

Inside Out Visual Place Recognition Supplementary Material

Sarah Ibrahim
Nanne van Noord
Tim Alpherts
Marcel Worring
{s.ibrahimi,n.j.e.vannoord,t.o.l.alpherts,m.worring}@uva.nl

University of Amsterdam
Science Park 904
Amsterdam
The Netherlands

A Amsterdam-XXXL

A.1 Images Indoor-Ams

Figure A.1 presents images from the several data sources in Amsterdam-XXXL.



Figure A.1: Examples of images from Amsterdam-XXXL. (a) processed street-view images, used for Outdoor-Ams and Ams30k. (b), (c) and (d) are images from the validation set of Indoor-Ams, where (b) has images of Trafficcam (upper two) and Unsplash (bottom two), (c) our selfmade pictures taken from public buildings, and (d) images from Flickr. (e) consists of images from the test set of Indoor-Ams, these are user images from TripAdvisor.

A.2 Data Augmentation on Ams30k

Figure A.2 presents the result of our data augmentation technique. These images are used as queries during training on Ams30k.

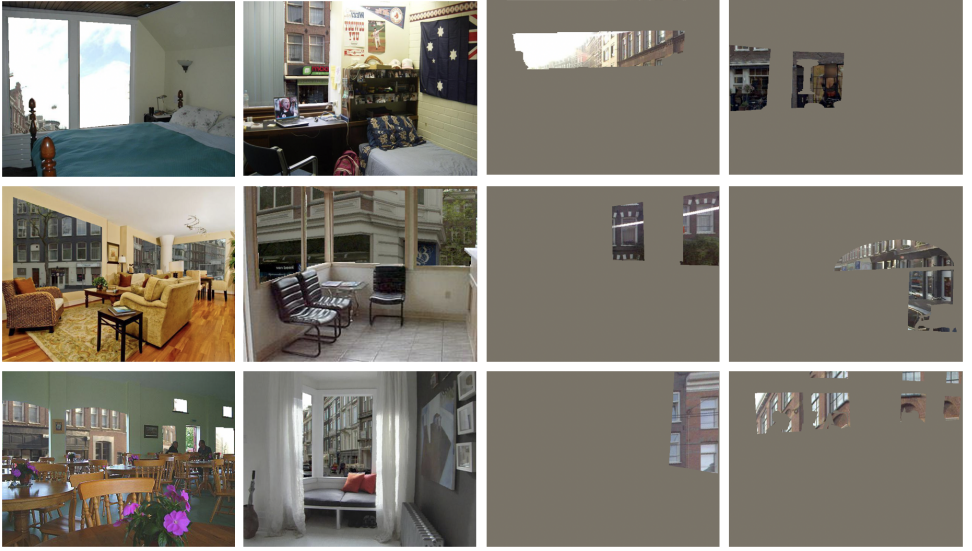


Figure A.2: Examples of query images of Ams30k with real and gray layouts that are used during training

A.3 Outdoor-Ams

For this collection we use images created in 2018, 2019, and 2020, and create the set in the following manner. First, we retrieve the metadata of all images taken in these years from the API of the Municipality of Amsterdam. With the help of shapefiles that indicate the boundaries of regions in Amsterdam¹, we select only the images that are part of central Amsterdam, that is surrounded by a highway. Then we use the DBScan algorithm [4] to select a set of panoramas such that at least every 5 meters of the selected area is covered by a panorama from our set. For this, we use PostGIS, a spatial database extender for PostgreSQL, and more specifically ST_ClusterDBSCAN. This option creates clusters of points with less than 5 meters distance. Note that the quality of the overall coverage depends on where the vehicle has passed for its recordings, but apart from some pedestrian streets in the central area of Amsterdam, this coverage is nearly complete. The panoramas we analyze for this set have resolution 2000×4000 and are processed similarly as was done in [4]. Firstly, the image is processed to exclude irrelevant content, such as the sky, the ground, and the vehicle that collected the images. This is done using an equilateral projection by which the panoramas are projected into six images of 960×960 : front, left, right, back, top, and bottom. As the top and bottom images only capture sky and ground respectively they are excluded. Subsequently the front, back, left, and right images are stitched together to create

