

LookCloser: Frequency-aware Radiance Field for Tiny-Detail Scene

Supplementary Material

This supplementary material includes video results for scenes from the Multi-frequency datasets. In the following sections, we first introduce additional implementation details (Sec. 7). Moreover, we provide additional experiments on frequency level evaluation from 2D to 3D on the real dataset Mip-NeRF360-v2 (Sec. 8). Finally, we present more experimental results (Sec. 9).

7. Additional Implementation Details

7.1. Dataset

We captured our dataset using two different cameras. We collected normal-resolution images with a panoramic camera to achieve a more comprehensive 360-degree field of view for better scene structure reconstruction. And high-resolution images were captured with a DSLR (Digital Single-Lens Reflex) camera. To obtain camera poses, we perform Structure from Motion reconstruction for both the panoramic and high-resolution images. We use shared intrinsic between all images of the same camera model in a scene, and calibrate using the OpenCV radial distortion model. We then project the panoramic images into six 600×600 perspective images, each with FOV of 60° to accommodate the perspective camera model commonly used in most NeRF models. We adopt a common dataset splitting method, selecting one out of every eight panoramic/high-resolution images as the test set, with the remainder constituting the training set.

7.2. Architecture details

We adopt a setup similar to Instant-NGP [31], utilizing 16 grid scales with the maximum resolution being $2048 \times \text{scene size}$ and the minimum resolution being 16, employing 2 feature channels per level. In our dataset, due to the larger scene sizes, we set the size of the hash table storing feature vector for each level to 2^{23} to mitigate the impact of hash collisions on scene representation. For other general scenes, we use an identical hash table size of 2^{19} to Instant-NGP [31]. The fetched hash feature vectors are down-weighted before being concatenated and fed to a one-layer MLP with 64 hidden units to get the scene features and the volume densities. Subsequently, the scene features are concatenated with the spherical harmonics encoding of the view directions, which is then input to a subsequent two-layer MLP of width 64 to yield the RGB colors.

7.3. Frequency Grid

To represent the frequency distribution in the 3D space, we maintain a frequency grid with a resolution of $128 \times \text{AABB}$,

where AABB, short for Axis-Aligned Bounding Box, denotes the scene size. For each scene, we adjust the AABB based on the 3D points from the SfM reconstruction to ensure it encompasses the majority of the 3D points. Each grid cell stores the frequency level as a uint8 number.

Initialization. Once we have the 2D frequencies of all training patches, we first calculate the 3D frequency of each 3D point p_i . After that, each 3D point is reprojected to obtain a set of observation patches $\{P_{ij} | j = 1, \dots, n\}$ and derive a set of 3D frequencies $\{f_{3D_{ij}} | j = 1, \dots, n\}$ with the depth of the point. To mitigate the influence of noisy patches, we take the median of this set as the 3D frequency f_{3D_i} for that point. Assuming that the frequencies at each level are $\{f_{3D_\ell} | \ell = 0, \dots, n_\ell\}$, we take the frequency level ℓ_i as $\arg \min_\ell (|f_{3D_\ell} - f_{3D_i}|)$. The frequency grid is then initialized to the maximum of the frequency levels of all 3D points within the grid.

Re-weighting. Unlike Instant-NGP [31], which directly concatenates feature vectors as the input for the tiny MLP, we take into account the 3D frequency at that point and re-weight different frequency components accordingly. Instead, we use the quantified frequency level ℓ as a threshold and apply a down-weighting to frequency components that are higher than ℓ . We compute the down-weighting factor w using an approximation for $\text{erf}(x)$:

$$\text{erf}(x) \approx \text{sign}(x) \sqrt{1 - \exp(-(4/\pi)x^2)} \quad (9)$$

Updating. We update the grids after every 1024 training iterations by the following steps. We first render the depth of the center pixel of a training patch P_i . Then, the 2D frequency of the patch is projected to the corresponding 3D point to obtain its 3D frequency f_{3D_i} and frequency level ℓ_i . Finally, the value ℓ of the frequency grid where the 3D point resides is then updated to $\max(\ell_i, \ell)$.

