Where am I looking at? Joint Location and Orientation Estimation by Cross-View Matching

Yujiao Shi,[1,2] Xin Yu,[1,2,3] Dylan Campbell,[1,2] Hongdong Li[1,2]

[1]Australian National University [2]Australian Centre for Robotic Vision [3]University of Technology Sydney

`firstname.lastname@anu.edu.au`

Abstract

Cross-view geo-localization is the problem of estimating the position and orientation (latitude, longitude and azimuth angle) of a camera at ground level given a large-scale database of geo-tagged aerial (e.g., satellite) images. Existing approaches treat the task as a pure location estimation problem by learning discriminative feature descriptors, but neglect orientation alignment. It is well-recognized that knowing the orientation between ground and aerial images can significantly reduce matching ambiguity between these two views, especially when the ground-level images have a limited Field of View (FoV) instead of a full field-of-view panorama. Therefore, we design a Dynamic Similarity Matching network to estimate cross-view orientation alignment during localization. In particular, we address the cross-view domain gap by applying a polar transform to the aerial images to approximately align the images up to an unknown azimuth angle. Then, a two-stream convolutional network is used to learn deep features from the ground and polar-transformed aerial images. Finally, we obtain the orientation by computing the correlation between cross-view features, which also provides a more accurate measure of feature similarity, improving location recall. Experiments on standard datasets demonstrate that our method significantly improves state-of-the-art performance. Remarkably, we improve the top-1 location recall rate on the CVUSA dataset by a factor of $1.5\times$ for panoramas with known orientation, by a factor of $3.3\times$ for panoramas with unknown orientation, and by a factor of $6\times$ for $180°$-FoV images with unknown orientation.

1. Introduction

Given an image captured by a camera at ground level, it is reasonable to ask: where is the camera and which direction is it facing? Cross-view image geo-localization aims to determine the geographical location and azimuth angle of a query image by matching it against a large geo-tagged
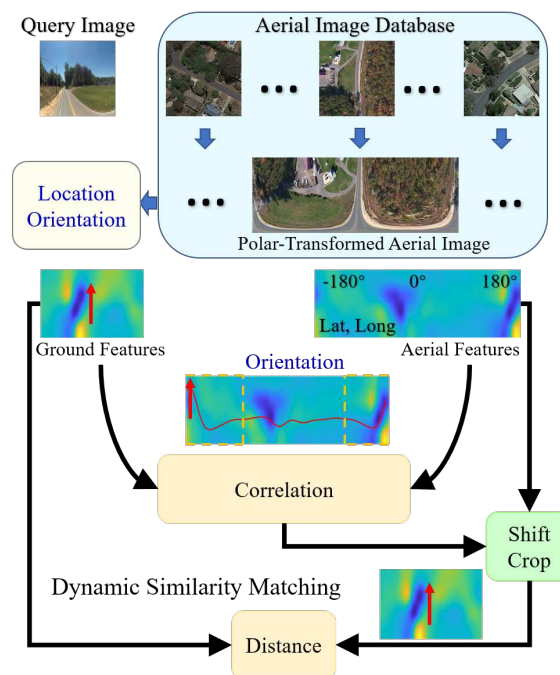

Figure 1. Given a query ground image, we aim to recover its geographical location (latitude and longitude) and orientation (azimuth angle) by matching it against a large database of geo-tagged aerial images. To do so, we first apply a polar transform to aerial images to approximately align them with the ground view images, up to an unknown azimuth angle. We next compute the correlation between ground and aerial features along the axis of azimuth angles (horizontal direction) to estimate the orientation alignment of the ground image with respect to the aerial image. We then shift and crop out the corresponding aerial features, and compute their distance to the ground features, which provides a similarity measure for location retrieval.

satellite map covering the region. Due to the accessibility and extensive coverage of satellite imagery, ground-to-aerial image alignment is becoming an attractive proposition for solving the image-based geo-localization problem.

However, cross-view alignment remains very difficult due to the extreme viewpoint change between ground and

aerial images. The challenges are summarized as follows.

(1) Significant visual differences between the views, including the appearance and projected location of objects in the scene, result in a large domain gap.

(2) The unknown relative orientation between the images, when the northward direction is not known in both, leads to localization ambiguities and increases the search space.

(3) Standard cameras have a limited Field of View (FoV), which reduces the discriminativeness of the ground-view features for cross-view localization, since the image region only covers local information and may match multiple aerial database images.

Existing methods cast this task as a pure location estimation problem and use deep metric learning techniques to learn viewpoint invariant features for matching ground and aerial images. Many approaches require the orientation to be provided, avoiding the ambiguities caused by orientation misalignments [15, 12, 10]. However, the orientation is not always available for ground images in practice. To handle this, some methods directly learn orientation invariant features [18, 5, 2], however they fail to address the large domain gap between ground and aerial images, limiting their localization performance.

