Fine-Grained Segmentation Networks: Self-Supervised Segmentation for Improved Long-Term Visual Localization

-

Supplementary Material

Måns Larsson Erik Stenborg Carl Toft Lars Hammarstrand Torsten Sattler Fredrik Kahl
Chalmers University of Technology
{mans.larsson, erik.stenborg, carl.toft, lars.hammarstrand, torsat, fredrik.kahl}@chalmers.se

This supplementary material provides details that could not be included in the paper submission due to space limitations: Sec. A provides details on the construction of the contingency tables used in Sec. 5.1 of the paper. Sec. B details the impact of using more fine-grained segmentations on the number of inliers and the inlier ratio in the context of visual localization (*c.f.* lines 738 to 741 in the paper). Finally, Sec. C describes the contents of the videos that are provided as part of the supplementary material.

A. Contingency Tables

As mentioned in the main paper, a contingency table displays the interrelation between two sets of assignments of the same data by forming a two-dimensional histogram, where each dimension corresponds to one of the assignments. In our case, the dimensions corresponds to the semantic class labels and cluster indices respectively. In practice, to create the tables visualized in Fig. 3 of the main paper, we take the index of the output cluster from the FGSN, c_i, and the semantic class of the annotation, t_i, for each pixel in each image of the test set. For each pair (c_i, t_i) we add one to value at row t_i and column c_i. A parallel can be drawn to a confusion matrix that is a special case of a contingency table, with true assignments for rows and predicted assignments for columns.

B. Visual Localization: Inlier counts and ratios

Fig. A shows cumulative distributions for the inlier count and inlier ratio for FGSNs with varying numbers of clusters. For this experiment, we use only the Simple Semantic Match Consistency (SSMC) approach. We compare using FGSNs to filtering with the 19 Cityscapes classes obtained from a network trained on Cityscapes, Vistas, and the correspondence datasets from [1]. In addition, we provide the results obtained without any semantic filtering as a baseline.

As can be seen from Table 2 of the main paper, SSMC benefits from using more fine-grained segmentations up to a certain point. For 100 and 200 clusters, the localization performance is considerably better compared to the baseline of using semantic classes. Fig. A shows that the inlier ratio CDF is lower for these, meaning that more outliers have been removed, thus increasing the probability that RANSAC finds the correct pose. For 1000 clusters however, the segmentations become too detailed. This results in a high inlier ratio since many outliers are removed. However, it also results in a lower absolute number of inlier since also correct matches are removed. This ultimately leads to a lower localization performance.

C. Supplementary Videos

C.1. Fine-Grained Segmentations

This supplementary video contain example outputs from the FGSNs for several traversal during different seasons and image conditions. The networks used to create the segmentation were trained with correspondence loss. The video is available at https://youtu.be/jXyA4wlm400.

C.2. Particle Filter-based Semantic Localization

The supplementary video compares the performance of the Particle Filter-based Semantic Localization (PFSL) approach [3] when using a semantic segmentation algorithm with 19 classes trained on Cityscapes, Vistas, and the correspondence datasets from [1] and when using a FGSN with 200 clusters, also trained on then correspondence datasets [1]. For both version we use only stationary classes in the localization filter. In Cityscapes' classes that means the 11 classes "road", "sidewalk", "building", "wall", "fence", "pole" "traffic light", "traffic sign", "vegetation", "terrain", and "sky". When using FGSN we can not assign stationary classes in this way, but instead we look at which classes have many correspondences in the training data, and use those as stationary. From the training data we

