2018

Leveraging Overhead Imagery for Localization, Mapping, and Understanding

Scott Workman
University of Kentucky, scottworkman@uky.edu
Digital Object Identifier: https://doi.org/10.13023/ETD.2018.128

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Scott Workman, Student

Dr. Nathan Jacobs, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

Leveraging Overhead Imagery for Localization, Mapping, and Understanding

---

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy in the
College of Engineering at the
University of Kentucky

By
Scott Workman
Lexington, Kentucky

Director: Dr. Nathan Jacobs
Associate Professor of Computer Science
Lexington, Kentucky 2018

ABSTRACT OF DISSERTATION

Leveraging Overhead Imagery for Localization, Mapping, and Understanding

Ground-level and overhead images provide complementary viewpoints of the world. This thesis proposes methods which leverage dense overhead imagery, in addition to sparsely distributed ground-level imagery, to advance traditional computer vision problems, such as ground-level image localization and fine-grained urban mapping. Our work focuses on three primary research areas: learning a joint feature representation between ground-level and overhead imagery to enable direct comparison for the task of image geolocalization, incorporating unlabeled overhead images by inferring labels from nearby ground-level images to improve image-driven mapping, and fusing ground-level imagery with overhead imagery to enhance understanding. The ultimate contribution of this thesis is a general framework for estimating geospatial functions, such as land cover or land use, which integrates visual evidence from both ground-level and overhead image viewpoints.

KEYWORDS: computer vision, machine learning, remote sensing, geospatial analysis

Author's signature: Scott Workman

Date: May 10, 2018

Leveraging Overhead Imagery for Localization, Mapping, and Understanding

By
Scott Workman

Director of Dissertation: Nathan Jacobs

Director of Graduate Studies: Miroslaw Truszczynski

Date: May 10, 2018

ACKNOWLEDGMENTS

I am immensely grateful to my advisor, Nathan Jacobs, who not only introduced me to research, but supported me when I was considering pursuing a doctoral degree. Over the years I have benefited from his guidance, support, patience, and knowledge; there is no question he has had a profound impact on my life, and my career. I honestly feel lucky to have had the opportunity to work alongside him and I am proud of what we have accomplished together.

I would like to take this opportunity to thank several individuals who helped set me on this path: Judy Goldsmith, who was the first person to encourage me to pursue a graduate education, Jerzy Jaromczyk, for helping set the ball rolling, and Jim Griffioen, for his insight and advice. In addition, I'd like to thank Grzegorz Wasilkowski for his kindness and mentorship, and Victor Marek for the wonderful discussions. A big thanks to the members of my advisory committee, Ruigang Yang, Judy Goldsmith, and Liang Liang, for their invaluable feedback throughout.

I have had the privilege of working and collaborating with many individuals, including: Mohammad Islam, Paul Mihail, Ryan Baltenberger, Connor Greenwell, Tawfiq Salem, Zach Bessinger, Weilian Song, Hui Wu, David Crandall, David Smith, Jim Knochelmann, Armin Hadzic, and others. A sincere thanks to you all. In particular, I would like to thank Richard Souvenir, for his longstanding help and support, and my friend and colleague, Menghua Zhai, with whom I worked closest and achieved a great deal.

I have been fortunate to make many lifelong friends during my time at the University of Kentucky. In lieu of trying to list everyone, please accept this thanks from the bottom of my heart. You have all helped to make this one of the best and most treasured periods of my life.

* * *

Last and most importantly I would like to recognize my parents, Joan and Bob. I feel it is nearly impossible to convey in words just how awesome and special these two individuals are. Everything I am is because of them. Thank you.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

> *"Of all of our inventions for mass communication, pictures still speak the most universally understood language."*
>
> – Walt Disney

Images of the natural world reveal a wide variety of subtle cues that allow humans to rapidly understand the semantic and geometric context of a scene. Studies regarding scene perception [82] have shown that humans can perceive at a glance an immense amount of visual information about a scene, from low-level details such as color and contours, to mid-level details about shape and texture, all the way to high-level details about semantics. For instance, it only takes a passing glance for a human to notice that the image in Figure 1.1 was captured during clear conditions, that the camera was positioned above the water, perhaps on a bridge, or that the geographic location is near a harbor due to the ships along the canal. These types of observations are critical to interpreting what is happening in an image.

The ultimate goal of computer vision is to teach a computer how to produce such observations; in other words, to develop methods and learn representations that allow for automatically understanding the contents of images and video. At first glance, this might not seem like a very complex problem, or at least it didn't to Marvin Minsky, who in 1966 asked a first-year undergraduate student "to spend the summer linking a camera to a computer and getting the computer to describe what it saw" [10]. While the student was unsuccessful, this overly ambitious summer project spawned decades of research that is making dramatic impacts in a wide variety of fields.