To reduce the cross-view domain gap, we explore the geometric correspondence between ground and aerial images. We observe that there are two geometric cues that are statistically significant in real ground images under an equirectangular projection: (i) horizontal lines in the image (parallel to the azimuth axis) have approximately constant depth and so correspond to concentric circles in the aerial image; and (ii) vertical lines in the image have depth that increases with the y coordinate and so correspond to radial lines in the aerial image. To be more specific, if the scene was flat, then a horizontal line in the ground image maps to a circle in the aerial image. We make use of these geometric cues by applying a polar coordinate transform to the aerial images, mapping concentric circles to horizontal lines. This reduces the differences in the projected geometry and hence the domain gap, as shown in Figure 2.

We then employ a two-stream CNN to learn feature correspondences between ground and aerial images. We extract feature volumes that preserve the spatial relationships between features, which is a critical cue for geo-localization. However, orientation misalignments lead to inferior results when using spatially-aware image features. Moreover, it is difficult to match features when a limited FoV is imaged, since the ground image contains only a small sector of the aerial image. Therefore, our intuition is to find the orientation alignment and thus facilitate accurate similarity matching.

As the polar transform projects the aerial image into the ground-view camera coordinate frame, it allows to estimate
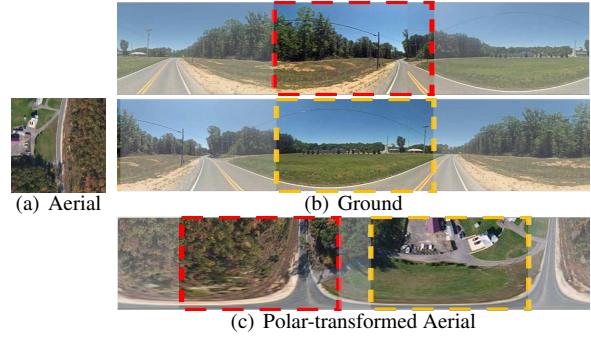

Figure 2. The challenge of cross-view image matching caused by the unknown orientation and limited FoV of the query ground image. The scene content in panoramas captured at the same location but with different azimuth angles (top and middle) is offset, and the image content in a limited FoV image can be entirely different to another image captured from the same location.

the orientation of each ground image with respect to its aerial counterpart by feature correlation. In this paper, we propose a Dynamic Similarity Matching (DSM) module to achieve that goal. To be specific, we compute the correlation between the ground and aerial features in order to generate a similarity score at each angle, marked by the red curve in Figure 1. The position of the similarity score maximum corresponds to the latent orientation of the ground image with respect to the aerial image. If the ground image has a limited FoV, we extract the appropriate local region from the aerial feature representation for localization. By using our DSM module, the feature similarity between the ground and aerial images is measured more accurately. Therefore, our method outperforms the state-of-the-art by a large margin.

The contributions of our work are:

- the first image-based geo-localization method to jointly estimate the position and orientation[1] of a query ground image regardless of its Field of View;

- a Dynamic Similarity Matching (DSM) module to measure the feature similarity of the image pair while accounting for the orientation of the ground image, facilitating accurate localization; and

- extensive experimental results demonstrating that our method achieves significant performance improvements over the state-of-the-art in various geo-localization scenarios.

2. Related Work

Existing cross-view image-based geo-localization aims to estimate the location (latitude and longitude) of a ground image by matching it against a large database of aerial images. Due to the significant viewpoint changes between

---

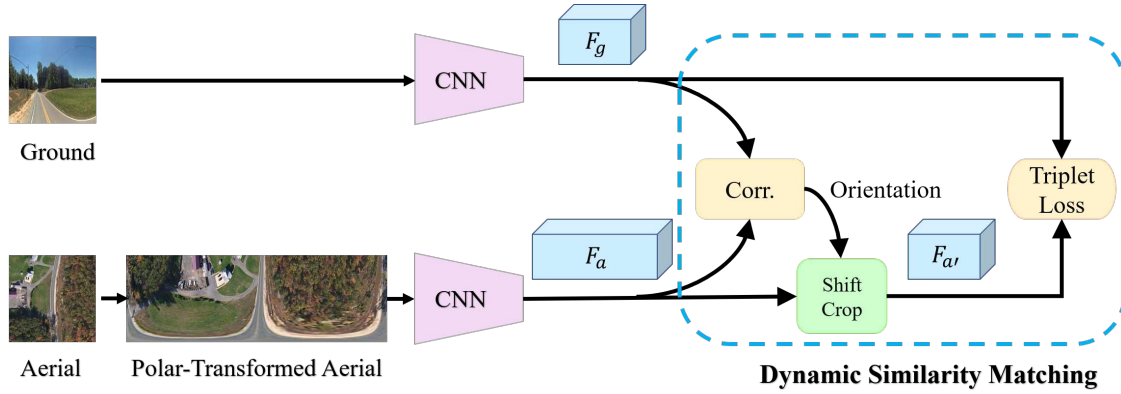[1]Throughout this paper, orientation refers to the 1-DoF azimuth angle.

