

Supplementary Materials: Exploring Matching Rates: From Key Point Selection to Camera Relocalization

Anonymous Authors

1 OVERVIEW

In this supplementary material, we furnish additional details of our network framework and augment the experimental data presented. Specifically, we commence by elucidating the intricacies of the network architecture employed in our study. Subsequently, we have supplemented our analysis with additional detailed experimental data and performance metrics on the 12scenes dataset to further elucidate our method. Finally, we illustrate the implicitly retained scene models inherent in our method through visualization techniques.

2 NETWORK ARCHITECTURE

2.1 Scene Coordinate Regression Head

In our research, we have employed a scene coordinate regression head that diverges from the one used in ACE [1]. This decision was informed by our observation that the parameterization of the scene coordinate regression head in ACE might struggle with mapping scene coordinates in more challenging scenarios, such as the stairs scene in the 7scenes dataset. To address this, we have increased the complexity of the regression head. In ACE, the impact of adding or subtracting MLP layers on performance was investigated, with experimental data suggesting that additional MLP layers do not enhance localization accuracy. Thus, we cautiously expanded the number of MLP layers using a residual connection approach, which allows the network to not only retain a substantial portion of the original feature information but also to conduct a more nuanced analysis of the subtle differences within the features.

We have visualized our network architecture in Figure 2. In deviation from ACE [1], we have incorporated four additional sets of residual blocks and introduced two extra MLP layers in the final scene coordinate regression segment. This augmentation has enhanced the head network’s capacity to process features, culminating in an overall improvement in the network’s performance.

2.2 Keypoint Classification Network

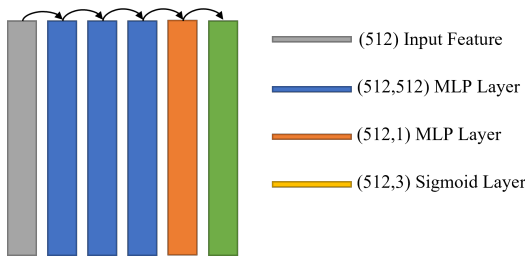


Figure 1: Keypoint Classification Network. We have developed a straightforward keypoint classification network that consists of only four MLP layers. At the end of the network, we employ a Sigmoid Layer to convert the estimations into probability values.

To augment the inlier ratio, we have devised a keypoint selection network, which essentially functions as a classification network tasked with discerning whether a point qualifies as an inlier. The input for this network is the feature set extracted by the feature backbone, identical to that used by the scene coordinate regression head. This implies that our approach does not significantly increase computational demands during testing. Moreover, since we classify features that have undergone rigorous training for scene coordinate prediction, this method is expected to substantially reduce training time.

Our network architecture is intentionally straightforward, composed of several MLP layers, with each blue MLP layer integrating a ReLU activation function. The final MLP layer in our design estimates the probability that scene coordinates correspond to keypoints, culminating with a sigmoid function that maps these estimations to probabilities. The network architecture is depicted in Figure 1.

During the training of our keypoint classification network, we adopt a scheme analogous to that employed for the scene coordinate regression head. We initiate by sampling features and their corresponding keypoint status. Sampling ceases upon reaching a count of 8 million. To decouple the sampled buffer, we employ a shuffling method, and then we proceed to rapidly train the network utilizing a large learning rate. Typically, our keypoint training network is capable of completing its training within a span of 5 minutes.

3 MORE TRAINING DETAILS

During initial training, we set the maximum loss for the scene coordinate regression head at 0.005 and the minimum loss at 0.0001. We set the initial learning rate for the training of the keypoint network to 0.0001. For the training of the keypoint selection network, the maximum loss was set at 0.0004 and the minimum at 0.0001. When fine-tuning the scene coordinate network using keypoints, we set our maximum loss at 0.0006 and the minimum loss at 0.00005. All of our training epochs were fixed at 32, a number determined to allow the scene coordinate estimation to reach a sufficiently good level of training. For all our training, we employed the AdamW [3] optimizer. During training, we uniformly applied the 1cycle learning rate policy as proposed by [5] for rapid convergence.

