This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Ground-to-Aerial Image Geo-Localization With a Hard Exemplar Reweighting Triplet Loss

Sudong Cai[1] Yulan Guo[2,1] Salman Khan[3] Jiwei Hu[4] Gongjian Wen[1,2]
[1]National University of Defense Technology [2]Sun Yat-Sen University
[3]Inception Institute of Artificial Intelligence [4]Wuhan University of Technology
781594648@whut.edu.cn yulan.guo@nudt.edu.cn salman.khan@inceptioniai.org

Abstract

The task of ground-to-aerial image geo-localization can be achieved by matching a ground view query image to a reference database of aerial/satellite images. It is highly challenging due to the dramatic viewpoint changes and unknown orientations. In this paper, we propose a novel in-batch reweighting triplet loss to emphasize the positive effect of hard exemplars during end-to-end training. We also integrate an attention mechanism into our model using feature-level contextual information. To analyze the difficulty level of each triplet, we first enforce a modified logistic regression to triplets with a distance rectifying factor. Then, the reference negative distances for corresponding anchors are set, and the relative weights of triplets are computed by comparing their difficulty to the corresponding references. To reduce the influence of extreme hard data and less useful simple exemplars, the final weights are pruned using upper and lower bound constraints. Experiments on two benchmark datasets show that the proposed approach significantly outperforms the state-of-the-art methods.

1. Introduction

Image-based geo-localization has attracted significant attention for its numerous applications in autonomous driving [21], augmented reality [22], and mobile robotics [30, 21]. Traditional methods mainly focus on ground-to-ground image matching, where both query and reference images are acquired from ground views [2, 14, 13, 27, 25, 21, 40, 6]. Although matching images that are all from ground-level is relatively easy, it is difficult to comprehensively cover a large area using only ground-view images obtained from vehicle-mounted cameras or crowd-sourcing websites [32, 11, 34]. Therefore, ground-to-ground image geo-localization tends to fail in places without available reference images. In contrast, images captured from a bird's eye view by satellites and aerial vehicles can densely and uniformly cover the Earth. Consequently, matching ground-level images to aerial images has gradually become an attractive approach for coarse-level geo-localization and place recognition [32], as shown in Figure 1. However, cross-view image matching is extremely challenging due to large viewpoint differences, lighting variations, and orientation (i.e., azimuth) uncertainty between ground and aerial images [11, 32, 17, 38, 34]. Despite numerous attempts, this problem is largely unsolved and needs new breakthroughs.

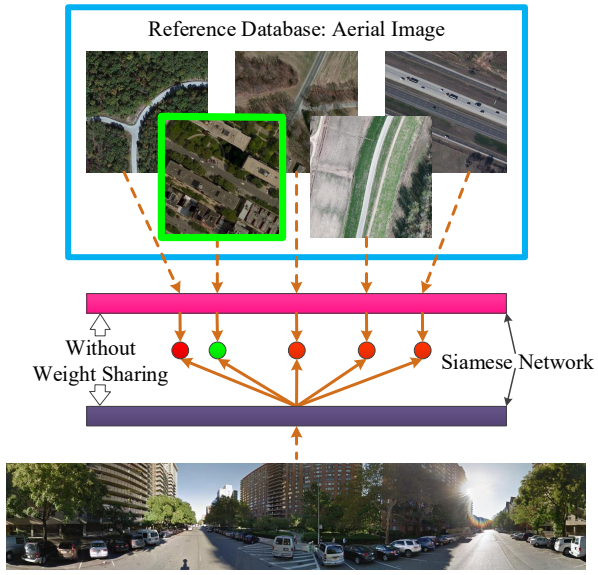

Figure 1. A demonstration of matching ground-view query to reference database consisting of aerial image. Here, "green" and "red" colors denote positive and negative matches, respectively.

Due to the low cross-view matching accuracy of hand-crafted features, existing methods generally compute similarities between features from CNN models trained independently for ground and aerial images. Inspired by face recognition [28, 31], Lin *et al*. [17] proposed '*Where-CNN*' using a Siamese architecture. However, it has been demonstrated that weight sharing in a Siamese architecture leads

to poor performance in cross-view image recognition. Vo and Hays [34] proposed a soft-margin distance-based loss and an auxiliary network branch to estimate the orientation. Their model was robust against random orientations. Recently, Hu *et al*. [11] proposed CVM-Nets based on a Siamese CNN with NetVLAD [2] (a learnable feature aggregation module). They also introduced a soft-margin loss with a manually given weight to speed up the training, and applied hard exemplar mining by directly using top-1 hard negatives, with state-of-the-art performance being achieved. Although the effectiveness of hard exemplar mining was investigated in [11], it is still difficult to locate informative hard exemplars appropriately. In contrast, we propose a hard exemplar mining strategy for cross-view image matching. Specifically, our approach automatically allocates weights to triplets according to their difficulty levels. This allows us to focus on informative hard exemplars for the improvement of cross-view image geo-localization.

Contributions: The contributions of this paper are as follows: **(1)** We propose a new triplet loss to improve the quality of network training for cross-view images. This loss can achieve online hard exemplar mining in an end-to-end manner. Experimental results on benchmark datasets demonstrate that our loss outperforms the soft-margin triplet loss [34]. **(2)** We propose a lightweight attention module (FCAM), and integrate it into a basic residual network to form our Siamese network, which achieves better performance than the plain model (without FCAM). **(3)** We train our Siamese network with our loss to obtain discriminative CNN features for cross-view image-based geo-localization. Experimental results demonstrate that our approach significantly outperforms the state-of-the-art approaches [11, 34] on benchmark datasets.

