

Supplementary Materials: InNeRF: Learning Interpretable Radiance Fields for Generalizable 3D Scene Representation and Rendering

Anonymous Authors

This supplementary material complements the main paper by providing detailed explanations and extended results. Sec. 1 analyzes the interpretability of the proposed InNeRF model. In Sec. 2, we describe in detail the proposed network structure and the training process. In Sec. 3, we provide additional novel view synthesis and interpretation results.

1 INNERF INTERPRETABILITY

The proposed InNeRF enhances the shape and appearance consistency in both surrounding view space and ray-cast space by utilizing the attention mechanism, hence strengthening the interpretability of our learned neural radiance field. Specifically, when rendering the density and radiance color of a query 3D point, InNeRF integrates information about the projected 2D pixels from the surrounding source views (in the $\text{Decoder}_\sigma^{\text{views}}$ and $\text{Decoder}_c^{\text{views}}$) and the neighboring 3D points along the query ray (in the $\text{Decoder}_\sigma^{\text{ray}}$ and $\text{Decoder}_c^{\text{ray}}$) by attention layers.

1.1 InNeRF interpretability in surrounding view space

The interpretability of InNeRF in surrounding view space is enhanced because its attention layers learn high-level relationships between observed views and the rendering view in the rendering process for a query 3D point. For attention layers in surrounding view space, the attention of different observed views for a query 3D point is learned to represent the contributions of observed views in rendering the query 3D point. Based on that, InNeRF can explain its rendering process in surrounding-view space by visualizing its attention layer in the $\text{Decoder}_\sigma^{\text{views}}$ and $\text{Decoder}_c^{\text{views}}$.

Mathematically, the attention function explores high-level relationships layer-by-layer between the query and the key,

$$\begin{aligned} & \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \\ &= \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,N_k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N_q,1} & \alpha_{N_q,2} & \cdots & \alpha_{N_q,N_k} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_{N_k} \end{bmatrix}, \quad (1) \\ & \text{where } \alpha_{i,j} = \frac{\exp(\mathbf{q}_i \mathbf{k}_j^T)}{\sqrt{d_k} \sum_{r=1}^{N_k} \exp(\mathbf{q}_i \mathbf{k}_r^T)}. \end{aligned}$$

The attention-score $\alpha_{i,j}$, stored in $\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})$, of a specific value \mathbf{v}_j is obtained by the match between the i -th query \mathbf{q}_i and the j -th key \mathbf{k}_j . The output for a query \mathbf{x}_i is computed by decoding the relationships between the query and key-value embeddings, i.e.

$[\mathbf{x}_1; \cdots \mathbf{x}_i \cdots; \mathbf{x}_{N_q}] = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}$, where

$$\mathbf{x}_i = \sum_{j=1}^{N_k} \alpha_{i,j} \mathbf{v}_j = \sum_{j=1}^{N_k} \frac{\exp(\mathbf{q}_i \mathbf{k}_j^T)}{\sqrt{d_k} \sum_{r=1}^{N_k} \exp(\mathbf{q}_i \mathbf{k}_r^T)} \mathbf{v}_j. \quad (2)$$

In the rendering process of the $\text{Decoder}_\sigma^{\text{views}}$ and $\text{Decoder}_c^{\text{views}}$, attention layers learn the density and color representation for a query 3D point by integrating features of its projected 2D pixels from the surrounding source views. Specifically, in the $\text{Decoder}_\sigma^{\text{views}}$, the output query \mathbf{x}_i^σ of the density for a query 3D point learns the relationships between the query density and source-view embeddings, and attention scores of source views on the query density describe the importance of source views on rendering the density of the query 3D point. Similarly, for a query 3D point, $\text{Decoder}_c^{\text{views}}$ explores the interconnections between the query color and source-view embeddings by attention scores and decodes them into the color query output.

For human perception in surrounding view space, an observed view capturing a query 3D point contains meaningful information for rendering the query 3D point and consequently contributes more than observed views that do not capture it. Accordingly, the importance of an observed view to a 3D-point rendering is positively correlated with the visibility of the 3D point in the observed view. It means the important source view should be the one that captures the target 3D point, while the trivial one does not. Therefore, we study InNeRF interpretability in surrounding view space by evaluating whether the attention in the $\text{Decoder}_\sigma^{\text{views}}$ and $\text{Decoder}_c^{\text{views}}$ is consistent with the above description of human perception.

1.2 InNeRF Interpretability in Ray-cast Space

The interpretability of InNeRF is improved in ray-cast space during the rendering of a query 3D point because attention layers take into account the 3D-scene consistency of adjacent 3D points on the target rendering ray. In the $\text{Decoder}_\sigma^{\text{ray}}$ and $\text{Decoder}_c^{\text{ray}}$, the attention of adjoining 3D points to the query 3D point is learned to represent the contributions of adjacent 3D points to the query 3D point rendering in ray-cast space. Therefore, we can visualize its attention layer in the $\text{Decoder}_\sigma^{\text{ray}}$ and $\text{Decoder}_c^{\text{ray}}$ to illustrate its rendering process in ray-cast space.

For the rendering process in ray-cast space, the attention layers of the $\text{Decoder}_\sigma^{\text{ray}}$ and $\text{Decoder}_c^{\text{ray}}$ learn the density and color representation of a query 3D point by integrating features of neighboring 3D points on the target rendering ray. For the query 3D point, the $\text{Decoder}_c^{\text{ray}}$ and $\text{Decoder}_\sigma^{\text{ray}}$ investigate the relationships between the query and adjacent 3D points on the target rendering ray by the attention function and decode them into the query output of color and density, respectively. The attention function in Eq. (1)

learns how the density and color of a query 3D point correlate with adjacent 3D points on the target rendering ray.

