mip-Grid - Supplementary Materials

A Source Code

We provide the source code to show the implementation details of our work. You can run the script files to reproduce each experiment presented in the main text and supplementary materials.

B Quantitative Comparisons on Forward-facing Scenes

We have conducted additional model evaluations on the multi-scale LLFF dataset [3]. The LLFF dataset consists of eight real-world scenes, with each scene containing multi-view images at a resolution of 1008×754 pixels. We downsampled the original images by a factor of 2, 4, and 8 as typically done in mip-NeRF [1], while rescaling the focal lengths accordingly. Tab. 1 compares the overall performance averaged across eight scenes at each resolution. Our method achieved the best results in all metrics, except for PSNR at the highest resolution, and also ours outperformed other methods by a large margin, especially at the lowest resolution. Note that we do not compare our method against mip-NeRF as it does not report evaluation results on the multi-scale LLFF dataset.

Table 1: Evaluation results on multi-scale LLFF dataset. We compare mip-TensoRF against the vanilla TensoRF and TensoRF (MS), a TensoRF trained on the multi-scale LLFF dataset.

	PSNR↑				SSIM↑				LPIPS↓			
	Full Res.	½ Res.	¼ Res.	⅛ Res.	Full Res.	½ Res.	¼ Res.	⅛ Res.	Full Res.	½ Res.	¼ Res.	⅛ Res.
TensoRF	26.73	27.89	26.70	22.81	0.8386	0.8932	0.8964	0.8063	0.2044	0.1069	0.1099	0.1685
TensoRF (MS)	25.16	27.17	29.10	25.26	0.7776	0.8665	0.9311	0.8784	0.2797	0.1508	0.0761	0.1118
mip-TensoRF	26.72	28.32	29.95	30.79	0.8397	0.8970	0.9398	0.9602	0.2001	0.1026	0.0586	0.0417

C Implementation Details

Baseline1, Baseline2, and mip-TensoRF. All three models are implemented on the top of TensoRF-VM-192 [2]. The number of channels of density and appearance grids are 16 and 48, respectively. Following the TensoRF approach, we begin training with an initial voxel size of 128^3 and progressively upsample at following steps: 2000, 3000, 4000, 5500, and 7000. We apply a binary occupancy mask [2] and update the mask at steps 2000 and 4000. We also scale the loss of each pixel at different resolution by a factor of 2^2, 4^2, and 8^2 following mip-NeRF [1]. However, when training mip-TensoRF, we do not scale the loss after the initial 10,000 iteration. In the case of mip-TensoRF, we use extra convolution kernels with a kernel size of 11. Since we have different kernel sets for each of the four scales, the number of channels is 64 for the density kernels and 192 for the appearance kernels. The input grid is repeated four times and convolved with the kernels to represent multi-scale feature grids.

Baseline2, Baseline3, and mip-*K*-Planes. We followed the experimental setting of *K*-planes and did not tuned any hyperparameters, with the exception of integrating our proposed method. As *K*-Planes are multi-resolution grid-based neural fields, we performed convolution operations on each 2D plane within every grid resolution. Furthermore, we also applied convolution on grids in proposal networks. For both the appearance and density grids, we employed 3x3-sized kernels for convolution.

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

Table 2: Total training hours for each scene in the NeRF synthetic dataset. We compare the runtime of mip-TensoRF and mip-K-Planes against the baseline models and mip-NeRF.

	Avg.	*chair*	*drums*	*ficus*	*hotdog*	*lego*	*materials*	*mic*	*ship*
TensoRF	0.17 ± 0.03	0.15	0.15	0.18	0.18	0.16	0.23	0.15	0.20
TensoRF (MS)	0.23 ± 0.04	0.20	0.19	0.22	0.25	0.23	0.29	0.20	0.26
mip-TensoRF	0.75 ± 0.08	0.70	0.68	0.69	0.86	0.72	0.85	0.69	0.83
K-Planes	0.66 ± 0.02	0.65	0.66	0.67	0.64	0.66	0.70	0.64	0.67
K-Planes (MS)	0.66 ± 0.00	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.67
mip-K-Planes	0.96 ± 0.03	0.96	0.96	1.00	0.95	0.98	0.95	0.95	0.91
mip-NeRF	>30.00	30.00							

Table 3: We provide the total training hours and PSNRs for *'hotdog'* object in the NeRF synthetic dataset, varying the kernel sizes and the number of generated multi-scale grids: K - kernel size, N - the number of multi-scale grids.

