Supplementary Materials for NeRF as Pretraining at Scale: Generalizable 3D-Aware Semantic Representation Learning from View Prediction

Wenyan Cong[1], Hanxue Liang[2], Zhiwen Fan[1], Peihao Wang[1], Yifan Jiang[1],
Dejia Xu[1], A. Cengiz Oztireli[2], Zhangyang Wang[1]
[1]University of Texas at Austin [2]University of Cambridge
{wycong, zhiwenfan, peihaowang, yifanjiang97, dejia, atlaswang}@utexas.edu, {hl589, aco41}@cam.ac.uk

1. Implementation Details

During the pretraining stage, we implement hardness-aware training strategies. Initially, we select the 10 nearest source views based on the camera positions relative to the target novel view. After 50,000 steps, we reduce the number of source views by one every 20,000 steps, until only two remain. Concurrently, we alter our sampling approach for source views, excluding the closest k views, where k begins at 0 and increments by 1 every 20,000 steps. In this process, there is an 80% chance of uniformly sampling from the available source views, and a 20% chance of selecting the remaining nearest ones.

2. Ablation Studies

Given that Figure 3 in the main script already illustrates the importance of scaling up both pretraining data and model size, we further conduct ablation studies on the hardness-aware training strategies. We evaluate these strategies in the context of both the pretraining task and few-shot downstream tasks.

2.1. Pretraining Cross-Scene Generalizaiton

The implementation of hardness-aware training strategies introduces additional challenges in the generalizable novel view synthesis pretraining. As shown in Table 1a, when compared to the GNT model scaled up in terms of pretraining data and model size but without hardness-aware training, those models employing hardness-aware strategies experience a slight performance degradation, attributable to the increased difficulty of the training process.

2.2. Few-Shot Generalization

To validate the advantages of hardness-aware training strategies in enhancing few-shot downstream task finetuning, we conduct ablation studies on multi-view semantic segmentation and depth estimation, detailed in Table 1b and Table 1c, respectively. In this context, 'few-shot' indicates the use of only a limited number of scenes for finetuning. Two key observations emerge from these studies for both downstream tasks: 1) the implementation of hardness-aware training strategies leads to improved performance in these tasks, and 2) there is also a noticeable enhancement in the performance of novel view synthesis specific to the downstream dataset. These results confirm that incorporating hardness-aware training in the pretraining phase not only boosts generalizability but also facilitates emergent capabilities in various downstream tasks.

Models				Local Light Field Fusion (LLFF)			NeRF Synthetic		
	dist.	# src	non-reg.	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Scaled GNT				**26.28**	**0.876**	**0.101**	**28.59**	**0.956**	**0.049**
Ours	✓			26.24	0.875	0.102	28.46	0.951	0.050
Ours	✓	✓		26.16	0.871	0.106	28.53	0.952	0.049
Ours	✓		✓	26.19	0.973	0.104	28.39	0.949	0.052
Ours	✓	✓	✓	26.17	0.873	0.104	28.41	0.951	0.052

(a) Zero-shot novel view synthesis on LLFF and NeRF Synthetic datasets.

Models				ScanNet					
	dist.	# src	non-reg.	PSNR↑	SSIM↑	LIPIPS↓	mIoU↑	Total Acc↑	Avg Acc↑
Scaled GNT				28.97	0.853	0.135	87.00	95.96	91.44
Ours	✓			29.14	0.860	0.130	90.99	97.45	94.27
Ours	✓	✓		29.35	0.868	0.122	92.34	97.87	95.06
Ours	✓		✓	29.27	0.866	0.128	91.70	97.67	94.76
Ours	✓	✓	✓	**29.53**	**0.873**	**0.119**	**93.12**	**98.46**	**95.26**

(b) Few-shot multi-view semantic segmentation on ScanNet dataset.

Models				ScanNet				
	dist.	# src	non-reg.	PSNR↑	SSIM↑	LIPIPS↓	AbsRel↓	RMSE↓
Scaled GNT				22.94	0.747	0.318	0.132	0.283
Ours	✓			23.13	0.750	0.315	0.125	0.264
Ours	✓	✓		23.18	0.751	0.313	0.120	0.253
Ours	✓		✓	23.20	0.752	0.312	0.121	0.259
Ours	✓	✓	✓	**23.26**	**0.754**	**0.312**	**0.119**	**0.249**

(c) Few-shot multi-view depth estimation on ScanNet dataset.

Table 1. Ablation analyses of hardness-aware training strategies on both the pretraining task and few-shot downstream tasks. 'Scaled GNT' represents the GNT model with scaled-up pretraining data and model size but without hardness-aware training. 'dist.' denotes increasing the distance between the target view and source views in the pretraining stage. '# src' denotes decreasing the number of source views. 'non-reg.' denotes using a non-regular sampling strategy, which contrasts with the typical approach of always selecting the closest source views.