Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization

Liu Liu [1,2], Hongdong Li [1,2] and Yuchao Dai [3]
[1] Australian National University, Canberra, Australia
[2] Australian Centre for Robotic Vision
[3] School of Electronics and Information, Northwestern Polytechnical University, Xian, China
{Liu.Liu; hongdong.li}@anu.edu.au; daiyuchao@nwpu.edu.cn

Abstract

This paper tackles the problem of large-scale image-based localization (IBL) where the spatial location of a query image is determined by finding out the most similar reference images *in a large database. For solving this problem, a critical task is to learn discriminative image representation that captures informative information relevant for localization. We propose a novel representation learning method having higher location-discriminating power.*

It provides the following contributions: 1) we represent a place (location) as a set of exemplar images depicting the same landmarks and aim to maximize similarities among intra-place images while minimizing similarities among inter-place images; 2) we model a similarity measure as a probability distribution on L_2-metric distances between intra-place and inter-place image representations; 3) we propose a new Stochastic Attraction and Repulsion Embedding (SARE) loss function minimizing the KL divergence between the learned and the actual probability distributions; 4) we give theoretical comparisons between SARE, triplet ranking [2] and contrastive losses [25]. It provides insights into why SARE is better by analyzing gradients.

Our SARE loss is easy to implement and pluggable to any CNN. Experiments show that our proposed method improves the localization performance on standard benchmarks by a large margin. Demonstrating the broad applicability of our method, we obtained the 3^{rd} place out of 209 teams in the 2018 Google Landmark Retrieval Challenge [1]. Our code and model are available at `https://github.com/Liumouliu/deepIBL`.

1. Introduction

The task of Image-Based Localization (IBL) is to estimate the geographic location of where a query image is taken, based on comparing it against geo-tagged images from a city-scale image database (*i.e.* a map). IBL has attracted considerable attention recently due to the wide-spread potential applications such as in robot navigation [20] and VR/AR [18, 39]. Depending on whether or not 3D point-clouds are used in the map, existing IBL methods can be roughly classified into two groups: *image-retrieval* based methods [2, 12, 30, 21, 40, 25] and *direct 2D-3D matching* based methods [27, 28, 14, 15, 5].

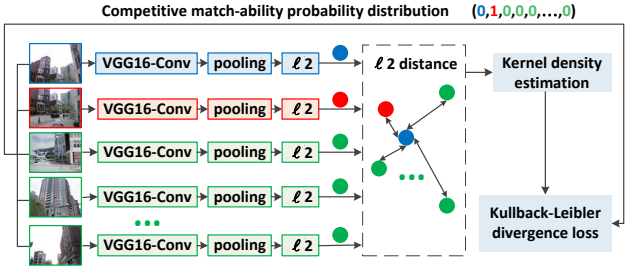


Figure 1: The pipeline of our method. We use the VGG16 net [35] with only convolution layers as our architecture. NetVLAD [2] pooling is used to obtain compact image representations. The feature vectors are post L_2 normalized. The L_2 distance between the query-positive and the query-negative images are calculated, and converted to a probability distribution. The estimated probability distribution is compared with the ground-truth match-ability distribution, yielding the Kullback-Leibler divergence loss.

This paper belongs to the *image-retrieval* group for its effectiveness at large scale and robustness to changing conditions [29]. For *image-retrieval* based methods, the main challenge is how to discriminatively represent images so that images depicting same landmarks would have similar representations while those depicting different landmarks would have dissimilar representations. The challenge is underpinned by the typically large-scale image database, in which many images may contain repetitive structures and similar landmarks, causing severe ambiguities.

Convolution Neural Networks (CNNs) have demonstrated great success for the IBL task [2, 12, 21, 8, 9, 25]. Typically, CNNs trained for image classification task are fine-tuned for IBL. As far as we know, all the state-of-the-art IBL methods focus on how to effectively aggregate a

CNN feature map to obtain discriminative image representation, but have overlooked another important aspect which can potentially boost the IBL performance markedly. The important aspect is how to effectively organize the aggregated image representations. So far, all state-of-the-art IBL methods use triplet ranking and contrastive embedding to supervise the representation organization process.

This paper fills this gap by proposing a new method to effectively organize the image representations (embeddings). We first define a "place" as a set of images depicting same location landmarks, and then directly enforce the intra-place image similarity and inter-place dissimilarity in the embedding space. Our goal is to cluster learned embeddings from the same place while separating embeddings from different places. Intuitively, we are organizing image representations using places as agents.

The above idea may directly lead to a multi-class classification problem if we can label the "place" tag for each image. Apart from the time-consuming labeling process, the formulation will also result in too many pre-defined classes and we need a large training image set to train the classification CNN net. Recently-proposed methods [40, 42] try to solve the multi-class classification problem using large GPS-tagged training dataset. In their setting, a class is defined as images captured from nearby geographic positions while disregarding their visual appearance information. Since images within the same class do not necessarily depict same landmarks, CNN may only learn high-level information [40] for each geographic position, thus inadequate for accurate localization.

Can we capture the intra-place image "attraction" and inter-place image "repulsion" relationship with limited data? To tackle the "attraction" and "repulsion" relationship, we formulate the IBL task as image similarity-based binary classification in feature embedding space. Specifically, the similarity for images in the same place is defined as 1, and 0 otherwise. This binary-partition of similarity is used to capture the intra-place "attraction" and inter-place "repulsion". To tackle the limited data issue, we use triplet images to train CNN, consisting of one query, positive (from the same place as the query), and negative image (from a different place). Note that a triplet is a minimum set to define the intra-place "attraction" and inter-place "repulsion".

