# Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition

Frederik Warburg[†,*], Søren Hauberg[†], Manuel López-Antequera[‡], Pau Gargallo[‡], Yubin Kuang[‡], and Javier Civera[§]

[†]Technical University of Denmark, [‡]Mapillary AB, [§]University of Zaragoza

[†]{frwa,sohau}@dtu.dk, [‡]{manuel, pau, yubin}@mapillary.com, [§]jcivera@unizar.es

## Abstract

*Lifelong place recognition is an essential and challenging task in computer vision, with vast applications in robust localization and efficient large-scale 3D reconstruction. Progress is currently hindered by a lack of large, diverse, publicly available datasets. We contribute with Mapillary Street-Level Sequences (MSLS), a large dataset for urban and suburban place recognition from image sequences. It contains more than 1.6 million images curated from the Mapillary collaborative mapping platform. The dataset is orders of magnitude larger than current data sources, and is designed to reflect the diversities of true lifelong learning. It features images from 30 major cities across six continents, hundreds of distinct cameras, and substantially different viewpoints and capture times, spanning all seasons over a nine-year period. All images are geo-located with GPS and compass, and feature high-level attributes such as road type.*

*We propose a set of benchmark tasks designed to push state-of-the-art performance and provide baseline studies. We show that current state-of-the-art methods still have a long way to go, and that the lack of diversity in existing datasets has prevented generalization to new environments. The dataset and benchmarks are available for academic research.*[1]

## 1. Introduction

Visual place recognition is essential for the long-term operation of Augmented Reality and robotic systems [31]. However, despite its relevance and vast research efforts, it remains challenging in practical settings due to the wide array of appearance variations in outdoor scenes, as seen in the examples extracted from our dataset in Figure 1.

Recent research on place recognition has shown that features learned by deep neural networks outperform traditional hand-crafted features, particularly for drastic appearance changes [5, 31, 55]. This has motivated the release of several datasets for training, evaluating and comparing deep learning models. However, such datasets are limited, in at least three aspects. First, none of them covers the many appearance variations encountered in real-world applications. Second, many of them have insufficient size for training large networks. Finally, most datasets are collected in small areas, lacking the geographical diversity needed for generalization.

Figure 1: Mapillary SLS contains imagery from 30 major cities around the world; red stands for training cities and blue for test cities. See four samples from San Francisco, Trondheim, Kampala and Tokyo with challenging appearance changes due to viewpoint, structural, seasonal, dynamic, and illumination.

This paper contributes to the progress of lifelong place recognition by creating a dataset addressing all the challenges described above. We present **Mapillary Street-Level Sequences (MSLS)**, the largest dataset for place recognition to date, with the widest variety of perceptual changes and the broadest geographical spread[2]. MSLS covers the following causes of appearance change: different seasons, changing weather conditions, varying illumination at different times of the day, dynamic

*The main part of this work was done while Frederik Warburg was an intern at Mapillary.

[1] www.mapillary.com/datasets/places

[2] See the video accompanying the paper for an overview and sample images.

| Name | Environment | Total length | Geographical coverage | Temporal coverage | Frames | Type of appearance changes | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Seasonal | Weather | Viewpoint | Dynamic | Day/night | Intrinsics | Structural |
| Nordland [36, 37] | Natural + urban | 728 km | 182 km | 1 year | ~115K | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| SPED [12] | Urban | - | - | 1 year | **~2.5M** | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| KITTI [20] | Urban + suburban | 39.2 km | 1.7 km | 3 days | ~13K | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| Eynsham [14] | Urban + suburban | 70 km | 35 km | 1 day | ~10K | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| St. Lucia [21] | Suburban | 47.5 km | 9.5 km | 1 day | ~33K | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| NCLT [9] | Campus | 148.5 km | 5.5 km | 15 mon. | ~300K | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| Oxford RobotCar [32] | Urban + suburban | 1.000 km | 10 km | 1 year | ~27K | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| VL-CMU [8] | Urban + suburban | 128 km | 8 km | 1 year | ~1.4K | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| FAS [34] | Urban + suburban | 120 km | 70 km | 3 years | ~43K | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| Garden Point [41] | Urban + campus | < 12 km | 4 km | 1 week | ~600 | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| SYNTHIA [44] | Urban | 6 km | 1.5 km | - | ~200K | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| GSV [56] | Urban | - | - | - | ~60K | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Pittsburgh 250k [51] | Urban | - | - | - | ~254K | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| TokyoTM/247 [50] | Urban | - | - | - | ~174K | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| TB-places [28] | Gardens | < 100m | < 100m | 1 year | ~60K | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| **Mapillary SLS (Ours)** | Urban + suburban | **11,560 km** | **4,228 km** | **7 years** | ~1.68M | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: **Summary of place recognition datasets**. Geographical coverage is the length of unique traversed routes. Total length is the geographical coverage multiplied by the number of times each route was traversed. Temporal coverage is the time span from the first recording of a route to the last recording. "–" stands for "not applicable".

objects such as moving pedestrians or cars, structural modifications such as roadworks or architectural work, camera intrinsics and viewpoints. Our data spans six continents, including diverse cities like Kampala, Zurich, Amman and Bangkok.

In addition to the dataset, we make several contributions related to its experimental validation. We benchmark particularly challenging scenarios such as day/night, seasonal and temporal changes. We tackle a wider set of problems not limited to image-to-image localization by proposing six variations of MultiViewNet [16] to model sequence-to-sequence place recognition. Moreover, we formulate two new research tasks: sequence-to-image and image-to-sequence recognition, and propose several feature descriptors that extend pretrained image-to-image models to these two new tasks.

