Supplementary Material for Enhancing NeRF akin to Enhancing LLMs: Generalizable NeRF Transformer with Mixture-of-View-Experts

Wenyan Cong[1*], Hanxue Liang[2,1*], Peihao Wang[1], Zhiwen Fan[1], Tianlong Chen[1],
Mukund Varma[3], Yi Wang[1], Zhangyang Wang[1]
[1]University of Texas at Austin, [2]University of Cambridge, [3]Indian Institute of Technology
{wycong, peihaowang, zhiwenfan, tianlong.chen, panzer.wy, atlaswang}@utexas.edu,
hl589@cam.ac.uk, mukundvarmat@gmail.com

1. More Training / Inference Details

The base learning rates for the feature extraction network and GNT-MOVE are 10^{-3} and 5×10^{-4}, respectively, which decay exponentially over training steps. For the zero-shot generalization experiments, we train the network for 330,000 steps with 4096 rays sampled from 4 different views in each iteration. In the few-shot setting, we further fine-tune the pretrained model on each scene for 2,4000 steps. During the inference, we sample 192 coarse points per ray in all experiments.

2. Cross-Scene Generalization

2.1. Testing Datasets

Local Light Field Fusion (LLFF) [2] consists of 8 forward-facing captures of real-world scenes using a smartphone. NeRF Synthetic dataset [3] consists of 8, 360° scenes of objects with complicated geometry and realistic material. Each scene consists of images rendered from viewpoints randomly sampled on a hemisphere around the object. Shiny-6 dataset [6] contains 8 forward-facing scenes with challenging view-dependent optical effects, such as the rainbow reflections on a CD, and the refraction through liquid bottles. Tanks-and-Temples [5] is a complex outdoor dataset and contains large unbounded scenes. Following NeRF++, we evaluate on four scenes, including M60, Train, Truck, and Playground, and use the same evaluate views as NeRF++ does. NMR [1] contains 360° views of various objects from unseen categories, which could be downloaded from NMR_Dataset.zip[1] (hosted by the authors of Differentiable Volumetric Rendering [4]). In the main paper, we report the average metrics across all eight scenes on each dataset for cross-scene generalization experiments.

2.2. Per-Scene Breakdown Results for Zero-Shot Generalization

To better demonstrate the effectiveness of our customized MoE, in Table 1 and Table 2, we pick few representative scenes for breakdown analysis of both GNT's and GNT-MOVE's quantitative results presented in Table 1a in the main paper. The scenes we choose mainly cover the complex geometries (e.g., leaves and orchids) and materials (e.g., room and materials). In both tables, our GNT-MOVE outperforms GNT by a significant margin in most scenes and achieves comparable results in the rest ones, demonstrating that with necessary customizations, MoE could be a strong tool to push the frontier of generalizable NeRF.

It is also worth mentioning that in Table 2, our GNT-MOVE has demonstrated superior performance, especially in scenes with complex materials (*e.g.*, Drums, Materials, Ship), showing that the customized MoE further enables cross-scene NeRF to generalize to difficult scenarios.

2.3. More Expert Selection Analyses

In Figure 1, we visualize more unseen scene rendering results and also the corresponding expert selections in the format of expert maps. It can be observed that our customized MoE is not only capable of keeping consistent selection across scenes (*e.g.*, white background in the left three scenes, leaves in the right two scenes), but also reacts properly to complex lighting effects and materials (*e.g.*, sparkling water in the left bottom scene Ship).

3. More Comparison: GNT v.s. GNT-MOVE

While the model size/speed is indeed not the main focus in this paper, GNT-MoE does generalize better, than the non-MoE counterpart with even heavier parameterization, while keeping per-instance inference low-cost.

Below, ▷ **1)** and ▷ **2)** demonstrate that the solid gain of MoE for generalizable NeRF goes way beyond naively

[*] Equal contribution.

