EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition (Supplementary)

Gabriele Berton*[1] Gabriele Trivigno*[1] Barbara Caputo[1] Carlo Masone[1]
[1] Politecnico di Torino
{gabriele.berton, gabriele.trivigno, barbara.caputo, carlo.masone}@polito.it

Supplementary

In this supplementary material we show:

- in Sec. 1 how different training methods use different images at train time;
- in Sec. 2 we provide further information regarding the datasets;
- in Sec. 3 further quantitative and qualitative results from our large set of experiments.

1. Data for Different Training Methods

Visualization of the training data used by different training methods is shown in Fig. 1.

In the image we can see that the query-positive pairs mined with NetVLAD [3] have very little viewpoint shift. NetVLAD [24] uses positive mining to obtain the most similar positive to the query, which is then used within a triplet loss. Note that this is different from negative mining (which in NetVLAD is also performed. This is the same (or very similar) approach used by most following works [13, 14, 10, 25, 18, 17, 11].

CosPlace [4] uses images with the same orientation for a given class, with images being just a few meters apart from each other. Conv-AP [1] and MixVPR [2] use a predefined set of classes by GSV-Cities [1], with little intra-class viewpoint variations.

In contrast with previous methods, EigenPlaces creates training data by ensuring large viewpoint shifts between images, as visually shown in the last row of Fig. 1, which in turn make the trained model more robust.

(a) Three Query-positive pairs mined as in **NetVLAD**

(b) Images within three classes, created with **CosPlace**

(c) Images within three classes, as used for training of **Conv-AP** and **MixVPR**

(d) Training data from three classes created with **EigenPlaces**

Figure 1: **Training data with different methods.** Only the images used by EigenPlaces provide large viewpoint shifts.

2. Datasets

We test all models on a large number of datasets, which helps to thoroughly understand each method's strength and weaknesses. To download a number of datasets (namely AmsterTime, Eynsham, San Francisco Landmark, Nordland, St Lucia and SVOX) we used the open-source automatic downloader from [6], as this ensures maximum reproducibility for future research. Below is a short description for each of the datasets.

AmsterTime [23] is a collection of over one thousand pairs of query-reference images from the city of Amsterdam. For

each pair, the query is a grayscale historical image, and its reference is a modern-day photo which represents the same place, as confirmed by human experts. The pairs provide multiple domain shifts: viewpoints, long-term temporal changes, modality (RGB vs grayscale), different cameras. This makes AmsterTime one of the most challenging dataset available, despite its relatively small scale.

Figure 2: Examples from AmsterTime query and database.

Eynsham [9] consists of images from cameras mounted on a car and GPS co-ordinates of the car going around around a loop twice. The original images are 360° panoramas, that we split in crops following standard practice [21, 20, 6]. The images are grayscale, and the car drives around the Oxford countryside, passing also through the city of Oxford.

Figure 3: Examples from Eynsham query and database.

Pitts30k and Pitts250k [21] are perhaps the most used dataset for VPR to date, on which a large number of works present their results [3, 13, 14, 10, 1, 2, 5, 11, 24, 4]. They are built with Google StreetView images from the city center of Pittsburgh, by ensuring that database and queries are taken in different years. They provide three splits for training, validation and test. The 6816 test queries used for Pitts30k are a subset of the 8280 used for Pitts250k, whereas the Pitts250k database is roughly 8 times larger.

Figure 4: Examples from Pitts30k query and database.

Tokyo 24/7 [20] is a challenging dataset from the center of Tokyo. The database is made from Google StreetView, whereas the queries are a collection of smartphone photos from 105 places, and each place is photographed during the day, at sunset and at night. This results in 315 queries, each to be geolocalized independently.

Figure 5: Examples from Tokyo 24/7 query and database.

San Francisco Landmark [8] is a large dataset from the center of San Francisco with a database of more than 1M images, and a set of 598 queries collected with a smartphone.

Figure 6: Examples from San Francisco Landmark query and database.