There are countless examples today highlighting the importance of research in this area. Of the most compelling are the autonomous robots and vehicles such as NASA's Mars Exploration Rover, which make use of a bevy of cameras critical for tasks such as terrain

Figure 1.1: (left) A personal photo, taken from the Nyhavn Bridge in Copenhagen, Denmark, overlooking the canal. (right) The corresponding overhead view, centered at the ground-level image capture location.

analysis and navigation [72]. Other successful applications include face detection software on mobile phones, automatic detection of tumors in medical imaging, and assistive technologies for individuals with vision impairments, to name only a few.

The aforementioned applications, and many more, are motivated by the capability images offer as a vast untapped resource of information about the world and the way it changes over time. The concept of teaching computers to see has endless real world functionality. In the realm of environmental monitoring, land-based cameras have been deployed to estimate atmospheric visibility [135], analyze beach usage [32], study nearshore oceanography [35], track leaf growth [77, 93], and estimate snow cover [96]. These are real world applications which use images to produce scientific measurements.

In this thesis, we explore how overhead imagery can be leveraged, in addition to sparsely distributed ground-level images, to improve solutions to problems in localization, mapping, and understanding. Though ground-level and overhead images provide complementary viewpoints of the world, ground-level imagery is not available at every location. Our work is motivated by the observation that you can often understand what would be present at a location from a ground-level viewpoint by looking at the corresponding overhead image.

1.1 Images for Geospatial Analysis

Traditionally research in computer vision has focused primarily on developing methods for ground-level image understanding, with great success. For ground-level imagery, there

now exist high-quality methods for face recognition and verification [113], object [92] and pedestrian detection [8], scene understanding [84], camera calibration [145], 3D reconstruction [107], image colorization [143] and synthesis [28], and much more. In general, these methods focus on extracting information from an individual ground-level image or a collection of ground-level images.

Recently, a large body of work has explored the use of geotagged social media, in particular ground-level imagery, to estimate geographic properties of the world, for example measuring snow fall and vegetation density [142]. These methods use volunteered geographic information obtained from sources such as blogs, social networks, and community contributed photo collections as a source of geospatial information to estimate some unobservable geospatial function. Each social media artifact (e.g., ground-level image) is an observation of this function at a particular geographic location.

This is a research direction often referred to as proximate sensing or image-driven mapping. Here, publicly available data from social media serves as a replacement for the expensive data collection process (e.g., field data collection, distributed sensor networks) typical in other fields that seek to create geospatial models, such as landscape phenology [64]. For example, Leung and Newsam [60] show that large collections of georeferenced images can be used to automatically estimate land cover. Similarly, Crandall et al. [15] analyze 35 million ground-level images and introduce methods for automatically identifying and classifying representative images using visual, textual, and temporal features. These works, and others, are motivated by the recent surge of publicly available geotagged imagery as a new source of data for solving existing problems.

Unfortunately, using ground-level images as the only source of information has its drawbacks. The most prominent challenge is that ground-level images aren't available at every location. As Crandall et al. [15] show, most images are captured in urban areas and around famous landmarks. Despite the existence of huge datasets of geotagged images [115], and billions more geotagged images publicly available online, there are still large geographic regions with little to no coverage. This is demonstrated empirically by Weyand et al. [126] who use 490 million geotagged images to partition the Earth's surface into a set of non-overlapping cells. Due to the non-uniform geographic distribution of photos, large regions such as central Africa and Northern Asia are either completely omitted, or collapsed into a single cell.

While the sparse and non-uniform distribution of ground-level images is the biggest difficulty, individual samples are often noisy due to incorrect metadata (e.g., incorrect geotags) or suffer from other issues such as manually manipulated visual content. Furthermore, despite massive improvements to existing recognition algorithms, they are still imperfect.

Therefore, in order to build geospatial models from both sparse and noisy observations, it is common for some form of simple local averaging to be applied. This process in turn results in coarse, low-resolution outputs. The primary challenge in this area is to produce accurate estimates over a large region, while maintaining a fine-grained high-resolution output.

In remote sensing, it is common to estimate physical properties of the Earth using satellite imagery. Satellite imagery has been used to monitor dust and ash from volcanoes as well as for preparing and responding to other natural hazards such as floods and landslides [118]. There are numerous examples demonstrating how remote sensed imagery is used to monitor the Earth, including: for global land cover classification [117], to aid precision agriculture [78], to analyze urban infrastructure [46], and for large scale monitoring of vegetation dynamics [144].