2. Related Work

Existing image geo-localization approaches can be broadly categorized into two classes according to their image representation methods.

2.1. Hand-Crafted Feature Based Approaches

Hand-crafted features were widely adopted to perform cross-view image matching before deep learning was introduced into this area [23, 4, 29, 16, 33]. Bansal *et al*. [4] extracted building facades from oblique aerial images and then performed geo-localization by matching building facade patches. Bansal *et al*. [3] further handled extreme viewpoint differences by encoding the self-similarity of patterns on facades at their corresponding scales using a Scale-selective Self-similarity (S4) descriptor. It was demonstrated that S4 feature achieved better performance than Scale Invariance Feature Transform (SIFT) [20]. Viswanathan *et al*. [33] improved the matching performance of local feature descriptors by converting ground images to their top-down view. Due to the significant differences in appearance, cross-view image matching performance achieved by hand-crafted features is relatively poor.

2.2. Deep Learning Based Approaches

Deep learning provides a more accurate alternative for cross-view image geo-localization and has recently dominated this area [11, 32, 17, 38, 41, 34]. Lin *et al*. [17] proposed the first deep learning method to achieve ground-to-aerial geo-localization based on two Siamese CNNs (namely, Where-CNN and Where-CNN-DS). Their Siamese CNNs were trained with a modified version of contrastive loss [7]. Comparative experiments demonstrated its significant improvement in performance as compared to hand-crafted descriptors. Workman *et al*. [38] introduced a deep learning method to learn semantic representations of aerial images. They also proposed a CNN model to fuse semantic features from different spatial scales. Their work demonstrated that features trained from cross-view image pairs significantly outperform off-the-shelf CNN features.

To further improve the robustness of CNN features, some methods exploit $attention$ for objects and patches of interest. Altwaijry *et al*. [1] integrated the Spatial Transformer (ST) module [12] into a Siamese network adapted from AlexNet. Similarity was calculated using features inferred from patches rather than whole images. Features produced by Siamese CNN with the ST module were demonstrated to outperform the original model. Tian *et al*. [32] proposed a two-stage framework to detect buildings using Faster R-CNN [26]. Then, images were represented by the dominant sets constructed by features inferred from patches of buildings. The pairwise similarity of dominant sets was learned from a Siamese network. Their method remarkably outperforms pre-trained CNN features. These methods [1, 32] impose robustness to CNN features with respect to visual transformations by utilizing the detection model to focus on specific landmark areas. However, their efficiency is limited. Instead, we emphasize informative features by designing a lightweight feature reweighting module for the $attention$ mechanism.

Recently, several methods have been proposed to address the learning of metrics and discriminative global image representations. Vo *et al*. [34] proposed a soft-margin triplet loss to reduce the distance between the anchor and positive exemplars while increasing the distance between the anchor and negative exemplars. Besides, an auxiliary orientation regression branch was added to achieve rotation invariance. It was shown that a well-designed learning metric can benefit cross-view image geo-localization. Hu *et al*. [11] proposed CVM-Net, which adopts an NetVLAD module [2] to aggregate CNN feature units for the generation of discriminative image representations. They also manually assigned a weight to the soft-margin triplet loss to speed up the train-

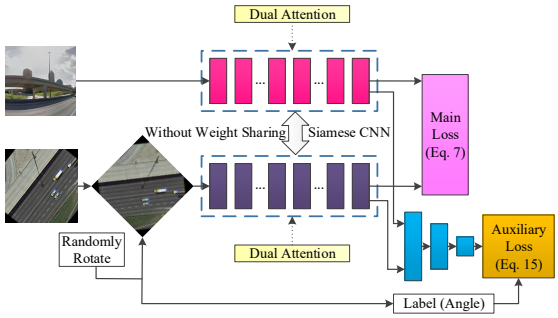

Figure 2. An overview of our approach.

ing. Besides, they applied a simple hard exemplar mining strategy by directly using top-1 hard negatives, which is similar to [10]. Although Hu *et al*. [11] achieved state-of-the-art results, using only top-1 negatives for hard exemplar mining tends to leave out some informative exemplars. Rather than using the top-1 hard sample, several works introduced adaptive sampling [8, 39] and gradient rescaling [18, 36, 24] methods to perform class-level retrieval, face verification and object detection. In this paper, we design an online exemplar reweighting triplet loss to allocate different weights to triplets according to their difficulty levels. Consequently, our loss can adaptively emphasize meaningful hard triplets during network training, rather than manually designating top-k hard exemplars. Experimental results show that our approach significantly outperforms existing methods.

3. The Proposed Method

In this section, we provide an overview of our method, and describe its two major contributions, i.e., the Feature Context-based Attention Module (FCAM) (Section 3.2) and the Hard Exemplar Reweighting (HER) triplet loss (Section 3.3).

3.1. Overview

As shown in Figure 1, given a ground-view query image, the retrieval of reference aerial images is achieved using pair-wise similarities between their CNN features.

Since convolution operations blend both channel and spatial information to generate informative features, we propose a lightweight dual attention module (i.e, FCAM) to improve feature discriminativeness by applying attention mechanism on both channel and spatial dimensions (Section 3.2). Two CNN feature extractors with the same structure are respectively constructed for ground-view and aerial images by integrating our attention module into the basic ResNet [9]. The overall Siamese network is built upon these two CNN feature extractors without weight sharing. It is further integrated with an auxiliary Orientation Regression (OR) learning branch.

To improve network training, we propose to fully use informative hard exemplars. Specifically, we introduce a new HER triplet loss to achieve online hard exemplar mining based on triplet reweighting (Section 3.3). We assign large weights to useful but hard triplets while allocating small weights to less informative but easy triplets. An overview of our approach is shown in Figure 2.

