Geographically Local Representation Learning with a Spatial Prior for Visual Localization

Zimin Xia[1][0000−0002−4981−9514], Olaf Booij[2], Marco Manfredi[2][0000−0002−2618−2493], and Julian F. P. Kooij[1][0000−0001−9919−0710]

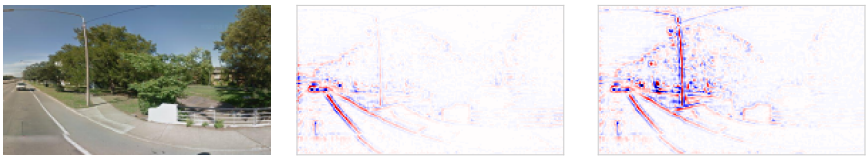
[1] Intelligent Vehicles Group, Technical University Delft, The Netherlands
{z.xia,j.f.p.kooij}@tudelft.nl
[2] TomTom, Amsterdam, The Netherlands
{olaf.booij,marco.manfredi}@tomtom.com

Abstract. We revisit end-to-end representation learning for cross-view self-localization, the task of retrieving for a query camera image the closest satellite image in a database by matching them in a shared image representation space. Previous work tackles this task as a global localization problem, i.e. assuming no prior knowledge on the location, thus the learned image representation must distinguish far apart areas of the map. However, in many practical applications such as self-driving vehicles, it is already possible to discard distant locations through well-known localization techniques using temporal filters and GNSS/GPS sensors. We argue that learned features should therefore be optimized to be discriminative within the geographic local neighborhood, instead of globally. We propose a simple but effective adaptation to the common triplet loss used in previous work to consider a prior localization estimate already in the training phase. We evaluate our approach on the existing CVACT dataset, and on a novel localization benchmark based on the Oxford RobotCar dataset which tests generalization across multiple traversals and days in the same area. For the Oxford benchmarks we collected corresponding satellite images. With a localization prior, our approach improves recall@1 by 9 percent points on CVACT, and reduces the median localization error by 2.45 meters on the Oxford benchmark, compared to a state-of-the-art baseline approach. Qualitative results underscore that with our approach the network indeed captures different aspects of the local surroundings compared to the global baseline.

Keywords: Visual localization, cross-view image matching, image retrieval, end-to-end representation learning

1 Introduction

Self-localization with respect to a known map is an indispensable part for navigation in mobile robotics and autonomous driving. With the rise of camera-equipped vehicles, visual localization provides an attractive approach to absolute positioning. Many visual localization methods construct a descriptor vector



(a) Input image (b) Geo-global features (c) Geo-local features (our)

Fig. 1: We reformulate visual localization in order to exploit coarse priors from GNSS/GPS, and thus end-to-end learn feature extractors specialized to discriminate nearby locations (in the order of meters). This is reflected in the visualized attention maps obtained by our approach (c), compared to (b) standard end-to-end learning [29]. Our features capture the location of nearby objects. The baseline focuses on the road, which is globally distinct, but locally ambiguous.

of a query camera image, and match those to a spatial map of descriptors. For instance, image retrieval-based localization simply represents the query image with a single descriptor vector, and the gallery is constructed from exemplar images with known geographic locations. A variant that has recently drawn a lot of attention is cross-view image retrieval-based localization [10], [16,17], [29], [36], [40], where the gallery is constructed from aerial or satellite images while a ground-level image is used as the query. Although this requires the construction of a shared feature space for both the ground level and satellite images, the satellite view provides a reliable representation of the local surroundings, plus large databases are nowadays readily available. As in other computer vision tasks, the feature extractors and their descriptor embeddings are nowadays often learned end-to-end [1], e.g. through a triplet loss, surpassing earlier hand-crafted methods [2], [11,12], [24].

In the robotics domain, localization is traditionally addressed using specialized sensors that provide noisy measurements of the absolute position in a fixed global coordinate frame directly, e.g. through Global Navigation Satellite Systems (GNSS) such as GPS, and through temporal filtering with odometry information. Unfortunately, the localization accuracy of GNSS can vary significantly near obstructions, buildings, trees and tunnels when less satellites are visible. The horizontal positioning error can easily reach tens of meters [3].

We observe however that there are several gaps in how the localization task is addressed in practice in mobile robotics and autonomous driving, and the state-of-the-art visual localization techniques based on deep representation learning.

First, while GNSS alone is not sufficiently accurate, it does provide a coarse estimate of the absolute position. End-to-end image representation learning approaches for visual localization do not consider the presence of such localization priors during training, and often during testing too, which does not reflect practice. We assert that both approaches should be used together, hence we should learn a feature representation that is locally discriminative within the er-

ror bounds of the coarse prior, rather than globally discriminative on the entirely mapped area, see Figure 1.