Frequency-averaged sampling (FAS). We divide the training batch into N segments based on the frequency quantization results. The sampling frequency is evenly distributed within a preset range of $[1, 3]$, meaning that the highest frequency content is sampled with a probability three times that of the lowest frequency. In our experiments, we found that this is a more stable setting compared to directly using the frequency ratio as the sampling proportion.

7.4. Loss Functions

As described in the main paper, the training loss is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{recon}(\hat{\mathbf{c}}, \mathbf{c}_{gt}) + \lambda_{depth} \mathcal{L}_{depth}(\hat{\mathbf{d}}, \mathbf{d}_{gt}) + \lambda_{dist} \mathcal{L}_{dist}(\mathbf{s}_{\mathbf{d}}, \mathbf{w}), \quad (10)$$

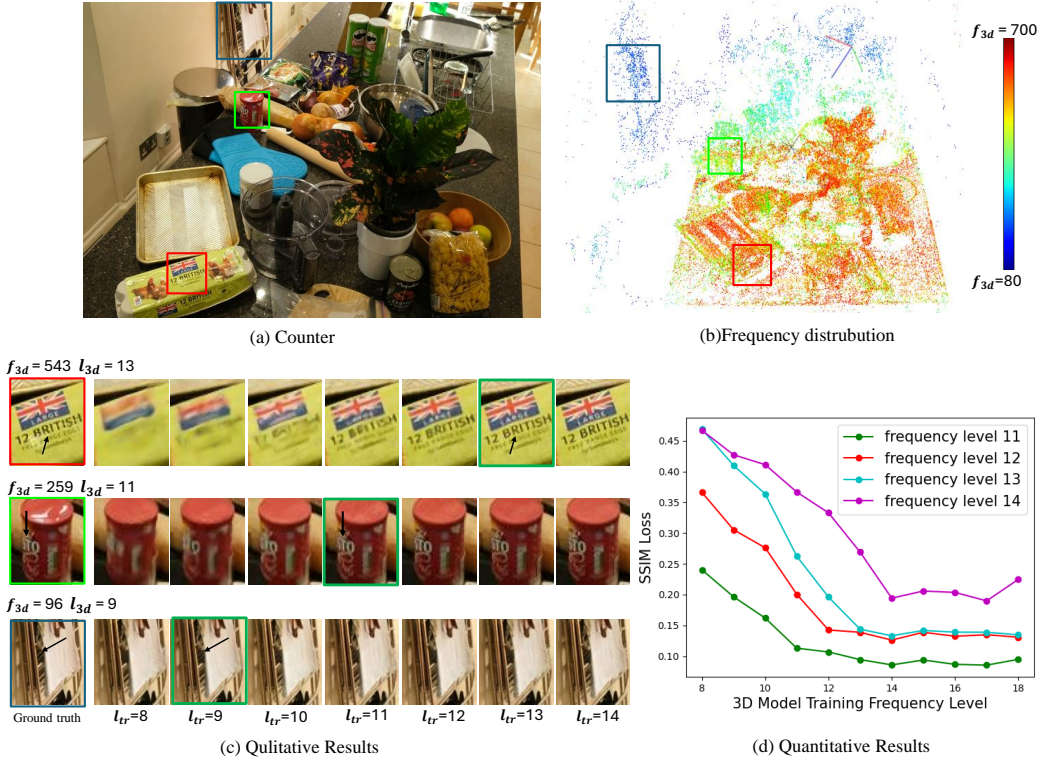


Figure 7. We test the effectiveness of frequency quantification on the real dataset NeRF-360-v2. (a) is a ground truth image from the *counter* dataset. (b) is the colored point cloud after the 3D frequency initialization of all 3D points, where colors leaning towards blue indicate lower frequencies and those towards red indicate higher frequencies. It can be observed that there is a trend of higher frequencies with smaller depths. Moreover, different regions at the same depth also exhibit varying distributions of 3D frequencies. In (a), we selected three patches with different frequency levels ℓ_{3D} 9, 11, and 13, represented by blue, green, and red, respectively. (c) is the rendering results of these three patches at training frequency level ℓ_{tr} ranging from 8 to 14. It can be noticed that details are not well recovered when ℓ_{tr} is lower than the quantified ℓ_{3D} , and when ℓ_{tr} exceeds ℓ_{3D} , there is no significant improvement in the quality of the rendering images. (d) shows the distribution of average SSIM Loss for patches of different frequency levels in the *counter* dataset at various training frequency levels ℓ_{tr} . Once ℓ_{tr} reaches the quantified ℓ_{3D} , there is no significant decrease in loss.

where the first term $\mathcal{L}_{recon}(\hat{\mathbf{c}}, \mathbf{c}_{gt}) = \sqrt{(\hat{\mathbf{c}} - \mathbf{c}_{gt})^2 + \epsilon}$ is a color reconstruction loss [2], $\hat{\mathbf{c}}$ is the rendered pixel color, \mathbf{c}_{gt} is the ground-truth pixel color, and $\epsilon = 10^{-4}$, and the last term is the regularization loss.