3.2. Feature Context-Based Attention Module

Attention has been demonstrated to be effective for improving the representation ability of CNNs by focusing on meaningful features while suppressing useless ones. Consequently, applying attention along both channel and spatial axes of feature maps can help a CNN to learn "which channels" and "which feature units" should be focused on. Our lightweight dual attention module can be sequentially decomposed into a channel attention submodule and a spatial attention submodule. Our channel attention is adopted from the Convolutional Block Attention Module (CBAM) [37], but a context-aware feature reweighting strategy is integrated into our spatial attention submodule. Given a feature map $\mathbf{U} \in \mathbb{R}^{W\times H\times C}$ as the input, the attention module jointly infers a 1D channel attention descriptor $\mathbf{Z}^C(\mathbf{U}) \in \mathbb{R}^{1\times 1\times C}$ and a 2D spatial attention mask $\mathbf{Z}^S(\mathbf{U}') \in \mathbb{R}^{W\times H\times 1}$. The overall attention process is defined as:

$$\mathbf{U}' = \mathbf{Z}^C(\mathbf{U}) \otimes \mathbf{U}, \tag{1}$$

$$\mathbf{U}'' = \mathbf{Z}^S(\mathbf{U}') \otimes \mathbf{U}', \tag{2}$$

where $\mathbf{U}' \in \mathbb{R}^{W\times H\times C}$ and $\mathbf{U}'' \in \mathbb{R}^{W\times H\times C}$ are the output feature maps of the channel and spatial attention submodules, respectively, $\otimes$ denotes element-wise multiplication.

Channel attention submodule. Channel attention is used to emphasize channels that are relatively informative. In this paper, we adopt the channel attention submodule to exploit inter-channel dependencies of CNN features. The channel attention submodule is shown in Figure 3. First, 1D global channel descriptors $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ are generated by applying max-pooling f_{max} and average-pooling f_{avg} to an input feature map $\mathbf{U}$. Then, both $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ are excited by a Multi-Layer Perceptron (MLP) to analyze their inter-channel dependencies. Consequently, the channel attention descriptor $\mathbf{Z}^C(\mathbf{U})$ is obtained by summing the excited descriptors with a sigmoid activation. The output channel attention map $\mathbf{U}'$ is generated by performing element-wise multiplication between the channel attention descriptor and input feature map $\mathbf{U}$. The channel attention descriptor is computed as:

$$\begin{aligned}\mathbf{Z}^C(\mathbf{U}) &= \delta\left(f_{ext}\left(f_{max}\left(\mathbf{U}\right)\right) + f_{ext}\left(f_{avg}\left(\mathbf{U}\right)\right)\right) \\ &= \delta\left(\mathbf{W}_2^e\sigma\left(\mathbf{W}_1^e\boldsymbol{v}^1\right)\right) + \delta\left(\mathbf{W}_2^e\sigma\left(\mathbf{W}_1^e\boldsymbol{v}^2\right)\right),\end{aligned} \tag{3}$$

where f_{ext} represents the MLP operation. $\mathbf{W}_1^e \in \mathbb{R}^{T\times C}$ and $\mathbf{W}_2^e \in \mathbb{R}^{C\times T}$ represent the MLP weights of the first

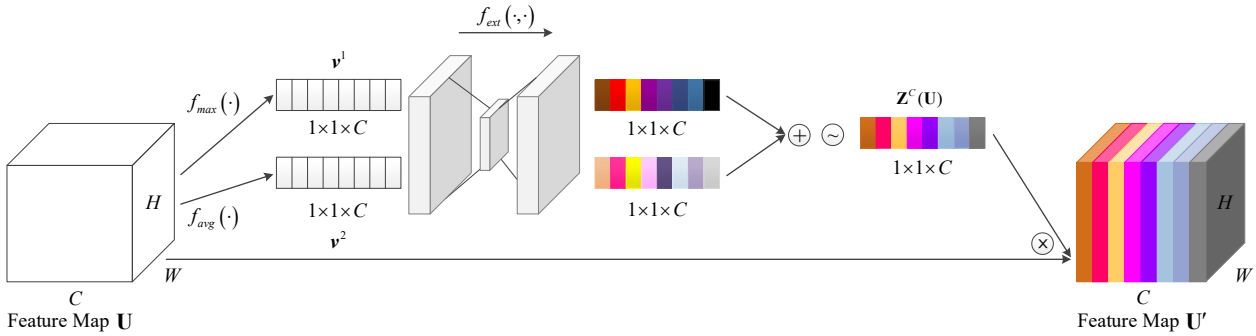

Figure 3. A diagram of the channel attention submodule.

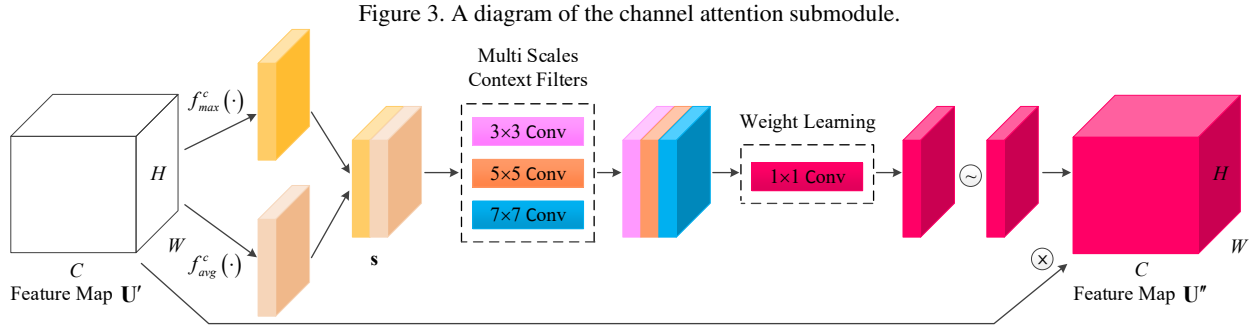

Figure 4. A diagram of the spatial attention submodule.

and the second fully-connected layer. δ and σ represent the sigmoid and the ReLU activation, respectively. Here, $T = C/r$ and r is the reduction level in the first learning layer of the MLP.

