Supplementary Details for the Proposed NeXT

A Comparisons to NeRF-ID and IBRNet

As far as we know, Transformer in NeRF is explored by NeRF-ID [1] and IBRNet [12]. NeRF-ID aims to learn to propose samples via a differentiable module (e.g., Transformer, Pool, MLPMix [11]), while still remains MLP as the query network. In contrast, this paper explores a pure Transformer-based paradigm with ray-level query for Novel View Synthesis (NVS), which is complementary to NeRF-ID [1].

More related to our work, Transformer is also utilized in IBRNet [12] for NVS. We point out that our method is essentially different from IBRNet:

- **Core spirit.** IBRNet [12] is proposed for synthesizing novel views of complex scenes by interpolating a sparse set of nearby views, where multi-view 2D image features are indispensable. By contrast, our proposed NeXT aims to design a general paradigm to achieve ray-level query for high quality renderings, which is complementary to NeRF [7] as well as its most follow-ups and can significantly boosts their performance.
- **Architecture.** IBRNet [12] utilizes image features and decouples the predictions of color and density, resulting in a hybrid network architecture, *i.e.*, an additional U-Net [8] based convolutional neural network to extract dense features, an MLP for color outputs and a Transformer for density outputs. Differently, NeXT is a pure Transformer-based paradigm to predict the RGB color and density concurrently, where multi-skip connection is proposed to enriches the position information for high quality renderings.
- **Quantitative results.** We conduct additional comparisons between IBRNet (with per-scene fine-tuning) and NeXT-S on DeepVoxels [9] and Blender [7] datasets. As shown in Table 1, NeXT-S outperforms IBRNet with **3.02** and **4.20** PSNR gain on DeepVoxels and Blender dataset.

B Calculation of Model FLOPs

Note that we adopt a coarse-to-fine sampling strategy following NeRF [7]. N_c and N_f points are sampled in the coarse and fine stage respectively. Thus, to render a pixel from the corresponding ray, model FLOPs is calculated by summing the query cost for coarse network $\mathbf{F}_c$ and fine network $\mathbf{F}_f$, formulated as:

$$\mathbf{FLOPs} = \mathbf{FLOPs}(\mathbf{F}_c(\mathbf{x}_1, ..., \mathbf{x}_c) + \mathbf{FLOPs}(\mathbf{F}_f(\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_f)), \tag{1}$$

where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are the input sample points of the coarse and fine stage, respectively.

Table 1: **Comparisons on DeepVoxels and Blender dataset.** NeXT-S significantly outperforms IBRNet by a large margin.

Method	Extra CNN	DeepVoxels		Blender	
		PSNR	SSIM	PSNR	SSIM
NeRF [7]	✗	40.15	0.991	31.01	0.947
IBRNet [12]	✓	42.93	**0.997**	28.14	0.942
NeXT-S	✗	**45.95**	0.996	**32.34**	**0.955**

C Additional Results

Blender dataset. Additional test SSIM metric for Blender dataset is shown in Table 2. NeXT variants achieve higher structural similarity than NeRF. Additional renderings produced by NeXT variants compared to the groundtruth and NeRF can be found in Fig. 1 and Fig. 2.

Table 2: **SSIM comparisons on Blender dataset.** "*" means adopting center pixel [2]. NeXT variants surpass previous state-of-the-art methods.

	#Params	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Avg.
SRN [10]	-	0.910	0.766	0.849	0.923	0.809	0.808	0.947	0.757	0.846
NV [5]	-	0.916	0.873	0.910	0.944	0.880	0.888	0.946	0.784	0.893
LLFF [6]	-	0.948	0.890	0.896	0.965	0.911	0.890	0.964	0.823	0.911
NSVF [4]	3.2M-16M	0.968	0.931	0.973	0.980	0.960	0.973	0.987	0.854	0.953
NeRF [7]	1,191K	0.967	0.925	0.964	0.974	0.961	0.949	0.980	0.856	0.947
NeRF (JAX) [3]	1,191K	0.975	0.927	0.967	0.979	0.968	0.952	0.987	0.868	0.953
vanilla Trans.	1,889K	0.970	0.923	0.964	0.971	0.966	0.968	0.981	0.845	0.949
NeXT-S	1,232K	0.971	0.927	0.975	0.979	0.971	0.969	0.983	0.864	0.955
NeXT-B	2,152K	0.977	0.934	0.981	0.982	0.978	0.972	0.986	0.876	0.961
NeXT-L	4,062K	0.985	0.943	0.986	0.983	0.982	0.980	0.988	0.887	0.967
NeXT-L*	4,062K	**0.986**	**0.945**	**0.987**	**0.984**	**0.984**	**0.980**	**0.991**	**0.891**	**0.969**

Multiscale Blender dataset. To evaluate NeXT variants versus NeRF [7] and Mip-NeRF[2] on each individual scene of multiscle Blender dataset, the PSNR and SSIM metrics are presented in Table 3. NeXT outperforms Mip-NeRF by a clear margin across all scenes.