The depth loss \mathcal{L}_{depth} of the sampled ray is defined by

$$\mathcal{L}_{depth}(\hat{\mathbf{d}}, \mathbf{d}_{gt}) = \sqrt{(\hat{\mathbf{d}} - \mathbf{d}_{gt})^2 + \epsilon} \quad (11)$$

where the depth of a ray is computed by the weighted sum of the sampled distance that $d = \sum_i w_i t_i$, and $\{w_i\}$ are the weights computed by the volume rendering. We only use the depth loss in early training for pixels with GT depth from the sparse point cloud to avoid incorrect geometry structure.

The regularization loss is proposed by Mip-NeRF360 [2]. We use it to prevent floaters and background collapse, which is defined as

$$\mathcal{L}_{dist}(\mathbf{s}_d, \mathbf{w}) = \sum_{i,j} w_i w_j \left| \frac{s_i + s_{i+1}}{2} - \frac{s_j + s_{j+1}}{2} \right| + \frac{1}{3} \sum_i w_i^2 (s_{i+1} - s_i), \quad (12)$$

where \mathbf{s}_d is the set of normalized ray distances and \mathbf{w} is the set of weights. It penalizes the discreteness to encourage the formation of thinner surfaces. In contrast to Mip-NeRF360 using a proposal network to obtain sampling suggestions, we compute this discrete version of sampling distribution regularization along the entire ray.

The hyperparameters λ_{dist} , λ_{depth} are used to balance the data terms and the regularize; we set $\lambda_{dist} = 0.01$, $\lambda_{depth} = 0.001$ for all experiments.



Figure 8. Qualitative comparisons with the Instant-NGP [31] that has a larger hash table size (Big) on the Mip-NeRF360-v2 [2] and Tank and Temples [19] dataset.

Method	room	counter	kitchen	bonsai	average
TensoRF	0.791	0.697	0.560	0.783	0.708
INGP-Base	0.893	0.845	0.857	0.924	0.879
INGP-Big	0.900	0.868	0.907	0.922	0.900
Mip-NeRF360	0.913	0.895	0.920	0.939	0.916
3D-GS	0.914	0.905	0.922	0.938	0.919
Ours	0.936	0.908	0.931	0.946	0.931

Table 4. SSIM on the Mip-NeRF360-v2 dataset

Method	room	counter	kitchen	bonsai	average
TensoRF	0.419	0.469	0.516	0.389	0.448
INGP-Base	0.242	0.255	0.170	0.198	0.216
INGP-Big	0.254	0.256	0.158	0.209	0.219
Mip-NeRF360	0.211	0.203	0.126	0.177	0.179
3D-GS	0.220	0.204	0.129	0.205	0.189
Ours	0.191	0.184	0.123	0.159	0.165

Table 5. LPIPS on the Mip-NeRF360-v2 dataset

8. Evaluate Frequency Level from 2D to 3D

In this section, we further demonstrate the effectiveness of frequency quantification from 2D to 3D using the real dataset Mip-NeRF360-v2.

