On the Estimation of Image-matching Uncertainty in Visual Place Recognition

Mubariz Zaffar
ME, TU Delft
The Netherlands
M.Zaffar@tudelft.nl

Liangliang Nan
ABE, TU Delft
The Netherlands
Liangliang.Nan@tudelft.nl

Julian F. P. Kooij
ME, TU Delft
The Netherlands
J.F.P.Kooij@tudelft.nl

Abstract

In Visual Place Recognition (VPR) the pose of a query image is estimated by comparing the image to a map of reference images with known reference poses. As is typical for image retrieval problems, a feature extractor maps the query and reference images to a feature space, where a nearest neighbor search is then performed. However, till recently little attention has been given to quantifying the confidence that a retrieved reference image is a correct match. Highly certain but incorrect retrieval can lead to catastrophic failure of VPR-based localization pipelines. This work compares for the first time the main approaches for estimating the image-matching uncertainty, including the traditional retrieval-based uncertainty estimation, more recent data-driven aleatoric uncertainty estimation, and the compute-intensive geometric verification. We further formulate a simple baseline method, "SUE", which unlike the other methods considers the freely-available poses of the reference images in the map. Our experiments reveal that a simple L2-distance between the query and reference descriptors is already a better estimate of image-matching uncertainty than current data-driven approaches. SUE outperforms the other efficient uncertainty estimation methods, and its uncertainty estimates complement the computationally expensive geometric verification approach. Future works for uncertainty estimation in VPR should consider the baselines discussed in this work.

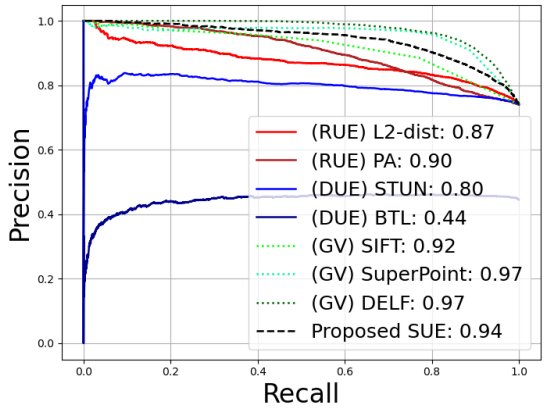

Figure 1. The Precision-Recall curves on the Pittsburgh dataset [4] for the three common categories of VPR uncertainty estimation methods (RUE, DUE, GV), and for our proposed baseline **SUE** which uniquely considers spatial locations of the top-K references. The global image descriptors [9] are fixed for all methods except BTL [50]. The *only* difference is the confidence given by each uncertainty estimation method to the best-matched reference descriptors for the corresponding queries. The legend lists the Area-under-the-Precision-Recall-curves. As *GV* methods are *two to three orders of magnitude* more computationally expensive than the others, they are plotted as dotted lines. Surprisingly, simple L2-distance in feature space is a better estimate of VPR uncertainty than recent deep learning-based uncertainty estimates. SUE outperforms all other efficient uncertainty estimation methods.

1. Introduction

Visual Place Recognition (VPR) is the problem of identifying a previously visited place given a query camera image and a map of geo-tagged reference images [28]. It has applications in vehicle localization [57], 3D modeling [1], image search [46], and loop-closure in Simultaneous Localization and Mapping (SLAM) [8, 28].

VPR is typically approached as an image retrieval problem, transforming images into feature vectors in a latent feature space where an efficient nearest neighbor search compares the query to all references. The pose of the query image is then approximated to be the same as that of the retrieved nearest neighbor references. Since successful VPR requires a good image representation that is robust to viewpoint and/or appearance changes [16, 28, 30], the field has benefited from advances in deep representation learning.

However, two images with similar visual content could still originate from geographically far-apart areas, a concept referred to as *perceptual-aliasing* in VPR [16]. For example, images with mostly sky could match many locations

on an outdoor map. This constitutes aleatoric uncertainty, i.e., inherent noise or ambiguity in the data which cannot be reduced, as opposed to epistemic uncertainty which could be addressed with more training data [23]. The close proximity of perceptually aliased images in the feature space can result in catastrophic failures For instance, a highly confident false-positive from VPR could result in an incorrect loop closure in a SLAM pipeline, leading to misaligned maps [8, 28]. Reliable uncertainty estimation on the quality of the match is therefore key to avoid such failures by, e.g., rejecting results above a certain uncertainty threshold. Moreover, uncertainty estimation can also be used to fuse multiple predictions in VPR ensemble methods [9].

From existing literature, we identify three categories of methods to estimate image-matching uncertainty in VPR: retrieval-based uncertainty estimation (*RUE*), data-driven aleatoric uncertainty estimation (*DUE*), and geometric verification (*GV*) by local feature matching. **RUE:** Traditionally in VPR, the L2-distance between the query and the best-matched reference in the feature space has been used as an estimate of uncertainty [35]. The ratio of L2-distance between the first and second nearest neighbour reference is also used [18]. **DUE:** On the other hand, several recent works, such as the Bayesian Triplet Loss [50] and the Self-teaching Uncertainty Estimation [9], have proposed to explicitly learn to predict the aleatoric uncertainty from the query's image content only. **GV:** Another way to assert matching confidence is to test for consistent geometry among matched local features between the query and the best matching reference in a RANSAC loop [33].

