CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization

Sixing Hu Mengdan Feng Rang M. H. Nguyen Gim Hee Lee
National University of Singapore
{hu.sixing,fengmengdan}@u.nus.edu {nguyenho,gimhee.lee}@comp.nus.edu.sg

Abstract

The problem of localization on a geo-referenced aerial/satellite map given a query ground view image remains challenging due to the drastic change in viewpoint that causes traditional image descriptors based matching to fail. We leverage on the recent success of deep learning to propose the CVM-Net for the cross-view image-based ground-to-aerial geo-localization task. Specifically, our network is based on the Siamese architecture to do metric learning for the matching task. We first use the fully convolutional layers to extract local image features, which are then encoded into global image descriptors using the powerful NetVLAD. As part of the training procedure, we also introduce a simple yet effective weighted soft margin ranking loss function that not only speeds up the training convergence but also improves the final matching accuracy. Experimental results show that our proposed network significantly outperforms the state-of-the-art approaches on two existing benchmarking datasets. Our code and models are publicly available on the project website[1].

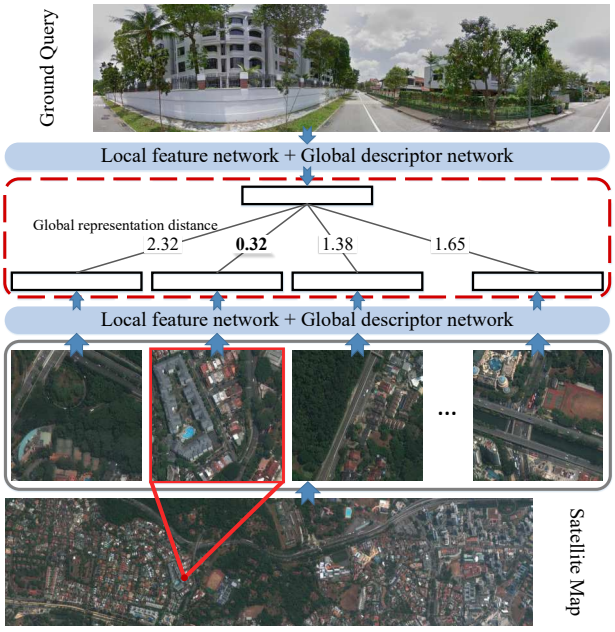

Figure 1. An illustration of the image based ground-to-aerial geo-localization problem, and our proposed framework.

1. Introduction

Image-based geo-localization has drawn a lot of attention over the past years in the computer vision community due to its potential applications in autonomous driving [26] and augmented reality [27]. Traditional image-based geo-localization is normally done in the context where both the query and geo-tagged reference images in the database are taken from the ground view [16, 52, 33, 45]. One of the major drawbacks of such approaches is that the database images, which are commonly obtained from crowd-sourcing, e.g. geo-tagged photos from Flickr etc, usually do not have a comprehensive coverage of the area. This is because the photo collections are most likely to be biased towards famous touristy areas. Consequently, ground-to-ground geo-localization approaches tend to fail in locations where reference images are not available. In contrast, aerial imagery taken from devices with bird's eye view, e.g. satellites and drones, densely covers the Earth. As a result, matching ground view photos to aerial imagery gradually becomes an increasingly popular geo-localization approach [5, 22, 37, 23, 49, 50, 46, 41, 53, 43]. However, cross-view matching still remains challenging because of the drastic change in viewpoint between ground and aerial images. This causes cross-view matching with traditional handcrafted features like SIFT [25] and SURF [7] to fail.

With the recent success of deep learning in many computer vision tasks, most of the existing works on cross-view image matching [49, 50, 46, 53] adopt the convolutional neural network (CNN) to learn representations for matching between ground and aerial images. To compensate for

[1] https://github.com/david-husx/crossview_localisation.git

the large viewpoint difference, Vo and Hays [46] use an additional network branch to estimate the orientation and utilize multiple possible orientations of the aerial images to find the best angle for matching across the two views. This approach causes significant overhead in both training and testing. In contrast, our work avoids the overhead by making use of the global VLAD descriptor that was shown to be invariant against large viewpoint and scene changes in the place recognition task [18]. Specifically, we embed the NetVLAD layer [3] on top of a CNN to extract descriptors that are invariant against large viewpoint changes. Figure 1 shows an illustration of our approach. The key idea is that NetVLAD aggregates the local features obtained from the CNN to form global representations that are independent of the locations of the local features.

Contributions In this paper, we propose a powerful network architecture: the CVM-Net for cross-view image-based ground-to-aerial geo-localization. Specifically, we combine NetVLAD layers with a Siamese network to jointly learn robust representations for cross-view image matching. Our CVM-Net learns local features and forms global descriptors that are invariant to large viewpoint change for ground-to-aerial geo-localization. As part of the training procedure, we also introduce a new weighted soft margin ranking loss that not only speeds up the training convergence but also improves the final retrieval accuracy. In addition, this new weighted soft margin can be embedded in both the triplet and quadruplet losses. Our extensive experiment results show that the proposed framework significantly outperform all state-of-the-art methods, especially on the panoramic CVUSA dataset [53].

2. Related Work

Most of the existing works on estimating the geographical location of a query ground image used image matching or image retrieval techniques. These works can be categorized based on the type of features.