Second, existing cross-view localization benchmarks, such as the CVACT dataset [17], split the train, validation, and test splits to different geographic regions of the overall map. This means that the standard split intends to demonstrate how well a learned feature representation generalizes to new areas, as neither satellite nor ground images from the validation or test set are available during training. However, in practice we *can* have satellite images of the test region available during training, especially for a navigation task with geo-localized road information, which already presupposes that the target region is known. Mapping companies may even have already collected ground images of the target region at some past date. An alternative but equally relevant question is therefore how well a learned representation generalizes to new observations of the same route, e.g. on a different time, day or even season.

To address the observed gaps, this paper presents the following contributions: (i) We propose a simple but effective adaptation to the commonly used triplet loss to learn an image representation that is specifically discriminative between images from geographically nearby locations, rather than for distant areas. Note that we include the term geographic to avoid potential confusion with image local features, i.e. which represent a local pixel neighborhood. (ii) To demonstrate the effectiveness of the approach compared to a state-of-the-art baseline, we extend the well-known Oxford RobotCar dataset with a map composed of satellite images to serve as a new dense cross-view localization benchmark to test generalization across recording days. We also test on data from the existing CVACT benchmark, for which we propose new splits. (iii) We report quantitative improvements on image retrieval results and qualitatively show that the proposed geographically local representation focuses on different structures in the environment than the baseline.

2 Related Work

Visual localization methods can be roughly divided into three categories. Camera pose regression [4], [13], [21], [37] uses the weights in the convolutional neural network (CNN) to implicitly describe the map by directly learning the complex function that converts the query image to map coordinates, but are in general not that accurate [28]. Structure-based visual localization [26, 27], [43] relies on extracting local image features from the query image, and matching these to an explicit spatial map of known features. Image retrieval-based localization [1], [10], [16], [25], [29, 30, 31], [36], [40] instead formulates the localization problem as simply matching the query image to gallery images and use the location of the matched gallery image as the location of the query image. Our work falls in this third category.

Image retrieval-based localization methods consist of two key steps. The first step is the image descriptor generation. Changes in illumination, appearance and viewing angle create challenges for this task. A good descriptor should be robust

against those changes and, in the meantime, should be discriminative enough to allow distinguishing different images. The second key step is the similarity measurement. Similar to the metric learning problem, we need a similarity/distance measurement to measure how similar is the query image to the database image. However, during the image retrieval, the query image needs to be compared to every (in extreme case) image in the gallery. So, a complicated similarity measurement is not desired w.r.t. the fast run time requirement.

Instead of learning a complex matching function [9], many recent image retrieval methods [1, 2], [11, 12], [24, 25], [30], [35] aim at the first key step. Those methods usually map the image to Euclidean space and use the L1/L2 distance or dot product as the similarity measurement for distinguishing different images. In the following we describe common approaches to build image/feature descriptors, used in image retrieval methods.

Traditional methods do not require any feature learning process, but usually aggregate hand-crafted local feature descriptors into a global descriptor of the image. For example bag-of-visual-words [25], [30], VLAD [2], [11] or Fisher vector [12], [24]. Recent works employ deep learning to obtain more informative image descriptors. Some works [1], [5], [10], [16, 17, 18], [27], [31], [33], [36], [40] use holistic features to construct the descriptor, and they are often more robust against different illumination and dynamic objects. Other works [6], [14], [22,23], [32], [34], [38], [41] try to let the network learn representative local features. Those learned descriptors are usually more robust against view-point changes comparing to learning on the whole image. PlaNet [39] is an exception since it does not learn a global or local image descriptor for metric learning but treats the image geolocalization problem as a classification problem.

One application of descriptor learning is the ground-to-aerial/satellite image retrieval [10], [16,17], [29], [36], [40]. Due to the superior representative capability of learned features, [40] proposes to reuse pre-existing CNNs for extracting ground-level image features, and then learn to predict such features from aerial images of the same location. They successfully used two separate CNNs to encode features from ground query images and aerial images, and matched them in feature space.

Cross-View Matching Network (CVM-Net) [10] has a two-branch CNN architecture to encode features from ground images and satellite images separately. It incorporates two NetVLADs to transform the features into a common space. The final matching score of input pairs is given by the distance measurement of two NetVLAD descriptors.

In [29] the Spatial-Aware Feature Aggregation (SAFA) network was proposed, for cross-view image-based geo-localization. Notably, [29] introduces a polar transformation pre-processing step, that warps satellite images in order to reduce the domain gap with respect to ground images.

Still, all these representation learning approaches focused on the challenge of learning *globally* discriminative localization features, without considering in the training task that in practice a good localization prior can be obtained from GPS and temporal filtering. Our work addresses this gap.

3 Methodology

In this section, we first shortly describe the cross-view matching and feature learning tasks, then summarize the common triplet loss found in the baseline and related work, and then discuss our proposed changes to the loss.

3.1 Cross-view matching task

In the image matching approach, the objective is to select for a given query image from the vehicle the closest image from the gallery with known geographic coordinates. We here consider the cross-view matching problem, where the query G_q is a ground-level (possibly panoramic) camera image from the ego-vehicle, and the target dataset $\mathbf{S} = (S_1, S_2, \cdots)$ contains top-down satellite images of the mapped environment. Each satellite image here shows a fixed-sized square area of the Earth's surface with a fixed image resolution. We assume that the 2D geographic position of the center of an image S is known, and given by $\pi(S) \in \mathcal{R}^2$ in meters in the map's coordinate system.

