Learned Contextual Feature Reweighting for Image Geo-Localization

Hyo Jin Kim
UNC Chapel Hill
hyojin@cs.unc.edu

Enrique Dunn
Stevens Institute of Technology
edunn@stevens.edu

Jan-Michael Frahm
UNC Chapel Hill
jmf@cs.unc.edu

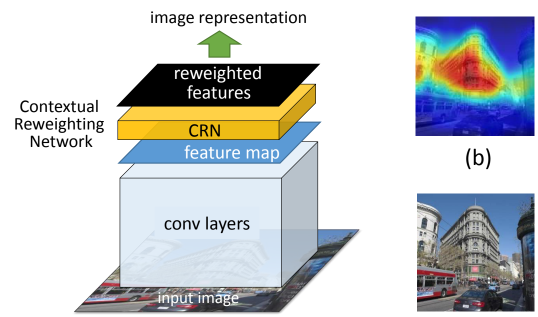
Abstract

We address the problem of large scale image geo-localization where the location of an image is estimated by identifying geo-tagged reference images depicting the same place. We propose a novel model for learning image representations that integrates context-aware feature reweighting in order to effectively focus on regions that positively contribute to geo-localization. In particular, we introduce a Contextual Reweighting Network (CRN) that predicts the importance of each region in the feature map based on the image context. Our model is learned end-to-end for the image geo-localization task, and requires no annotation other than image geo-tags for training. In experimental results, the proposed approach significantly outperforms the previous state-of-the-art on the standard geo-localization benchmark datasets.We also demonstrate that our CRN discovers task-relevant contexts without any additional supervision.

1. Introduction

Visual image geo-localization has been an active research area for the past decade [2, 28, 48, 53], owing to its wide range of applications including augmented reality [36], autonomous driving [33], adding and refining geo-tags in image collections [19, 65], large-scale 3D reconstruction [11], and photo editing [68].

Finding regions of interest has long been of great interest in computer vision. Much research has been done in the areas of feature selection, attention, and saliency [20, 63, 37]. Because task-relevant information is not generally uniformly distributed throughout an image [1], focusing on "interesting" areas, as opposed to "irrelevant" or even "distracting" areas, can often achieve better performance [51, 28, 13]. This is especially true for image geo-localization, where challenges come not only from photometric and geometric changes between the query and the database images, but also from confusing visual elements [29]. For instance, features extracted from time-varying objects such as pedestrians and trees, or ubiquitous objects, like vehicles and fences, can introduce misleading cues into the geo-localization process.



(a) (b) (c)

Figure 1. Image representation with contextual feature reweighting. (a) A contextual reweighting network takes convolutional features of a deep CNN as input to produce a spatial weighting mask (b) based on the learned contexts. The mask is used for weighted aggregation of input features to produce the representation of the input image (c).

To address this problem, there has been a recent push to intelligently select or reweight local features for image geo-localization [4, 49, 29, 28]. These methods focus on features with high distinctiveness in feature space [4] or in geographical space [49, 29]. Recently, Kim et al. [28] proposed a data-driven notion of "good" features as features that offer relatively high matching score to correct locations.

However, these approaches focus their analysis on individual local features in general. What is often overlooked is that a feature's usefulness depends largely on the context in the scene. For example, signage on buildings is useful for geo-localization, while signage on buses and t-shirts is misleading. There have been attempts to use top-down information such as semantic segmentation to restrict features to man-made structures [38], or repetitive structure detection to avoid over-counting of visual words in the bag-of-words representation [58]. We point out that such supervised priors are limited and do not capture all relevant contexts about which regions to focus on for image geo-localization.

Hence, we aim to provide contextual guidance for reweighting features in an unsupervised data-driven manner. To this end, we propose a novel end-to-end convolutional neural network (CNN) model for learning image representations that adaptively reweight features based on the image context. In particular, we introduce a *Contextual Reweighting Network* (CRN) that sits on the top of the convolutional layers in a standard deep architecture (Fig. 1 (a)). The CRN takes the feature maps of the base convolutional layers and estimates a weight for each feature based on its surrounding region (Fig. 1 (b)). By feature, we mean a column of activations at each spatial location of the feature maps. These weights are then applied to each feature as they are aggregated to produce an overall image representation.

We cast the image geo-localization problem as an image retrieval task and optimize the network with a triplet ranking loss based on the generated image representations. As a result, task-relevant contexts are discovered in an unsupervised manner, as the network learns in which context certain features should be emphasized or suppressed to better produce the spatial weighting, even though no ground-truth for the weighting nor the context information is provided. Visualizations of these learned contexts illustrate that the discovered contextual information contains rich high-level information that is not restricted to semantic cues, such as different types of buildings, vehicles, vegetation, and ground, but includes structural cues like lattice structure, different perspectives of buildings, and architectural styles.

Our training pipeline requires no training labels other than image geo-tags, which are commonly available. We propose to use geometric verification, that is, the verification of a valid two-view relationship of the views, and the convex hull of matched inlier points to generate positive reference images for training. In addition, we introduce efficient hard negative mining for image geo-localization which can be seen as mimicing the image geo-localization process within a training batch.

To summarize, our innovations are as follows: (1) We propose a novel end-to-end, fully-convolutional CNN for learning image representations that integrates context-aware feature reweighting. In particular, we introduce a contextual reweighting network that predicts weights for each region in the feature map based on its context. We experimentally validate that our pipeline significantly boosts the performance of the state-of-the-art methods. (2) We also show that unsupervised context discovery is acheived as a byproduct of training our network. The visualizations of these learned contexts illustrate that they capture rich high-level information. (3) We propose a training pipeline where only image geo-tags are required to automatically generate training data, with an efficient hard negative mining solution for image geo-localization.

