Understanding the Limitations of CNN-based Absolute Camera Pose Regression - Supplementary Material

Torsten Sattler[1] Qunjie Zhou[2] Marc Pollefeys[3,4] Laura Leal-Taixé[2]
[1]Chalmers University of Technology [2]TU Munich [3]ETH Zürich [4]Microsoft

The supplementary material of two parts: **i)** The accompanying video shows how the base translations estimated by MapNet [1] are coupled to the image content and illustrates the poses predicted for the test images in some of the scenes shown in the paper. Sec. A gives a short overview over the video. **ii)** Sec. B presents an additional experiment on the DeepLoc dataset [6] that was left out of the paper due to space constraints.

A. Supplementary Video

The video consists of two parts: The first part shows how the impact of each estimated base translation on the predicted pose depends on the image content. This is shown for the training images from the scene from Fig. 2(right) in the paper.

The second part shows the positions estimated for the test images. We show the test image itself, the most similar training image (where similarity is measured based on the embeddings in the high-dimensional space), the base translations for the two images, and a 2D top-down view of the camera trajectories. In the 2D view, we show the ground truth training and testing positions, the pose of the current test image predicted by an absolute pose regression technique, the ground truth pose of the test image, and the pose of the most similar training images.

For all experiments shown in the video, the absolute pose regression technique used was MapNet [1]. Only test images that can be localized by Active Search [7] are shown.

B. Experiments on the DeepLoc Dataset [6]

The DeepLoc[1] dataset [6] was captured from a robot driving a triangular-shaped trajectory multiple times (*c.f.* Fig. 1). In contrast to the RobotCar dataset [4], which was captured in an urban environment, the DeepLoc dataset shows a significant amount of vegetation.

Tab. 1 (first row) compares the results obtained with DenseVLAD [9] without (*DenseVLAD*) and with interpolation (*DenseVLAD+Inter.*) with the results for various absolute pose regression techniques reported in [6]. Again, DenseVLAD significantly outperforms pose regression approaches based on a single image [2, 3, 5]. The table also compares DenseVLAD and DenseVLAD+Inter. against three sequence-based approaches, VLocNet [10], VLocNet++$_{\text{STL}}$ [6], and VLocNet++$_{\text{MTL}}$ [6]. All three directly fuse feature map responses from the previous time step $t-1$ into the CNN that predicts the pose at time t. VLocNet++$_{\text{MTL}}$ also integrates some form of higher-level scene understanding through semantic segmentation. All three methods operate on image sequences and thus use more information compared to DenseVLAD, which only uses a single image for localization. Still, DenseVLAD outperforms VLocNet [10].

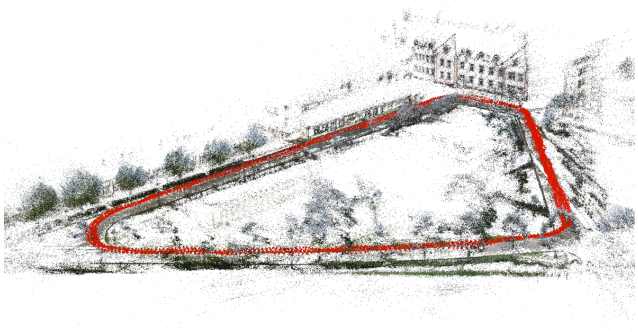

Figure 1. Visualization of the SfM model of the DeepLoc dataset [6] that we constructed from the training images (red).

The ground truth for the DeepLoc dataset was created using LIDAR-based SLAM. The dataset only provides the poses of the LIDAR sensor and not the cameras. This is not an issue for pose regression techniques as the camera and the LIDAR are related by a fixed (but unknown) transformation and it is irrelevant for the regressor which of the two local coordinate systems is used. However, not knowing the relative transformation from the LIDAR to the camera coor-

[1] http://deeploc.cs.uni-freiburg.de/

Pose Net [3]	Bay. PoseNet [2]	SVS Pose [5]	VLocNet [10]	DenseVLAD [9]	DenseVLAD +Inter.	VLocNet++$_{\text{STL}}$ [6]	VLocNet++$_{\text{MTL}}$ [6]	Active Search [7]
2.42m, 3.66°	2.24m, 4.31°	1.61m, 3.52°	0.68m, 3.43°	0.57m, 3.15 °	0.48m, 3.14°	0.37m, 1.93°	0.32m, 1.48°	
				0.51m, 2.57°	0.44m, 2.52°			0.01m, 0.04°

Table 1. Median position and orientation errors on the **DeepLoc** dataset [6]. *DenseVLAD+Inter.* uses the top-15 retrieved images for interpolation. We show results for (top row) the original dataset and (bottom) our SfM version of the dataset.

dinate system prevents us from easily creating a 3D model for structure-based methods. In order to be able to compare against Active Search [7], we thus created a second version of the dataset using SfM [8]. To this end, we ran SfM on both the training and test images together. We then registered the SfM model against the LIDAR ground truth poses[2] to recover the scale of the model. This provided us with ground truth poses for the training and test images. Finally, we used the ground truth poses of the training images and the feature matches between them to triangulate the 3D model used by Active Search[3]. This ensures that the 3D model used for localization only contains information from the training images.

The second row of Tab. 1 shows the results obtained by Active Search on our version of the dataset. As can be seen, Active Search is significantly more accurate than all pose regression techniques, including VLocNet++$_{\text{MTL}}$, even though it only uses a single image for localization. For reference, we also include results obtained with DenseVLAD and DenseVLAD+Inter. on this new version of the dataset. As can be seen, the results obtained via DenseVLAD and DenseVLAD+Inter. do not change significantly between both versions of the datasets. This shows that the results obtained by Active Search and the pose regression algorithms on the two variants of the dataset are comparable.

References

[1] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-Aware Learning of Maps for Camera Localization. In *CVPR*, 2018. 1

[2] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *ICRA*, 2016. 1, 2

[3] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 1, 2

[4] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The Oxford RobotCar dataset. *IJRR*, 36(1):3–15, 2017. 1

[5] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-DoF global localization in outdoor environments. In *IROS*, 2017. 1, 2

[6] Noha Radwan, Abhinav Valada, and Wolfram Burgard. VLocNet++: Deep Multitask Learning For Semantic Visual Localization And Odometry. *RA-L*, 3(4):4407–4414, 2018. 1, 2

[7] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *PAMI*, 39(9):1744–1756, 2017. 1, 2

[8] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 2

[9] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 1, 2

[10] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep Auxiliary Learning For Visual Localization And Odometry. In *ICRA*, 2018. 1, 2

[2] There seems to be some drift in the vertical direction for the LIDAR poses while there seems to be little height variation in the scene. We thus use a variant of the original ground truth positions, where all heights are set to the same value, for computing the alignment between the SfM model and the positions.

[3] As was done for the datasets used in the paper.