Spatial attention submodule. Spatial attention is used to highlight meaningful feature units. In CBAM [37], Woo *et al*. utilize a 7×7 convolution to learn the spatial attention mask after concatenating the max-pooled and average-pooled spatial descriptors. Inspired by the Contextual Reweighting Network (CRN) [14], we integrate feature context-aware learning into the basic spatial attention submodule of CBAM. That is, rather than using a 7×7 convolution only, we use the convolutions with different receptive fields to generate intermediate feature masks. Then, we concatenate these intermediate masks and use a 1×1 convolution to learn weights. The spatial attention map can be considered as the weighted sum of feature masks. An illustration of our spatial attention submodule is shown in Figure 4. $\mathbf{s} \in \mathbb{R}^{W \times H \times 2}$ is generated by concatenating squeezed feature masks $f^c_{max}(\mathbf{U}')$ and $f^c_{avg}(\mathbf{U}')$. Here, f^c_{max} and f^c_{avg} denote max-pooling and average-pooling along the channel axis, respectively. To exploit the contextual information of feature units, three different scales of context filters (3×3, 5×5, and 7×7) are used. The feature mask $\mathbf{p} \in \mathbb{R}^{W \times H \times 3}$ is produced by concatenating these channel masks generated by the 3×3, 5×5, and 7×7 context filters. Then, we use a 1×1 convolution to learn and to accumulate weights. The spatial attention mask can be computed as:

$$\mathbf{Z}^S(\mathbf{U}') = \delta\left(f^{1\times1}\left(f^{3\times3}(\mathbf{s}); f^{5\times5}(\mathbf{s}); f^{7\times7}(\mathbf{s})\right)\right), \quad (4)$$

where $f^{n\times n}$ denotes an $n \times n$ convolution. Consequently, the spatial attention map $\mathbf{U}''$ is obtained by element-wise multiplication between the channel attention map $\mathbf{U}'$ and the final spatial attention mask $\mathbf{Z}^S(\mathbf{U}')$.

Our overall attention module is formed by a sequential combination of the channel attention submodule and the spatial attention submodule (as shown in Figure 5). We integrate our attention module into each building block of the basic ResNet [9] to form the CNN feature extractor (namely, FCANet) for cross-view matching.

3.3. Hard Exemplar Reweighting Triplet Loss

To improve network training with instance-wise exemplars, we propose an online hard exemplar mining strategy based on triplet reweighting. We then integrate this strategy into the soft-margin triplet loss [34], resulting in the hard exemplar reweighting triplet loss. Given a triplet of anchor A_i, its corresponding positive exemplar P_i, and negative exemplar $N_{i,k}$ (i.e, the k-th negative exemplar of A_i) in an mini-batch, the original triplet loss [28] is defined as:

$$L_{tri}(A_i, P_i, N_{i,k}) = \max(0, m + d_p(i) - d_n(i,k)), \quad (5)$$

where m is the max-margin, $d_p(i)$ represents the squared Euclidean distance between A_i and P_i, $d_n(i,k)$ represents the squared Euclidean distance between A_i and $N_{i,k}$. This loss has achieved remarkable success in instance-wise recognition tasks [2, 14, 5, 19, 35]. However, since this loss relies on a max-margin to truncate the penalization, the gap between distances of negative pairs and distances of positive pairs is limited.

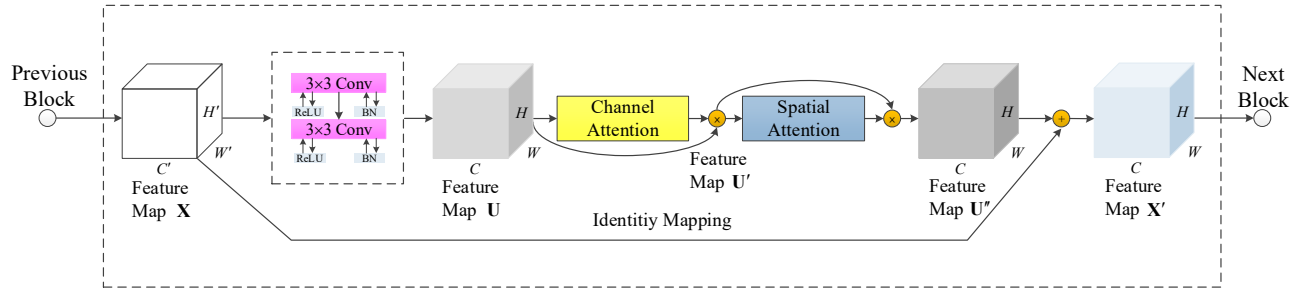

Figure 5. A diagram of the modified building block. This building block is produced by integrating the proposed FCAM into the basic residual block.