Our CNN architecture is given in Fig. 1. We name our metric-learning objective as Stochastic Attraction and Repulsion Embedding (SARE) since it captures pairwise image relationships under the probabilistic framework. Moreover, our SARE objective can be easily extended to handle multiple negative images coming from different places, *i.e*. enabling competition with multiple other places for each place. In experiments, we demonstrate that, with SARE, we obtain improved performance on various IBL benchmarks. Validations on standard image retrieval benchmarks further justify the superior generalization ability of our method.

2. Related Work

There is a rich family of work in IBL. We briefly review CNN-based image representation learning methods. Please refer to [43, 44] for an overview.

While there have been many works [26, 8, 9, 25, 2, 21, 12, 30] in designing effective CNN feature map aggregation methods for IBL, they almost all exclusively use triplet or contrastive embedding objective to supervise CNN training. Both of these two objectives in spirit pulling the L_2 distance of matchable image pair while pushing the L_2 distance of non-matching image pair. While they are effective, we will show that our SARE objective outperforms them in the IBL task later. Three interesting exceptions which do not use triplet or contrastive embedding objective are the planet [42], IM2GPS-CNN [40], and CPlaNet [34]. They formulate IBL as a geographic position classification task. They first partition a 2D geographic space into cells using GPS-tags and then define a class per-cell. CNN training process is supervised by the cross-entropy classification loss which penalizes incorrectly classified images. We also show that our SARE objective outperforms the multi-class classification objective in the IBL task.

Although our SARE objective enforces intra-place image "attraction" and inter-place image "repulsion", it differs from traditional competitive learning methods such as Self-Organizing Map [13] and Vector Quantization [19]. They are both devoted to learning cluster centers to separate original vectors. No constraints are imposed on original vectors. Under our formulation, we directly impose the "attraction-repulsion" relationship on original vectors to supervise the CNN learning process.

3. Problem Definition and Method Overview

Given a large geotagged image database, the IBL task is to estimate the geographic position of a query image q. Image-retrieval based method first identifies the most visually similar image from the database for q, and then use the location of the database image as that of q. If the identified most similar image comes from the same place as q, then we deem that we have successfully localized q, and the most similar image is a positive image, denoted as p. If the identified most similar image comes from a different place as q, then we have falsely localized q, and the most similar image is a negative image, denoted as n.

Mathematically, an image-retrieval based method is executed as follows: First, query image and database images are converted to compact representations (vectors). This step is called image feature embedding and is done by a CNN network. For example, query image q is converted to

a fixed-size vector $f_\theta(q)$, where f is a CNN network and θ is the CNN weight. Second, we define a similarity function $S(\cdot)$ on pairwise vectors. For example, $S\left(f_\theta(q), f_\theta(p)\right)$ takes vectors $f_\theta(q)$ and $f_\theta(p)$, and outputs a scalar value describing the similarity between q and p. Since we are comparing the entire large database to find the most similar image for q, $S(\cdot)$ should be simple and efficiently computed to enable fast nearest neighbor search. A typical choice for $S(\cdot)$ is the L_2-metric distance, or functions monotonically increase/decrease with the L_2-metric distance.

Relying on feature vectors extracted by un-trained CNN to perform nearest neighbor search would often output a negative image n for q. Thus, we need to train CNN using easily obtained geo-tagged training images (Sec.7.1). The training process in general defines a loss function on CNN extracted feature vectors, and use it to update the CNN weight θ. State-of-the-art triplet ranking loss (Sec.4.1) takes triplet training images q, p, n, and imposes that q is more similar to p than n. Another contrastive loss (Sec.4.2) tries to separate $q \sim n$ pair by a pre-defined distance margin (see Fig.2). While these two losses are effective, we construct our metric embedding objective in a substantially different way.

Given triplet training images q, p, n, we have the prior knowledge that $q \sim p$ pair is matchable and $q \sim n$ pair is non-matchable. This simple match-ability prior actually defines a probability distribution. For $q \sim p$ pair, the match-ability is defined as 1. For $q \sim n$ pair, the match-ability is defined as 0. Can we respect this match-ability prior in feature embedding space? Our answer is yes. To do it, we directly fit a kernel on the L_2-metric distances of $q \sim p$ and $q \sim n$ pairs and obtain a probability distribution. Our metric-learning objective is to minimize the Kullback-Leibler divergence of the above two probability distributions (Sec.4.3).

What's the benefit of respecting the match-ability prior in feature embedding space? Conceptually, in this way, we capture the intra-place (defined by $q \sim p$ pair) "attraction" and inter-place (defined by $q \sim n$ pair) "repulsion" relationship in feature embedding space. Potentially, the "attraction" and "repulsion" relationship balances the embedded positions of the entire image database well. Mathematically, we use gradients of the resulting metric-learning objective with respect to triplet images to figure out the characteristics, and find that our objective adaptively adjusts the force (gradient) to pull the distance of $q \sim p$ pair, while pushing the distance of $q \sim n$ pair (Sec.5).

4. Deep Metric Embedding Objectives in IBL

In this section, we first give the two widely-used deep metric embedding objectives in IBL - the triplet ranking and contrastive embedding, and they are facilitated by minimizing the triplet ranking and contrastive loss, respectively. We then give our own objective - Stochastic Attraction and Repulsion Embedding (SARE).

