Efficient Large-Scale Semantic Visual Localization in 2D Maps - Supplementary Material

Tomas Vojir[1,2][0000−0001−7324−5883], Ignas Budvytis[1][0000−0002−4278−198X], and Roberto Cipolla[1][0000−0002−8999−2151]

[1] University of Cambridge, Cambridge, UK
[2] Czech Technical University in Prague, Prague, CZ

1 Introduction

This document provides supplementary material for the main paper. Section 2 describes the process of creating alternative trajectories for a 2D map and query trajectories. Section 3 contains the explanation of the qualitative visualizations of a single point and trajectory retrieval shown in the accompanying videos. Section 4 provides additional results of our method for different sources of depth.

2 Map Alternative and Query Trajectories Details

There are two main stages to generate alternative and query trajectories. First is to create a connectivity graph between map locations, and secondly, random sampling of trajectories of a given length.

Connectivity graph. Given a map locations (GPS locations of available Google StreetView images), a connection graph is computed as follows: (i) for each location create an edge (connection) to all neighbor locations within 20m and (ii) remove the edge if it intersects a building.

Trajectories generation. Given the connection graph, we generated 200k unique random trajectories for Cambridge and 500k random unique trajectories for San Francisco as follows:

1. Sample uniformly map location.
2. Obtain neighbours (nodes with a direct edge in the connectivity graph) which are in range from 7m to 18m.
3. Remove neighbours that form angle outside the interval of $-120°$ to $120°$ with the trajectory. The angle is computed between the vector formed from the two most recently added locations and the vector formed from the most recent location and the node of interest. This constraint prevents from having backward turns which are not likely in real trajectories.
4. From the remaining neighbors, choose one in random if the last 5 locations created a straight line ($< 15°$) else choose one with minimum angle. This condition prevents trajectories from forming multiple turns in a row (e.g. constant switching of lanes in the multi-lane road).

Fig. 1. This figure shows the query trajectories for the Cambridge (left) and San Francisco (right). A different color visualizes the different trajectories. Note that, especially for San Francisco, the query trajectories are evenly distributed mainly because of the larger number of buildings per image in the San Francisco city.

5. Repeat from (2) until trajectory is of desired length (e.g. 32).

The test trajectories were generated using the same protocol, with the exception of the following restrictions. In particular, all locations in a trajectory have to have at least three close visible buildings in the corresponding images. Note that even using these restrictions, the test trajectories are evenly spaced in the cities, see Figure 1.

3 Additional Result Visualizations

This section provides the explanation of the visualisations used in the supplementary videos.

3.1 Single Query Localization

Videos *single_point_loc_videos_Cambridge.mp4* and *single_point_loc_videos_SanFrancisco.mp4* show qualitative results for a single point localization of several sample trajectories from Cambridge and San Francisco cities respectively. Figure 2 illustrates a single frame from one of the aforementioned videos and explains it.

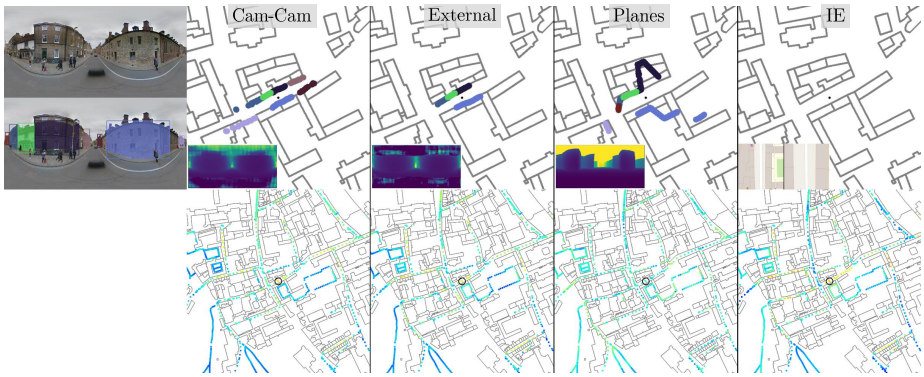

Fig. 2. The left column shows a query image and the corresponding segmentation. The first row of the second to fifth columns shows a zoomed-in map with overlayed descriptors extracted from different depth sources (i.e. same city trained depth, external depth, and depth normals from Google StreetView) and results for Image Embedding method. The bottom row shows the localization score for each map location, using "jet" colormap (i.e. red corresponds to the most similar location and blue to the least similar location) and the correct location marked by the black circle.