Remarkably, none of the three categories exploit the spatial locations of matched images in the actual reference map, which we hypothesize can be an important source of information for estimating VPR matching uncertainty. To test this hypothesis, we formulate a new simple baseline, Spatial Uncertainty Estimation (SUE). SUE is a straightforward and efficient approach to estimating uncertainty for a query image's match, using the spatial spread of the physical poses for the most similar references in the map as a proxy. A high spatial spread indicates perceptual aliasing leading to high matching uncertainty, while a low spread indicates a distinct area is matched. An overview of the sources of information employed by all categories of methods and by SUE is provided in Table 1.

While all categories of uncertainty estimation methods aim for the same task, i.e., rejecting false positives in VPR, previous evaluations did not include all categories, providing an incomplete picture of the state-of-the-art. This work therefore compares the three existing categories and SUE on a levelled playing field, to provide recommendations for future research, and insights on the strengths/weaknesses of each category. For instance, as the preview of the experimental results in Fig. 1 indicates, SUE outperforms other efficient methods (this and other experiments will be discussed in more detail in Section 4).

Categ.	Descr.?	Poses?	Images?	Efficient?
RUE	Top-K	No	No	Yes
DUE	No	Only train	Yes	Yes
GV	No	No	Yes	No
SUE	*Top-K*	*Top-K*	*No*	*Yes*

Table 1. Overview of the sources of information needed by the current main categories for VPR uncertainty estimation, and by the proposed method *SUE*: the query/reference global image descriptors, the reference poses, or complete query/reference images. Efficiency refers to the inference time needed by each approach.

Concretely, our contributions are:

1. A comparison of three different categories of uncertainty estimation methods in VPR.
2. A new simple baseline method, SUE, that considers the spatial locations of the reference images, a source of information not used by existing categories.
3. Since *GV* gives the best uncertainty estimates albeit at a higher computational cost, we investigate whether the other methods are complementary to *GV*.

2. Related work

Visual place recognition was first surveyed in the seminal work of Lowry *et al.* [28]. The three fundamental VPR challenges identified by Lowry *et al.* are viewpoint changes, appearance changes, and perceptual-aliasing.

The concept of matching images for VPR dates back to before the deep-learning era, when handcrafted methods were used to perform VPR [12, 21, 42, 44]. However, with the rise of deep learning, many deep learning-based methods were proposed to solve the first two challenges in VPR. A broad categorization of these methods can be done based on their underlying novelty, such as the use of a novel loss function [26, 39, 45], better training data [2, 6], new architectures [48, 51, 56], and new methods for feature aggregation [3, 4, 19, 37]. A number of benchmarks have been proposed in VPR, for example, the recent Deep Visual Geo-localization benchmark [7], VPR-Bench [52] and similar benchmarks in the image retrieval community [38, 41]. From these benchmarks, it is clear that the deep learning-based VPR techniques outperform handcrafted techniques by a significant margin on most datasets.

We focus on the third challenge identified in Lowry *et al.*, i.e., perceptual aliasing, which arises from aleatoric uncertainty in the data. This challenge has received less attention in VPR literature compared to viewpoint and appearance changes. Most works in VPR use the distance (e.g., L2 or Cosine) in feature space between a query and the nearest neighbor as the uncertainty estimate [7], or the distance

between the retrieved nearest neighbors [18]. Some more recent works model the aleatoric uncertainty in image retrieval, e.g., the Bayesian Triplet Loss (BTL) [50] and the Self-Teaching Uncertainty Estimation (STUN) [9]. Both BTL and STUN estimate the aleatoric uncertainty in the training data by representing images as distributions instead of point estimates in the feature space. Each image thus has an associated mean and variance for a feature descriptor.

Gronat *et al.* [17] treat VPR as a classification problem by training place-specific classifiers, one for each place, where each classifier naturally outputs a confidence estimate for the corresponding pose. Pion *et al.* [36] approximate the pose of the query image by aggregating the pose hypotheses from the top-retrieved nearest neighbors, weighing each hypothesis based on the distance in the feature space. The variance of the aggregated pose represents uncertainty over the pose space. Notably, this concept of pose uncertainty has been modeled in these existing works [17, 36, 53] and other related tasks such as classical Particle Filters [14], but, to the authors' best knowledge, the uncertainty estimates derived based on the distribution of pose hypotheses has not yet been studied as a proxy for image-matching uncertainty.

Beyond global descriptors-based VPR, in local feature matching-based image retrieval the inlier count (aka. geometric verification) has been used as an estimate of confidence [33]. Zeisl *et al.* [55] perform 2D-to-3D local feature matching to estimate a distribution over the possible query poses. The work of [54] uses such inlier count from local feature matching and combines it with the pose distribution of retrieved images to estimate the confidence of localization. Since local features can appear in similar geometric configurations (geometric burstiness) across unrelated images, [40] proposes to use the pose information to downweight such matches in the inlier count. However, retrieving images based on local feature descriptors is computationally expensive, whereas VPR instead only efficiently compares global image descriptors. [1] Absolute Pose Regression (APR) directly regresses the absolute pose given a camera image, and has also considered pose uncertainty estimation. Some approaches to uncertainty-aware APR include CoordiNet [32], Bayesian PoseNet [22], and HydraNet [34]. Unlike VPR, APR approaches do not generalize to new environments. In this work, we focus on estimating the image-matching uncertainty for VPR.

3. Methodology

This section first introduces the task of uncertainty estimation for VPR. We then formalize VPR, and describe the three main categories of uncertainty estimation methods. Next, we formulate the proposed baseline approach, SUE, which unlike the other three categories uses the freely available reference poses information. Finally, we outline how we combine the different categories of methods with the computationally expensive geometric verification to investigate if the uncertainty estimates are complementary.