2. Related Work

Image geo-localization has been cast as an image retrieval task [2, 10, 19, 28, 66], a 2D to 3D registration task [17, 21, 32, 47, 46], and a per-location classification task [16, 62]. In this paper, we address image geo-localization as an instance retrieval task, given an archive of previously localized images.

Retrieval-based solutions must be robust to a wide range of variability between the query and the reference dataset. Building image representations from local invariant features have been shown to be effective as they provide robustness to photometric and geometric changes [58, 4, 67, 31, 66]. However, not all content in the image is relevant to geo-localization [29, 28], requiring the discrimination of uninformative or misleading information. There have been attempts that try to address this problem by casting feature selection as a classification problem [28], reweighting features by computing their distinctiveness in feature space [4], and discarding features that appear in multiple geo-locations [29, 9, 45, 49]. There is also much work on feature selection and weighting for general image retrieval tasks [41, 55, 59, 69]. In contrast to these methods, which focus on local features themselves, we propose a model that adjusts feature weights based on contextual information, as it provides important cues about a feature's usefulness.

Recently, approaches were proposed to learn task-relevant features for image geo-localization and retrieval using end-to-end CNNs [2, 43, 56, 60, 34]. However, the contextual content discrimination was not addressed. In our work, we propose the incorporation of a context-adaptive feature preponderance into a CNN framework.

Our work is most closely related to the work of Gordo et al. [15] that uses a region proposal network [44] to learn which regions should be pooled to form a global image descriptor. Their method uses an explicit region proposal loss that requires bounding box annotations for training, in addition to a ranking loss for retrieval. In contrast, components in our network are optimized under one triplet ranking loss for image geo-localization. Furthermore, their work outputs regions of interest as a bounding box, whereas our proposed CRN produces a weighted mask that provides much more flexibility for focusing on relevant features.

Contextual information has been widely used for object recognition [39, 42, 64, 8, 13, 27]. However, the use of context has received significantly less attention for image geo-localization. Mousavian et al. [38], used semantic segmentation to filter out local features based on the notion that reliable features are likely to occur on buildings. Arandjelovic and Zisserman [5] also embedded semantic information to disambiguate matching between local features. Torii et al. [58], lowers the weights for features occurring in repetitive structure, as they violate feature independence and lead to over counting in bag-of-words representations. In our model, we learn the relevant contexts and generate weights for features in a strictly data-driven manner.

There have also been attempts to automatically generate training data for image retrieval for CNNs. Radenovic et al.

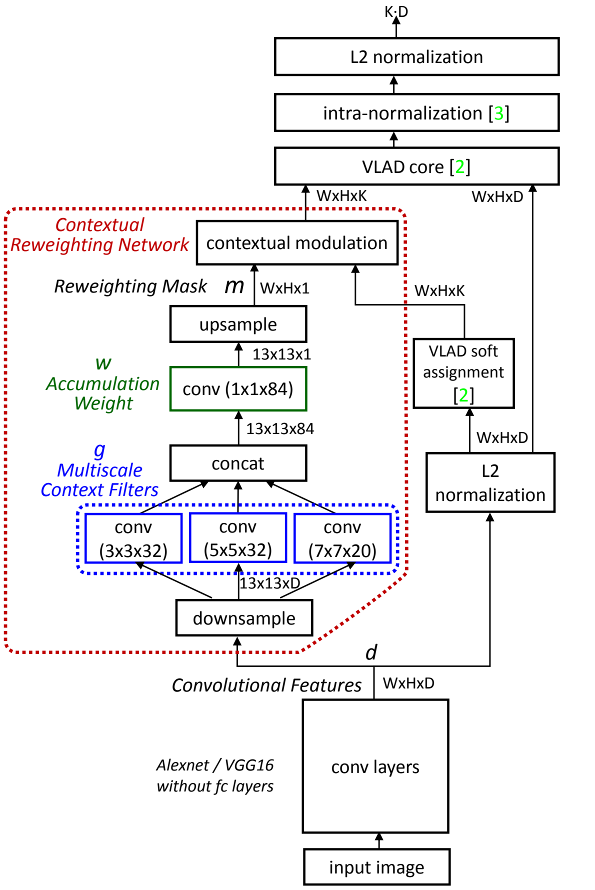


Figure 2. Overall network architecture. A CRN is a shallow network that takes the feature maps of convolutional layers as input and outputs a weighted mask indicating the importance of spatial regions in the feature maps. The resulting mask is used for performing context modulation for feature aggregation to create a global representation of the input image.

[43] exploit 3D reconstruction for selecting training data, and Gordo et al. [15] uses landmark graphs obtained from pairwise matching of images in the dataset. However, both their methods requires a dense image distribution in order to construct 3D models or scene graphs. We acquire training data similarly to the work of Arandjelović et al. [2], using GPS-tags. As geo-location itself is not sufficient to determine image overlap due to different camera orientations and occlusions, they choose positive images based on the current representation during learning. This depends on the quality of the learned representation, and does not contribute much to the current network status. On the other hand, we not only use GPS-tags, but use geometric verification to once verify positive images and refine them, thus reducing the memory and compute requirements. Also, while these methods [43, 2] perform periodic full retrieval for hard negative mining, we introduce within-batch hard negative mining for image geo-localization which is computationally less expensive. This is similar to within-batch hard negative mining in [61], and stochastic sampling used in [60]. However, we use GPS-tags to determine negatives.