K	N	Full res.	½ Res.	¼ Res.	⅛ Res.	Avg.	Time (hours)
3	2	37.52	39.02	38.71	36.58	37.96	0.36
	4	37.56	39.05	38.75	36.46	37.96	0.37
	5	37.58	39.13	38.91	36.74	38.09	0.41
5	2	37.56	39.08	38.94	37.04	38.16	0.37
	4	37.52	39.05	39.02	37.26	38.21	0.45
	5	37.55	39.09	39.05	37.24	38.23	0.50
11	2	37.54	39.13	39.22	37.60	38.38	0.57
	4	37.56	39.15	39.30	37.94	38.49	1.08
	5	37.48	39.04	39.14	37.85	38.38	1.01

D Runtime comparisons

We have recognized that the runtime evaluations in the main text were not carried out in fully controlled environments. To ensure more rigorous comparisons, we re-evaluated the runtime of each model using a single NVIDIA A100 GPU. Tab. 2 shows the elapsed time for the total training iterations of each model. Please note that, for mip-NeRF, estimated values are provided due to the limited computational resources. We measured the time elapsed for 100 iterations and multiplied it by 10,000 to get the total runtime for 1 million iterations. While our method requires a longer training time compared to the baseline models, both mip-TensoRF and mip-K-Planes can be trained in less than an hour. Moreover, our method can be sped up by decreasing the kernel size or the number of multi-scale grids. Specifically, if we use the kernel size of 5 and two multi-scale grids, our method can achieve PSNR of 38.16 in around 20 minutes Tab. 3.

E Per-scene Results

Fig. 1 and Fig. 2 show the qualitative results on NeRF synthetic dataset and LLFF dataset. Tab. 4 and Tab. 5 provide the per-scene evaluations on NeRF synthetic dataset and LLFF dataset.

