NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior (Supplementary Material)

Wenjing Bian Zirui Wang Kejie Li Jia-Wang Bian
Victor Adrian Prisacariu
Active Vision Lab, University of Oxford
{wenjing, ryan, kejie, jiawang, victor}@robots.ox.ac.uk

A. Implementation Details

The following sections include more details about the datasets we use, our training procedure and evaluation metrics.

A.1. Dataset

We select sequences containing dramatic camera motions from ScanNet [1] and Tanks and Tamples [3] for training and evaluation. Tab. 1 lists details about these sequences, where *Max rotation* denotes the maximum relative rotation angle between any two frames in a sequence. The sampled images are further split into training and test sets. Starting from the 5*th* image, we sample every 8*th* image in a sequence as a test image. However, this leads to a change in the sampling rate in the temporal domain among training images. We found that the rotation errors are often higher than average at these positions where the sampling rate changes. In order to study the effect of the sampling rate changes, for scene *Family* in Tanks and Temples [3], we sample every other image as test images, i.e. training on images with odd frame ids and testing on images with even frame ids.

	Scenes	Type	Seq. length	Frame rate	Max. rotation (deg)
ScanNet	0079_00	indoor	90	30	54.4
	0418_00	indoor	80	30	27.5
	0301_00	indoor	100	30	43.7
	0431_00	indoor	100	30	45.8
Tanks and Temples	Church	indoor	400	30	37.3
	Barn	outdoor	150	10	47.5
	Museum	indoor	100	10	76.2
	Family	outdoor	200	30	35.4
	Horse	outdoor	120	20	39.0
	Ballroom	indoor	150	20	30.3
	Francis	outdoor	150	10	47.5
	Ignatius	outdoor	120	20	26.0

Table 1. **Details of selected sequences.** We downsample several videos to a lower frame rate. FPS denote frame per second. *Max rotation* denotes the maximum relative rotation angle between any two frames in a sequence. We show our method can handle dramatic camera motion (large maximum rotation angle) whereas previous methods can only handle forward-facing scenes.

A.2. Training Details

During training, we sample 1024 pixels/rays for an image and we sample 128 points along each ray for our approaches and all baselines. For all approaches, we use the same pre-defined sampling range (i.e., near and far) and sample uniformly between this range. During refinement, the learning rate of NeRF model decays every 10 epochs with 0.9954, and the learning rate for the camera poses decays every 100 epochs with 0.9. Note that we only use RGB loss during refinement and depth distortion parameters are no longer optimised. As the scene scales can be arbitrary, the optimised scale parameter of the depth map during training is also arbitrary. To avoid scale collapsing (all scales reduced to 0.0) during training, we manually set the scale of the depth map for the last frame to 1.0. We also use the normalised point clouds when computing the inter-frame point cloud loss.

A.3. Test-time Optimisation

During the evaluation for novel view synthesis, following our baselines NeRFmm [8], BARF [4] and SC-NeRF [2], we run a test-time optimisation to align the camera poses of the test set by minimising the photometric error on the synthesised images, while keeping the trained NeRF model froze. Although all these baseline methods have their own way to align camera poses (discussed below), all of them fail to align camera poses in complex camera trajectories in ScanNet and Tanks and Temples.

To fairly evaluate all methods in challenging camera trajectories, we propose to align test camera poses by first initialising from learned poses of adjacent training images, followed by a test-time optimisation. We shorthand this alignment as **Neighbour + opt**. In practice, we find this initialisation is robust and provides the best alignment for all approaches. All results in our main paper are evaluated in this way.

	Sim(3) + no opt.			Identity + opt.			Sim(3) + opt.			(4) Neighbour + opt		
	PSNR↑	SSIM ↑	LPIPS ↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Ours	17.24	0.62	0.58	13.38	0.39	0.70	32.47	0.84	0.41	32.47	0.84	0.41
BARF	14.68	0.55	0.66	19.56	0.65	0.57	17.82	0.60	0.61	32.31	0.83	0.43
NeRFmm	11.28	0.40	0.80	30.59	0.81	0.49	12.46	0.43	0.80	30.59	0.81	0.49
SC-NeRF	10.68	0.38	0.80	22.39	0.71	0.55	11.25	0.40	0.80	31.33	0.82	0.46

Table 2. **Comparison of various pose alignment methods during test-time optimisation (ScanNet 0079_00).**

The following paragraphs outline previous alignment methods, and we show a comparison for all method with a ScanNet scene in Tab. 2.

