NeRFMatch | Supplementary

In this supplementary document, we provide further details regarding our proposed method and qualitative results. We describe our implementation details for NeRF in Appendix A, and for NeRFMatch in Appendix B. Then, we present additional analysis and discussion of our method in Appendix C.

Fig. 1: Example of masking on Kings College scene. Top images - original images, bottom - semantic segmentation using [4].

A NeRF Implementation Details

Handling challenges in outdoor scenes. Outdoor reconstruction in the wild has a lot of challenges including illumination changes, transient objects, and distant regions. For the task of localization, we are interested in reconstructing only the static scene elements, *e.g.*, roads, buildings, and signs.

To properly train NeRF in such a scenario, we use a pre-trained semantic segmentation model [4] and mask out the sky and transient objects: pedestrians, bicycles, and vehicles. These objects occupy only a minor part of the captured images and are excluded from the loss computation during the training process. Analogous methods for masking in sky regions and/or dynamic object areas have been implemented in other works focused on the reconstruction of urban scenes [10, 14, 16]. We show examples of semantic segmentation in Fig. 1 and its effect on synthesized views in Fig. 2.

Fig. 2: Example of masking on the King's College scene of Cambridge Landmarks [5]. The bottom row are rendered with NeRF, and the top row - ground truth images.

Table 1: NeRF PSNR scores. We present the PSNR scores for our trained MipNeRF models on each scene of Cambridge Landmarks [5] and 7-Scenes [13].

Cambridge Landmarks - Outdoor						7-Scenes - indoor							
Kings	Hospital	Shop	StMary	Court	Average	Chess	Fire	Heads	Office	Pump.	Kitchen	Stairs	Average
22.9	22.1	24.0	23.0	23.2	23.1	29.6	30.0	32.5	30.2	31.4	27.9	34.7	30.9

To account for illumination changes, we use an appearance vector that we concatenate together with the view direction as input, similar to [8]. The appearance vector changes across sequences but stays the same for all frames in one sequence since appearance does not drastically change inside a sequence.

NeRF architecture. Our NeRF model consists of a MipNeRF [2] architecture with both coarse and fine networks. We utilize the final outputs from the fine network to render RGB, depth maps, and 3D features.

NeRF training. For each scene, we load a subset of 900 training images and 8 validation images and train each model for 15 epochs. From the set of all pixels in all training samples, we randomly sample a batch of 9216 rays. Subsequently, for each ray, we sample 128 points for the coarse network and an additional 128 for the fine network. We use the Adam optimizer [6] with a learning rate 1.6×10^{-3} and cosine annealing schedule [7]. In Tab. 1, we present the per-scene PSNR scores for our trained models on the training images.

B NeRFMatch Implementation Details

We summarize average runtime performance for NeRF and both matching models in Tab. 2.

Training pairs. We use the same training pairs[1] generated by PixLoc [12] which were computed based on image covisibility within the training split. During training, for each train image we load its top-20 covisible pairs. For each

[1] Image pairs are available from `https://cvg-data.inf.ethz.ch/pixloc_CVPR2021/`

Table 2: Runtime. We show runtime of NeRFMatch-Mini and NeRFMatch. For pose refinement we are using optimization refinement for NeRFMatch-Mini and iterative refinement for NeRFMatch.

NeRF type	NeRFMatch-Mini	NeRFMatch
NeRF render	141ms	141ms
Image-to-NeRF matching	37ms	157ms
Pose refinement	398ms	141ms

training epoch, we then randomly sample 10000 training pairs from those co-visible pairs for each scene. In the case, we train multiple scenes, we merge those pairs across scenes which allows us to balance the training samples across different scenes.

Image retrieval. We adopt the retrieval pairs pre-computed by PixLoc [12] using NetVLAD [1] for Cambridge Landmarks [5] and DenseVLAD [15] for 7 Scenes [13] during inference. We use those retrieval pairs for all experiments by default except for the NeRF-only localization experiment in Sec. 5.1. That experiment is to confirm the feasibility of NeRF-only localization, therefore we run NetVLAD [1] to extract retrieval pairs at image resolution 480×480 between the real query images and the training images synthesized by NeRF.

During inference, we noticed applying top$-k$ retrieval pairs with $k > 1$ show evident improvement for NeRFMatch on Cambridge Landmarks. Thus, we set $k = 10$ following the common localization practice [11, 12]. For NeRFMatch-Mini, setting $k > 1$ did not change much the performance. We suspect this is due to its less accurate matches, which makes the outlier rejection harder when merging noisy correspondences from more pairs. For the indoor 7 Scenes dataset, we use $k = 1$ which is sufficient for relatively small-size scenes.

