Wide-Area Image Geolocalization with Aerial Reference Imagery

Scott Workman[1] scott@cs.uky.edu
Richard Souvenir[2] souvenir@uncc.edu
Nathan Jacobs[1] jacobs@cs.uky.edu
[1]University of Kentucky [2]University of North Carolina at Charlotte

Abstract

We propose to use deep convolutional neural networks to address the problem of cross-view image geolocalization, in which the geolocation of a ground-level query image is estimated by matching to georeferenced aerial images. We use state-of-the-art feature representations for ground-level images and introduce a cross-view training approach for learning a joint semantic feature representation for aerial images. We also propose a network architecture that fuses features extracted from aerial images at multiple spatial scales. To support training these networks, we introduce a massive database that contains pairs of aerial and ground-level images from across the United States. Our methods significantly out-perform the state of the art on two benchmark datasets. We also show, qualitatively, that the proposed feature representations are discriminative at both local and continental spatial scales.

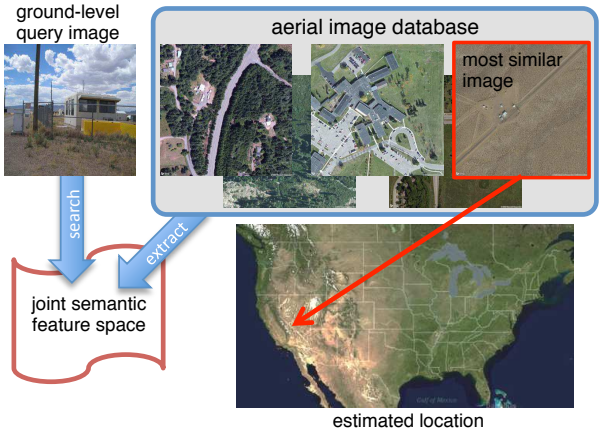


Figure 1: We learn a joint semantic feature representation for aerial and ground-level imagery and apply this representation to the problem of cross-view image geolocalization.

1. Introduction

We address the problem of cross-view image geolocalization, which aims to localize ground-level query images by matching against a database of aerial images (Figure 1). This contrasts with the majority of existing image localization methods which infer location using visual similarity between the query image and a database of other ground-level images. The inherent limitation with these approaches is that they fail in locations where ground-level images are not accessible. Even with hundreds of millions of geo-tagged ground-level images available via photo-sharing websites and social networks, there are still very large geographic regions with few images; most images are captured in cities and around famous landmarks [6].

Cross-view image geolocalization is motivated by the observation that the distribution of geo-tagged ground-level imagery is relatively sparse in comparison to the abundance of high-resolution aerial imagery. The underlying idea is to learn a mapping between ground-level and aerial image viewpoints, such that a ground-level query image can be directly matched against an aerial image reference database. In contrast to previous work [25] which used hand-engineered features, we propose to learn feature representations using deep convolutional neural networks (CNNs). Our methods build upon recent success in using CNNs for ground-level image understanding [20, 44].

We refer to our approach as cross-view training. The idea is take advantage of existing CNNs for interpreting ground-level imagery and use a large database of ground-level and aerial image pairs of the same location to learn to extract semantic, geo-informative features from aerial images. This is a general strategy with many potential applications but we demonstrate it in the context of cross-view geolocalization.

Our work makes the following main contributions: (1) an extensive evaluation of off-the-shelf CNN network architectures and target label spaces for the problem of cross-view localization; (2) cross-view training for learning a joint semantic feature space from different image sources; (3) a massive new dataset with multi-scale aerial imagery; (4) state-of-the-art performance on two smaller-scale evaluation benchmarks for cross-view geolocalization; and (5) extensive qualitative evaluation, including visualizations,