For 3D scene understanding, the shape and appearance consistency of a 3D scene determines the density and color correlation of neighboring 3D points on the target rendering ray. As a result, in a neural radiance field, neighboring 3D points with comparable features contribute more to rendering the query 3D point than those with less similar features. Based on the consistency in a 3D representation, the $\text{Decoder}_\sigma^{\text{ray}}$ and $\text{Decoder}_c^{\text{ray}}$ should pay more attention to the neighboring 3D points with similar representations when rendering a query 3D point in ray-cast space. We analyze InNeRF interpretability in ray-cast space to examine whether the attention in the $\text{Decoder}_\sigma^{\text{ray}}$ and $\text{Decoder}_c^{\text{ray}}$ is in accord with the geometry consistency.

In sum, the proposed InNeRF enhances interpretability from the following aspects:

- In contrast to prior work that used separate fusion and rendering modules, the proposed framework unites source-view fusion and target-view rendering processes via our Transformer-based network. Here, the Transformer attention of various source views for a query 3D point is used to learn the contribution of source views to the rendering of the query 3D point.
- Our Transformer-based rendering ($\text{Decoder}_\sigma^{\text{views}}$ and $\text{Decoder}_c^{\text{views}}$) investigates deep relationships between the target rendering view and source views, simultaneously decoding observed multi-view representations and target-rendering spatial information for a query 3D point into density and color representations, whereas previous MLP-based rendering requires an auxiliary pooling-based fusion to aggregate multiple features for source views.
- Self-attention layers in the $\text{Decoder}_\sigma^{\text{views}}$ learn source-view representations by considering global correlations among source views rather than learning each source view separately.
- In contrast to prior work that processed each 3D point separately, our InNeRF takes into account nearby 3D points on the target rendering ray by using attention layers when rendering a query 3D point.

2 IMPLEMENTATION DETAILS

2.1 Network details

We use a shared network to extract features of each source view and the shared network is of a U-Net architecture, as in [6]. Our InNeRF network contains four sub-modules: $\text{Decoder}_\sigma^{\text{views}}$, $\text{Decoder}_\sigma^{\text{ray}}$, $\text{Decoder}_c^{\text{views}}$, and $\text{Decoder}_c^{\text{ray}}$, where the first two are for the density representation by considering surrounding source views, and the last two are for the color representation by considering neighboring points along the target ray. The sub-modules (Sec. 3.2, 3.3, 3.4, 3.5 in the main text) are adapted from the vanilla Transformer network, and each of them contains $L = 2$ blocks. The first two sub-modules have $H = 4$ heads in attention layers and $d = 32$ dimensions for each token feature while the last two have $H = 5$ heads in attention layers and $d = 45$ dimensions for token features.

2.2 Rendering and Training

As in NeRF [4], a differentiable ray marching rendering is utilized to render a 2D image from our radiance field scene representation F_{InNeRF} . To calculate a pixel color at a rendered image, classical volume rendering marches a ray via the pixel and accumulates radiance at sampled 3D points along the ray. Specifically, the rendered color $C(\mathbf{r})$ of the camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with near and far bounds $[t_n, t_f]$ is computed as:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t); \mathbf{I}, \Theta) \mathbf{c}(\mathbf{r}(t), \mathbf{d}; \mathbf{I}, \Theta) dt, \quad (3)$$

where the function $T(t) = \exp(-\int_{t_n}^t \sigma(\mathbf{r}(s); \mathbf{I}, \Theta) ds)$ calculates the accumulated transmittance along the ray between depth bounds $[t_n, t_f]$.

The rendering loss between the rendered and ground-truth pixel colors for all camera rays of the target view with camera pose Θ is

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}(\Theta)} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|_2^2, \quad (4)$$

where $\mathcal{R}(\Theta)$ is the set of all camera rays of the desired virtual camera with pose Θ .

Given a set of observed source views, the rendering loss \mathcal{L} between the ground truth and the prediction is minimized by optimizing the overall network parameters. Here, our InNeRF is fully differentiable and can be trained end-to-end. After being trained on a large dataset containing diverse scenes, the proposed InNeRF effectively generalizes to unseen scenes even without per-scene finetuning. Accordingly, given the camera pose of a query view of a scene and its multiple source views, InNeRF can render the query view from the pre-trained general neural radiance field by casting rays from the target camera center towards the 3D space using volume rendering in Eq. (3).

We present a generic Transformer-based NeRF framework with a high capacity in modeling radiance field scene representation from observed source view images. Similar to MLP-based NeRFs, our framework can be easily extended to advanced InNeRF derivatives, e.g. from NeRF to NeRF-W [2] by adding appearance and transient variables as inputs to model uncertainty in the wild scenarios.

2.3 Training details

We train and evaluate the proposed network on a collection of several multi-view datasets containing both synthetic data and real data, as in IBRNet [6].

- **Real datasets for training** include the Spaces dataset [1], RealEstate10K [7], and the handheld-cellphone-captured scene dataset. The Spaces dataset has 100 scenes and each scene is collected with a 16-camera rig at 3 to 10 rig positions. RealEstate10K, a large indoor-scene dataset, is captured from around 80K video clips with camera poses. The cellphone-captured scene dataset contains 95 real scenes (36 from LLFF [3] and 59 from IBRNet [6]), where each scene consists of 20 to 60 forward-facing images. COLMAP [5] is adopted to estimate camera poses, intrinsic parameters and scene bounds for real data.

- **Synthetic dataset for training** is generated by IBRNet from Google Scanned Objects, which contains 1,030 models with a variety of view density rates.
- **Real dataset for evaluation** collects 8 complex real-world scenes captured with a handheld cellphone (5 from LLFF [3] and 3 from NeRF [4]). Each scene consists of 20 to 62 forward-facing images with 1008×756 pixels, and 1/8th of these is held out as the test set (7/8 for per-scene fine-tuning).
- **Synthetic dataset for evaluation**, adopted from NeRF [4], includes 8 objects with complicated geometry and realistic non-Lambertian materials which are rendered at 800×800 pixels from viewpoints sampled either on the upper hemisphere or full sphere (100 views for per-scene finetuning and 200 views for testing).

