Supplementary Material : Direct Alignment for Robust NeRF Learning

Ravi Garg[1,2], Shin Fang Chng[1,2], and Simon Lucey[1,2]

[1] Adelaide University
[2] Australian Institute of Machine Learning
ravi.garg@adelaide.edu.au

In this supplementary material, we provide additional details about our approach, with clarifications on experimental setting and ablations followed by further results with comparison of our method against state of the art beyond BARF [5].

1 Experimental Details

We use the code provided by the authors [5] to create baseline results and adapt the training and test methodology specified in [5]. For completeness, we include some important hyper-parameters used and the evaluation protocol here for different experiments, see [5] for full details.

1.1 LLFF experiments

Hyper-parameters: We used the learning rate for radiance filed estimation network to decay from 1×10^{-3} to 1×10^{-4}. Neural-net free pose estimation uses the learning rates starting from 3×10^{-3} and gradually decays to 1×10^{-5}. For the frequencies used for positional encoding, we use $L = 10$ for 3D points and $L = 4$ for the viewing direction. The coarse-to-fine scheduler for BARF is linearly adjusted from iteration $20K$ to $100K$.
Training and Evaluation Protocol We kept the training-test split for the full LLFF experiments same as [5] whereas the five frame sequences results are reported only on the first test-image used in the full LLFF experiments. We found evaluating view synthesis measures on the test images far away from training is less meaningful and have omitted these results form the main manuscript. We computed error metrics on all test images here these results are shown in table 2. As specified in the paper, we omit the test-time pose optimization used in [5] from our experiments. For completeness we include the evaluation for joint pose and structure estimation trained with full LLFF data with default runtime optimization in table 1. It can be seen that for quite a few sequences the view synthesis with test time pose optimization can hide structural errors.

2 DTU experiments

Hyper parameters: As BARF [5] has not been tested on DTU datasets, we adapt the LLFF settings with minimal required changes. We found that both

Sequence	**BARF**		**+Alignment**		**BARF**			**+Alignment**		
	E Rot	E Trans	E Rot	E Trans	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Fern	0.197	1.82	**0.172**	**1.78**	23.7	0.71	**0.31**	**23.81**	0.71	0.33
Leaves	1.311	2.65	**1.019**	**2.29**	18.87	0.55	0.35	**18.90**	0.55	**0.34**
Orchid	0.580	3.94	**0.508**	**3.42**	**19.54**	**0.59**	**0.28**	19.33	0.56	0.29
T-rex	1.198	7.51	**0.185**	**2.373**	**23.11**	**0.77**	**0.21**	22.82	0.76	0.23
Flower	0.212	2.26	**0.202**	**1.88**	23.88	0.71	0.20	**24.95**	**0.73**	**0.19**
Fortress	0.372	3.14	**0.253**	**1.97**	**29.08**	**0.83**	**0.12**	28.82	0.82	0.13
Horns	2.950	13.97	**0.120**	**1.41**	20.72	0.69	0.30	**22.57**	**0.71**	**0.34**
Room	0.375	2.93	**0.071**	**1.07**	31.91	0.94	0.10	31.06	0.92	0.13
Mean	0.899	4.777	**0.3162**	**2.0241**	23.851	**0.723**	**0.234**	**24.03**	0.72	0.2475

Table 1: View Synthesis evaluation comparison for joint radiance field and pose optimizations with test time pose optimization by peaking into target image. Should be contrasted with table in main manuscript.

Seq.	**BARF**		**+Alignment**		**BARF**			**+Alignment**		
	E Rot	E Trans	E Rot	E Trans	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Fern	4.10	8.70	**3.13**	**6.65**	11.83	**0.33**	0.68	**12.45**	**0.32**	**0.63**
Leaves	2.00	4.30	**1.13**	**2.41**	11.25	**0.12**	0.51	**11.40**	**0.11**	**0.47**
Orchid	0.74	5.58	**0.38**	**2.15**	12.31	0.17	0.46	**13.36**	**0.22**	**0.43**
T-rex	8.22	40.82	**0.34**	**1.71**	10.28	0.28	0.71	**15.48**	**0.47**	**0.39**
Flower	**0.55**	2.40	0.60	**1.14**	**17.03**	**0.38**	**0.32**	**16.12**	**0.30**	**0.34**
Fortress	11.84	63.84	**2.19**	**13.31**	11.10	0.29	0.67	**14.71**	**0.35**	**0.46**
Horns	3.39	28.26	**1.71**	**13.96**	11.96	0.29	**0.63**	**13.14**	**0.31**	**0.57**
Room	0.65	4.49	**0.15**	**1.17**	16.27	0.63	0.40	**18.94**	**0.72**	**0.39**
Mean	3.99	19.93	**1.20**	**5.31**	12.75	0.31	0.55	**14.45**	**0.35**	**0.46**

Table 2: Results on LLLF sequence with first five images used for training. Note that this experiment do not correspond to the once reported in DS-NeRF [4] and other literature where training views are sampled uniformly to create with large baselines. Should be compared with the corresponding table in main manuscript.

BARF and proposed method required lowering the learning rate for pose estimation. We set the initial learning rate for pose to be 10^{-4} and gradually decrease this to minimum value of 10^{-5}. Instead of using the far depth to be infinity as in the case of LLFF, for DTU we use the provided range by the dataset. We run all methods for a fix 100K iterations. We gradually introduce the higher frequency in positional encoding from iteration 10K onwards and have full positional encoding enabled at 60K iterations. These hyper parameters are not tuned to suit any method but are coarsely selected to minimize pose estimation divergence for all the methods.