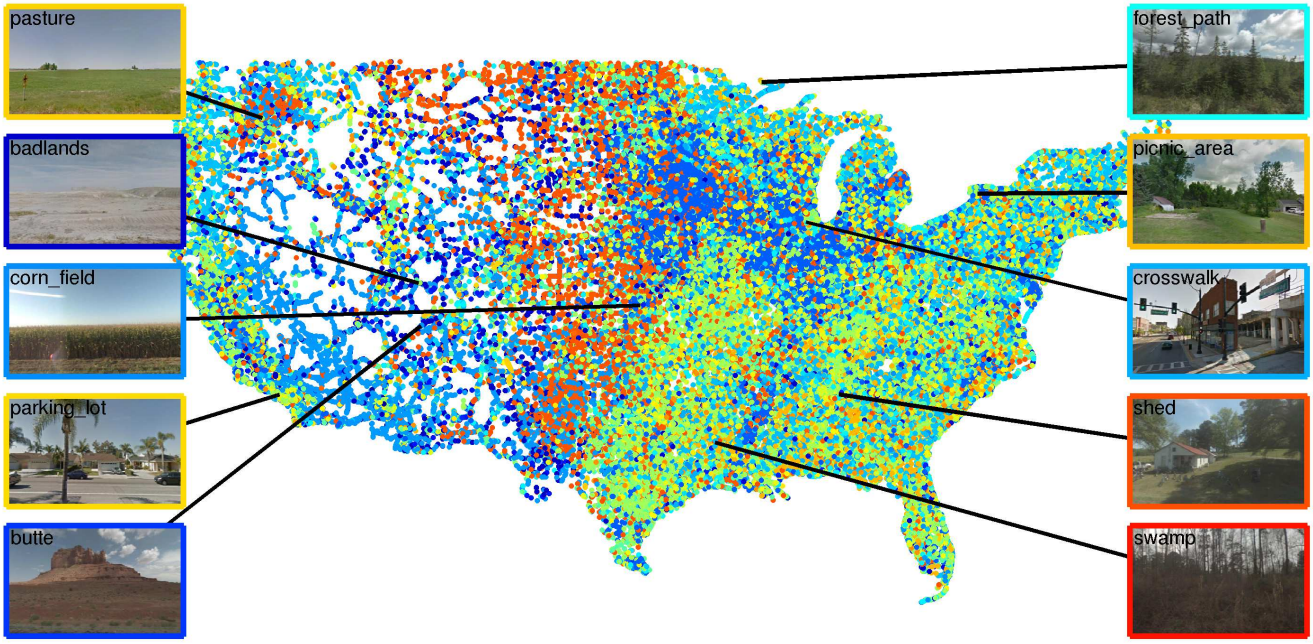


Figure 2: Existing CNNs trained on ground-level imagery provide high-level semantic representations which can be location dependent. Each point represents a geo-tagged image extracted from a Google Street View panorama, colored according to the predicted scene category from the Places [44] network.

which highlights the utility of cross-view training.

2. Related Work

Estimating the geographic location at which an image was captured based on its appearance is a problem of great interest to the vision community. In recent years, a plethora of methods for automatic image geolocalization have been introduced [1, 8, 11, 19, 23, 45]. A wide variety of visual cues have been investigated, including photometric and geometric properties such as sun position [5, 21, 41], shadows [17, 32, 42], and weather [13, 14, 36].

Despite this breadth, the dominant paradigm is to formulate the localization problem as image retrieval. The premise is to take advantage of the ever-increasing number of publicly available geo-tagged images by building a large reference dataset of ground-level images with known location. Then, given a query image, infer its location by finding visually similar images in the dataset. These methods generally fall into one of two categories. The first category of methods infer location by matching using local image features [1, 4, 6, 33, 35, 38, 43]. The second category of methods match using global image features [11, 15, 45]. Matching with local image descriptors is advantageous in that a more precise location estimate is possible, but often requires additional computational resources and fails when no visual overlap exists with the reference dataset. Conversely, whole image descriptors provide a weaker prior over location but require less computation and provide a foundation for many other image understanding tasks.

Estimating geographic information from a single image match requires learning geographically discriminative, location-dependent features [8, 9, 12, 29]. The recent surge of deep learning in computer vision has shown that convolutional neural networks can learn feature hierarchies that perform well for a wide variety of tasks, including object recognition [20], object detection [10], and scene classification [44]. Razavian et al. [30] further show that these feature hierarchies are useful as generic descriptors. Lee et al. [22] estimate geo-informative attributes from an image using convolutional neural network classifiers.

Only recently has aerial imagery been discovered as a valuable resource for ground-level image understanding [2, 27]. Shan et al. [34] geo-register ground-level multi-view stereo models using ground-to-aerial image matching. Viswanathan et al. [39] evaluate a number of hand-engineered feature descriptors for the task of ground-to-aerial image matching in robot self-localization. The cross-view image geolocalization problem was introduced by Lin et al. [25]. Workman et al. [40] show that features extracted from convolutional neural networks are useful for problems in geospatial image analysis. Most akin to our work, Lin et al. [26] apply a siamese CNN architecture for learning a joint feature representation between ground-level images and 45° oblique aerial imagery. Our approach is more gen-

eral; we operate on orthorectified aerial imagery, do not require scale and depth metadata for each query, and our joint feature representation is semantic.