## 2. Related Works

**Place Recognition.** Place recognition consists of finding the most similar place of a query image within a database of registered images [31, 55]. Traditional visual place descriptors are based on aggregating local features using bag-of-words [45], Fischer vectors [39] or VLAD [25]. Other hand-crafted approaches exploit geometric and/or temporal consistency [15, 17, 33] in image sequences. Torii et al. [50] synthesizes viewpoint changes from panorama images with associated depth. These synthetic images make the place descriptor, DenseVLAD [4, 26], more robust to viewpoint and day/night changes.

As in other computer vision tasks, deep features have demonstrated better performance than hand-crafted ones [55]. Initially, features from existing pre-trained networks were used for single-view place recognition [7, 11, 46–48]. Later works demonstrated that the performance improves if the networks are trained for the specific task of place recognition [5, 22, 30]. One of the recent successes is NetVLAD [5, 55], which uses a base network (e.g. VGG16) followed by a generalized VLAD layer (NetVLAD) as an image descriptor. Other works, such as R-MAC [49] and Chen et al. [13], extract regions directly from the CNN response maps to form place descriptors.

Recent deep-learning-based methods exploit the temporal, spatial, and semantic information in images or image sequences. Radenovic et al. [42] proposes a pipeline to obtain large 3D scene reconstructions from unordered images and uses these 3D reconstructions as ground truth for training a Generalized Mean (GeM) layer with hard positive and negative mining. Garg et al. [18], on the other hand, uses single-view depth predictions to recognize places revisited from opposite directions. Also, addressing extreme viewpoint changes, Garg et al. [19] suggests semantically aggregating salient visual information. The 3D geometry of a place is also used by PointNetVLAD [2] that combines PointNet and NetVLAD to form a global place descriptor from LiDAR data. MultiViewNet [16] investigates different pooling strategies, descriptor fusion and LSTMs to model temporal information in image sequences. This research is, however, hindered by the lack of appropriate datasets.

**Place Recognition Datasets.** Table 1 summarizes a set of relevant place recognition datasets. Below we highlight more details and compare our contributions against existing datasets.

**Nordland** [36, 37] contains 4 sequences of a 182km-long train journey, traversed once per season. It captures seasonal changes but contains small variations in viewpoint, camera intrinsics, time of day or structural changes.

**SPED** [12] was curated from images taken by 2.5K static surveillance cameras over 1 year. It contains dynamic, illumination, weather and seasonal changes. However, it does not include viewpoint changes or ego-motion.

**KITTI** [20], **Eynsham** [14] and **St. Lucia** [21] were all recorded by car-mounted cameras. In all three cases the cars drove in urban environments within a few days, capturing dynamic elements and slight viewpoint and weather changes, but no long-term variations. There are several other datasets oriented to autonomous driving collected over longer periods: **NCLT** [9] (recorded over a period of 15 months in a campus environment), **Oxford RobotCar** [32] (recorded from a car traversing the same 10 km route twice every week for a year), **VL-CMU** [8] (composed by $16 \times 8$ km street-view videos captured over one year) and **Freiburg Across Seasons (FAS)** [34] (composed of $2 \times 60$ km summer videos and $1 \times 10$ km winter video over a period of three years). None of these have geographical diversity, nor do they have variations in the camera intrinsics. Moreover, their viewpoint, structural and weather changes are minor.

**Gardens Point** [41] was recorded with a hand-held iPhone. It contains day/night and significant viewpoint changes, but a small representation of other appearance changes and has a small size. **SYNTHIA** [44] contains 4 synthetic image sequences along the same route. It includes varying viewpoints, seasonal, weather, dynamic and day/night changes.

**GSV** [56] compiled a street-level image dataset from Google Street View. However, it is relatively small at 60,000 images. It is limited to a few US cities with no temporal changes and it is composed of still images instead of sequences. **Pittsburgh250k** [51] was also extracted from Google Street View panoramas in Pittsburgh ($10,586$ of them specifically, using two yaw directions and 12 pitch directions). The limited geographical span of these datasets results in a low number of unique places compared to ours. **Tokyo** [50] comes in two versions: The Tokyo Time Machine dataset ($\sim$98K images) and Tokyo 24/7 ($\sim$75K images). Tokyo 24/7 has significant day/night changes. However, [5] comments that the trained model with the Tokyo datasets shows signs of overfitting, probably caused by their limited geographical coverage and size.

Notice that **GSV**, **Pittsburgh250k** and **Tokyo** have significant viewpoint variation but do not include information on viewing direction for the images, and hence positive mining of images with overlaps in view point is not straightforward. In our **Mapillary SLS**, we include viewing direction information for each image (see details in section 3).

Image retrieval is a similar task to place recognition, aiming to find an image in a database that is the most similar to a query image. There exist several image-retrieval datasets (typically created from Flickr images) and established benchmarks, *e.g.*, Holidays [24], Oxford5k, Paris6k [40], Revisited Oxford5k and Paris6k [43], San Francisco Landmarks [10] and Google Landmarks [35, 38]. They usually focus on single-image retrieval and have a very large set of images from the same place, which limits their application in benchmarking lifelong place recognition.

## 3. The Mapillary SLS Dataset

To push the state-of-the-art in lifelong place recognition, there is a need for a larger and more diverse dataset. With this in mind, we have created a new dataset comprised of 1.6 million images from Mapillary[3]. In this section, we present an overview of the curation process, characteristics, and statistics of the dataset. With the available sequential information of the dataset, we additionally propose two new research benchmark tasks.

### 3.1. Data Curation

Our goal is to create a dataset for place recognition with images that (1) have wide geographical reach, reducing bias towards highly populated cities in developed countries (2) are visually diverse, capturing scenarios under varying weather, lighting, and time, and (3) are tagged with reliable geometric and sequential information, enabling new research and practical applications.