The potential of overhead imagery has been recognized for over one hundred years. In military scenarios, overhead imagery is thought to have been captured and used for reconnaissance as early as 1859 during the Battle of Solferino, when the French Army took images of the Austrian troops using balloons, as well as during the American Civil War (1861-1865) [86]. Stichelbaut et al. [109] present an extensive and informative discussion of the development of overhead photography and its early dependence on military reconnaissance and the birth of aviation, especially during the First World War. Naturally, the rise of overhead imagery presented unique research challenges.

More recently, efforts have been made to automate overhead image analysis. As early as 1970 [37] methods were introduced for classifying terrain types from a single overhead image, with the goal of automatically generating terrain maps. Similarly, in 1976 Bajcsy et al. [5] described a system for recognizing roads, intersections, and other road-like objects in overhead imagery. However, as overhead imaging differs drastically from ground-level imaging, the majority of techniques that have been developed have occurred independently and in task-specific ways [95].

When available, overhead imagery offers dense coverage compared to other sparsely available measurements, such as ground-level images. However, high-resolution overhead imagery has traditionally only been available through commercial vendors. As such, Wulder and Coops [134] recently argued that satellite imagery should be made freely available for its potential to improve science and environmental monitoring. Ignoring cost of access, only recently has overhead imagery become more widely available at higher spatial and temporal resolutions. This is largely due to a surge in the number of microsatellites [22], miniaturized satellites that fly in low-orbit and are more cost-effective to launch, and the so-called commercialization of space.

Overhead imagery presents a complementary viewpoint to ground-level imagery which

can aid understanding. Now, for most locations, overhead imagery is freely available. If the location of a ground-level image is known, an overhead image can be used to provide additional context. For instance, it becomes immediately obvious that the ground-level image shown in Figure 1.1 (left) was captured from a bridge over the canal when examining the co-located overhead image in Figure 1.1 (right). Despite this, overhead images have largely been ignored in the computer vision community; very little work has explored how overhead imagery can aid existing computer vision algorithms that target ground-level image understanding.

1.2 Our Approach

Our thesis focuses on the joint understanding of ground-level and overhead image viewpoints to improve geospatial modeling. The main research question we will address is, "How can geotagged ground-level imagery and overhead imagery be exploited in unison to address problems in localization, mapping and understanding?" Our proposed methods take advantage of the fact that high-resolution overhead imagery now exists across the globe and is updated regularly. Below, we highlight three main areas of our work.

- **Learning a Joint Feature Representation:** We investigate learning a joint feature representation between ground-level and overhead images, such that images from differing viewpoints can be directly compared. Our insight is that, while the relationship between ground-level and overhead image viewpoints is complex, overhead imagery is densely available and a joint feature representation would enable several potential methods for extending existing approaches in ground-level image understanding (e.g., image geolocalization).

- **Inferring Labels for Overhead Imagery:** We explore how overhead imagery can be used to directly drive predictions when ground-level imagery is not available. Our motivation is that there is a plethora of unlabeled overhead imagery which could be used to augment existing techniques. To do this, our main insight is that we can leverage nearby ground-level images, and existing image recognition algorithms, to infer labels for overhead imagery. As a result, unlabeled overhead imagery can be used to train models by inferring the target label from a nearby ground-level image, enabling direct predictions from overhead imagery.

- **Fusing Ground-level and Overhead Imagery:** Finally, we investigate several methods for fusing ground-level imagery and overhead imagery together in order to

estimate general geospatial functions. An integral part of this process is exploring how overhead imagery can be used to bias the interpolation of sparse ground-level image samples. Our key insight is that overhead imagery is essentially a source of latent information that can, for example, be used for adaptive kernel bandwidth estimation.

To support our research efforts, we construct large datasets containing both geotagged ground-level images and overhead images. In general, each ground-level image is paired with a co-located overhead image, centered at the ground-level image capture location. Typically, the overhead imagery is collected at multiple spatial resolutions. This work describes multiple such datasets, all of which have been made available to the computer vision community.

The ultimate contribution of this thesis is a general framework for estimating geospatial functions which integrates visual evidence from both ground-level and overhead image viewpoints. Our approach combines the strengths of proximate sensing and remote sensing, resulting in a general architecture that can be trained end-to-end such that it learns to extract the optimal features from each viewpoint. Further, the proposed framework is general and can be adapted to use any sparsely distributed measurements.