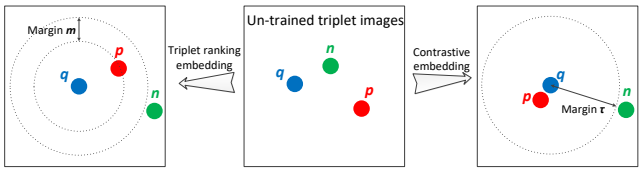


Figure 2: Triplet ranking loss imposes the constraint $\|f_\theta(q) - f_\theta(n)\|^2 > m + \|f_\theta(q) - f_\theta(p)\|^2$. Contrastive loss pulls the L_2 distance of $q \sim p$ pair to infinite-minimal, while pushing the L_2 distance of $q \sim n$ pair to at least τ-away.

4.1. Triplet Ranking Loss

The triplet ranking loss is defined by

$$L_\theta\left(q, p, n\right) = \max\left(0, m + \|f_\theta(q) - f_\theta(p)\|^2 - \|f_\theta(q) - f_\theta(n)\|^2\right), \tag{1}$$

where m is an empirical margin, typically $m = 0.1$ [3, 2, 8, 25]. m is used to prune out triplet images with $\|f_\theta(q) - f_\theta(n)\|^2 > m + \|f_\theta(q) - f_\theta(p)\|^2$.

4.2. Contrastive Loss

The contrastive loss imposes constraint on image pair $i \sim j$ by:

$$\begin{aligned} L_\theta\left(i, j\right) = & \frac{1}{2}\eta \left\|f_\theta(i) - f_\theta(j)\right\|^2 + \\ & \frac{1}{2}(1-\eta)\left(\max\left(0, \tau - \left\|f_\theta(i) - f_\theta(j)\right\|\right)^2\right) \end{aligned} \tag{2}$$

where for $q \sim p$ pair, $\eta = 1$, and for $q \sim n$ pair, $\eta = 0$. τ is an empirical margin to prune out negative images with $\|f_\theta(i) - f_\theta(j)\| > \tau$. Typically, $\tau = 0.7$ [25].

Intuitions to the above two losses are compared in Fig.2.

4.3. SARE-Stochastic Attraction and Repulsion Embedding

In this subsection, we present our Stochastic Attraction and Repulsion Embedding (SARE) objective, which is optimized to learn discriminative embeddings for each "place". A triplet images q, p, n define two places, one defined by $q \sim p$ pair and the other defined by n. The intra-place and inter-place similarity are defined in a probabilistic framework.

Given a query image q, the probability q picks p as its match is conditional probability $h_{p|q}$, which equals to 1 based on the co-visible or matchable prior. The conditional probability $h_{n|q}$ equals to 0 following above definition. Since we are interested in modeling pairwise similarities, we set $h_{q|q} = 0$. Note that the triplet probabilities $h_{q|q}, h_{p|q}, h_{n|q}$ actually define a probability distribution (summing to 1).

In the feature embedding space, we would like CNN extracted feature vectors to respect the above probability distribution. We define another probability distribution $c_{q|q}, c_{p|q}, c_{n|q}$ in the embedding space, and try to minimize the mismatch between the two distributions. The Kullback-Leibler divergence is employed to describe the cross-entropy loss and is given by:

$$\begin{aligned} L_\theta(q,p,n) &= h_{p|q}\log\left(\frac{h_{p|q}}{c_{p|q}}\right) + h_{n|q}\log\left(\frac{h_{n|q}}{c_{n|q}}\right) \\ &= -\log\left(c_{p|q}\right), \end{aligned} \tag{3}$$

In order to define the probability q picks p as its match in the feature embedding space, we fit a kernel on pairwise L_2-metric feature vector distances. We use three typical-used kernels to compare their effectiveness: Gaussian, Cauchy, and Exponential kernels. In next paragraphs, we use the Gaussian kernel to demonstrate our method. Loss functions defined by using Cauchy and Exponential kernels are given in Supplementary Material.

For the Gaussian kernel, we have:

$$c_{p|q} = \frac{\exp\left(-\left\|f_\theta(q) - f_\theta(p)\right\|^2\right)}{\exp\left(-\left\|f_\theta(q) - f_\theta(p)\right\|^2\right) + \exp\left(-\left\|f_\theta(q) - f_\theta(n)\right\|^2\right)}. \tag{4}$$

In the feature embedding space, the probability of q picks n as its match is given by $c_{n|q} = 1 - c_{p|q}$. If the embedded feature vectors $f_\theta(q)$ and $f_\theta(p)$ are sufficiently near, and $f_\theta(q)$ and $f_\theta(n)$ are far enough under the L_2-metric, the conditional probability distributions $c_{\cdot|q}$ and $h_{\cdot|q}$ will be equal. Thus, our SARE objective aims to find an embedding function $f_\theta(\cdot)$ that pulls the L_2 distance of $f_\theta(q) \sim f_\theta(p)$ to infinite-minimal, and that of $f_\theta(q) \sim f_\theta(n)$ to infinite-maximal.

Note that although ratio-loss [10] looks similar to our Exponential kernel $\exp(-||x-y||)$ defined loss function, they are theoretically different. The building block of ratio-loss is $\exp(||x-y||)$, and it directly applies $\exp()$ to distance $||x-y||$. This is problematic since it is not positive-defined (Please refer to Proposition 3&4 [32] or [31]).

5. Comparing the Three Losses

In this section, we illustrate the connections between the above three different loss functions. This is approached by deriving and comparing their gradients, which are key to the back-propagation stage in networks training. Note that gradient may be interpreted as the resultant force created by a set of springs between image pair [16]. For the gradient with respect to the positive image p, the spring pulls the $q \sim p$ pair. For the gradient with respect to the negative image n, the spring pushes the $q \sim n$ pair.