#### 3.1.1 Image Selection

**Geographical Diversity.** To ensure geographical diversity, we start with a set of candidate cities for image selection. For each candidate city, we create a regular grid of 500m$^2$ cell size and process each of the cells independently. For each cell, we extract a series of image sequences recorded within this cell. Each sequence contains the image keys and their associated GPS coordinates and raw compass angles (indicating viewing direction).

MSLS contains data from 30 cities spread over 6 continents. See Figure 1 and Table 2 for details. It covers diverse urban and suburban environments, as indicated by the distribution of corresponding OpenStreetMap (OSM)[4] road attributes (Figure 3).

**Unique User and Capture Time.** To ensure variation in the scene structure, time of day, camera intrinsics and view points within each geographical cell, we only keep one sequence per photographer and pick sequences from different days.

**Consistent Viewing Direction.** To ensure that viewing direction measurement is reliable for selecting matching images, we enforce consistency between raw compass angles (measured by the capturing device) and the estimated viewing direction computed with Structure from Motion (SfM)[5]. We select only sequences in which at least 80% of the images' computed angles agree ($\leq 30°$ difference) with the raw compass angle.

#### 3.1.2 Sequence Clustering

To maximize the variety of the dataset, we opt for a larger number of short sequences. The sequence length is curated to

[3] www.mapillary.com

[4] www.openstreetmap.org

[5] The estimated viewing directions were computed based on the relative camera poses estimated using the default OpenSfM [1] pipeline with the camera positions aligned with the GPS measurements

Figure 2: Mapillary SLS pairs showing day/night, weather, seasonal, structural, viewpoint and domain changes.

| Continent | # Frames | # Night Frames | Geo. Coverage [km] | Total Coverage [km] | #Clusters |
|---|---|---|---|---|---|
| Europe | 516 K | 1,098 | 1,052 | 2,985 | 8,654 |
| Asia | 468 K | 9,820 | 965 | 2,729 | 5,483 |
| North America | 431 K | 3,968 | 171 | 4,616 | 6,504 |
| South America | 61 K | 1,177 | 214 | 599 | 1,065 |
| Australia | 200 K | 0 | 259 | 568 | 1,493 |
| Africa | 5 K | 0 | 28 | 63 | 108 |
| **Total** | **1,681 K** | **16,063** | **4,228** | **11,560** | **23,307** |

Table 2: Continental coverage in Mapillary SLS.

match what researchers are currently using for sequence-based place recognition.

Given this initial set of sequences from the image selection process, we generate clusters of sequences that are candidates for place recognition. To avoid sequences where the distance between consecutive images is large, we first split each raw sequence into subsequences if there is more than 30 m between two consecutive frames. Then, we pairwise-match these sub-sequences based on their distance, viewing direction, and motion direction [6]. This is done by searching among all the sub-sequences and forming candidate clusters (sub-sequence pairs) based on their distances to all other neighboring sub-sequences.

To form a candidate cluster, we use the following criteria: Frames from sub-sequences $A$ and $B$ are clustered together if: **1)** Their distance is less than 30 m. **2)** The difference between their viewing directions is less than $40°$. **3)** The difference between their moving directions is less than $40°$. In practice, we use a k-d tree to efficiently discover these pairwise correspondences. The above criteria sometimes skip intermediate images in a sequence, *e.g.*, a sub-sequence might have the images $\{1,2,4,5\}$, thus missing image 3. To avoid this effect, we add all such skipped images back into the sequence.

After matching sub-sequence pairs into potential clusters, we prune them to obtain the frames where both subsequences overlap and hence can be used for sequence-to-sequence place recognition. Since there might be more matching sequences, we merge all pairwise clusters (e.g., we merge clusters A, B and C if there are images that belong to clusters AB, AC and BC.)

We end up with clusters of sequences that have the same geographical coverage and the same moving and viewing direction. The sequences in the clusters are relatively short (5-300 frames), providing a very diverse set of sequential examples for training and development of multi-view place descriptors.

Finally, we filter the resulting clusters enforcing: **1)** that each subsequence has 5 or more frames for proper evaluation of multi-view place recognition models; and **2)** that each cluster has at least two sub-sequences, in order to have a sufficient number of positive training and test samples.

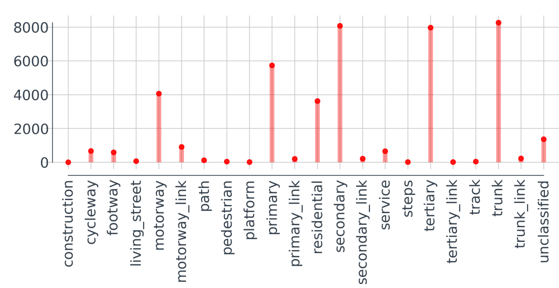



Figure 3: Distribution of OSM road attributes for Mapillary SLS.

### 3.2. Image Attributes

For each image, we additionally provide several raw metadata, post-processed metadata and image attributes that are relevant for further research.

[6]The motion direction for each image is calculated using the GPS measurement and the capture times of consecutive images in a sequence.

**Metadata.** We provide the raw GPS coordinates, capture time (timestamp) and compass angle (which corresponds to the absolute orientation) for each image. We also include UTM coordinates and binary labels indicating the presence or absence of the car control panel (calculated using a semantic segmentation network).

**Day and Night.** We provide an attribute indicating whether a sequence is captured during day or night time. We verified that the day/night attribute could not be robustly estimated from the capture time of the images. Therefore, we implemented a day/night classifier based on the hue distribution of the entire image and of the sky region identified using semantic segmentation. Given the prediction of each image, we then performed a majority voting across the entire sequence to provide consistent day/night tags. To obtain the sky region, we used a semantic segmentation mask provided by Mapillary's API. By manual inspection, we found that such a classifier is sufficient.