During training, we randomly sample M source views for each target view from a pool of $m \times M$ views where M is sampled uniformly at random from [8, 12] and m is from [1, 5]. We adopt the hierarchical volume sampling strategy from NeRF [4] to achieve more efficient sampling. For the hierarchical volume sampling, we construct a coarse InNeRF and a fine InNeRF with identical network architecture. At each optimization iteration, we randomly sample a batch of camera rays from the set of all pixels in the dataset (batch size = 2048 rays), and then sample $N_c = 64$ points at the coarse scale and $N_f = 64$ additional points at the fine-scale for each ray. After rendering the color of each ray from both scale sets of samples by the volume rendering, we can train the feature extraction network as well as the coarse and fine InNeRF simultaneously by minimizing the mean squared error between ground-truth pixel colors and the rendered colors in coarse and fine scales. Both the feature extraction network and the InNeRF are optimized using AdamW. The initial learning rates of the feature extraction network and InNeRF are 10^{-3} and 5×10^{-4} , respectively.

3 RESULTS

3.1 Qualitative and quantitative results

We provide more novel view synthesis results of competing methods. Fig. 1 and 2 show the results for the Trex and Fern scenes in the scene-agnostic setting, and Fig. 3 and 4 show the synthesized views in the per-scene fine-tuning setting. In these figures, the i -th column exhibits the rendering results based on source view set S_i . As highlighted in the colored frames, the qualitative comparison suggests the superiority of InNeRF in terms of the rendering appearance and details.

We also provide more comprehensive quantitative results. Tab. 1 and 2 show quantitative results for the synthetic and real datasets in the scene-agnostic setting. The quantitative results of the per-scene finetuning setting for the synthetic and real datasets are shown in Tab. 3 and 4, respectively. The best score for each scene is highlighted in bold.

3.2 Interpretation results

This section includes more interpretation results of InNeRF.

In Fig. 5, we give an interpretation for the same target rendering view of the chair scene based on the source-view set S_1 . As shown

in Fig. 5 (a), the difference between source views in S_1 and the rendering view is smaller than that between S_4 and the rendering view (in the main text). Fig. 5 (b) and (c) show density attention and color attention to different source views for the identical red-box region in the rendering view. Similarly, source views (2, 72, 28, and 95) with high attention values are consistent with those with the visible region of interest shown in Fig. 5 (a), indicating that the attentions in $\text{Decoder}_\sigma^{\text{views}}$ and $\text{Decoder}_c^{\text{views}}$ are in accordance with human perception. Fig. 5 (d) shows that $\text{Decoder}_\sigma^{\text{ray}}$ and $\text{Decoder}_c^{\text{ray}}$ utilize the consistency of 3D points on the target rendering ray for the query point rendering.

REFERENCES

- [1] John Flynn, Michael Broxton, Paul E Debevec, Matthew DuVall, Graham Fyffe, Ryan S Overbeck, Noah Snavely, and Richard Tucker. 2019. DeepView: View Synthesis With Learned Gradient Descent.. In *CVPR*.
- [2] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7210–7219.
- [3] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *ACM Transactions on Graphics (TOG)* (2019).
- [4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 405–421.
- [5] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.
- [6] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. 2021. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4690–4699.
- [7] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–12.



Figure 1: Qualitative results for the Trex scene under the scene-agnostic setting.