Training and evaluation protocol We use the training and test spit used in SPARF [8] to evaluate variant of proposed methods. We use the background masks at the training and the test-time to (i) replace the background color to be white and (ii) ignore pixels on the background for depth evaluation. For the evaluation of depth, given that the optimized scene is subject to a 3D similarity, we align the scale of the predicted depth with the scale determined from the alignment procedure. We compute the mean absolute difference between the scaled predicted depth and ground-truth depth. We consider only those areas where valid ground-truth depth data is available in our evaluation.

Sampling surrogate views: Selecting pixels in a random frame and matching it with every other frame in the training set worked reasonably well in LLFF dataset. However due to large baselines in DTU, we observed that postponing to match far apart image to a later stage helps. To keep the number of loss terms for rendering and matching same, we following the following procedure to sample surrogate views to match. For any image, we define a random image withing permissible distance to be the surrogate view. Every sampled pixel for which we are minimizing rendering loss, we have a candidate frame to match it in. We minimize the relevant alignment loss for each of these sample-pixel and surrogate view pair. We start with matching images with only neighbouring views and gradually increasing the matching window to a maximum of 4 frames on the either side of the reference frame.

3 Comparison with other state of art approaches on LLFF dataset

BARF was selected as the natural baseline for our work because it is one of the few joint pose and radiance field optimization method that does not use outside packages trained on large datasets such as depth or correspondences estimators and is well adapted by current mainstream NeRF works.

Broadly, two major research directions have emerged post BARF: First, are regularization techniques [1,4,6,8] often using externally leaned priors in the form of depth/correspondence. This methods improve over BARF significantly when only few views are available to learn. Most of these methods show diminishing utility as training views increase with or without pose optimization. Second, are approaches that inherently change the pose representation or optimization routine - e.g. GARF [3], L2G [2], CamP [7] to name a few. These methods often show consistent improvements in the joint pose and NeRF estimation with dense views and struggle with large baseline captures without initialization. Note however that the contribution in these papers are orthogonal to our work.

For completeness, we include results of BARF [5], GARF [3], SPARF [8], L2G [2] and nope-nerf [1] on LLFF datasets in Tab 3. These results use same test splits evaluation protocol as described in the main paper.

Our approach outperforms all of the aforementioned baselines despite some of the baselines using off the self pre-trained models to provide depth / correspondences in terms of estimating camera poses. Many of the orthogonal improvements on BARF can be used with the direct alignment loss to futher improve camera pose estimation.

We report view synthesis results with and without test time pose optimization for the approaches whose code is publicly available. BARF, L2G and nope-nerf all show inferior view synthesis results on LLFF data to proposed approach without test time pose optimization. Many of the baseline see artificial boost in view synthesis accuracies when they are allowed to peak into test images. These results highlight how large structure from motion errors are masked by test time optimization. We notice large errors of our nope-nerf runs. To ensure correctness, we compare our relative pose estimation and test-time optimised view synthesis performance with reported result in Tab 4. While minor pose estimation differences could be due to train-test split change, large view synthesis performance mismatch happens due us using "sim(3)+opt" instead of "neighbour+opt" scheme, later being unemployable on our test split. We hope that these results will challenge the notion that more recent approach is better and increase appreciation of solid ablations.

Table 3: Comparison on LLFF. {*} and {#} denote results reported by authors and reproduced by us using public code. All pose errors are absolute pose errors and TTO represents test time pose optimization

	Pose errors		**no TTO**			**with TTO**		
Method	Trans	Rot	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Ours	**0.20**	**0.32**	**20.16**	**0.53**	**0.27**	24.03	0.72	0.247
BARF	0.48	0.90	17.04	0.42	0.32	23.85	0.72	0.23
GARF*	0.29	0.33	-	-	-	24.54	0.74	0.22
L2G#	0.30	0.48	19.31	0.52	0.25	24.35	0.75	0.21
SPARF*	0.23	0.77	-	-	-	25.18	0.78	0.20
Nope-nerf#	0.27	2.432	13.09	0.27	0.54	23.64	0.76	0.26

References

1. Bian, W., Wang, Z., Li, K., Bian, J.W., Prisacariu, V.A.: Nope-nerf: Optimising neural radiance field with no pose prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4160–4169 (2023)
2. Chen, Y., Chen, X., Wang, X., Zhang, Q., Guo, Y., Shan, Y., Wang, F.: Local-to-global registration for bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8264–8273 (2023)
3. Chng, S.F., Ramasinghe, S., Sherrah, J., Lucey, S.: Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In: European Conference on Computer Vision. pp. 264–280. Springer (2022)
4. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
5. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021)

Table 4: Relative pose and view synthesis of our nope-nerf run and reported results by the authors. Test pose are initialized with sim3 alignment with COLMAP and optimized during test time to maximize synthesis performance.

	Pose errors		Novel view synthesis		
	Rotation (°)	Translation (×100)	PSNR	SSIM	LPIPS
Reported	0.452	0.172	25.9	0.75	0.33
Reproduced	0.488	0.187	23.64	0.76	0.26

6. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
7. Park, K., Henzler, P., Mildenhall, B., Barron, J.T., Martin-Brualla, R.: Camp: Camera preconditioning for neural radiance fields. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
8. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4190–4200 (2023)