4 EXPERIMENTS ON 12SCENES

The 12Scenes [6] dataset, consisting of indoor scenes, served as another testing ground for the validation of our method to demonstrate its efficacy in more generalized settings. We trained our model based on poses obtained via Structure from Motion (SfM) [2], a more ubiquitous approach when depth-based SLAM systems are inapplicable. The results are presented in Table 1. In the 12Scenes dataset, since the localization accuracy at the precision of (5°, 5cm) has nearly reached 100%, we have also listed the accuracies at the

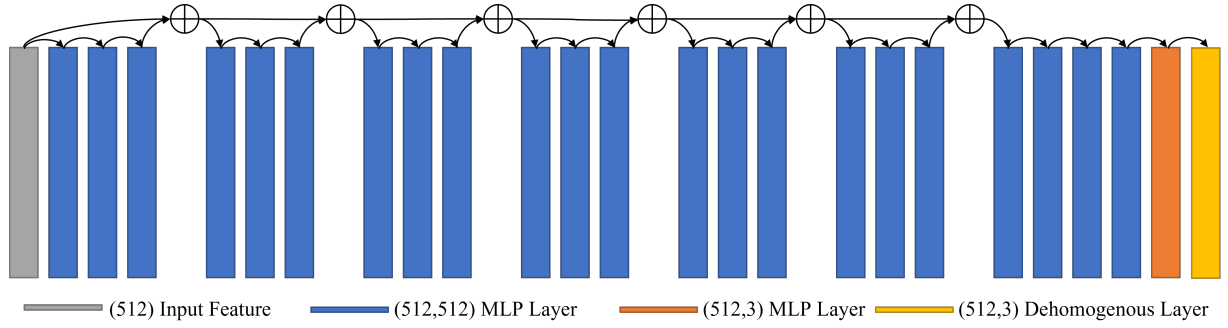


Figure 2: Network Architecture of Head. We have designed a deeper regression head for the estimation of scene coordinates, employing six residual blocks, each comprising three layers of Multi-Layer Perceptrons (MLPs). Five MLPs are dedicated to the final regression of the scene coordinates, and we have retained the Dehomogenization Layer from the ACE framework. Except for the last and the first layers, which may be adjusted according to specific requirements—for instance, when adopting a Homogeneous estimation approach, the output of the penultimate orange-red layer consists of four channels—the remaining layers uniformly feature 512 channels.

Table 1: Pose Estimation Results on 12scenes.

Scene	Within (5°,5cm)		Within (2°,2cm)		Within (1°,1cm)		Median Err.(°,cm)	
	ACE	Ours	ACE	Ours	ACE	Ours	ACE	Ours
apt1_kitchen	100.00%	100.00%	99.40%	100.00%	88.00%	92.70%	0.5,0.5	0.5,0.4
apt1_living	100.00%	100.00%	99.80%	100.00%	80.10%	89.50%	0.5,0.6	0.5,0.5
apt2_bed	100.00%	100.00%	97.50%	99.60%	80.30%	84.80%	0.5,0.5	0.5,0.5
apt2_kitchen	100.00%	100.00%	99.10%	99.10%	76.10%	84.80%	0.4,0.7	0.4,0.6
apt2_living	100.00%	100.00%	98.90%	99.40%	80.50%	88.30%	0.5,0.7	0.5,0.5
apt2_luke	100.00%	100.00%	97.80%	99.80%	65.50%	80.90%	0.5,0.8	0.5,0.6
office1_gates362	100.00%	100.00%	99.00%	100.00%	65.00%	74.40%	0.5,0.8	0.4,0.7
office1_gates381	99.90%	100.00%	96.70%	98.10%	64.10%	72.40%	0.6,0.8	0.6,0.7
office1_lounge	100.00%	100.00%	91.70%	97.60%	58.10%	67.30%	0.5,0.9	0.5,0.8
office1_manolis	100.00%	100.00%	93.90%	96.20%	70.60%	72.70%	0.5,0.8	0.5,0.7
office2_5a	99.40%	100.00%	92.00%	95.60%	51.90%	66.40%	0.5,1.0	0.5,0.8
office2_5b	99.50%	99.50%	91.60%	97.50%	61.50%	83.00%	0.4,0.9	0.5,0.6
Average	99.90%	99.96%	96.45%	98.58%	70.14%	79.77%	0.5,0.8	0.5,0.6

higher precisions of (2°, 2cm) and (1°, 1cm). The data clearly indicate that our method has significantly improved the accuracy at these more stringent thresholds of (2°, 2cm) and (1°, 1cm).