**Qualitative View Direction.** We additionally include the facing direction of the camera: forward, backward or sideways, which is the relative orientation of the camera to its movement.

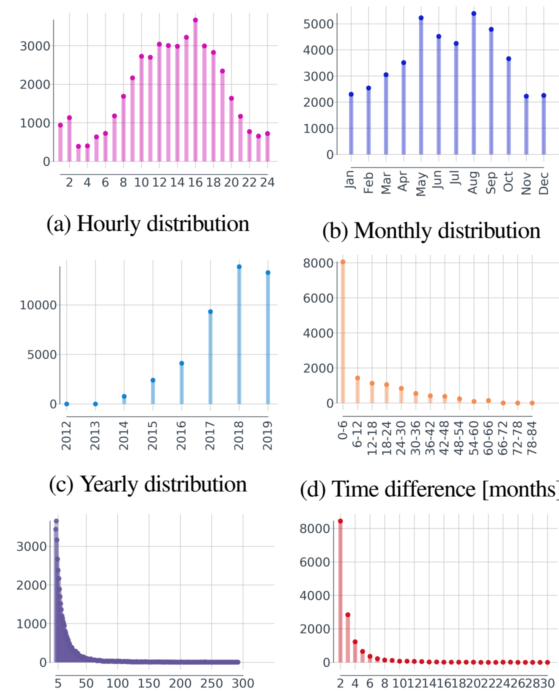
**Road Attributes.** Based on the GPS locations of the images, we also tagged each sequence with road attributes (*e.g.,* residential, motorway, path or others), which were obtained from OpenStreetMap[7] (OSM).

### 3.3. Data Overview

In this section, we provide an overview of the Mapillary SLS dataset in terms of its diversity. In Figures 4a and 4b, we show that the dataset covers all times of the day and months of the year. Figures 4c and 4d show that the dataset spans nine years and that the same places have been revisited with up to seven years time difference, making MSLS the dataset with largest time span for lifelong place recognition. Figures 4e and 4f show large variety in sequence length and number of recordings for the same places.

To highlight the broad variety and challenge, Figure 2 shows image samples from our dataset, where each column contains a query and a database image at a nearby location. In the first column, the query images are taken during the day, whereas the database image is taken at night. The second column shows an example of drastic weather changes as well as a new roadworks traffic sign. The third column shows images from Kampala; a drastic change in environment compared to the images in the first two columns from Copenhagen and San Francisco. Seasonal and structural changes are visible in the two last columns, as the sky-scraper on the left side of the road is under construction in the bottom image and stands finished in the top one. More visual examples of vast variety of changes between query and database images are available in the supplementary material.

[7] https://wiki.openstreetmap.org




(a) Hourly distribution (b) Monthly distribution

(c) Yearly distribution (d) Time difference [months]

(e) # of frames per sequence (f) # of sequences per cluster

Figure 4: Distribution of image sequences in Mapillary SLS on a daily, monthly, yearly scale, time variation, and sequence-related characteristics.

### 3.4. Data Partition and Evaluation

We divide the dataset into a training set (roughly 90%) and a test set (the remaining 10%) containing disjointed sets of cities. Specifically, the test set consists of images collected from Miami, Athens, Buenos Aires, Stockholm, Bengaluru and Kampala. We phrase four place recognition tasks combining single images and sequences in the query and database. These tasks will herefter be referred to as **im2im**, **seq2seq**, **im2seq**, and **seq2im** (**x2y** stands for query **x** and database **y**), respectively.

In addition to evaluating on the whole test set, we suggest the following three research challenges and provide a separate scoreboard for each: **Day/Night** (how well the model recognizes places from day and night and vice versa), **Seasonal** (how well the model recognizes places between seasons, Summer/Winter and vice versa being the most challenging) and **New/Old** (how well the model recognizes places after several years).

Similar to previous works, we cast place recognition as an image retrieval problem and use the top-5 recall as the evaluation metric. For each cluster, we choose one sequence to be the query and the remaining ones to be the database. In the following we will use the **query example** to describe either a query image or a query sequence. The query example is chosen as the center frame(s) in the chosen query sequence. Only one query example is chosen per query sequence, ensuring an equal weight of every place in the evaluation independently of its number of frames. We define the ground-truth matches as those images within a radius of 25m of the query image with a viewing angle difference to it smaller than 40°. A correct sequence match is when any of the frames in the query sequence is less than 25 meters away from any of the frames in the database sequence. This definition also explains why seq2im is harder than im2seq, as the area of correct matching is larger in the latter.

To avoid overfitting to the test set, we withhold the metadata for the test set, except the ordering of the sequences. The test set is divided into query and database sets to ease evaluation. The test set is geographically far from the training set, ensuring that there is no shared visual content, which is a problem for existing datasets (Pittsburgh250k and Tokyo TM/Tokyo 24/7).

# 4. Experiments

In this section, we first present our training procedure for the baseline methods. We show experimental results on the Mapillary SLS dataset in both single-view and multi-view settings.

## 4.1. Training

For the baseline method, we have used NetVLAD [5] and followed a similar training procedure and hyper-parameter selection scheme. The model is trained with the triplet loss [52], for which presenting hard triplets is critical to learn a good embedding. We apply a simple, yet effective sub-caching method with constant time and space use similar to Arandjelovic et al. [6]. Both the query and positive images can be sampled from the cache as well as negatives. It is important to keep the subset size large enough to find adequately hard triplets. In our experiments, we use 10,000 query images and refresh the cache every 1,000 iterations. We use 5 negative examples per triplet instead of 10 [5] as this allows us to fit a batch size of 4 into the memory.