Visualization of Frequency Distribution. As described in the main paper, we reproject each 3D point from the sparse point cloud back into all the observation images. Then we calculate the 3D frequency set S based on all the corresponding patches. The median of S is taken as the 3D frequency for that point. Fig. 7(b) shows a visualization of the 3D frequency distribution of all 3D points after initialization for the dataset *counter* in Mip-NeRF360-v2, where the color of the points indicates the corresponding 3D frequency, with points closer to blue indicating a lower frequency and those closer to red indicating a higher frequency. Fig. 7(a) represents the ground truth image, where the blue, green, and red boxes represent three patches with 3D frequencies from low to high as shown in Fig. 7(b).

Qualitative Results. Fig. 7(c) depicts the visual comparison of the rendering results under varying training frequencies of the three patches mentioned above, where the boxed patches represent the rendering results under the quantified

Method	room	counter	kitchen	bonsai	average
TensoRF	26.88	23.39	23.12	25.46	24.71
INGP-Base	30.31	26.21	29.00	31.08	29.15
INGP-Big	30.19	27.27	30.86	30.57	29.72
Mip-NeRF360	31.40	29.44	32.02	33.11	31.49
3D-GS	30.63	28.70	30.32	31.98	30.95
Ours	31.45	29.19	31.41	32.75	31.20

Table 6. PSNR on the Mip-NeRF360-v2 dataset

3D frequency level ℓ_{3D} . It is clearly demonstrated that when the training frequency level is lower than ℓ_{3D} , the network is unable to fully recover the detailed information. Conversely, when the training frequency exceeds the quantified 3D frequency, the network does not yield better results either.

Quantitative Results. Furthermore, in Fig. 7(d), the lines in green, red, blue, and purple correspond to patches with 3D frequency levels of 11, 12, 13, and 14, respectively. With the escalation of the training frequency from 8 to 14, there is a progressive reduction in the SSIM loss for the generated patches. Upon reaching the quantified 3D frequency for each patch with the training frequency, the SSIM loss reduction becomes more consistent. This observation suggests two key points: firstly, the necessary minimum NeRF frequency level for the complete reconstruction of the scene’s diverse 3D frequency structures and textures is variable; secondly, the 3D frequency estimation we employ for the content provides an accurate reflection of their actual frequencies.

9. More Experimental Details

9.1. Quantitative Results on Standard Datasets

We compare our methods against our baselines on the standard datasets whose scenes have a smaller frequency span and size. The quantitative results are shown in the Tab. 2. Here we show the qualitative comparisons with the Instant-NGP [31], as depicted in Fig. 8. Our methods render sharper and clearer high-frequency contents than the Instant-NGP, indicating that while our frequency-aware framework is designed to handle high-quality model scene structures and details in scenarios with significant frequency disparities, it still generalizes well on standard datasets, enhancing rendering quality, particularly in high-frequency details.

9.2. More Ablation Studies

Component Ablation. We conducted ablation experiments in each scene, and the results are shown in Tab. 7. The results indicate that the impact of different features on overall performance varies across scenes of different scales. In particular, in high-frequency scenes captured at close range, such as the “Flower Shop”, the sampling interval adjustment has a more pronounced effect. This tendency is par-