3.1. Uncertainty estimation in VPR

Typically VPR is considered as an image retrieval task: finding the most similar reference images to the query by Euclidean distance in some feature space. The poses associated with the images however distinguish VPR from other image retrieval tasks, such as web search, where matches are correct if their image content should be judged as the same. In VPR we often instead refer to the location of the query and references to judge matches: a retrieved reference is only acceptable if its pose is within a maximum distance threshold of the (unknown) true pose of the query [7, 16, 52]. Ideally, the closest matches in the feature space thus also have the poses closest to the query pose. However, this is often not the case in VPR due to *perceptual aliasing*, a form of aleatoric uncertainty since it cannot be reduced by choosing a different feature encoder or by using more training data.

It is therefore desirable to obtain some *uncertainty score* s_q for a query and the retrieved nearest neighbor, where a low score expresses confidence that the nearest neighbor is a correct match. A threshold τ on the score could then reject a query ($s_q > \tau$) for which the best match is at risk of being incorrect to prevent failures of the downstream application [16]. The objective of VPR uncertainty estimation is thus to score queries, such that queries with reliable matches can be distinguished from those with possible incorrect matches. Note that while an uncertainty estimation method could provide scores with an explicit probabilistic interpretation, this is not a strict requirement to apply an acceptance threshold.

3.2. Formalizing VPR

Given a set of reference images $\mathcal{I}$ with known poses $\mathcal{P}$, the goal of VPR is to find one or multiple reference images $I_i \in \mathcal{I}$ that match the place of a query image I_q. The unknown pose p_q for the query I_q can then be approximated from the poses of the matched references $p_i \in \mathcal{P}$, since correct matches should have been taken in the same area. The exact formulation of a pose generally depends on the localization source and the task, for example, 2D GPS coordinates for visual geo-localization [7], or 6D pose [25]. In this research, we follow a general task-independent formulation and only assume that a pose p_i consists of 2D or 3D spatial coordinates in some global coordinate system.

In the offline map preparation phase of VPR, before accepting queries, a feature extractor G is applied to every ref-

[1] We study the relation between geometric burstiness and SUE in the supplementary materials.

erence image $I_i \in \mathcal{I}$ to obtain D-dimensional reference feature descriptors $f_i = G(I_i)$. Usually G is a trained neural network [30] or a handcrafted feature descriptor [13]. The resulting VPR map $\mathcal{M} = (\mathcal{R}, \mathcal{P})$ contains the reference feature descriptors set $\mathcal{R} = \{f_1, \cdots f_N\}$, where each descriptor f_i is associated with a corresponding pose $p_i \in \mathcal{P}$.

At test time, the same feature extractor G is applied to the query image I_q, and its query descriptor $f_q = G(I_q)$ is compared to the reference descriptors in the map $\mathcal{M}$. This can be achieved through an efficient K-nearest neighbor lookup, considering the L2-distances $d_i = ||f_i - f_q||_2$ between each reference i and the query. This gives an ordered list of K nearest neighbor references $\mathcal{R}_{\text{nn}} = [f_{(1)}, \cdots, f_{(K)}]$, ranked by increasing distance $d_{(1)} \leq \cdots \leq d_{(K)}$ and with corresponding poses $\mathcal{P}_{\text{nn}} = [p_{(1)}, \cdots, p_{(K)}]$. Here we use bracketed subscript (j) to indicate j-th item in the ranked order, i.e., $f_{(1)} = \text{argmin}_{f_i \in \mathcal{R}} ||f_i - f_q||_2$ is the descriptor with the smallest distance to the query in the feature space.

Each corresponding pose $p_{(i)} \in \mathcal{P}$ can be considered as a hypothesis to estimate the query's true pose p_q, though usually only the pose of the best matching reference feature descriptor $f_{(1)}$ is considered as the VPR pose estimate p'_q for the query, i.e., $p'_q = p_{(1)}$ [52]. We follow this best-match-based query pose estimation in this work. In benchmarks, a match is considered correct if p'_q is 'physically near' to p_q. The threshold on what distance is still accepted as the same 'place' depends on the scale of each localization task [28].

3.3. Current VPR uncertainty estimation categories

We now describe various representative uncertainty estimation methods for the three common categories.

Retrieval-based uncertainty estimation (*RUE*): Commonly, the matching uncertainty in VPR is considered proportional to the L2-distance from the best match $d_{(1)}$, so $s_q = d_{(1)}$ [7, 52], as this distance indicates relevant differences between the visual content of the query and match.

An alternative is to consider the distance ratio between the first and second nearest neighbor, $s_q = d_{(1)}/d_{(2)}$. This ratio is quite similar to the perceptual aliasing score (PA score) [18] and the false-positive rejection criterion in the popular local feature descriptor SIFT [27].

Data-driven uncertainty estimation (*DUE*): State-of-the-art VPR encoders are typically deep neural networks trained on a labeled VPR dataset. The labeled training data contains the ground-truth poses $\mathcal{P}_{\text{train}}$ for the training references and query images $\mathcal{I}_{\text{train}}$. A deep encoder G can be adapted to also predict the aleatoric uncertainty of matching a nearby pose, by learning from the training query image in $\mathcal{I}_{\text{train}}$ when an image is distinctive and obtains good pose matches within $\mathcal{P}_{\text{train}}$, and when not (e.g., images of trees, uniform walls, or sky). Methods in this category include the Bayesian Triplet Loss (BTL) [50], and STUN [9]. Note that the learned uncertainty is based on the training images and poses, not those in the test-time reference map $\mathcal{M}$.