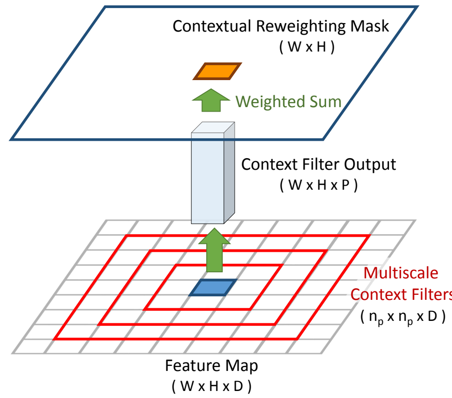


Figure 3. Contextual Reweighting Network. For each $1 \times 1 \times D$ convolutional feature, multi-scale contextual information is captured by P context filters with different window sizes ($n_p \times n_p \times D$). The filter output is then accumulated with learned weights to produce a reweighting value for the feaure in question.

3. Method

In this section, we first describe our model for learning image representations that integrate contextual reweighting of features (Sec. 3.1). We then illustrate our learning objective and the overall training process (Sec. 3.2).

3.1. Contextual Reweighting Network

In order to integrate a learned contextual reweighting, we begin with a standard representation approach and add an auxiliary context reweighting network (CRN). This network takes features produced by the original representation as input and outputs a spatial weighting over those features. The overall representation consists of a base network that generates mid-level features, a CRN, and a feature aggregation layer, as illustrated in Fig. 2.

Local Feature Representation: We treat activations of convolutional layers as local features. This has been shown to be effective for image geo-localization and image retrieval [56, 43, 6, 2, 15]. In particular, the $W \times H \times D$ dimensional feature maps of the last convolutional layer of the base network are treated as a set of D-dimensional local descriptors at $W \times H$ spatial locations. In our experiments, we used the conv5 output of AlexNet [30] and VGG16 [52]

Contextual Reweighting Mask: The CRN captures context information by using its hidden *context filters*, denoted as g_p, to explicitly look at $n_p \times n_p$ spatial windows around a local feature, as illustrated in Fig. 3. This is implemented using a convolution layer with kernel size n (Fig. 2). To obtain multi-scale contextual information, we use context filters with three different kernel sizes. These filters produce an activation map of $W \times H \times P$, where P is the total number of filters across scale.

The contextual reweighting mask m is computed as a weighted sum of these filter outputs, which is implemented with a $1 \times 1 \times P$ convolutional layer (Fig. 2) so that weights $\{w_p\}$ and the bias c are also learned during the training:

$$m = \sum_{p} w_p \cdot g_p(d) + c, \tag{1}$$

where d and $g_p(d)$ denote the feature maps and the output of filter g_p, respectively. Convolution layers in CRN are followed by ReLU non-linearity. The resulting weighting mask is of size $W \times H$, and the values in the mask indicate which spatial regions of the feature maps are important.

As our model is fully convolutional, it is not restricted by image size. However, it is difficult to train context filters on different image resolutions. To bypass this issue, we scale the spatial range of the feature maps to a fixed scale (e.g., 13×13) before applying the filters, and then rescale the computed contextual reweighting mask to the original feature map's spatial size. This provides image-level scale normalization, and is implemented with a matching downsampling (pooling)/upsampling layer pair which is applied at the beginning and end of the mask computation (Fig. 2).

Feature Aggregation via Contextual Modulation: Having a compact, fixed-length, global representation is necessary for efficient searching and limiting memory requirements. We utilize the reweighting mask from our CRN to generate a fixed-length image representation f for the query and the geo-tagged reference images. The contextual modulation layer (Fig. 2) in the CRN adjusts the impact of a local feature d_l at spatial location l to the global representation f based on the reweighting mask value m_l at that location.

In our experiments, we use the trainable vector of locally aggregated descriptors (VLAD) layer [2], which has been shown to be the state-of-the-art for place recognition and image retrieval tasks. The subvector of v that corresponds to the visual word k, denoted as v_k, is obtained as the accumulation of differences between local features d_l and the centroid c_k, weighted by the soft assignment a_l^k of d_l belonging to k, such that $v_k = \sum_{l \in R} a_l^k (d_l - c_k)$, where R denotes the set of spatial locations in the feature map.

Applying our context modulation, we obtain the reweighted VLAD representation f as follows:

$$f = [f_1, f_2, ..., f_K], \tag{2}$$

where

$$f_k = \sum_{l \in R} m_l \cdot a_l^k (d_l - c_k), \tag{3}$$

and K and R denote the number of visual words and a set of spatial locations in the feature map, respectively. This can be seen as a weighted pooling of features d_l. Following [2], f is intra-normalized [3], then L_2 normalized. The similarity between the two representations is computed as the inner product of the two. In this case, the contextual layer is implemented as a simple layer that takes m and performs element-wise multiplication across all channels of the soft assignment output a ($W \times H \times K$).

Note that these local features, when learned end-to-end, inhibit activations on task-irrelevant visual elements based on the local appearance. On the other hand, our contextual modulation produces a spatially varying weighting based on semi-global context. Therefore, our overall model can be seen as combining both top-down and bottom-up guidance to determine on which areas to focus.

3.2. Training

Training Objective: In our setting, the geo-location of a query image I_q is approximated by finding the nearest neighbor reference images $\{I_r\}$ in feature space. Thus, the objective for learning our image representation f is to ensure that matching reference images I_r^+ are closer to the query image than non-matching ones I_r^-. To this end, we use a triplet ranking loss [50, 60, 15]. During training, we provide image triplets to the network, each consisting of a training query image I_t, a positive reference image I_r^+, and a negative reference image I_r^-.

$$\begin{aligned} L_f(I_t, I_r^+, I_r^-) = \\ \max(0, \|f(I_t) - f(I_r^+)\|_2 - \|f(I_t) - f(I_r^-)\|_2 + \delta) \end{aligned} \tag{4}$$

By minimizing the triplet ranking loss (Eqn. (4)), the network learns in which context certain features should be emphasized or suppressed to better generate the spatial weighting mask m (Eqn. (1) and (3)), even though no ground-truth for the mask or the context is provided. The visualizations for the learned contexts and the masks are shown in Fig. 8 and Fig. 6, respectively (Sec. 4.3).

