University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization

Zhedong Zheng Yunchao Wei Yi Yang*
SUSTech-UTS Joint Centre of CIS, Southern University of Science and Technology
ReLER, AAII, University of Technology Sydney
zhedong.zheng@student.uts.edu.au,yunchao.wei@uts.edu.au,yi.yang@uts.edu.au

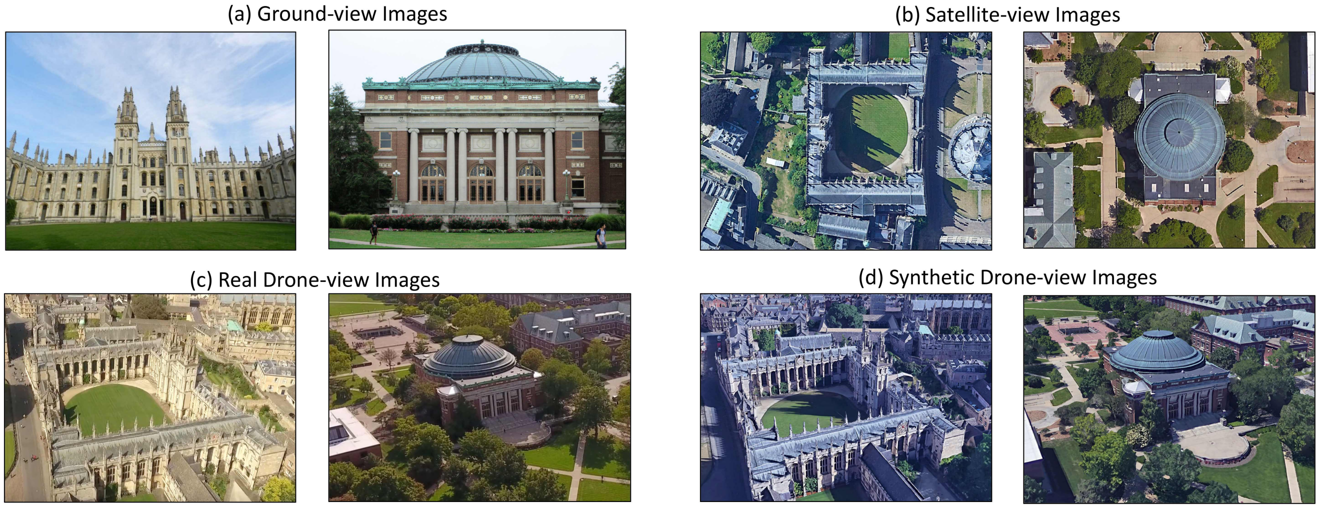
(a) Ground-view Images
(b) Satellite-view Images
(c) Real Drone-view Images
(d) Synthetic Drone-view Images

Figure 1: It is challenging, even for a human, to associate (a) ground-view images with (b) satellite-view images. In this paper, we introduce a new dataset based on the third platform, *i.e.*, drone, to provide real-life viewpoints and intend to bridge the visual gap against views. (c) Here we show two real drone-view images collected from public drone flights on Youtube [1, 8]. (d) In practice, we use the synthetic drone-view camera to simulate the real drone flight. It is based on two concerns. First, the collection expense of real drone flight is unaffordable. Second, the synthetic camera has a unique advantage in the manipulative viewpoint. Specifically, the 3D engine in Google Earth is utilized to simulate different viewpoints in the real drone camera.

ABSTRACT

We consider the problem of cross-view geo-localization. The primary challenge is to learn the robust feature against large viewpoint changes. Existing benchmarks can help, but are limited in the number of viewpoints. Image pairs, containing two viewpoints, *e.g.*, satellite and ground, are usually provided, which may compromise the feature learning. Besides phone cameras and satellites, in this paper, we argue that drones could serve as the third platform to deal with the geo-localization problem. In contrast to traditional ground-view images, drone-view images meet fewer obstacles, *e.g.*, trees, and provide a comprehensive view when flying around the target place. To verify the effectiveness of the drone platform, we introduce a new multi-view multi-source benchmark for drone-based geo-localization, named University-1652. University-1652 contains data from three platforms, *i.e.*, synthetic drones, satellites and ground cameras of 1, 652 university buildings around the world. To our knowledge, University-1652 is the first drone-based geo-localization dataset and enables two new tasks, *i.e.*, drone-view target localization and drone navigation. As the name implies, drone-view target localization intends to predict the location of the target place via drone-view images. On the other hand, given a satellite-view query image, drone navigation is to drive the drone to the area of interest in the query. We use this dataset to analyze a variety of off-the-shelf CNN features and propose a strong CNN baseline on this challenging dataset. The experiments show that University-1652 helps the model to learn viewpoint-invariant features and also has good generalization ability in real-world scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; **Image representations**.

KEYWORDS

Drone, Geo-localization, Benchmark, Image Retrieval

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413896

ACM Reference Format:
Zhedong Zheng Yunchao Wei Yi Yang* . 2020. University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413896

1 INTRODUCTION

The opportunity for cross-view geo-localization is immense, which could enable subsequent tasks, such as, agriculture, aerial photography, navigation, event detection and accurate delivery [3, 39, 45]. Most previous works regard the geo-localization problem as a subtask of image retrieval [2, 16, 18, 21, 31, 32, 36, 38]. Given one query image taken at one view, the system aims at finding the most relevant images in another view among large-scale candidates (gallery). Since candidates in the gallery, especially aerial-view images, are annotated with the geographical tag, we can predict the localization of the target place according to the geo-tag of retrieval results.

In general, the key to cross-view geo-localization is to learn a discriminative image representation, which is invariant to visual appearance changes caused by viewpoints. Currently, most existing datasets usually provide image pairs and focus on matching the images from two different platforms, *e.g.*, phone cameras and satellites [18, 40]. As shown in Figure 1 (a) and (b), the large visual difference between the two images, *i.e.*, ground-view image and satellite-view image, is challenging to matching even for a human. The limited two viewpoints in the training set may also compromise the model to learn the viewpoint-invariant feature.