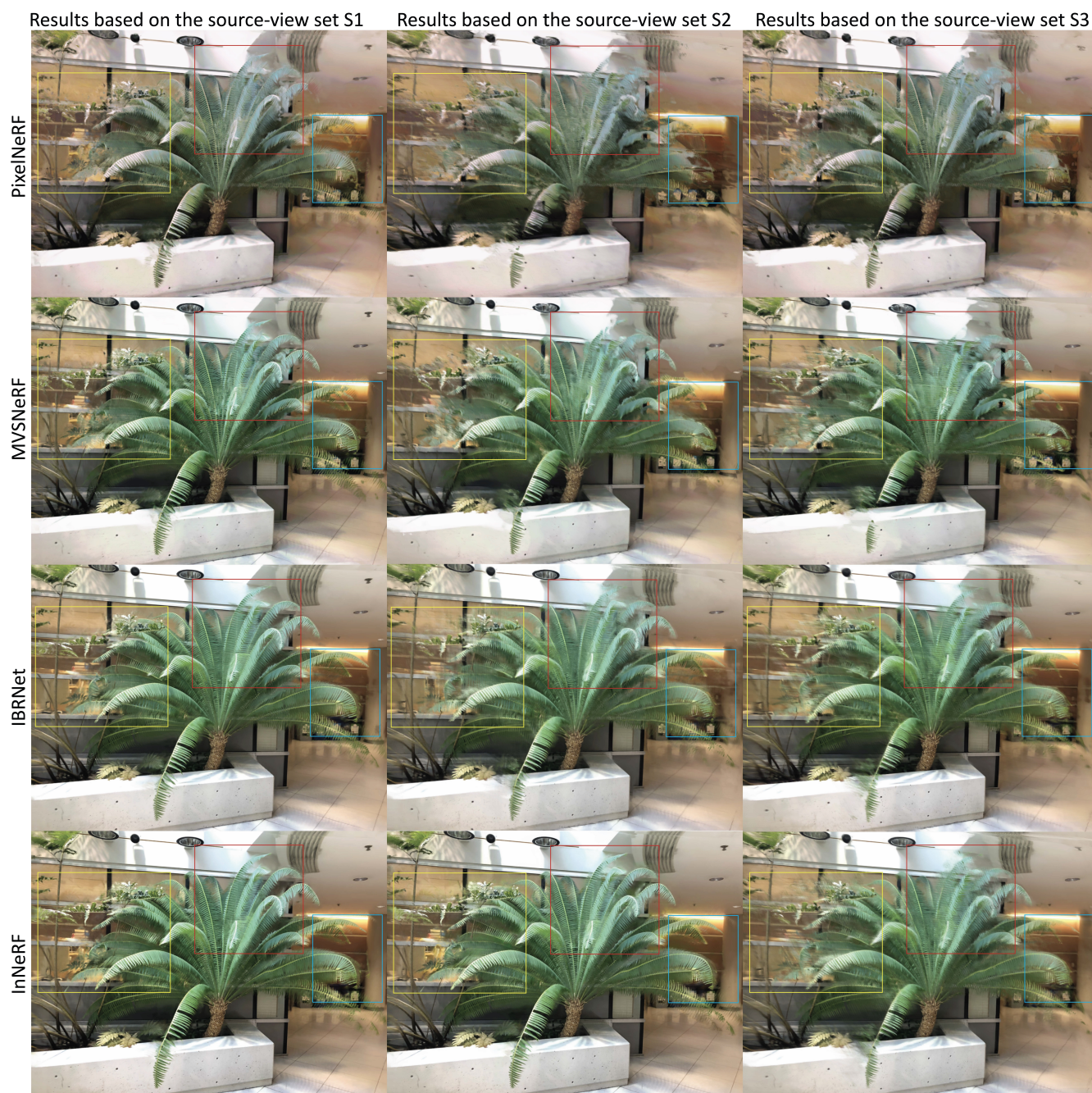


Figure 2: Qualitative results for the Fern scene under the scene-agnostic setting.

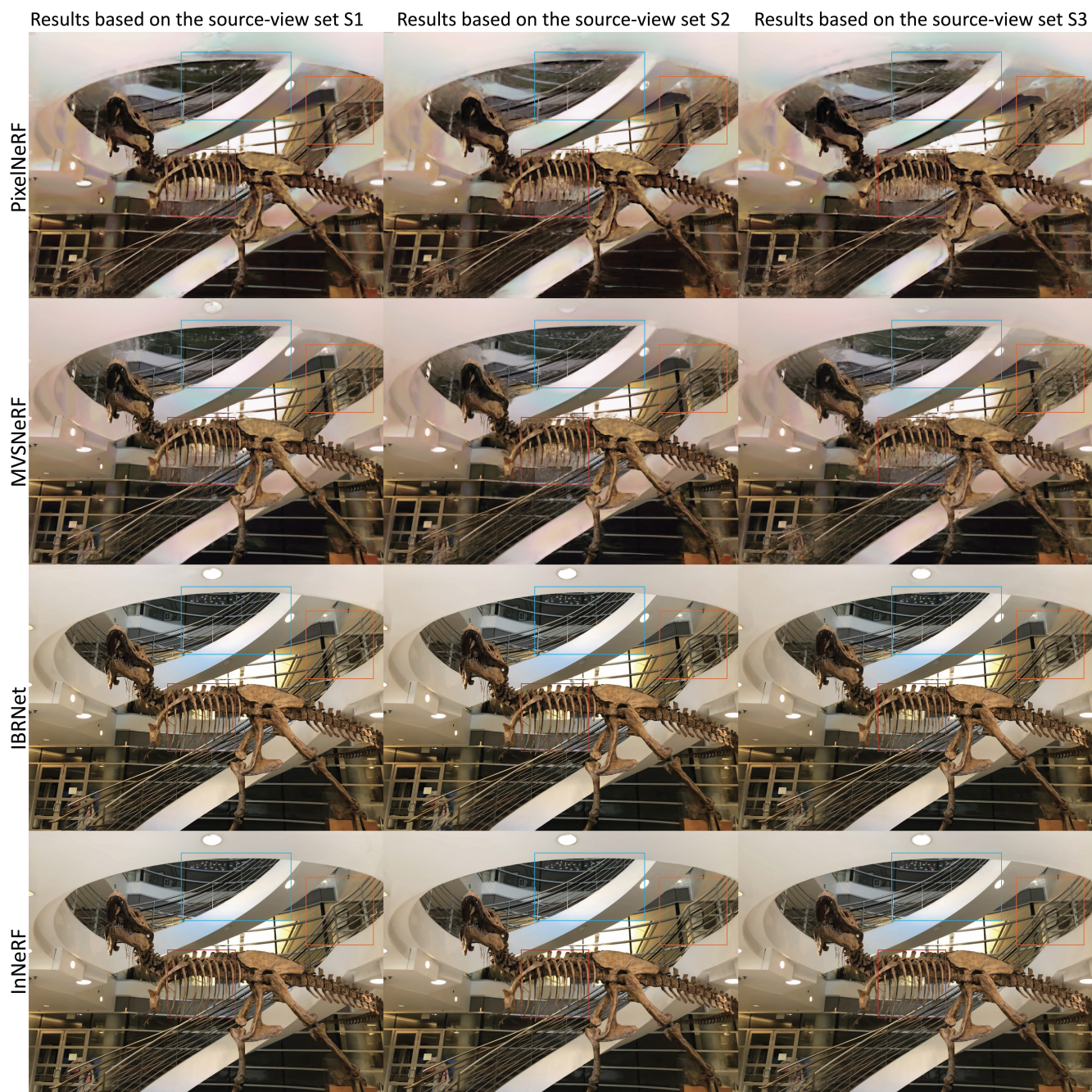


Figure 3: Qualitative results for the Trex scene under the per-scene fine-tuning setting.