Training Data Generation: We used images with GPS-tags as our training query image set $\{I_t\}$. We collected 6K Flickr images with GPS-tags, and 17K Google Streetview Research Dataset images, both covering the same region as the reference images in the evaluation benchmark [10]. The images collected from Flickr are particularly challenging as they are unconstrained Internet photos that greatly vary from the reference images. Images from the Google Streetview Research Dataset are comparatively less challenging as they are also street view images taken from a vehicle, but they still differ significantly from reference images in terms of illumination, viewing angle, occlusion, and season. In order to further increase the size of our training data, we also added a randomly selected subset of $\{I_r\}$ to $\{I_t\}$. Standard data augmentation techniques such as random cropping and re-lighting [30] were applied.

Given an image set with only GPS-tags, we want to automatically generate image triplets $\{I_t, I_r^+, I_r^-\}$ for training. To verify positive images, we use geometric verification. Following Kim et al. [28], for each training query image

I_t, we define positive reference images I_r^+ as reference images that fall within 50m from the given GPS location and that pass geometric verification with respect to I_t by fitting a fundamental matrix [18] using RANSAC [12] over SIFT [35] matches. We select the top k_t images with the highest number of inliers. For I_r^-, we take the top k_t reference images that have the smallest distance to I_q based on the initialized image representation f (see implementation details) and are at least 225m away from the given GPS location, in order to mitigate possible geo-tag errors and lessen the effect of landmarks being visible from a large distance. The resulting I_r^+ and I_r^- are paired randomly to form k_t triplets $\{I_t, I_r^+, I_r^-\}$. In our experiments, k_t was set to 4.

To refine our training data, we perform ROI-based cropping and a scale test in order to account for $\{I_t, I_r^+\}$ pairs with small overlaps and large scale changes. A small overlap between I_t and I_r^+ could make the problem too difficult by giving the network misleading information that features on non-overlapping sides are not useful for image geo-localization. Therefore, we first perform a scale test using the area of the convex hull of the feature inliers from the geometric verification. Then we approximate the scale difference as the ratio of the areas of the two convex hulls. If they differ by more than a factor of 2, we exclude the triplet from the training set. Also, if a pair passes the scale test, but the difference is more than a factor of 1.5, we crop and rescale the area around the center of the convex hull of I_t. In addition to making our training data more robust, this framework also expands training data by generating more training triplets. After the refinement, we ended up with 36K training query images each producing four triplets, giving a total of 144K training triplets for the San Francisco city area.

Hard Negative Mining: While previous work performed periodic full retrieval to update hard negatives [2, 43], we propose an efficient way of mining hard negatives for image geo-localization. We mimic the image geo-localization process within the training batch. For every iteration, we perform image retrieval for I_q within the batch. We then select hard negatives I_r^- from the top retrieved images that are at least 225m away from the GPS location of I_t, similarly to how we selected initial I_r^-. As learning only based on the hardest negatives can lead to a bad local minima [50], we also select some I_r^-'s randomly from the batch. In our experiments we used two each of the hard and the random negatives, and average their triplet losses and the corresponding gradients. We used the accumulation of gradients as a proxy for having large batches, averaging gradients of 25 batches.

Implementation Detail: Following [2], we chose the number of centroids K in the VLAD representation to 64. The margin δ for the triplet ranking loss was set to 0.25. To initialize our model, we used Xavier random initialization [14] for our CRN. For the other layers, we used the parameters of NetVLAD models [2] fine-tuned on our data using the same training procedure. In practice, we found that it was crucial to train the base convolutional layers, CRN, and VLAD layer jointly for convergence. We used a learning rate of 0.005 for the CRN, and 0.0005 for other layers except conv1, which we fixed to its pretrained state. We used a batch size of 24, and trained for approximately 10 epochs. For VGG16-based models, we fixed accumulation weights w_p to 1. We used images with a resolution of 480×480. For testing, we averaged the image similarity computed by three patches (left, center, and right) similar to [30] for both our approach and NetVLAD, unless otherwise specified. Our implementation used Caffe [25].

4. Experiments

4.1. Image Geo-Localization

Evaluation Dataset: We used the San Francisco 1.2M benchmark dataset for image geo-localization from Chen et al. [10]. It consists of query images taken with different mobile cameras in various settings, and reference images taken from vehicle-mounted wide-angle cameras. The ground-truth annotations for correct matches for each test query image are given in the benchmark.

We also evaluated our method on Tokyo 24/7 [57] and Pittsburgh 250K test [2], where a retrieved image is deemed to be correct if it is within 25m from the ground-truth position of the query. We used the corresponding training and validation sets for these datasets, namely, Tokyo Time Machine data [2] and Pittsburgh 250K train/validation set [2].

Evaluation Metric: We follow the evaluation protocol of [58, 4, 10, 7, 2, 54], where performance is measured by the recall given the top N candidates in the shortlist. We also performed PCA whitening (learnt on the reference database for the San Francisco and on training images for Tokyo and Pittsburgh) on the obtained image representation $f(I)$ for both our method and NetVLAD [2], reducing the dimension by half. In all of our experiments, we did not use post-processing such as geometric re-ranking.