In Fig. 3, we compare the magnitudes of gradients with respect to p and n for different objectives. The mathematical equations of gradients with respect to p and n for different objectives are given in Table 1. For each objective, the gradient with respect to q is given by $\partial L/\partial f_\theta(q) = -\partial L/\partial f_\theta(p) - \partial L/\partial f_\theta(n)$.

In the case of triplet ranking loss, $\|\partial L/\partial f_\theta(p)\|$ and $\|\partial L/\partial f_\theta(n)\|$ increase linearly with respect to the distance $\|f_\theta(q) - f_\theta(p)\|$ and $\|f_\theta(q) - f_\theta(n)\|$, respectively. The saturation regions in which gradients equal to zero correspond to triplet images producing a zero loss (Eq. (1)). For triplet images producing a non-zero loss, $\|\partial L/\partial f_\theta(p)\|$ is independent of n, and vice versa. Thus, the updating of $f_\theta(p)$ disregards the current embedded position of n and vice versa.

For the contrastive loss, $\|\partial L/\partial f_\theta(p)\|$ is independent of n and increase linearly with respect to distance $\|f_\theta(q) - f_\theta(p)\|$. $\|\partial L/\partial f_\theta(n)\|$ decreases linearly with respect to distance $\|f_\theta(q) - f_\theta(n)\|$. The area in which $\|\partial L/\partial f_\theta(n)\|$ equals zero corresponds to negative images with $\|f_\theta(q) - f_\theta(n)\| > \tau$.

For all kernel defined SAREs, $\|\partial L/\partial f_\theta(p)\|$ and $\|\partial L/\partial f_\theta(n)\|$ depend on distances $\|f_\theta(q) - f_\theta(p)\|$ and $\|f_\theta(q) - f_\theta(n)\|$. The implicitly respecting of the distances comes from the probability $c_{p|q}$ (Eq. (4)). Thus, the updating of $f_\theta(p)$ and $f_\theta(n)$ considers the current embedded positions of triplet images, which is beneficial for the possibly diverse feature distribution in the embedding space.

The benefit of kernel defined SARE-objectives can be better understood when combined with hard-negative mining strategy, which is widely used in CNN training. The strategy returns a set of hard negative images (*i.e.* nearest negatives in L_2-metric) for training. Note that both the triplet ranking loss and contrastive loss rely on empirical parameters (m, τ) to prune out negatives (*c.f.* the saturation regions). In contrast, our kernel defined SARE-objectives do not rely on these parameters. They preemptively consider the current embedded positions. For example, hard negative with $\|f_\theta(q) - f_\theta(p)\| > \|f_\theta(q) - f_\theta(n)\|$ (top-left-triangle in gradients figure) will trigger large force to pull $q \sim p$ pair while pushing $q \sim n$ pair. "semi-hard" [33] negative with $\|f_\theta(q) - f_\theta(p)\| < \|f_\theta(q) - f_\theta(n)\|$ (bottom-right-triangle in gradients figure) will still trigger force to pull $q \sim p$ pair while pushing $q \sim n$ pair, however, the force decays with increasing $\|f_\theta(q) - f_\theta(n)\|$. Here, large $\|f_\theta(q) - f_\theta(n)\|$ may correspond to well-trained samples or noise, and the gradients decay ability has the potential benefit of reducing over-fitting.

To better understand the gradient decay ability of kernel defined SARE objectives, we fix $\|f_\theta(q) - f_\theta(p)\| = \sqrt{2}$, and compare $\|\partial L/\partial f_\theta(n)\|$ for all objectives in Fig. 4. Here, $\|f_\theta(q) - f_\theta(p)\| = \sqrt{2}$ means that for uniformly distributed feature embeddings, if we randomly sample $q \sim p$

Table 1: Comparison of gradients with respect to p and n for different objectives. Note that $\hat{c}_{p|q}$ and $\bar{c}_{p|q}$ are different from $c_{p|q}$ since they are defined by Cauchy and Exponential kernels, respectively. $\hat{c}_{p|q}$ and $\bar{c}_{p|q}$ share similar form as $c_{p|q}$.

Loss \ Gradients	$\partial L/\partial f_\theta(p)$	$\partial L/\partial f_\theta(n)$
Triplet ranking	$2\left(f_\theta(p)-f_\theta(q)\right)$	$2\left(f_\theta(q)-f_\theta(n)\right)$
Contrastive	$f_\theta(p)-f_\theta(q)$	$-\left(1-\tau/\left\|f_\theta(q)-f_\theta(n)\right\|\right)\left(f_\theta(q)-f_\theta(n)\right)$
Gaussian SARE	$2\left(1-c_{p\|q}\right)\left(f_\theta(p)-f_\theta(q)\right)$	$2\left(1-c_{p\|q}\right)\left(f_\theta(q)-f_\theta(n)\right)$
Cauchy SARE	$2\left(1-\hat{c}_{p\|q}\right)\frac{f_\theta(p)-f_\theta(q)}{1+\left\|f_\theta(p)-f_\theta(q)\right\|^2}$	$2\left(1-\hat{c}_{p\|q}\right)\frac{f_\theta(q)-f_\theta(n)}{1+\left\|f_\theta(q)-f_\theta(n)\right\|^2}$
Exponential SARE	$\left(1-\bar{c}_{p\|q}\right)\frac{f_\theta(p)-f_\theta(q)}{\left\|f_\theta(p)-f_\theta(q)\right\|}$	$\left(1-\bar{c}_{p\|q}\right)\frac{f_\theta(q)-f_\theta(n)}{\left\|f_\theta(q)-f_\theta(n)\right\|}$