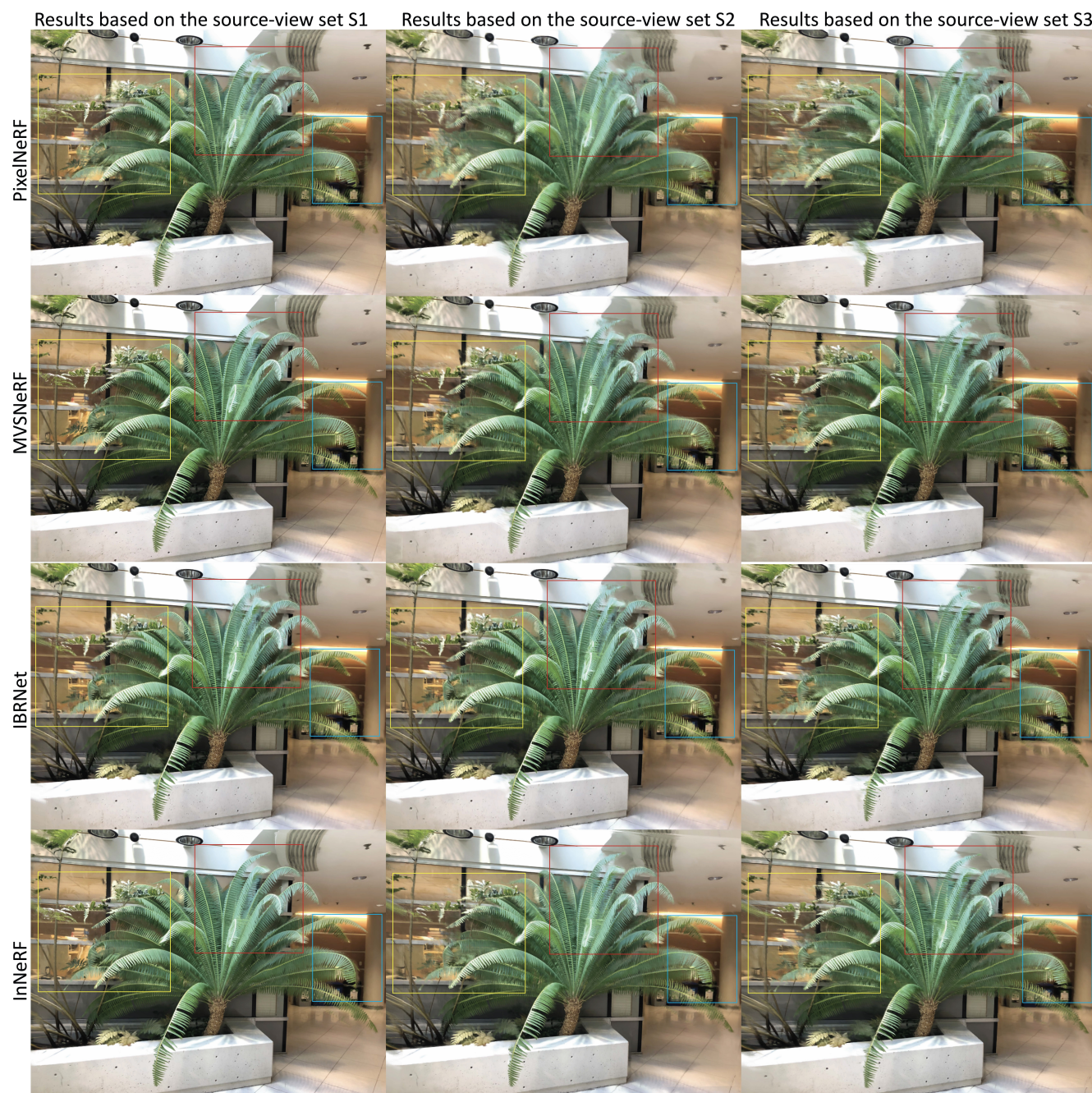


Figure 4: Qualitative results for the Fern scene under the per-scene fine-tuning setting.

Table 1: Quantitative comparisons of competing methods for the scene-agnostic setting on the realistic synthetic dataset.

		PSNR \uparrow				SSIM \uparrow				LPIPS \downarrow			
Scene	S_i	PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF
Chair	S1	21.21	23.50	28.55	29.06	0.890	0.910	0.942	0.954	0.135	0.108	0.066	0.055
	S2	16.98	19.38	24.93	25.79	0.784	0.811	0.854	0.903	0.240	0.224	0.173	0.135
	S3	15.76	18.37	24.12	25.17	0.712	0.745	0.798	0.870	0.298	0.267	0.206	0.164
	S4	14.51	17.21	23.12	24.85	0.612	0.647	0.714	0.823	0.387	0.338	0.259	0.194
Lego	S1	19.51	21.13	24.64	26.13	0.834	0.862	0.917	0.923	0.174	0.148	0.106	0.090
	S2	15.30	17.08	21.14	22.80	0.685	0.717	0.787	0.851	0.331	0.311	0.257	0.179
	S3	14.21	16.08	20.33	22.18	0.613	0.650	0.731	0.819	0.390	0.354	0.291	0.207
	S4	13.04	14.92	19.33	21.10	0.532	0.571	0.665	0.768	0.471	0.421	0.339	0.246
Ship	S1	21.31	21.79	22.92	24.57	0.803	0.808	0.825	0.836	0.267	0.256	0.227	0.211
	S2	17.04	17.69	19.37	21.25	0.650	0.658	0.689	0.725	0.443	0.438	0.397	0.305
	S3	15.98	16.69	18.56	20.64	0.577	0.591	0.633	0.693	0.503	0.480	0.431	0.333
	S4	14.83	15.59	17.63	19.34	0.481	0.497	0.552	0.637	0.599	0.565	0.495	0.370
Drums	S1	17.60	18.74	21.28	21.99	0.771	0.807	0.893	0.902	0.184	0.164	0.113	0.105
	S2	13.85	15.14	18.21	18.69	0.642	0.683	0.782	0.802	0.321	0.312	0.252	0.201
	S3	12.72	14.16	17.40	18.49	0.570	0.616	0.725	0.770	0.383	0.354	0.285	0.230
	S4	11.53	12.99	16.40	18.18	0.480	0.528	0.652	0.724	0.474	0.434	0.346	0.254
Mic	S1	26.51	27.28	28.93	29.32	0.929	0.938	0.951	0.968	0.071	0.062	0.045	0.033
	S2	22.64	23.61	25.80	26.09	0.852	0.864	0.891	0.914	0.165	0.165	0.135	0.135
	S3	21.46	22.60	24.99	25.85	0.779	0.797	0.835	0.881	0.226	0.207	0.169	0.164
	S4	20.53	21.61	24.15	25.65	0.686	0.706	0.758	0.816	0.295	0.265	0.207	0.197
Ficus	S1	21.86	22.73	24.72	25.14	0.899	0.904	0.919	0.923	0.125	0.117	0.089	0.077
	S2	18.60	19.67	22.20	22.70	0.793	0.801	0.830	0.831	0.227	0.228	0.189	0.159
	S3	17.51	18.65	21.39	22.56	0.720	0.735	0.774	0.799	0.288	0.270	0.222	0.187
	S4	16.05	17.27	20.15	22.14	0.626	0.642	0.695	0.746	0.368	0.339	0.271	0.235
Materials	S1	19.47	19.93	20.98	22.67	0.828	0.847	0.895	0.903	0.195	0.174	0.124	0.110
	S2	14.41	14.91	16.48	19.15	0.654	0.683	0.739	0.807	0.357	0.348	0.288	0.198
	S3	13.15	13.92	15.66	18.54	0.581	0.615	0.683	0.774	0.420	0.390	0.322	0.227
	S4	11.89	12.66	14.58	16.25	0.486	0.521	0.603	0.729	0.514	0.472	0.386	0.267
Hotdog	S1	22.14	24.70	30.45	32.70	0.902	0.919	0.958	0.968	0.135	0.115	0.066	0.054
	S2	17.20	20.02	26.29	28.82	0.797	0.817	0.871	0.904	0.275	0.259	0.198	0.149
	S3	16.22	18.99	25.48	28.20	0.724	0.750	0.814	0.871	0.332	0.301	0.232	0.178
	S4	14.98	17.79	24.43	26.27	0.636	0.662	0.740	0.825	0.410	0.365	0.276	0.216
Ave	S1	21.20	22.47	25.31	26.45	0.857	0.874	0.913	0.922	0.161	0.143	0.104	0.092
	S2	17.00	18.44	21.80	23.16	0.732	0.755	0.805	0.842	0.295	0.286	0.236	0.183
	S3	15.88	17.43	20.99	22.70	0.660	0.687	0.749	0.810	0.355	0.328	0.270	0.211
	S4	14.67	16.25	19.97	21.72	0.567	0.597	0.672	0.758	0.440	0.400	0.322	0.248