3. Cross-View Training for Aerial Image Feature Extraction

We propose a cross-view training strategy that uses deep convolutional neural networks to extract features from aerial imagery. The key idea is to use pre-existing CNNs for extracting ground-level image features and then learn to predict these features from aerial images of the same location. This is a general approach that could be useful in a wide variety of domains. It is conceptually similar to domain adaptation [7], where the source domain is the ground-level view and the target domain is aerial imagery. The end result of cross-view training is a CNN that is able to extract semantically meaningful features from aerial images without manually specifying semantic labels.

3.1. Cross-View Feature Representations

We assume the existence of two functions: $f_a(l;\Theta_a)$, which extracts features from the aerial imagery centered at location, l, and $f_g(I;\Theta_g)$, which extracts features from a ground-level image. Here, Θ_g and Θ_a are the parameters for feature extraction. We propose to use deep feed-forward convolutional neural networks as the feature extraction functions, f_a and f_g. In this framework, the parameters of these functions, Θ_a and Θ_g, include both the network architecture and the weights.

Our main insight is that we can take advantage of the significant progress that has been made applying CNNs to ground-level image understanding in the past several years by *transferring* feature representations to aerial images. This is possible if the location of the ground-level imagery is known. For example, in Figure 2, we show the estimated label from the Places [44] network, trained for the task of scene classification, on a set of images extracted from Google Street View panoramas captured across the United States. The predicted label is clearly location dependent. For the purposes of learning a useful aerial image feature function, what matters is that the ground-level features are geo-informative, not necessarily that the ground-level detector is perfect.

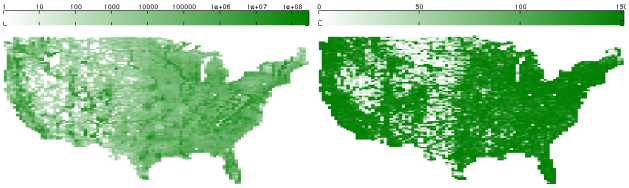
We compare alternative choices for ground-level feature extraction in Section 4 for the problem of cross-view image geolocalization. In the remainder of this section, we describe our cross-view training approach to adapt a network trained for ground-level feature extraction to aerial imagery.

3.2. Cross-View Training a Single-Scale Model

Given a semantically meaningful feature representation for ground imagery, we propose to extract features from aerial imagery, which we refer to as cross-view training. Given a set of ground-level training images, $\{I_i\}$, with known location, $\{l_i\}$ and known ground-level feature extractor parameters, Θ_g, we seek a set of parameters, Θ_a, that minimize the following objective function:

$$J(\Theta_a) = \sum_i \|f_a(l_i;\Theta_a) - f_g(I_i;\Theta_g)\|_2. \quad (1)$$

Intuitively, the objective is to learn to extract features from the aerial imagery that match those from a corresponding ground-level image.



(a) Google Street View (b) Flickr

Figure 3: The distribution of ground-level images in the CVUSA dataset.

3.3. Cross-View Training a Multi-Scale Model

The view frustum of ground-level imagery can vary dramatically from image to image. It is possible that the nearest object in the scene is hundreds of meters away or that the furthest object is tens of meters. This introduces ambiguity when matching the location observed by a ground-level image to the known geolocation of the aerial imagery. To address this issue, we extend our aerial image feature function, f_a, to support extracting features at multiple spatial scales. Rather than mapping a single ground-level image to a single aerial image, the multi-scale approach allows for a ground-level image to be matched to aerial images at multiple scales. In support of multi-scale, cross-view training, we introduce a large dataset of ground-level and aerial image pairs.

3.4. A Large Cross-View Training Dataset

Previous cross-view datasets have been limited in spatial scale and number of training images. The largest dataset [40] contains 174 217 training image pairs sampled from a $200km \times 200km$ area around San Francisco. Features learned using such a dataset are unlikely to be as effective when applied to another location. In an effort to broaden the applicability of the learned feature extractor, we constructed a massive dataset of pairs of ground-level and aerial images from across the United States, called the Cross-View USA (CVUSA) dataset.

Geo-tagged, ground-level images were collected from both Google Street View and Flickr. For Google Street

Figure 4: Example matched ground-level and aerial images from the CVUSA dataset.