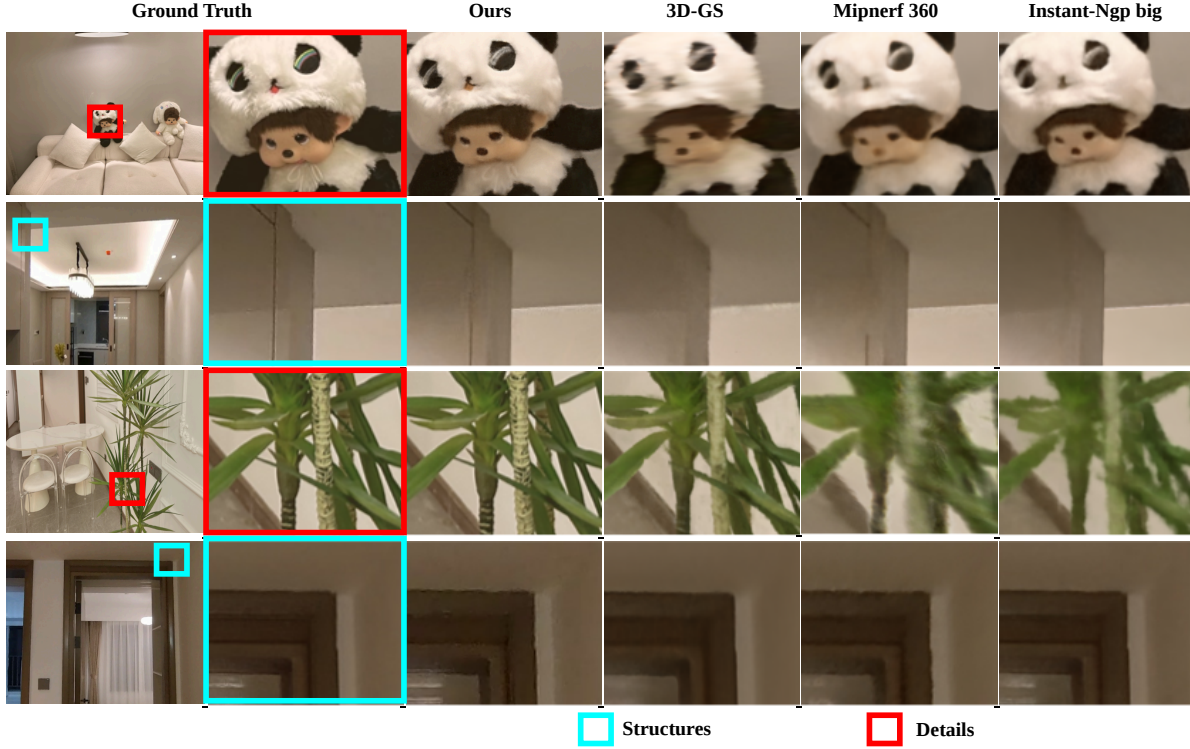


Figure 9. Additional visual comparisons on the Multi-frequency dataset.

Setting	FlowerShop	Home	DollsRoom	MusicRoom	PlantRoom	Average
normal-res(600×600)						
w/o Feature Re-weighting	28.19	32.93	34.12	33.46	33.42	32.42
w/o FAS	26.36	32.93	34.18	33.50	33.59	32.11
w/o Interval Adjustment	27.12	32.62	34.09	32.30	32.96	31.82
Our Complete Model	28.23	32.91	34.20	33.52	33.36	32.44
high-res(4032×3024)						
w/o Feature Re-weighting	24.62	26.14	28.41	26.54	24.55	26.05
w/o FAS	24.01	26.47	28.74	26.84	23.98	25.97
w/o Interval Adjustment	23.82	25.79	28.31	26.02	23.81	25.55
Our Complete Model	24.86	26.24	28.75	26.97	24.63	26.29

Table 7. Ablation Studies on Multi-frequency dataset

ticularly evident in large-scale datasets or close-range captures. As shown in Figure. 10, close-range high-frequency content becomes blurred in the absence of sampling interval adjustment, which aligns with the description in Section 4.2 of the paper. Due to variations in the proportion of high-frequency data within scenes, the efficacy of FAS also varies. Balancing training batches sometimes enhances high-frequency effects, while at other times it may diminish them, depending on the distribution of scene data. Feature re-weighting enhances the network’s efficiency in utilizing various frequency ranges, particularly when there is abundant scene content and limited network capacity.

