Supplementary Materials for

ImPosing: Implicit Pose Encoding for Efficient Visual Localization

This document presents further analysis on our method. We present additional ablation studies, latent space visualization, results of the attached video and reproducibility details. We invite readers to view the supplementary video where localization results are shown on a wide range of scenarios.

1 Ablation study on the pose encoder capacity

All experiments in the main paper report results with 4 layers in the pose encoder MLP network. We evaluate the localization results with different pose encoder capacity on Neighborhood and Countryside scenes from the 4seasons dataset[2].

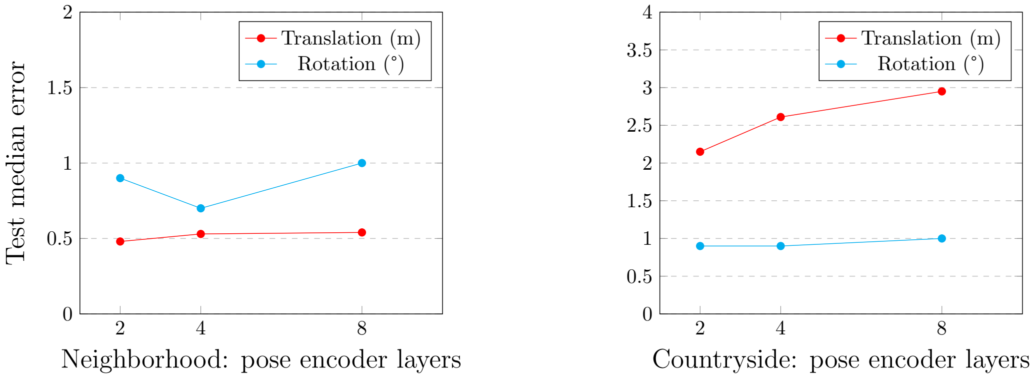

Figure 1: Localization accuracy depending on pose encoder capacity

Surprisingly, we observe that MLPs with a single hidden layer perform better on both scenes. The reason is not very clear: more capacity should not degrade performance except in case of overfitting, which is not the case here because the training loss is lower for smaller models as well. It might be that bigger MLPs just take more time to converge, we stopped the experiment after 250 epochs.

2 Ablation study on the similarity score

We tried to alternatives to cosine similarity for computing the score between image and camera pose latent vectors. A first alternative is based on L2 distance between the image and map signatures:

$$s(I,p) = 1 - \|f_I(I) - f_M(p)\|_2 \quad \mathbb{1}_{1-\|f_I(I)-f_M(p)\|_2>0} \tag{1}$$

Then, we also tried to learn this step with a 2 layers MLP, which takes $f_I(I) - f_M(p)$ as input, uses a ReLU activation in the hidden layer and outputs a score through the sigmoid activation. These solutions are compared on the Neighborhood scene from the 4seasons dataset[2] on figure 2. These scores are supervised with the target scores described in section 3.2 on the main paper.

The ablation confirms that cosine similarity performs better than other alternatives to compute the score between image-pose pairs.

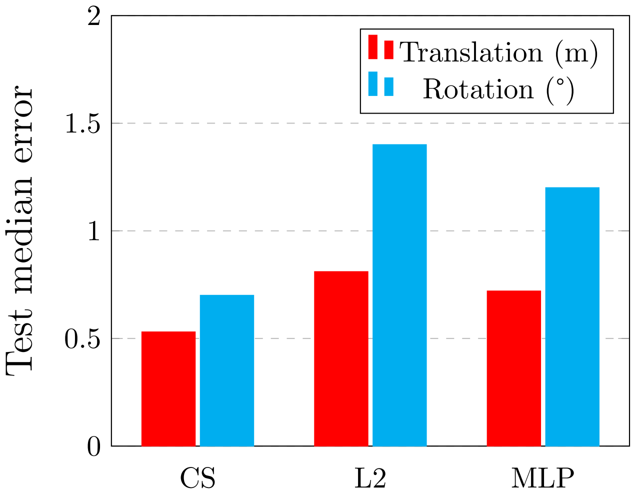

Figure 2: Localization accuracy depending on similarity score computation. CS stands for cosine similarity, L2 for the distance-based similarity and MLP for the learned score computation

3 Supplementary video

The attached video file shows sequential qualitative results of single scenes ImPosing models (corresponding to quantitative results of tables 1 and 2 in the main paper) on Daoxiang Lake [3] and 4Seasons datasets [2].