Hand-crafted features In the early stage, traditional features that were commonly used in the computer vision community were utilized to do the cross-view image matching [30, 6, 35, 36, 22, 44]. However, due to the huge difference in viewpoint, the aerial image and ground view image of the same location appeared to be very different. This caused direct matching with traditional local features to fail. Therefore, a number of approaches warped the ground image to the top-down view to improve feature matching [30, 35, 44]. In cases where the aerial image was oblique where building facades are visible, geo-localization could be achieved with facade patch-matching [6].

Learnable features As the deep learning approaches were proven to be extremely successful in image/video classification and recognition tasks, some efforts were taken to introduce deep learning into the domain of cross-view image matching and retrieval. Workman and Jacobs [49] conducted experiments on the AlexNet [21] model trained on ImageNet [13] and Places [54]. They showed that deep features for common image classification significantly outperformed hand-crafted features. Later on, Workman et al. [50] further improved the matching accuracy by training the convolutional neural network on aerial branch. Vo and Hays [46] conducted thorough experiments on existing classification and retrieval networks, including binary classification network, Siamese network and Triplet network. With the novel soft-margin triplet loss and exhausting mini-batch training strategy, they achieved a significant improvement on the retrieval accuracy. On the other hand, Zhai et al. [53] proposed a weakly supervised training network to obtain the semantic layout of satellite images. These layouts were used as image descriptors to do retrieval from database.

The most important part of image retrieval is to find a good descriptor of an image which is discriminative and fast for comparison. Sivic and Zisserman [39, 40] proposed the Bag-of-Visual-Word descriptors to aggregate a set of local features into a histogram of visual words, i.e. the global descriptor. They showed that the descriptor was partially viewpoint and occlusion invariant, and outperformed local feature matching. Nister and Stewenius [29] created a tree structure vocabulary to support more visual words. Jegou et al. [18] proposed VLAD descriptor. Instead of histogram, they aggregated the residuals of the local features to cluster centroids. Based on that work, Arandjelovic et al. [3] proposed a learnable layer of VLAD, i.e. NetVLAD, that could be embedded into the deep network for end-to-end training. In their extended paper [4], they illustrated that NetVLAD was better than multiple fully connected layers, max pooling and VLAD. Due to the superior performance of NetVLAD, we adopt the NetVLAD layer in our proposed network.

Image retrieval loss Our work is also related to metric learning via deep networks. The most widely used loss function in image retrieval task is the max-margin triplet loss [9, 3, 8, 11, 24, 42, 47, 14, 32, 19] that enforces the distances of positive pairs to be less than the distances of negative pairs. The work in [17] concluded that this margin value required to be carefully selected. To overcome this issue, Vo and Hays [46] proposed a soft-margin triplet loss which was proven to be effective [17]. Since the triplet loss has no constraint on irrelevant pairs, it will cause the inter-class variation to be small when decreasing the intra-class variation during training. To alleviate this problem, the quadruplet [10] and angular [48] losses were proposed to further improve the training of triplet network and the performance of image retrieval.

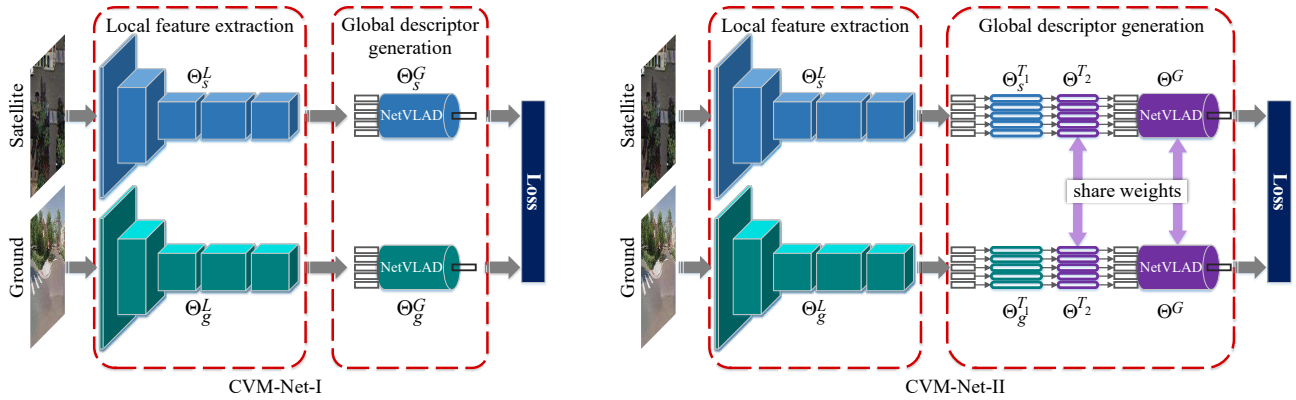

Figure 2. Overview of our proposed CVM-Nets. **CVM-Net-I**: The deep network with two aligned (no weight-shared) NetVLADs which are used to pool the local features from different views into a common space. **CVM-Net-II**: The deep network with two weight-shared NetVLADs that transform the local features into a common space before aggregating to obtain the global descriptors.

3. Our Approach

Similar to the existing works on image-based ground-to-aerial geo-localization [50, 46, 53], our goal is to find the closest match of a query ground image from a given database of geo-tagged satellite images (see Figure 1), i.e. cross-view image retrieval. To this end, we propose the CVM-Net.

3.1. System Overview

To learn the joint relationship between satellite and ground images, we adopt the Siamese-like architecture that has been shown to be very successful in image matching and retrieval tasks. In particular, our framework contains two network branches of the same architecture. Each branch consists of two parts: local feature extraction and global descriptor generation. In the first part, CNNs are used to extract the local features. See Section 3.2 for the details. In the second part, we encode the local features into a global descriptor that is invariant across large viewpoint changes. Towards this goal, we adopt the VLAD descriptor by embedding NetVLAD layers on top of each CNN branch. See Section 3.3 for the details.