3.2 Trajectory Localization

Videos *trajectory_loc_videos_Cambridge.mp4* and *trajectory_loc_videos_SanFrancisco.mp4* show qualitative results for trajectory level localization. Figure 3 shows single frame from one of the aforementioned videos and explains it.

4 Additional Results for Different Depth Sources

In this section, we provide additional quantitative results for the trajectory localization and single point localization when different sources of depth are used. We tested three sources of depth information: (i) mono-depth network trained from images from the same city (reported in the main paper), (ii) mono-depth network trained on different cities than the tested one (denoted as external) and (iii) depth accompanying Google StreetView images. The results for the external depth are reported in the Table 4 and results for Google StreetView depth are reported in Table 5.

References

1. Abonce, O.S., Zhou, M., Calway, A.: You Are Here: Geolocation by Embedding Maps and Images. arXiv:1911.08797 (2019)

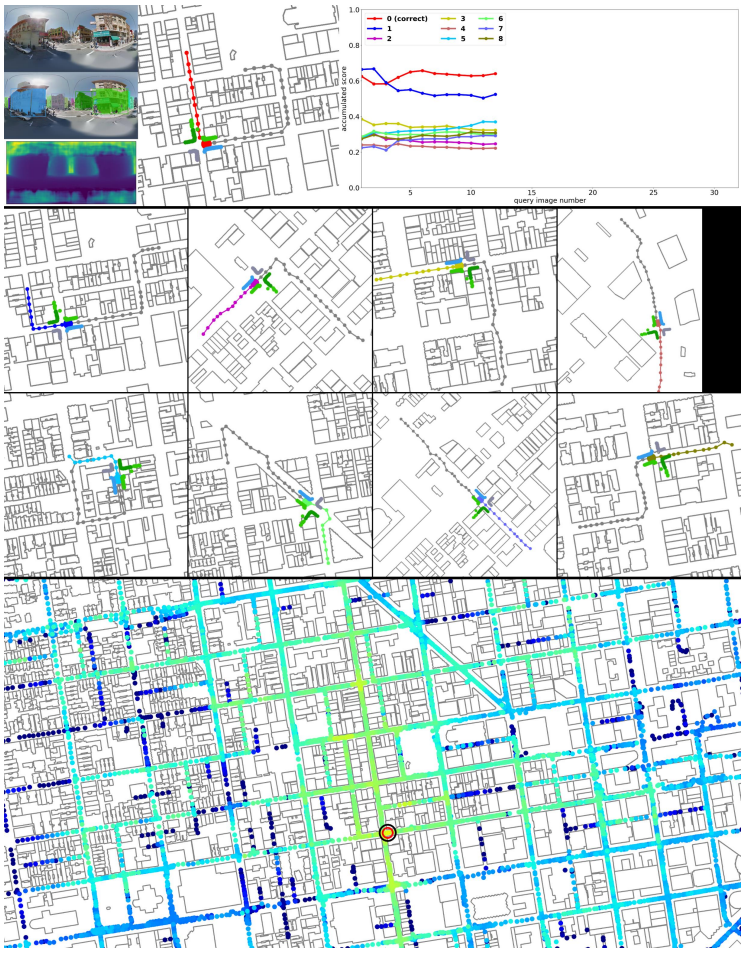

Fig. 3. This figure shows a single frame from the supplementary video, which illustrates the qualitative results of the trajectory level localization. The video shows the progress of the trajectory localization accuracy while increasing the length of the trajectory (from 1 to 32). The top left column of three images shows the most recent query image within a trajectory, corresponding segmentation labels and depth values. The middle image of the top row illustrates the query trajectory where the red color shows already traveled frames. The final image in the top row shows the graph, which plots the similarity score between the query trajectory and (i) its correct location (red), (ii) the top-2 retrieved trajectory (blue) and (iii) 7 random alternative trajectories from the map. The second and third rows of the figure the alternative trajectories (note that blue trajectory is always second best). The last row plots similarity score for each Google StreetView map location of the best matching trajectory which ends at that point. The similarity is encoded using "jet" colormap (i.e. red corresponds to high similarity and blue to low similarity). The black circle marks the correct ground-truth location and the red circle marks the current most likely endpoint of the trajectory.