View, we randomly sampled from locations within the continental United States. At each location, we obtained the corresponding panoramic image and extracted two perspective images from viewpoints separated by 180° along the roadway. For Flickr, we divided the area of the United States into a 100×100 grid and downloaded up to 150 images from each grid cell (from 2012 onwards, sorted by the Flickr "interesting" score). As Flickr images are overrepresented in urban areas, this binning step ensures a more even sampling distribution. From this set, we automatically filtered out images of indoor scenes using the *Places* [44] scene classification network by retaining images that match to one of the outdoor scene categories.

This process resulted in 1 036 804 Street View images and 551 851 Flickr images. Figure 3 visualizes the relative density of each set of images. For each ground-level image, we downloaded an 800×800 aerial image centered at that location from Bing Maps, at multiple spatial scales (zoom levels 14, 16 and 18). After accounting for overlap, this results in 879 318 unique aerial image locations and a total of 1 588 655 million geo-tagged, image matched pairs. Figure 4 shows several example matched ground-level and aerial images from our dataset.

4. Application to Cross-View Localization

We focus on the problem of cross-view image geolocalization [25] in which the goal is to use a database of aerial images, with known location, to estimate the geographic location of a ground-level query image in that region. This is a challenging problem because of the dramatic appearance differences between ground-level and aerial viewpoints.

4.1. Evaluation Datasets

We evaluate our proposed cross-view training approach on two existing benchmark datasets. The first dataset, Charleston, was introduced by Lin et al. [25] and contains imagery from a $40km \times 40km$ region around Charleston, South Carolina. In total, there are 6 756 ground-level images collected from Panoramio, each with an associated aerial image and land-cover attribute map centered at its location. The aerial image reference database contains 182 988 images. The second benchmark dataset, San Francisco, is introduced by Workman et al. [40] and contains imagery from a $200km \times 200km$ region around San Francisco, California. Ground-level imagery consists of 74 217 images from Flickr and 100 000 Street View cutouts. Similar to Charleston, each ground-level image is accompanied by a corresponding aerial image centered at the ground-level image location. The aerial image reference database contains 278 561 images. Each dataset identifies a set of "hard to localize" ground-level images, with no nearby ground-level reference imagery, to be used for evaluation.

4.2. Localization Method and Performance Metric

The process for localizing a ground-level query image, $\hat{I}$, is straightforward. We directly compare the ground-level feature, $f_g(\hat{I}; \Theta_g)$, for the query image against a reference aerial image feature, $f_a(l; \Theta_a)$, at location l, using Euclidean distance $\|f_a(l; \Theta_a) - f_g(\hat{I}; \Theta_g)\|_2$. If a single pinpoint match is needed, we return the geolocation of the image that is the nearest neighbor of the ground-level image in feature space; otherwise we return a list of candidate regions sorted by distance in feature space. As described by Lin et al. [25], the performance metric for this problem is the rank of the ground truth location in the sorted list of localization scores, for a set of aerial image reference locations. We represent the localization results using a cumulative graph of the percentage of correctly localized images as a function of the percentage of candidates searched.

4.3. Localization using Off-The-Shelf CNN Features

As a baseline to our cross-view training approach, we evaluated the localization performance of "off-the-shelf" CNN features on Charleston. We extracted features from both the aerial and ground-level query image using a variety of network architectures trained for different target label spaces. The network architectures used included GoogleNet [37], AlexNet [20], NIN [24], and VGG 19 [3]. Training databases included Places [44], ImageNet [31], Hybrid [44], Oxford Flowers [28], and Flickr Style [18]. We evaluated multiple such configurations, all publicly available as Caffe [16] model files.

Our findings from this experiment are visualized in Figure 5. The top two performing configurations in terms of top 5% accuracy are trained for the task of scene classification on the Places [44] database, which contains over two million images labeled from 205 different categories. These two networks vastly outperform the next best net-

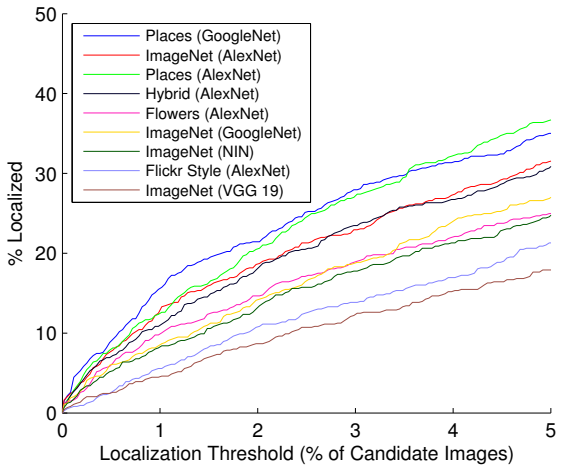


Figure 5: Comparison of several off-the-shelf CNN features in terms of localization accuracy on the Charleston dataset.