3.2. Local Feature Extraction

We use a fully convolutional network (FCN) f^L to extract local feature vectors of an image. For a satellite image I_s, the set of local features is given by $U_s = f^L(I_s; \Theta_s^L)$, where Θ_s^L is the parameters of the FCN of the satellite branch. For a ground image I_g, the set of local features $U_g = f^L(I_g; \Theta_g^L)$, where Θ_g^L is the parameters of the FCN of the ground view branch. In this work, we compare the results of our network using the convolutional part of AlexNet [21] and VGG16 [38] as f^L. Details of the implementation and comparison are shown in Section 5.

3.3. Global Descriptor Generation

We feed the set of local feature vectors obtained from the FCN into a NetVLAD layer to get the global descriptor. NetVLAD [3] is a trainable deep network version of VLAD [18], which aggregates the residuals of the local feature vectors to their respective cluster centroid to generate a global descriptor. The centroids and distance metrics are trainable parameters in NetVLAD. In this paper, we try two strategies, i.e. CVM-Net-I and CVM-Net-II, to aggregate local feature vectors from the satellite and ground images into their respective global descriptors that are in a common space for similarity comparison.

CVM-Net-I: Two independent NetVLADs As shown in Figure 2, we use a separate NetVLAD layer for each branch to generate the respective global descriptors of a satellite and ground image. The global descriptor of an image can be formulated as $v_i = f^G(U_i; \Theta_i^G)$, where $i \in \{s, g\}$ represents the satellite or ground branch. There are two groups of parameters in Θ_i^G - (1) K cluster centroids $C_i = \{c_{i,1}, ..., c_{i,K}\}$, and (2) a distance metric $W_{i,k}$ for each cluster. The number of clusters in both NetVLADs are set to be same. Each NetVLAD layer produces a VLAD vector, i.e. global descriptor, for the respective views v_s and v_g that are in the same space, which can then be used for direct similarity comparison. More details are given in the next paragraph. To keep computational complexity low, we reduce the dimension of the VLAD vectors before feeding them into the loss function for end-to-end training, or using them for similarity comparison.

In addition to the discriminative power, the two NetVLAD layers with the same number of clusters that are trained together in a Siamese-like architecture, are able to output two VLAD vectors that are in a common space. Given a set of local feature vectors $U = \{u_1, ..., u_N\}$ (we drop the index i in U_i for brevity), the k^{th} element of the

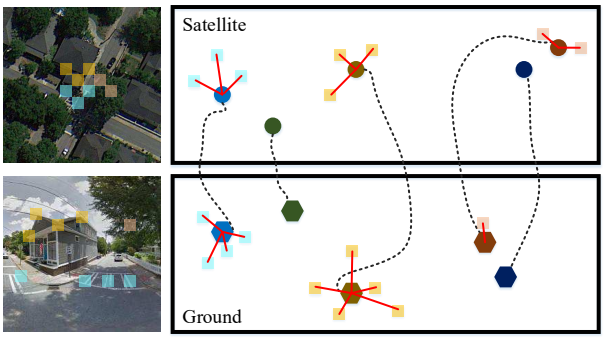

Figure 3. An illustration of how NetVLAD achieves cross-view matching. (Top): satellite view, (Bottom): ground view. In each view, there are a set of local features (colorful squares) and their associated centroids (hexagons and circles). After training, each centroid of satellite view is associated with the unique centroid of ground view (dotted lines). The residuals (red lines) are independent to their own views and comparable to the other view because they are only relative to the centroids. Thus, the global descriptors, i.e. aggregated residuals, of two views are in the common space.

VLAD vector V is given by

$$V(k) = \sum_{j=1}^{N} \bar{a}_k(u_j)(u_j - c_k), \tag{1}$$

where $\bar{a}_k(u_j)$ is the soft-assignment weight determined by the distance metric parameters and input local feature vectors. Refer to [3] for more details of $\bar{a}_k(u_j)$. As shown in Equation 1, the descriptor vector of each centroid is the summation of residuals to the centroid. The residuals to the centroids of two views are in a new common space, independent to the domain of two centroids. Therefore, they can be regarded as in a common "residual" space with respect to the pair of centroids in two views. The comparison of satellite and ground view descriptors is the centroid-wise comparison. It makes the VLAD descriptors of two views comparable. Figure 3 shows an illustration of this concept.

The complete model of our CVM-Net-I is shown in Figure 2. The global descriptor of the satellite image is given by $v_s = f^G(f^L(I_s;\Theta_s^L);\Theta_s^G)$ and ground image is given by $v_g = f^G(f^L(I_g;\Theta_g^L);\Theta_g^G)$. The two branches have identical structures with different parameters. Finally, the dimensions of the global descriptors from the two views are reduced by a fully connected layer.

CVM-Net-II: NetVLADs with shared weights Instead of having two independent networks of similar structure in CVM-Net-I, we propose a second network - CVM-Net-II with some shared weights across the Siamese architecture. Figure 2 shows the architecture of our CVM-Net-II. Specifically, the CNN layers for extracting local features U_s and U_g remain the same. These local features are then passed through two fully connected layers - the first layer with independent weights $\Theta_s^{T_1}$ and $\Theta_g^{T_1}$, and the second layer with shared weights Θ^{T_2}. The features U'_s and U'_g after the two fully connected layers are given by

$$u'_{s,j} = f^T(u_{s,j};\Theta_s^{T_1},\Theta^{T_2}), \tag{2a}$$
$$u'_{g,j} = f^T(u_{g,j};\Theta_g^{T_1},\Theta^{T_2}). \tag{2b}$$

where $u_{s,j} \in U_s$, $u_{g,j} \in U_g$ and $u'_{s,j} \in U'_s$, $u'_{g,j} \in U'_g$.