To address the limitation of generating penalization with max-margin, Vo *et al*. [34] proposed a soft-margin triplet loss to enforce penalization according to the current performance of the network. This loss has demonstrated a better performance than the original triplet loss, that is:

$$L_{soft}\left(A_i, P_i, N_{i,k}\right) = \log(1+\exp(d_p(i)-d_n(i,k))). \quad (6)$$

Leveraging soft-margin triplet loss, we propose a new loss to integrate hard exemplar mining in an end-to-end manner. Our online hard exemplar mining strategy is based on triplet reweighting, and the weight allocated to each triplet is computed according to its difficulty level. Suppose the weight computed by our hard exemplar mining method for a triplet is $w_{hard}(A_i, P_i, N_{i,k})$, our loss is then defined as:

$$L_{hard}\left(A_i, P_i, N_{i,k}\right) = w_{hard}(A_i, P_i, N_{i,k}) * \log(1 + \exp(d_p(i) - d_n(i,k))). \quad (7)$$

The computation of weight $w_{hard}(A_i, P_i, N_{i,k})$ will be described in the following sections.

Distance rectified logistic regression. We propose a distance rectified logistic regression to estimate the difficulty of current triplets. The most difficult negative exemplar for a given anchor A_i is defined as the negative exemplar in the current batch that has the smallest distance to the anchor.

It has been demonstrated that extremely hard exemplars decrease the quality of training by leading it to local minimum at early stage [28]. Let $gap(i,k) = d_n(i,k) - d_p(i)$, then, extremely hard exemplars satisfy the following condition: $C_h : gap(i,k) \leq 0$.

Meanwhile, less informative simple triplets satisfy the condition: $C_s : gap(i,k) \geq m$. Considering both conditions C_h and C_s, we define a reference negative distance D_{ref} for each anchor, based on its distance to the positive exemplar:

$$D_{ref}(A_i) = d_p(i) + \frac{m}{2}. \quad (8)$$

The negative exemplar N_i^r with distance to anchor A_i being equal to $D_{ref}(A_i)$ is considered as the reference negative. For the given triplet of A_i, P_i, and $N_{i,k}$, we suppose that when $N_{i,k}$ is the reference negative N_i^r, the weight allocated for the current triplet is 1 (i.e., without emphasizing or suppression). Then, we consider $gap(i,k)$ as a random variable, and define the weight as:

$$w\left(A_i, P_i, N_{i,k}\right) = -\log_2\left(p_{match}\left(A_i, P_i, N_{i,k}\right)\right), \quad (9)$$

where $p_{match}\left(A_i, P_i, N_{i,k}\right)$ denotes the correct matching probability of the triplet:

$$p_{match}\left(A_i, P_i, N_{i,k}\right) = \frac{1}{1+\exp\left(-gap(i,k)+\beta\right)}, \quad (10)$$

where $\beta = {}^{m}/_{2}$ is the distance rectification factor. That is, when a reference negative N_i^r occurs, the probability of correctly matching A_i to P_i equals the probability of mistakenly matching A_i to N_i^r (i.e., $p_{match} = 0.5$), and the weight for the current triplet is $w\left(A_i, P_i, N_i^r\right) = 1$. Therefore, when the distance of an anchor to its negative exemplar $d_n(i,k)$ is smaller than the corresponding reference distance $D_{ref}(A_i)$, it will be emphasized by allocating a weight larger than 1. Note that, for un-normalized features, the margin m can be computed as:

$$m = \frac{\gamma}{2B}\sum_{i=1}^{B}\left(|\, f\left(A_i\right)|^2 + |\, f\left(P_i\right)|^2\right), \quad (11)$$

where $f(\cdot)$ represents network inference, γ is a ratio ranging from 0 to 1, and B is the number of anchors in the current mini-batch.

We further propose an upper and lower weight limit to reduce the influence of extremely hard exemplars while eliminating the negative neutralization effect of simple exemplars. According to C_h , the critical condition for extremely hard exemplars is defined by $gap(i,k) = 0$, and the upper bound for the weight truncation can be computed by:

$$w_{high} = -\log_2\left(\frac{1}{1+\exp\left(\beta\right)}\right). \quad (12)$$

Then, if the weight for a triplet is larger than the upper bound w_{high}, it is thresholded to w_{high}. Rather than directly discarding extremely hard exemplars, our method

fully uses these exemplars by adding a constraint to prevent them from assigning an overlarge weight.

Similarly, when a triplet satisfies the critical condition $gap(i,k) = m$, the threshold for weight truncation is computed as:

$$w_{low} = -\log_2\left(\frac{1}{1+\exp\left(-m+\beta\right)}\right). \quad (13)$$

Triplets with weights lower than threshold w_{low} are assigned with a small weight $\frac{\varepsilon}{B}$. Here, ε is a small value. Specifically, we assign small weights to suppress meaningless simple triplets rather than directly discarding them, because the current mini-batch may do not have any hard triplet at all according to our criterion.

Therefore, the weight assigned to a triplet is defined as:

$$w_{hard}\left(A_i, P_i, N_{i,k}\right) := \begin{cases} \frac{\varepsilon}{B}, & gap(i,k) \geq m \\ w_{high}, & gap(i,k) \leq 0. \\ w\left(A_i, P_i, N_{i,k}\right), & \text{otherwise} \end{cases} \quad (14)$$

Orientation regression. In existing cross-view geo-localization benchmark datasets [38, 34], the orientations of an anchor and its corresponding positive exemplar are fixed in the training sets, while shuffled in the test sets. Therefore, angles generated by random rotation can serve as labels for training. To address the problem of unknown orientation, adding an additional orientation regression branch to the main model is shown to be effective for matching ground-level images to aerial images [34]. In this paper, the reweighted orientation regression loss is defined as:

$$L_{OR}(A_i, P_i, N_{i,k}) = w_{hard}(A_i, P_i, N_{i,k}) * (d_R^1(i) + d_R^2(i)). \quad (15)$$

This auxiliary branch is used to regress the sine and cosine value of randomly generated angle θ_i (i.e., the orientation difference between P_i and A_i, deliberately given during training), where $d_R^1(i)$ and $d_R^2(i)$ denote the regression errors of sine and cosine values, respectively. $w_{hard}(A_i, P_i, N_{i,k})$ is the weight defined by Eq. 14.