Results on San Francisco 1.2M benchmark [10]: To demonstrate the benefits of our context-aware image representation, we first compare our result to NetVLAD [2]. The only difference between our architecture and that of NetVLAD is the existence of our proposed CRN that performs contextual feature reweighting. Fig. 4 depicts the recall curves of our method and NetVLAD based on AlexNet [30] and VGG16 [52]. Our system with contextual feature reweighting consistently outperforms the systems without it. Our margin over NetVLAD at the top ($N = 1$) retrieved result is 4.8% for the AlexNet-based architecture, and 2.9% for the VGG16-based architecture. We implemented the VLAD layer proposed by Arandjelović et al. [2] using Caffe [25] and used it in both our method and NetVLAD. Both networks were optimized in the same training pipeline (Sec. 3.2) for a fair comparison.

Method	% Correct
Ours (VGG16)	**83.2**
ASMK* [54]	80.6
NetVLAD [2] fine-tuned (VGG16)	80.3
Ours (AlexNet)	**78.7**
DisLoc [4]	74.6
NetVLAD [2] fine-tuned (AlexNet)	73.9
HE-BURST [54]	71.9
Repttile [58]	65.4
NoGPS [10]	41.2
tf-idf [58]	23.2

Table 1. Proportion of correctly localized images at top 1

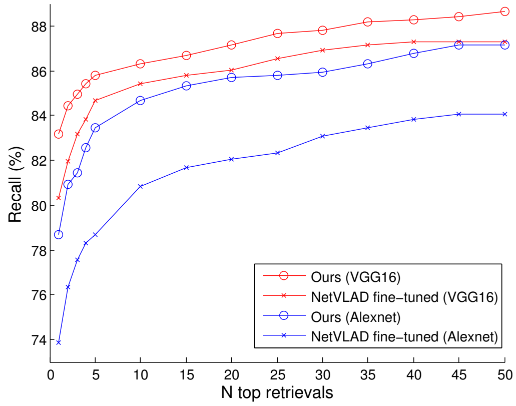


Figure 4. Recalls with and without contextual feature reweighting.

In Fig. 5, we compare our performance with other state-of-the-art methods. Our method achieves the best performance with a 83.2% recall at $N = 1$, exceeding the recall of the previous state-of-the-art [54] by 2.6%. The full comparison of recalls at the top of the shortlist ($N = 1$) with baseline methods is displayed in Table 1. The compared methods include binarized aggregated selective match kernel (ASMK*) [54], local distinctiveness based feature weighting (DisLoc) [4], repetitive feature re-weighting (Repttile) [58], Hamming embedding [24] with burstiness normalization [23] (HE-BURST) [22], vocaburary tree with histogram equalization (NoGPS) [10], and tf-idf weighting [40]. For ASMK*, DisLoc, Repttile, and NoGPS, we use the recall values reported by the authors. For HE-BURST, we used the recall reported in [54] using binary signatures of 128 bits. For tf-idf, we used the recalls reported in [58].

We show examples of our results in Fig. 6, where the top retrieved images for each query image are displayed for our method and the NetVLAD [2]. As can be seen, our method retrieves correct reference images despite the existence of confusing objects, such as trees and cars (Fig. 6 (a)-(c),(f)), focuses its attention on signage on stationary objects such as buildings (Fig. 6 (d)-(e), (j)-(k)), and distinguishes similar places with different details (Fig. 6 (e), (g)-(i), (l)-(n)).

To demonstrate the benefit of our learnt CRN, we also compared our method with Crow [26] which performs feature reweighting in a predefined way. Crow creates a spatial weighting mask by computing the L_2 norms of the features

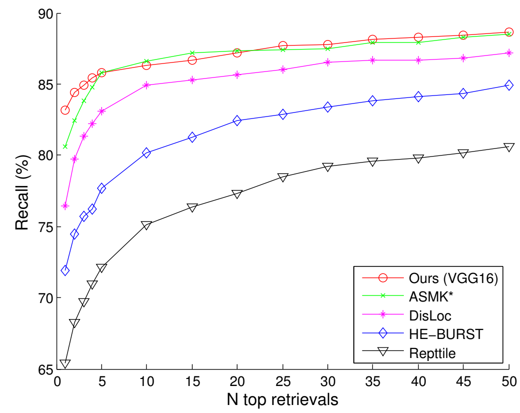


Figure 5. Comparison of recalls with the state-of-the-arts methods.

	top-1	top-5	top-10	top-25
CRN+NetVLAD (V)	**83.2**	**85.8**	**86.3**	**87.7**
CroW+NetVLAD (V)	80.1	84.3	85.3	86.5
CRN+NetVLAD (A)	**78.7**	**83.4**	**84.7**	**85.8**
CroW+NetVLAD (A)	74.1	79.3	80.8	82.3

Table 2. Comparison of our proposed CRN and CroW [26] with (V)GG16 and (A)lexnet base architectures.

data	set	method	top-1	top-5	top-10
Tokyo 24/7 [57]	all	**Ours**	**75.2**	**83.8**	**87.3**
		NetVLAD	71.8	82.5	86.4
	sunset /night	**Ours**	**66.7**	**76.7**	**81.9**
		NetVLAD	61.4	75.7	81.0
Pittsburgh 250K [58]	test [2]	**Ours**	85.5	**93.5**	**95.5**
		NetVLAD	**86.0**	93.2	95.1

Table 3. Recalls on Tokyo 24/7 [57] and Pittsburgh 250k test [2] datasets. All models are based on VGG16 architecture. For NetVLAD, we used the recalls reported by authors of [2]. We used the full resolution images for evaluation as in [2].

at each spatial location, which results in emphasizing regions with high activations. Table 2 shows the performance of NetVLAD when CRN is replaced with CroW for spatial reweighting, all of which underperform our proposed CRN.