9.3. Per-Scene Metrics

We provide the per-scene results on the Multi-frequency dataset, Tanks&Temples dataset, and Mip-NeRF360-v2 dataset in Tab. 14 - 5. The results are reported in the metrics of PSNR, SSIM, and LPIPS. We provide more visual comparisons on the Multi-frequency dataset in Fig. 9.

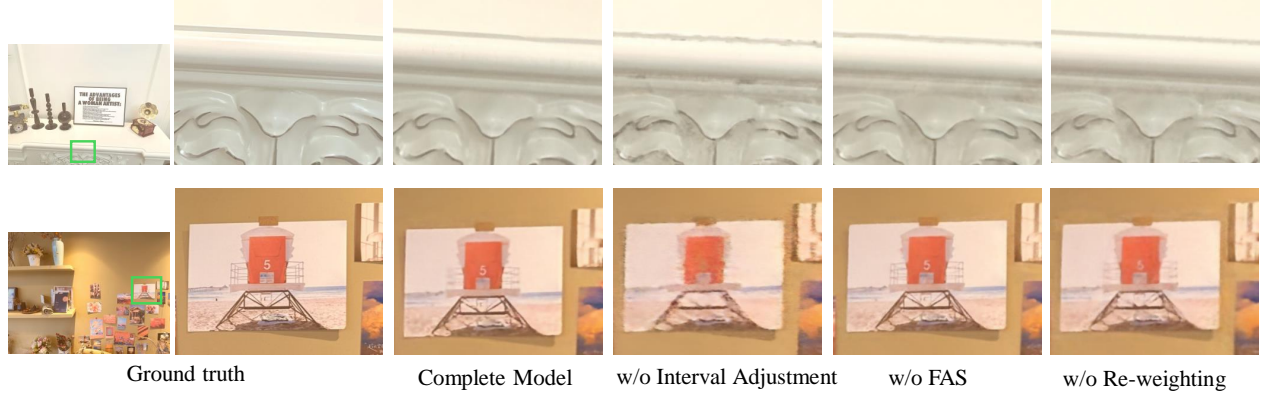


Figure 10. Additional visual comparisons with different settings of our method.

Table 8. PSNR on the Multi-frequency dataset(structure view)

PSNR	FlowerShop	Home	DollsRoom	MusicRoom	PlantRoom	Average
TensorRF	24.36	29.63	30.78	30.69	28.94	28.88
INGP-Base	25.31	30.10	32.07	31.78	32.07	30.27
INGP-Big	26.22	30.94	32.75	32.25	32.67	30.97
Mip-NeRF360	26.90	32.01	31.28	32.59	31.15	30.79
3D-GS	27.02	32.08	31.30	32.74	31.17	30.85
Ours	28.23	32.91	34.2	33.52	33.36	32.44

Table 9. SSIM on the Multi-frequency dataset(structure view)

SSIM	FlowerShop	Home	DollsRoom	MusicRoom	PlantRoom	Average
TensorRF	0.740	0.884	0.893	0.890	0.862	0.854
INGP-Base	0.791	0.902	0.925	0.921	0.924	0.893
INGP-Big	0.830	0.918	0.934	0.930	0.933	0.909
Mip-NeRF360	0.851	0.935	0.903	0.923	0.896	0.906
3D-GS	0.846	0.931	0.900	0.910	0.899	0.897
Ours	0.890	0.942	0.946	0.931	0.937	0.929

Table 10. LPIPS on the Multi-frequency dataset(structure view)

