

A APPENDIX

A.1 THE SNAPSHOT OF MSLS LEADERBOARD

The MSLS place recognition challenge ¹ is an authoritative competition for VPR with over 100 participants. Fig. 4 shows a snapshot of the MSLS challenge leaderboard at the time of submission. The proposed method (named “SuperPlace” due to the double-blind review policy) ranks first.



The screenshot shows the CodaLab interface for the MSLS challenge. The user 'SuperPlace' is highlighted in the top right corner. The leaderboard table is as follows:

#	User	Entries	Date of Last Entry	recall@5 ▲
1	SuperPlace	5	07/07/24	0.94 (1)
2	amaralibey CVPR'24 BoQ	1	07/07/24	0.90 (2)
3	LiQQQQQQQ	6	06/25/24	0.90 (2)
4	mapillary_challenge	8	04/17/24	0.90 (2)
5	Skyxuan	8	07/04/24	0.90 (3)
6	anonymous123	8	07/08/24	0.90 (4)
7	magnus	1	06/05/24	0.89 (5)
8	ningzuotao	16	12/20/23	0.89 (5)
9	razor	1	06/05/24	0.89 (6)
10	izquierdo CVPR'24 SALAD	25	11/15/23	0.89 (7)
11	uno	30	06/12/24	0.89 (8)
12	qixi	6	12/19/23	0.89 (8)
13	anonymous02 ICLR'24 SelaVPR	1	09/17/23	0.89 (8)

Figure 4: A snapshot of MSLS leaderboard. The upper-right corner of the screenshot indicates our username. By consulting the supplementary materials of SelaVPR (Lu et al., 2024b), we confirm that ‘anonymous02’ corresponds to SelaVPR.

A.2 COMPARISON OF DIFFERENT DIMENSIONS

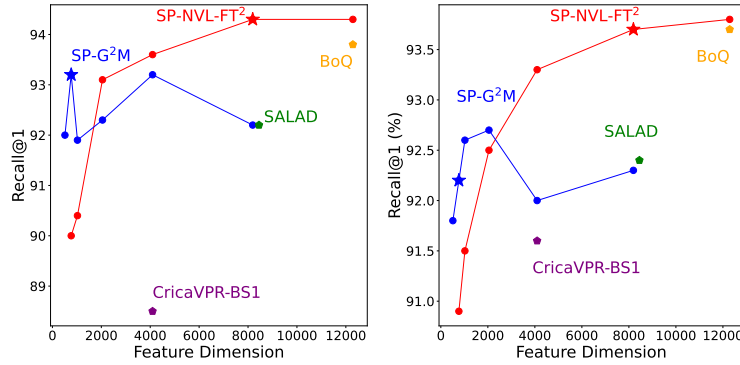


Figure 5: The Recall@1 and descriptor dimensionality comparison of different methods on MSLS-Val (left) and Pitts-30k (right).

We further explored the performance of G²M and NVL-FT² in different dimensions, with other parameters consistent with the SP in Tab. 3. As shown in Fig. 5, NVL-FT² shows a significant performance improvement as the dimension increases, but the growth is relatively weak after exceeding 8000 dimensions. The performance trend of G²M is less consistent, and we recommend maintaining a feature dimension aligned with the number of channels in the extracted feature map.

¹<https://codalab.lis.upsaclay.fr/competitions/865>

A.3 DATASET DETAILS

Pittsburgh-250k (Arandjelovic et al., 2016) is collected from Google Street View and provides 24 images with different viewpoints at each place. The images in this dataset have large viewpoint variations and moderate condition variations.

Tokyo24/7 (Torii et al., 2017) includes 75,984 database images and 315 query images captured from urban scenes. The query images are selected from 1,125 images taken at 125 distinct places with three different viewpoints and at three different times of day. Significant viewpoint and condition changes (e.g., day-night transitions) are present.

Mapillary Street-Level Sequences (MSLS) (Warburg et al., 2020) is a large-scale VPR dataset containing over 1.6 million images labeled with GPS coordinates and compass angles, captured from 30 cities in urban, suburban, and natural scenes over seven years. It covers various challenging visual changes due to illumination, weather, season, viewpoint, and dynamic objects. It includes subsets of training, public validation (MSLS-val), and withheld test (MSLS-challenge).

Nordland (Sünderhauf et al., 2013) primarily consists of suburban and natural place images captured from the same viewpoint in the front of a train across four seasons, which results in severe condition changes (e.g., seasons and lighting) but no viewpoint variations. Its ground truth is provided by the frame-level correspondence. Following previous work (Sünderhauf et al., 2013; Wang et al., 2022), we use the dataset partition first presented in (Sünderhauf et al., 2013) for our experiments.

AmsterTime (Yildiz et al., 2022) is a collection of over one thousand pairs of query-reference images of Amsterdam. For each pair, the query is a grayscale historical image, and its reference is a modern-day photo that represents the same place, as confirmed by human experts. The pairs exhibit multiple domain shifts, including changes in viewpoint, long-term temporal variations, modality differences (RGB vs. grayscale), and different camera systems. Despite its relatively small scale, AmsterTime is one of the most challenging datasets available.

SPED (Chen et al., 2017) comprises low-quality, high-scene-depth images taken from CCTV cameras around the globe. The images in this dataset show various condition variations, such as lighting, weather, and seasonal changes. This dataset covers various outdoor scenes, including forest landscapes, country roads, and urban environments.

SF-XL (Berton et al., 2022a) is a huge dataset covering 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.

SVOX (Berton et al., 2021b) is a cross-domain dataset built from cross-domain VPR that evaluates multiple weather conditions. It spans the city of Oxford, with a large (single domain) database from GSV images: the queries are instead from the Oxford RobotCar dataset (Maddern et al., 2017), providing several weather conditions, such as overcast, rainy, sunny, snowy, and night domains.