Identity + opt. BARF [4] uses test-time optimisation to identify poses for the test frames, where all poses are initialised with identity matrices. This initialisation works well for simple forward-facing scenes, but not for complex trajectories. The optimisation is sensitive to the learning rate, and can easily fall into local minima when the target pose is far from the identity initialisation.

Sim(3) + opt. In NeRFmm [8], the poses are first initialised using Sim(3) alignment with an ATE toolbox [10]. Then, an additional test-time optimisation is used to further adjust the test poses. This initialisation works well when the learned poses can be aligned precisely to COLMAP poses (Ours in Tab. 2). However, incorrect pose estimations can affect the Sim(3) alignment.

Sim(3) + no opt. In SC-NeRF [2], the test poses are identified using a Sim(3) alignment between COLMAP poses and the learned poses. And no test-time optimisation is used. However, the results are biased toward COLMAP estimations, and misalignment can affect the view synthesis quality significantly.

A.4. Evaluation Metrics

Novel View Synthesis. We use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [7] and Learned Perceptual Image Patch Similarity (LPIPS) [9] to measure the novel view synthesis quality. For LPIPS, we use a VGG architecture [6].

Depth. The error metrics we use for depth evaluation include Abs Rel, Sq Rel, RMSE, RMSE log, δ_1, δ_2 and δ_3. The definitions are as follows:

- Abs Rel: $\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|d - d_{gt}\|/d_{gt}$;
- Sq Rel: $\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|d - d_{gt}\|_2^2/d_{gt}$;
- RMSE: $\sqrt{\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|d - d_{gt}\|_2^2}$;
- RMSE log: $\sqrt{\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|\log d - \log d_{gt}\|_2^2}$;
- δ_i: % of y s.t. $max(\frac{d}{d_{gt}}, \frac{d_{gt}}{d}) = \delta < i$;

where d is the estimated depth, d_{gt} is the ground truth depth, and $\mathcal{V}$ is the collection of all valid pixels on a depth map.

scenes	Ours			NeRFmm		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR	SSIM	LPIPS
Fern	**23.01**	**0.71**	**0.38**	20.58	0.59	0.50
Flower	**29.39**	**0.86**	**0.19**	27.02	0.76	0.32
Fortress	**29.38**	**0.80**	**0.28**	24.94	0.57	0.57
Horns	**25.24**	**0.73**	**0.37**	23.67	0.66	0.48
Leaves	**19.85**	**0.60**	**0.40**	19.46	0.55	0.46
Orchids	**19.51**	**0.56**	**0.43**	16.77	0.40	0.55
Room	**28.54**	**0.89**	**0.28**	26.14	0.84	0.39
Trex	**25.82**	**0.84**	**0.29**	24.13	0.77	0.39
mean	**25.09**	**0.75**	**0.33**	22.84	0.64	0.46

Table 3. **Novel view synthesis results on LLFF-NeRF dataset.**

scenes	Ours			NeRFmm		
	RPE_t ↓	RPE_r ↓	ATE↓	RPE_t	RPE_r	ATE
Fern	**0.252**	**0.993**	**0.003**	0.706	1.816	0.007
Flower	**0.035**	**0.096**	**0.001**	0.086	0.418	**0.001**
Fortress	**0.081**	**0.296**	**0.001**	0.233	0.739	0.004
Horns	**0.217**	**0.452**	**0.004**	0.321	0.850	0.008
Leaves	0.218	0.143	0.002	**0.138**	**0.051**	**0.001**
Orchids	**0.203**	**0.383**	**0.003**	0.686	2.030	0.010
Room	**0.244**	**0.936**	**0.004**	0.670	1.664	0.011
Trex	**0.219**	**0.319**	**0.004**	0.542	0.775	0.009
mean	**0.184**	**0.452**	**0.003**	0.423	1.043	0.006

Table 4. **Pose accuracy on LLFF-NeRF dataset.**

B. Additional Results

LLFF-NeRF Dataset. We compare our approach against NeRFmm on the LLFF-NeRF dataset [5] in terms of novel view synthesis quality (Tab. 3) and pose accuracy (Tab. 4). We show better performances than NeRFmm in both pose accuracy and synthesis quality. We use the normalized device coordinate (NDC) for both approaches.