San Francisco eXtra Large (SF-XL) [4] is a huge dataset covering the whole city of San Francisco with over 41M images. Its test set covers the same with a less dense set of 2.8M images. Two sets of queries are used: the first (*test v1*) is a challenging set of 1000 images from Flickr, with multiple challenges like night images and photos from the sidewalk. *Test v2* uses the same set of queries from San Francisco Landmark.

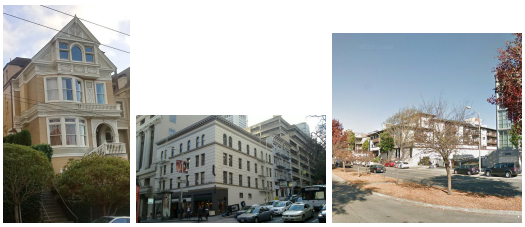

Figure 7: Examples from SF-XL: (left to right) a query from SF-XL *test v1*, a query from SF-XL *test v2* and an example from database.

MSLS [22] is the Mapillary Street-Level Sequences dataset, which has been created for image and sequence-based VPR. The dataset consists of more than 1M images from multiple cities, although only a small subset is used for evaluation. Following common practice [11, 2, 1] we evaluate on their validation set, as the labels for the test set have not been released. The test set is from the cities of Copenhagen and

San Francisco, and although being mostly single-domain, it provides a small number of night and lateral-view images.

Figure 8: Examples from MSLS query and database.

Nordland [19] was collected by recording a video from a train riding through the Norwegian countryside, and traversing the same path across four seasons. Images are then extracted at 1FPS. Following previous works [12, 11] we use the winter traverse as queries and summer as database, which have been post-processed to ensure alignment of the frames. Unlike in most other VPR datasets, a query is considered correctly localized if the matched database image is less than 10 frames away.

Figure 9: Examples from Nordland query and database.

St Lucia [16] is a dataset collected with a car-mounted camera, with long videos from multiple drives along the same area: the St Lucia suburb of Brisbane. Following [6], we use the first and last drive (of the nine available) as queries and database, and we sample one frame every 5 meters of driving.

Figure 10: Examples from St Lucia query and database.

SVOX [7] is a cross-domain dataset built from cross-domain VPR, that allows to evaluate on multiple weather conditions. It spans the city of Oxford, with a large (single-domain) database from Google StreetView images: the queries are instead from the Oxford RobotCar dataset [15], providing a number of weather conditions, such as overcast, rainy, sunny, snowy and night domains.

Figure 11: Examples from SVOX: in the top row are an image from the database, and queries from the night and overcast domain; in the bottom row are queries from rain, snow, and sun domains.

3. Experiments

3.1. Further results

In this section we report different values of recalls for the same datasets as in the main paper. Results on multi-view datasets are in Tab. 1 , whereas on frontal-view datasets in Tab. 2.

3.2. Qualitative results

Some qualitative results are shown in Fig. 12. The figure allows to understand the strengths of EigenPlaces in a more visual and intuitive way. We can see that EigenPlaces is able to handle difficult points of view, such as photos taken from the sidewalk.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 1, 2, 4

[2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 1, 2, 4

[3] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1437–1451, 2018. 1, 2, 4

[4] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *CVPR*, June 2022. 1, 2, 4

[5] Gabriele Berton, Carlo Masone, Valerio Paolicelli, and Barbara Caputo. Viewpoint invariant dense matching for visual geolocalization. In *IEEE International Conference on Computer Vision*, pages 12169–12178, October 2021. 2