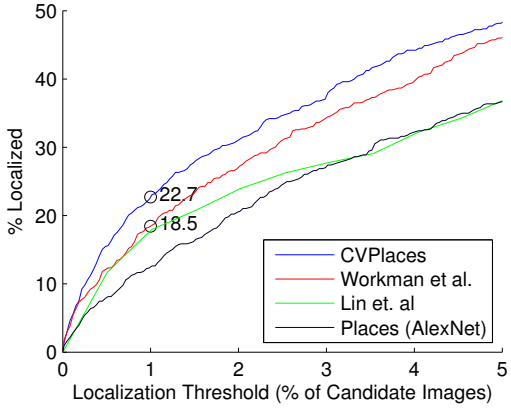
work, which was trained on ImageNet for the task of object recognition. These results are interesting, but unsurprising, as scenes are more likely to be visible from aerial imagery. For the rest of the experiments, we apply cross-view training to learn an aerial image feature extractor for Places features using the *AlexNet* architecture [44], which we refer to as *Places*.

4.4. Localization using Cross-View Features

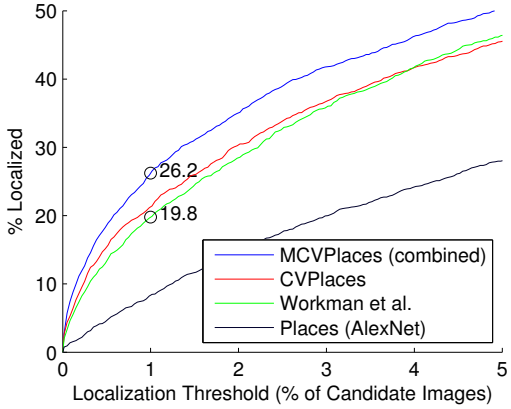
The *AlexNet* architecture [20] consists of five convolutional layers (interspersed with dropout, pooling, and local response normalization layers) and three fully-connected layers (called 'fc6', 'fc7', and, the output layer, 'fc8'). The only difference with *Places* is the dimensionality of the output layer (205 versus 1000 possible categorical labels).

Given the architecture and weights, Θ_g, of *Places*, we apply the cross-view training approach described in Section 3 to train a model to predict the 'fc8' features. In practice, we fix the network architecture and optimize the weights. For training, we use pairs of ground-level images and the highest-resolution aerial images in our CVUSA dataset (zoom level 18). We refer to this model as *CVPlaces*. Figure 6 shows the improvement in localization of our single-scale model, with and without cross-view training, on Charleston and San Francisco.

Initial experiments showed that initializing the solver with $\Theta_a^0 = \Theta_g$ worked well, therefore we use that strategy throughout. We reserve 1000 matched pairs of images from each benchmarks training set as a validation set for model selection. Our models are implemented using the Caffe toolbox [16] and trained using stochastic gradient descent with a Euclidean loss for parameter fitting to reflect (1). The full model file, solver definition, and learned network



(a) Charleston



(b) San Francisco

Figure 6: Accuracy of localization as a function of retrieved candidate locations on two benchmark datasets.

weights are available online.[1]

4.5. Evaluating Multi-Scale Cross-View Training

Our multi-scale model architecture consists of three single-scale *CVPlaces* networks with untied weights, each taking as input a different spatial resolution of aerial imagery. The top feature layer from each individual network is concatenated and used as input to a final fully-connected layer with a 205 dimensional output. The resulting model has approximately 180 million parameters. For training, we initialize each of the sub-networks with the weights for our best single-scale network and randomly initialize the output layer. We refer to our multi-scale model as *MCVPlaces*.

To evaluate *MCVPlaces*, we augmented San Francisco with additional multi-scale aerial imagery (zoom levels 16 and 14). Figure 6 shows a comparison of our multi-scale approach versus our single-scale approach and a recent

[1] http://cs.uky.edu/~scott/

method on San Francisco. The features learned via multi-scale cross-view training significantly out-perform all others. In terms of top 1% accuracy, we improve the state-of-the art by 6.4%, a percentage change of 32.32%.

5. Discussion

The evaluation suggests that the cross-view training procedure learns features that are effective for localization. In the remainder of this section, we explore this representation in more depth.