## 4.2. Single-View Place Recognition

In Table 3, we benchmark the most common deep models for **im2im** recognition, reporting their top-5 recall on several challenging recognition cases, as well as their top-1/5/10 recall on the entire test set. These challenging cases include summer to winter (Su/Wi), day to night (Da/Ni), old to new (Ol/Ne) and vice versa. We define old images as those taken between 2011–2016 and new images as those taken since 2018. The goal is to separately evaluate the performance of each method when exposed to seasonal, day/night and structural changes.

We evaluate two early models: Amosnet and Hybridnet [12] and two more recent ones: NetVLAD [5] and GeM (Generalized Mean) [42]. Amosnet and Hybridnet have a Caffe-net backbone followed by two fully connected layers. NetVLAD [5] consists of a VGG16 core with a trainable VLAD layer, and is the state-of-the-art on several place recognition datasets. We evaluated the variant of GeM with VGG16 backbone architecture that is trained on 3d reconstructions of 120k images from Flikr (SfM-120k).

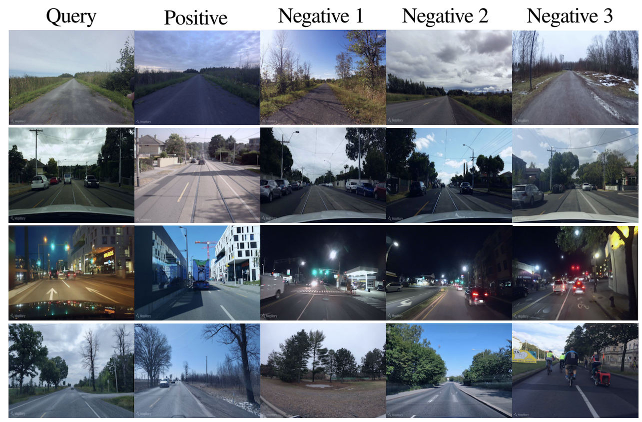



Figure 5: Triplets with multiple negatives. Hard-negatives are mined during training using our proposed sub-caching methodology. For each query-positive pair, the negatives that most violate the triplet constraint, $||q-p||_2^2+m<||q-n||_2^2$, are chosen. Here $q$, $p$, $n$ refer to the cached embeddings for the query, positive and negative images. $m$ is the margin.

Table 3 shows that training on the diverse MSLS improves the overall performance. The performance boost is mainly caused by improved capabilities to recognize places that have undergone seasonal and temporal changes. All models are especially challenged by the night to day changes. Figure 6 shows a detailed comparison of **im2im** models with varying distance and number of image candidates.

## 4.3. Multi-View Place Recognition

We propose to reformulate MultiViewNet [16] to address **seq2seq**, **seq2im**, and **seq2im** place recognition. To the best of our knowledge, no previous work has addressed these two latter cases. We propose two novel architectures based on NetVLAD and show the results in Table 3.

**seq2seq.** We propose six variations of a MultiViewNet [16], specifically, three pooling techniques for NetVLAD and three for GeM. The motivation is to adapt embeddings that are known to work well for single-view place recognition. The first technique, NetVLAD/GeM-MAX, performs max pooling across the embeddings of each image in the sequence. The second variation, NetVLAD/GeM-AVG, does average pooling. The last technique, NetVLAD/GeM-CAT, concatenates the embeddings. Results are reported in Table 3.

**seq2im.** In the sequence-to-image case, we propose to make a majority voting across the sequence, i.e. select the image in the database that most images are nearest to in the query sequence. Given a query sequence of $N$ frames, we calculate the distance from each frame to each database image. We then look at the closest $k$ distances for each of our $N$ frames in our query sequence. This gives a total of $k\times N$ closest database images. We then select the most frequently occurring. The intuition is that if all the frames in a sequence are close to a database