[6] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3

Method	Backbone	Desc. Dim.	AmsterTime	Eynsham	Pitts30k	Pitts250k	Tokyo 24/7	San Francisco Landmark	SF-XL test v1	SF-XL test v2
CosPlace [4]	VGG-16	512	38.7/61.3/67.3/72.9	88.3/92.7/94.1/95.1	88.4/94.6/95.7/96.5	89.7/96.6/97.8/98.4	81.9/90.2/92.4/95.9	80.8/87.5/89.6/91.0	65.9/75.3/77.4/80.4	83.1/91.3/94.8/95.7
EigenPlaces (Ours)	VGG-16	512	38.0/59.2/64.8/71.9	89.4/93.6/94.8/95.7	89.7/95.0/96.4/97.4	91.2/96.8/97.9/98.6	82.2/90.8/93.3/94.3	83.8/90.6/91.8/93.0	69.4/78.4/82.0/84.8	86.3/93.6/95.3/96.2
NetVLAD [3]	VGG-16	4096	16.3/29.8/36.9/46.4	77.7/87.8/90.5/92.5	85.0/92.1/94.4/95.9	85.9/93.1/95.0/96.3	69.8/81.3/82.9/85.7	79.1/87.6/89.6/90.8	40.0/52.9/57.8/61.9	76.9/88.8/91.1/92.8
SFRS [10]	VGG-16	4096	29.7/48.5/55.6/63.4	72.3/83.5/87.1/89.8	89.1/94.6/96.1/97.0	90.4/96.3/97.6/98.2	80.3/88.6/91.7/92.7	83.1/90.0/91.8/92.8	50.3/60.0/64.9/68.5	83.8/90.5/92.8/94.3
CosPlace [4]	ResNet-50	128	39.9/61.8/67.9/74.2	88.6/93.0/94.5/95.4	89.0/94.7/96.1/97.1	89.6/96.0/97.5/98.3	81.0/90.8/93.7/94.6	82.9/89.6/91.1/91.8	69.1/76.5/79.0/82.2	86.5/92.6/94.8/96.7
MixVPR [2]	ResNet-50	128	23.1/40.1/49.4/56.9	84.8/90.6/92.1/93.4	87.7/94.3/95.7/96.9	88.7/95.8/97.2/98.2	56.8/73.3/80.0/84.1	66.9/76.3/80.1/83.3	36.7/49.6/55.3/60.1	68.4/81.9/87.5/90.6
EigenPlaces (Ours)	ResNet-50	128	37.9/57.0/65.1/72.9	89.1/93.7/94.8/95.8	89.6/95.6/96.7/97.3	90.2/96.4/97.7/98.4	79.4/89.5/93.7/95.6	85.5/91.5/92.5/93.3	72.4/79.4/82.3/84.5	86.6/94.3/95.3/96.7
CosPlace [4]	ResNet-50	512	46.4/67.5/73.3/78.3	89.9/93.8/94.8/95.6	90.2/95.2/96.3/97.1	91.7/97.0/98.1/98.7	89.5/94.9/96.5/97.5	85.6/90.3/92.3/93.5	76.7/82.5/85.6/87.4	89.0/95.3/96.3/96.8
Conv-AP [1]	ResNet-50	512	28.4/46.5/52.8/60.4	86.2/91.5/93.1/94.3	89.1/94.6/96.1/97.0	90.4/96.7/97.8/98.4	61.3/77.8/82.5/87.3	68.4/78.4/81.6/84.6	41.8/53.1/58.0/62.7	64.0/74.6/79.1/84.1
MixVPR [2]	ResNet-50	512	35.8/52.8/60.0/65.9	87.6/92.0/93.3/94.3	90.4/95.4/96.3/97.2	93.0/97.8/98.6/99.0	78.4/86.7/90.2/93.0	79.4/86.1/88.3/89.6	57.7/70.3/74.2/77.4	84.3/91.6/94.0/94.5
EigenPlaces (Ours)	ResNet-50	512	45.7/68.5/74.6/80.1	90.5/94.3/95.3/96.2	91.9/96.4/97.4/97.9	93.5/97.8/98.7/99.0	89.8/95.2/95.9/96.5	89.5/**94.5/95.5/96.2**	82.6/87.6/90.3/91.9	90.6/95.5/**97.2/97.8**
CosPlace [4]	ResNet-50	2048	47.7/**69.8**/75.8/81.0	90.0/93.9/94.9/95.7	90.9/95.7/96.7/97.4	92.3/97.4/98.4/98.9	87.3/94.0/95.6/97.1	87.1/91.1/92.1/92.8	76.4/83.3/85.5/88.2	88.8/95.0/96.8/97.5
Conv-AP [1]	ResNet-50	2048	31.3/49.6/58.1/64.9	86.6/91.7/93.1/94.3	90.4/95.1/96.4/97.2	92.3/97.5/98.4/99.0	71.1/81.0/84.8/87.3	71.7/81.4/83.9/85.6	47.8/58.3/63.1/67.3	68.1/80.9/83.9/87.3
EigenPlaces (Ours)	ResNet-50	2048	**48.9**/69.5/**76.0/81.4**	**90.7/94.4/95.4/96.3**	**92.5/96.8/97.6/98.2**	**94.1**/97.9/98.7/99.1	**93.0/96.2/97.5/97.8**	**89.6**/94.3/95.3/95.8	**84.1/89.1/90.7/92.6**	**90.8/95.7**/96.7/97.5
Conv-AP [1]	ResNet-50	4096	33.9/53.0/59.1/66.8	87.5/92.2/93.5/94.6	90.5/95.3/96.6/97.5	92.3/97.8/98.6/99.0	76.2/85.1/87.3/89.2	73.7/81.6/84.6/86.3	47.5/59.7/63.8/67.8	74.4/86.6/89.0/90.8
MixVPR [2]	ResNet-50	4096	40.2/59.1/64.6/72.5	89.4/93.2/94.3/95.1	91.5/95.5/96.3/97.5	**94.1/98.2/98.9/99.3**	85.1/91.7/94.3/95.6	83.8/90.3/91.1/92.5	71.1/78.2/79.7/82.3	88.5/93.6/94.5/96.0
Conv-AP [1]	ResNet-50	8192	35.0/53.8/60.9/68.2	87.6/92.4/93.6/94.5	90.5/95.2/96.4/97.3	92.6/97.5/98.4/99.0	72.1/84.1/87.6/90.5	74.4/82.9/85.5/87.8	49.3/61.0/64.8/69.8	75.8/85.1/89.0/91.3