Table 2: Quantitative comparisons of methods for the scene-agnostic setting on the real forward-facing dataset.

Scene	S_i	PSNR \uparrow				SSIM \uparrow				LPIPS \downarrow			
		PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF
Fern	S1	20.65	21.12	23.69	23.70	0.671	0.696	0.767	0.771	0.355	0.322	0.250	0.247
	S2	18.69	19.40	22.18	22.35	0.602	0.639	0.709	0.720	0.427	0.389	0.307	0.295
	S3	17.99	18.80	21.96	22.11	0.574	0.607	0.703	0.704	0.460	0.410	0.316	0.304
Trex	S1	18.63	19.24	23.83	23.84	0.705	0.722	0.849	0.850	0.392	0.377	0.239	0.237
	S2	16.03	16.88	21.68	21.85	0.633	0.660	0.788	0.797	0.467	0.448	0.299	0.287
	S3	13.80	15.18	19.94	20.51	0.561	0.611	0.737	0.764	0.544	0.516	0.352	0.328
Horns	S1	20.01	21.35	26.34	26.35	0.702	0.739	0.866	0.868	0.356	0.323	0.179	0.179
	S2	16.78	18.37	23.56	23.74	0.626	0.680	0.802	0.812	0.438	0.401	0.247	0.236
	S3	14.06	15.99	21.32	21.71	0.537	0.610	0.734	0.760	0.531	0.482	0.316	0.293
Fortress	S1	22.37	24.31	29.97	29.98	0.719	0.732	0.879	0.880	0.327	0.287	0.155	0.152
	S2	20.14	22.36	28.19	28.40	0.683	0.701	0.853	0.862	0.368	0.325	0.180	0.170
	S3	15.15	17.97	23.68	24.36	0.560	0.595	0.751	0.767	0.498	0.446	0.287	0.267
Leaves	S1	15.56	16.37	20.30	20.31	0.513	0.561	0.722	0.724	0.418	0.378	0.228	0.226
	S2	13.07	14.23	18.26	18.54	0.395	0.456	0.615	0.627	0.529	0.484	0.324	0.309
	S3	11.20	12.73	16.87	17.39	0.303	0.394	0.544	0.578	0.607	0.550	0.378	0.355
Orchids	S1	15.57	16.14	19.25	19.26	0.464	0.492	0.629	0.631	0.461	0.422	0.294	0.291
	S2	13.64	14.58	17.76	18.07	0.371	0.410	0.547	0.558	0.560	0.517	0.378	0.363
	S3	12.28	13.42	16.89	17.26	0.299	0.345	0.496	0.507	0.627	0.575	0.421	0.399
Room	S1	21.52	22.74	29.70	29.71	0.820	0.835	0.941	0.944	0.318	0.293	0.155	0.153
	S2	17.81	19.43	26.45	26.78	0.775	0.788	0.907	0.916	0.379	0.352	0.202	0.191
	S3	13.71	15.64	22.83	23.33	0.699	0.738	0.852	0.870	0.479	0.442	0.278	0.259
Flower	S1	17.84	19.46	26.61	26.63	0.614	0.663	0.854	0.858	0.414	0.376	0.165	0.157
	S2	14.21	16.23	23.43	23.76	0.524	0.580	0.775	0.787	0.507	0.464	0.243	0.229
	S3	10.28	11.97	19.15	19.85	0.374	0.446	0.646	0.662	0.662	0.607	0.374	0.336
Ave	S1	19.02	20.09	24.96	24.97	0.651	0.680	0.813	0.816	0.380	0.347	0.208	0.205
	S2	16.30	17.68	22.69	22.94	0.576	0.614	0.749	0.760	0.459	0.422	0.273	0.260
	S3	13.56	15.21	20.33	20.81	0.489	0.543	0.683	0.701	0.551	0.504	0.340	0.318