¹https://data.lab.fiware.org/dataset/gebiedsindeling_amsterdam

one image of 960×3840 pixels. From this stitched image the bottom 240 pixels are then removed to remove the area that captured the vehicle, resulting in a single panoramic image of 720×3840 that only captures the surroundings of the vehicle which might be used for Visual Place Recognition. To produce the perspective images that make up the Outdoor-Ams set, this resulting image is cut into 24 partially overlapping images of 480×640 for two pitch levels and twelve different yaws. Note that there is an overlap of 240 vertical and 320 horizontal pixels between the perspective images to increase the likelihood of objects being fully visible, and not split by the boundaries between images.

A.4 Indoor-Ams

The validation set of Indoor-Ams is created collecting images with a suitable license from websites such as Flickr. Since the majority of the images does not contain GPS coordinates in the metadata, we focused on searching by keywords. We searched for generic keywords such as 'restaurant', 'cafe', 'hotel', 'coffee', 'beer' combined with 'Amsterdam'. But also more specific keywords, such as names of streets, hotels, restaurants and shops. From the retrieved results, we manually selected the images with windows and added only directly the images to our dataset that had an exact address or a store name. For images with only a neighborhood indication, we manually tried to verify the location with the help of Google Maps and added the image to our dataset whenever we had a match with a nearby panorama image in Outdoor-Ams.

For the test set of Indoor-Ams we used images of TripAdvisor, which were all tagged with a location name. With the help of the Google Maps API, we restricted this set to images in the area from Outdoor-Ams. Consecutively, a window segmentation network, which is explained in Section 5, is used to detect images with at least 5% of their pixels labeled as window. From these remaining images we manually selected 500 images which were most suitable for the task and which had a nearby panorama image in the Outdoor-Ams set. During the creation process of the Indoor-Ams test set, multiple types of images were detected by the window segmentation network that are not suitable for Inside Out Visual Place Recognition. Examples of such images are shown in Figure A.3.



Figure A.3: Examples of images not usable for for Inside Out Visual Place Recognition because they contain incorrectly detected windows, views which are "non-detectable", or outdoor scenes with extreme exposure (i.e., underexposure due to nighttime, or overexposure due to bright sunlight).

A.5 Ams30k

In Table 1, the setup of the Ams30k partition is presented. For each split, the number of unique locations is equal to the number of panoramic street-view images. By processing

these street-view images to 24 images of size 480×640 each, we obtain the total number of images.

| Subset | # Unique locations | # Images |
|---------------|--------------------|----------|
| Train query | 417 | 10008 |
| Train gallery | 466 | 11184 |
| Val query | 366 | 8784 |
| Val gallery | 385 | 9240 |
| Test query | 387 | 9288 |
| Test gallery | 422 | 10128 |

Table 1: Statistics of Ams30k

B Results

Figure B.1 presents the results corresponding to Figure 3(a) and 3(b) in the main paper, but for window ratios of $>10\%$ and $>30\%$.

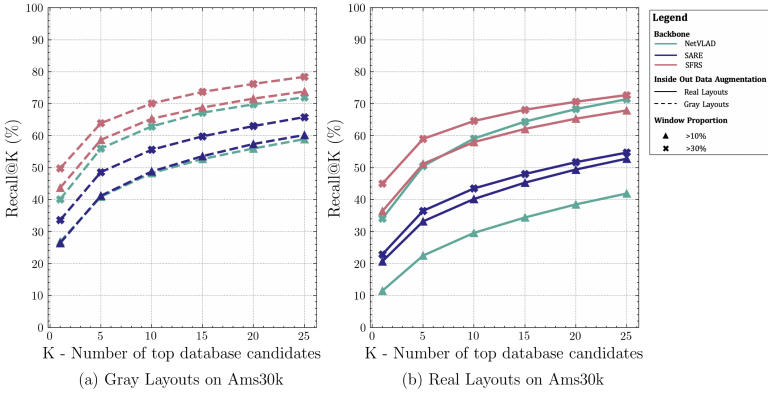


Figure B.1: Results of Inside Out Visual Place Recognition on the Ams30k. Figure (a) and (b) present the results for gray and real layouts on Ams30k for 10% and 30%. This is best viewed in color.

