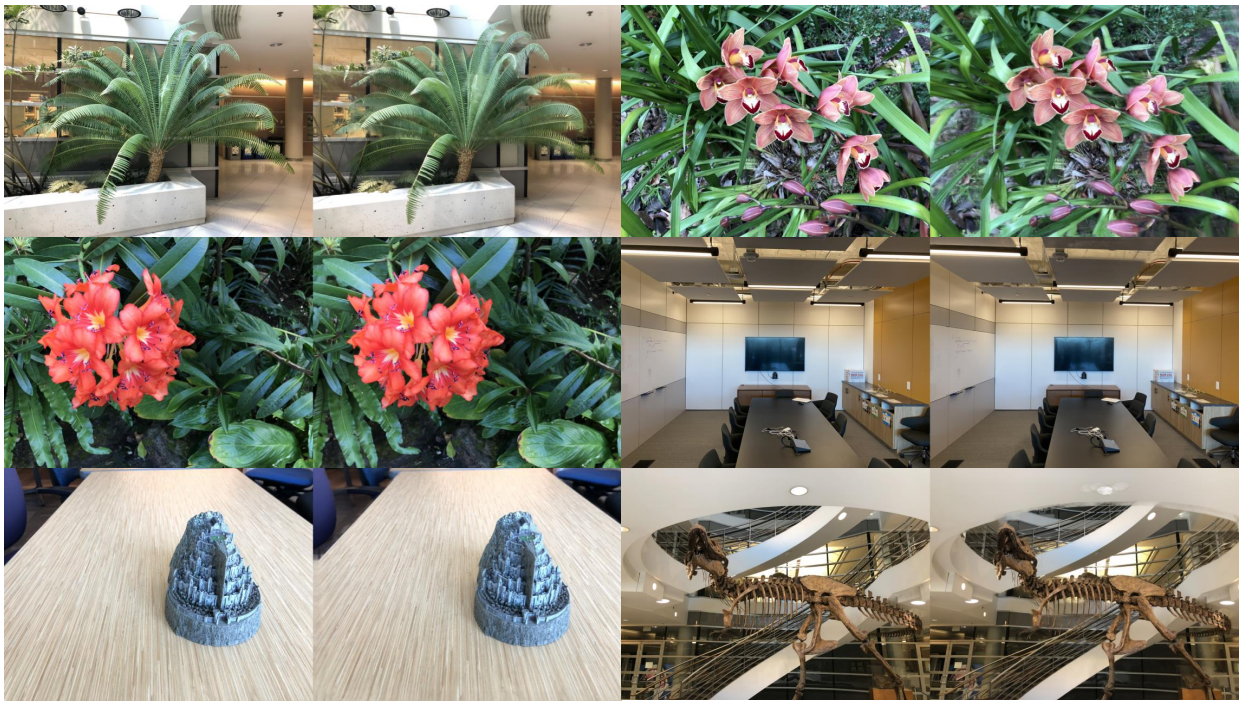
Tab.2 provides the quantitative results of our model on each scene of the NeRF Synthetic dataset [7] and the LLFF dataset [6] in terms of PSNR↑, SSIM↑, and LPIPS↓. We also display more qualitative results in Fig.2, 3. Fig.2 shows more comparisons between the ground truth and our rendered images on the LLFF dataset [6] while Fig.3 shows extra rendered results on the NeRF Synthetic dataset [7].

2.3. Qualitative Results for Our Ablation Studies

In the main paper, we mentioned that we conducted our ablation studies with the following variants: (a) the baseline model without any of our novel modules, (b) the baseline model + our WMVS module, (c) the baseline model + our WMVS module + our FSS sampling strategy, (d) the baseline model + all three of our proposed modules but without the WFL loss, and (e) our complete model. Here we present some qualitative results of our ablation studies on the DTU datset [4] in Fig.1.

2.4. Budget Analysis

Due to the different experiment settings between the few-shot NeRF methods and the fast NeRF methods such as Instant-NGP [9] and TensoRF [1], we only conducted the budget analysis and compare our work with GeoNeRF [5]. The table below shows the comparisons in parameter numbers, training time, and inference time (on a single RTX 3090 GPU). Our model has fewer parameters but longer training and inference time due to wavelet-related operations. The training and inference time could be shortened



Reference *WaveNeRF* *Reference* *WaveNeRF*

Figure 2: Additional qualitative results rendered by our WaveNeRF. The images are from the LLFF dataset [6]

	NeRF Synthetic [7]							
	chair	drums	ficus	hotdog	lego	materials	mic	ship
PSNR↑	30.00	22.22	23.44	32.50	26.18	23.43	27.36	23.85
SSIM ↑	0.963	0.897	0.906	0.965	0.932	0.898	0.957	0.826
LPIPS↓	0.060	0.124	0.110	0.086	0.097	0.143	0.069	0.214

	LLFF [6]							
	fern	flower	fortress	horns	leaves	orchids	room	trex
PSNR↑	23.16	27.25	29.09	25.36	19.10	18.68	28.33	23.26
SSIM ↑	0.748	0.863	0.855	0.860	0.662	0.586	0.928	0.849
LPIPS↓	0.249	0.147	0.146	0.184	0.258	0.313	0.160	0.240

Table 2: Quantitative results of our WaveNeRF on each scene of the NeRF Synthetic dataset [7] and the LLFF dataset [6] in terms of PSNR↑, SSIM↑, and LPIPS↓.

with faster wavelet transform calculations or utilizing a tri-plane structure to replace feature volumes. Note the training time is not a crucial factor as the primary objective of this work is a few-shot generalizable NeRF that requires no additional training or fine-tuning for new scenes. Hence, we can conclude that our method trains better few-shot generalizable NeRF with a comparable budget to GeoNeRF.