In light of the above discussions, it is of importance to (1) introduce a multi-view dataset to learn the viewpoint-invariant feature and bridge the visual appearance gap, and (2) design effective methods that fully exploit the rich information contained in multi-view data. With the recent development of the drone [11, 15, 45], we reveal that the drone could serve as a primary data collection platform for cross-view geo-localization (see Figure 1 (c) and (d)). Intuitively, drone-view data is more favorable because drones could be motivated to capture rich information of the target place. When flying around the target place, the drone could provide comprehensive views with few obstacles. In contrast, the conventional ground-view images, including *panorama*, inevitably may face occlusions, *e.g.*, trees and surrounding buildings.

However, large-scale real drone-view images are hard to collect due to the high cost and privacy concerns. In light of the recent practice using synthetic training data [14, 17, 26, 37], we propose a multi-view multi-source dataset called University-1652, containing synthetic drone-view images. University-1652 is featured in several aspects. First, it contains multi-view images for every target place. We manipulate the drone-view engine to simulate images of different viewpoints around the target, which results in 54 drone-view images for every place in our dataset. Second, it contains data from multiple sources. Besides drone-view images, we also collect satellite-view images and ground-view images as reference. Third, it is large-scale, containing 50, 218 training images in total, and has 71.64 images per class on average. The images in the benchmark are captured over 1, 652 buildings of 72 universities. More detailed descriptions will be given in Section 3. Finally, University-1652 enables two new tasks, *i.e.*, drone-view target localization and drone navigation.

Task 1: Drone-view target localization. (Drone → Satellite) Given one drone-view image or video, the task aims to find the most similar satellite-view image to localize the target building in the satellite view.

Task 2: Drone navigation. (Satellite → Drone) Given one satellite-view image, the drone intends to find the most relevant place (drone-view images) that it has passed by. According to its flight history, the drone could be navigated back to the target place.

In the experiment, we regard the two tasks as cross-view image retrieval problems and compare the generic feature trained on extremely large datasets with the viewpoint-invariant feature learned on the proposed dataset. We also evaluate three basic models and three different loss terms, including contrastive loss [16, 35, 43], triplet loss [5, 6], and instance loss [42]. Apart from the extensive evaluation of the baseline method, we also test the learned model on real drone-view images to evaluate the scalability of the learned feature. Our results show that University-1652 helps the model to learn the viewpoint-invariant feature, and reaches a step closer to practice. Finally, the University-1652 dataset, as well as code for baseline benchmark, will be made publicly available for fair use.

2 RELATED WORK

2.1 Geo-localization Dataset Review

Most previous geo-localization datasets are based on image pairs, and target matching the images from two different platforms, such as phone cameras and satellites. One of the earliest works [16] proposes to leverage the public sources to build image pairs for the ground-view and aerial-view images. It consists of 78k image pairs from two views, *i.e.*, 45° bird view and ground view. Later, in a similar spirit, Tian *et al.* [31] collect image pairs for urban localization. Differently, they argue that the buildings could serve as an important role to urban localization problem, so they involve building detection into the whole localization pipeline. Besides, the two recent datasets, *i.e.*, CVUSA [40] and CVACT [18], study the problem of matching the panoramic ground-view image and satellite-view image. It could conduct user localization when Global Positioning System (GPS) is unavailable. The main difference between the former two datasets [16, 31] and the later two datasets [18, 40] is that the later two datasets focus on localizing the user, who takes the photo. In contrast, the former two datasets and our proposed dataset focus on localizing the target in the photo. Multiple views towards the target, therefore, are more favorable, which could drive the model to understand the structure of the target as well as help ease the matching difficulty. The existing datasets, however, usually provide the two views of the target place. Different from the existing datasets, the proposed dataset, University-1652, involves more views of the target to boost the viewpoint-invariant feature learning.

2.2 Deeply-learned Feature for Geo-localization

Most previous works treat the geo-localization as an image retrieval problem. The key of the geo-localization is to learn the

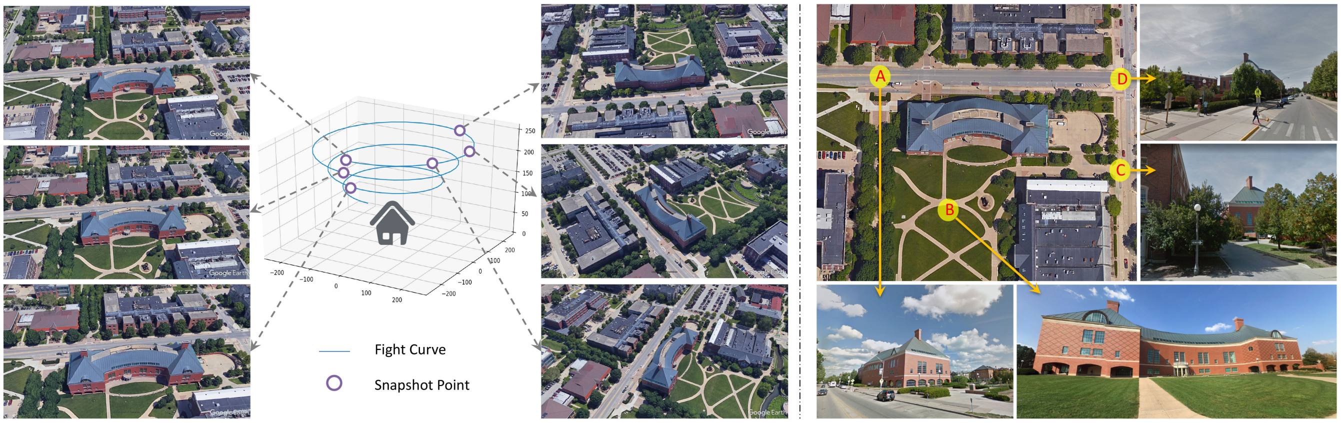

Figure 2: (a) The drone flight curve toward the target building. When flying around the building, the synthetic drone-view camera could capture rich information of the target, including scale and viewpoint variants. (b) The ground-view images are collected from street-view cameras to obtain different facets of the building as well. It simulates real-world photos when people walk around the building.