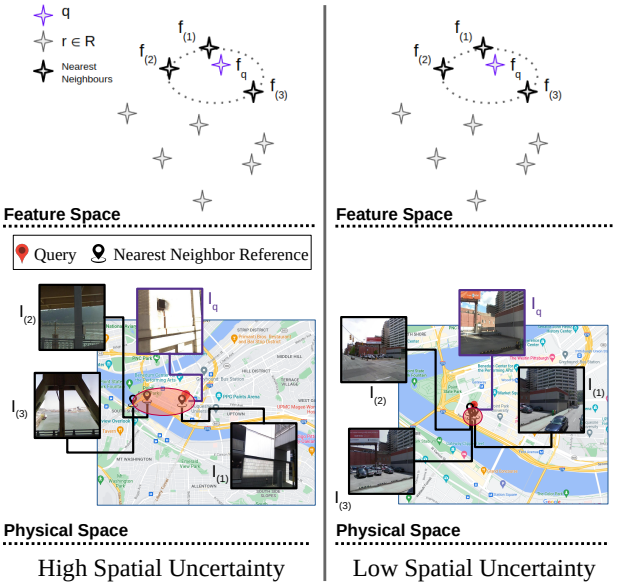

Figure 2. In VPR, a query q is compared in feature space to features $f_i \in \mathcal{R}$ of reference images with known poses. The nearest neighbors $f_{(1)}, \cdots, f_{(K)}$ are retrieved as matches. Left: The retrieved references $I_{(1)}, I_{(2)}, I_{(3)}$ share similar visual content with the query (walls, pillars, and blobs), but are geographically far apart, reflecting high uncertainty that the matched reference is correct. Right: For another query, the retrieved references are geographically close together, indicating low uncertainty.

In general, an uncertainty-aware encoder $(\bar{f}_i, \sigma_i^2) = G'(I_i)$ predicts for an image I_i not only the expected feature $\bar{f}_i$, but also the *variance in the feature space*, i.e., $f_i \sim \mathbf{N}(\bar{f}_i, \sigma_i^2)$. The total variance in σ_i^2 can be used as a proxy for the image-matching uncertainty, $s_q = ||\sigma_i^2||_1$. The computational overhead of the deep network producing an additional output σ_i^2 is low.

Geometric verification (*GV*): Another way to estimate image-matching uncertainty is to compare the query and the best-matched reference image in more detail through local feature matching and geometric verification in a RANSAC loop, e.g., through the use of SIFT [27], DELF [33], and SuperPoint [15]. All the matched local features that satisfy a geometric transformation estimated from the randomly sampled set of matched local features between the query image and the reference image are considered inliers. The confidence is indicated by the number of inliers c_{gv}, which could be expressed as a matching uncertainty estimate, i.e., $s_q = -c_{gv}$. While geometric verification yields high-quality uncertainty estimates, such post-processing is computationally expensive compared to the other methods.

3.4. Spatial uncertainty estimation (*SUE*) for VPR

We observe that the poses in the reference set $\mathcal{P}$ are a potentially powerful and freely available source of information at *test time*, which current uncertainty estimation methods do not exploit (more details will be presented in Sec. 3.1). The intuition behind this is illustrated in Fig. 2, where we show that if the nearest neighbors in the feature space are spatially far apart in their respective 2D/3D world coordinates, it indicates that such a feature suffers from perceptual aliasing: various areas in the test reference set contain the queried appearance, thus uncertainty on the pose estimate should be high. On the other hand, if the nearest neighbors in the feature space are also spatially close together, there is agreement among the matching pose hypotheses that the matched area is distinct within that given reference set, thus the uncertainty should be low.

To test this insight, we now formulate SUE, a purposefully simple image-matching uncertainty estimation method. Given the K-best retrieved references, fit a 2D or 3D multivariate Gaussian distribution $\mathbf{N}(\mu_p, \Sigma_p)$ over their 2D or 3D poses $\mathcal{P}_{\text{nn}}$,

$$\mu_p = \frac{1}{\sum_i w_{(i)}} \sum_{i=1}^{K} w_{(i)} \cdot p_{(i)}, \tag{1}$$

$$\Sigma_p = \frac{1}{\sum_i w_{(i)}} \sum_{i=1}^{K} w_{(i)} \cdot (p_{(i)} - \mu_p)(p_{(i)} - \mu_p)^\top, \tag{2}$$

where the relative contribution $w_{(i)}$ of the i-th best reference pose $p_{(i)}$ decreases as its L2-distance $d_{(i)}$ to the query in the feature space increases,

$$w_{(i)} = e^{-\lambda \cdot d_{(i)}}, \quad \text{where} \quad d_{(i)} = ||f_q - f_{(i)}||_2. \tag{3}$$

The total variance across the spatial pose dimensions could then serve as a proxy for image-matching uncertainty, i.e., $s_q = \text{trace}(\Sigma_p)$.

The hyper-parameter λ controls the non-linear relative contribution of a pose p_i for the nearest neighbor $f_{(i)} \in \mathcal{R}_{\text{nn}}$ given its distance $d_{(i)}$ in the feature space. This hyper-parameter can be optimized on training data, though our experiments will show that its choice is remarkably robust across various real-world benchmark datasets.

3.5. Complementing geometric verification

To study to what extent SUE's (or another method's) s_q provides information not captured by the c_{gv} metric from geometric verification, we treat both scores as a 2D feature vector and train a classifier to predict if a best-matched reference should be accepted as a true-positive, or rejected as a false-positive. The regular rejection threshold is extended from a single score ($s_q > \tau$) to a linear weighted sum of both scores ($s_q/\tau_1 + c_{gv}/\tau_2 > 1$), by the use of a regular linear Support Vector Machine (SVM) as a classifier.