Depth Estimation. We show detailed depth evaluation results for ScanNet scenes in Tabs. 5 to 8. Our depth estimation accuracy outperforms other baselines by a large margin.

0079_00	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Ours	**0.099**	**0.047**	**0.335**	**0.128**	**0.904**	**0.995**	**1.000**
BARF	0.208	0.165	0.588	0.263	0.639	0.896	0.983
NeRFmm	0.494	1.049	1.419	0.534	0.378	0.567	0.765
SC-NeRF	0.360	0.450	0.902	0.396	0.407	0.730	0.908
DPT	0.149	0.095	0.456	0.173	0.818	0.978	0.999

Table 5. **Depth map evaluation on ScanNet 0079_00**.

Pose Estimation. We visualise additional results for pose estimation on Tanks and Temples (Fig. 3) and ScanNet

0418_00	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Ours	**0.152**	**0.137**	**0.645**	**0.185**	**0.738**	**0.988**	**0.997**
BARF	0.718	1.715	1.563	0.630	0.205	0.569	0.769
NeRFmm	0.907	3.650	2.176	0.769	0.240	0.456	0.621
SC-NeRF	0.319	0.441	0.898	0.377	0.456	0.792	0.930
DPT	0.190	0.187	0.745	0.211	0.719	0.965	**0.997**

Table 6. **Depth map evaluation on ScanNet 0418_00**.

0301_00	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Ours	0.185	0.252	0.711	**0.233**	**0.792**	**0.918**	**0.958**
BARF	**0.179**	**0.146**	**0.502**	0.268	0.736	0.883	0.938
NeRFmm	0.444	0.830	1.239	0.481	0.397	0.680	0.845
SC-NeRF	0.383	0.378	0.810	0.452	0.360	0.663	0.846
DPT	0.317	0.568	1.133	0.350	0.597	0.821	0.914

Table 7. **Depth map evaluation on ScanNet 0301_00**.

0431_00	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Ours	**0.127**	**0.111**	**0.579**	**0.160**	**0.877**	**0.978**	**0.994**
BARF	0.398	0.710	1.307	0.444	0.381	0.655	0.847
NeRFmm	0.514	1.354	1.855	0.562	0.250	0.539	0.742
SC-NeRF	0.608	1.300	1.706	0.677	0.225	0.446	0.645
DPT	0.132	0.135	0.670	0.171	0.855	0.973	0.991

Table 8. **Depth map evaluation on ScanNet 0431_00**.

(Fig. 4).

More Visualisations. We present additional qualitative results for novel view synthesis and depth estimation on Tanks and Temples (Fig. 1) and ScanNet (Fig. 2).