| | Model | Training set | Base | Input Size | Dim | Su/Wi | Wi/Su | Da/Ni | Ni/Da | Ol/Ne | Ne/Ol | All (@1/5/10) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| im2im | Amos | SPED | CaffeNet | 227x227 | 2543 | 0.17 | 0.09 | 0.20 | 0.09 | 0.17 | 0.14 | 0.06/0.11/0.14 |
| | Hybrid | SPED | CaffeNet | 227x227 | 2543 | 0.13 | 0.11 | 0.14 | 0.11 | 0.18 | 0.17 | 0.08/0.13/0.15 |
| | NetVLAD | Pitts250k | VGG16 | 480x640 | 512 | 0.43 | 0.44 | 0.37 | 0.09 | 0.49 | 0.50 | 0.28/0.35/0.39 |
| | GeM | SfM-120k | VGG16 | 480x640 | 2048 | 0.51 | 0.48 | 0.37 | 0.20 | 0.55 | 0.56 | 0.30/0.40/0.44 |
| | NetVLAD | MSLS | VGG16 | 480x640 | 512 | **0.76** | **0.74** | **0.49** | **0.23** | **0.71** | **0.75** | **0.48/0.58/0.64** |
| seq2seq | NetVLAD + MAX | Pitts250k | VGG16 | 480x640 | 512 | 0.40 | 0.51 | 0.37 | 0.09 | 0.55 | 0.57 | 0.23/0.32/0.36 |
| | NetVLAD + AVG | Pitts250k | VGG16 | 480x640 | 512 | 0.41 | 0.39 | 0.37 | 0.09 | 0.54 | 0.54 | 0.20/0.31/0.34 |
| | NetVLAD + CAT | Pitts250k | VGG16 | 480x640 | 512 | 0.44 | 0.47 | 0.37 | 0.14 | 0.57 | 0.56 | 0.23/0.33/0.37 |
| | GeM + MAX | SfM-120k | VGG16 | 480x640 | 2048 | 0.53 | 0.54 | 0.43 | **0.26** | 0.67 | 0.57 | 0.29/0.43/0.48 |
| | GeM + AVG | SfM-120k | VGG16 | 480x640 | 2048 | 0.60 | 0.52 | 0.40 | 0.14 | 0.66 | 0.57 | 0.29/0.42/0.46 |
| | GeM + CAT | SfM-120k | VGG16 | 480x640 | 2048 | 0.55 | 0.46 | 0.46 | **0.26** | 0.65 | 0.53 | 0.28/0.42/0.46 |
| | NetVLAD + MAX | MSLS | VGG16 | 480x640 | 512 | 0.75 | **0.79** | 0.51 | 0.14 | **0.80** | **0.76** | 0.42/0.58/0.63 |
| | NetVLAD + AVG | MSLS | VGG16 | 480x640 | 512 | 0.75 | 0.78 | 0.51 | 0.06 | 0.78 | 0.73 | 0.37/0.56/0.60 |
| | NetVLAD + CAT | MSLS | VGG16 | 480x640 | 512 | **0.84** | 0.76 | **0.57** | 0.20 | **0.80** | 0.72 | **0.41/0.60/0.65** |
| seq2im | NetVLAD + MIN | Pitts250k | VGG16 | 480x640 | 512 | 0.53 | 0.53 | 0.37 | 0.03 | 0.60 | 0.62 | 0.30/0.37/0.40 |
| | NetVLAD + MODE | Pitts250k | VGG16 | 480x640 | 512 | 0.53 | 0.51 | 0.46 | 0.06 | 0.61 | 0.59 | 0.28/0.37/0.41 |
| | GeM + MIN | SfM-120k | VGG16 | 480x640 | 2048 | 0.62 | 0.62 | 0.37 | 0.23 | 0.71 | 0.67 | 0.38/0.47/0.50 |
| | GeM + MODE | SfM-120k | VGG16 | 480x640 | 2048 | 0.59 | 0.52 | 0.46 | **0.26** | 0.67 | 0.66 | 0.32/0.45/0.51 |
| | NetVLAD + MIN | MSLS | VGG16 | 480x640 | 512 | **0.86** | **0.86** | **0.54** | 0.20 | **0.83** | **0.81** | **0.56/0.68/0.71** |
| | NetVLAD + MODE | MSLS | VGG16 | 480x640 | 512 | 0.53 | 0.51 | 0.46 | 0.06 | 0.61 | 0.59 | 0.28/0.37/0.41 |
| im2seq | NetVLAD + MIN | Pitts250k | VGG16 | 480x640 | 512 | 0.20 | 0.30 | 0.29 | 0.14 | 0.33 | 0.28 | 0.12/0.20/0.26 |
| | GeM + MIN | SfM-120k | VGG16 | 480x640 | 2048 | 0.24 | 0.22 | 0.26 | **0.31** | 0.37 | 0.29 | 0.13/0.22/0.31 |
| | NetVLAD + MIN | MSLS | VGG16 | 480x640 | 512 | **0.45** | **0.39** | **0.31** | 0.23 | **0.48** | **0.37** | **0.23/0.34/0.48** |

Table 3: Evaluation of different im2im, seq2seq, seq2im and im2seq models on Mapillary SLS test set. We report the models recall@5 on several challenging recognition cases as well as their overall recall@1/5/10. For a fair comparison, we compare models with similar backbone architecture.

image, then we are more confident that this database image is indeed close to the query sequence. In Table 3, we refer to this majority voting as +MODE. We also test the selection of the closest image in the database among all the images in the query sequence, which we will refer to as +MIN in Table 3. Again, we test these methods using both the VGG16 + GeM and VGG + NetVLAD embeddings (See Table 3).

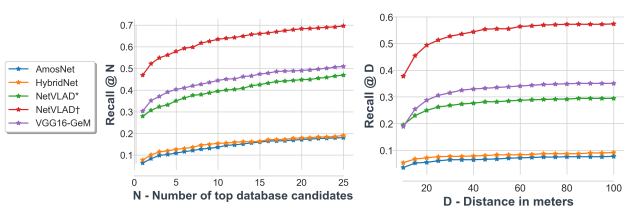



Figure 6: Recall of different methods on MSLS as a function of number of nearest neighbors (left) and distance threshold (right). $*$/† indicates pretrained/trained models, respectively.

**im2seq.** In the image-to-sequence case, we test the selection of the sequence containing the image with the nearest query. In practice, we calculate the distances from the query image to all the frames in the database sequences, and select the sequence that contains the nearest frame.

Table 3 shows that these simple pooling strategies do not improve model performance much compared to single-view models. The reason is that MSLS sequences are captured at different frame-rates and user velocities, requiring the model to learn a time-independent relation between the frames. These complex relations cannot be captured by simple pooling strategies. This motivates the development and further research in multi-view methods, which is accommodated by release of the MSLS.

### 4.4. Further Analysis

In this section, to understand better the strength of the diversity of the dataset, we present qualitative and geographical analyses of the recognition results.

**Qualitative Model Comparison:** In Figure 7, we qualitatively evaluate AmosNet, HybridNet, VGG16-GeM and NetVLAD trained on Pittsburgh250k and SfM120K, and NetVLAD trained on Mapillary SLS. Notice how the diversity of the MSLS data makes NetVLAD more robust to viewpoint and weather changes compared with models that trained on other datasets that do not encapsulate as much diversity as our dataset.

**Geographical Bias:** State-of-the-art place recognition networks are trained on images from developed countries. Figure 8 shows the performance of several models on individual cities from MSLS, confirming their geographical bias. Notice that the GeM model has slightly less bias, the performance drop on Asian cities being relatively low. This could be related to the fact that this model is trained on Flickr images, which are more diverse than other datasets. Figure 9 shows that this geographical bias is reduced by training on MSLS for both AmosNet and NetVLAD.