Table 2-5 correspond to Figure 3(a) - 3(d) from the main paper respectively.

| Model | R@1 | R@5 | R@10 | R@15 | R@20 | R@25 |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| NetVLAD >10% | 26.8 | 40.9 | 48.3 | 52.7 | 56.0 | 58.9 |
| NetVLAD >20% | 29.8 | 45.1 | 52.5 | 57.0 | 60.4 | 63.2 |
| NetVLAD >30% | 40.1 | 56.0 | 62.9 | 67.2 | 69.8 | 72.0 |
| SARE >10% | 26.4 | 41.2 | 48.8 | 53.6 | 57.4 | 60.2 |
| SARE >20% | 29.0 | 43.4 | 50.5 | 55.4 | 58.6 | 61.4 |
| SARE >30% | 33.6 | 48.6 | 55.6 | 59.8 | 63.0 | 65.8 |
| SFRS >10% | 43.7 | 58.7 | 65.3 | 68.8 | 71.6 | 73.8 |
| SFRS >20% | 52.0 | 65.9 | 71.4 | 74.6 | 76.9 | 78.8 |
| SFRS >30% | 49.8 | 63.9 | 70.1 | 73.7 | 76.2 | 78.4 |
| NetVLAD - No Augmentation | 25.5 | 38.4 | 44.8 | 49.3 | 52.8 | 55.9 |
| Patch-NetVLAD - No Augmentation | 19.3 | 33.3 | 41.6 | 47.0 | 51.4 | 54.5 |
| Patch-NetVLAD >20% | 25.0 | 41.1 | 49.9 | 55.3 | 59.7 | 63.0 |
| SFRS - No Augmentation | 26.5 | 40.4 | 47.6 | 52.2 | 55.6 | 58.5 |
| Patch-SFRS - No Augmentation | 20.4 | 34.7 | 43.2 | 49.0 | 53.1 | 57.2 |
| Patch-SFRS >20% | 32.3 | 47.5 | 56.7 | 62.6 | 67.2 | 70.9 |

Table 2: Results for Inside Out Data Augmentation with gray layouts on Ams30k, corresponding to Fig 3(a) and B.1(a)

| Model | R@1 | R@5 | R@10 | R@15 | R@20 | R@25 |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| NetVLAD >10% | 11.5 | 22.5 | 29.6 | 34.4 | 38.5 | 41.9 |
| NetVLAD >20% | 32.5 | 48.2 | 55.1 | 59.0 | 62.3 | 65.2 |
| NetVLAD >30% | 34.1 | 50.5 | 59.1 | 64.4 | 68.3 | 71.4 |
| SARE >10% | 20.7 | 33.2 | 40.2 | 45.3 | 49.4 | 52.8 |
| SARE >20% | 25.9 | 40.2 | 47.6 | 52.3 | 55.8 | 58.4 |
| SARE >30% | 22.9 | 36.5 | 43.5 | 48.0 | 51.7 | 54.7 |
| SFRS >10% | 36.4 | 51.2 | 58.0 | 62.1 | 65.3 | 67.9 |
| SFRS >20% | 47.4 | 61.2 | 67.2 | 71.0 | 73.3 | 75.5 |
| SFRS >30% | 45.0 | 59.0 | 64.6 | 68.1 | 70.6 | 72.7 |
| NetVLAD - No Augmentation | 20.4 | 31.7 | 38.4 | 42.7 | 46.3 | 48.9 |
| Patch-NetVLAD - No Augmentation | 14.7 | 26.3 | 33.9 | 39.0 | 43.3 | 47.0 |
| Patch-NetVLAD >20% | 21.0 | 39.7 | 49.9 | 55.6 | 59.9 | 62.9 |
| SFRS - No Augmentation | 19.6 | 30.5 | 36.6 | 41.0 | 44.9 | 48.4 |
| Patch-SFRS - No Augmentation | 13.6 | 24.0 | 31.3 | 36.4 | 40.8 | 44.4 |
| Patch-SFRS >20% | 22.0 | 38.3 | 48.2 | 54.8 | 60.2 | 64.6 |