Model	Para.#	Training Time	Inference Time
GeoNeRF	1203485	86.1h	32s
WaveNeRF	1185058	125.05h	47s

2.5. Cross-dataset Generalization Ability

To evaluate the generalization ability of WaveNeRF, we trained a new model solely using the DTU dataset [4] (i.e.,

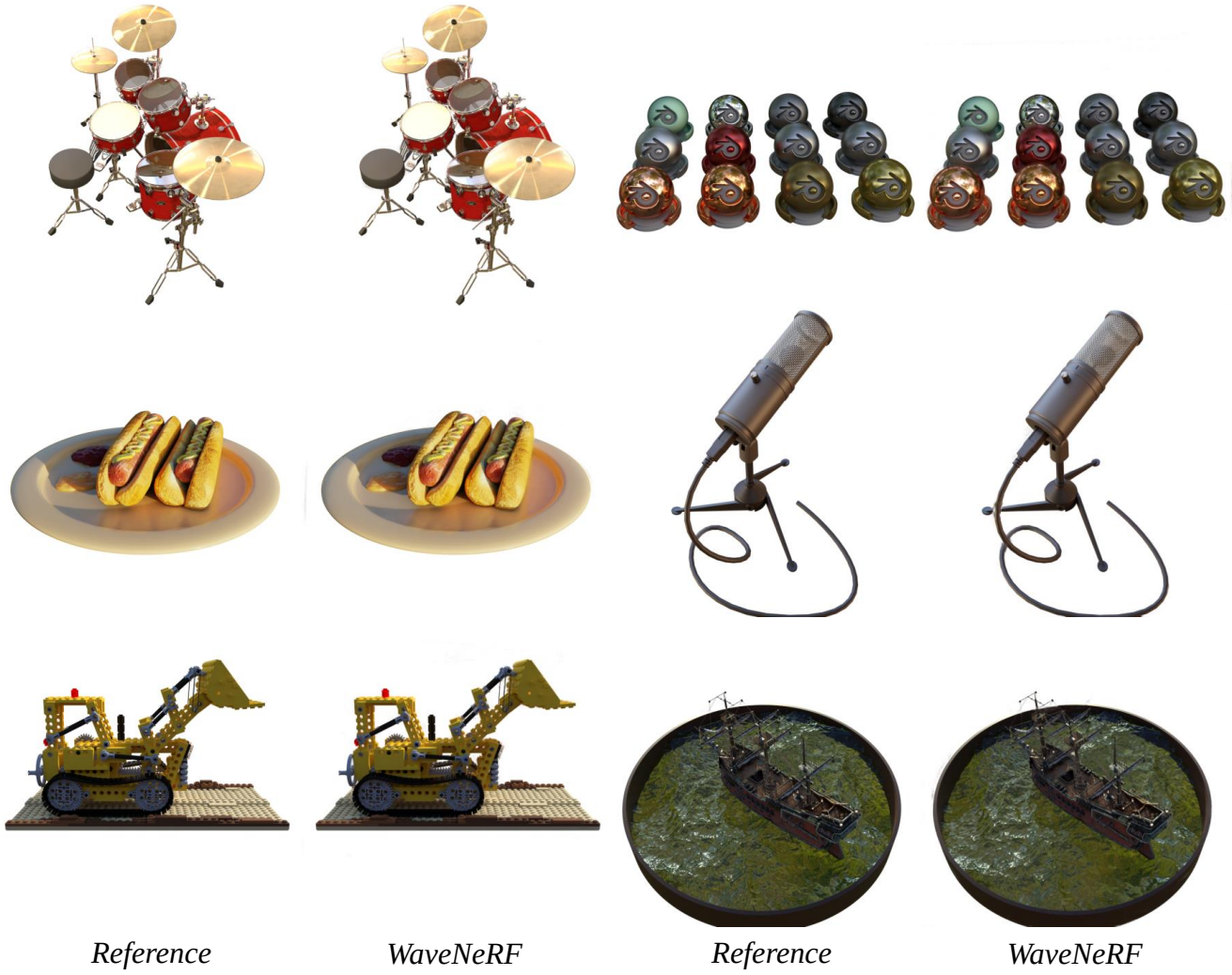


Figure 3: Additional qualitative results rendered by our WaveNeRF. The images are from the NeRF synthetic dataset [7]

WaveNeRF (DTU) shown in the table below) and evaluated it on LLFF dataset [6]. Compared with DTU and NeRF synthetic datasets [7], LLFF comprises more diverse and complex scenes with many non-centric objects, posing strong inter-domain gaps. We compared WaveNeRF (DTU) with MVSNeRF, GeoNeRF, and our model in the manuscript (all trained using LLFF). We can observe WaveNeRF (DTU) obtains lower PSNR than WaveNeRF due to domain gaps. However, WaveNeRF (DTU) still achieves competitive performance as compared with MVSNeRF and GeoNeRF (trained using LLFF), demonstrating the good generalization ability of WaveNeRF.

3. Ethical Consideration

The proposed model aims to synthesize novel view images from three-shot source views without any finetuning. It could have a negative impact when it is used to boost illegal 3D data collection. Thus, some watermarking technologies or detection methods [8] could be employed to identify the synthesized 3D asset.

References

[1] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022.

[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.

[3] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.

[4] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

[5] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022.

[6] 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.

[7] 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.

[8] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

[9] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.