4. Experiments

We first present the setup for our experiments. Then, we compare the performance of all the image-matching uncertainty estimation methods on multiple benchmark datasets. Next, we test if the methods are complementary to geometric verification. Finally, we present an ablation over the hyper-parameters of SUE and provide a discussion.

4.1. Experimental setup

This section describes the datasets, baselines, evaluation metrics, and implementation details of our work.

Datasets: We use six public VPR datasets in this work: Pittsburgh-250k [4], Sanfrancisco [10, 47], Stlucia [31], Eysham [11], MSLS [49] and Nordland [43]. Details of these datasets and their respective ground-truths in [5].

Baselines: Our primary baselines for uncertainty estimation include the L2-distance in feature space d_q, the perceptual aliasing score (PA score [18]), the Bayesian Triplet Loss (BTL) [50] and STUN [9]. As the code for BTL is not open-source, we implement it following the pseudo-code and the network details provided in the original paper.

For geometric verification, we test three types of local feature descriptors, namely the handcrafted SIFT [27], the deep-learning-based DELF [33], and SuperPoint [15], which we refer to as SIFT-RANSAC, DELF-RANSAC and Superpoint-RANSAC, respectively.

Evaluation metrics: The precision-recall (PR) curves have been widely used in VPR for estimating the retrieval quality [52]. However, they can also be used to estimate the uncertainty estimate in VPR as widely used in existing uncertainty estimation tasks in deep learning [20, 29]. The choice of PR-curves over the Receiver Operating Characteristic (ROC) curve is due to the absence of true-negatives in employed VPR datasets. Given a fixed list of retrieved images, the Precision-Recall curves can reflect the technique with the better uncertainty estimates s_q. A technique that can perfectly classify between true-positives (TP) and false-positives (FP), given the uncertainty estimates, achieves an Area-under-the-Precision-Recall-Curve (AUC-PR) of 1.

For the combination of uncertainty estimates with geometric verification, the task is formulated as binary classification and we use accuracy as an evaluation metric based on the ground-truth true-positives and false-positives [5].

Implementation details: For SUE and all the other baselines except BTL, we use a ResNet-50 backbone with GeM pooling trained in a self-teaching manner in [9] on the training split of the Pittsburgh dataset. Each feature vector f_i is 2048 dimensional. For BTL, the same backbone and training data are used, but using the training procedure specified in the original BTL paper [50]. For DELF and SuperPoint, the implementations are open-sourced by the respective authors, and the default settings are employed. For SIFT-RANSAC we use the OpenCV implementation with

Method	↑ Pitts.	↑ Sanfr.	↑ Stluc.	↑ Eyn.	↑ MSLS	↑ Nordland	↑ Average	↓ Time
(RUE) L2-distance	0.87	0.76	0.79	0.87	0.64	0.18	0.69	**0.05**
(RUE) PA-Score [18]	0.90	0.65	0.77	0.88	0.68	0.21	0.68	**0.05**
(DUE) BTL [50]	0.44	0.17	0.34	0.45	0.21	0.07	0.28	0.20
(DUE) STUN [9]	0.79	0.57	0.66	0.71	0.44	0.05	0.54	0.10
SUE	**0.94**	**0.84**	**0.88**	**0.93**	**0.77**	**0.26**	**0.77**	1.08
(GV) SIFT-RANSAC [27]	0.92	0.89	0.93	**0.96**	0.70	0.15	0.76	**129**
(GV) DELF-RANSAC [33]	0.97	**0.92**	**0.97**	0.95	**0.95**	**0.84**	**0.93**	1587
(GV) Super-RANSAC [15]	0.95	**0.95**	**0.97**	**0.96**	0.87	0.50	0.87	848

Table 2. The AUC-PR of all the compared methods. Higher AUC-PR is better, and best is in Bold. The bottom rows are the computationally expensive geometric verification methods. The last column lists the time (msec) to give an uncertainty estimate for a single query image.

the number of extracted features set to 5000, the Lowe test ratio to 0.6, and the number of RANSAC iterations to 1000.

The hyper-parameters in SUE are fined-tuned only on the Pittsburgh dataset and then fixed as $\lambda = 350$ and $K = 10$ for all datasets and experiments. An ablation over these parameters is given later in section 4.4. The SVM is trained with stochastic gradient descent with hinge loss and an L1-penalty, and a maximum of 1000 training iterations.

4.2. Performance comparison

We first compare all the uncertainty estimation methods formulated in this work, both qualitatively and quantitatively, and in terms of their computational overhead.

Area-under-the-Precision-Recall-curves: The AUC-PR for all the methods on all the datasets are summarized in Table 2. SUE outperforms other efficient methods by a clear margin, even on the Pittsburgh dataset which was used for training STUN and BTL. It is also important to note that a basic L2-distance-based uncertainty already outperforms BTL and STUN. Moreover, geometric verification outperforms all other uncertainty estimates although SUE achieves comparable performance. The precision-recall curves for the Pittsburgh dataset are shown in Fig. 1, and for the remainder datasets are provided in the accompanying supplementary materials.

Computational requirements: We further report the time taken to compute the *GV* confidence s_q and the uncertainty estimates in Table 2. Although *GV* gives useful uncertainty estimates, the high computational cost of these *GV* methods may be prohibitive for real-time online applications. Our implementation of SUE is about *three orders of magnitude* faster than *GV* using DELF-RANSAC.