1.3 Synopsis

The remainder of this work is organized as follows:

- **Chapter 2 - Are Deep Image Representations Geo-Informative?**

  In this chapter, we investigate the usefulness of deep image representations, extracted from convolutional neural networks applied to traditional vision tasks, for problems in geospatial image analysis. In particular, we analyze their discriminative ability with regard to location through several problem settings, including region identification in ground-level imagery, understanding and interpreting overhead images, and cross-view image matching. Our results demonstrate the effectiveness of deep image representations extracted from CNNs, on both ground-level and overhead imagery, for capturing geographically discriminative features relating image appearance to geographic location. This points to a promising direction for future research in building deep-learning based models that are directly targeted at problems of localization and location-related feature extraction from ground-level

and overhead imagery. This work was originally reported in [127].

- **Chapter 3 - Wide-Area Image Geolocalization with Overhead Reference Imagery**

  In this chapter, we propose an approach for learning a joint feature representation between ground-level and overhead imagery and demonstrate its application for the task of image geolocalization. In our cross-view problem formulation we match against georeferenced overhead images, as opposed to standard image geolocalization techniques, which infer the location of a ground-level query image from a reference database of ground-level images with known location. Densely available overhead imagery enables fine-grained geolocalization results at varying spatial scales. Our proposed methods take advantage of deep convolutional neural networks; 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 overhead images. We also propose a variant of our network architecture that fuses features extracted from overhead images at multiple spatial scales. This work was originally reported in [129].

- **Chapter 4 - Understanding and Mapping Natural Beauty**

  In this chapter, we show how overhead imagery, in particular unlabeled overhead imagery, can be exploited to improve image-driven mapping. To begin, we focus on the subjective property of image scenicness and propose an approach to predict scenicness which explicitly accounts for the variance of human ratings. Then, given a method for predicting the scenicness of an individual ground-level image, we explore methods for mapping image scenicness over a large spatial region. To learn to predict image scenicness from unlabeled overhead imagery, we apply a cross-view training approach. Instead of predicting the scenicness of the overhead image, we predict the scenicness of a ground-level image captured at the same location. Our results demonstrate that quantitative measures of scenicness can benefit semantic image understanding, content-aware image processing, and a novel application of cross-view mapping, where the sparsity of ground-level images can be addressed by incorporating unlabeled overhead images in the training and prediction steps. This work was originally reported in [130].

- **Chapter 5 - A Unified Model for Near and Remote Sensing**

  In this chapter, we propose a framework for fusing together ground-level and overhead images for geospatial modeling. Specifically, we describe a novel convolutional neural network architecture for estimating geospatial functions such as population density, land cover, or land use. Our approach uses neural networks to extract features from both overhead and ground-level imagery. For the ground-level images, we use kernel regression and density estimation to convert the sparsely distributed feature samples into a dense feature map spatially consistent with the overhead image. This ground-level feature map is then fused with an overhead image feature map at an intermediate layer. The output of our network is a dense estimate of the geospatial function in the form of a pixel-level labeling of the overhead image. This work was originally reported in [131].

- **Chapter 6 - Discussion**

  In this chapter, we summarize the contributions of this thesis and our most important findings. In addition, we discuss possible future research directions that will lead to improved methods for geospatial analysis.

Chapter 2

Are Deep Image Representations Geo-Informative?

2.1 Introduction

The relationship between image appearance and geographic location is complex, fascinating, and well studied. As such, a significant amount of work has focused on extracting geographically discriminative location-dependent features from images. For example, recent work has attempted to characterize the relationship between facial appearance and geographic location [39], learn attributes for recognizing the identify of a city [149], and relate the visual aesthetics and perception of fashion to geographic location [105].

The common underlying objective of such methods is to learn geographically discriminative attributes [23, 39, 59, 89] from images. The most influential work in this area is that by Doersch et al. [19] who introduced a method for automatically finding geographically distinctive visual elements for a region using a discriminative clustering approach applied to a large repository of geotagged imagery. Patterson and Hays [84], and Laffont et al. [55] learn high level scene attributes for scene recognition [55, 84]. Recently, Zhou et al. [149] introduced a method for characterizing the identity of a city from a data-driven attribute analysis.

Modern advances in deep learning, specifically convolutional neural networks (CNNs), have lead to significant performance improvements for a wide variety of vision tasks, including: object classification and detection [30], face recognition and verification [113], image super resolution [20], and scene recognition [146]. This is in large part due to their ability to learn custom feature hierarchies directly from raw image data. However, despite this demonstrated success, their performance in many vision problem domains has yet to

be established.

Several works have explored how higher-level features extracted from CNNs can be used as generic descriptors. Razavian et al. [91] show that feature hierarchies from CNNs are useful as generic image descriptors for object recognition. Penatti et al. [85] examine the generalization of deep features to remote sensing, with a focus on image classification. Recently, Lee et al. [59] applied CNNs for recognizing general geo-informative attributes such as population density from ground-level imagery. Inspired by Fischer et al. [26], who show that mid-level features compare favorably to a hand-engineered feature for descriptor matching, our work extends this line of research to include problems relating to geospatial image analysis.