Finally, the transformed local features are fed into the NetVLAD layers with shared weights Θ^G. The global descriptors of the satellite and ground images are given by

$$v_s = f^G(U'_s;\Theta^G), \tag{3a}$$
$$v_g = f^G(U'_g;\Theta^G). \tag{3b}$$

The complete model of our CVM-Net-II is illustrated in Figure 2. We adopted weight sharing in our CVM-Net-II network because weight sharing has been proven to improve metric learning in many of the Siamese network architectures [12, 34, 15, 51, 31].

4. Weighted Soft-Margin Ranking Loss

The triplet loss is often used as the objective function to train deep networks for image matching and retrieval tasks. The goal of the triplet loss is to learn a network that brings positive examples closer to a chosen anchor point than the negative examples. The simplest triplet loss is the max-margin triplet loss: $\mathcal{L}_{max} = max(0, m + d_{pos} - d_{neg})$, where d_{pos} and d_{neg} are the distances of all the positive and negative examples to the chosen anchor. m is the margin and it has been shown in [17] that m has to be carefully selected for best results. A soft-margin triplet loss was proposed in [46, 17] to avoid the need to determine the margin in the triplet loss: $\mathcal{L}_{soft} = ln(1 + e^d)$, where $d = d_{pos} - d_{neg}$. We use the soft-margin triplet loss to train our CVM-Nets, but noted that this loss resulted in slow convergence. To improve the convergence rate, we propose a weighted soft-margin ranking loss which scales d in $\mathcal{L}_{soft}$ by a coefficient α:

$$\mathcal{L}_{weighted} = ln(1 + e^{\alpha d}). \tag{4}$$

Our weighted soft-margin ranking loss becomes the soft-margin triplet loss when $\alpha = 1$. We made the observation through experiments that the rate of convergence and results improve as we increase α. The gradient of the loss increases with α, which might cause the network to improve the weights faster so as to reduce the larger errors.

Our proposed loss can also be embedded into other loss functions with the triplet loss component. The quadruplet loss [10] is the improved version of the triplet loss which also tries to force the irrelevant negative pairs further away from the positive pairs. The quadruplet loss is given by

$$\begin{aligned}\mathcal{L}_{quad} =& max(0, m_1 + d_{pos} - d_{neg}) + \\ & max(0, m_2 + d_{pos} - d^*_{neg}),\end{aligned} \tag{5}$$

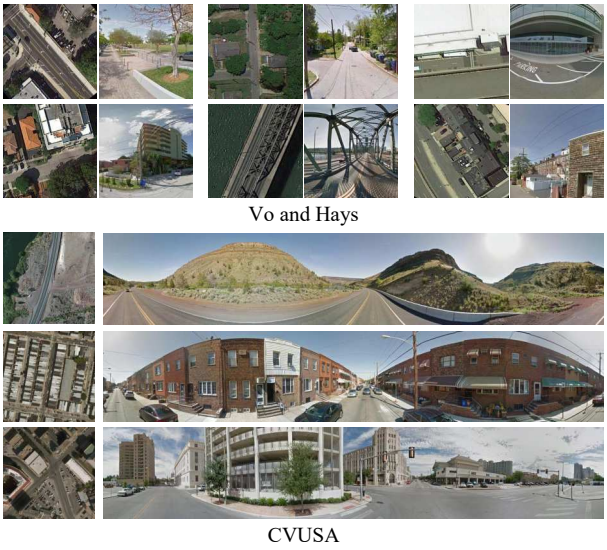

Figure 4. Sample images from the Vo and Hays [46], and CVUSA [53].

where m_1 and m_2 are the margins and d^*_{neg} is distance of another example that is outside of the chosen set of positive, negative and anchor examples. We note that the margins are no longer needed with our weighted soft-margin component. Our weighted quadruplet loss is given by

$$\mathcal{L}_{quad,weighted} = ln(1 + e^{\alpha(d_{pos} - d_{neg})}) + ln(1 + e^{\alpha(d_{pos} - d^*_{neg})}). \quad (6)$$

5. Experiments and Discussions

5.1. Dataset

We evaluate our proposed deep networks on two existing datasets - CVUSA [53] and Vo and Hays [46]. The CVUSA dataset contains 35,532 image pairs for training and 8,884 image pairs for testing. All ground images are panoramas. Vo and Hays' dataset consists of around one million image pairs from 9 different cities. All ground images are cropped from panoramic images to a fixed size. We use all image pairs from 8 of the 9 cities to train the networks and use the image pairs from the 9th city, i.e. Denver city, for evaluation. Figure 4 shows some examples of the two datasets.

5.2. Implementation and Training

We use the VGG16 [38] architecture with 13 convolutional layers to extract local features, and a NetVLAD with 64 clusters to generate the global descriptors. We set $\alpha = 10$ for both the weighted triplet and weighted quadruplet losses. We use the squared Euclidean distance in our loss functions. The parameters in VGG16 are initialized with a pre-trained model on ImageNet [13]. All the parameters in NetVLAD and fully connected layers are randomly initialized.