view-point invariant representation, which intends to bridge the gap between images of different views. With the development of the deeply-learned model, convolutional neural networks (CNNs) are widely applied to extract the visual features. One line of works focuses on metric learning and builds the shared space for the images collected from different platforms. Workman *et al.* show that the classification CNN pre-trained on the Place dataset [44] can be very discriminative by itself without explicitly fine-tuning [34]. The contrastive loss, pulling the distance between positive pairs, could further improve the geo-localization results [16, 35]. Recently, Liu *et al.* propose Stochastic Attraction and Repulsion Embedding (SARE) loss, minimizing the KL divergence between the learned and the actual distributions [19]. Another line of works focuses on the spatial misalignment problem in the ground-to-aerial matching. Vo *et al.* evaluate different network structures and propose an orientation regression loss to train an orientation-aware network [33]. Zhai *et al.* utilize the semantic segmentation map to help the semantic alignment [40], and Hu *et al.* insert the NetVLAD layer [2] to extract discriminative features [12]. Further, Liu *et al.* propose a Siamese Network to explicitly involve the spatial cues, *i.e.*, orientation maps, into the training [18]. Similarly, Shi *et al.* propose a spatial-aware layer to further improve the localization performance [29]. In this paper, since each location has a number of training data from different views, we could train a classification CNN as the basic model. When testing, we use the trained model to extract visual features for the query and gallery images. Then we conduct the feature matching for fast geo-localization.

3 UNIVERSITY-1652 DATASET

3.1 Dataset Description

In this paper, we collect satellite-view images, drone-view images with the simulated drone cameras, and ground-view images for every location. We first select 1, 652 architectures of 72 universities around the world as target locations. We do not select landmarks as the target. The two main concerns are: first, the landmarks usually contain discriminative architecture styles, which may introduce some unexpected biases; second, the drone is usually forbidden to fly around landmarks. Based on the two concerns, we select the buildings on the campus as the target, which is closer to the real-world practice.

It is usually challenging to build the relation between images from different sources. Instead of collecting data and then finding the connections between various sources, we start by collecting the metadata. We first obtain the metadata of university buildings from Wikipedia [1], including building names and university affiliations. Second, we encode the building name to the accurate geo-location, *i.e.*, latitude and longitude, by Google Map. We filter out the buildings with ambiguous search results, and there are 1, 652 buildings left. Thirdly, we project the geo-locations in Google Map to obtain the satellite-view images. For the drone-view images, due to the unaffordable cost of the real-world flight, we leverage the 3D models provided by Google Earth to simulate the real drone camera. The 3D model also provides manipulative viewpoints. To enable the scale changes and obtain comprehensive viewpoints, we set the flight curve as a spiral curve (see Figure 2(a)) and record the flight video with 30 frames per second. The camera flies around the target with three rounds. The height gradually decreases from 256 meters to 121.5 meters, which is close to the drone flight height in the real world [3, 27].

For ground-view images, we first collect the data from the street-view images near the target buildings from Google Map. Specifically, we manually collect the images in different aspects of the building (see Figure 2(b)). However, some buildings do not have the street-view photos due to the accessibility, *i.e.*, most street-view images are collected from the camera on the top of the car. To tackle this issue, we secondly introduce one extra source, *i.e.*, image search engine. We use the building name as keywords to retrieve the relevant images. However, one unexpected observation is that the

[1] https://en.wikipedia.org/wiki/Category:Buildings_and_structures_by_university_or_college

Datasets	University-1652	CVUSA [40]	CVACT [18]	Lin *et al.*[16]	Tian *et al.*[31]	Vo *et al.*[33]
#training	701 × 71.64	35.5k × 2	35.5k × 2	37.5k × 2	15.7k × 2	900k × 2
Platform	Drone, Ground, Satellite	Ground, Satellite	Ground, Satellite	Ground, 45° Aerial	Ground, 45° Aerial	Ground, Satellite
#imgs./location	54 + 16.64 + 1	1 + 1	1+1	1+1	1+1	1+1
Target	Building	User	User	Building	Building	User
GeoTag	✓	✓	✓	✓	✓	✓
Evaluation	Recall@K & AP	Recall@K	Recall@K	PR curves & AP	PR curves & AP	Recall@K

Table 1: Comparison between University-1652 and other geo-localization datasets. The existing datasets usually consider matching the images from two platforms, and provide image pairs. In contrast, our dataset focuses on multi-view images, providing 71.64 images per location. For each benchmark, the table shows the number of training images and average images per location, as well as the availability of collection platform, geo-tag, and evaluation metric.

Split	#imgs	#classes	#universities
Training	50,218	701	33
Query$_{drone}$	37,855	701	39
Query$_{satellite}$	701	701	
Query$_{ground}$	2,579	701	
Gallery$_{drone}$	51,355	951	
Gallery$_{satellite}$	951	951	
Gallery$_{ground}$	2,921	793	

Table 2: Statistics of University-1652 training and test sets, including the image number and the building number of training set, query set and gallery set. We note that there is no overlap in the 33 universities of the training set and 39 universities of test sets.

retrieved images often contain lots of noise images, including indoor environments and duplicates. So we apply the ResNet-18 model trained on the Place dataset [44] to detect indoor images, and follow the setting in [13] to remove the identical images that belong to two different buildings. In this way, we collect 5,580 street-view images and 21,099 common-view images from Google Map and Google Image, respectively. It should be noted that images collected from Google Image only serve as an extra training set but a test set.