Matching is done by using two learned functions $f(\cdot)$ and $g(\cdot)$ to respectively map the targets and query to an n-dimensional feature space, where the correct target is expected to have the shortest Euclidean distance to the query. In other words, to localize a given ground-based vehicle image G_q, the location $\widehat{p}$ of the best matched target $\widehat{S}$ is returned,

$$\widehat{S} = \text{argmin}_{S \in \mathbf{S}} \, ||f(S) - g(G_q)||_2, \qquad \widehat{p} = \pi(\widehat{S}). \tag{1}$$

In practice, $f(\cdot)$ and $g(\cdot)$ are implemented as deep convolutional neural networks, and trained on training data with known pairs $\mathcal{X} = \{(S_1, G_1), (S_2, G_2), \cdots\}$. This task is typically addressed using the triplet loss.

3.2 Baseline global triplet loss

We define $d_{i,j} = ||f(S_i) - g(G_j)||_2$ as the Euclidean distance between satellite image S_i and ground image G_j in the embedding space. Ideally, the learned embedding minimizes the *positive* distance term $d_{i,i}$ between a correctly matched satellite (S_i) and ground (G_i) image pair. Meanwhile, it should maximize any *negative* distance term $d_{i,j}$ between a mismatched pair, i.e. where $i \neq j$. This objective is captured by the weighted soft-margin triplet loss, of which we can formulate two versions,

$$l_1(i,j) = \log(1 + e^{\gamma(d_{i,i} - d_{i,j})}), \qquad \text{(satellite-to-ground)} \tag{2}$$

$$l_2(i,j) = \log(1 + e^{\gamma(d_{i,i} - d_{j,i})}), \qquad \text{(ground-to-satellite)} \tag{3}$$

where γ is a scalar parameter to adjust the gradient of the loss. The two versions differ in whether we select a mismatched ground image, Eq. (2), or satellite image, Eq. (3), to form the negative term.

For a minibatch $\mathcal{B} = (P_1, P_2, ..., P_N) \subseteq \mathcal{X}$ of N pairs, the baseline implementation [29] computes the final loss as

$$\mathcal{L}(\mathcal{B}) = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1 \bigwedge j \neq i}^{N} l_1(i,j) + l_2(i,j). \tag{4}$$

These loss terms can be efficiently computed by performing the forward passes $f(S_i)$ and $g(G_i)$ only once for all samples, and then just computing N^2 Euclidean distances $d_{i,j}$ of all combinations i, j.

An important aspect of the baseline is that it selects minibatches from the training data by randomly shuffling *all* samples in each epoch, thus any two pairs are equally likely to co-occur in the batches, independently if their actual geographic coordinates are close together or far away. This triplet loss thus learns a *globally* discriminative representation.

3.3 Proposed local triplet loss

In many outdoor localization applications, GNSS or temporal filtering can already provide a good estimate of the approximate location. We will assume that the worst-case error in this coarse prior describes a geospatial circle with max. radius of r meters, and more distant locations can be discard a-priori. This leads us to propose two effective but simple to implement adaptations to the original loss, namely *geo-distance weighted loss terms* and *local minibatches*.

Geo-distance weighted loss terms We add a weighting term $w(i,j)$ to the triplet losses that adapts their contribution based on the Euclidean distance in meters between the two geographic positions $\pi(S_i)$ and $\pi(S_j)$,

$$l_1(i,j) = w(i,j) \cdot \log(1 + e^{\gamma(d_{i,i} - d_{i,j})}), \qquad \text{(satellite to ground)} \tag{5}$$

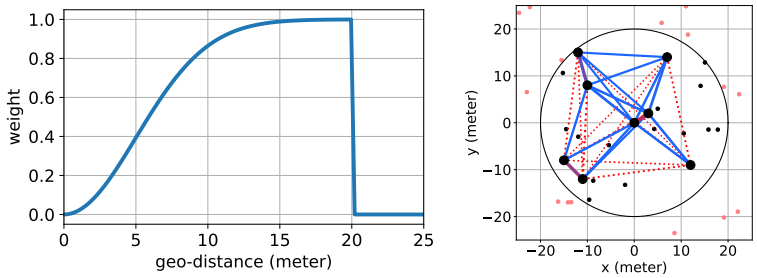
$$l_2(i,j) = w(i,j) \cdot \log(1 + e^{\gamma(d_{i,i} - d_{j,i})}). \qquad \text{(ground to satellite)} \tag{6}$$

We define the weighting term using hyperparameters r and σ in meters,

$$w(i,j) = \begin{cases} 0 & \text{iff} \quad ||\pi(S_i) - \pi(S_j)||_2 > r \\ 1 - e^{-||\pi(S_i) - \pi(S_j)||_2^2 / (2\sigma^2)} & \text{otherwise.} \end{cases} \tag{7}$$

The weighting term considers two cases. First, it cancels any triplet term between pairs that are further away than the maximally assumed prior localization error, given by the maximum distance r in meters. Second, if the pairs are within the acceptable distance, a positive weight should be assigned, though we smoothly reduce the weight to zero if the samples are too close together. The smoothness of this reduction is controlled by σ, see Figure 2a for an example.