Method	Single Query Localization [Top x%]						Trajectory Localization [Top 1]								
	Cambridge		San Francisco		Average		Cambridge			San Francisco			Average		
	1%	10%	1%	10%	1%	10%	80m	160m	320m	80m	160m	320m	80m	160m	320m
BSD (75%)	9.6	45.1	13.1	52.9	11.4	49.0	6.9	27.2	70.0	1.3	6.2	21.3	4.1	16.7	45.7
Ours: Hand-Crafted	28.4	81.5	34.6	76.4	31.5	79.0	10.6	20.3	52.1	9.9	21.9	51.4	10.3	21.1	51.8
Ours: Pittsburgh	29.8	77.0	34.5	70.9	32.2	74.0	14.3	34.1	65.0	16.1	38.1	75.1	15.2	36.1	70.1
Ours: Pittsburgh + Manhattan	33.5	80.4	48.1	83.2	40.8	81.8	21.2	41.9	71.9	41.1	66.5	89.2	31.2	54.2	80.6
Ours: Cambridge	33.2	75.5	40.7	80.2	37.0	77.9	20.7	41.0	64.5	28.4	52.7	81.5	24.6	46.9	73.0
Ours: San Francisco	33.5	78.8	49.3	85.6	41.4	82.2	24.0	49.8	78.3	42.2	**74.0**	**94.2**	33.1	61.9	86.3
Ours: All (P+M+C+SF)	**36.1**	80.1	**51.7**	**87.1**	43.9	**83.6**	**28.6**	**58.1**	**90.3**	**46.9**	**80.0**	**95.7**	**37.8**	**69.1**	**93.0**
IE: Pittsburgh	**49.3**	**93.4**	46.5	85.6	**47.9**	**89.5**	**40.1**	**65.9**	**85.7**	34.4	65.8	90.3	**37.3**	**65.9**	**88.0**
IE: Pittsburgh + Manhattan	35.9	**87.8**	**53.9**	**91.1**	**44.9**	**89.5**	16.1	35.9	63.6	**42.6**	71.4	92.0	29.4	53.7	77.8

Fig. 4. The table shows the retrieval accuracy for top 1% and 10% best ranked single locations and the accuracy of retrieving the right trajectory at different lengths of 80m, 160m and 320m when **external depth** source is used for query images for our methods. The "average" column shows a combined score over Cambridge and San Francisco cities. The IE stands for the Image Embedding method [1]. The best and second best results are in bold green and blue respectively.

Method	Single Query Localization [Top x%]						Trajectory Localization [Top 1]								
	Cambridge		San Francisco		Average		Cambridge			San Francisco			Average		
	1%	10%	1%	10%	1%	10%	80m	160m	320m	80m	160m	320m	80m	160m	320m
BSD (75%)	9.6	45.1	13.1	52.9	11.4	49.0	6.9	27.2	70.0	1.3	6.2	21.3	4.1	16.7	45.7
Ours: Hand-Crafted	29.9	72.4	35.6	75.8	32.8	74.1	10.1	12.4	42.9	10.1	15.9	47.7	10.1	14.2	45.3
Ours: Pittsburgh	30.7	73.1	40.2	76.9	35.5	75.0	18.0	39.2	63.6	22.2	40.4	78.9	20.1	39.8	71.3
Ours: Pittsburgh + Manhattan	**37.3**	78.4	45.2	81.9	41.3	**80.2**	**24.9**	**50.2**	**77.0**	28.8	54.0	84.9	26.9	52.1	**81.0**
Ours: Cambridge	32.5	74.5	36.2	75.0	34.4	74.8	24.0	41.9	61.8	18.5	36.6	66.7	21.3	39.3	64.3
Ours: San Francisco	34.0	76.5	42.4	82.2	38.2	79.4	24.4	42.4	71.0	23.7	46.9	80.0	24.1	44.7	75.5
Ours: All (P+M+C+SF)	36.2	76.2	45.9	82.3	41.1	79.3	**24.9**	48.8	75.6	29.5	58.3	86.2	27.2	53.6	80.9
IE: Pittsburgh	**49.3**	**93.4**	**46.5**	**85.6**	**47.9**	**89.5**	**40.1**	**65.9**	**85.7**	**34.4**	**65.8**	**90.3**	**37.3**	**65.9**	**88.0**
IE: Pittsburgh + Manhattan	35.9	**87.8**	**53.9**	**91.1**	**44.9**	**89.5**	16.1	35.9	63.6	**42.6**	**71.4**	**92.0**	**29.4**	**53.7**	77.8

Fig. 5. The table shows the retrieval accuracy for top 1% and 10% best ranked single locations and the accuracy of retrieving the right trajectory at different lengths of 80m, 160m and 320m when **Google StreetView depth** source is used for query images for our methods. The "average" column shows a combined score over Cambridge and San Francisco cities. The IE stands for the Image Embedding method [1]. The best and second best results are in bold green and blue respectively.