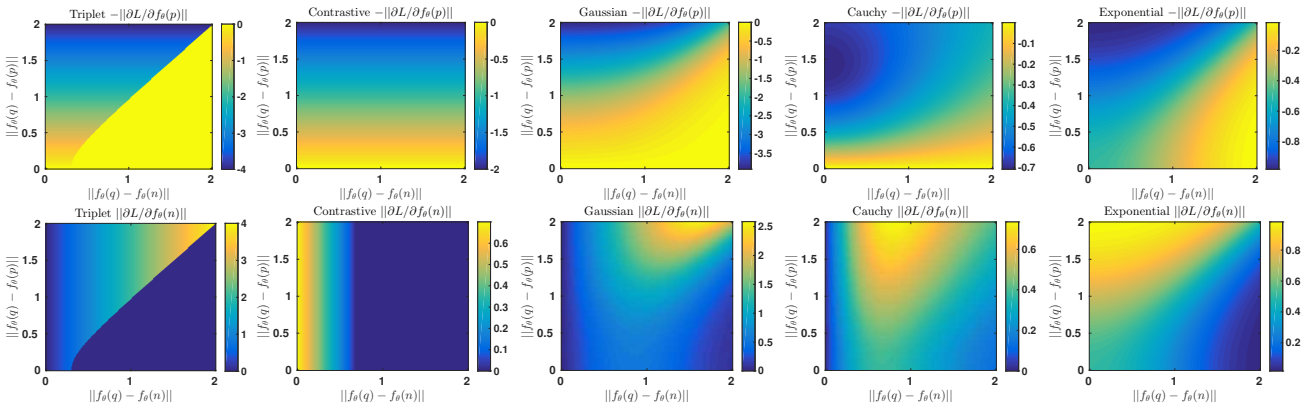


Figure 3: Comparison of gradients with respect to p and n for different objectives. $m = 0.1, \tau = 0.7$. (Best viewed in color on screen)

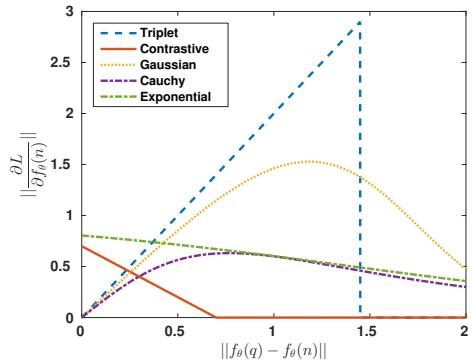


Figure 4: Comparison of the gradients with respect to n for different objectives. $m = 0.1, \tau = 0.7$.

pair, we are likely to obtain samples that are $\sqrt{2}$-away [17]. Uniformly distributed feature embeddings correspond to an initial untrained/un-fine-tuned CNN. For triplet ranking loss, Gaussian SARE and Cauchy SARE, $\|\partial L/\partial f_\theta(n)\|$ increases with respect to $\|f_\theta(q) - f_\theta(n)\|$ when it is small. In contrast to the gradually decay ability of SAREs, triplet ranking loss suddenly "close" the force when the triplet images produce a zero loss (Eq. (1)). For contrastive loss and Exponential SARE, $\|\partial L/\partial f_\theta(n)\|$ decreases with respect to $\|f_\theta(q) - f_\theta(n)\|$. Again, the contrastive loss "close" the force when the negative image produces a zero loss.

6. Handling Multiple Negatives

In this section, we give two methods to handle multiple negative images in CNN training stage. Equation (3) defines a SARE loss on a triplet and aims to shorten the embedded distance between the query and positive images while enlarging the distance between the query and negative images. Usually, in the task of IBL, the number of positive images is very small since they should depict same landmarks as the query image while the number of negative images is very big since images from different places are negative. At the same time, the time-consuming hard negative images mining process returns multiple negative images for each query image [2, 12]. There are two ways to handle these negative images: one is to treat them independently and the other is to jointly handle them, where both strategies are illustrated in Fig. 5.

Given N negative images, treating them independently results in N triplets, and they are substituted to Eq. (3) to calculate the loss to train CNN. Each triplet focuses on the competitiveness of two places (positive *VS* negative). The repulsion and attractive forces from multiple place pairs are averaged to balance the embeddings.

Jointly handling multiple negatives aims to balance the distance of positives over multiple negatives. In our formulation, we can easily construct an objective function to push

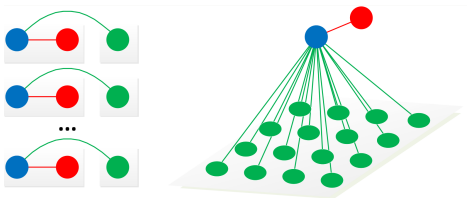


Figure 5: Handling multiple negative images. **Left**: The first method treats multiple negatives independently. Each triplet focuses on the competitiveness over two places, one defined by query ● and positive ●, and the other one defined by negative ●. **Right**: The second strategy jointly handles multiple negative images, which enables competitiveness over multiple places.

N negative images simultaneously. Specifically, the matchability priors for all the negative images are defined as zero, *i.e.* $h_{n|q} = 0, n = 1, 2, ..., N$. The Kullback-Leibler divergence loss over multiple negatives is given by:

$$L_\theta(q, p, n) = -\log\left(c^*_{p|q}\right), \tag{5}$$

where for Gaussian kernel SARE, $c^*_{p|q}$ is defined as:

$$c^*_{p|q} = \frac{\exp\left(-\|f_\theta(q) - f_\theta(p)\|^2\right)}{\exp\left(-\|f_\theta(q) - f_\theta(p)\|^2\right) + \sum_{n=1}^{N} \exp\left(-\|f_\theta(q) - f_\theta(n)\|^2\right)}. \tag{6}$$

The gradients of Eq. (5) can be easily computed to train CNN.