Qualitative results: To obtain insight into how the different methods interpret the visual content in query images and what they are sensitive to, we show in Fig. 3 examples of the most and the least uncertain query images for different methods in the Pittsburgh dataset. While all methods usually consider feature-rich and distinctive buildings as least uncertain for VPR, differences between the methods lie in the most uncertain images. Highly saturated test images are considered most uncertain by L2-distance-based uncertainty because such saturation did not exist during the reference traversals of the same scene. On the other hand, STUN considers images of trees and walls that usually contribute to perceptual aliasing as the most uncertain for VPR. SUE considers traffic squares and common building patterns as the most uncertain. Note that because SUE uses the freely available pose information in the test reference set; whether a traffic square or a building is considered uncertain is specific to this test reference set and not due to a generally-applicable visual property.

We further show in Fig. 4 several images that illustrate failure cases of *RUE* and *DUE* in comparison to *SUE*. Images of walls generally contribute to high aleatoric uncertainty (*DUE*) and are closer together in the feature space in terms of L2-distance (*RUE*). However, we note that the query in Fig. 4 Top is correctly matched since only a unique wall with this pattern exists in the test reference set. SUE and L2-distance correctly give this query a low uncertainty, but STUN fails. The query image in Fig. 4 Bottom is given low uncertainty by L2-distance than ranking with STUN and SUE. This is because images with large portions of sky contribute to aleatoric uncertainty but they are close in terms of the feature space L2-distance. This query is mismatched and identifies where L2-distance-based uncertainty fails in comparison to STUN and SUE.

4.3. Complementing geometric verification

Finally, we test if efficient uncertainty estimation can complement geometric verification, as outlined in Sec. 3.5. We show in Fig. 5 the relation between the different types of uncertainties with the uncertainty from geometric verification. As we note from Table 2, STUN outperforms BTL, and L2-distance is on average better than the PA-score, thus we only combine STUN, L2-distance, and SUE with geometric-verification for this analysis.

SUE provides complementary performance by giving low uncertainty s_q to images that are correctly matched but

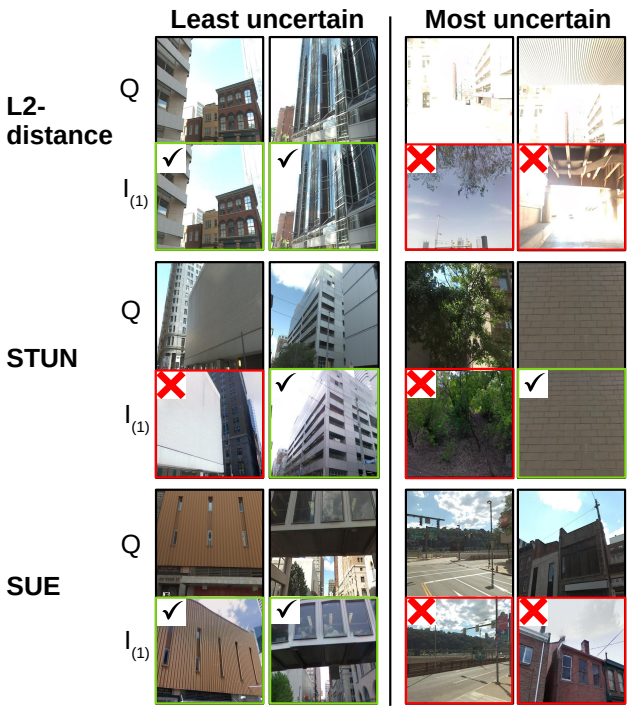

Figure 3. Examples of the two least and the two most uncertain query images with the corresponding nearest neighbor on the Pittsburgh dataset. The colors/symbols indicate whether the retrieved image is a correct match.

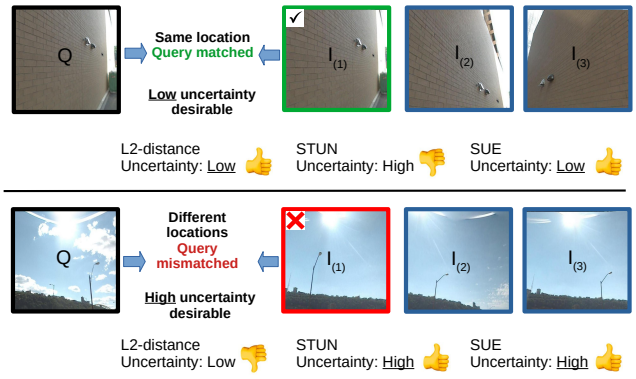

Figure 4. Two queries and their nearest neighbor reference images that illustrate cases where SUE outperforms other methods. Ideally a method assigns high uncertainty to the mismatched query and low uncertainty to the correct match, as SUE does here.

which were given high *GV* uncertainty. Some of these complementary queries are shown in Fig. 6, where it can be seen that these queries are images that are generally difficult to match local feature descriptors, such as facades, trees, and other repetitive features within the image [24]. We also show the linear boundaries learned by SVM to classify between true-positives and false-positives. The classification accuracy of the different methods is reported in Table 3.

Method	Pitts.	San.	Stlu.	Eyn.	MSLS
Superpoint	85.1	53.2	76.4	67.5	36.9
DELF	86.0	86.6	85.3	78.3	**80.2**
L2-distance	75.7	57.3	56.2	67.7	36.8
STUN	74.0	54.0	58.0	67.6	37.4
SUE	78.9	70.7	72.8	77.3	46.0
DELF+L2-di.	85.7	86.1	82.3	77.3	72.0
DELF+STUN	85.4	81.6	80.1	75.0	68.2
DELF+SUE	**87.1**	**89.6**	**88.7**	**82.1**	**73.4**

Table 3. Binary classification accuracy given the uncertainty estimates of various methods, using a linear SVM trained *only* on the Pittsburgh dataset. The combination *DELF + SUE* generalizes better than baseline combinations, except on the MSLS dataset where although *DELF+SUE* is better than the other combinations, the SVM boundaries learned from Pittsburgh are not the best.