LPIPS	FlowerShop	Home	DollsRoom	MusicRoom	PlantRoom	Average
TensorRF	0.310	0.230	0.233	0.240	0.266	0.256
INGP-Base	0.267	0.205	0.206	0.207	0.196	0.216
INGP-Big	0.210	0.169	0.181	0.184	0.172	0.183
Mip-NeRF360	0.181	0.158	0.208	0.187	0.206	0.188
3D-GS	0.177	0.164	0.217	0.189	0.211	0.191
Ours	0.148	0.130	0.146	0.162	0.152	0.148

Table 11. PSNR on the Multi-frequency dataset(detail view)

PSNR	FlowerShop	Home	DollsRoom	MusicRoom	PlantRoom	Average
TensorRF	23.23	16.54	26.93	25.19	21.91	22.76
INGP-Base	22.27	22.82	26.75	24.49	21.83	23.63
INGP-Big	22.65	23.36	27.02	24.90	22.06	24.00
Mip-NeRF360	23.27	24.28	25.98	25.28	21.97	24.16
3D-GS	23.42	24.46	26.11	25.41	22.01	24.29
Ours	24.86	26.24	28.75	26.97	24.64	26.29

Table 12. SSIM on the Multi-frequency dataset(detail view)

SSIM	FlowerShop	Home	DollsRoom	MusicRoom	PlantRoom	Average
TensoRF	0.726	0.655	0.791	0.882	0.852	0.781
INGP-Base	0.716	0.717	0.769	0.868	0.848	0.784
INGP-Big	0.720	0.722	0.770	0.870	0.849	0.786
Mip-NeRF360	0.686	0.731	0.795	0.884	0.863	0.792
3D-GS	0.735	0.758	0.791	0.893	0.833	0.802
Ours	0.770	0.813	0.817	0.924	0.892	0.843

Table 13. LPIPS on the Multi-frequency dataset(detail view)

LPIPS	FlowerShop	Home	DollsRoom	MusicRoom	PlantRoom	Average
TensoRF	0.459	0.592	0.415	0.316	0.367	0.430
INGP-Base	0.466	0.500	0.404	0.323	0.346	0.408
INGP-Big	0.485	0.449	0.401	0.316	0.337	0.398
Mip-NeRF360	0.421	0.459	0.413	0.292	0.343	0.383
3D-GS	0.422	0.454	0.423	0.306	0.347	0.390
Ours	0.384	0.361	0.367	0.250	0.302	0.332

Table 14. Results on the Tanks&Temples dataset

Dataset	Train			Truck			Average		
Method—Metric	PSNR [↑]	SSIM [↑]	LPIPS _(VGG) [↓]	PSNR [↑]	SSIM [↑]	LPIPS _(VGG) [↓]	PSNR [↑]	SSIM [↑]	LPIPS _(VGG) [↓]
TensoRF	18.73	0.569	0.490	20.30	0.657	0.411	19.52	0.613	0.451
INGP-Base	20.43	0.684	0.367	22.69	0.777	0.269	21.56	0.731	0.318
INGP-Big	20.39	0.711	0.332	22.98	0.803	0.227	21.69	0.757	0.280
Mip-NeRF360	19.52	0.660	0.354	24.91	0.857	0.159	22.22	0.759	0.257
3D-GS	23.06	0.813	0.200	25.66	0.849	0.226	24.36	0.831	0.213
Ours	23.13	0.802	0.197	25.77	0.841	0.210	24.45	0.821	0.205

Table 15. Quantitative comparisons

	Structure View(600×600)			Detail View(4032× 3024)		
Method(Mem)	PSNR [↑]	SSIM [↑]	LPIPS [↓]	PSNR [↑]	SSIM [↑]	LPIPS [↓]
BungeeNeRF	22.21	0.524	0.428	19.41	0.401	0.587
BungeeNeRF(adapted)	30.42	0.903	0.194	24.07	0.770	0.379
Mip-NeRF360	30.79	0.906	0.188	24.16	0.792	0.383
Ours	32.44	0.929	0.148	26.29	0.843	0.332