7. Experiments

This section mainly discusses the performance of SARE objectives for training CNN. We show that with SARE, we can improve the IBL performance on various standard place recognition and image retrieval datasets.

7.1. Implementation Details

Datasets. Google Street View Time Machine datasets have been widely-used in IBL [37, 2, 12]. It provides multiple street-level panoramic images taken at different times at close-by spatial locations on the map. The panoramic images are projected into multiple perspective images, yielding the training and testing datasets. Each image is associated with a GPS-tag giving its approximate geographic location, which can be used to identify nearby images not necessarily depicting the same landmark. We follow [2, 40] to identify the positive and negative images for each query image. For each query image, the positive image is the closest neighbor in the feature embedding space at its nearby geoposition, and the negatives are far away images. The above positive-negative mining method is very efficient despite some outliers may exist in the resultant positive/negative images. If accurate positives and negatives are needed, pairwise image matching with geometric validation [12] or SfM reconstruction [25] can be used. However, they are time-consuming.

The Pitts30k-training dataset [2] is used to train CNN, which has been shown to obtain best CNN [2]. To test our method for IBL, the Pitts250k-test [2], TokyoTM-val [2], 24/7 Tokyo [37] and Sf-0 [6, 30] datasets are used. To show the generalization ability of our method for image retrieval, the Oxford 5k [23], Paris 6k [24], and Holidays [11] datasets are used.

CNN Architecture. We use the widely-used compact feature vector extraction method NetVLAD [2, 21, 12, 30, 29] to demonstrate the effectiveness of our method. Our CNN architecture is given in Fig. 1.

Evaluation Metric. For the place recognition datasets Pitts250k-test [2], TokyoTM-val [2], 24/7 Tokyo [37] and Sf-0 [6], we use the Precision-Recall curve to evaluate the performance. Specifically, for Pitts250k-test [2], TokyoTM-val [2], and 24/7 Tokyo [37], the query image is deemed correctly localized if at least one of the top N retrieved database images is within $d = 25$ meters from the ground truth position of the query image. The percentage of correctly recognized queries (Recall) is then plotted for different values of N. For the large-scale Sf-0 [6] dataset, the query image is deemed correctly localized if at least one of the top N retrieved database images shares the same building IDs (manually labeled by [6]). For the image-retrieval datasets Oxford 5k [23], Paris 6k [24], and Holidays [11], the mean-Average-Precision (mAP) is reported.

Training Details. We use the training method of [2] to compare different objectives. For the state-of-the-art triplet ranking loss, the off-the-shelf implementation [2] is used. For the contrastive loss [25], triplet images are partitioned into $q \sim p$ and $q \sim n$ pairs to calculate the loss (Eq. (2)) and gradients. For our method which treats multiple negatives independent (*Our-Ind.*), we first calculate the probability $c_{p|q}$ (Eq. (4)). $c_{p|q}$ is then used to calculate the gradients (Table 1) with respect to the images. The gradients are back-propagated to train CNN. For our method which jointly handles multiple negatives (*Our-Joint*), we use Eq.(5) to train CNN. Our implementation is based on MatConvNet [38].

7.2. Kernels for SARE

To assess the impact of kernels on fitting the pairwise L_2-metric feature vector distances, we compare CNNs trained by Gaussian, Cauchy and Exponential kernel defined SARE-objectives, respectively. All the hyper-parameters are the same for different objectives, and the results are given in Fig. 6. CNN trained by Gaussian kernel defined SARE generally outperforms CNNs trained by others.

We find that handling multiple negatives jointly (*Gaussian-Joint*) leads to better training and validation per-

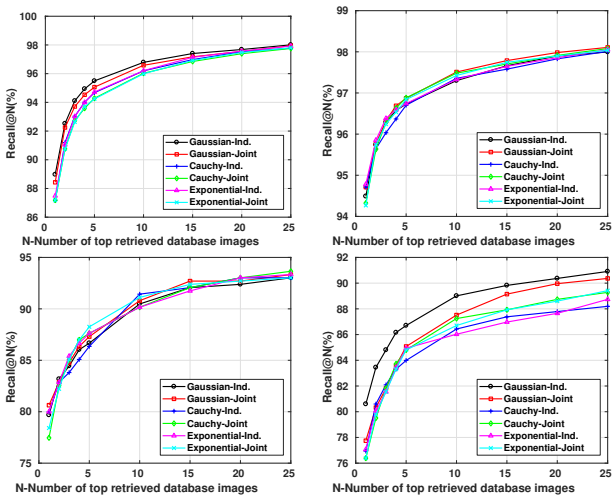


Figure 6: Comparison of recalls for different kernel defined SARE-objectives. From left to right and top to down: Pitts250k-test, TokyoTM-val, 24/7 Tokyo and Sf-0. (Best viewed in color on screen)

formances than handling multiple negatives independently (*Gaussian-Ind.*). However, when testing the trained CNNs on Pitts250k-test, TokyoTM-val, and 24/7 Tokyo datasets, the recall performances are similar. The reason may come from the negative images sampling strategy. Since the negative images are dropped randomly from far-away places from the query image using GPS-tags, they potentially are already well-balanced in the whole dataset, thus the repulsion and attractive forces from multiple place pairs are similar, leading to a similar performance of the two methods. *Gaussian-Ind.* behaves surprisingly well on the large-scale Sf-0 dataset.