4.4. Ablation study

SUE requires two hyper-parameters, the number of nearest neighbors K and the decay parameter λ that controls the relative contribution of the poses of the nearest neighbors. We show the ablation over these parameters in Fig. 7 by plotting the corresponding AUC-PR values for all datasets given a set of values for each parameter. The trend remains primarily the same across all datasets. We note that the AUC-PR increases by considering more nearest neighbors but the curves mostly plateau after $K = 5$, since poses from low-ranked neighbors contribute less to the overall pose hypothesis. For λ, we see that the range $200 - 400$ is generally stable and gives reliable uncertainty estimates. We also note here (with details in the appendix) that SUE generalizes to different backbones (CosPlace [6]), and that exponential weighing of SUE in Equation (2) performs better than uniform weighing (an average AUC of 0.87 vs 0.70).

4.5. Discussion

We can now make several **recommendations** for estimating the image-matching uncertainty in VPR. First, future works evaluating image-matching uncertainty estimation should include diverse baselines such as SUE, even if they are simple. As our intra-category comparison revealed, even a common L2-distance-based image-matching uncertainty estimation may outperform data-driven techniques. Second, aleatoric uncertainty from training data does not necessarily generalize to the test data, so learning-based approaches should consider that perceptual aliasing is not just a property of the image content, but also the reference map at test time. Referring back to the example of the sky in images being ambiguous for an outdoor map; in an indoor map containing just one open-air patio, such images with sky might instead be considered distinctive for their loca-

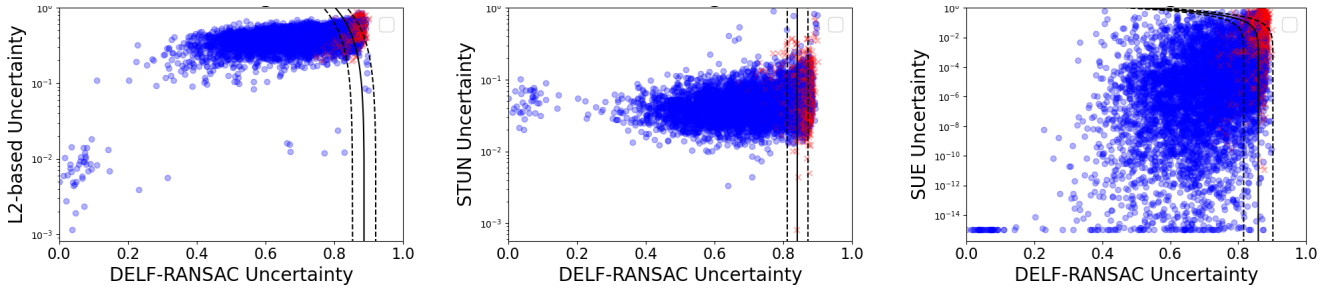

Figure 5. The relation between geometric verification uncertainty (x-axis) and the L2/STUN/SUE uncertainty (y-axis) on the Pittsburgh dataset [4]. Each point represents a query, with *blue* indicating a correct match, and *red* otherwise. The linear SVM boundaries are shown as black lines, while the dashed lines are the SVM margins. Scores have been linearly scaled to the $[0, 1]$ range based on the min/max value in the training data, and for better visualization the vertical scale is in log-space, hence the SVM boundaries appear non-linear. The class distributions in the right-most plot reveal that SUE complements geometric-verification, especially when the latter has low confidence.

Figure 6. Correctly matched queries that are given high uncertainty by DELF-RANSAC and low uncertainty by SUE.

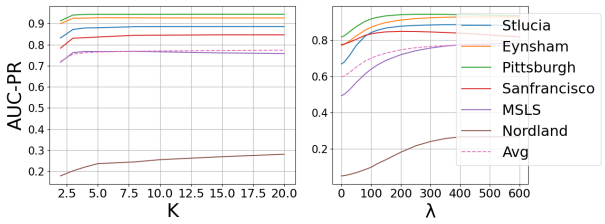

Figure 7. Effect of changing SUE's hyper-parameters K and λ on the AUC-PR. For each curve, the other fixed hyper-parameter is chosen as $K = 10$ or $\lambda = 350$.

tion. Third, while *GV* gives the best uncertainty estimates at the expense of high computational needs, it is still susceptible to aleatoric uncertainty within the image, as repetitive structures, trees, and walls may also lead to incorrect matches of local features. In VPR, *GV* methods can still benefit from complementary uncertainty estimates provided by other methods, such as SUE.

We also note some potential **limitations of SUE**. SUE may underestimate the uncertainty if K is too small to retrieve aliased references from multiple locations. Selecting K for maps with mixed scene depths can therefore be challenging. Images in areas with low scene depth will already be perceptually distinct at small spatial offsets, whereas at high scene depth even images further apart may suffer from perceptual aliasing. A K that suffices for small scene depths could be too small for areas with high scene depths. This could be mitigated by dynamically incrementing K till $w_{(K)}$ becomes nearly zero. Now consider reference locations A and B which are perceptually aliased, i.e. all their image descriptors are similar. If A has 1000 references and B has one, even with $K \geq 1001$, SUE will always be confident about queries from either A or B as nearly all retrieved matches are spatially close. The high coverage of A over B thus presents an unwanted confidence bias, unless the chance of visiting A over B at test time is also $1000\times$ higher. Nevertheless, we have shown that despite these assumptions SUE performs well on many real-world datasets. We study this more in depth in the supplementary materials, and there also present a possible solution.