Table 1: **Recalls (R@1 / R@5 / R@10 / R@20) on multi-view datasets**, split according to the utilized backbone and descriptors dimension. Best overall results on each dataset are in **bold**, best results for each group are underlined.

Method	Backbone	Desc. Dim.	MSLS Val	Nordland	St Lucia	SVOX Night	SVOX Overcast	SVOX Rain	SVOX Snow	SVOX Sun
CosPlace [4]	VGG-16	512	82.6/89.9/92.0/94.3	58.5/73.7/79.4/84.8	95.3/97.9/98.9/99.5	44.8/63.5/70.0/77.6	88.5/93.9/95.2/96.7	85.2/91.7/93.8/95.3	89.0/94.0/94.6/96.0	67.3/79.2/83.8/88.4
EigenPlaces (Ours)	VGG-16	512	84.2/90.0/91.8/94.1	54.5/70.1/76.4/82.4	95.4/98.1/99.5/99.7	42.3/61.0/68.5/75.8	89.4/94.4/95.6/97.4	83.5/91.6/92.8/94.6	89.2/94.4/95.5/96.1	69.7/82.2/86.1/89.8
NetVLAD [3]	VGG-16	4096	58.9/70.8/75.0/79.1	13.1/21.1/26.1/32.0	64.6/80.3/85.8/91.3	8.0/17.4/23.1/29.6	66.4/81.5/85.7/89.3	51.5/69.3/74.7/80.4	54.4/71.8/77.2/82.4	35.4/52.7/58.8/65.8
SFRS [10]	VGG-16	4096	70.0/80.0/83.5/86.1	16.0/24.1/28.7/34.4	75.9/86.6/91.2/94.3	28.6/40.6/46.4/52.1	81.1/88.4/91.2/92.9	69.7/81.5/84.6/87.7	76.0/86.1/89.4/91.6	54.8/68.3/74.1/78.5
CosPlace [4]	ResNet-50	128	85.5/92.3/93.2/94.6	54.7/70.9/77.9/83.4	98.7/99.8/99.9/**100.0**	35.4/55.4/63.8/71.0	88.5/96.0/96.9/97.5	80.4/90.3/94.1/95.9	86.6/95.1/96.4/97.4	65.2/80.3/84.4/88.4
MixVPR [2]	ResNet-50	128	79.1/87.4/90.3/92.0	47.8/66.5/73.9/80.5	99.0/**99.9**/99.9/99.9	25.9/43.3/50.9/59.2	92.3/96.6/97.4/97.7	80.9/91.2/93.8/94.9	87.7/94.6/95.6/96.9	73.5/88.1/91.2/94.3
EigenPlaces (Ours)	ResNet-50	128	83.4/90.9/93.5/95.1	50.5/66.8/73.6/80.0	98.8/99.7/99.9/**100.0**	29.0/48.5/57.7/65.4	90.9/96.2/97.6/98.3	83.8/92.8/94.6/96.7	91.1/97.0/97.9/99.0	68.5/83.7/88.2/91.8
CosPlace [4]	ResNet-50	512	86.9/93.2/94.2/95.5	66.5/79.7/84.8/88.9	99.1/**99.9**/**100.0**/**100.0**	51.6/68.8/76.1/80.9	90.0/96.6/97.2/97.6	87.3/94.7/95.7/97.3	89.5/97.0/98.0/98.2	75.9/88.3/92.2/94.6