This chapter investigates the value of deep image representations captured by CNNs for geospatial image analysis. In particular we ask the question, "Are deep image representations geo-informative?" In other words, are the features learned by CNNs capturing information related to location? Our goal here is not to propose new learning algorithms for CNNs. Instead, we focus on image representations extracted from existing state-of-the-art CNNs that have been trained for various tasks. Through numerous experiments, we find that CNNs are valuable as geospatial feature extractors. The primary contributions of this work can be summarized as follows:

- We introduce a large new dataset that includes hundreds of thousands of pairs of ground-level and overhead images.
- Using our proposed dataset, we demonstrate that deep image representations captured by CNNs have sufficient discriminative power to distinguish between two geographic regions in ground-level imagery, and that we can extract iconic images from the learned models.
- Further, we show that deep image representations are useful for interpreting and understanding overhead images, despite the CNNs being trained solely on images from a ground-level perspective.
- In addition, we demonstrate that deep image representations extracted from ground-level images are closely related to the features extracted from overhead images captured at the same location.
- Finally, we show how to use deep image representations for cross-view image geolocalization, improving the state-of-the-art relative to previous methods that use hand-engineered features.

Together these results demonstrate the effectiveness of deep image representations extracted from CNNs, on both ground-level and overhead imagery, for capturing geographically discriminative features relating image appearance to geographic location.

The remainder of this chapter is organized as follows. First we describe our process for extracting deep image representations (Section 2.2). Then, the following sections detail our experiments in three domains: (1) distinguishing regions in ground-level imagery (Section 2.3), (2) understanding and interpreting overhead images, including visualizing land cover differences, improving geospatial modeling, and image-based search (Section 2.4), and (3) cross-view image matching, in which pairs of overhead and ground-level images are used to localize images in regions without ground-level reference imagery (Section 2.5).

2.2 Deep Features for Geospatial Image Analysis

In this section we provide an overview of convolutional neural networks and the notion of deep image representations. We start with a brief history on applying CNNs for image understanding, then highlight details about our process for extracting deep image representations, and finally introduce the datasets that support our experiments.

2.2.1 Background: CNNs for Image Understanding

Deep learning is an emergent area of machine learning research that uses artificial neural networks to model high-level abstractions of input data for some task (e.g., vision and speech recognition). Artificial neural networks are inspired by the biological neural networks that exist in the brain. Biological neural networks are composed of a series of interconnected neurons, where a neuron is a cell that receives and transmits information through electrical signals. Practically, a single neuron operates by receiving a set of input signals, checking if the sum of the input exceeds a certain threshold and if so, the neuron "activates", transmitting a signal forward to other neurons.

The study of artificial neural networks dates back to the early 1940s when McCulloch and Pitts [74] used logical calculus to create a computation model of nervous system activity. In 1958, Frank Rosenblatt [94] introduced the perceptron, an algorithm to model the hypothetical nervous system. The perceptron was implemented as a machine, the "Mark 1 perceptron", designed for image recognition. Today the perceptron is known as a single layer neural network acting as a linear classifier. Due primarily to research by Minsky and Seymour [76] showing the limits of single layer neural nets and the lack of available

computational power, neural network research stagnated until the introduction of the now famous backpropagation algorithm by Paul Werbos [125] in 1974.

It wasn't until the late 1980s that the backpropagation algorithm was applied to multi-layer, or "deep", neural nets. In 1989 LeCun et al. [58] showed how backpropagation could be used to train a neural network with multiple layers for the task of handwritten digit recognition. In the modern sense, deep learning refers to an artificial neural network that combines several stages of non-linear feature transformations into a deep architecture, where a stage, or layer, consists of many neurons followed by a non-linear activation function. Effectively, this allows a deep network to create a high-level abstraction of the input, or in other words, build a hierarchy of feature representations for the given task. Each layer transforms its input into a higher-level feature.

At the same time, LeCun et al. [58] introduced layers of convolution in these models, a seminal idea that led to the widespread use of neural network variants called convolutional neural networks by the vision community. As opposed to typical "fully connected" networks, where each layer input is connected to each output, the spatial information inherently contained in images is exploited by CNNs through connectivity constraints that enforce locally contiguous receptive fields. A side benefit of this local connectivity and sharing of convolution functions (filters), is a reduction in the total number of model parameters. Typically, CNN architectures use several convolutional layers, each followed by some form of regularization or normalization (e.g., sub-sampling (pooling), dropout [108]). The last layers are often similar to standard fully connected neural networks, and act as non-linear classifiers or regressors.