We implement our CVM-Nets using Tensorflow [2] and train using the Adam optimizer [20] with the learning rate of 10^{-5} and dropout ($= 0.9$) for all fully connected layers. The training is divided into two stages. In the first stage, we adopt the exhaustive mini-batch strategy [46] to maximize the number of triplets within a batch. We feed pairs of corresponding satellite and ground images into our Siamese-like architecture. We have a total of $M \times 2(M-1)$ triplets for M positive pairs of ground-to-satellite images. This is because for each ground or satellite image in M positive pairs, there are $M-1$ corresponding negative pairs from all the other images, i.e. $2(M-1)$ for both the ground and satellite images in a positive pair. Once the loss stops decreasing, we start the second stage with in-batch hard negative mining. For each positive pair, we choose the negative pair with smallest distance in current batch.

	Recall @top 1%	
	Cropped [46]	Panorama [53]
Siamese (AlexNet)	1.1%	4.7%
Siamese (VGG)	1.3%	9.9%
Workman et al. [50]	15.4%	34.3%
Vo and Hays [46]	59.9%	63.7%
Zhai et al. [53]	—	43.2%
CVM-Net-I	**67.9%**	**91.4%**
CVM-Net-II	66.6%	87.2%

Table 1. Comparison of top 1% recall on our CVM-Nets with other existing approaches [53, 46, 50] and two baselines, i.e. Siamese network with AlexNet and VGG.

5.3. Comparison and Results

Evaluation metrics We follow Vo and Hays [46], and Workman et al. [50] in using the recall accuracy at top 1% as the evaluation metric for our networks. For a query ground view image, we retrieve the top 1% closest satellite images with respect to the global descriptor distance. It is regarded as correct if the corresponding satellite image is inside the retrieved set.

Comparison to existing approaches We compare our proposed CVM-Nets to three existing works [46, 50, 53] on the two datasets [46, 53]. We used the implementations provided in the authors' webpages. Furthermore, we take the Siamese network with both AlexNet [21] and VGG [38] as the baseline in our comparisons, since these networks are widely used in image retrieval tasks. We use our weighted soft-margin ranking loss in our CVM-Nets. The soft-margin triplet loss is used on the network from Vo and Hays [46], as suggested by the authors in the paper. We also apply the soft-margin triplet loss on the two baseline Siamese networks - AlexNet and VGG since the soft-margin triplet loss

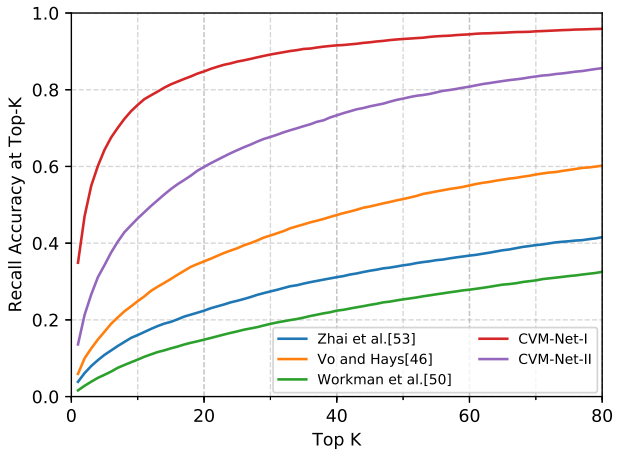

Figure 5. Comparison of our CVM-Nets and other existing approaches [53, 46, 50]: All models are trained on CVUSA [53].

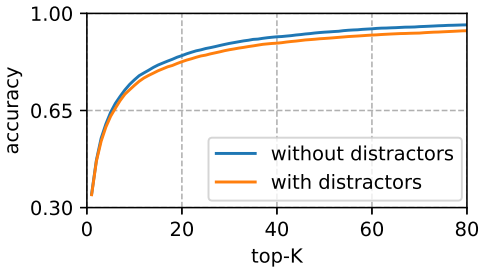

Figure 6. Top-K recall accuracy on the evaluation dataset with and without distractor images.The model is trained on CVM-Net-I on CVUSA dataset [53].

	Triplet	Quadruplet
CVM-Net-I (AlexNet)	65.4%	73.7%
CVM-Net-I (VGG16)	91.4%	89.9%
CVM-Net-II (AlexNet)	63.0%	83.9%
CVM-Net-II (VGG16)	87.2%	88.7%

Table 2. Performance of different architectures and losses on the CVUSA dataset [53]: AlexNet [21] and VGG16 [38] are used as the local feature extraction network.

produces the state-of-the-art results in [46]. The Euclidean loss is used on the network proposed by Workman et al. [50] since their network is trained on only positive pairs.

Table 1 shows the top 1% recall accuracy results of our CVM-Nets compared to all the other approaches on the two datasets - which we called "Cropped" [46] and "Panorama" [53] in the table for brevity. It can be seen that both our proposed networks - CVM-Net-I and CVM-Net-II significantly outperform all the other approaches. This suggests that NetVLAD used in both our CVM-Nets is capable of learning much more discriminative features compared to the CNN and/or fully connected layers architectures utilized by the other approaches. Furthermore, it can be seen that CVM-Net-I outperforms CVM-Net-II in both datasets. This result suggests that weight sharing, a technique commonly used in Siamese network based architectures, is not necessary for our network for cross-view image retrieval. It is not surprising that all networks perform better on the panorama images since these images contain more information from the wide field-of-views. The low accuracies of the Siamese networks indicate that they are not suitable for cross-view image retrieval, although they performed well in traditional image retrieval tasks, e.g. face identification.