Finally, every building has 1 satellite-view image, 1 drone-view video, and 3.38 real street-view images on average. We crop the images from the drone-view video every 15 frames, resulting in 54 drone-view images. Overall, every building has totally 58.38 reference images. Further, if we use the extra Google-retrieved data, we will have 16.64 ground-view images per building for training. Compared with existing datasets (see Table 1), we summarize the new features in University-1652 into the following aspects:

1) Multi-source: University-1652 contains the data from three different platforms, *i.e.*, satellites, drones and phone cameras. To our knowledge, University-1652 is the first geo-localization dataset, containing drone-view images.

2) Multi-view: University-1652 contains the data from different viewpoints. The ground-view images are collected from different facets of target buildings. Besides, synthetic drone-view images capture the target building from various distances and orientations.

3) More images per class: Different from the existing datasets that provide image pairs, University-1652 contains 71.64 images per location on average. During the training, more multi-source & multi-view data could help the model to understand the target structure as well as learn the viewpoint-invariant features. At the testing stage, more query images also enable the multiple-query setting. In the experiment, we show that multiple queries could lead to a more accurate target localization.

3.2 Evaluation Protocol

The University-1652 has 1,652 buildings in total. There are 1,402 buildings containing all three views, *i.e.*, satellite-view, drone-view and ground-view images, and 250 buildings that lack either 3D model or street-view images. We evenly split the 1,402 buildings into the training and test sets, containing 701 buildings of 33 Universities, 701 buildings of the rest 39 Universities. **We note that there are no overlapping universities in the training and test sets.** The rest 250 buildings are added to the gallery as distractors. More detailed statistics are shown in Table 2. Several previous datasets [18, 33, 40] adopt the Recall@K, whose value is 1 if the first matched image has appeared before the K-th image. Recall@K is sensitive to the position of the first matched image, and suits for the test set with only one true-matched image in the gallery. In our dataset, however, there are multiple true-matched images of different viewpoints in the gallery. The Recall@K could not reflect the matching result of the rest ground-truth images. We, therefore, also adopt the average precision (AP) in [16, 31]. The average precision (AP) is the area under the PR (Precision-Recall) curve, considering all ground-truth images in the gallery. Besides Recall@K, we calculate the AP and report the mean AP value of all queries.

4 CROSS-VIEW IMAGE MATCHING

Cross-view image matching could be formulated as a metric learning problem. The target is to map the images of different sources to a shared space. In this space, the embeddings of the same location should be close, while the embeddings of different locations should be apart.

4.1 Feature Representations

There are no “standard” feature representations for the multi-source multi-view dataset, which demands robust features with good scalability towards different kinds of input images. In this work, we mainly compare two types of features: (1) the generic deep-learned features trained on extremely large datasets, such as ImageNet [7], Place-365 [44], and SfM-120k [24]; (2) the learned feature on our dataset. For a fair comparison, the backbone of all networks is ResNet-50 [9] if not specified. More details are in Section 5.2. Next, we describe the learning method on our data in the following section.

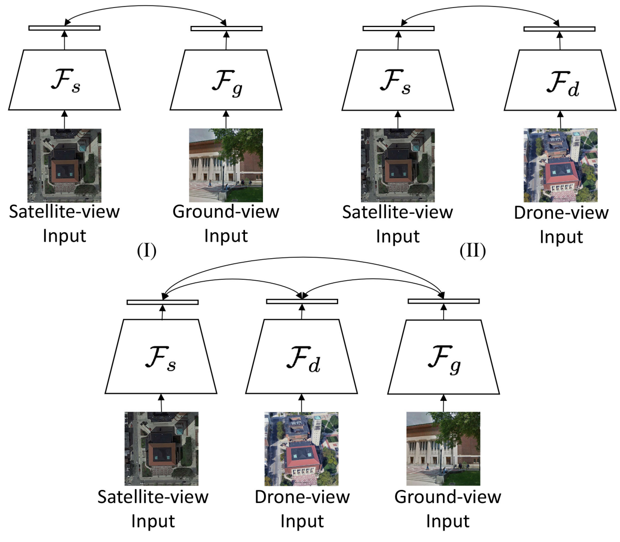

Figure 3: The basic model architectures for cross-view matching. Since the low-level patterns of different data are different, we apply multi-branch CNN to extract high-level features and then build the relation on the high-level features. (I) Model-I is a two-branch CNN model, which only considers the satellite-view and ground-view image matching; (II) Model-II is a two-branch CNN model, which only considers the satellite-view and drone-view image matching; (III) Model-III is a three-branch CNN model, which fully utilizes the annotated data, and considers the images of all three platforms. There are no "standard" methods to build the relationship between the data of multiple sources. Our baseline model applies the instance loss [42] and we also could adopt other loss terms, *e.g.*, triplet loss [5, 6] and contrastive loss [16, 35, 43].

4.2 Network Architecture and Loss Function

The images from different sources may have different low-level patterns, so we denote three different functions $\mathcal{F}_s$, $\mathcal{F}_g$, and $\mathcal{F}_d$, which project the input images from satellites, ground cameras and drones to the high-level features. Specifically, to learn the projection functions, we follow the common practice in [16, 18], and adopt the two-branch CNN as one of our basic structures. To verify the priority of the drone-view images to the ground-view images, we introduce two basic models for different inputs (see Figure 3 (I),(II)). Since our dataset contains data from three different sources, we also extend the basic model to the three-branch CNN to fully leverage the annotated data (see Figure 3 (III)).