Given weights λ_1 and λ_2, the overall HER loss is defined as the combination of the main loss and the auxiliary loss:

$$L_{HER}\left(A_i, P_i, N_{i,k}\right) = \lambda_1 * L_{hard}\left(A_i, P_i, N_{i,k}\right) + \lambda_2 * L_{OR}\left(A_i, P_i, N_{i,k}\right). \quad (16)$$

4. Experiments and Discussions

4.1. Evaluation Dataset

Our approach is evaluated on two benchmark cross-view datasets [38, 34]. The CVUSA dataset [38] contains 35532 ground-aerial image pairs for training and 8884 image pairs for testing. All ground images are panoramas, both street-view and overhead-view images are high in resolution. The

CVUSA

Vo and Hays (VH)

Figure 6. Ground-to-aerial sample images from the CVUSA dataset [38] and the VH dataset [34].

Vo and Hays' (VH) dataset [34] contains more than 1 million cross-view image pairs collected by Google Map from 11 different cities in the US. 8 subsets are used as the training set and the remaining 3 subsets captured from Denver, Detroit, and Seattle are used for evaluation. All street-view query images are cropped to a fixed size of 230×230. The azimuth angles of cross-view image pairs in the training sets are fixed, while those of the test subsets are unknown. Example images in these two datasets are shown in Figure 6.

4.2. Implementation Details

We integrated our lightweight attention module (i.e., FCAM) into the basic ResNet [9] to obtain the FCANet feature extractor. Specifically, two versions of ResNet with 18 and 34 learning layers were adopted, where each network was formed using only part of ResNet, before the global average pooling layer. Therefore, our Siamese networks are named Siam-FCANet18 and Siam-FCANet34, respectively. To fully use global information in CNN feature maps, a CRN module [14] and an FC layer of proper size were directly connected to FCANet to generate feature vectors without global average pooling. Additionally, we also tested the performance of FCANets with an NetVLAD feature aggregation module [2], and compared it to the original FCANets. The parameter setting of the NetVLAD layer is the same as [11]. Our Siam-FCANets were constructed using two arms of FCANets without weight sharing, which generate CNN features from both ground-view images and aerial images. SGD was used to train our models, with a momentum of 0.9 and a weight decay of 0.0005. The learning rate was started from 0.5×10^{-5} with polynomial decay.

Parameter γ in our loss was set to 0.15, and an exhaustive mini-batch strategy [34] was used to maximize the number of triplets under limited computing resources.

4.3. Comparative Results

To demonstrate the effectiveness of our approach, we compare our approach to existing methods [38, 34, 11, 17, 41] on two benchmark datasets [38, 34]s. Results of existing methods under comparison are obtained from their original publications [38, 11, 41] or open-sourced implementations [34].

Evaluation metrics. Following [38, 34, 11, 17], we use the recall at top 1% as the performance evaluation metric in our experiments. That is, for a given ground-level query image, we retrieve the top 1% closest satellite images from the reference database according to their feature distances. The localization is considered to be successful if the correctly matched satellite image of the current ground-view query image is ranked within the top 1% of the retrieval results.

Comparison to existing approaches. In our approach, feature vectors of Siam-FCANet18 and Siam-FCANet34 were trained with our proposed HER triplet loss. Siamese and triplet AlexNets [15] evaluated in [34] are used as the baseline methods. They were trained with contrastive loss [7] and triplet loss [28], respectively. Siam DBL-Net and Tri DBL-Net proposed in [34] were trained with their soft-margin contrastive and triplet losses, respectively. The Vo method [34] has an additional OR branch. It was trained with an exhaustive soft-margin triplet loss and an auxiliary OR square loss. The CNN model of [38] was trained with the Euclidean loss using only positive pairs. Both CVM-Net-I and CVM-Net-II [11] were trained with their scaled soft-margin ranking loss.

Comparative results are shown in Table 1. It can be seen that our Siam-FCANet18 and Siam-FCANet34 networks trained with HER triplet loss significantly outperform existing methods on benchmark datasets, including the Vo method [34] and CVM-Nets [11]. Note that, the test set of VH [27] consists of three subsets: Denver, Detroit, and Seattle. This is mainly because our approach can produce more discriminative features than state-of-the-art methods [34, 11].

It is also observed that all learning-based methods achieve higher performance on CVUSA [38] than on the VH dataset [34]. For example, Siam-FCANet18 achieves a recall of 77.2% on the Denver subset, but a much higher recall of 98.3% on the CVUSA dataset. This is because panoramas with wide fields-of-view and high resolutions provide more meaningful information for networks to learn better representations for cross-view image recognition. In contrast, it is challenging for deep learning models to learn cross-view features from coarsely cropped images with drastic orientation variations (e.g., the VH dataset).