We show the recall accuracy from top 1 to top 80 (top 0.9%) of our CVM-Nets with all the other approaches on CVUSA dataset [53] in Figure 5. It illustrates that our proposed networks outperform all the other approaches. In Figure 9, we show some retrieval examples on two benchmark datasets [46, 53].

Adding distractor images We add 15,643 distractor satellite images in Singapore to our original test database which has 8,884 satellite images in USA. Figure 6 shows the top-K recall accuracy curve. The result is from the model trained on CVM-Net-I on the CVUSA [53] dataset. There is only a marginal difference between the results with and without distractor images. This proves the robustness of our proposed network.

5.4. Discussions

Local feature extraction In Table 2, we compare several variations on our proposed architecture. We conduct experiments to investigate AlexNet and VGG16 for local feature vectors extraction. It can be seen from the table that VGG16 performs better than AlexNet in both our CVM-Nets. This result is not surprising because VGG16 is a deeper network compared to Alexnet, hence is able to extract richer local features.

Cross-view matching It can be seen from Table 2 that CVM-Net-I outperforms CVM-Net-II for both the VGG16 and AlexNet implementations for local features extraction. This further reinforces our analysis in the previous paragraph that shared weights implemented on CVM-Net-II is not necessary for our cross-view image based retrieval task. It is also interesting to note from the results in Table 2 that our CVM-Nets implemented with both VGG16 and AlexNet outperform all other approaches in Table 1.

Ranking loss The triplet loss has been widely used in image retrieval for a long time, while the quadruplet loss [10] was introduced recently to further improve the triplet loss. We train our CVM-Nets implemented with AlexNet and VGG16 for local feature extraction on both the triplet and

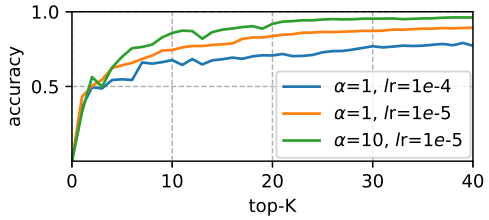

Figure 7. Performance of our weighted soft-margin triplet loss with different parameters. *lr* is short for learning rate. It takes about 1 hour to train each epoch.

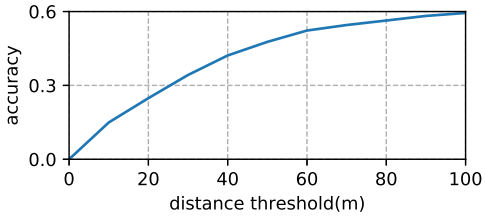

Figure 8. The retrieval accuracy on distance error threshold.

quadruplet losses for comparison. As can be seen from the results in Table 2, quadruplet loss outperforms triplet loss significantly on both our CVM-Nets with AlexNet. However, only minor differences in performances of the triplet and quadruplet losses can be observed for our CVM-Nets with VGG16. These results suggest that quadruplet loss has a much larger impact on shallower networks, i.e. AlexNet for feature extraction.

Contrastive loss was widely used in the past as well. To test contrastive loss, we train our CVM-Net-I and II on CVUSA dataset [53]. The top 1% recall accuracy is 87.8% and 79.8% respectively. It is not as good as the triplet loss or the quadruplet loss whose results are shown in Table 2.

Weighted soft-margin We also compare the performance of our CVM-Nets on different α values in our weighted soft-margin triplet loss $\mathcal{L}_{weighted}$ (see Equation 4). Specifically, we conducted experiments on $\alpha = 10$ with learning rate 10^{-5}, $\alpha = 1$ (soft-margin triplet loss) with learning rate 10^{-5}. In addition, we also tested on $\alpha = 1$ with learning rate 10^{-4} to compare the convergence speed with our weighted loss. The accuracies from the respective parameters with respect to the number of epochs are illustrated in Figure 7. As can be seen, our loss function makes the network converge to higher accuracies in a shorter amount of time. We choose $\alpha = 10$ in our experiments since the larger value of α does not make much different.

5.5. Cross-view Localization

We perform image-based geo-localization with respect to a geo-referenced satellite map with our cross-view image retrieval CVM-Nets. Our geo-referenced satellite map covers a region of 10×5 km of the South-East Asian country - Singapore. We collect the ground panoramic images of Singapore from Google Street-view [1]. We choose to test our CVM-Nets on Singapore because it is easy to get the datasets online for the highly developed country. Furthermore, we want to show that our CVM-Nets trained on the North American based CVUSA datasets generalize well on a drastically different area. We tessellate the satellite map into grids at 5 m intervals. Each image patch is 512×512 pixels and the latitude and longitude coordinates of the pixel center give the location of the image patch. We use our CVM-Net-I trained on the CVUSA dataset to extract global descriptors from our Singapore dataset. We visualize the heatmap of the similarity scores on the reference satellite map of two examples in Figure 10. We apply the exponential function to improve the contrast of the similarity scores. We can see that our CVM-Net-I is able to recover the ground truth locations for both examples in Figure 10. It is interesting to see that our street-view based query image generally return higher similarity scores on areas that correspond to the roads on the satellite map.

We conduct a metric evaluation on geo-localization. A query is regarded as correctly localized if the distance to the ground truth location is less than the threshold. We show the recall accuracy with respect to the distance threshold in Figure 8. The accuracy on 100 m threshold is 67.1%. The average localization error is 676.7 m. As can be seen from the metric evaluation result, there is large room for ground-to-aerial geo-localization study despite state-of-art retrieval performance.