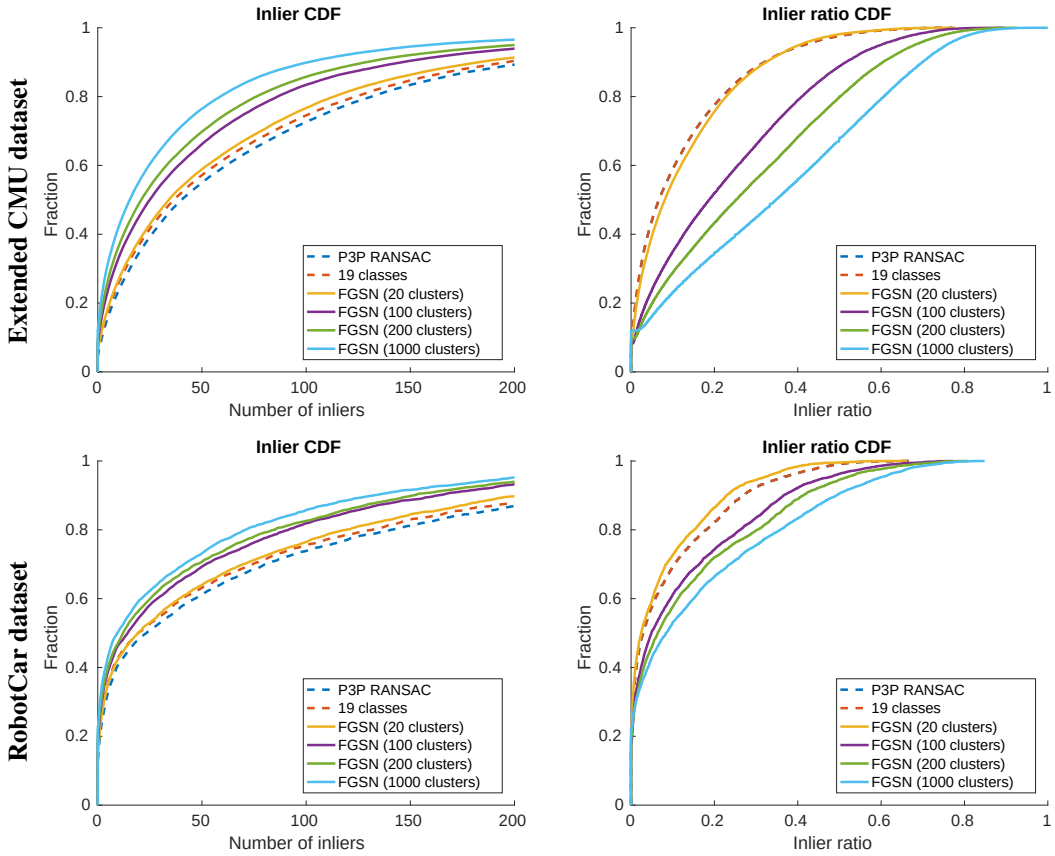

Figure A. Inlier count and inlier ratio on the Extended CMU dataset (above) and the RobotCar dataset (below) using SSMC. FGSNs with varying amount of clusters are evaluated against two baselines. For the for the "19 classes" [1], the Cityscapes classes are used for match consistency, while for the "P3P RANSAC" no filtering is done. Ideal curves are flat for a small number of inliers / inlier ratio and the quickly grow for a larger number of inliers / inlier ratio.

obtain discrete probability mass functions over the classes both for how the correspondences are distributed, $p_c(c)$, and for how all pixels in the images are distributed, $p_p(c)$. If the ratio $p_c(c)/p_p(c) > 0.2$ we select the class c as stationary, and use it in the localization.

The top row shows results obtained with semantic segmentation and the bottom row shows results obtained via our FGSN. The left and right columns show segmentations of the left and right camera of the vehicle used to capture the CMU dataset, respectively. In addition, the points in the point cloud visible in the camera are shown in the image. Gray pixels indicate non-stationary classes or clusters and are hence not used for localization. The middle column shows the semantically labeled 3D point cloud of part of the extended CMU dataset (obtained by backprojecting the segmentations of the database images onto the 3D points) and the reference poses for the vehicle[1] (orange dots). The reference pose corresponding to the current images is marked with a cross. We also show the position estimated by PFSL (black dot) and the covariance ellipse of PFSL's estimate. The video is available at https://youtu.be/-HoLNolQKoM.

[1]The authors of [2] provided reference poses for a subset of the extended CMU dataset to aid this visualization.

References

[1] Måns Larsson, Erik Stenborg, Lars Hammarstrand, Torsten Sattler, Marc Pollefeys, and Fredrik Kahl. A Cross-Season Correspondence Dataset for Robust Semantic Segmentation. In *CVPR*, 2019. 1, 2

[2] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *CVPR*, 2018. 2

[3] Erik Stenborg, Carl Toft, and Lars Hammarstrand. Long-term visual localization using semantically segmented images. *ICRA*, 2018. 1