## 5. Conclusions and Future Work

We have presented Mapillary SLS, a large collection of image sequences for training and evaluating place recognition algo-


Figure 7: Qualitative comparison of different pre-trained networks as well as our NetVLAD model trained on Pittsburgh250k and Mapillary SLS. Training on MSLS improves robustness towards weather changes and diverse vegetation such as palm trees. Green: true positive; Red: false positive.

rithms. The data has been collected from Mapillary and contains over 1.6 million frames from 30 different cities over six continents. The gathered sequences span a period of seven years and the places experience large perceptual changes due to seasons, construction, dynamic objects, cameras, weather and lighting.

MSLS contains the largest geographical and temporal coverage of all publicly available datasets; and it is among the ones with the widest range of appearance variations and the largest number of images. All these features make our dataset a valuable addition to the available data corpus for training place recognition algorithms. The many variation modes and the considerable size of realistic urban data make it particularly appealing for deep-learning approaches and autonomous car applications.

We have also run extensive benchmarks on our dataset with previous state-of-the-art methods to illustrate the difficulty of our dataset. We also introduce two new tasks: seq2im and im2seq. We propose new techniques to solve these tasks using models trained for im2im place recognition and to evaluate several pre-trained models as well as models trained on MSLS.

Although, the focus of the present paper is place recognition, Mapillary SLS is also useful for other computer vision tasks such as pose regression [27, 54], image synthesis (e.g., night-to-day translation [3]), image-to-gps [53, 57]), change detection, feature learning, scene classification using the OSM

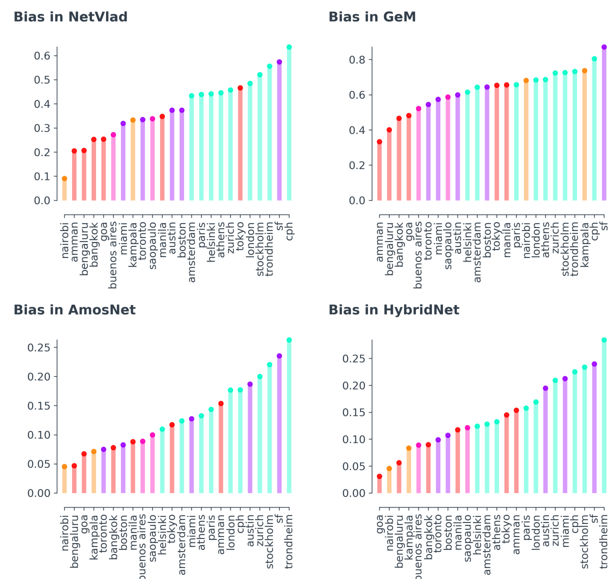
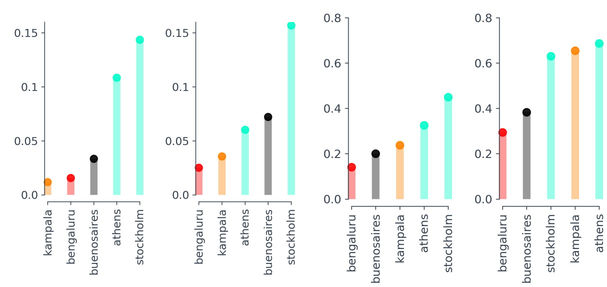



Figure 8: Geographical bias in 4 place recognition models. y-axis shows the top-5 recall. Cities are colored depending on their location: Africa (orange), Asia (red), South America (pink), North America (purple), Europe (turquoise).




(a) AmosNet∗ (b) AmosNet† (c) NetVLAD∗ (d) NetVLAD†

Figure 9: Bias reduction for both AmosNet and NetVLAD when trained on MSLS and evaluated on the MSLS test set. ∗/† indicates pretrained/trained models, respectively.

road tags and unsupervised depth learning [23, 29].

**Acknowledgements.** The majority of this project was carried out as an internship at Mapillary. It received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement nº 757360), from the Spanish Government (PGC2018-096367-B-I00) and the Aragón Government (DGA T45 17R/FSE). SH was supported in part by a research grant (15334) from VILLUM FONDEN. We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPU hardware.

## References


[1] OpenSfM. https://github.com/mapillary/OpenSfM. 3

[2] M. Angelina Uy and G. Hee Lee. PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4470–4479, 2018. 2

[3] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019. 8

[4] R. Arandjelovic and A. Zisserman. All About VLAD. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013. 2

[5] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 1, 2, 3, 6

[6] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 40(06):1437–1451, jun 2018. ISSN 1939-3539. doi: 10.1109/TPAMI.2017.2711011. 6

[7] A. Babenko and V. Lempitsky. Aggregating local deep features for image retrieval. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1269–1277, Dec 2015. doi: 10.1109/ICCV.2015.150. 2

[8] H. Badino, D. Huber, and T. Kanade. Visual topometric localization. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 794–799, June 2011. doi: 10.1109/IVS.2011.5940504. 2, 3

[9] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *International Journal of Robotics Research*, 35(9): 1023–1035, 2015. 2, 3

[10] 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 *CVPR 2011*, pages 737–744, June 2011. doi: 10.1109/CVPR.2011.5995610. 3

[11] Z. Chen, O. Lam, A. Jacobson, and M. Milford. Convolutional neural network-based place recognition. *CoRR*, abs/1411.1509, 2014. URL http://arxiv.org/abs/1411.1509. 2

[12] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. D. Reid, and M. Milford. Deep learning features at scale for visual place recognition. *CoRR*, abs/1701.05105, 2017. URL http://arxiv.org/abs/1701.05105. 2, 3, 6

[13] Z. Chen, F. Maffra, I. Sa, and M. Chli. Only look once, mining distinctive landmarks from ConvNet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9–16, Sep. 2017. doi: 10.1109/IROS.2017.8202131. 2