To learn the semantic relationship, we need one objective to bridge the gap between different views. Since our datasets provide multiple images for every target place, we could view every place as one class to train a classification model. In light of the recent development in image-language bi-directional retrieval, we adopt one classification loss called instance loss [42] to train the baseline. The main idea is that a shared classifier could enforce the images of different sources mapping to one shared feature space. We denote x_s, x_d, and x_g as three images of the location c, where x_s, x_d, and x_g are the satellite-view image, drone-view image and ground-view image, respectively. Given the image pair $\{x_s, x_d\}$ from two views, the basic instance loss could be formulated as:

$$p_s = softmax(W_{share} \times \mathcal{F}_s(x_s)), \tag{1}$$

$$L_s = -\log(p_s(c)), \tag{2}$$

$$p_d = softmax(W_{share} \times \mathcal{F}_d(x_d)), \tag{3}$$

$$L_d = -\log(p_d(c)), \tag{4}$$

where W_{share} is the weight of the last classification layer. $p(c)$ is the predicted possibility of the right class c. Different from the conventional classification loss, the shared weight W_{share} provides a soft constraint on the high-level features. We could view the W_{share} as one linear classifier. After optimization, different feature spaces are aligned with the classification space. In this paper, we further extend the basic instance loss to tackle the data from multiple sources. For example, if one more view is provided, we only need to include one more criterion term:

$$p_g = softmax(W_{share} \times \mathcal{F}_g(x_g)), \tag{5}$$

$$L_g = -\log(p_g(c)), \tag{6}$$

$$L_{total} = L_s + L_d + L_g. \tag{7}$$

Note that we keep W_{share} for the data from extra sources. In this way, the soft constraint also works on extra data. In the experiment, we show that the instance loss objective L_{total} works effectively on the proposed University-1652 dataset. We also compare the instance loss with the widely-used triplet loss [5, 6] and contrastive loss [16, 35, 43] with hard mining policy [10, 20] in Section 5.3.

5 EXPERIMENT

5.1 Implementation Details

We adopt the ResNet-50 [9] pretrained on ImageNet [7] as our backbone model. We remove the original classifier for ImageNet and insert one 512-dim fully-connected layer and one classification layer after the pooling layer. The model is trained by stochastic gradient descent with momentum 0.9. The learning rate is 0.01 for the new-added layers and 0.001 for the rest layers. Dropout rate is 0.75. While training, images are resized to 256 × 256 pixels. We perform simple data augmentation, such as horizontal flipping. For satellite-view images, we also conduct random rotation. When testing, we use the trained CNN to extract the corresponding features for different sources. The cosine distance is used to calculate the similarity between the query and candidate images in the gallery. The final retrieval result is based on the similarity ranking. If not specified, we deploy the Model-III, which fully utilizes the annotated data as the baseline model. We also share the weights of $\mathcal{F}_s$ and $\mathcal{F}_d$, since the two sources from aerial views share some similar patterns.

5.2 Geo-localization Results

To evaluate multiple geo-localization settings, we provide query images from source A and retrieve the relevant images in gallery B. We denote the test setting as $A \rightarrow B$.

Generic features vs. learned features. We evaluate two categories of features: 1) the generic CNN features. Some previous

Training Set	Feature	Drone → Satellite		Satellite → Drone	
	Dim	R@1	AP	R@1	AP
ImageNet [7]	2048	10.11	13.04	33.24	11.59
Place365 [44]	2048	5.21	6.98	20.40	5.42
SfM-120k [24]	2048	12.53	16.08	37.09	10.28
University-1652	512	58.49	63.13	71.18	58.74

Table 3: Comparison between generic CNN features and the learned feature on the University-1652 dataset. The learned feature is shorter than the generic features but yields better accuracy. R@K (%) is Recall@K, and AP (%) is average precision (high is good).

Query → Gallery	R@1	R@5	R@10	AP
Ground → Satellite	1.20	4.61	7.56	2.52
Drone → Satellite	58.49	78.67	85.23	63.13
*m*Ground → Satellite	1.71	6.56	10.98	3.33
*m*Drone → Satellite	69.33	86.73	91.16	73.14

Table 4: Ground-view query vs. drone-view query. *m* denotes multiple-query setting. The result suggests that drone-view images are superior to ground-view images when retrieving satellite-view images.

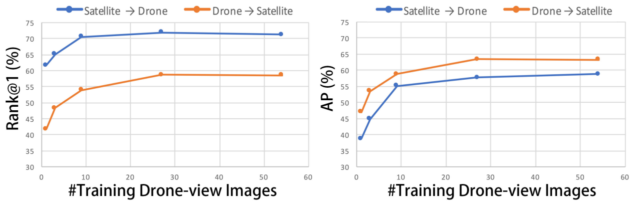

Figure 4: The test accuracy curves when using n training drone-view images per class, $n \in \{1, 3, 9, 27, 54\}$. The two subfigures are the Rank@1 (%) and AP (%) accuracy curves, respectively. The orange curves are for the drone navigation (Satellite → Drone), and the blue curves are for the drone-view target localization (Drone → Satellite).

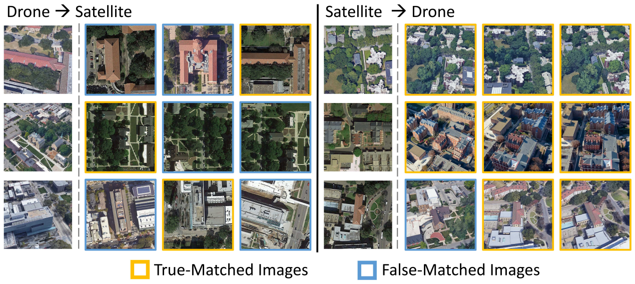

Figure 5: Qualitative image retrieval results. We show the top-3 retrieval results of drone-view target localization (left) and drone navigation (right). The results are sorted from left to right according to their confidence scores. The images in yellow boxes are the true matches, and the images in the blur boxes are the false matches. (Best viewed when zoomed in.)

works [35] show that the CNN model trained on either ImageNet [7] or PlaceNet [44] has learned discriminative feature by itself. We extract the feature before the final classification layer. The feature dimension is 2048. Besides, we also test the widely-used place recognition model [24], whose backbone is ResNet-101. 2) the CNN features learned on our dataset. Since we add one fully-connected layer before the classification layer, our final feature is 512-dim. As shown in Table 3, our basic model achieves much better performance with the shorter feature length, which verifies the effectiveness of the proposed baseline.