5. Conclusions

We have compared different approaches for estimating the image-matching uncertainty in VPR, which provided (surprising) insights into this task, e.g. existing methods that learn aleatoric uncertainty from the training dataset often do not generalize well to the reference map at test time, and the common L2-distance in the feature space can be a more reliable indicator of matching uncertainty. We have shown that matching uncertainty in VPR is tightly related to the reference set at test time. Our new baseline SUE uniquely considers the spatial locations of the references, and outperforms all but the computationally expensive geometric verification. Its uncertainty estimates complement those of geometric verification. The choices for SUE's hyper-parameters generalize for most queries across the tested datasets. We made recommendations for future research in this area.

Acknowledgement. This work was supported by the 3D Urban Understanding Lab established under the TU Delft AI Initiative, and the EU Horizon 2020 programme under grant number 964505 (Epistemic AI).

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54 (10):105–112, 2011.

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

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

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef 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.

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

[6] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022.

[7] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5407, 2022.

[8] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[9] Kaiwen Cai, Chris Xiaoxuan Lu, and Xiaowei Huang. STUN: Self-teaching uncertainty estimation for place recognition. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6614–6621. IEEE, 2022.

[10] David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al. City-scale landmark identification on mobile devices. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 737–744. IEEE, 2011.

[11] Mark Cummins. Highly scalable appearance-only slam-fab-map 2.0. *In Proceedings of the Robotics: Sciences and Systems (RSS) Conference*, 2009.

[12] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 886–893. IEEE, 2005.

[14] Frank Dellaert, Wolfram Burgard, Dieter Fox, and Sebastian Thrun. Using the condensation algorithm for robust, vision-based mobile robot localization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 588–594. IEEE, 1999.

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018.

[16] Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[17] Petr Gronat, Guillaume Obozinski, Josef Sivic, and Tomas Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 907–914, 2013.

[18] Stephen Hausler, Tobias Fischer, and Michael Milford. Unsupervised complementary-aware multi-process fusion for visual place recognition. *arXiv preprint arXiv:2112.04701*, 2021.

[19] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-CNN: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021.

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2016.

[21] Kin Leong Ho and Paul Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286, 2007.

[22] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016.

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

[24] Jan Knopp, Josef Sivic, and Tomas Pajdla. Avoiding confusing features in place recognition. In *Proceedings of the European Conference on Computer Vision*, pages 748–761. Springer, 2010.

[25] Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera relocalization by computing pairwise relative poses using convolutional neural network. In *IEEE International Conference on Computer Vision Workshops*, pages 929–938, 2017.

[26] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Generalized contrastive optimization of siamese networks for place recognition. *arXiv preprint arXiv:2103.06638*, 2021.

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[28] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015.

[29] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[30] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021.

[31] Michael J Milford and Gordon F Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 24(5):1038–1053, 2008.

[32] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Coordinet: uncertainty-aware pose regressor for reliable vehicle localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2229–2238, 2022.

[33] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3456–3465, 2017.

[34] Valentin Peretroukhin, Brandon Wagstaff, Matthew Giamou, and Jonathan Kelly. Probabilistic regression of rotations using quaternion averaging and a deep multi-headed network. *arXiv preprint arXiv:1904.03182*, 2019.

[35] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109, 2018.

[36] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *International Conference on 3D Vision (3DV)*, pages 483–494. IEEE, 2020.

[37] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018.

[38] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2018.

[39] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116, 2019.

[40] Torsten Sattler, Michal Havlena, Konrad Schindler, and Marc Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1582–1590, 2016.

[41] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018.

[42] Stephen Se, David Lowe, and Jim Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research*, 21(8):735–758, 2002.

[43] Sindre Skrede. Nordland dataset. https://bit.ly/2QVBOym, 2013.

[44] Elena Stumm, Christopher Mei, and Simon Lacroix. Probabilistic place recognition with covisibility maps. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4158–4163. IEEE, 2013.

[45] Janine Thoma, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Geometrically mappable image features. *IEEE Robotics and Automation Letters*, 5(2):2062–2069, 2020.

[46] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, 116(3):247–261, 2016.

[47] Akihiko Torii, Hajime Taira, Josef Sivic, Marc Pollefeys, Masatoshi Okutomi, Tomas Pajdla, and Torsten Sattler. Are large-scale 3d models really necessary for accurate visual localization? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[48] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13648–13657, 2022.

[49] Frederik Warburg, Soren Hauberg, Manuel López-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2626–2635, 2020.

[50] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian Triplet Loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12158–12168, 2021.

[51] Jun Yu, Chaoyang Zhu, Jian Zhang, Qingming Huang, and Dacheng Tao. Spatial pyramid-enhanced CNN with weighted triplet loss for place recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2):661–674, 2019.

[52] Mubariz Zaffar, Sourav Garg, Michael Milford, Julian Kooij, David Flynn, Klaus McDonald-Maier, and Shoaib Ehsan. VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision*, 129(7): 2136–2174, 2021.

[53] Mubariz Zaffar, Liangliang Nan, and Julian Francisco Pieter Kooij. CoPR: Toward accurate visual localization with con-

tinuous place-descriptor regression. *IEEE Transactions on Robotics*, 2023.

[54] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *Proceedings of the European Conference on Computer Vision*, pages 255–268. Springer, 2010.

[55] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2704–2712, 2015.

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

[57] Jianliang Zhu, Yunfeng Ai, Bin Tian, Dongpu Cao, and Sebastian Scherer. Visual place recognition in long-term and large-scale environment based on CNN feature. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1679–1685. IEEE, 2018.