6. Conclusion

In this paper, we propose two cross-view matching networks - CVM-Net-I and CVM-Net-II, which are able to match ground view images with satellite images in order to achieve cross-view image localization. We introduce the weighted soft-margin ranking loss, and show that it notably accelerates training speed and improves the performance of our networks. We demonstrate that our approach significantly outperforms state-of-the-art approaches with experiments on large datasets.

Possible extensions Our proposed CVM-Nets can also be trained to work in other cross-domain image retrieval tasks, e.g. matching hand sketches to natural photos, day and night images, paintings with different styles etc. Furthermore, our networks can be extended to general cross-domain information retrieval tasks. An example is the word-to-image retrieval where we can replace the local image feature extraction component with e.g. the Word2Vec model [28] branch, while keeping VGG16 in the other branch.

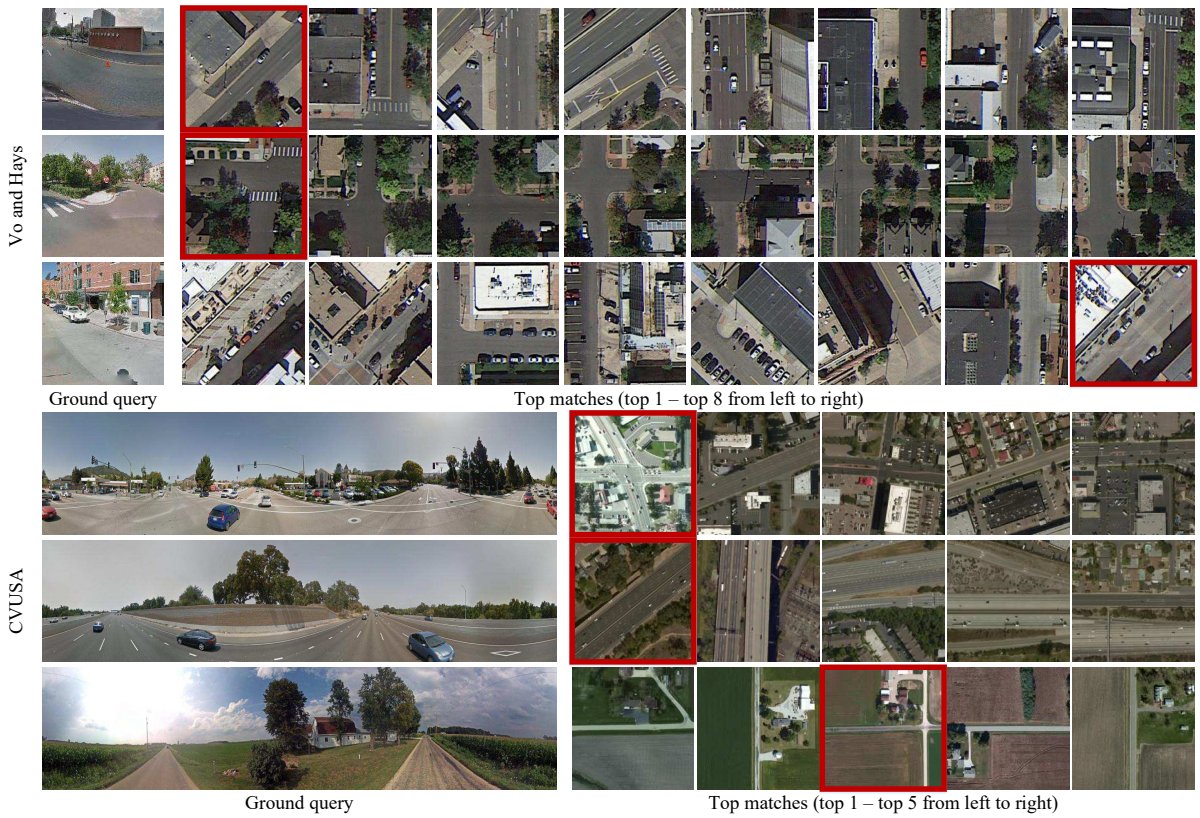

Figure 9. Image retrieval examples on Vo and Hays dataset [46] and CVUSA dataset [53]. The satellite image bordered by red square is the groundtruth.