Table 3: Results for Inside Out Data Augmentation with real layouts on Ams30k, corresponding to Fig 3(b) and B.1(b)

| Model | Size | R@1 | R@5 | R@10 | R@15 | R@20 | R@25 | R@50 | R@75 | R@100 |
|---------------------------------|------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SFRS - No Augmentation | 12k | 4.4 | 11.0 | 16.8 | 19.4 | 22.2 | 25.8 | 36.2 | 41.6 | 49.2 |
| SFRS - Gray Layouts | 12k | 2.2 | 7.6 | 9.8 | 13.2 | 16.0 | 18.4 | 27.6 | 37.0 | 44.6 |
| SFRS - Real Layouts | 12k | 12.2 | 32.0 | 43.0 | 52.8 | 59.0 | 63.4 | 80.0 | 88.4 | 92.4 |
| Patch-SFRS - No Augmentation | 12k | 5.4 | 12.6 | 18.0 | 21.0 | 23.8 | 26.6 | 37.6 | 44.6 | 49.2 |
| Patch-SFRS - Real Layouts | 12k | 7.6 | 19.8 | 30.0 | 37.4 | 44.2 | 50.8 | 71.6 | 84.6 | 92.4 |
| Patch-NetVLAD - No Augmentation | 12k | 5.2 | 12.0 | 17.2 | 20.8 | 23.6 | 27.0 | 37.4 | 44.4 | 49.8 |
| Patch-NetVLAD - Real Layouts | 12k | 5.4 | 10.8 | 16.2 | 20.8 | 24.0 | 26.4 | 40.0 | 46.8 | 49.6 |
| SFRS - No Augmentation | 100k | 1.8 | 4.8 | 6.0 | 7.2 | 8.4 | 10.8 | 16.2 | 19.6 | 24.0 |
| SFRS - Gray Layouts | 100k | 0.8 | 1.4 | 2.4 | 3.4 | 3.8 | 4.8 | 7.0 | 9.4 | 10.8 |
| SFRS - Real Layouts | 100k | 6.0 | 11.2 | 17.4 | 22.0 | 24.0 | 27.6 | 37.6 | 43.2 | 49.2 |
| Patch-SFRS - No Augmentation | 100k | 4.8 | 7.6 | 11.4 | 13.4 | 14.8 | 15.8 | 20.6 | 22.8 | 24.0 |
| Patch-SFRS - Real Layouts | 100k | 4.2 | 12.2 | 17.0 | 22.2 | 25.0 | 28.4 | 38.8 | 44.8 | 49.2 |
| Patch-NetVLAD - No Augmentation | 100k | 2.8 | 6.0 | 8.2 | 10.0 | 12.4 | 14.8 | 20.8 | 23.0 | 25.0 |
| Patch-NetVLAD - Real Layouts | 100k | 3.4 | 5.0 | 5.8 | 7.2 | 9.0 | 10.2 | 14.6 | 16.6 | 17.2 |
| SFRS - No Augmentation | 1M | 1.2 | 2.0 | 2.4 | 2.8 | 3.6 | 4.4 | 7.2 | 8.2 | 9.8 |
| SFRS - Gray Layouts | 1M | 0.2 | 0.4 | 0.6 | 1.4 | 1.8 | 1.8 | 2.0 | 2.4 | 3.2 |
| SFRS - Real Layouts | 1M | 2.8 | 4.6 | 6.2 | 8.2 | 8.8 | 9.2 | 12.8 | 15.4 | 18.4 |
| Patch-SFRS - No Augmentation | 1M | 2.2 | 3.6 | 5.0 | 6.2 | 7.0 | 7.0 | 8.2 | 9.4 | 9.8 |
| Patch-SFRS - Real Layouts | 1M | 2.8 | 5.8 | 7.8 | 8.6 | 9.6 | 10.6 | 15.6 | 17.4 | 18.4 |
| Patch-NetVLAD - No Augmentation | 1M | 2.2 | 3.2 | 4.6 | 5.6 | 6.6 | 7.2 | 9.4 | 11.0 | 12.2 |
| Patch-NetVLAD - Real Layouts | 1M | 2.2 | 3.4 | 3.8 | 4.6 | 5.0 | 5.0 | 7.2 | 8.0 | 8.4 |
| SFRS - No Augmentation | Full | 0.8 | 2.0 | 2.4 | 2.6 | 3.2 | 3.8 | 5.0 | 6.2 | 7.6 |
| SFRS - Gray Layouts | Full | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.8 | 1.0 | 2.0 |
| SFRS - Real Layouts | Full | 1.4 | 2.8 | 3.8 | 4.0 | 4.0 | 4.2 | 5.8 | 7.6 | 8.0 |
| Patch-SFRS - No Augmentation | Full | 3.4 | 3.6 | 4.2 | 4.8 | 5.0 | 5.6 | 6.4 | 7.6 | 7.6 |
| Patch-SFRS - Real Layouts | Full | 2.2 | 3.0 | 3.6 | 4.4 | 4.4 | 4.6 | 6.8 | 7.6 | 8.0 |
| Patch-NetVLAD - No Augmentation | Full | 2.4 | 3.4 | 4.2 | 5.4 | 6.0 | 6.2 | 7.6 | 8.2 | 8.4 |
| Patch-NetVLAD - Real Layouts | Full | 1.8 | 3.2 | 3.8 | 4.0 | 4.2 | 4.4 | 5.2 | 5.4 | 5.4 |