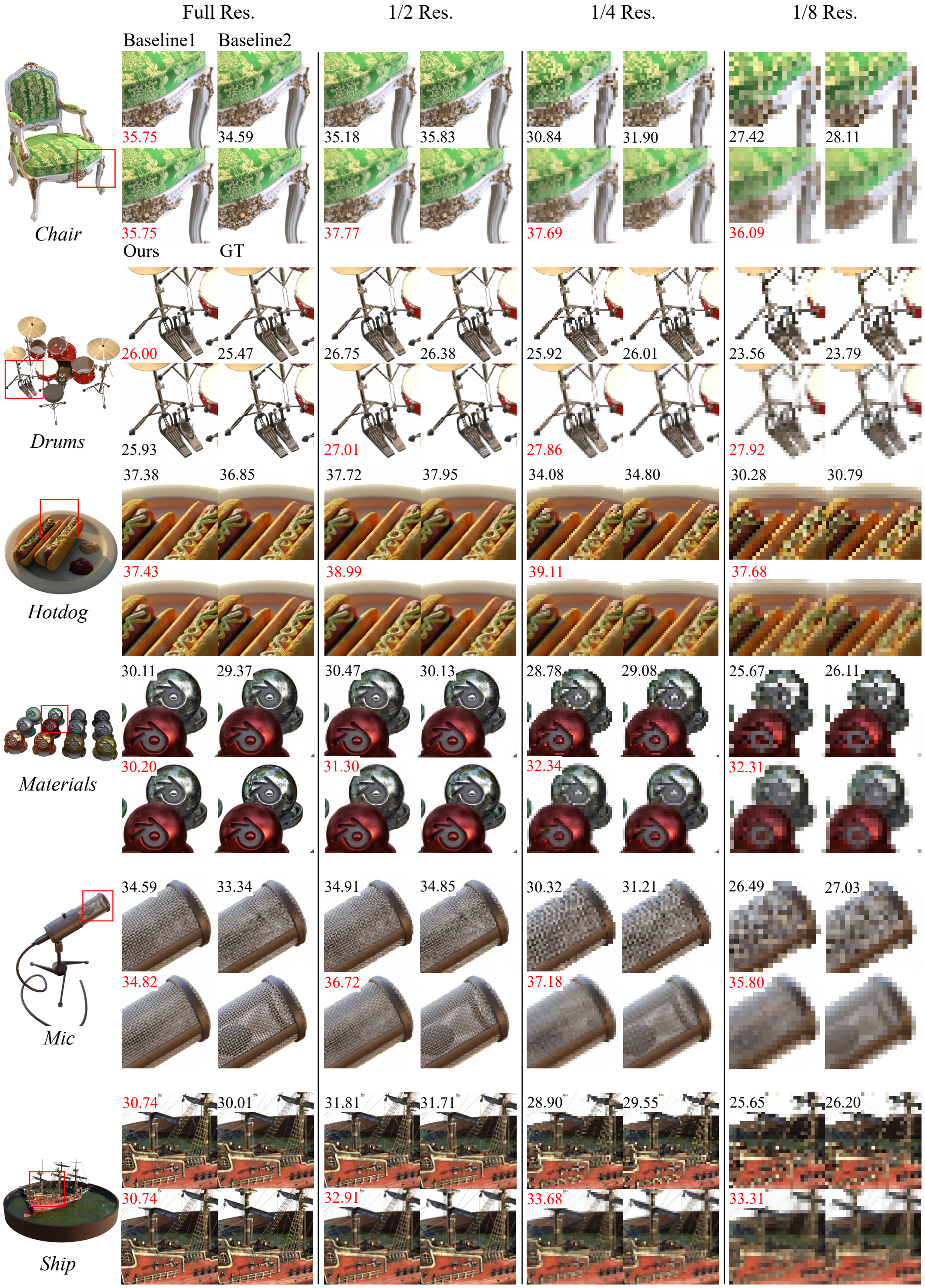

Figure 1: Qualitative comparison between Baseline1, Baseline2 and mip-TensoRF on the NeRF synthetic dataset. The cropped region and PSNR (the highest one was highlighted in red color) of each scene at four different scales are shown. Best viewed in color and zoom-in.

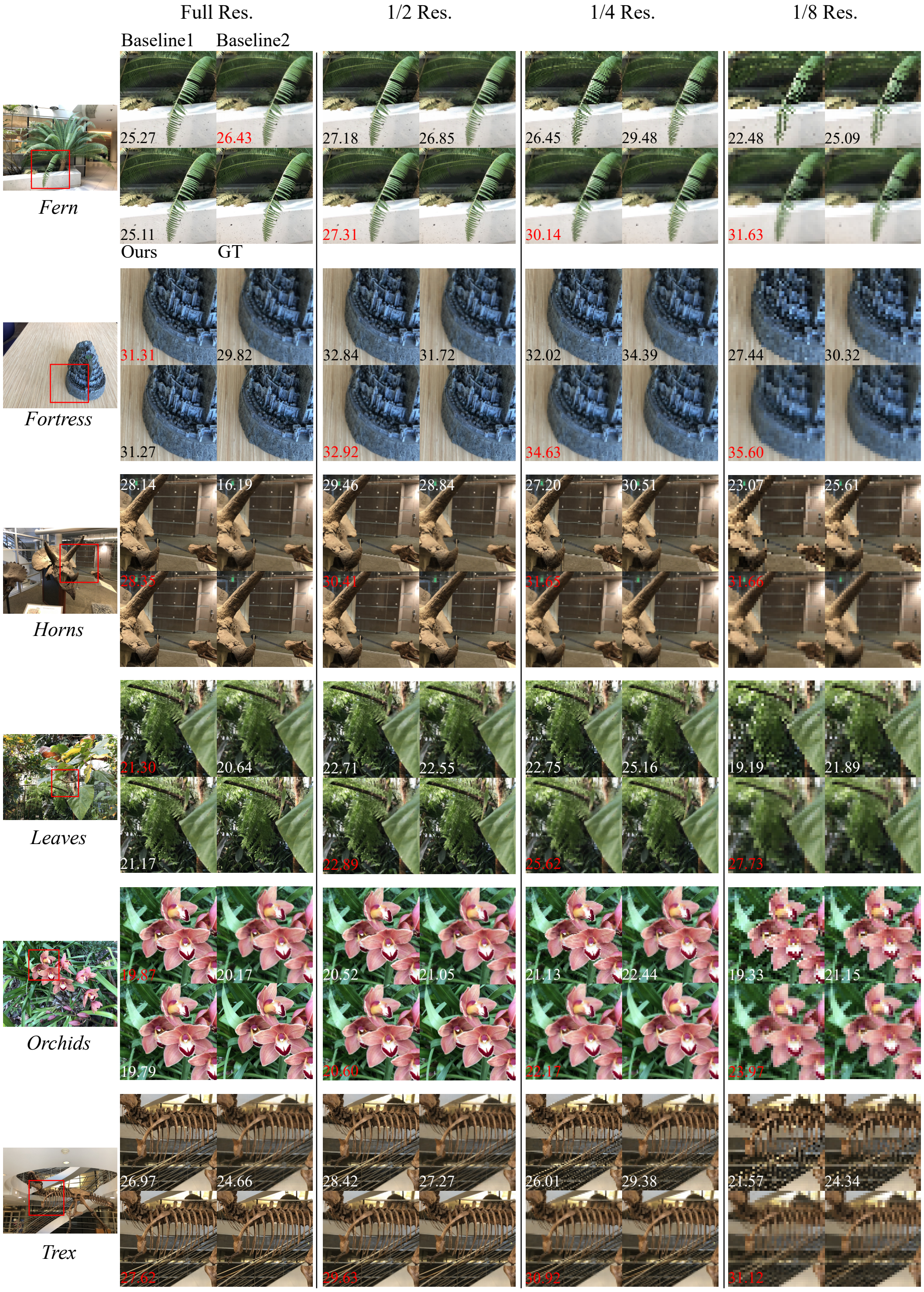

Figure 2: Qualitative comparison between Baseline1, Baseline2 and mip-TensoRF on the LLFF dataset. The cropped region and PSNR (the highest one was highlighted in red color) of each scene at four different scales are shown. Best viewed in color and zoom-in.