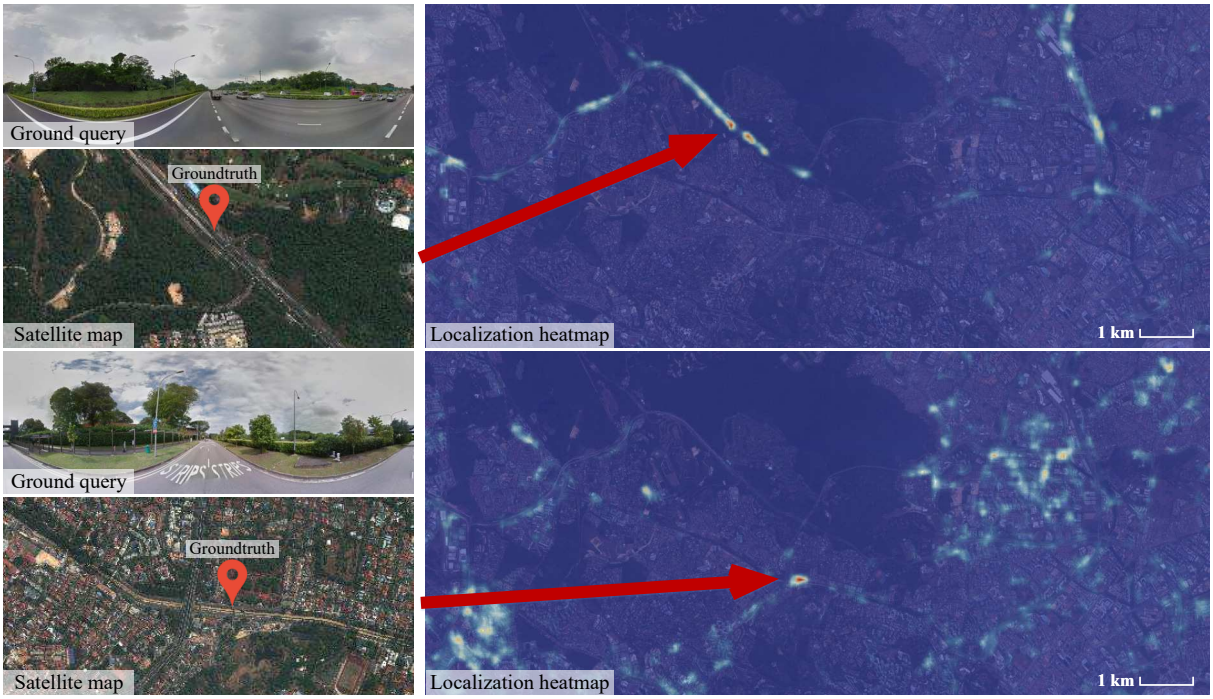

Figure 10. Large-scale geo-localization examples on our dataset.

References

[1] Google street view image api. https://developers.google.com/maps/documentation/streetview/intro, 2017 (accessed September 4, 2017). 7

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016. 5

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 2, 3, 4

[4] 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), 2017. 2

[5] M. Bansal, K. Daniilidis, and H. Sawhney. Ultra-wide baseline facade matching for geo-localization. In *European Conference on Computer Vision*, October 2012. 1

[6] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis. Geo-localization of street views with aerial image databases. In *ACM International Conference on Multimedia*, November 2011. 2

[7] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, May 2006. 1

[8] Y. Cao, M. Long, and J. Wang. Correlation hashing network for efficient cross-modal retrieval. In *British Machine Vision Conference*, September 2017. 2

[9] Y. Cao, M. Long, J. Wang, and S. Liu. Deep visual-semantic quantization for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 2

[10] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 2, 4, 6

[11] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 2

[12] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2005. 4

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2009. 2, 5

[14] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10), 2015. 2

[15] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 4

[16] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2008. 1

[17] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 2, 4

[18] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010. 2, 3

[19] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis. Joint learning for attribute-consistent person re-identification. In *European Conference on Computer Vision Workshops*, September 2014. 2

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, December 2012. 2, 3, 5, 6

[22] T.-Y. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2013. 1, 2

[23] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 1

[24] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet cnn for person re-identification. In *ACM International Conference on Multimedia*, October 2016. 2

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004. 1

[26] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *IEEE International Conference on Robotics and Automation*, May 2014. 1

[27] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-dof localization on mobile devices. In *European Conference on Computer Vision*, September 2014. 1

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. 7

[29] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2006. 2

[30] M. Noda, T. Takahashi, D. Deguchi, I. Ide, H. Murase, Y. Kojima, and T. Naito. Vehicle ego-localization by matching in-vehicle camera images to an aerial image. In *Asian Conference on Computer Vision Workshops*, November 2010. 2

[31] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 4

[32] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 2

[33] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *Conference on Computer Vision and Pattern Recognition*, June 2016. 1

[34] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 4

[35] T. Senlet and A. Elgammal. A framework for global vehicle localization using stereo images and satellite and road maps. In *IEEE International Conference on Computer Vision Workshops*, November 2011. 2

[36] T. Senlet and A. Elgammal. Satellite image based precise robot localization on sidewalks. In *IEEE International Conference on Robotics and Automation*, May 2012. 2

[37] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz. Accurate geo-registration by ground-to-aerial image matching. In *International Conference on 3D Vision*, December 2014. 1

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3, 5, 6

[39] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, October 2003. 2

[40] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*. Springer, 2006. 2

[41] E. Stumm, C. Mei, S. Lacroix, J. Nieto, M. Hutter, and R. Siegwart. Robust visual place recognition with graph kernels. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 1

[42] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, October 2016. 2

[43] Y. Tian, C. Chen, and M. Shah. Cross-view image matching for geo-localization in urban environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 1

[44] A. Viswanathan, B. R. Pires, and D. Huber. Vision based robot localization by ground to satellite matching in gps-denied situations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2014. 2

[45] N. Vo, N. Jacobs, and J. Hays. Revisiting im2gps in the deep learning era. In *IEEE International Conference on Computer Vision*, October 2017. 1

[46] N. N. Vo and J. Hays. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision*, October 2016. 1, 2, 3, 4, 5, 6, 8

[47] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 2

[48] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep metric learning with angular loss. In *IEEE International Conference on Computer Vision*, October 2017. 2

[49] S. Workman and N. Jacobs. On the location dependence of convolutional neural network features. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2015. 1, 2

[50] S. Workman, R. Souvenir, and N. Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, December 2015. 1, 2, 3, 5, 6

[51] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 4

[52] A. R. Zamir and M. Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8), August 2014. 1

[53] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs. Predicting ground-level scene layout from aerial imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 1, 2, 3, 5, 6, 7, 8

[54] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, December 2014. 2