Table 4: Results of Indoor-Ams validation set, evaluated on the full Outdoor-Ams and its subsets of 12k, 100k, and 1M images, corresponding to Fig 3(c)

| Model | Size | R@1 | R@5 | R@10 | R@15 | R@20 | R@25 | R@50 | R@75 | R@100 |
|---------------------------------|------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SFRS - No Augmentation | 12k | 3.4 | 9.0 | 11.8 | 17.2 | 21.6 | 24.2 | 35.0 | 41.8 | 46.0 |
| SFRS - Gray Layouts | 12k | 3.0 | 7.4 | 10.8 | 16.8 | 19.6 | 22.0 | 34.0 | 44.4 | 50.2 |
| SFRS - Real Layouts | 12k | 20.6 | 41.6 | 53.8 | 62.6 | 69.6 | 73.6 | 86.6 | 91.2 | 93.4 |
| Patch-SFRS - No Augmentation | 12k | 5.4 | 12.0 | 16.4 | 21.0 | 24.0 | 26.4 | 34.0 | 42.0 | 46.0 |
| Patch-SFRS - Real Layouts | 12k | 10.8 | 24.6 | 34.2 | 43.8 | 48.20 | 54.0 | 72.6 | 84.4 | 93.4 |
| Patch-NetVLAD - No Augmentation | 12k | 6.2 | 14.2 | 19.8 | 23.8 | 26.6 | 28.8 | 39.6 | 46.0 | 52.0 |
| Patch-NetVLAD - Real Layouts | 12k | 7.6 | 15.4 | 21.2 | 23.8 | 28.4 | 31.8 | 42.8 | 49.6 | 54.8 |
| SFRS - No Augmentation | 100k | 1.8 | 4.4 | 6.0 | 7.6 | 7.8 | 9.0 | 14.8 | 19.2 | 23.2 |
| SFRS - Gray Layouts | 100k | 0.8 | 2.4 | 3.4 | 4.6 | 5.4 | 6.4 | 10.0 | 13.4 | 16.2 |
| SFRS - Real Layouts | 100k | 7.0 | 18.4 | 24.2 | 28.6 | 33.2 | 36.0 | 47.4 | 54.8 | 62.4 |
| Patch-SFRS - No Augmentation | 100k | 4.6 | 7.6 | 10.4 | 11.4 | 13.2 | 14.2 | 18.8 | 22.0 | 23.2 |
| Patch-SFRS - Real Layouts | 100k | 8.4 | 18.0 | 23.0 | 28.8 | 32.4 | 36.4 | 49.8 | 56.6 | 62.4 |
| Patch-NetVLAD - No Augmentation | 100k | 5.2 | 8.4 | 10.8 | 13.0 | 15.0 | 16.6 | 20.4 | 23.2 | 24.8 |