Results on Tokyo 24/7 [57] and Pittsburgh 250K test [2]: Table 3 displays the results of our method evaluated on the Tokyo 24/7 and Pittsburgh 250K test datasets. We used the same training and validation sets as in [2], but with our training pipeline (Sec. 3.2). Our method consistently outperforms the state-of-the-art NetVLAD [2] on Tokyo 24/7, with a margin of 3.4% for all test images and 5.3% for the challenging sunset/night-time images at $N = 1$. Our performance on Pittsburgh 250K test data is similar to that of NetVLAD. We suspect this is due to lower variablity between the query and the reference images, in which case it may not be beneficial to down weight certain features. Also, while the query images for Pittsburgh 250K consists of randomly sampled streetviews, our training pipeline may have introduced a bias to the network by dropping training query images that do not pass geometric verification, utilizing only 72% of the training query set.

4.2. Comparison of the Emphasized Features

We qualitatively compare the emphasis on the features of our context-aware image representation and NetVLAD [2] in Fig. 7. We visualize the weighted mask generated by CRN for our method. To measure which regions of the feature maps were emphasized for NetVLAD, we computed the change in representation in Euclidean distance when leaving out each 1×1 spatial window in the convolutional feature maps. As can be seen, our method focuses on regions that are useful for image geo-localization while avoiding confusing visual elements. Moreover, it is capable of emphasizing the distinctive details on buildings. On the other hand, the NetVLAD [2] is inherently limited as it emphasizes local features independently; many features on confusing scene elements such as vegetation, pedestrians, and vehicles are emphasized.

4.3. Unsupervised Discovery of Contexts for Image Geo-Localization

To visualize the learned contexts, we display the image patches with the highest responses. That is, for each learned context filter g_p, we collect the strongest responses in each sampled image from the database. We crop out the square image patch from the original image at the center of the scaled feature map with the width of $n_p \frac{W_I}{W}$, where n_p, W_I, and W are kernel size of g_p, image width, and feature map width, respectively. Although the network is optimized only under the loss for image geo-localization, we observe that interesting contexts were captured by our contextual filters through the learning process. The results are shown in Fig. 8, where visualization of the context filters are aligned based on the sign of the accumulation weights w_p. If w_p is positive, it means the context captured contributes to assigning positive weights on the feature in question. The reverse is true for negative w_p. These contexts not only highlight semantic cues like buildings, vehicles, and pedestrians (Fig. 8 (a),(i-k),(o)), but also structural information such as the geometric changes in buildings (Fig. 8 (d-e)), sky lines (Fig. 8 (h)), architectural styles (Fig. 8 (b-c)), and buildings with signs (Fig. 8 (f-g)). Notably, even without supervision, our model assigns negative accumulation weights to lattice structures (Fig. 8 (n)), which is similar to what has previously been achieved with supervision [58] that lowers weights for features occurring in repetitive structures.

4.4. Image Retrieval

To assess generalizability of our approach, we evaluated our image representation using CRN trained on San Francisco on standard image retrieval benchmarks [40] without any fine-tuning. The results are shown in Table 4. We compared with NetVLAD [2] trained in the same pipeline as ours on San Francisco (SF), and the one that is trained on Pittsburgh (PGH) as reported in [2]. For all methods, we did not perform cropping of the ROI in the query, spatial re-ranking, or query expansion. Our model outperforms both representations. Especially, it consistently exceeds the mAP of NetVLAD trained on the same dataset by 2-4% margins.

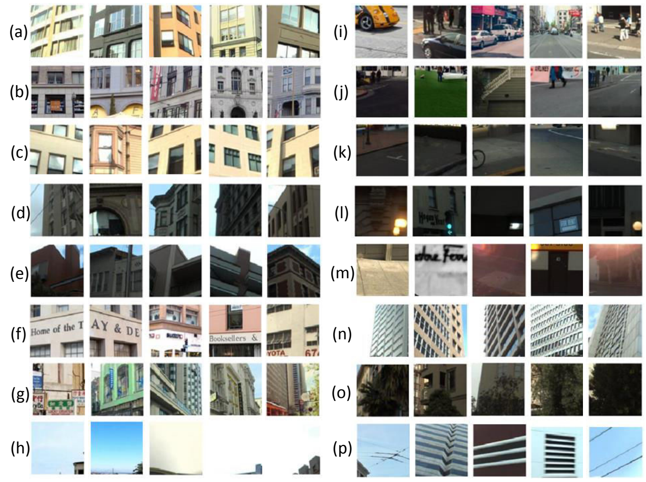


Figure 8. Discovered data-driven contexts for image geo-localization. For each learned context filters g_p, we display image patches with top responses (Sec. 4.3). (Left) Filters assigned positive weights $w_p > 0$. (Right) Filters assigned negative weights $w_p < 0$. Results are based on our AlexNet-based model.

	Oxford 5K [40]			Oxford 105K [40]	
Method	Ours	NetVLAD [2]		Ours	NetVLAD[2]
Train / Dim	SF	SF	PGH	SF	SF
16384	**0.704**	0.683	-	**0.685**	0.664
8192	**0.699**	0.682	-	**0.680**	0.660
4096	**0.692**	0.672	0.691	**0.671**	0.651
2048	**0.683**	0.660	0.677	**0.662**	0.633
1024	0.667	0.650	**0.669**	**0.644**	0.625
512	0.645	0.626	**0.656**	**0.622**	0.598
256	**0.642**	0.608	0.625	**0.617**	0.579
128	**0.615**	0.569	0.604	**0.586**	0.540

Table 4. Retrieval performance of our model trained on San Francisco on image retrieval benchmarks. No cropping of ROI in the query, spatial re-ranking, or query expansion was performed. The accuracy is measured by the mean Average Precision (mAP). All compared models are based on VGG16 architecture.