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1

[2] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *ICCV*, 2021. 1, 2

[3] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 2017. 1

[4] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 1, 2

[5] 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)*, 2019. 2

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 2

[8] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 1, 2

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

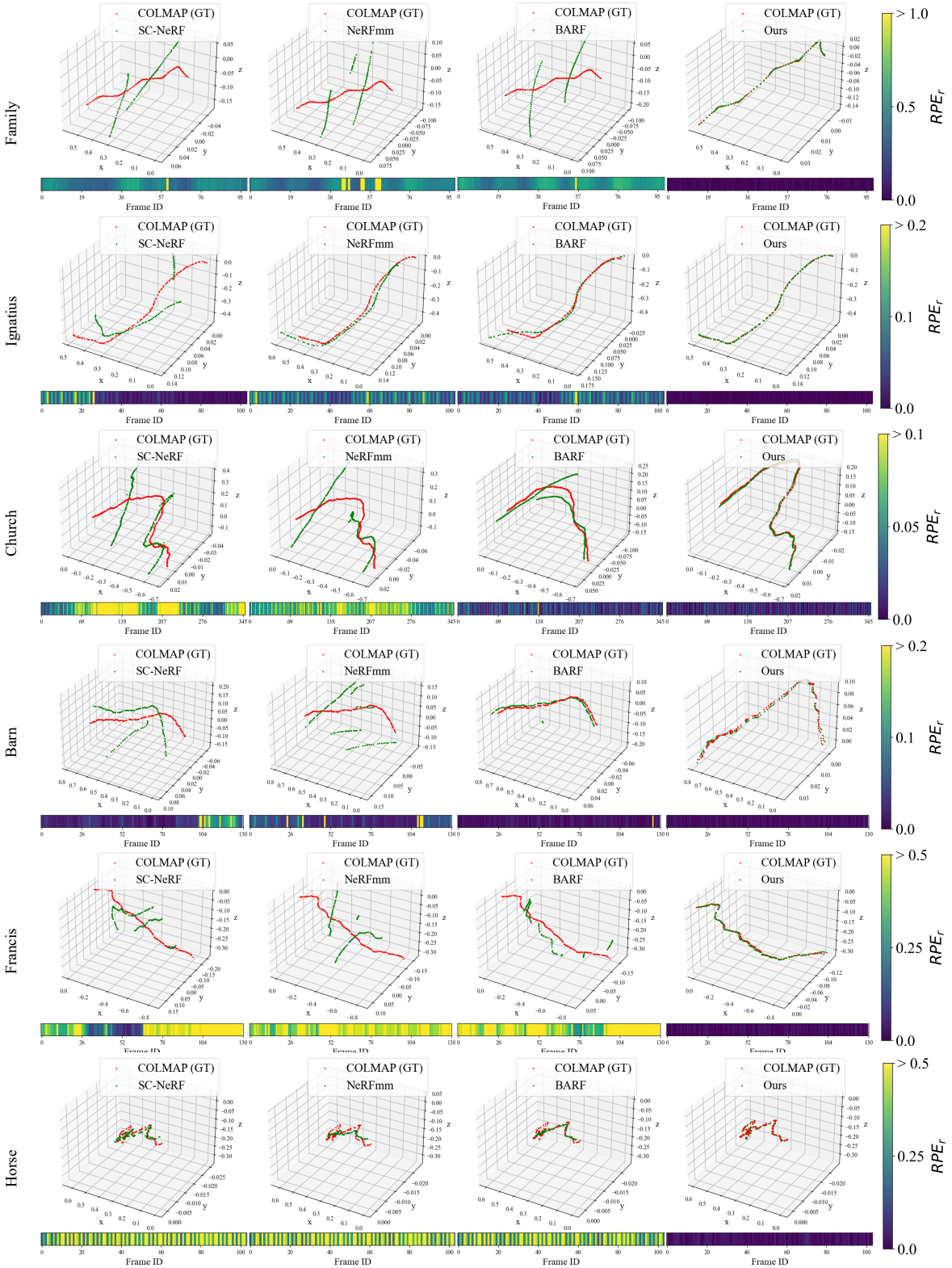
[10] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *IROS*. IEEE, 2018. 2

(a) BARF (b) NeRFmm (c) SC-NeRF (d) Ours (e) Ground Truth

Figure 1. **Qualitative results of novel view synthesis and depth prediction on Tanks and Temples.** We visualise the synthesised images and the rendered depth maps (top left of each image) for all methods. NoPe-NeRF is able to recover details for both colour and geometry.