7.3. Comparison with state-of-the-art

We use Gaussian kernel defined SARE objectives to train CNNs, and compare our method with state-of-the-art NetVLAD [2] and NetVLAD with Contextual Feature Reweighting [12]. The complete *Recall@N* performance for different methods are given in Table 2.

CNNs trained by Gaussian-SARE objectives consistently outperform state-of-the-art CNNs by a large margin on almost all benchmarks. For example, on the challenging 24/7 Tokyo dataset, *our-Ind.* trained NetVLAD achieves recall@1 of 79.68% compared to the second-best 75.20% obtained by CRN [12], *i.e.* a relative improvement in recall of 4.48%. On the large-scale challenging Sf-0 dataset, *our-Ind.* trained NetVLAD achieves recall@1 of 80.60% compared to the 75.58% obtained by NetVLAD [2], *i.e.* a relative improvement in recall of 5.02%. Note that we do not use the Contextual Reweighting layer to capture the "context" within images, which has been shown to be more effective than the original NetVLAD structure [12]. Similar improvements can be observed in other datasets. This confirms the important premise of this work: formulating the IBL problem in competitive learning framework, and using SARE to supervise the CNN training process can learn discriminative yet compact image representations for IBL. We visualize 2D feature embeddings of query images from 24/7 Tokyo and Sf-0 datasets. Images taken from the same place are mostly embedded to nearby 2D positions despite the significant variations in viewpoint, pose, and configuration.

7.4. Qualitative Evaluation

To visualize the areas of the input image which are most important for localization, we adopt [7] to obtain a heat map showing the importance of different areas of the input image. The results are given in Fig. 7. As can be seen, our method focuses on regions that are useful for image geo-localization while emphasizing the distinctive details on buildings. On the other hand, the NetVLAD [2] emphasizes local features, not the overall building style.

7.5. Generalization on Image Retrieval Datasets

To show the generalization ability of our method, we compare the compact image representations trained by different methods on standard image retrieval benchmarks (Oxford 5k [23], Paris 6k [24], and Holidays [11]) without any fine-tuning. The results are given in Table 3 . Comparing the CNN trained by our methods and the off-the-shelf NetVLAD [2] and CRN [12], in most cases, the mAP of our methods outperforms theirs'. Since our CNNs are trained using a city-scale building-oriented dataset from urban areas, it lacks the ability to understand the natural landmarks (*e.g.* water, boats, cars), resulting in a performance drop in comparison with the city-scale building-oriented datasets. CNN trained by images similar to images encountered at test time can increase the retrieval performance [4]. However, our purpose here is to demonstrate the generalization ability of SARE trained CNNs, which has been justified.

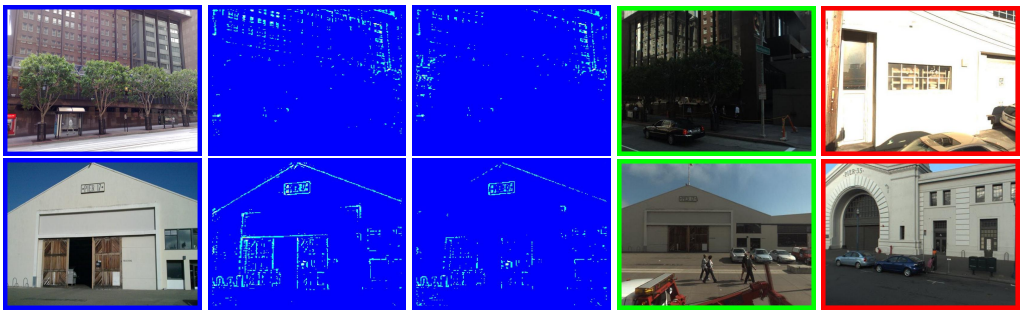
7.6. Comparison with Metric-learning Methods

Although deep metric-learning methods have shown their effectiveness in classification and fine-grain recognition tasks, their abilities in the IBL task are unknown. As another contribution of this paper, we show the performances of six current state-of-the-art deep metric-learning methods in IBL, and compare our method with : (1) Contrastive loss used by [25]; (2) Lifted structure embedding [22]; (3) N-pair loss [36]; (4) N-pair angular loss [41]; (5) Geo-classification loss [40]; (6) Ratio loss [10].

Fig. 8 shows the results of the quantitative comparison between our method and other deep metric learning methods. Our theoretically-grounded method outperforms the Contrastive loss [25] and Geo-classification loss [40], while remains comparable with other state-of-the-art methods.

Table 2: Comparison of Recalls on the Pitts250k-test, TokyoTM-val, 24/7 Tokyo and Sf-0 datasets.

Method \ Dataset	Pitts250k-test			TokyoTM-val			24/7 Tokyo			Sf-0		
	r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10
Our-Ind.	**88.97**	**95.50**	**96.79**	94.49	96.73	97.30	79.68	86.67	90.48	**80.60**	**86.70**	**89.01**
Our-Joint	88.43	95.06	96.58	**94.71**	**96.87**	97.51	**80.63**	**87.30**	**90.79**	77.75	85.07	87.52
CRN [12]	85.50	93.50	95.50	-	-	-	75.20	83.80	87.30	-	-	-
NetVLAD [2]	85.95	93.20	95.13	93.85	96.77	**97.59**	73.33	82.86	86.03	75.58	83.31	85.21



(a) Query (b) Our-heatmap (c) NetVLAD-heatmap (d) Our-top1 (e) NetVLAD-top1

Figure 7: Example retrieval results on Sf-0 benchmark dataset. From left to right: query image, the heat map of *Our-Ind*, the heat map of NetVLAD [2], the top retrieved image using our method, the top retrieved image using NetVLAD. Green and red borders indicate correct and incorrect retrieved results, respectively. (Best viewed in color on screen)

Table 3: Retrieval performance of CNNs on image retrieval benchmarks. No spatial re-ranking or query expansion is performed. The accuracy is measured by the mean Average Precision (mAP).