	Recall@ Top 1%			
	CVUSA	Denver	Detroit	Seattle
Siam-Net [34]	—	21.6%	21.9%	17.7%
Tri-Net [34]	—	43.2%	39.5%	35.5%
Siam DBL-Net [34]	—	48.4%	45.0%	41.8%
Tri DBL-Net [34]	—	49.3%	47.1%	40.0%
Workman *et al.* [38]	34.3%	15.4%	—	—
Zhai *et al.* [41]	43.2%	—	—	—
Vo [34]	83.9%	62.4%	55.8%	48.1%
CVM-Net-I [11]	91.4%	67.9%	—	—
CVM-Net-II [11]	87.2%	66.6%	—	—
Our approach				
Siam-FCANet18	**98.3%**	77.2%	71.5%	68.1%
Siam-FCANet34	**98.3%**	**78.3%**	**71.9%**	**71.1%**

Table 1. Comparative results on two datasets. Here, "-" denotes that the results on the corresponding test set are unavailable.

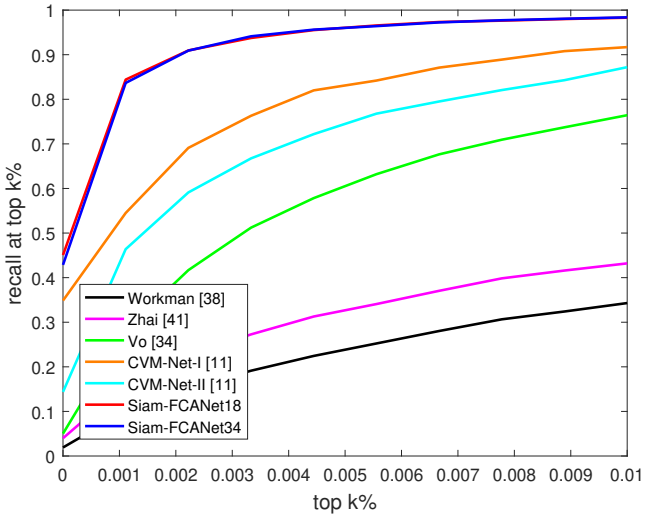

Figure 7. The recall@top k% curves of our approach and existing methods.

Our Siam-FCANet18 model performs closely to Siam-FCANet34. This means the direct stacking of more learning layers in our model cannot bring significant improvement in feature discriminativeness for cross-view recognition.

In Figure 7, we further show the recall at top k% (i.e., from top 1 to top 1%) achieved by our FCANets and existing methods on the CVUSA dataset [38]. It is clear our Siam-FCANet18 and Siam-FCANet34 models trained with HER triplet loss achieve comparable performance and outperform all existing methods by a large margin.

4.4. Ablation Studies

In this section, we conduct ablation studies to test the effectiveness of our designs.

FCAM. To evaluate the effectiveness of our dual attention module, we removed FCAM from each building block of the FCANet18 and FCANet34 extractors, resulting in two plain Siamese networks (namely, Siam-PNet18

	Recall@ Top 1%			
	CVUSA	Denver	Detroit	Seattle
Without FCAM				
Siam-PNet18	97.7%	76.7%	70.6%	68.6%
Siam-PNet34	98.0%	77.0%	71.8%	69.5%
With FCAM				
Siam-FCANet18	98.3%	77.2%	71.5%	68.1%
Siam-FCANet34	98.3%	78.3%	71.9%	71.1%

Table 2. Ablation study on the FCAM module.

and Siam-PNet34). Both Siam-FCANets and Siam-PNets were trained with our hard exemplar reweighting triplet loss. Comparative results on the CVUSA [38] and VH datasets [34] are shown in Table 2.

It can be seen that models integrated with FCAM outperform plain models on most of the test sets. Comparing Siam-FCANet34 to Siam-PNet34, the recalls are improved by 0.3% and 1.3% on the CVUSA [38] and Denver [34] test sets, respectively.

FC layer with orientation regression vs NetVLAD layer. We used two different ways to aggregate CNN features. In the original Siam-FCANets, we used an FC layer in each arm, and further added an auxiliary orientation regression branch to impose orientation invariance. Alternatively, we used the NetVLAD layer [2] to form two new models, i.e., Siam-VFCANet18 and Siam-VFCANet34. The results are shown in Table 3.

	Recall@ Top 1%			
	CVUSA	Denver	Detroit	Seattle
NetVLAD				
Siam-VFCANet18	93.9%	70.4%	63.6%	60.3%
Siam-VFCANet34	92.6%	67.0%	59.1%	60.4%
FC layer with orientation regression				
Siam-FCANet18	98.3%	77.2%	71.5%	68.1%
Siam-FCANet34	98.3%	78.3%	71.9%	71.1%

Table 3. Comparison of two different CNN feature aggregation methods.

Comparative results demonstrate that models with an FC layer and additional orientation regression significantly outperform models with an NetVLAD layer on all datasets. This means, directly using FC layer with an extra auxiliary OR learning branch can learn orientation invariance and generate better representations than the clustering-based feature aggregation module for cross-view image matching.

Our HER triplet loss vs exhaustive soft-margin triplet loss. To further demonstrate the effectiveness of our HER triplet loss, we compare it to the exhaustive soft-margin triplet loss [34] on two datasets [38, 34]. For fair comparison, both losses were supplemented with an auxiliary orientation regression loss, and were used to train two versions of Siam-FCANets separately. Additionally, we also trained the Vo network [34] with our loss, and compare it to its original method trained with the exhaustive soft-margin triplet loss.

	Recall@ Top 1%			
	CVUSA	Denver	Detroit	Seattle
Exhaustive soft-margin triplet loss [34]				
Vo Network [34]	83.9%	62.4%	55.8%	48.1%
Siam-FCANet18	95.1%	75.3%	69.5%	66.2%
Siam-FCANet34	96.5%	76.9%	70.6%	69.7%
Our loss				
Vo network [34]	86.7%	65.9%	58.9%	51.7%
Siam-FCAMNet18	98.3%	77.2%	71.5%	68.1%
Siam-FCAMNet34	98.3%	78.3%	71.9%	71.1%

Table 4. Comparative results achieved by our loss and the exhaustive soft-margin triplet loss.