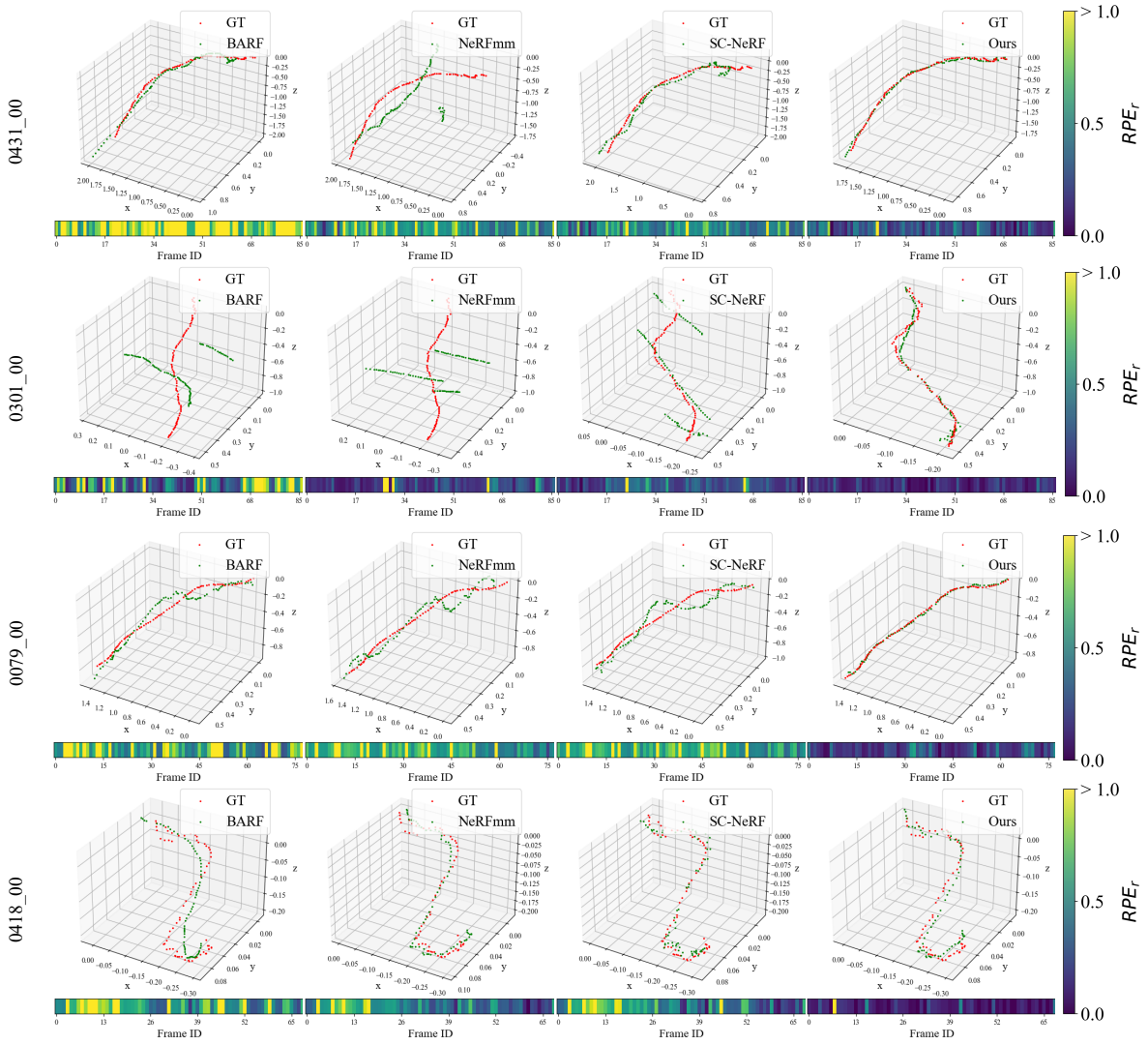
(a) BARF (b) NeRFmm (c) SC-NeRF (d) Ours (e) Ground Truth

Figure 2. **Qualitative results of novel view synthesis and depth prediction on ScanNet.** We visualise the synthesised images and the rendered depth maps (top left of each image) for all methods. NoPe-NeRF is able to recover details for both colour and geometry.



(a) BARF (b) NeRFmm (c) SC-NeRF (d) Ours

Figure 3. **Pose Estimation Comparison on Tanks and Temples.** We visualise the trajectory (3D plot) and relative rotation errors RPE_r (bottom colour bar) of each method on *Ballroom* and *Museum*. The colour bar on the right shows the relative scaling of colour.



(a) BARF (b) NeRFmm (c) SC-NeRF (d) Ours

Figure 4. **Pose Estimation Comparison on ScanNet.** We visualise the trajectory (3D plot) and relative rotation errors RPE_r (bottom colour bar) of each method on *Ballroom* and *Museum*. The colour bar on the right shows the relative scaling of colour.