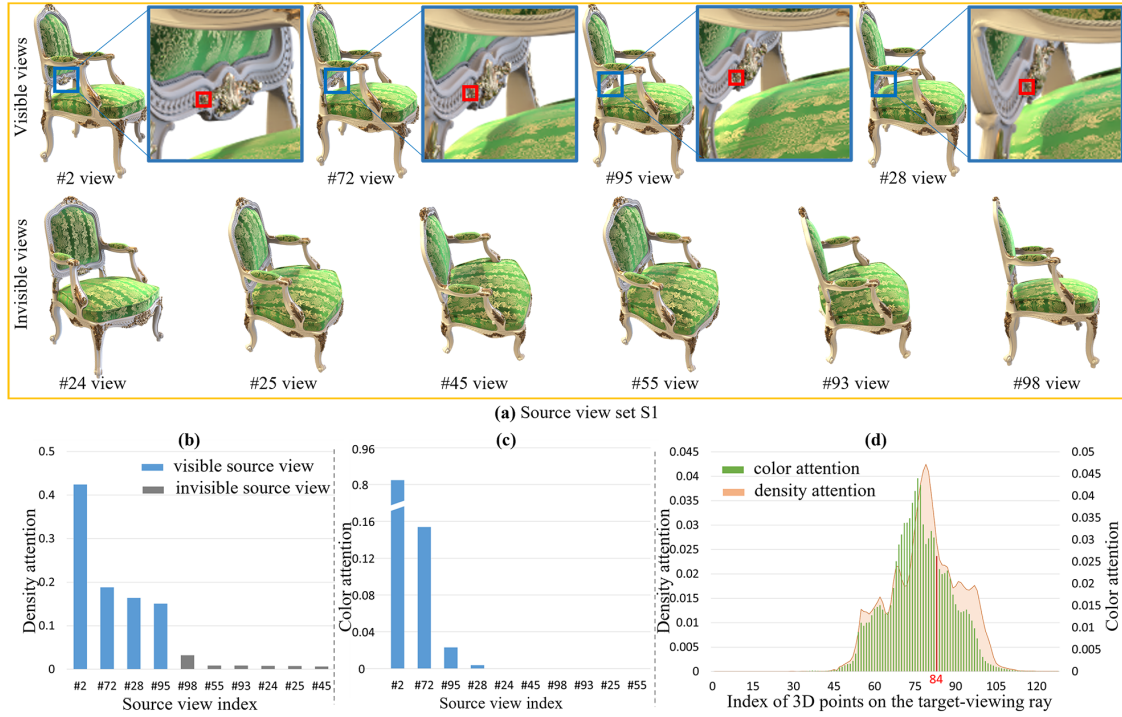
**Figure 5: Interpretation results of finetuned InNeRF for a target view of Chair scene based on source-view set S1.**

Table 3: Quantitative comparisons of methods for the per-scene fine-tuning setting on the realistic synthetic dataset.

		PSNR \uparrow				SSIM \uparrow				LPIPS \downarrow			
Scene	S_i	PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF
Chair	S1	23.95	29.48	32.62	33.49	0.910	0.950	0.972	0.989	0.114	0.071	0.044	0.038
	S2	19.84	25.54	29.06	30.45	0.820	0.870	0.896	0.942	0.205	0.156	0.123	0.096
	S3	18.93	24.78	28.35	30.21	0.764	0.817	0.852	0.911	0.244	0.192	0.153	0.118
	S4	17.90	23.80	27.48	29.84	0.683	0.738	0.780	0.868	0.300	0.241	0.190	0.143
Lego	S1	22.50	26.34	28.65	30.46	0.854	0.916	0.947	0.959	0.153	0.099	0.068	0.056
	S2	18.58	22.58	25.28	27.38	0.722	0.790	0.829	0.890	0.304	0.246	0.207	0.140
	S3	17.67	21.82	24.57	26.91	0.664	0.734	0.782	0.859	0.343	0.283	0.237	0.162
	S4	16.64	20.84	23.69	26.02	0.601	0.676	0.731	0.814	0.394	0.327	0.269	0.194
Ship	S1	24.22	25.33	26.83	28.90	0.823	0.838	0.855	0.870	0.246	0.216	0.188	0.177
	S2	20.34	21.61	23.49	25.84	0.688	0.707	0.733	0.763	0.416	0.381	0.346	0.265
	S3	19.45	20.85	22.80	25.15	0.631	0.653	0.687	0.731	0.454	0.418	0.376	0.287
	S4	18.48	19.94	21.98	24.26	0.551	0.578	0.619	0.680	0.523	0.479	0.425	0.317
Drums	S1	20.27	22.96	25.20	26.32	0.791	0.861	0.922	0.935	0.164	0.116	0.076	0.053
	S2	16.67	19.53	22.15	23.27	0.677	0.757	0.823	0.840	0.302	0.248	0.202	0.161
	S3	15.96	18.96	21.64	23.22	0.622	0.705	0.779	0.808	0.341	0.284	0.232	0.183
	S4	14.93	17.98	20.76	23.11	0.551	0.637	0.718	0.767	0.404	0.339	0.276	0.201
Mic	S1	29.39	31.65	32.77	33.64	0.949	0.970	0.981	0.983	0.050	0.039	0.033	0.027
	S2	25.84	28.26	29.76	30.68	0.889	0.914	0.934	0.952	0.124	0.109	0.095	0.084
	S3	25.11	27.67	29.24	30.66	0.831	0.860	0.887	0.921	0.156	0.138	0.118	0.114
	S4	24.24	26.85	28.51	30.57	0.756	0.790	0.824	0.861	0.202	0.177	0.145	0.139
Ficus	S1	24.74	26.68	28.54	29.46	0.919	0.933	0.949	0.957	0.104	0.076	0.051	0.044
	S2	21.80	23.90	26.15	27.29	0.825	0.844	0.869	0.872	0.203	0.171	0.137	0.119
	S3	21.07	23.32	25.62	27.19	0.771	0.792	0.826	0.838	0.242	0.206	0.167	0.141
	S4	19.81	22.11	24.51	27.06	0.695	0.721	0.762	0.791	0.295	0.253	0.202	0.184
Materials	S1	22.27	23.32	24.98	26.99	0.848	0.889	0.925	0.938	0.174	0.130	0.098	0.077
	S2	17.27	18.49	20.54	23.74	0.688	0.740	0.781	0.845	0.325	0.275	0.237	0.158
	S3	16.44	17.80	19.90	22.95	0.633	0.686	0.735	0.814	0.364	0.310	0.267	0.181
	S4	15.33	16.73	18.94	21.18	0.556	0.612	0.669	0.774	0.432	0.371	0.316	0.216
Hotdog	S1	25.14	30.58	34.54	37.03	0.922	0.948	0.973	0.986	0.114	0.077	0.048	0.041
	S2	20.90	26.51	30.85	33.40	0.850	0.881	0.915	0.943	0.226	0.185	0.147	0.110
	S3	19.56	25.32	29.71	32.54	0.792	0.826	0.867	0.910	0.264	0.220	0.177	0.132
	S4	18.49	24.30	28.79	31.19	0.719	0.758	0.806	0.869	0.314	0.262	0.207	0.164
Ave	S1	24.06	27.04	29.27	30.79	0.877	0.913	0.940	0.952	0.140	0.103	0.076	0.064
	S2	20.15	23.30	25.91	27.76	0.770	0.813	0.847	0.881	0.263	0.221	0.187	0.142
	S3	19.27	22.56	25.23	27.35	0.714	0.759	0.802	0.849	0.301	0.256	0.216	0.165
	S4	18.23	21.57	24.33	26.65	0.639	0.689	0.739	0.803	0.358	0.306	0.254	0.195