5. Conclusions

We introduced a novel Contextual Reweighting Network that learns image representations incorporating context-aware feature preponderance. We demonstrated that our CRN-based representation improves upon the existing state-of-the-art accuracy for geo-localization. The visualization of the outputs of our CRN shows that the relevant contexts for image geo-localization are captured as a byproduct of training our network. We also provide an efficient training pipeline only using geo-tags. Our proposed CRN can be combined with other feature aggregation methods, and can be applied to other problems such as object recognition.

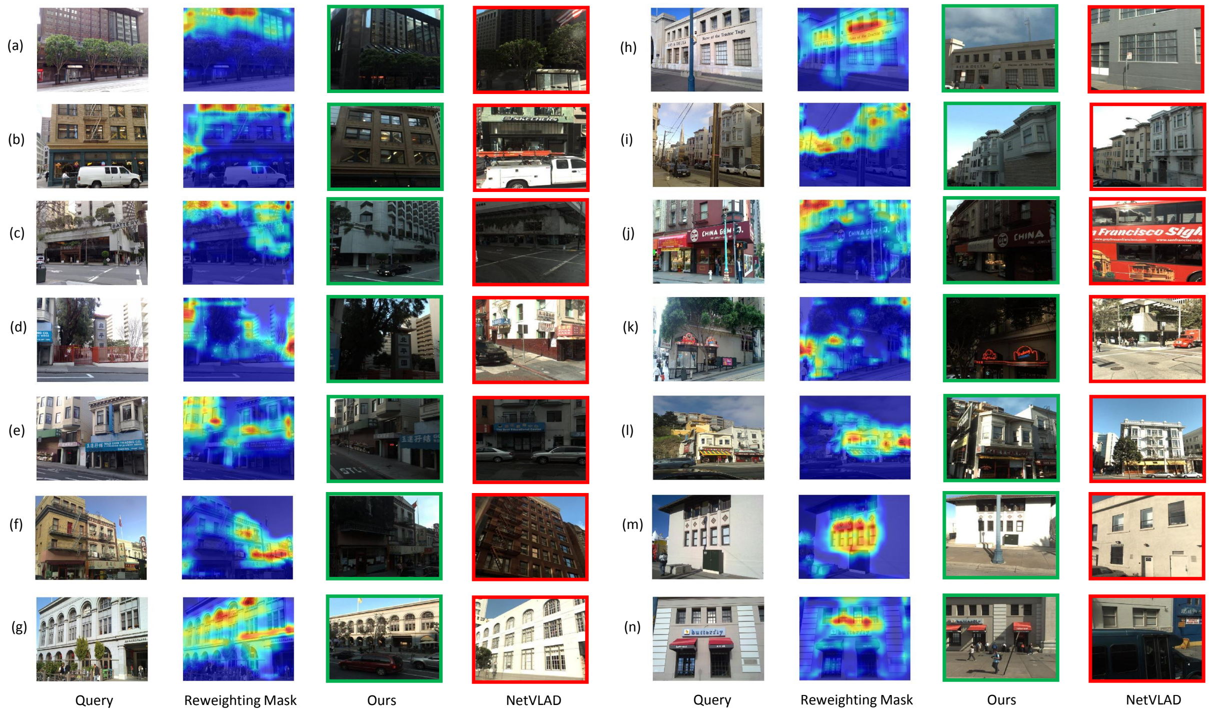


Figure 6. Example retrieval results on San Francisco benchmark dataset. From left to right: query image, our contextual reweighting mask in heat map, the top retrieved image using our method, the top retrieved image using NetVLAD [2]. Green and red borders indicate correct and incorrect retrieved results, respectively. Results are based on our AlexNet-based model.

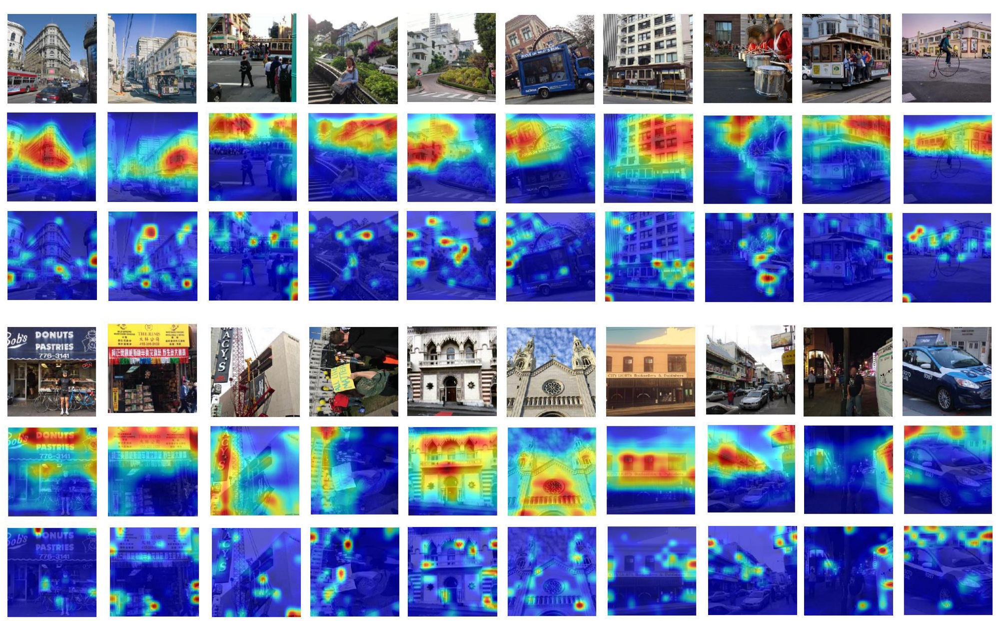


Figure 7. Comparison of emphasis on features. (top) Our contextual reweighting mask. (bottom) NetVLAD [2] emphasis on features. Both models are based on AlexNet architecture.