Until recently, training large CNN models was impractical due to the large number of parameters (in the tens of millions) and small training sets. Improvements in processing power, for example graphical processing units (GPUs), along with the introduction of large training datasets and superior optimization algorithms, led to significant performance improvements for CNNs in several traditional computer vision tasks. The seminal work of Krizhevsky et al. [54], winner of the 2012 Large Scale Visual Recognition Challenge [98], demonstrated the first successful application of a CNN architecture for object recognition. This architecture is often referred to as AlexNet, and we adopt it for this work.

The AlexNet architecture consists of eight layers with trainable parameters. Five convolutional layers are connected in a feed-forward manner, interspersed with pooling layers and regularization layers such as local response normalization and dropout. The convolutional layers are followed by three fully connected layers. Rectified linear units (ReLU) are used as the non-linear activation function. The model is trained by minimizing a multinomial logistic loss function. Intuitively, the convolutional layers extract local features across

the image and the fully connected layers combine them to make a prediction.

2.2.2 Pretrained CNN Models

The recent success of CNNs for image understanding has sparked the development of several modular and expandable deep learning libraries (e.g., Caffe [47], Theano [9]) that have made it easier for researchers to share findings in this domain. In this work we use the Caffe [47] library due to its widespread adoption and a large database of pretrained models. Pretrained models contain a network architecture and corresponding model parameters.

As a beneficial side effect, common naming schemes of the various structural elements of deep neural networks have also been widely adopted in the vision community. Convolutional layers, whose output are feature maps, are referred to as *convX*, where *X* denotes the layer's depth away from the data input layer. Similarly for other layer types such as pooling layers, *poolX*, and fully connected layers, *fcX*. Using Caffe or another similar framework it is straightforward to extract the feature corresponding to a layer. Given a mean subtracted image, we resize it to the input size of the network and make a feed-forward pass through the network.

To investigate CNN features as geospatial information predictors, we make use of two publicly available pretrained models, both of which use the AlexNet architecture. The first is trained on ImageNet [17] for detecting object categories and is available through Caffe [47]. The second is trained on the recently introduced Places Database [146] with the goal of scene recognition. We refer to these as ImageNet and Places throughout. Other than having differing model parameters, the only difference between the ImageNet and Places models is the dimensionality of the final fully connected layer, which is dependent on the number of target labels in the original classification task (ImageNet has 1,000 object classes, Places 205 scene classes). In line with recent work [62, 91], we focus our experiments on the features corresponding to the fully connected layers, *fc6 – fc8*.

2.2.3 Datasets

We perform our experiments using two datasets. The first dataset, Charleston, was introduced by Lin et al. [66] and contains 6,756 ground-level images with corresponding overhead and land cover images, and a reference map database of overhead and land cover images, without corresponding ground-level images, for a 40km $\times$ 40km region around Charleston, SC. Of the ground-level images, 737 are considered isolated because there are no other ground-level images nearby.

(a) (b) (c)

Figure 2.1: Overview of the introduced San Francisco dataset. (a) A coverage map where red indicates the spatial coverage of overhead imagery, overlaid with Street View (green) and Flickr (blue) image locations. (b) Example Street View panoramas (top) and their corresponding cutouts (bottom). (c) Example Flickr images after filtering.

We introduce a new dataset, San Francisco, containing ground-level and overhead images collected in a 200km $\times$ 200km region around San Francisco, CA. We collected overhead imagery for the entire region from Bing Maps, each image of size 256 $\times$ 256 and covering a 480m $\times$ 480m area, as the reference map database. Ground-level images from the region were collected from both Flickr and Google Street View. For Flickr, we queried and downloaded images from 2013 onwards, totaling 114,384 images. We used the pre-trained Places model to filter images that were unlikely to be images of outdoor scenes by manually assigning a label of indoor/outdoor to each of the 205 scene categories. This resulted in a final set of 74,217 images. For Street View, we downloaded 50,000 street-level panoramas from which we extracted two side-facing perspective images of size 800 $\times$ 600, totaling 100,000 images. Finally, for each ground-level image we downloaded its corresponding overhead image, centered at the same location.

Our proposed dataset, while similar in conception to Charleston, has several benefits. These include a significantly larger region of interest for localization, many more images, a different region of the country with different land cover attributes, automatic filtering of non-outdoor images, and a large number of images with accurate GPS tags (by virtue of Google Street View). In total, the dataset contains 278,561 map images and 174,217 ground-level images and their associated overhead images. As with the Charleston dataset, we identify a set of isolated images, totaling 2,245 ground-level images. Figure 2.1 visualizes the coverage of our dataset and shows several example images.