Ground-view query vs. drone-view query. We argue that drone-view images are more favorable comparing to ground-view images, since drone-view images are taken from a similar viewpoint, *i.e.*, aerial view, with the satellite images. Meanwhile, drone-view images could avoid obstacles, *e.g.*, trees, which is common in the ground-view images. To verify this assumption, we train the baseline model and extract the visual features of three kinds of data. As shown in Table 4, when searching the relevant satellite-view images, the drone-view query is superior to the ground-view query. Our baseline model using drone-view query has achieved 58.49% Rank@1 and 63.13% AP accuracy.

Multiple queries. Further, in the real-world scenario, one single image could not provide a comprehensive description of the target building. The user may use multiple photos of the target building from different viewpoints as the query. For instance, we could manipulate the drone fly around the target place to capture multiple photos. We evaluate the multiple-query setting by directly averaging the query features [41]. Searching with multiple drone-view queries generally arrives higher accuracy with about 10% improvement in Rank@1 and AP, comparing with the single-query setting (see Table 4). Besides, the target localization using the drone-view queries still achieves better performance than ground-view queries by a large margin. We speculate that the ground-view query does not work well in the single-query setting, which also limits the performance improvement in the multiple-query setting.

Does multi-view data help the viewpoint-invariant feature learning? Yes. We fix the hyper-parameters and only modify the number of drone-view images in the training set. We train five models with n drone-view images per class, where $n \in \{1, 3, 9, 27, 54\}$. As shown in Figure 4, when we gradually involve more drone-view training images from different viewpoints, the Rank@1 accuracy and AP accuracy both increase.

Does the learned model work on the real data? Yes. Due to the cost of collecting real drone-view videos, here we provide a qualitative experiment. We collect one 4K real drone-view video of University-X from Youtube granted by the author. University-X is one of the schools in the test set, and the baseline model has not seen any samples from University-X. We crop images from the video to evaluate the model. In Figure 6, we show the two retrieval results, *i.e.*, Real Drone → Synthetic Drone, Real Drone → Satellite. The first retrieval result is to verify whether our synthetic data well simulates the images in the real drone cameras. We show the top-5 similar images in the test set retrieved by our baseline model. It demonstrates that the visual feature of the real drone-view query is close to the feature of our synthetic drone-view images. The second result on the Real Drone → Satellite is to verify the generalization of our trained model on the real drone-view data. We observe that

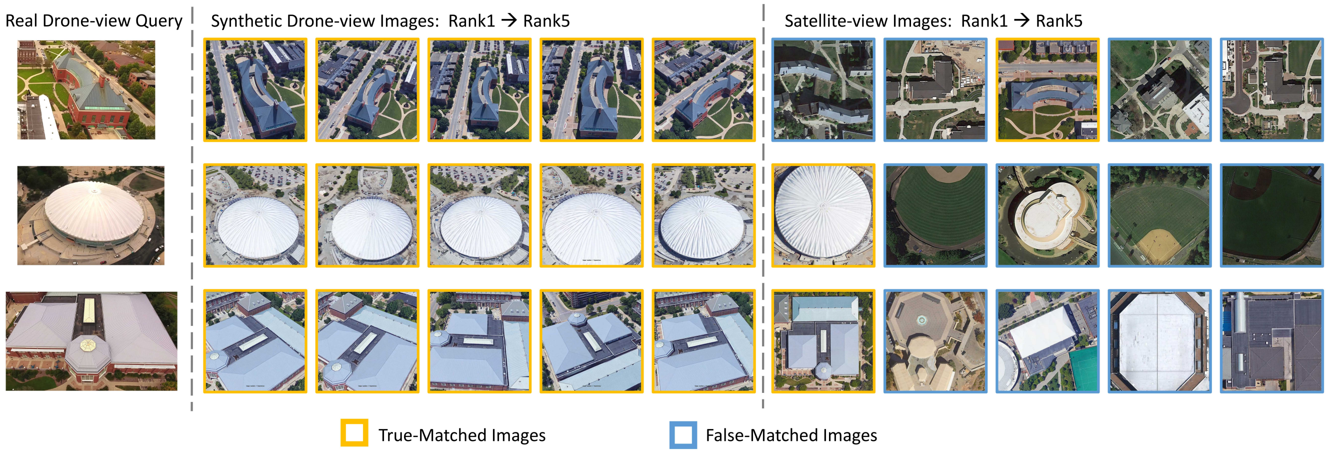
Real Drone-view Query
Synthetic Drone-view Images: Rank1 → Rank5
Satellite-view Images: Rank1 → Rank5
True-Matched Images
False-Matched Images

Figure 6: Qualitative image search results using real drone-view query. We evaluate the baseline model on an unseen university. There are two results: (I) In the middle column, we use the real drone-view query to search similar synthetic drone-view images. The result suggests that the synthetic data in University-1652 is close to the real drone-view images; (II) In the right column, we show the retrieval results on satellite-view images. It verifies that the baseline model trained on University-1652 has good generalization ability and works well on the real-world query.

Loss	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
Contrastive Loss	52.39	57.44	63.91	52.24
Triplet Loss (margin=0.3)	55.18	59.97	63.62	53.85
Triplet Loss (margin=0.5)	53.58	58.60	64.48	53.15
Weighted Soft Margin Triplet Loss	53.21	58.03	65.62	54.47
Instance Loss	58.23	62.91	74.47	59.45

Table 5: Ablation study of different loss terms. To fairly compare the five loss terms, we trained the five models on satellite-view and drone-view data, and hold out the ground-view data. For contrastive loss, triplet loss and weighted soft margin triplet loss, we also apply the hard-negative sampling policy.