Figure 3. Flowchart of the proposed method. A polar transform is first applied to the aerial image, and then a two-stream CNN is employed to extract features from ground and polar-transformed aerial images. Given the extracted feature volume representations, the correlation between the two is used to estimate the orientation of the ground image with respect to the aerial image. Next, the aerial features are shifted and cropped to obtain the section that (potentially) corresponds to the ground features. The similarity of the resulting features is then used for location retrieval.

ground and aerial images, hand-crafted feature matching [3, 9, 11] becomes the bottleneck of the performance of cross-view geo-localization. Deep convolutional neural networks (CNNs) have proven their powerful capability on image representations [13]. This motivates recent geo-localization works to extract features from ground and aerial images with CNNs.

Workman and Jacobs [19] first introduced deep features to the cross-view matching task. They used an AlexNet [7] network fine-tuned on Imagenet [13] and Places [22] to extract deep features for cross-view image matching. They demonstrated that further tuning of the aerial branch by minimizing the distance between matching ground and aerial pairs led to better localization performance [20]. Vo and Hays [18] investigated a set of CNN architectures (classification, hybrid, Siamese and triplet CNNs) for matching cross-view images. Considering the orientation misalignments between ground and aerial images, they proposed an auxiliary orientation regression block to let the network learn orientation-aware feature representations, and used multiple aerial images with different orientations during testing phase. To learn orientation invariant features, Hu *et al.* [5] embedded a NetVlad layer [1] on top of a two-branch CNN for cross-view image matching. Cai *et al.* [2] introduced a lightweight attention module to reweight spatial and channel features to obtain more representative descriptors, and then proposed a hard exemplar reweighting triplet loss to improve the quality of network training. They also employed an orientation regression block to force the network to learn orientation-aware features. Sun *et al.* [17] employed capsule networks to encode spatial feature hierarchies for feature representations. Although these methods learned orientation-aware descriptors for localization, they overlooked the domain difference between ground and aerial images.

To bridge the large domain gap between ground and aerial images, Zhai *et al.* [21] learned a transformation matrix between aerial and ground features for predicting ground semantic information from aerial images. Regmi and Shah [12] synthesized an aerial image from a ground one using a generative model, and then fused features of the ground image and the synthesized aerial image as the descriptor for retrieval. Shi *et al.* [15] proposed a feature transport module to map ground features into the aerial domain and then conducted similarity matching. Shi *et al.* [14] also used polar transform to first bridge the geometric domain difference and then a spatial-aware feature aggregation module to select salient features for global feature descriptor representation. However, all those methods require ground images to be panoramas or orientation-aligned. Finally, Liu & Li [10] found that the orientation provided important clues for determining the location of a ground image, and thus explicitly encoded the ground-truth orientation as an additional network input.

In contrast to existing works, we aim to estimate the location and orientation of ground images jointly, since exploring orientation information can facilitate cross-view matching for both panoramas and images with limited FoV.

3. Location and Orientation Estimation by Cross-view Image Matching

In the cross-view image-based geo-localization task, ground images are captured by a camera whose image plane is perpendicular to the ground plane and y axis is parallel to the gravity direction, and aerial images are captured from a camera whose image plane is parallel to the ground plane. Since there are large appearance variations between these two image domains, our strategy is to first reduce the projection differences between the viewpoints and then to

extract discriminative features from the two domains. Furthermore, inspired by how humans localize themselves, we use the spatial relationships between objects as a critical cue for inferring location and orientation. Therefore, we enable our descriptors to encode the spatial relationship among the features, as indicated by F_g and F_a in Figure 3.

Despite the discriminativeness of the spatially-aware features, they are very sensitive to orientation changes. For instance, when the azimuth angle of a ground camera changes, the scene contents will be offset in the ground panorama, and the image content may be entirely different if the camera has a limited FoV, as illustrated in Figure 2. Therefore, finding the orientation of the ground images is crucial to make the spatially-aware features usable. To this end, we propose a dynamic similarity matching (DSM) module, as illustrated in Figure 3. With this module, we not only estimate the orientation of the ground images but also achieve more accurate feature similarity scores, regardless of orientation misalignments and limited FoVs, thus enhancing geo-localization performance.