The input image is displayed on the top left corner. The right part shows the current predicted trajectory in red, ground truth poses in green and training trajectories in gray. The bottom left corner displays the 256 best candidates selected for pose averaging in red, the predicted pose in black and the groundtruth pose in green. Finally, the last plot shows the score of all candidates in the entire map from transparent ($s = 0$) to red ($s = 1$).

Scenes are displayed in the following order : Daoxiang Lake (00:00 to 01:03), Neighborhood (01:04 to 02:09), Office Loop (02:10 to 02:40), Business campus (02:41 to 03:38), City Loop (03:39 to 03:59), Countryside (04:00 to 04:44), Old Town (04:45 to 05:00).

These video samples show clearly advantages and limitations of our method:

- Coarse localization is correct most of the time, even in large maps with repetitive and featureless environments (see figure 3).
- In ambiguous scenarios, our method provides a multimodal distribution of scores in first iterations and then solves the ambiguity in further steps (see figure 4).
- Sequences of predictions are not temporally smooth, because each frame is treated independently in this experiment. In practice, this can be solved by filtering with a motion model, similar to [1].
- Precise pose estimation is sometimes inaccurate but sufficient to provide a lane level localization for navigation of autonomous vehicles.

It should also be noted that experiments on 4seasons dataset are extreme scenarios where the quantity of available data is small w.r.t. to the challenges introduced by weather conditions.

4 Latent space visualization

We attempt to visualize the structure of the latent space learned by ImPosing. We compute the latent vector of all reference poses of the Daoxiang Lake map. Then we compute a PCA of

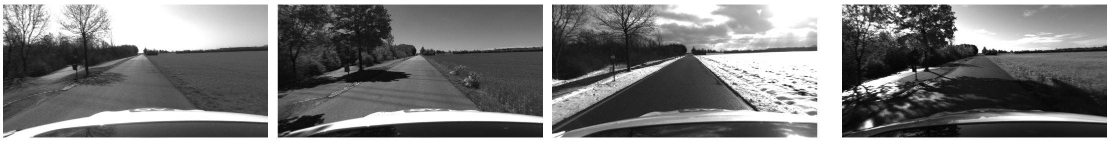

Figure 3: Featureless environments and varying weather conditions. Test is performed on the image on the right, while the network has been trained with 3 recordings with different lightning conditions. Our method is able to provide a coarse localization in these scenarios, where as image retrieval and pose regression competitors fail.

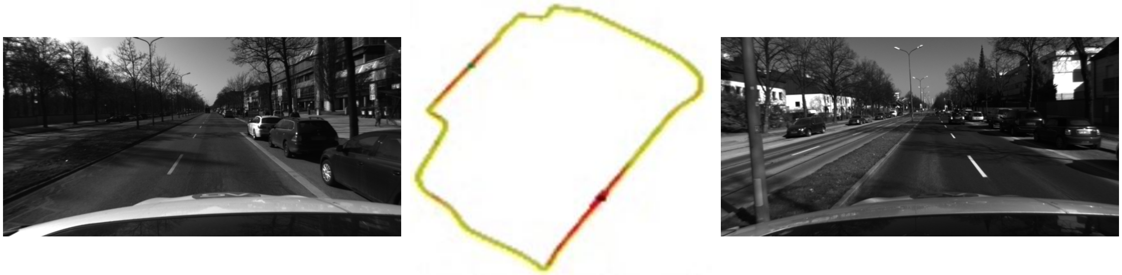

Figure 4: Multimodal score distribution in ambiguous cases. Many road areas present similar structure and appearance, introducing ambiguities in the localization task. In this scenario from the City Loop scene, the model outputs high scores for areas depicted in left and right images, which are very far one from each other. By refining the estimate in further steps, the model is able to solve this ambiguity in most cases.

the 256 dimensional vectors and display it on the map in figure 5. We observe that our pose encoder learns a smooth representation of the map, where close representations share similar visual content.