We find that down-weighting the loss on geographically nearby samples is crucial to learn a good representation in densely populated data sets. For instance, consider pairs i and j with 1 meter interval then without down-weighting the optimization requires both minimizing the embedding distance of positive match S_i and G_i, while maximizing the embedding distance of the almost identical satellite image S_j and ground image G_i leading to severe overfitting.



(a) Weight function (b) Local minibatch creation

Fig. 2: (a) Weight function of Equation (7), here shown as a function of geo-distance $||\pi(S_i) - \pi(S_j)||_2$ with $r = 20m$ and $\sigma = 5m$. (b) Example of generating a local minibatch of size 8. The black/red dots indicate geographic locations in/outside $\mathcal{N}_r$ of the first picked sample (dot in center). Thick dots mark samples picked for the batch. Edges between picked dots indicate the weight w: red for low weight, blue for high weight, and dashed edges for distances larger than r.

Local minibatches Using the geo-distance weighted loss term, most randomly picked pairs from the training data would have zero weight as they are likely to be at distant geographic locations, especially when the mapped area is large. We therefore construct local minibatches that only contain pairs from nearby geographic locations, maximizing the impact of each sample per epoch:

1. pre-compute before training for each pair $P_i = (S_i, G_i)$ the local neighborhood of pairs within geographic radius of r meters, i.e.
$$\mathcal{N}_r(i) = \{(S_j, G_j) \mid i \neq j \wedge ||\pi(S_i) - \pi(S_j)||_2 \leq r\} \subset \mathcal{X}. \tag{8}$$
2. At the start of an epoch, create a fresh set $\tilde{\mathcal{X}}$ containing all training samples, $\tilde{\mathcal{X}} \leftarrow \mathcal{X}$, representing the still unused samples in this epoch.
3. To create a new minibatch $\mathcal{B}$ of size N, first randomly pick a pair P_i from pool $\tilde{\mathcal{X}}$, and then uniformly pick without replacement the remaining $N-1$ samples from the neighborhood set $\mathcal{N}_r(i)$. All picked samples are removed from the epoch's pool, $\tilde{\mathcal{X}} \leftarrow \tilde{\mathcal{X}}/\mathcal{B}$. Once $\tilde{\mathcal{X}}$ is empty, a new epoch is started.

Note that overall each pair occurs in *at most* one minibatch per epoch, and pairs without enough neighbors will not be used. Since all pairs j in the batch are per definition within distance r from the first sampled pair i, two samples j and j' in the minibatch can be *at most* a distance of $2r$ meters geographically apart. Our local minibatch formulation thus maximizes the chance that many pairs in the minibatch are also within each other's r-meter radius, and thus minimizes the chance of near-zero geo-distance weighted loss terms, see Figure 2b for an example. Contrast this to the standard minibatches, where the maximum distance is bounded by the geographic size of the mapped area, which is potentially several orders of magnitude larger than $2r$ meters.

4 Experiments

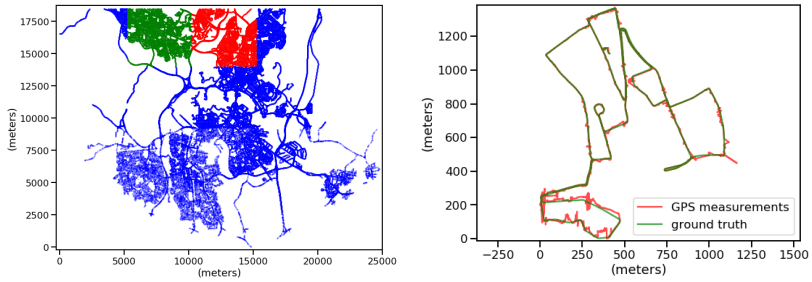
We perform various cross-view matching experiments to compare our geo-distance weighted loss to the standard loss used in the SAFA baseline method [29]. Using two datasets, we explore generalization to new areas, generalization to new traversals (e.g. on different days), and provide qualitative results to demonstrate to what image properties the attention maps in our trained model responds to, and how our approach affects localization uncertainty.

4.1 Datasets

We will first review our two adapted and novel localization benchmarks.

CVACT Dataset: CVACT [17] is a large cross-view dataset with GPS footprint for image retrieval. It contains 35532 ground panorama-and-satellite image pairs, denoted as CVACT_train, and 92802 pairs as CVACT_test. Notably, the validation set CVACT_val of 8884 pairs is a subset of CVACT_test, and [29] reported their quantitative results on the CVACT_val rather than CVACT_test. We will not directly follow the data split in [17], [29] since CVACT_val is rather sparse and distributed over too large an area, which we found trivialized localization with a prior too much as it discarded all negative samples. Furthermore, we only split the ground images into training, validation and test set, and follow the target use-case where all satellite images are available during training.