3.1. A Polar Transform to Bridge the Domain Gap

Since ground panoramas[2] project 360-degree rays onto an image plane using an equirectangular projection, and are orthogonal to the satellite-view images, vertical lines in the ground image correspond to radial lines in the aerial image, and horizontal lines correspond approximately to circles in the aerial image, assuming that the pixels along the line have similar depths, which occurs frequently in practice. This layout correspondence motivates us to apply a polar transform to the aerial images. In this way, the spatial layouts of these two domains can be roughly aligned, as illustrated in Figure 2(b) and Figure 2(c).

To be specific, the polar origin is set to the center of each aerial image, corresponding to the geo-tag location, and the $0°$ angle is chosen as the northward direction, corresponding to the upwards direction of an aligned aerial image. In addition, we constrain the height of the polar-transformed aerial images to be the same as the ground images, and ensure that the angle subtended by each column of the polar transformed aerial images is the same as in the ground images. We apply a uniform sampling strategy along radial lines in the aerial image, such that the innermost and outermost circles of the aerial image are mapped to the bottom and top line of the transformed image respectively.

Formally, let $S_a \times S_a$ represent the size of an aerial image and $H_g \times W_g$ denote the target size of polar transform. The polar transform between the original aerial image points

(x_i^a, y_i^a) and the target polar transformed ones (x_i^t, y_i^t) is

$$
\begin{aligned}
x_i^a &= \frac{S_a}{2} - \frac{S_a}{2}\frac{(H_g - x_i^t)}{H_g}\cos\left(\frac{2\pi}{W_g}y_i^t\right), \\
y_i^a &= \frac{S_a}{2} + \frac{S_a}{2}\frac{(H_g - x_i^t)}{H_g}\sin\left(\frac{2\pi}{W_g}y_i^t\right).
\end{aligned}
\tag{1}
$$

By applying a polar transform, we coarsely bridge the projective geometry domain gap between ground and aerial images. This allows the CNNs to focus on learning the feature correspondences between the ground and polar-transformed aerial images without consuming network capacity on learning the geometric relationship between these two domains.

3.2. A Spatially-Aware Feature Representation

Applying a translation offset along the x axis of a polar-transformed image is equivalent to rotating the aerial image. Hence, the task of learning rotational equivariant features for aerial images becomes learning translational equivariant features, which significantly reduces the learning difficulty for our network since CNNs inherently have the property of translational equivariance [8]. However, since the horizontal direction represents a rotation, we have to ensure that the CNN treats the leftmost and rightmost columns of the transformed image as adjacent. Hence, we propose to use circular convolutions with wrap-around padding along the horizontal direction.

We adopt VGG16 [16] as our backbone network. In particular, the first ten layers of VGG16 are used to extract features from the ground and polar-transformed aerial images. Since the polar transform might introduce distortions along the vertical direction, due to the assumption that horizontal lines have similar finite depths, we modify the subsequent three layers which decrease the height of the feature maps but maintain their width. In this manner, our extracted features are more tolerant to distortions along the vertical direction while retaining information along the horizontal direction. We also decrease the feature channel number to 16 by using these three convolutional layers, and obtain a feature volume of size $4 \times 64 \times 16$. Our feature volume representation is a global descriptor designed to preserve the spatial layout information of the scene, thus increasing the discriminativeness of the descriptors for image matching.

3.3. Dynamic Similarity Matching (DSM)

When the orientation of ground and polar-transformed aerial features are aligned, their features can be compared directly. However, the orientation of the ground images is not always available, and orientation misalignments increase the difficulty of geo-localization significantly, especially when the ground image has a limited FoV. When humans are using a map to relocalize themselves, they determine their location and orientation jointly by compar-

---

[2]Although we use panoramic images as an example, the correspondence relationships between ground and aerial images also apply to images with limited FoV.

ing what they have seen with what they expect to see on the map. In order to let the network mimic this process, we compute the correlation between the ground and aerial features along the azimuth angle axis. Specifically, we use the ground feature as a sliding window and compute the inner product between the ground and aerial features across all possible orientations. Let $F_a \in R^{H \times W_a \times C}$ and $F_g \in R^{H \times W_g \times C}$ denote the aerial and ground features respectively, where H and C indicate the height and channel number of the features, W_a and W_g represent the width of the aerial and ground features respectively, and $W_a \geq W_g$. The correlation between F_a and F_g is expressed as

$$[F_a * F_g](i) = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W_g} F_a(h, (i+w)\%W_a, c)F_g(h, w, c), \quad (2)$$

where $F(h, w, c)$ is the feature response at index (h, w, c), and $\%$ denotes the modulo operation. After correlation computation, the position of the maximum value in the similarity scores is the estimated orientation of the ground image with respect to the polar-transformed aerial one.