5.1. Understanding Network Activations

To understand what the network is learning, we analyze the node-level activations for a large set of images on the *Places* network and our *CVPlaces* network. We randomly sampled 20 000 pairs of ground-level/aerial images from CVUSA and recorded the activations for each. Figure 7 shows a set of images that resulted in the maximum activation for particular 'fc8' nodes of each network. We selected the 'fc8' nodes because they are the last layer before the *softmax* output and are therefore semantically meaningful. The ground-level images that result in high activations on the *Places* network are good exemplars of their corresponding category. However, using the same network, high-activation aerial images are often semantically incorrect. For example the "wheat field" image is actually a forest and the "airport" image is a highway. When passed through our *CVPlaces* network, the high-activation images are much more semantically plausible. These results highlight that the cross-view training process is learning to recognize locations in aerial images where particular scene categories are likely to be observed from a ground-level viewpoint.

5.2. Geospatial Visualization of Aerial Image Features

We visualize the geospatial distribution of high-level features extracted from the high-resolution aerial reference imagery from the Charleston dataset [25]. The result is a coarse-resolution false-color image that summarizes the semantic information extracted by a particular CNN from the aerial images. To support this, we computed the 'fc8' features from two networks, *Places* and our *CVPlaces*. For visualization purposes, we choose three high-level categories (urban, rural, and water-related) and assign a set of representative scene categories to each. The false-color image is generated as follows: for the red channel, we compute the average activation for the set of categories defined as urban on the aerial imagery under each pixel. The same procedure is applied for rural (green) and water-related (blue). We then linearly scale the averaged activations to the range $[0, 1]$. The result is a false-color aerial image (Figure 8) with semantically meaningful colors. For example, a bright red pixel identifies an urban area and a purple pixel is an urban area near the water, etc. Our *CVPlaces* network results in a clearer distinction between regions, highlighting the urban core of Charleston and distinguishing water regions from rural. This demonstrates that the cross-view training procedure enables the *CVPlaces* network to extract semantically meaningful features from aerial imagery. This is especially interesting because the network was trained using the entire CVUSA dataset and was not fine-tuned specifically for the Charleston area.

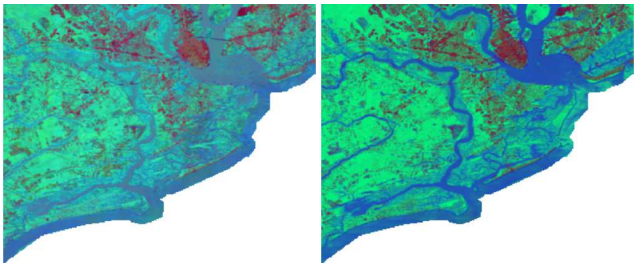


Figure 8: (left) A false-color image generated by applying the *Places* network to aerial imagery. In both images the colors are semantically meaningful (red=urban, green=rural, blue=water-related). (right) The same as (left) but with our *CVPlaces* network (trained on the entire USA dataset, with no Charleston-specific fine tuning).

5.3. Localization at Dramatically Different Spatial Scales

The quantitative evaluation shows that by using our *CVPlaces* network, we obtain state-of-the-art localization performance at the scale of a major metropolitan area (approx. $100km$ across). In this section, we explore whether *CVPlaces* might work at larger and smaller spatial scales. We begin at the continental scale: given a ground-level query image from CVUSA, we compute the feature distance between the *Places* 'fc8' feature vector of the query image and *CVPlaces* 'fc8' feature vector of all aerial images in the dataset. Figure 9 shows qualitative results as a heatmap that represents the distance between the query and corresponding aerial image. The black dot represents the ground truth location of the query images. In the first example, our method clearly identifies the image as having been captured in the desert southwest. The second example, of a suburban neighborhood, results in a heatmap that highlights urban areas. The third example identifies the query image as having been captured on a coast.

We also explore whether the proposed method can be used for localization at a much smaller scale. Figure 10 shows examples where the method is able to distinguish between locations a few decameters apart. To accomplish this, we implemented a system that takes as input a query image and an initial location estimate. It samples a grid of nearby geographic locations and computes the distance be-