Table 3: **Per-scene comparisons on multiscale Blender dataset.** NeXT outperforms Mip-NeRF by a clear margin across all scenes.

		Average PSNR								
	#Params	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Avg.
NeRF (JAX) [3]	1,191K	29.92	23.27	27.15	32.00	27.75	26.30	28.40	26.46	27.66
Mip-NeRF [2]	612K	37.14	27.02	33.19	39.31	35.74	32.56	38.04	33.08	34.51
NeXT-S	616K	37.60	27.54	33.20	40.36	36.30	35.36	37.72	32.92	35.13
NeXT-B	1,076K	38.32	27.91	34.05	40.84	37.26	35.85	38.27	33.54	35.76
NeXT-L	2,031K	**39.73**	**28.85**	**35.74**	**41.74**	**38.76**	**37.41**	**39.87**	**34.56**	**37.08**

		Average SSIM								
	#Params	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Avg.
NeRF (JAX) [3]	1,191K	0.944	0.891	0.942	0.959	0.926	0.934	0.958	0.861	0.927
Mip-NeRF [2]	612K	0.988	0.945	0.984	0.988	0.984	0.977	0.993	0.922	0.973
NeXT-S	616K	0.986	0.948	0.982	0.988	0.984	0.986	0.990	0.915	0.972
NeXT-B	1,076K	0.989	0.953	0.985	0.990	0.987	0.988	0.992	0.923	0.976
NeXT-L	2,031K	**0.993**	**0.961**	**0.990**	**0.992**	**0.991**	**0.991**	**0.995**	**0.935**	**0.981**

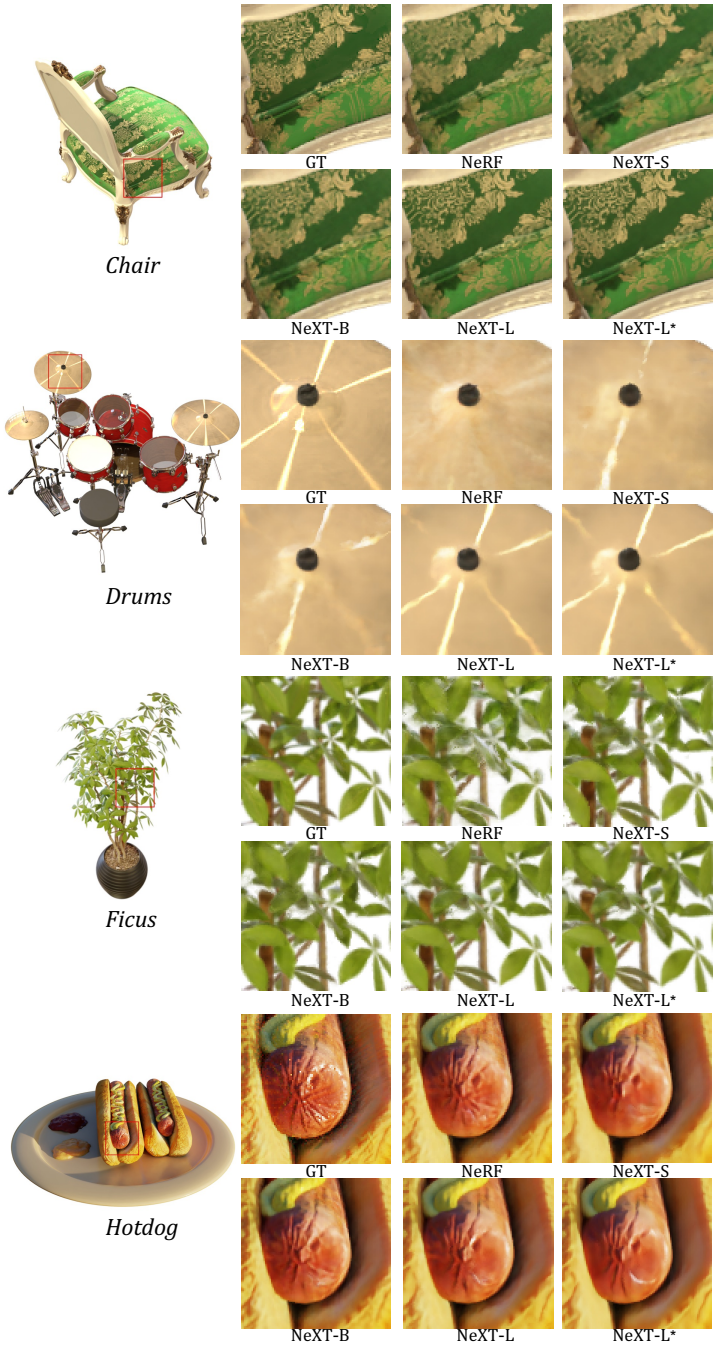

Fig. 1: Visualization of renderings from NeXT variants versus groundtruth and NeRF. Cropped regions on four scenes of Blender dataset are presented.