2.3 Distinguishing Regions in Ground-Level Imagery

Our first experiment explores whether or not deep image representations are geo-informative by formulating a supervised learning task to distinguish between images captured in Charleston and San Francisco, two cities in very diverse regions. In other

Table 2.1: Region classification accuracy.

Feature	Accuracy
GIST [83]	81.7 %
ImageNet *fc6*	82.7 %
ImageNet *fc7*	82.2 %
ImageNet *fc8*	80.9 %
Places *fc6*	85.1 %
Places *fc7*	85.1 %
Places *fc8*	84.5 %

words, we examine the task of classifying dataset membership, i.e., was the picture taken in San Francisco or Charleston? This is similar to recent work by Zhou et al. [149] that attempts to recognize the identity of a city given a ground-level image. However, instead of using high-level attributes to relate images to location, we examine deep image representations extracted from pretrained CNN models.

To begin, we process each ground-level image to extract deep image representations corresponding to the *fc6*, *fc7*, and *fc8* layers of the ImageNet and Places models, as described in Section 2.2.2. Using these features as input, we train an SVM classifier with an RBF kernel, on a set of randomly selected ground-level images from each dataset, one for each feature and model. This results in six independent SVM classifiers. For evaluation, we use an equal number of images from the isolated set of images defined in both datasets.

Table 2.1 gives a comparison of the accuracy of the different features, trained using 10,000 images from each dataset. As a baseline, we compare against the GIST descriptor [83], a common hand-engineered feature used in scene recognition tasks. We find that in general features from both the Places and ImageNet models are superior to GIST, and that Places is superior to ImageNet. However, the differences between the various feature levels is minor.

Figure 2.2 shows montages of ground-level images classified by the SVM model trained on Places *fc8* features with very high and very low confidences. Many of the detected images are iconic images of the corresponding region, for example the Golden Gate Bridge in San Francisco. In addition, Figure 2.2 shows montages of images in San Francisco that the classifier determines look most like Charleston, and vice-versa. Finally, Figure 2.3 shows a montage of the most ambiguous images, many of which would be very difficult for a person to label correctly. This experiment demonstrates that CNN features capture subtle characteristics of various areas from ground-level imagery.

Figure 2.4 shows how the accuracy of our region classification approach depends on the number of training examples. We select N training examples, 50% from each region,

(a) Most Charleston-like in either region

(b) Most San Francisco-like in either region

(c) Most Charleston-like in San Francisco

(d) Most San Francisco-like in Charleston

Figure 2.2: Montages of ground-level images with high, or low, SVM scores for a model trained on Places *fc8* features (see Section 2.3 for details). (a, b) Images with the highest and lowest SVM scores. (c, d) Images from the respective regions with the most incorrect SVM scores.

Figure 2.3: The most ambiguous images based on the SVM score for a region classifier trained on Places *fc8* features, as described in Section 2.3.

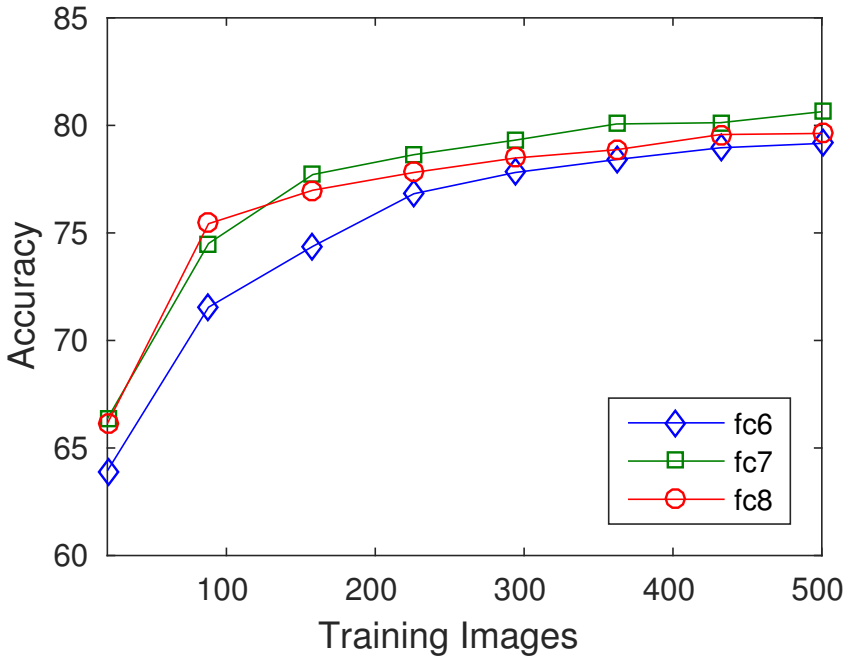


Figure 2.4: Region classification accuracy versus training set size.

at random and train a classifier, using the Places *fc8* features. We repeat this 60 times for each N and show the average. The results show that with a relatively small number of training examples (only 250 per region) we obtain relatively high accuracy. Increasing the number of training examples results in only minor improvements. This highlights that the feature space is geo-informative.