Table 4: Quantitative comparisons of methods for the per-scene finetuning setting on the real forward-facing dataset.

		PSNR \uparrow				SSIM \uparrow				LPIPS \downarrow			
Scene	S_i	PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF	PixelNeRF	MVSNeRF	IBRNet	InNeRF
Fern	S1	22.35	23.43	24.89	24.92	0.713	0.749	0.803	0.812	0.300	0.262	0.212	0.208
	S2	20.52	21.86	23.60	23.89	0.653	0.696	0.752	0.775	0.351	0.308	0.253	0.236
	S3	19.77	21.06	23.51	23.85	0.630	0.680	0.749	0.769	0.378	0.324	0.257	0.240
Trex	S1	20.33	23.21	26.43	26.47	0.747	0.817	0.896	0.900	0.338	0.271	0.202	0.198
	S2	17.74	20.87	24.38	24.68	0.684	0.759	0.843	0.864	0.394	0.323	0.248	0.235
	S3	15.76	19.20	23.07	23.76	0.622	0.705	0.801	0.841	0.455	0.375	0.287	0.268
Horns	S1	21.71	24.34	28.27	28.31	0.744	0.801	0.898	0.907	0.300	0.241	0.145	0.142
	S2	18.92	21.79	26.01	26.31	0.679	0.742	0.842	0.864	0.361	0.295	0.195	0.183
	S3	16.18	19.34	23.94	24.61	0.595	0.667	0.779	0.829	0.435	0.359	0.247	0.228
Fortress	S1	24.07	26.94	30.99	31.05	0.761	0.822	0.894	0.901	0.272	0.216	0.143	0.139
	S2	22.51	25.60	29.96	30.27	0.738	0.804	0.881	0.896	0.295	0.235	0.155	0.144
	S3	17.94	21.30	26.05	26.72	0.625	0.698	0.787	0.830	0.400	0.330	0.238	0.217
Leaves	S1	17.26	19.22	21.65	21.69	0.555	0.637	0.773	0.776	0.363	0.283	0.189	0.186
	S2	14.78	16.97	19.70	19.99	0.432	0.523	0.659	0.691	0.450	0.364	0.267	0.256
	S3	12.80	15.33	18.39	19.10	0.351	0.447	0.598	0.634	0.512	0.415	0.305	0.284
Orchids	S1	17.27	19.03	20.70	20.74	0.506	0.596	0.682	0.691	0.406	0.316	0.233	0.230
	S2	15.51	17.50	19.46	19.77	0.420	0.514	0.605	0.632	0.485	0.390	0.302	0.289
	S3	14.04	16.30	18.67	19.32	0.349	0.452	0.555	0.589	0.539	0.434	0.333	0.312
Room	S1	23.22	27.59	32.03	32.08	0.862	0.904	0.955	0.960	0.263	0.205	0.140	0.137
	S2	19.95	24.56	29.30	29.59	0.828	0.875	0.930	0.947	0.304	0.243	0.172	0.160
	S3	16.56	21.42	26.58	27.20	0.766	0.820	0.888	0.920	0.378	0.306	0.222	0.203
Flower	S1	19.54	22.77	27.89	27.93	0.656	0.740	0.872	0.878	0.359	0.284	0.148	0.145
	S2	16.27	19.74	25.15	25.45	0.567	0.657	0.792	0.818	0.429	0.348	0.209	0.196
	S3	12.24	15.00	20.79	21.46	0.414	0.512	0.659	0.721	0.565	0.473	0.322	0.299
Ave	S1	20.72	23.32	26.61	26.65	0.693	0.758	0.847	0.853	0.325	0.260	0.177	0.173
	S2	18.28	21.11	24.69	24.99	0.625	0.696	0.788	0.811	0.384	0.313	0.225	0.212
	S3	15.66	18.62	22.62	23.25	0.544	0.623	0.727	0.767	0.458	0.377	0.276	0.256