Method	Oxford 5K		Paris 6k		Holidays
	full	crop	full	crop	
Our-Ind.	**71.66**	**75.51**	**82.03**	81.07	80.71
Our-Joint	70.26	73.33	81.32	**81.39**	**84.33**
NetVLAD [2]	69.09	71.62	78.53	79.67	83.00
CRN [12]	69.20	-	-	-	-

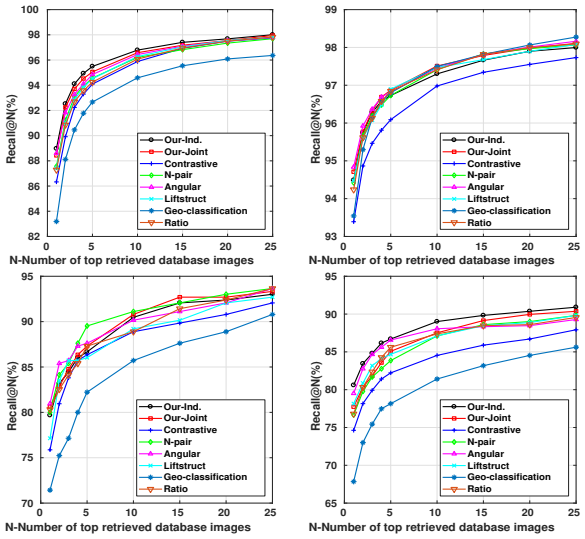


Figure 8: Comparison of recalls for deep metric learning objectives. From left to right and top to down: Pitts250k-test, TokyoTM-val, 24/7 Tokyo and Sf-0. (Best viewed in color on screen)

8. Conclusion

This paper has addressed the problem of learning discriminative image representations specifically tailored for the task of Image-Based Localization (IBL). We have proposed a new Stochastic Attraction and Repulsion Embedding (SARE) objective for this task. SARE directly enforces the "attraction" and "repulsion" constraints on intra-place and inter-place feature embeddings, respectively. The "attraction" and "repulsion" constraints are formulated as a similarity-based binary classification task. It has shown that SARE improves IBL performance, outperforming other state-of-the-art methods.

Acknowledgement

This research was supported in part by the Australian Research Council (ARC) grants (CE140100016), Australia Centre for Robotic Vision, and the Natural Science Foundation of China grants (61871325, 61420106007, 61671387, 61603303). Hongdong Li is also funded in part by ARC-DP (190102261) and ARC-LE (190100080). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU. We thank all anonymous reviewers for their valuable comments.

References

[1] Google landmark retrieval challenge leaderboard. https://www.kaggle.com/c/landmark-retrieval-challenge/leaderboard.

[2] 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.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[4] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.

[5] Toft Carl, Stenborg Erik, Hammarstrand Lars, Brynte Lucas, Pollefeys Marc, Sattler Torsten, and Kahl Fredrik. Semantic match consistency for long-term visual localization. *ECCV*, 2018.

[6] 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 *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 737–744. IEEE, 2011.

[7] Nassir Navab Federico Tombari Felix Grün, Christian Rupprecht. A taxonomy and library for visualizing learned features in convolutional neural networks. In *ICML Visualization for Deep Learning Workshop*, 2016.

[8] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257. Springer, 2016.

[9] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[10] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[11] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. *Computer Vision–ECCV 2008*, pages 304–317, 2008.

[12] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *CVPR*, 2017.

[13] Teuvo Kohonen. The self-organizing map. *Neurocomputing*, 21(1):1–6, 1998.

[14] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. In *European conference on computer vision*, pages 791–804. Springer, 2010.

[15] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[17] R Manmatha, Chao-Yuan Wu, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2859–2867. IEEE, 2017.

[18] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-dof localization on mobile devices. In *European conference on computer vision*, pages 268–283. Springer, 2014.

[19] José Muñoz-Perez, José Antonio Gómez-Ruiz, Ezequiel López-Rubio, and M Angeles Garcia-Bernal. Expansive and competitive learning for vector quantization. *Neural processing letters*, 15(3):261–273, 2002.

[20] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[21] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[22] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[23] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[24] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[25] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*, pages 3–20. Springer, 2016.

[26] Ali Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. A baseline for visual instance retrieval with deep convolutional networks. 4, 12 2014.

[27] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 667–674. IEEE, 2011.

[28] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2017.

[29] 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 *Proc. CVPR*, volume 1, 2018.

[30] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *CVPR 2017-IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[31] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[32] Bernhard Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[34] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. *ECCV*, 2018.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[37] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015.

[38] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

[39] Jonathan Ventura, Clemens Arth, Gerhard Reitmayr, and Dieter Schmalstieg. Global localization from monocular slam on a mobile phone. *IEEE transactions on visualization and computer graphics*, 20(4):531–539, 2014.

[40] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[41] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[42] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer, 2016.

[43] Yihong Wu. Image based camera localization: an overview. *CoRR*, abs/1610.03660, 2016.

[44] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1224–1244, 2018.