Conv-AP [1]	ResNet-50	512	82.3/90.3/91.6/93.5	59.2/74.6/80.1/85.2	99.2/**99.9**/99.9/99.9	36.0/52.5/61.2/67.9	90.5/95.9/96.9/98.2	80.3/90.0/93.0/95.4	86.4/95.3/96.6/98.3	75.3/88.1/91.5/93.1
MixVPR [2]	ResNet-50	512	83.6/91.5/93.4/94.3	67.2/81.0/85.9/90.0	99.2/**99.9**/**100.0**/**100.0**	44.8/63.2/71.0/77.0	93.9/97.7/98.3/98.7	86.4/93.9/96.3/97.4	93.9/97.6/97.9/98.5	78.7/91.2/93.6/95.4
EigenPlaces (Ours)	ResNet-50	512	**89.5**/93.6/94.5/96.1	67.9/81.1/85.6/89.6	99.5/**99.9**/**100.0**/**100.0**	51.5/70.8/78.4/84.0	92.8/97.6/97.9/98.4	89.0/95.5/97.1/98.1	92.0/97.5/98.3/98.7	83.1/93.8/95.7/**97.1**
CosPlace [4]	ResNet-50	2048	87.4/**94.1**/94.9/95.9	71.9/83.8/88.1/91.5	99.6/**99.9**/**100.0**/**100.0**	50.7/67.4/74.8/80.2	92.2/97.7/97.9/98.7	87.0/95.1/96.8/97.5	92.0/**98.4/98.9/99.1**	78.5/89.7/93.1/94.8
Conv-AP [1]	ResNet-50	2048	81.2/89.5/91.6/93.6	62.3/76.9/82.0/86.7	99.3/**99.9**/**100.0**/**100.0**	37.9/57.1/65.4/72.8	92.0/96.1/97.2/98.5	83.7/93.4/95.2/97.2	90.2/95.7/97.5/98.4	80.3/90.5/93.8/95.4
EigenPlaces (Ours)	ResNet-50	2048	89.1/93.8/**95.0/96.2**	71.2/83.8/88.1/91.6	99.6/**99.9**/**100.0**/**100.0**	58.9/76.9/82.6/87.0	93.1/97.8/98.3/98.7	90.0/96.4/98.0/**98.5**	93.1/97.6/98.2/98.6	**86.4/95.0/96.4**/96.8
Conv-AP [1]	ResNet-50	4096	82.8/89.9/91.8/94.5	59.6/74.4/79.7/84.9	99.6/**99.9**/**100.0**/**100.0**	41.9/61.4/68.7/76.5	91.2/95.8/97.1/98.1	81.9/92.6/95.2/96.9	87.9/95.7/97.7/98.7	82.0/91.7/94.5/96.0
MixVPR [2]	ResNet-50	4096	87.2/93.1/94.3/95.4	**76.2/86.9/90.3/93.3**	99.6/**99.9**/**100.0**/**100.0**	**64.4/79.2/83.1/87.7**	**96.2/98.3/98.9/99.2**	**91.5/97.2/98.1/98.5**	**96.8/98.4/98.9**/99.0	84.8/93.2/94.7/95.9
Conv-AP [1]	ResNet-50	8192	82.4/90.4/92.0/94.3	62.9/77.3/82.5/86.8	**99.7/99.9**/99.9/99.9	43.4/63.1/71.6/79.1	91.9/96.6/98.3/98.6	82.8/93.0/95.6/96.1	91.0/96.7/97.6/98.4	80.4/90.3/93.2/95.0