When a ground image is a panorama, regardless of whether the orientation is known, the maximum value in the correlation results is directly converted to the L_2 distance by computing $2(1 - \max([F_a * F_g](i))$, where F_a and F_g are L_2-normalized. When a ground image has a limited FoV, we crop the aerial features corresponding to the FoV of the ground image at the position of the maximum similarity score. Then we re-normalize the cropped aerial features and calculate the L_2 distance between the ground and aerial features as the similarity score for matching. Note that if there are multiple maximum similarity scores, we choose one randomly, since this means that the aerial images has symmetries that cannot be disambiguated.

3.4. Training DSM

During the training process, our DSM module is applied to all ground and aerial pairs, whether they are matching or not. For matching pairs, DSM forces the network to learn similar feature embeddings for ground and polar-transformed aerial images with discriminative feature representations along the horizontal direction (*i.e.*, azimuth). In this way, DSM is able to identify the orientation misalignment as well as find the best feature similarity for matching. For non-matching pairs, as it is the most challenging case when they are aligned (*i.e.*, their similarity is larger), our DSM is also used to find the most feasible orientation for a ground image aligning to a non-matching aerial one, and we minimize the maximum similarity of non-matching pairs to make the features more discriminative. Following traditional cross-view localization methods [5, 10, 15], we employ the weighted soft-margin triplet loss [5] to train our

network

$$\mathcal{L} = \log\left(1 + e^{\alpha\left(\|F_g - F_{a'}\|_F - \|F_g - F_{a*'}\|_F\right)}\right), \quad (3)$$

where F_g is the query ground feature, $F_{a'}$ and $F_{a*'}$ indicate the cropped aerial features from the matching aerial image and a non-matching aerial image respectively, and $\|\cdot\|_F$ denotes the Frobenius norm. The parameter α controls the convergence speed of training process; following precedents we set it to 10 [5, 10, 15].

3.5. Implementation Details

We use the first ten convolutional layers in VGG16 with pretrained weights on Imagenet [4], and randomly initialize the parameters in the following three layers for global feature descriptor extraction. The first seven layers are kept fixed and the subsequent six layers are learned. The Adam optimizer [6] with a learning rate of 10^{-5} is employed for training. Following [18, 5, 10, 15], we adopt an exhaustive mini-batch strategy [18] with a batch size of $B = 32$ to create the training triplets. Specifically, for each ground image within a mini-batch, there is one matching aerial image and $(B-1)$ non-matching ones. Thus we construct $B(B-1)$ triplets. Similarly, for each aerial image, there is one matching ground image and $(B-1)$ non-matching ones within a mini-batch, and thus we create another $B(B-1)$ triplets. Hence, we have $2B(B-1)$ triplets in total.

4. Experiments

4.1. Datasets

We carry out the experiments on two standard cross-view datasets, CVUSA [21] and CVACT [10]. They both contain $35,532$ training ground and aerial pairs and $8,884$ testing pairs. Following an established testing protocol [10, 15], we denote the test sets in CVUSA and CVACT as CVUSA and CVACT_val, respectively. CVACT also provides a larger test set, CVACT_test, which contains $92,802$ cross-view image pairs for fine-grained city-scale geo-localization. Note that the ground images in both of the two datasets are panoramas, and all the ground and aerial images are north aligned. Figure 4 presents samples of cross-view image pairs from the two datasets.

Furthermore, we also conduct experiments on ground images with unknown orientation and limited FoV. We use the image pairs in CVUSA and CVACT_val, and randomly rotate the ground images along the azimuth direction and crop them according to a predetermined FoV. The constructed test set with different FoVs as well as our source code are available via https://github.com/shiyujiao/cross_view_localization_DSM.git.

Figure 4. Cross-view image pairs from the CVUSA (top two rows) and CVACT (bottom two rows) datasets. The aerial images are on the left and the ground panoramas are on the right.

4.2. Evaluation Metrics

Location estimation: Following the standard evaluation procedure for cross-view image localization [18, 5, 10, 15, 2, 17, 12], we use the top K recall as the location evaluation metric to examine the performance of our method and compare it with the state-of-the-art. Specifically, given a ground image, we retrieve the top K aerial images in terms of L_2 distance between their global descriptors. The ground image is regarded as successfully localized if its corresponding aerial image is retrieved within the top K list. The percentage of correctly localized ground images is recorded as recall at top K (r@K).

Orientation estimation: The predicted orientation of a query ground image is meaningful only when the ground image is localized correctly. Hence, we evaluate the orientation estimation accuracy of our DSM only on ground images that have been correctly localized by the top-1 recall. In this experiment, when the differences between the predicted orientation of a ground image and its ground-truth orientation is within ±10% of its FoV, the orientation estimation of this ground image is deemed as a success. We record the percentage of ground images for which the orientation is correctly predicted as the orientation estimation accuracy (orien_acc). Since aerial images are often rotationally symmetric, orientation estimation can yield large errors, such as 180° for scenes that look similar in opposite directions. Hence we report the robust median orientation error, denoted as median_error, instead of the mean.