The overview of our data split is shown in Figure 3a. In total, there are 128334 satellite images and the number of ground images is 86469, 21249 and 20616 in training, validation and test set respectively. The data is relatively sparse, using a localization prior of $r = 100m$ most samples having between 25 and 100 other pairs in their local neighborhood.



(a) CVACT (b) Oxford RobotCar

Fig. 3: (a) Our used CVACT train (blue) / validation (green) / test (red) data splits. (b) One traversal from Oxford RobotCar, with raw GPS (red) and ground truth RTK (green). Raw GPS can have large errors over extended periods.

Oxford RobotCar Dataset: Oxford RobotCar [20], [19] is a dataset targeted at autonomous driving which contains images, LiDAR measurements and GPS recordings under different lighting and weather conditions collected over a year over multiple traversals in the Oxford region. The ground truth location is acquired via GPS-RTK. We note that the recordings reveal the limitations of raw GNSS/GPS and the necessity of our research. As shown from a sample traversal in Figure 3b, the raw GPS error can reach 50 meters. This highlights the practical application of our proposed approach as a refining step on the inaccurate GNSS/GPS measurements.

The dataset has not been used for cross-view image matching-based localization, as it does not contain satellite/aerial images. To construct a novel benchmark we collected 600×600 pixel satellite images at zoom level 20 ($\sim 0.0924m$ per pixel) from Google Maps Static API [1] for each ground front-viewing image. For now, we do not target the most extreme lighting and weather condition and select the traversals recorded in day time with label "sun" or "overcast" and which contain both raw GPS and accurate RTK localization measurements. In the dataset the front-viewing images are taken at 16Hz. To make sure the consecutive ground images do not look too similar in appearance, we sample the images to make sure there is at least $5m$ between two consecutive frames in each traversal. Finally, we acquire the corresponding satellite images centered at the ground truth locations to formulate the ground-to-satellite pairs. In total we acquire 23554 pairs from 13 traversals. We always keep all the satellite images, and use the ground image from 11 traversals as the training set (19707), 1 traversal as the validation set (1953), and 1 traversal as the test set (1894). In this dense dataset, almost all images have more than 200 pairs in a $r = 50m$ neighborhood. Some example ground and satellite pairs are shown in Figure 4.

4.2 Network architecture and implementation details

In our experiments we apply our new loss to the baseline SAFA method for cross-view matching of ground images to a map of satellite images [29]. We here shortly discuss pre-processing, and the neural network architectures for the functions $g(\cdot)$ and $f(\cdot)$ from Equation (1). [2]

First, when the ground images are 360° panoramic views, as is the case for the CVACT dataset, [29] proposed to use a polar image transformation as a pre-processing step on the satellite images, as this can make the two image domains more similar, and simplifies the learning task for the network somewhat. In our experiments on CVACT, we will use these pre-processed images too.

Second, the networks $f(\cdot)$ and $g(\cdot)$ are both structured the same. Each network starts with the first 16 layers of a VGG network as feature extractor, and the extracted features are then input into the 8 separate spatial-aware position embedding modules [29], the results of which are concatenated resulting in a

[1] `https://developers.google.com/maps/documentation/maps-static/dev-guide`

[2] Code of our implementation is available at `https://github.com/tudelft-iv/Visual-Localization-with-Spatial-Prior`

(a) (b) (c) (d)

Fig. 4: Four sample pairs in the proposed Oxford RobotCar cross-view localization benchmark to highlight some local and global differences. (a) and (b) are 5 meters apart, (c) and (d) are 20 meters apart. Ground images are from different traversals and recording days in the original dataset, resulting in variations in cars and lighting conditions. Note that the presence of a white road marking would be informative to globally discriminate between locations (a) and (c), but not to locally discriminate between (a) and (b), nor between (c) and (d).

4096-length descriptor in the shared space of $f(\cdot)$ and $g(\cdot)$. During training, both networks are optimized jointly without weight sharing.

We trained the baseline model on our proposed data split using the code released by the author of [29]. For our method, we do not change the baseline architecture but directly replace the loss with our proposed geo-distance weighted loss. Similar to [29], the VGG model is pre-trained on Imagenet [7]. For the triplet loss, γ is set to 10. Both models are trained with Adam optimizer [15]. On the CVACT dataset, we use a batch size of 16 for the baseline. Since some images do not have more than 15 neighbors, we use batch size of 4 for our model, and the learning rate is set to 10^{-5}. On the Oxford RobotCar dataset, the batch size is set to 16. A learning rate of 5×10^{-5} works well for our model, but we find that a learning rate of 10^{-5} works better for the baseline. Due to the dense geospatial distribution of this dataset, many satellite images are very similar. We employ two strategies to combat overfitting. First, we use dropblock [8] with block size of 11 and keep probability of 0.8 for our method. We also tested this on the baseline but did not find that it improved its results. Second, we perform data augmentation by selecting for a query ground image a random satellite image at a small geospatial offsets of maximally 5 meter radius for additional robustness.