Table 2: **Recalls (R@1 / R@5 / R@10 / R@20) on frontal-view datasets**, split according to the utilized backbone and descriptors dimension. Best overall results on each dataset are in **bold**, best results for each group are underlined.

[7] Gabriele Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: A hybrid approach. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2918–2927, January 2021. 3

[8] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 737–744, 2011. 2

[9] M. Cummins and P. Newman. Highly scalable appearance-only slam - FAB-MAP 2.0. In *Robotics: Science and Systems*, 2009. 2

[10] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 369–386, Cham, 2020. Springer International Publishing. 1, 2, 4

[11] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 1, 2, 3

[12] S. Hausler, A. Jacobson, and M. Milford. Multi-process fusion: Visual place recognition using multiple image processing methods. *IEEE Robotics and Automation Letters*, 4(2):1924–1931, 2019. 3

[13] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3251–3260, 2017. 1, 2

[14] Liu Liu, Hongdong Li, and Yuchao Dai. Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization. In *IEEE International Conference on Computer Vision*, 2019. 1, 2

[15] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research*, 2017. 3

[16] Michael Milford and G. Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 24:1038–1053, 2008. 3

[17] Guohao Peng, Yufeng Yue, Jun Zhang, Zhenyu Wu, Xiaoyu Tang, and Danwei Wang. Semantic reinforced attention learning for visual place recognition. In *IEEE In-*

Figure 12: **Qualitative results with most popular methods.** Each column represents a query (top row) and the first predicted image from the database. We can see that EigenPlaces is able to better handle challenging viewpoints than previous methods.

ternational Conference on Robotics and Automation, pages 13415–13422. IEEE, 2021. 1

[18] Guohao Peng, Jun Zhang, Heshan Li, and Danwei Wang. Attentional pyramid pooling of salient visual residuals for place recognition. In *IEEE International Conference on Computer Vision*, pages 885–894, October 2021. 1

[19] N. Sünderhauf, P. Neubert, and P. Protzel. Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons. In *Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation*, page 2013, 2013. 3

[20] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):257–271, 2018. 2

[21] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2346–2359, 2015. 2

[22] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong

place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[23] B. Yildiz, S. Khademi, R. Siebes, and J. Van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2749–2755, Los Alamitos, CA, USA, aug 2022. IEEE Computer Society. 1

[24] Jian Zhang, Yunyin Cao, and Qun Wu. Vector of locally and adaptively aggregated descriptors for image feature representation. *Pattern Recognition*, 116:107952, 2021. 1, 2

[25] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. Attention-based pyramid aggregation network for visual place recognition. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 99–107. ACM, 2018. 1