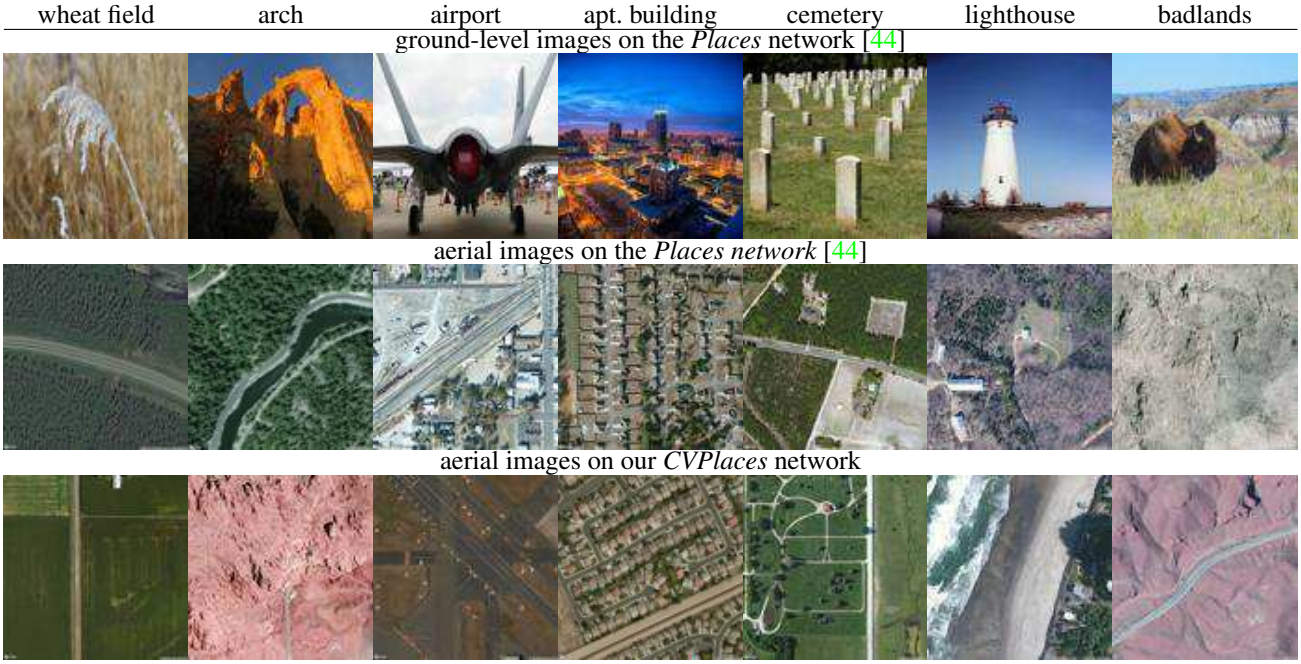


Figure 7: Images that result in high activations for particular scene categories. (top) The high-activation ground-level images are exemplars for the corresponding semantic class. (middle) The high-activation aerial images for the network trained on ground-level images are, not surprisingly, less semantically correct. For example, in the "arch" category the image may look like an arch, but is not a location you are likely to see an arch from the ground. (bottom) After fine-tuning for the aerial domain, the high-activation images are a better match to the respective categories.

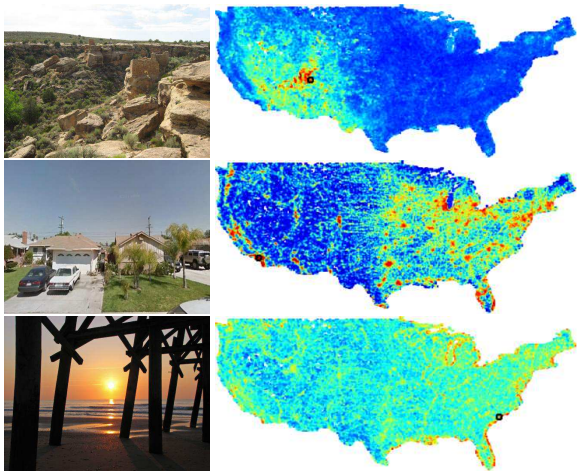


Figure 9: Localization examples at a continental scale. (left) A ground-level query image. (right) A heatmap of the distance between the *Places* 'fc8' feature of the query image and the corresponding *CVPlaces* feature of an aerial image at that location (red: more likely location, blue: less likely location). The black circle marks the true location of the camera.

tween the *Places* 'fc8' feature vector of the query image and the corresponding *CVPlaces* feature of the sub-window of the aerial imagery. Note that sampling on the grid could be accelerated by computing it convolutionally on the GPU. These results show that in some cases, such as the American football example, it can identify a football stadium given an image of players. In the other examples, the heatmaps reflect the inherent uncertainty of localization. The lake-shore example is particularly interesting because even though the shore is not visible, the heatmap correctly reflects that the photographer is less likely to be standing in the middle of the lake than on its shore.