Method	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
Not sharing weights	39.84	45.91	50.36	40.71
Sharing weights	58.49	63.31	71.18	58.74

Table 6: Ablation study. With/without sharing CNN weights on University-1652. The result suggests that sharing weights could help to regularize the CNN model.

Image Size	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
256	58.49	63.31	71.18	58.74
384	62.99	67.69	75.75	62.09
512	59.69	64.80	73.18	59.40

Table 7: Ablation study of different input sizes on the University-1652 dataset.

the baseline model has good generalization ability and also works on the real drone-view images for drone-view target localization. The true-matched satellite-view images are all retrieved in the top-5 of the ranking list.

Visualization. For additional qualitative evaluation, we show retrieval results by our baseline model on University-1652 test set (see Figure 5). We can see that the baseline model is able to find the relevant images from different viewpoints. For the false-matched images, although they are mismatched, they share some similar structure pattern with the query image.

5.3 Ablation Study and Further Discussion

Effect of loss objectives. The triplet loss and contrastive loss are widely applied in previous works [5, 6, 16, 35, 43], and the weighted soft margin triplet loss is deployed in [4, 12, 18]. We evaluate these three losses on two tasks, *i.e.*, Drone → Satellite and Satellite → Drone and compare three losses with the instance loss used in our baseline. For a fair comparison, all losses are trained with the same backbone model and only use drone-view and satellite-view data as the training set. For the triplet loss, we also try two common margin values $\{0.3, 0.5\}$. In addition, the hard sampling policy is also applied to these baseline methods during training [10, 20]. As shown in Table 5, we observe that the model with instance loss arrives better performance than the triplet loss and contrastive loss on both tasks.

Effect of sharing weights. In our baseline model, $\mathcal{F}_s$ and $\mathcal{F}_d$ share weights, since two aerial sources have some similar patterns. We also test the model without sharing weights (see Table 6). The performance of both tasks drops. The main reason is that limited satellite-view images (one satellite-view image per location) are prone to be overfitted by the separate CNN branch. When sharing weights, drone-view images could help regularize the model, and the model, therefore, achieves better Rank@1 and AP accuracy.

Model	Training	Drone → Satellite			Satellite → Drone			Ground → Satellite			Satellite → Ground		
	Set	R@1	R@10	AP	R@1	R@10	AP	R@1	R@10	AP	R@1	R@10	AP
Model-I	Satellite + Ground	-	-	-	-	-	-	0.62	5.51	1.60	0.86	5.99	1.00
Model-II	Satellite + Drone	58.23	84.52	62.91	74.47	83.88	59.45	-	-	-	-	-	-
Model-III	Satellite + Drone + Ground	58.49	85.23	63.13	71.18	82.31	58.74	1.20	7.56	2.52	1.14	8.56	1.41

Table 8: Comparison of the three CNN models mentioned in Figure 3. R@K (%) is Recall@K, and AP (%) is average precision (high is good). Model-III that utilizes all annotated data outperforms the other two models in the three of four tasks.

Methods	R@1	R@5	R@10	R@Top1%
Workman [35]	-	-	-	34.40
Zhai [40]	-	-	-	43.20
Vo [33]	-	-	-	63.70
CVM-Net [12]	18.80	44.42	57.47	91.54
Orientation [18]†	27.15	54.66	67.54	**93.91**
Ours	**43.91**	**66.38**	**74.58**	91.78

Table 9: Comparison of results on the two-view dataset CVUSA [40] with VGG-16 backbone. †: The method utilizes extra orientation information as input.

Method	Oxford	Paris	ROxf (M)	RPar (M)	ROxf (H)	RPar (H)
ImageNet	3.30	6.77	4.17	8.20	2.09	4.24
$\mathcal{F}_s$	9.24	13.74	5.83	13.79	2.08	6.40
$\mathcal{F}_g$	25.80	28.77	15.52	24.24	3.69	10.29

Table 10: Transfer learning from University-1652 to small-scale datasets. We show the AP (%) accuracy on Oxford [21], Paris [22], ROxford and RParis [23]. For ROxford and RParis, we report results in both medium (M) and hard (H) settings.

Effect of the image size. Satellite-view images contain the fine-grained information, which may be compressed with small training size. We, therefore, try to enlarge the input image size and train the model with the global average pooling. The dimension of the final feature is still 512. As shown in Table 7, when we increase the input size to 384, the accuracy of both task, drone-view target localization (Drone → Satellite) and drone navigation (Satellite → Drone) increases. However, when we increase the size to 512, the performance drops. We speculate that the larger input size is too different from the size of the pretrained weight on ImageNet, which is 224×224. As a result, the input size of 512 does not perform well.

Different baseline models. We evaluate three different baseline models as discussed in Section 4. As shown in Table 8, there are two main observations: 1). Model-II has achieved better Rank@1 and AP accuracy for drone navigation (Satellite → Drone). It is not surprising since Model-II is only trained on the drone-view and satellite-view data. 2). Model-III, which fully utilizes all annotated data, has achieved the best performance in the three of all four tasks. It could serve as a strong baseline for multiple tasks.

Proposed baseline on the other benchmark. As shown in Table 9, we also evaluate the proposed baseline on one widely-used two-view benchmark, *e.g.*, CVUSA [40]. For fair comparison, we also adopt the 16-layer VGG [30] as the backbone model. We do not intend to push the state-of-the-art performance but to show the flexibility of the proposed baseline, which could also work on the conventional dataset. We, therefore, do not conduct tricks, such as image alignment [28] or feature ensemble [25]. Our intuition is to provide one simple and flexible baseline to the community for further evaluation. Compared with the conventional Siamese network with triplet loss, the proposed method could be easily extended to the training data from N different sources ($N \geq 2$). The users only need to modify the number of CNN branches. Albeit simple, the experiment verifies that the proposed method could serve as a strong baseline and has good scalability toward real-world samples.