4.3 Evaluation metrics

For our main task we assume at test time a known (worst-case) prior localization error of radius r, and thus directly discard for both methods any false negatives

beyond r meters of the true location. Still, for reference we also review the case when no such prior would be available (i.e. an *infinite* test radius).

The recall@1 is our main quantitative metric. The reported percentage indicates how often the the top-1 retrieved satellite image exactly corresponds to the test query location. On the dense Oxford RobotCar dataset, we also report recall@x-meters, where any satellite image within that radius is counted as correct since these are nearly identical. Our maximal acceptable offset is $x = 5m$, the same distance used to select the camera frames (see Sec. 4.1).

4.4 Experiment on CVACT Dataset

For the CVACT datases, we use $r = 100$ meters as the localization prior for training our model, and testing. Both the baseline and our model are trained for 100 epochs, and we keep the best model for both according to validation split performance. Results are reported on the test split.

The test split is from a region not seen during training, hence the recall@1 is indicative how well the learned feature representation also generalizes to ground images in new areas. With a test radius of 100m, recall@1 for our method is 74.0%, and for the baseline approach 65.0%, which demonstrates that our representation indeed exploits the availability of a localization prior. For reference, with an infinite test radius (no prior), our recall@1 is 54.5% compared to 58.4% for the baseline. As expected, in this case our model perform indeed somewhat worse than the globally trained baseline. Still, in real world applications where a prior is feasible, this suggests our model outperforms the baseline by 9% points.

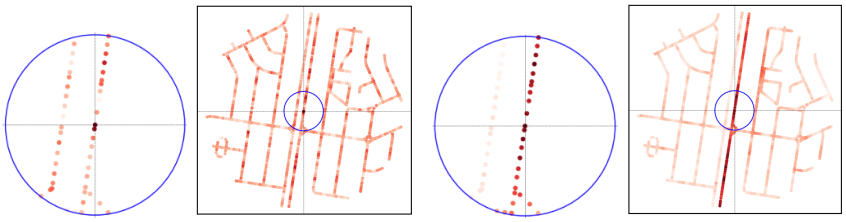
To provide a more intuitive view of the difference of the behaviour of the baseline and our model, we visualized the location heat map of a given query using the similarity score provided by the models during inference. As shown in Figure 5, our model is less certain outside the prior area, but it is capable to localize the image along a road, where the baseline shows high uncertainty.

The advantage of our model comes from the geographically local representation it used. We verified this by comparing the encoded features from the baseline and our model. Similar to [29], we follow [42] to back-propagate the spatial embedding maps to the input image to show where the model extracts features from. As shown in Figure 6, our model pays attention at poles and streetlights. Such objects are repeated at many different places but they are quite useful in distinguishing other images along this road. The baseline model, on the other hand, ignores these objects and pays more attention on the road structure, which is more useful in finding out the global location.

4.5 Experiment on Oxford RobotCar Dataset

On the Oxford RobotCar dataset, the baseline and our model are trained for 200 epochs. Since the images are distributed much denser here, we can use a more realistic hypothetical GPS prior with the location uncertainty at $r = 50$ meters.

Table 1 summarizes the image matching results. Our model surpasses the baseline by a large margin when tested with location prior of 50 meters. The



(a) Geo-local representation (ours) (b) Geo-global representation (baseline)

Fig. 5: Example of localization heat maps on CVACT dataset. Each dot represents a satellite image and the ground truth location is indicated by the cross in the heat map. Darker colors indicate smaller embedding distance between the satellite images at those locations and the ground query taken at the center location. The circle indicates the local neighborhood with 100m radius, the boxed image the surrounding $1km^2$ area. Within the local neighborhood, our approach results in a single peak, while the baseline distribution is more spread.

satellite images are densely distributed along the roads, and there are on average 26 satellite images within a 5 meter neighborhood. Therefore our model can successfully locate over 99% of query ground images in the test split.

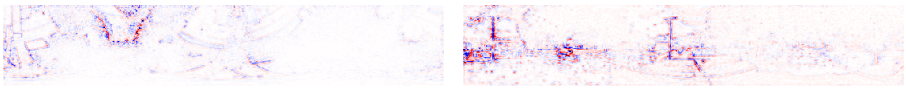
Surprisingly, our model also shows better result in recall@1,1m,2m,3m than the baseline without location prior. A possible reason is that images outside the prior area do not share common local features with the ground truth satellite images. Consequently, our model gains global localization ability with those prominent features. Besides, as seen in Figure 7c, the localization uncertainty of the baseline approach barely benefits from discarded negatives outside the localization prior. This validates our original hypothesis that exploiting available localization priors during training directly improves the utility of the learnt representations.