4.3. Localizing Orientation-Aligned Panoramas

We first investigate the location estimation performance of our method and compare it with the state-of-the-art on the standard CVUSA and CVACT datasets, where ground images are orientation-aligned panoramas. In Table 1, we

Table 1. Comparison with existing methods on the CVUSA [21] dataset. Here, "–" denotes that the results on the corresponding evaluation metric are not available as some of the works only use r@1% as the evaluation metric.

Methods	CVUSA			
	r@1	r@5	r@10	r@1%
Workman *et al.* [20]	–	–	–	34.3
Zhai *et al.* [21]	–	–	–	43.2
Vo and Hays [18]	–	–	–	63.7
CVM-NET [5]	22.47	49.98	63.18	93.62
Liu & Li [10]	40.79	66.82	76.36	96.12
Regmi and Shah [12]	48.75	–	81.27	95.98
Siam-FCANet34 [2]	–	–	–	98.3
CVFT [15]	61.43	84.69	90.49	99.02
Ours	**91.96**	**97.50**	**98.54**	**99.67**

Table 2. Comparison with existing methods on the CVACT_val [10] dataset by re-training existing networks using the code provided by the authors.

Methods	CVACT_val			
	r@1	r@5	r@10	r@1%
CVM-NET [5]	20.15	45.00	56.87	87.57
Liu & Li [10]	46.96	68.28	75.48	92.01
CVFT [15]	61.05	81.33	86.52	95.93
Ours	**82.49**	**92.44**	**93.99**	**97.32**

present our results on the CVUSA dataset with the recall rates reported in other works [10, 15, 5, 12, 2]. We also retrain existing networks [5, 10, 15] on the CVACT dataset using source code provided by the authors. The recall results at top-1, top-5, top-10 and top-1% on CVACT_val are presented in Table 2, and the complete r@K performance curves on CVUSA and CVACT_val are illustrated in Figure 5(a) and Figure 5(b), respectively.

Among those comparison methods, [20, 21, 18] are first explorers to apply deep-based methods to cross-view related tasks. CVM-NET [5] and Siam-FCANet34 [2] focus on designing powerful feature extraction networks. Liu & Li [10] introduce the orientation information to networks so as to facilitate geo-localization. However, all of them ignore the domain difference between ground and aerial images, thus leading to inferior performance. Regmi and Shah [12] adopt a conditional GAN to generate aerial images from ground panoramas. Although it helps to bridge the cross-view domain gap, undesired scene contents are also induced in this process. Shi *et al.* [15] propose a cross view feature transport module (CVFT) to better align ground and aerial features. However, it is hard for networks to learn geometric and feature response correspondences simultaneously. In contrast, our polar transform explicitly reduces the projected geometry difference between ground and aerial images, and thus eases the burden of networks. As is clear in Table 1 and Table 2, our method significantly outperforms
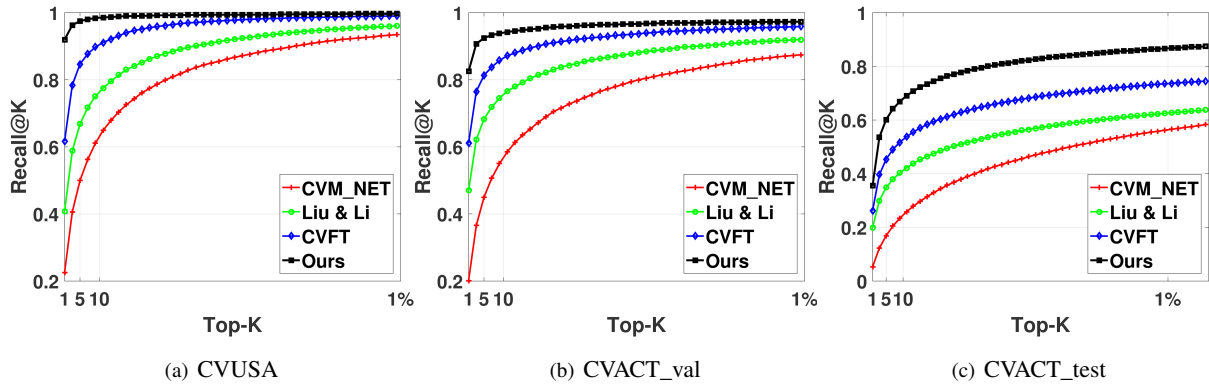
(a) CVUSA   (b) CVACT_val   (c) CVACT_test

Figure 5. Recall comparison at different values of K on the CVUSA, CVACT_val and CVACT_test datasets.

the state-of-the-art methods by a large margin.