Table 4: Per-scene evaluations on NeRF synthetic dataset.

	Avg.	*chair*	*drums*	*ficus*	*hotdog*	*lego*	*materials*	*mic*	*ship*
Average PSNR↑									
TensoRF	30.70	32.30	25.55	31.47	34.86	31.83	28.76	31.58	29.28
TensoRF (MS)	30.72	32.61	25.41	31.08	35.10	31.89	28.67	31.61	29.37
mip-TensoRF	33.94	36.83	27.18	33.22	38.30	35.65	31.54	36.13	32.66
K-Planes	29.76	31.67	25.29	29.53	33.89	30.87	28.19	30.45	28.17
K-Planes (MS)	30.37	32.53	25.65	29.66	34.59	31.29	28.62	31.56	29.09
mip-K-Planes	32.27	34.91	26.55	30.71	36.02	33.97	30.04	34.39	31.57
Average SSIM↑									
TensoRF	0.9579	0.9654	0.9319	0.9758	0.9775	0.9661	0.9608	0.9787	0.9069
TensoRF (MS)	0.9568	0.9676	0.9301	0.9743	0.9779	0.9669	0.9586	0.9786	0.9007
mip-TensoRF	0.9730	0.9884	0.9471	0.9835	0.9873	0.9858	0.9728	0.9904	0.9289
K-Planes	0.9542	0.9661	0.9321	0.9673	0.9754	0.9615	0.9578	0.9751	0.8984
K-Planes (MS)	0.9575	0.9713	0.9341	0.9684	0.9780	0.9658	0.9589	0.9789	0.9044
mip-K-Planes	0.9676	0.9828	0.9435	0.9748	0.9822	0.9804	0.9661	0.9877	0.9234
Average LPIPS↓									
TensoRF	0.0525	0.0426	0.0740	0.0291	0.0356	0.0371	0.0581	0.0388	0.1050
TensoRF (MS)	0.0536	0.0418	0.0791	0.0322	0.0351	0.0357	0.0600	0.0380	0.1067
mip-TensoRF	0.0296	0.0145	0.0550	0.0180	0.0169	0.0132	0.0301	0.0113	0.0782
K-Planes	0.0565	0.0414	0.0727	0.0357	0.0385	0.0457	0.0625	0.0435	0.1120
K-Planes (MS)	0.0529	0.0367	0.0764	0.0345	0.0352	0.0413	0.0581	0.0368	0.1046
mip-K-Planes	0.0358	0.0211	0.0594	0.0254	0.0229	0.0198	0.0405	0.0141	0.0836

Table 5: Per-scene evaluations on LLFF dataset.

	Avg.	*chair*	*drums*	*ficus*	*hotdog*	*lego*	*materials*	*mic*	*ship*
Average PSNR↑									
TensoRF	26.03	25.35	28.11	30.90	26.97	21.48	20.21	29.48	25.74
TensoRF (MS)	26.67	26.43	27.97	31.56	27.79	22.56	21.20	29.46	26.41
mip-TensoRF	28.94	28.55	29.98	33.61	30.52	24.35	21.63	33.10	29.82
Average SSIM↑									
TensoRF	0.8586	0.8425	0.9033	0.8909	0.8805	0.8076	0.7452	0.9171	0.8818
TensoRF (MS)	0.8634	0.8499	0.8923	0.8987	0.8878	0.8103	0.7611	0.9255	0.8815
mip-TensoRF	0.9092	0.8934	0.9306	0.9437	0.9397	0.8705	0.7875	0.9615	0.9466
Average LPIPS↓									
TensoRF	0.1474	0.1723	0.0945	0.1133	0.1505	0.1581	0.1987	0.1293	0.1627
TensoRF (MS)	0.1546	0.1749	0.1171	0.1109	0.1492	0.1721	0.1956	0.1464	0.1705
mip-TensoRF	0.1008	0.1247	0.0728	0.0634	0.0849	0.1113	0.1572	0.0883	0.1034

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 (ICCV)*, pages 5855–5864, October 2021.

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*. Springer, 2022.

[3] 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 *European conference on computer vision*, pages 405–421. Springer, 2020.