[1] https://s3.eu-central-1.amazonaws.com/avg-projects/differentiable_volumetric_rendering/data/NMR_Dataset.zip

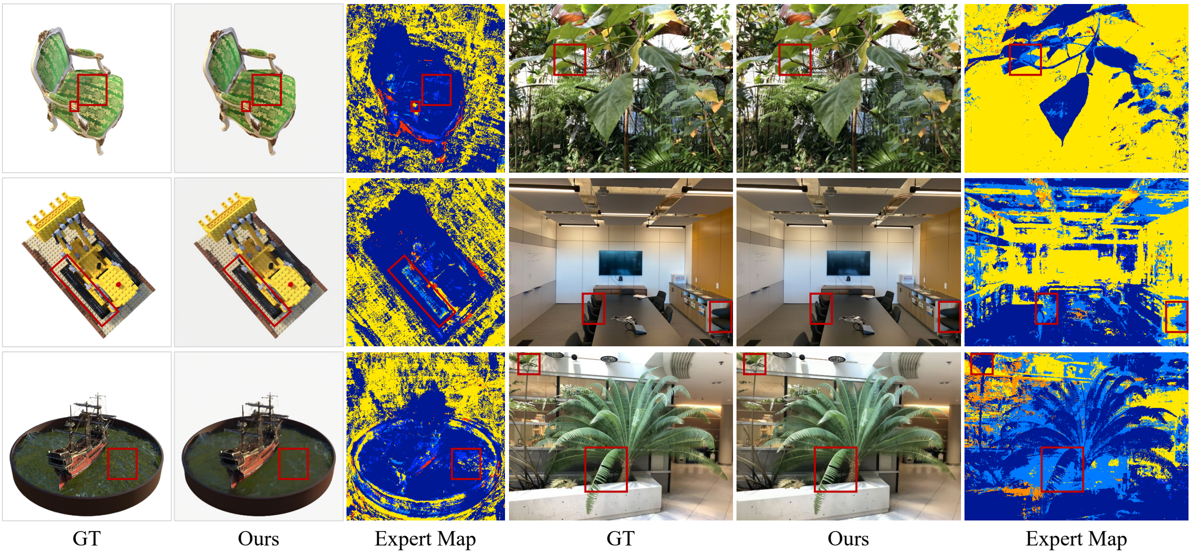

Figure 1: Results of unseen scene rendering and visualization of expert selection using different colors.

Models	Room	Leaves	Orchids	Flower	T-Rex	Horns
GNT	29.63	19.98	18.84	25.86	24.56	26.34
Ours	**29.94**	**20.45**	**19.38**	**27.04**	**24.58**	**26.87**

(a) PSNR↑

Models	Room	Leaves	Orchids	Flower	T-Rex	Horns
GNT	0.940	0.756	0.661	0.859	0.885	0.892
Ours	**0.946**	**0.770**	**0.668**	**0.871**	**0.878**	**0.894**

(b) SSIM↑

Models	Room	Leaves	Orchids	Flower	T-Rex	Horns
GNT	0.091	0.183	0.216	0.108	0.127	0.118
Ours	**0.087**	**0.173**	**0.209**	**0.101**	**0.123**	**0.113**

(c) LPIPS↓

Models	Room	Leaves	Orchids	Flower	T-Rex	Horns
GNT	0.031	0.097	0.119	0.048	0.054	0.046
Ours	**0.029**	**0.093**	**0.115**	**0.043**	**0.054**	**0.044**

(d) Avg↓

Table 1: Comparison between our GNT-MOVE and GNT for cross-scene generalization under zero-shot setting on the LLFF Dataset (scene-wise).

Models	Chair	Drums	Materials	Mic	Ship
GNT	29.17	22.83	23.80	29.61	25.99
Ours	**29.64**	**23.19**	**24.16**	**30.30**	**26.48**

(a) PSNR↑

Models	Chair	Drums	Materials	Mic	Ship
GNT	0.959	0.927	0.931	0.977	0.836
Ours	**0.962**	**0.979**	**0.935**	**0.982**	**0.845**

(b) SSIM↑

Models	Chair	Drums	Materials	Mic	Ship
GNT	0.038	0.059	0.058	0.017	0.154
Ours	**0.038**	**0.057**	**0.056**	**0.015**	**0.149**

(c) LPIPS↓

Models	Chair	Drums	Materials	Mic	Ship
GNT	0.021	0.044	0.040	0.014	0.054
Ours	**0.021**	**0.042**	**0.040**	**0.013**	**0.051**