Fine-grained localization: We also compare our method with state-of-the-art methods on the CVACT_test dataset. This dataset provides fine-grained geo-tagged aerial images that densely cover a city, and the localization performance is measured in terms of distance (meters). Specifically, a ground image is considered as successfully localized if one of the retrieved top K aerial images is within 5 meters of the ground truth location of the query ground image. Following the evaluation protocol in Liu & Li [10], we plot the percentage of correctly localized ground images (recall) at different values of K in Figure 5(c). Our method achieves superior results compared to state-of-the-art on this extremely challenging test set.

4.4. Localizing With Unknown Orientation and Limited FoV

In this section, we test the performance of our algorithm and other methods, including CVM-NET [5] and CVFT [15], on the CVUSA and CVACT_val datasets in a more realistic localization scenario, where the ground images do not have a known orientation and have a limited FoV. Recall that Liu & Li [10] require orientation information as an input, so we cannot compare with this method.

Location estimation: Since existing methods are only designed to estimate the location of ground images, we only evaluate their location recall performance. In order to evaluate the impact of orientation misalignments and limited FoVs on localization performance, we randomly shift and crop the ground panoramas along the azimuth direction for the CVUSA and CVACT_val datasets. In this manner, we mimic the procedure of localizing images with limited FoV and unknown orientation. The first results column in Table 3 demonstrates the performance of localizing panoramas with unknown orientation. It is clear that our method significantly outperforms all the comparison algorithms, obtaining a $2.34\times$ improvement on CVUSA and a $2.72\times$ improvement on CVACT in terms of r@1. We also conduct

comparisons with the other methods on ground images with FoVs of $180°$ (fish-eye camera), $90°$ (wide-angle camera) and $70°$ (general phone camera) respectively in Table 3. Note that the orientation is also unknown. As illustrated in Figure 6(b), as the FoV of the ground image decreases, the image become less discriminative. This increases the difficulty of geo-localization especially when the orientation is unknown. As indicated in the second, third and fourth results column of Table 3, our method, benefiting from its DSM module, significantly reduces the ambiguity caused by unknown orientations and measures feature similarity more accurately, achieving better performance than the state-of-the-art.

Orientation estimation: As previously mentioned, the experiments of orientation estimation is conducted on ground images which are correctly localized in terms of top-1 retrieved candidates. The first row in Table 4 presents the orientation prediction accuracy of ground images with different FoVs. As indicated in the table, the orientation of almost all ground images with $360°$ and $180°$ FoV is predicted correctly, demonstrating the effectiveness of our DSM module for estimating the orientation of ground images. It is also clear that the matching ambiguity increases as the FoV decreases. Considering that scene contents in an aerial image might be very similar in multiple directions, the orientation estimation can be inaccurate while the estimated location is correct. For instance, a person standing on a road is able to localize their position but will find it difficult to determine their orientation if the view is similar along the road in both directions. We provide an example of this in the supplementary material. Therefore, even when our method estimates orientation inaccurately, it is still possible to localize the position correctly using our DSM module. We also report the median value of the errors (in degrees) between the estimated and ground truth orientation in the second row of Table 4. The estimated errors are very small with respect to the FoV of the image, and so will not negatively affect the localization performance. Figure 6 shows the estimated

Table 3. Comparison of recall rates for localizing ground images with unknown orientation and varying FoVs.

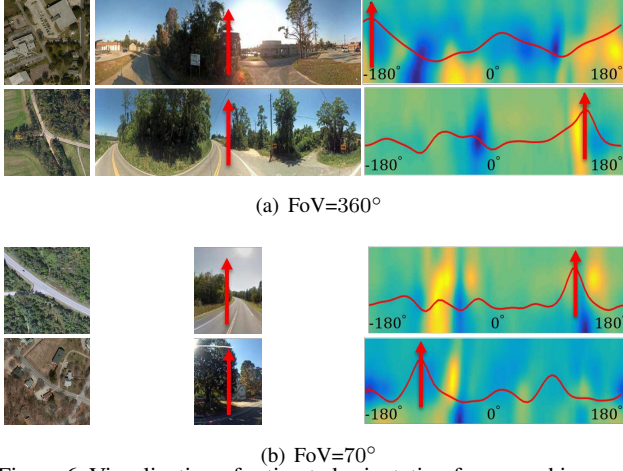| Dataset	Comparison Algorithms	FoV=360°				FoV=180°				FoV=90°				FoV=70°			
		r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
CVUSA	CVM-NET [5]	16.25	38.86	49.41	88.11	7.38	22.51	32.63	75.38	2.76	10.11	16.74	55.49	2.62	9.30	15.06	21.77
	CVFT [15]	23.38	44.42	55.20	86.64	8.10	24.25	34.47	75.15	4.80	14.84	23.18	61.23	3.79	12.44	19.33	55.56
	Ours	**78.11**	**89.46**	**92.90**	**98.50**	**48.53**	**68.47**	**75.63**	**93.02**	**16.19**	**31.44**	**39.85**	**71.13**	**8.78**	**19.90**	**27.30**	**61.20**
CVACT_val	CVM-NET [5]	13.09	33.85	45.69	81.80	3.94	13.69	21.23	59.22	1.47	5.70	9.64	38.05	1.24	4.98	8.42	34.74
	CVFT [15]	26.79	46.89	55.09	81.03	7.13	18.47	26.83	63.87	1.85	6.28	10.54	39.25	1.49	5.13	8.19	34.59
	Ours	**72.91**	**85.70**	**88.88**	**95.28**	**49.12**	**67.83**	**74.18**	**89.93**	**18.11**	**33.34**	**40.94**	**68.65**	**8.29**	**20.72**	**27.13**	**57.08**