Test Radius	50m						infinite (no prior)					
Recall@	1	1m	2m	3m	4m	5m	1	1m	2m	3m	4m	5m
Our method (%)	**8.9**	**54.7**	**78.0**	**89.3**	**95.1**	**99.2**	**5.4**	**34.1**	**46.6**	**52.4**	55.6	57.7
Baseline [29](%)	2.4	22.3	35.9	47.3	56.2	70.2	2.4	22.3	35.9	47.3	**56.2**	**70.2**

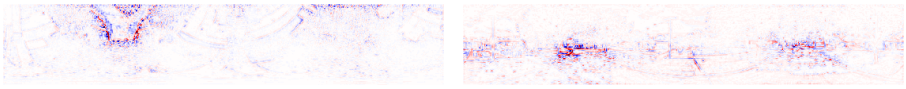
Table 1: Recall comparison on Oxford RobotCar (best results in bold).

Many image retrieval-based localization methods do not report a metric evaluation of their localization capability due to the sparsity of the datasets. We report the distance error of geolocation of the top-1 retrieved satellite image from our model and the baseline on the Oxford RobotCar dataset in Table 2. With 50 meter location prior, our model achieved a median localization error of 0.86 meters on the test split, which is 2.45 meters lower than the baseline.

(a) Input satellite and ground image



(b) Geo-local attention maps (ours)



(c) Geo-global attention maps (baseline)

Fig. 6: Visualized Back-propagated Encoded Feature [42] attention maps for two CVACT inputs. Saturated red/blue areas indicate strong positive/negative activation in the maps, bright areas indicate low absolute activation.

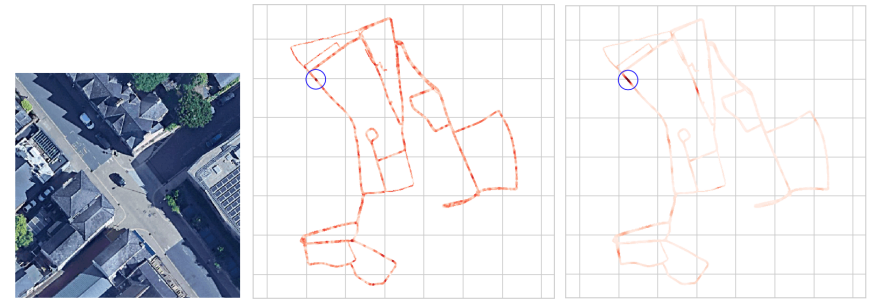
Top1 distance	median	80%quantile	90%quantile	95%quantile	mean
Our method (meter)	**0.86**	**2.13**	**3.07**	**3.97**	**1.27**
Baseline [29] (meter)	3.31	5.44	6.74	9.63	3.62

Table 2: Geo-distance error of top-1 result on Oxford RobotCar (best in bold).

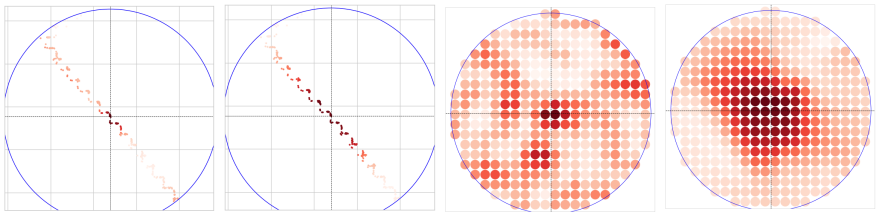
To provide more intuition about how the baseline and our models work on this novel denser dataset, we visualized the localization heat map on the regular grid of satellite images in Figure 7. In Figure 7, the cross indicates the center of the circle, which is also the ground truth location. Notice that the ground truth location is always not on the grid point. The color means the probability of the query image located at that grid point. The darker the color the higher the probability. In most of cases, the baseline is quite discriminative globally, but has local uncertainty in around 20 meters by 20 meters area. For our model, although it has no global localization ability, it is more accurate in local area.

5 Conclusions

Our experiments show that there is a clear quantitative and qualitative difference between learned image representations that must distinguish between either only geographically nearby locations, i.e. a 'geo-local' representation, or that must also distinguish between geographically distant locations, a 'geo-global' representation. While previous work only focused on learning geo-global representations, we have shown that a geo-local representation can already be obtained with easy to implement adjustments to the triplet loss. We find an improvement of 2.45 meters and 2.35 meters in terms of median and mean localization accuracy given



(a) Satellite image (b) Ours (full map) (c) Baseline (full map)



(d) Ours (local) (e) Baseline (local) (f) Ours (local grid) (g) Baseline (l. grid)

Fig. 7: Example localization heat maps comparing our method to the baseline in the Oxford RobotCar benchmark, the circle marks the $r = 50$ meter test radius around the query. In the heat maps, darker colors indicate smaller embedding distance between the the satellite images at those locations and the ground query. (a) Satellite image of the query. (b), (c) Response to the full map. Our approach also shows matches outside the test radius, since those are ignored during training and testing. The baseline matches the same region as the coarse prior, adding little more information. (d), (e) Within the test radius, our method has less uncertainty. (f), (g) Matching satellite images at regular grid locations reveals the structure of the learned embedding in more detail.

a weak localization prior during inference. Our qualitative visualizations show that the proposed modifications result in different attention patterns. In particular, our method focuses on surrounding trees and lamp posts, which would be at distinct positions when moving only a few meters away. The baseline global approach instead focuses on the road layout that distinguishes distant map regions, but is less discriminative for nearby locations. The proposed geographic localized triplet loss is general, and in future work we will investigate how it affects other learned map representations.