It can be seen from Table 4 that our loss significantly outperforms the exhaustive soft-margin triplet loss on all datasets. These results demonstrate that our hard exemplar reweighting loss can improve network training and learn more discriminative features for cross-view image geo-localization.

Multiple rotation samples. In the VH dataset [34], all matched cross-view image pairs in the training subsets are well aligned in azimuth. However, images in test subsets are completely shuffled by full degrees of random rotation. We further tested our methods using features averaged from multiple samples with different rotations. Table 5 shows the recall achieved by our approach with multiple rotation samples. It can be seen that using multiple rotation samples improves performance under unknown orientations.

	Recall@ Top 1%		
	Denver	Detroit	Seattle
Original samples			
Siam-FCANet18	77.2%	71.5%	68.1%
Siam-FCANet34	78.3%	71.9%	71.1%
16 rotation samples			
Siam-FCANet18	79.0%	73.0%	69.2%
Siam-FCANet34	80.1%	72.7%	73.3%

Table 5. Comparative results achieved by different networks trained with original samples and 16 rotated samples.

5. Conclusion

In this paper, we have proposed a cross-view geo-localization method by matching ground-view images to aerial images. We proposed a new triplet loss to achieve online end-to-end hard exemplar mining based on exemplar reweighting. Our loss can adaptively focus on useful hard triplets while suppressing useless simple triples. Besides, we introduced a lightweight dual attention module to further improve the representation capability of CNN features. We tested our method on two existing benchmark datasets. Experimental results demonstrated that our method significantly outperforms the state-of-the-art approaches.

References

[1] Hani Altwaijry, Eduard Trulls, James Hays, Pascal Fua, and Serge J. Belongie. Learning to match aerial images with deep attentive architectures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3547, 2016. 2

[2] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 1, 2, 4, 6, 8

[3] Mayank Bansal, Kostas Daniilidis, and Harpreet S. Sawhney. Ultra-wide baseline facade matching for geo-localization. In *IEEE International Conference on Computer Vision*, pages 175–186, 2012. 2

[4] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with aerial image databases. In *ACM International Conference on Multimedia*, pages 1125–1128. ACM, 2011. 2

[5] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Deep visual-semantic quantization for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 916–925, 2017. 4

[6] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian D. Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *IEEE International Conference on Robotics and Automation*, pages 3223–3230, 2017. 1

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–546, 2005. 2, 7

[8] Ben Harwood, Vijay Kumar B. G, Gustavo Carneiro, Ian D. Reid, and Tom Drummond. Smart mining for deep metric learning. In *IEEE International Conference on Computer Vision*, pages 2840–2848, 2017. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 4, 6

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[11] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 6, 7

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2

[13] Hyo Jin Kim, Enrique Dunn, and Jan Michael Frahm. Predicting good features for image geo-localization using per-bundle VLAD. In *IEEE International Conference on Computer Vision*, pages 1170–1178, 2015. 1

[14] Hyo Jin Kim, Enrique Dunn, and Jan Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3251–3260, 2017. 1, 4, 6

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 7

[16] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2

[17] Tsung-Yi Lin, Yin Cui, Serge J. Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5007–5015, 2015. 1, 2, 7

[18] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2999–3007, 2017. 3

[19] Jiawei Liu, Zheng-Jun Zha, Q. I. Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet CNN for person re-identification. In *ACM international conference on Multimedia*, pages 192–196, 2016. 4

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

[21] Colin McManus, Winston Churchill, William P. Maddern, Alexander D. Stewart, and Paul Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *IEEE International Conference on Robotics and Automation*, pages 901–906, 2014. 1

[22] 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, 2014. 1

[23] Masafumi Noda, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Yoshiko Kojima, and Takashi Naito. Vehicle ego-localization by matching in-vehicle camera images to an aerial image. In *IEEE International Conference on Computer Vision*, pages 163–173, 2010. 2

[24] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. 3

[25] Filip Radenovic, 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, 2016. 1

[26] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2

[27] Olivier Saurer, Georges Baatz, Kevin Köser, L'ubor Ladický, and Marc Pollefeys. Image based geo-localization in the alps. *International Journal of Computer Vision*, 116(3):213–225, 2016. 1

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1, 4, 5, 7

[29] Turgay Senlet and Ahmed M. Elgammal. A framework for global vehicle localization using stereo images and satellite and road maps. In *IEEE International Conference on Computer Vision Workshops*, pages 2034–2041, 2011. 2

[30] Niko Suenderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robotics: Science and Systems*, volume 11, 2015. 1

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1

[32] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1998–2006, 2017. 1, 2

[33] Anirudh Viswanathan, Bernardo Rodrigues Pires, and Daniel Huber. Vision based robot localization by ground to satellite matching in gps-denied situations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 192–198, 2014. 2

[34] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision*, pages 494–509, 2016. 1, 2, 4, 5, 6, 7, 8

[35] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1288–1296, 2016. 4

[36] Xun Wang, Xintong Han, Weiling Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 3

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *European Conference on Computer Vision*, pages 3–19, 2018. 3, 4

[38] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 1, 2, 6, 7, 8

[39] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *IEEE International Conference on Computer Vision*, pages 2859–2867, 2017. 3

[40] Amir Roshan Zamir and Mubarak Shah. Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1546–1558, 2014. 1

[41] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4132–4140, 2017. 2, 7