(a) FoV=360°

(b) FoV=70°

Figure 6. Visualization of estimated orientation for ground images with FoV = 360° and 70°. In each of the subfigures, the aerial images are on the left and the ground images are in the middle. We visualize the polar transformed-aerial features and the correlation results (red curves) in the right column. The positions of the correlation maxima in the curves corresponds to the orientation of the ground images.

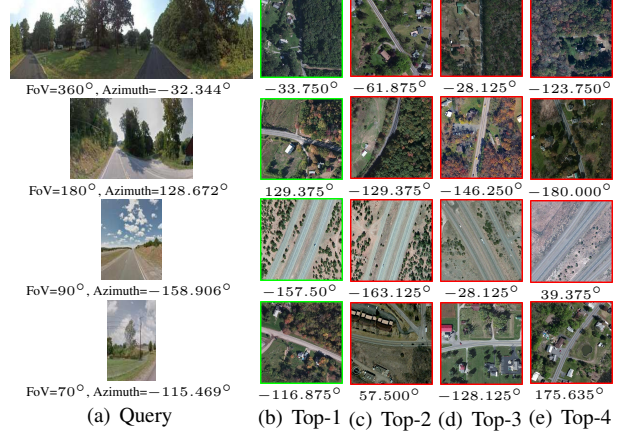(a) Query (b) Top-1 (c) Top-2 (d) Top-3 (e) Top-4

Figure 7. Visualization of joint location and orientation estimation results by our method on the CVUSA dataset. The FoV and ground truth azimuth angle are presented under each query image. The angle under each aerial image is the estimated relative orientation of the query image with respect to this aerial image. Green and red borders indicate correct and wrong retrieved results, respectively.

Table 4. Orientation prediction performance on correctly localized ground images.

Dataset	CVUSA				CVACT_val			
FoV	360°	180°	90°	70°	360°	180°	90°	70°
orien_acc	99.41	98.54	76.15	61.67	99.84	99.10	74.51	55.18
median_error	2.38	2.38	4.50	4.88	1.97	2.89	5.21	6.22

orientation of ground images with 360° and 70° FoVs, and Figure 7 presents some qualitative examples on joint location and orientation estimation. More visualization results on orientation estimation are provided in the supplementary material.

5. Conclusion

In this paper we have proposed an effective algorithm for image-based geo-localization, which can handle complex situation when neither location nor orientation is known. Contrast to many existing methods, our algorithm recovers both location and orientation by joint cross-view image matching. Key components of our framework include a polar-transformation to bring different domains closer and a novel Dynamic Similarity Matching module (DSM) to regress on relative orientation. Benefited from the two items, our network is able to extract appropriate aerial features if the ground image is disoriented and has a limited FoV. We obtained higher location recalls for cross-view image matching, significantly improve the state-of-the-art in multitude practical scenarios.

6. Acknowledgments

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 3

[2] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 6

[3] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 9–17, 2015. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[5] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 5, 6, 7, 8

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[7] 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. 3

[8] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015. 4

[9] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 3

[10] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 5, 6, 7

[11] Arsalan Mousavian and Jana Kosecka. Semantic image based geolocation given a map. *arXiv preprint arXiv:1609.00278*, 2016. 3

[12] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 6

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[14] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems*, pages 10090–10100, 2019. 3

[15] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. *arXiv preprint arXiv:1907.05021*, 2019. 2, 3, 5, 6, 7, 8

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4

[17] Bin Sun, Chen Chen, Yingying Zhu, and Jianmin Jiang. Geocapsnet: Aerial to ground view image geo-localization using capsule network. *arXiv preprint arXiv:1904.06281*, 2019. 3, 6

[18] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision*, pages 494–509. Springer, 2016. 2, 3, 5, 6

[19] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 70–78, 2015. 3

[20] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 3, 6

[21] 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*, volume 3, 2017. 3, 5, 6

[22] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 3