Optimization refinement. Similar to iNeRF [17], we are doing a forward pass through frozen NeRF MLP layers using an estimated pose as the initial camera pose. Instead of rendering the entire image, we sample and render 3600 rays, which are equally spread in a grid structure across the image plane. The we apply a regular photometric loss between the query image and the rendered image and backpropagate to update the initial camera pose. Instead of using the raw updated camera pose, we render the NeRF features and match them with the NeRFMatch to obtain the final camera pose.

C Additional Details

NeRF backbones. In this section, we evaluate additional NeRF type - Instant NGP [9] in comparison to MipNeRF [2]. We use MipNeRF for our experiments in the main paper . As shown in Tab. 3, Instant NGP performs significantly worse. We hypothesize that this is due to noisy depth reconstruction that is typical for Instant NGP.

Impact of scene sizes. Scene size affects both NeRF and localization performance, often coupled with scene content and camera pose distribution. Ranking

Table 3: NeRF backbone ablation on Cambridge Landmarks. We compare NeRFMatch-Mini and NeRFMatch performances using Instant NGP.

NeRF type	Avg. Med ($cm/°$) ↓/Recall (%) ↑	
	NeRFMatch-Mini	NeRFMatch
Instant NGP	41.1/0.7/44.4	28.1/0.5/61.3
MipNeRF	20.0/0.4/69.7	13.3/0.3/80.8

scenes by localization errors (lower is better) leads to OldHospital ($50 \times 40m^2$) > KingsCollege ($140 \times 40m^2$) > ShopFacade ($35 \times 25m^2$) for outdoor and stairs ($2.5 \times 2 \times 1.5m^3$) > pumpkin ($2.5 \times 2 \times 1m^3$) > redkitchen ($4 \times 4 \times 1.5m^3$) > chess ($3 \times 2 \times 1m^3$) for indoor. This suggests that smaller scenes (OldHospital, stairs) can be more challenging than larger scenes (KingsCollege, redkitchen) due to challenging contents like repetitive structures and texture-less regions.

Image retrieval on synthesized views. The goal of NeRF-only experiment is to verify the possibility to use NeRF as the only scene representation removing the need to maintain the original image collection. Our experiments show a slight performance decrease due to the domain gap between rendered and real images. Yet, we did not claim an efficient solution for online image retrieval and NeRF rendering. Future research is needed to improve its runtime efficiency either via caching scene reference poses in a hierarchical tree structure to fasten the searching process or leveraging any available prior information such as GPS coordinates to quickly find a subset of poses.

Indoor performance bottleneck. NeRF predicted depth maps are used to compute pseudo ground-truth for matching supervision. Incorrect depth predictions can lead to misaligned feature correspondences. In contrast, image matching, SCR, and APR methods use more accurate labels like Colmap camera poses or 3D maps. For small-scale indoor scenes, precise supervision is essential to achieve centimeter-level errors. Our method based on feature matching, however, scales better than regression-based approaches in larger outdoor scenes. Introducing uncertainty measures to ignore inaccurate matches, as in [3], and improved NeRF reconstructions with accurate depth maps will benefit our method.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
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. Chen, L., Chen, W., Wang, R., Pollefeys, M.: Leveraging neural radiance fields for uncertainty-aware visual localization. arXiv preprint arXiv:2310.06984 (2023)

4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
5. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, (ICLR) 2015 (2015)
7. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
8. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7210–7219 (2021)
9. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
10. Rematas, K., Liu, A., Srinivasan, P.P., Barron, J.T., Tagliasacchi, A., Funkhouser, T., Ferrari, V.: Urban radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12932–12942 (2022)
11. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12716–12725 (2019)
12. Sarlin, P.E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., et al.: Back to the feature: Learning robust camera localization from pixels to pose. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3247–3257 (2021)
13. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2930–2937 (2013)
14. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)
15. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1808–1817 (2015)
16. Xie, Z., Zhang, J., Li, W., Zhang, F., Zhang, L.: S-nerf: Neural radiance fields for street views. arXiv preprint arXiv:2303.00749 (2023)
17. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: inerf: Inverting neural radiance fields for pose estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1323–1330. IEEE (2021)