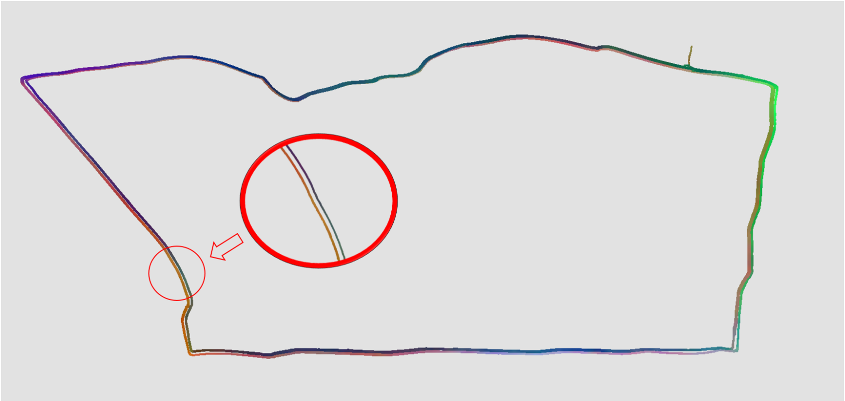

Figure 5: Latent space visualization. Training poses, colored by the 3 principal components of map descriptors. Poses with similar colors are close in the latent space. Opposite ways of the same road are represented by dissimilar representations. Best viewed in color

5 Datasets preparation

This section contains dataset splits used in our experiments to ensure reproducibility. To the best of our knowledge, 4Seasons [2] and Daoxiang Lake datasets [3] had not been used previously to evaluate direct learning-based methods. After preparing all the absolute poses of a map, we normalize the positions between -0.5 and 0.5, such that the networks converge faster in kilometers scale maps.

5.1 Oxford RobotCar dataset

The dataset can be downloaded here. We replicate experiments from previous methods using undistorted front camera images:

	Oxford Loop	Oxford Full
Training set	2014-06-26-09-24-58 2014-06-23-15-41-25	2014-11-28-12-07-13 2014-12-02-15-30-08
Test set	2014-06-26-08-53-56 2014-06-23-15-36-04	2014-12-09-13-21-02

5.2 Daoxiang Lake dataset

The dataset can be downloaded here. Vehicles are equipped with multiple sensors but we only use the front cameras images with associated vehicle poses. We don't use the train/test split provided by the dataset because the test set is not an entire held out sequence (images from the same sequence has been observed during training) and then is not a realistic test scenario.

	Daoxiang Lake dataset
Training set	20191216123346 20191130112819 20191025104732 20191021162130 20191014142530 20190924124848 20190918143332
Test set	20191225153609

5.3 4seasons dataset

The dataset can be downloaded here. Absolute poses are generated using available Python tools. We use keyframes from the left camera only.

	Neighborhood	Office Loop	Countryside	Bus. campus	City Loop	Old Town
Train	2020-03-26_13-32-55 2020-10-07_14-47-515 2020-10-07_14-53-52 2020-12-22_11-54-24 2021-02-25_13-25-15 2021-05-10_18-02-12	2020-03-24_17-36-22 2020-03-24_17-45-31 2020-04-07_10-20-32 2020-06-12_10-10-57 2021-01-07_12-04-03	2020-04-07_11-33-45 2020-06-12_11-26-43 2021-01-07_13-30-07	2020-10-08_09-30-57 2021-01-07_13-12-23	2020-12-22_11-33-15 2021-01-07_14-36-17	2020-10-08_11-53-41 2021-01-07_10-49-45 2021-05-10_21-32-00
Test	2021-05-10_18-32-32	2021-02-25_13-51-57	2020-10-08_09-57-28	2021-02-25_14-16-43	2021-02-25_11-09-49	2021-02-25_12-34-08

References

[1] Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: Coordinet: uncertainty-aware pose regressor for reliable vehicle localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2229–2238 (2022)

[2] Wenzel, P., Wang, R., Yang, N., Cheng, Q., Khan, Q., von Stumberg, L., Zeller, N., Cremers, D.: 4Seasons: A cross-season dataset for multi-weather SLAM in autonomous driving. In: Proceedings of the German Conference on Pattern Recognition (GCPR) (2020)

[3] Zhou, Y., Wan, G., Hou, S., Yu, L., Wang, G., Rui, X., Song, S.: Da4ad: End-to-end deep attention-based visual localization for autonomous driving. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)