| Patch-NetVLAD - Real Layouts | 100k | 3.4 | 5.8 | 8.2 | 9.0 | 10.2 | 11.8 | 15.4 | 18.0 | 19.8 |
| SFRS - No Augmentation | 1M | 1.0 | 2.4 | 3.0 | 3.8 | 4.8 | 5.2 | 6.8 | 8.0 | 9.4 |
| SFRS - Gray Layouts | 1M | 0.2 | 0.6 | 1.8 | 2.2 | 2.6 | 2.6 | 3.2 | 4.4 | 5.2 |
| SFRS - Real Layouts | 1M | 2.8 | 6.6 | 8.2 | 9.8 | 11.8 | 13.6 | 19.2 | 22.8 | 25.4 |
| Patch-SFRS - No Augmentation | 1M | 2.8 | 4.4 | 5.0 | 5.6 | 5.8 | 6.2 | 7.0 | 8.6 | 9.4 |
| Patch-SFRS - Real Layouts | 1M | 5.6 | 8.8 | 11.4 | 12.6 | 13.6 | 15.6 | 22.0 | 24.4 | 25.4 |
| Patch-NetVLAD - No Augmentation | 1M | 3.2 | 5.0 | 5.6 | 6.2 | 7.2 | 8.0 | 10.6 | 11.8 | 12.4 |
| Patch-NetVLAD - Real Layouts | 1M | 1.0 | 2.8 | 3.8 | 4.4 | 4.8 | 6.0 | 8.4 | 9.2 | 9.6 |
| SFRS - No Augmentation | Full | 1.2 | 2.2 | 3.6 | 4.0 | 4.6 | 4.8 | 5.8 | 8.0 | 9.8 |
| SFRS - Gray Layouts | Full | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.4 | 0.4 | 0.8 | 1.4 |
| SFRS - Real Layouts | Full | 1.4 | 3.0 | 4.2 | 4.8 | 5.4 | 5.6 | 7.0 | 8.6 | 10.4 |
| Patch-SFRS - No Augmentation | Full | 3.2 | 4.4 | 5.4 | 5.6 | 5.8 | 6.2 | 8.0 | 8.6 | 9.8 |
| Patch-SFRS - Real Layouts | Full | 4.2 | 5.2 | 6.4 | 6.8 | 7.0 | 7.4 | 9.8 | 10.2 | 10.4 |
| Patch-NetVLAD - No Augmentation | Full | 2.4 | 3.4 | 4.0 | 4.0 | 4.6 | 4.8 | 6.6 | 8.8 | 9.4 |
| Patch-NetVLAD - Real Layouts | Full | 1.4 | 1.8 | 2.6 | 3.2 | 3.2 | 3.2 | 4.4 | 4.8 | 5.0 |

Table 5: Results of Indoor-Ams test set, evaluated on the full Outdoor-Ams and its subsets of 12k, 100k, and 1M images, corresponding to Fig 3(d)

References

- [1] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016.
- [2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.