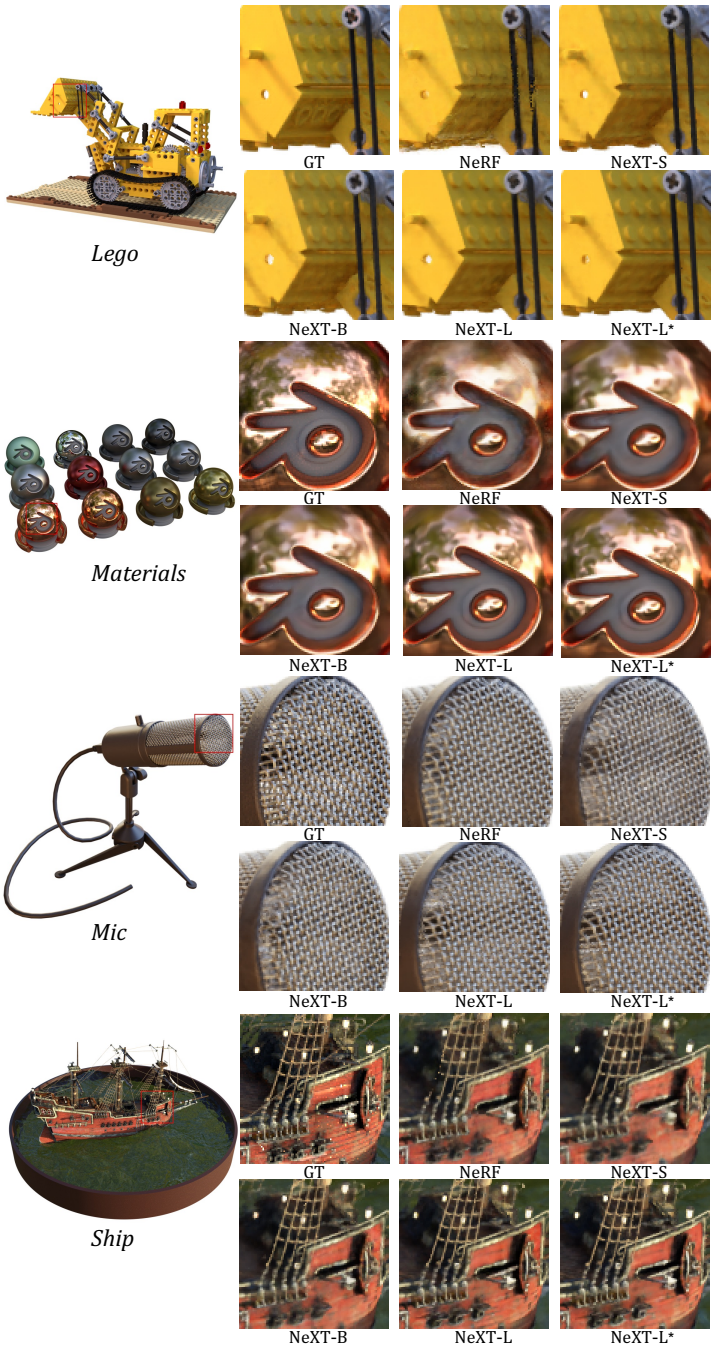

Fig. 2: Additional visualization of renderings from NeXT variants versus groundtruth and NeRF. The other four scenes on Blender dataset are presented in the same format as Fig. 1.

References

1. Arandjelović, R., Zisserman, A.: Nerf in detail: Learning to sample for view synthesis. arXiv preprint arXiv:2106.05264 (2021)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Deng, B., Barron, J.T., Srinivasan, P.P.: Jaxnerf: an efficient jax implementation of nerf (2020)
4. Liu, L., Gu, J., Zaw Lin, K., Chua, T.S., Theobalt, C.: Neural sparse voxel fields. Advances in Neural Information Processing Systems **33**, 15651–15663 (2020)
5. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)
6. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
7. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
9. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2437–2446 (2019)
10. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019)
11. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems **34** (2021)
12. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)