Transfer learning from University-1652 to small-scale datasets. We evaluate the generalization ability of the baseline model on two small-scale datasets, *i.e.*, Oxford [21] and Pairs [22]. Oxford and Pairs are two popular place recognition datasets. We directly evaluate our model on these two datasets without finetuning. Further, we also report results on the revised Oxford and Paris datasets (denoted as ROxf and RPar) [23]. In contrast to the generic feature trained on ImageNet [7], the learned feature on University-1652 shows better generalization ability. Specifically, we try two different branches, *i.e.*, $\mathcal{F}_s$ and $\mathcal{F}_g$, to extract features. $\mathcal{F}_s$ and $\mathcal{F}_g$ share the high-level feature space but pay attention to different low-level patterns of inputs from different platforms. $\mathcal{F}_s$ is learned on satellite-view images and drone-view images, while $\mathcal{F}_g$ learns from ground-view images. As shown in Table 10, $\mathcal{F}_g$ has achieved better performance than $\mathcal{F}_s$. We speculate that there are two main reasons. First, the test data in Oxford and Pairs are collected from Flickr, which is closer to the Google Street View images and the images retrieved from Google Image in the ground-view data. Second, $\mathcal{F}_s$ pay more attention to vertical viewpoint changes instead of horizontal viewpoint changes, which are common in Oxford and Paris.

6 CONCLUSION

This paper contributes a multi-view multi-source benchmark called University-1652. University-1652 contains the data from three platforms, including satellites, drones and ground cameras, and enables the two new tasks, *i.e.*, drone-view target localization and drone navigation. We view the two tasks as the image retrieval problem, and present the baseline model to learn the viewpoint-invariant feature. In the experiment, we observe that the learned baseline model has achieved competitive performance towards the generic feature, and shows the feasibility of drone-view target localization and drone navigation. In the future, we will continue to investigate more effective and efficient feature of the two tasks.

Acknowledgement. We would like to thank the real drone-view video providers, Regal Animus and FlyLow. Their data helps us verify our dataset and models.

REFERENCES

[1] Regal Animus. 2015. Fly High 1 "UIUC" - Free Creative Commons Download. https://www.youtube.com/watch?v=jOC-WJW7GAg.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5297–5307.

[3] Simran Brar, Ralph Rabbat, Vishal Raithatha, George Runcie, and Andrew Yu. 2015. Drones for Deliveries. *Sutardja Center for Entrepreneurship & Technology, University of California, Berkeley, Technical Report* 8 (2015), 2015.

[4] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. 2019. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proceedings of the IEEE International Conference on Computer Vision*. 8391–8400.

[5] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11, Mar (2010), 1109–1135.

[6] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. 2018. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing* 27, 8 (2018), 3893–3903.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[8] FlyLow. 2016. Oxford / Amazing flight. https://www.youtube.com/watch?v=bs-rwVI_big.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[11] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. 2017. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE International Conference on Computer Vision*. 4145–4153.

[12] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. 2018. CVM-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7258–7267.

[13] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. 2016. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*. Springer, 301–320.

[14] Peike Li, Yunchao Wei, and Yi Yang. 2020. Meta Parsing Networks: Towards Generalized Few-shot Scene Parsing with Adaptive Metric Learning. In *Proceedings of the 28th ACM international conference on Multimedia*.

[15] Siyi Li and Dit-Yan Yeung. 2017. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[16] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. 2015. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5007–5015.

[17] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. 2018. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4099–4108.

[18] Liu Liu and Hongdong Li. 2019. Lending Orientation to Neural Networks for Cross-view Geo-localization. *CVPR* (2019).

[19] Liu Liu, Hongdong Li, and Yuchao Dai. 2019. Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 2570–2579.

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

[21] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*.

[22] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*.

[23] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *CVPR*.

[24] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1655–1668.

[25] Krishna Regmi and Mubarak Shah. 2019. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 470–479.

[26] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for Data: Ground Truth from Computer Games. In *European Conference on Computer Vision (ECCV) (LNCS)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9906. Springer International Publishing, 102–118.

[27] Troy A Rule. 2015. Airspace in an Age of Drones. *BUL Rev.* 95 (2015), 155.

[28] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. 2019. Spatial-Aware Feature Aggregation for Image based Cross-View Geo-Localization. In *Advances in Neural Information Processing Systems*. 10090–10100.

[29] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. 2020. Optimal Feature Transport for Cross-View Image Geo-Localization. *AAAI Conference on Artificial Intelligence* (2020).

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

[31] Yicong Tian, Chen Chen, and Mubarak Shah. 2017. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3616.

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

[33] Nam N Vo and James Hays. 2016. Localizing and orienting street views using overhead imagery. In *European conference on computer vision*. Springer, 494–509.

[34] Scott Workman and Nathan Jacobs. 2015. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 70–78.

[35] Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*. 3961–3969.

[36] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wei Bian, and Yi Yang. 2019. Progressive learning for person re-identification with one example. *IEEE Transactions on Image Processing* 28, 6 (2019), 2872–2881. https://doi.org/10.1109/TIP.2019.2891895

[37] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. 2018. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 518–534.

[38] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang. 2009. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the 17th ACM international conference on Multimedia*. 175–184.

[39] Qian Yu, Chaofeng Wang, Barbaros Cetiner, Stella X Yu, Frank Mckenna, Ertugrul Taciroglu, and Kincho H Law. 2019. Building Information Modeling and Classification by Visual Learning At A City Scale. *NeurlPS Workshop* (2019).

[40] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. 2017. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 867–875.

[41] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[42] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-Path Convolutional Image-Text Embeddings with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23. https://doi.org/10.1145/3383184

[43] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A Discriminatively Learned CNN Embedding for Person Re-identification. *ACM Transactions on Multimedia Computing Communications and Applications* (2017). https://doi.org/10.1145/3159171

[44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[45] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. 2018. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437* (2018).