(d) Avg↓

Table 2: Comparison between our GNT-MOVE and GNT for cross-scene generalization under zero-shot setting on the NeRF Synthetic Dataset (scene-wise).

larger model size; and ▷ **3)** demonstrates that the gain can only be unleashed with PE and SR. Detailed results could be found in Table 3. As preliminary, every expert in GNT-MOVE is half the size of GNT's same layer. The default GNT-MOVE (row 4) selects E = 2 such experts from K = 4 candidates, plus 1 permanent expert. Hence, if we treat the total parameter and inference FLOPs of GNT both as unit (**"1"**), then the default GNT-MOVE has **"2.5"** total parameter and **"1.5"** inference FLOPs. We construct the following comparison groups:

▷ **1) the same FLOPs at inference.** Rows 1-2 compare GNT (FLOPs **"1"**) versus GNT-MOVE using only one selectable expert (E = 1) + one PE (0.5 + 0.5 = **"1"**). Despite the same inference complexity, the extra flexibility to "select" endows GNT-MOVE with superior performance.

▷ **2) the same total parameter.** Row 3 widens GNT by 2.5 times to match the total parameter size **"2.5"** of GNT-MOVE, called "GNT (Large)". Compared to Row 4 (default GNT-MOVE), they have the same total parameter counts; meanwhile, GNT-MOVE has smaller per-inference FLOPs. However, GNT (Large) performs worse - and that

Models	Local Light Field Fusion (LLFF)				NeRF Synthetic				Tanks-and-Temples (Truck)		
	PSNR↑	SSIM↑	LPIPS↓	Avg↓	PSNR↑	SSIM↑	LPIPS↓	Avg↓	PSNR↑	SSIM↑	LPIPS↓
GNT	25.86	0.867	0.116	0.047	27.29	0.937	0.056	0.030	17.39	0.561	0.429
GNT-MOVE (E=1, K=4, w. PE)	25.94	0.868	0.111	0.043	27.43	0.939	0.057	0.029	19.08	0.611	0.393
GNT (Large)	25.89	0.867	0.113	0.046	27.37	0.936	0.058	0.033	18.26	0.579	0.405
GNT-MOVE (E=2, K=4, w. PE)	**26.02**	**0.869**	**0.108**	**0.043**	**27.47**	**0.940**	**0.056**	**0.029**	**19.71**	**0.628**	**0.379**
GNT-MOVE **w.o. PE** (E=3, K=5)	25.81	0.867	0.114	0.047	27.32	0.933	0.059	0.031	18.11	0.570	0.414

Table 3: Comparisons to illustrate the solid gain of MoE and PE in GNT-MOVE.

clearly indicates for generalizable NeRF, “the more parameter the better” is NOT the right quote, and per-scene specialization is necessary.

▷ **3) Does PE undermine MoE claim? NO.** First, the above two points already justified the necessity of MoE and disapprove “natural to have better performance with more parameters”. Second, our core claim is NEVER “MoE shall work out of box for NeRF”. Instead, while MoE is promising to balance “generality” and “specialization”, making it work for generalizable NeRF demands customized tactics to inject the key priors of cross-view consistency & cross-scene commodity - PE is one such tactic.

To explain the second note, we stress that learning MoEs over NeRFs differs greatly from over standard image sets. If treating each view observation as an image sample, a “NeRF dataset” would exhibit significant clustering due to different views of the same scene, and even different scenes will bear natural scene similarity. The highly non-i.i.d distribution, with multi-dimensional similarity entangled across views and scenes, can make naive MoE training more prone to collapse - see our ablation in Supplement sec. 3. Our important contribution is to show one can reap the benefit of MoE with proper regularizations including PE.

To directly show PE values beyond just “more parameters”, we compare Row 5 in Table (replacing GNT-MOVE’s PE with a selectable expert, and making E = 3), which has same total parameter & inference FLOPs with our default GNT-MOVE setting (Row 4). Having PE helps evidently.

References

[1] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 1

[2] 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. 1

[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, 2020. 1

[4] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 1

[5] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 623–640. Springer, 2020. 1

[6] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021. 1