6. Conclusion

We proposed a cross-view training approach, in which we learn to predict features extracted from ground-level imagery from aerial imagery of the same location. We introduced a massive dataset of such pairs and proposed single and multi-scale networks for extracting aerial image features, obtaining state-of-the-art results for cross-view localization on two benchmark datasets.

Our focus was learning the optimal parameters, Θ_a, for extracting features from aerial imagery. We tried fixing the aerial parameters, Θ_a, using pre-existing networks, and optimizing over Θ_g, but the performance was poor. We also

Figure 10: Examples of localization at finer spatial scales. (top) The ground-level query image. (middle) An aerial image centered at the ground location. (bottom) An overlay showing the distance between the ground-level image feature and the aerial image features at each location, computed using a sliding window approach (red: more likely, blue: less likely).

attempted jointly optimizing over Θ_a and Θ_g but the results did not improve over exclusively optimizing for Θ_a. We suspect both of these results are because existing ground-level image feature extractors are better suited for cross-view localization than aerial image feature extractors. However, finding better initial values for Θ_a is an interesting area for future work.

When the ground-level query image was captured in a location that is distinctive from above, such as an outdoor football stadium or an intersection with a unique pattern of intersecting roads, it is possible to obtain a precise estimate of the geographic location using the cross-view localization approach. However, many locations are not so distinctive. Therefore, it is useful to consider the proposed approach as a pre-processing step to a more expensive matching process. Such a matching process might be purely computational, as with sparse keypoint matching, or may involve manual human search.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *European Conference on Computer Vision*, 2012.

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

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[4] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[5] F. Cozman and E. Krotkov. Robot localization using a computer vision sextant. In *International Conference on Robotics and Automation*, 1995.

[6] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *International World Wide Web Conference*, 2009.

[7] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006.

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 31(4), 2012.

[9] Q. Fang, J. Sang, and C. Xu. Discovering geo-informative attributes for location recognition and exploration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1s):19, 2014.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic

segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

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

[12] M. T. Islam, S. Workman, H. Wu, N. Jacobs, and R. Souvenir. Exploring the geo-dependence of human face appearance. In *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[13] N. Jacobs, K. Miskell, and R. Pless. Webcam geolocalization using aggregate light levels. In *IEEE Workshop on Applications of Computer Vision*, 2011.

[14] N. Jacobs, N. Roman, and R. Pless. Toward fully automatic geo-location and geo-orientation of static outdoor cameras. In *IEEE Workshop on Applications of Computer Vision*, 2008.

[15] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *IEEE International Conference on Computer Vision*, 2007.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[17] I. N. Junejo and H. Foroosh. Gps coordinates estimation and camera calibration from solar shadows. *Computer Vision and Image Understanding*, 2010.

[18] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *British Machine Vision Conference*, 2014.

[19] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *European Conference on Computer Vision*, 2010.

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

[21] J.-F. Lalonde, S. G. Narasimhan, and A. A. Efros. What do the sun and the sky tell us about the camera? *International Journal of Computer Vision*, 2010.

[22] S. Lee, H. Zhang, and D. J. Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[23] Y. Li, D. Crandall, and D. Huttenlocher. Landmark classification in large-scale image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[24] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations*, 2014.

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

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

[27] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In *ACM International Conference on Multimedia*, 2008.

[28] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[29] D. Quercia, N. K. O'Hare, and H. Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2014.

[30] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE CVPR Workshop: DeepVision: Deep learning in Computer Vision*, 2014.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[32] F. E. Sandnes. Determining the geographical location of image scenes based on object shadow lengths. *Journal of Signal Processing Systems*, 2011.

[33] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

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

[35] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Transactions on Graphics*, 25(3):835–846, 2006.

[36] K. Sunkavalli, F. Romeiro, W. Matusik, T. Zickler, and H. Pfister. What do color changes reveal about an outdoor scene? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[38] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

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

[40] S. Workman and N. Jacobs. On the location dependence of convolutional neural network features. In *IEEE/ISPRS Workshop: Looking From Above: When Earth Observation Meets Vision*, 2015.

[41] S. Workman, R. P. Mihail, and N. Jacobs. A pot of gold: Rainbows as a calibration cue. In *European Conference on Computer Vision*, 2014.

[42] L. Wu and X. Cao. Geo-location estimation from two shadow trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[43] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, 2010.

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

[45] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *European Conference on Computer Vision*, 2014.