Acknowledgments Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government. The authors would also like to thank Relja Arandjelović and Akihiko Torii for providing data, code, and sharing insights, and Alex Berg for helpful discussions.

References

[1] A. Almahairi, N. Ballas, T. Cooijmans, Y. Zheng, H. Larochelle, and A. C. Courville. Dynamic capacity networks. In *CoRR*, 2015. 1

[2] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[3] R. Arandjelovic and A. Zisserman. All about vlad. In *CVPR*, 2013. 4

[4] R. Arandjelović and A. Zisserman. DisLocation: Scalable descriptor distinctiveness for location recognition. In *ACCV*, 2014. 1, 2, 5, 6

[5] R. Arandjelović and A. Zisserman. Visual vocabulary with a semantic twist. In *ACCV*, 2014. 2

[6] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson. From generic to specific deep representations for visual recognition. In *CVPRW*, 2015. 3

[7] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *ECCV*. 2012. 5

[8] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *CVPR*, 2016. 2

[9] S. Cao and N. Snavely. Graph-based discriminative learning for location recognition. In *CVPR*, 2013. 2

[10] D. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR*, 2011. 2, 4, 5, 6

[11] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *CVPR*, 2011. 1

[12] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 5

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *PAMI*, 2016. 1, 2

[14] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, 2010. 5

[15] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *ECCV*, 2016. 2, 3, 4

[16] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *CVPR*, 2013. 2

[17] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, and F. Wu. 3d visual phrases for landmark recognition. In *CVPR*, 2012. 2

[18] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5

[19] J. Hays and A. Efros. Large-scale image geolocalization. In *Multimodal Location Estimation of Videos and Images*. 2015. 1, 2

[20] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 1

[21] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009. 2

[22] M. Jain, H. Jégou, and P. Gros. Asymmetric hamming embedding: taking the best of our bits for large scale image search. In *ACMMM*, 2011. 6

[23] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*. IEEE, 2009. 6

[24] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 2010. 6

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, 2014. 5

[26] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional weighting for aggregated deep convolutional featuresoo. In *ECCVW*, 2016. 6

[27] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, 2016. 2

[28] H. J. Kim, E. Dunn, and J.-M. Frahm. Predicting good features for image geo-localization using per-bundle vlad. In *ICCV*, 2015. 1, 2, 4

[29] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *ECCV*. 2010. 1, 2

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 4, 5

[31] Y. Li, N. Snavely, and D. Huttenlocher. Location recognition using prioritized feature matching. In *ECCV*. 2010. 2

[32] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *ECCV*. 2012. 2

[33] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele. Real-time image-based 6-dof localization in large-scale environments. In *CVPR*, 2012. 1

[34] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In *CVPR*, 2015. 2

[35] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 5

[36] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-dof localization on mobile devices. In *ECCV*, 2014. 1

[37] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014. 1

[38] A. Mousavian, J. Košecka, and J.-M. Lien. Semantically guided location recognition for outdoors scenes. In *ICRA*, 2015. 1, 2

[39] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in cognitive sciences 11*, 2007. 2

[40] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 6, 7

[41] D. Qin, C. Wengert, and L. Van Gool. Query adaptive similarity for large scale object retrieval. In *CVPR*, 2013. 2

[42] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 2

[43] F. Radenović, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016. 2, 3, 5

[44] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[45] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys. Hyperpoints and fine vocabularies for large-scale location recognition. In *ICCV*, 2015. 2

[46] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *ICCV*. 2

[47] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, 2012. 2

[48] O. Saurer, G. Baatz, K. Köser, M. Pollefeys, et al. Image based geo-localization in the alps. *IJCV*, 2016. 1

[49] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007. 1, 2

[50] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. 2015. 4, 5

[51] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. Efros. Data-driven visual similarity for cross-domain image matching. In *ACMTOG*, 2011. 1

[52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3, 5

[53] A. Taneja, L. Ballan, and M. Pollefeys. Never get lost again: Vision based navigation using streetview images. In *ACCV*, 2014. 1

[54] G. Tolias, Y. Avrithis, and H. Jégou. Image search with selective match kernels: aggregation across single and multiple images. *IJCV*, 2016. 5, 6

[55] G. Tolias and H. Jégou. Visual query expansion with or without geometry: refining local descriptors by feature aggregation. *Pattern Recognition*, 2014. 2

[56] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*, 2016. 2, 3

[57] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 5, 6

[58] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual place recognition with repetitive structures. 2015. 1, 2, 5, 6, 7

[59] P. Turcot and D. G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV Workshops*, 2009. 2

[60] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 2, 3, 4

[61] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 3

[62] T. Weyand, I. Kostrikov, and J. Philbin. Planet-photo geolocation with convolutional neural networks. *ECCV*, 2016. 2

[63] D. Yoo, S. Park, J.-Y. Lee, A. Paek, and I. S. Kweon. Attentionnet: Aggregating weak directions for accurate object detection. In *ICCV*, 2015. 1

[64] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016. 2

[65] A. R. Zamir, S. Ardeshir, and M. Shah. Gps-tag refinement using random walks with an adaptive damping factor. In *CVPR*, 2014. 1

[66] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *ECCV*, 2010. 2

[67] A. R. Zamir and M. Shah. Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. *TPAMI*, 2014. 2

[68] C. Zhang, J. Gao, O. Wang, P. Georgel, R. Yang, J. Davis, J. Frahm, and M. Pollefeys. Personal photograph enhancement using internet photo collections. *TVCG*, 2014. 1

[69] C.-Z. Zhu, H. Jégou, and S. Satoh. Query-adaptive asymmetrical dissimilarities for visual object retrieval. In *ICCV*, 2013. 2