Table 2: Median Error on 7Scenes Dataset.

Scene	Init.	Init. With K.S.	K.F.T with K.S.
chess	0.5,0.6	0.5,0.6	0.6,0.5
fire	0.5,0.3	0.5,0.6	0.5,0.6
heads	-, -	0.5,0.5	0.5,0.5
office	0.4,1.5	0.5,1.2	0.5,1.1
pumpkin	0.5,1.2	0.5,1.1	0.5,0.9
redkitchen	0.5,0.4	0.5,1.0	0.5,0.8
stairs	-, -	0.7,1.4	0.9,1.2

5 MORE RESULTS

In Table 2, we present the median pose errors obtained at different stages of training on the 7Scenes [4] dataset. Our approach yielded only a slight improvement in the median error for scene coordinate estimation. In certain scenarios, there was a minimal reduction in

median error, but these reductions were almost imperceptible. Concurrently, the probability of improvement in our method was higher than the probability of experiencing a decrease in performance.

6 RECONSTRUCTION

Our task entails the input of an image and the estimation of pixel-wise scene coordinates, the aggregation of which from multiple viewpoints yields a model of the scene. In essence, our relocalization method involves an implicit exploration of the scene structure within the network model, thereby facilitating scene reconstruction. We have conducted a demonstration of scene reconstruction on the 7Scenes dataset. The reconstruction results have been exhibited in other studies as well [7, 8]. We primarily focus on comparing the reconstruction performance between the main comparative method ACE and our approach, followed by an analysis based on these findings. The reconstruction outcomes are illustrated in Figures 3. The left half of each figure displays the reconstruction results from ACE, while the right half showcases the results from our reconstruction method.



(a) Chess Reconstruction with ACE



(b) Chess Reconstruction with Ours

Figure 3: Scene Reconstruction of Chessc on 7Scenes dataset.

In the reconstruction process using the ACE method, we directly reconstructed using inliers and discarded scene coordinate points that were more than 5 meters from the scene's origin. For scene reconstruction based on our method, we filtered the scene coordinates used for reconstruction. Initially, we selected key points through the estimated keypoint probabilities (greater than 0.9). Subsequently, we rotated the images (20°) and extracted features using a feature network. Afterward, by estimating homography transformations, we obtained another set of corresponding points. Points with scene coordinate distances less than 0.1 meters were considered for scene reconstruction. We selected points that satisfied both criteria as the reconstruction points. Our method evidently filters out most outliers, making the reconstructed scene more congruent with the actual scene. As can be seen from Figures 3, our method, after filtering out outliers, results in many voids, indicating regions where scene coordinate estimation is unfavorable have been eliminated.

Additionally, we observed a phenomenon where depth estimation tends to be challenging around edge regions, often resulting in depth discontinuities or 'holes' at the edges. This suggests that although these locations may have been successful in minimizing reprojection error, they may have erroneously estimated scene coordinates, particularly incorrect depth values from the current viewpoint. This observation will be addressed in future research.

Furthermore, despite achieving better reconstruction results with our designed scheme, the discarding of a substantial number of points has led to a decrease in the overall accuracy of pose estimation. This issue will also be addressed in subsequent research.

REFERENCES

- [1] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. 2023. Accelerated Coordinate Encoding: Learning to Relocalize in Minutes Using RGB and Poses. In *CVPR*. 5044–5053.
- [2] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. 2021. On the limits of pseudo ground truth in visual camera re-localisation. In *ICCV*. 6218–6228.
- [3] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *ICLR*.

- [4] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. 2013. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*. 2930–2937.
- [5] Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *AI/ML for MDO applications*, Vol. 11006. SPIE, 369–386.
- [6] Julien Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. 2016. Learning to navigate the energy landscape. In *3DV*. IEEE, 323–332.
- [7] Ruihong Yin, Sezer Karaoglu, and Theo Gevers. 2023. Geometry-guided Feature Learning and Fusion for Indoor Scene Reconstruction. In *ICCV*. 3652–3661.
- [8] Lei Zhou, Zixin Luo, Tianwei Shen, Jiahui Zhang, Mingmin Zhen, Yao Yao, Tian Fang, and Long Quan. 2020. Kfnet: Learning temporal camera relocalization using kalman filtering. In *CVPR*. 4919–4928.