[14] M. Cummins. Highly scalable appearance-only SLAM - FAB-MAP 2.0. *Proc. Robotics: Sciences and Systems (RSS), 2009*, 2009. 2, 3

[15] M. Cummins and P. Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011. 2

[16] J. M. Fácil, D. Olid, L. Montesano, and J. Civera. Condition-invariant multi-view place recognition. *CoRR*, abs/1902.09516, 2019. URL http://arxiv.org/abs/1902.09516. 2, 6

[17] D. Gálvez-López and J. D. Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012. 2

[18] S. Garg, V. M. Babu, T. Dharmasiri, S. Hausler, N. Sünderhauf, S. Kumar, T. Drummond, and M. Milford. Look no deeper: Recognizing places from opposing viewpoints under varying scene appearance using single-view depth estimation. *CoRR*, abs/1902.07381, 2019. URL http://arxiv.org/abs/1902.07381. 2

[19] S. Garg, N. Sünderhauf, and M. Milford. Semantic–geometric visual place recognition: a new perspective for reconciling opposing views. *The International Journal of Robotics Research*, page 027836491983976, 04 2019. doi: 10.1177/0278364919839761. 2

[20] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2, 3

[21] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth. FAB-MAP+ RatSLAM: Appearance-based SLAM for multiple times of day. In *2010 IEEE international conference on robotics and automation*, pages 3507–3512. IEEE, 2010. 2, 3

[22] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. G. Jiménez. Training a convolutional neural network for appearance-invariant place recognition. *CoRR*, abs/1505.07428, 2015. URL http://arxiv.org/abs/1505.07428. 2

[23] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 8

[24] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-88681-5. doi: 10.1007/978-3-540-88682-2_24. URL http://dx.doi.org/10.1007/978-3-540-88682-2_24. 3

[25] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*, pages 3304–3311. IEEE Computer Society, 2010. 2

[26] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, June 2010. doi: 10.1109/CVPR.2010.5540039. 2

[27] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 8

[28] M. Leyva-Vallina, N. Strisciuglio, M. L. Antequera, R. Tylecek, M. Blaich, and N. Petkov. TB-places: A data set for visual place recognition in garden environments. *IEEE Access*, 7: 52277–52287, 2019. 2

[29] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages

2041–2050, 2018. 8

[30] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez. Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognition Letters*, 92:89–95, 2017. 2

[31] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016. 1, 2

[32] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. doi: 10.1177/0278364916679498. URL http://dx.doi.org/10.1177/0278364916679498. 2, 3

[33] M. J. Milford and G. F. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation*, pages 1643–1649. IEEE, 2012. 2

[34] T. Naseer, W. Burgard, and C. Stachniss. Robust visual localization across seasons. *IEEE Transactions on Robotics*, 34(2):289–302, April 2018. ISSN 1552-3098. doi: 10.1109/TRO.2017.2788045. 2, 3

[35] H. Noh, A. Araujo, J. Sim, and B. Han. Image retrieval with deep local features and attention-based keypoints. *CoRR*, abs/1612.06321, 2016. URL http://arxiv.org/abs/1612.06321. 3

[36] NRK. Nordlandsbanen: minute by minute, season by season, 2013. URL https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/. 2

[37] D. Olid, J. M. Fácil, and J. Civera. Single-view place recognition under seasonal changes. *CoRR*, abs/1808.06516, 2018. URL http://arxiv.org/abs/1808.06516. 2

[38] K. Ozaki and S. Yokoo. Large-scale landmark retrieval/recognition under a noisy and diverse dataset. *CoRR*, abs/1906.04087, 2019. URL http://arxiv.org/abs/1906.04087. 3

[39] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, June 2010. doi: 10.1109/CVPR.2010.5540009. 2

[40] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383172. 3

[41] A. Queensland University of Technology, Brisbane. Day and night with lateral pose change datasets, 2014. URL https://wiki.qut.edu.au/display/cyphy/Day+and+Night+with+Lateral+Pose+Change+Datasets. 2, 3

[42] F. Radenovic, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *CoRR*, abs/1711.02512, 2017. URL http://arxiv.org/abs/1711.02512. 2, 6

[43] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. *CoRR*, abs/1803.11285, 2018. URL http://arxiv.org/abs/1803.11285. 3

[44] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3

[45] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, page 1470. IEEE, 2003. 2

[46] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford. On the performance of convnet features for place recognition. *CoRR*, abs/1501.04158, 2015. URL http://arxiv.org/abs/1501.04158. 2

[47] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of convnet features for place recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4297–4304. IEEE, 2015.

[48] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Proceedings of Robotics: Science and Systems XII*, 2015. 2

[49] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *International Conference on Learning Representations*, 2016. 2

[50] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015. 2, 3

[51] 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, Nov 2015. doi: 10.1109/TPAMI.2015.2409868. 2, 3

[52] D. P. Vassileios Balntas, Edgar Riba and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.119. URL https://dx.doi.org/10.5244/C.30.119. 6

[53] N. Vo, N. Jacobs, and J. Hays. Revisiting IM2GPS in the Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2621–2630, 2017. 8

[54] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization using LSTMs for structured feature correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 627–637, 2017. 8

[55] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. D. McDonald-Maier. Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions. *CoRR*, abs/1903.09107, 2019. URL http://arxiv.org/abs/1903.09107. 1, 2

[56] A. R. Zamir and M. Shah. Image geo-localization based on multiplenearest neighbor feature matching using generalized graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1546–1558, 2014. 2, 3

[57] E. Zemene, Y. T. Tesfaye, H. Idrees, A. Prati, M. Pelillo, and M. Shah. Large-scale image geo-localization using dominant sets. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):148–161, 2018. 8