Acknowledgements. This work is part of the research programme Efficient Deep Learning (EDL) with project number P16-25, which is (partly) financed by the Dutch Research Council (NWO).

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
2. Arandjelovic, R., Zisserman, A.: All about VLAD. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1578–1585 (2013)
3. Ben-Moshe, B., Elkin, E., Levi, H., Weissman, A.: Improving accuracy of GNSS devices in urban canyons. In: Proceedings of the Canadian Conference on Computational Geometry (2011)
4. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Geometry-aware learning of maps for camera localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2616–2625 (2018)
5. Chen, Z., Jacobson, A., Sünderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I., Milford, M.: Deep learning features at scale for visual place recognition. In: IEEE International Conference on Robotics and Automation. pp. 3223–3230 (2017)
6. Chen, Z., Maffra, F., Sa, I., Chli, M.: Only look once, mining distinctive landmarks from convnet for visual place recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 9–16 (2017)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
8. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. In: Advances in Neural Information Processing Systems. pp. 10727–10737 (2018)
9. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3279–3286 (2015)
10. Hu, S., Feng, M., Nguyen, R.M., Hee Lee, G.: CVM-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7258–7267 (2018)
11. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3304–3311 (2010)
12. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(9), 1704–1716 (2011)
13. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2938–2946 (2015)
14. Kim, H.J., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3251–3260 (2017)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2014)
16. Lin, T.Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5007–5015 (2015)

17. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5624–5633 (2019)
18. Lopez-Antequera, M., Gomez-Ojeda, R., Petkov, N., Gonzalez-Jimenez, J.: Appearance-invariant place recognition by discriminatively training a convolutional neural network. Pattern Recognition Letters **92**, 89–95 (2017)
19. Maddern, W., Pascoe, G., Gadd, M., Barnes, D., Yeomans, B., Newman, P.: Real-time kinematic ground truth for the oxford robotcar dataset. arXiv preprint arXiv: 2002.10152 (2020)
20. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
21. Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.: Image-based localization using hourglass networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 879–886 (2017)
22. Naseer, T., Oliveira, G.L., Brox, T., Burgard, W.: Semantics-aware visual localization under challenging perceptual conditions. In: IEEE International Conference on Robotics and Automation. pp. 2614–2620 (2017)
23. Neubert, P., Protzel, P.: Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments. IEEE Robotics and Automation Letters **1**(1), 484–491 (2016)
24. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3384–3391 (2010)
25. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007)
26. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12716–12725 (2019)
27. Sarlin, P.E., Debraine, F., Dymczyk, M., Siegwart, R., Cadena, C.: Leveraging deep visual descriptors for hierarchical efficient localization. Conference on Robot Learning (2018)
28. Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixe, L.: Understanding the limitations of cnn-based absolute camera pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3302–3312 (2019)
29. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. In: Advances in Neural Information Processing Systems, pp. 10090–10100. Curran Associates, Inc. (2019)
30. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the IEEE International Conference on Computer Vision. vol. 2, pp. 1470–1477 (2003)
31. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4297–4304 (2015)
32. Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M.: Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. Robotics: Science and Systems XI: pp. 1–10 (2015)
33. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: Inloc: Indoor visual localization with dense matching and view syn-

thesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7199–7209 (2018)
34. Tian, Y., Chen, C., Shah, M.: Cross-view image matching for geo-localization in urban environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3616 (2017)
35. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1808–1817 (2015)
36. Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: European Conference on Computer Vision. pp. 494–509. Springer (2016)
37. Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using LSTMs for structured feature correlation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 627–637 (2017)
38. Weinzaepfel, P., Csurka, G., Cabon, Y., Humenberger, M.: Visual localization by learning objects-of-interest dense match regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5634–5643 (2019)
39. Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. In: European Conference on Computer Vision. pp. 37–55. Springer (2016)
40. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
41. Xin, Z., Cai, Y., Lu, T., Xing, X., Cai, S., Zhang, J., Yang, Y., Wang, Y.: Localizing discriminative visual landmarks for place recognition. In: IEEE International Conference on Robotics and Automation. pp. 5979–5985 (2019)
42. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. pp. 818–833. Springer (2014)
43. Zeisl, B., Sattler, T., Pollefeys, M.: Camera pose voting for large-scale image-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2704–2712 (2015)