2.4 Overhead Imagery Analysis

We now present insights into the performance of representations captured by ImageNet and Places, but applied to overhead imagery. Despite being trained solely on ground-level imagery, our experiments show that both the ImageNet and Places CNNs extract strongly location-related features from overhead imagery.

2.4.1 Visualizing Deep Image Representations

To understand the differences between deep image representations for ground-level and overhead images, we visualize the response of the units of various layers in the Places model. Our strategy for generating the visualization of these responses follows that of Zhou et al. [146]: given a unit (i.e., node, neuron) at a layer, push a set of images through

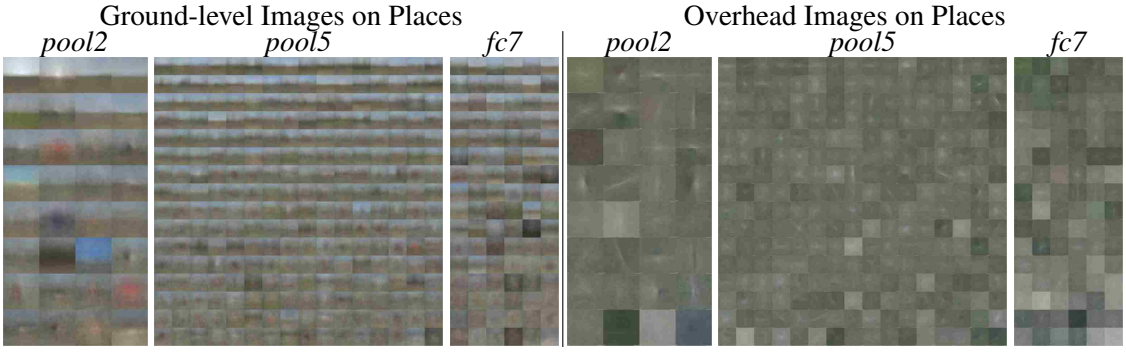


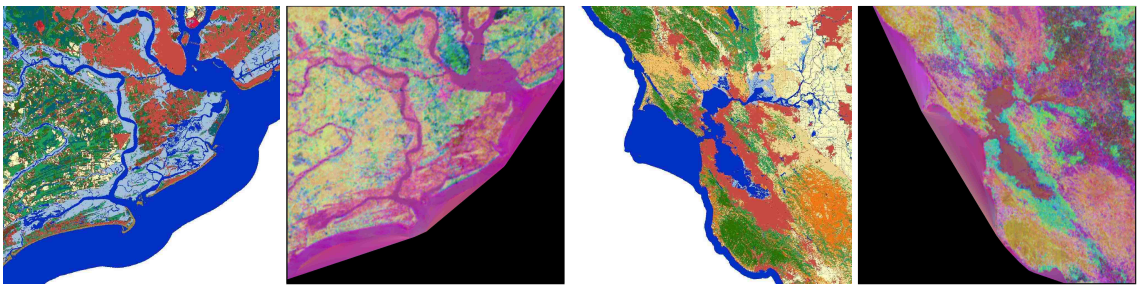
Figure 2.5: (left) Averages of the top 100 images that activate a subset of *pool2*, *pool5*, and *fc*7 layers of ground-level images on the Places model. Each montage is sorted by the first PCA coefficient of the corresponding image. (right) The result of the same procedure applied to overhead images. Note that unlike ground-level imagery, the average images of the overhead imagery are more uniform due to the nature of the viewpoint.

the network, sort the images by their activation response at that unit, and average the top 100 images. The result is a visualization that captures the receptive field of that unit. To produce the final visualization for a layer, the mean images from different units are sorted by the first principal component coefficient computed using PCA. Figure 2.5 visualizes the learned representations of the *pool2*, *pool5*, and *fc7* layers, for a set of ground-level and overhead images.

There is a drastic difference between the representations for ground-level and overhead images. Similar to that shown by Zhou et al. [146], the receptive fields for the ground-level set of images look like landscapes and other spatial structures, for instance in several of the mean images you can make out the sky or buildings. This is not the case for the receptive fields from the overhead set of images, where some filters seem to correspond to road orientation, terrain, and vegetation. These visualizations indicate that deep image representations are informative for overhead images despite being trained on images of a different viewpoint.

2.4.2 Exploring Relationship with Land Cover

We begin with an analysis of deep image representations for visualizing and understanding land cover. For this experiment, we use *fc8* features from the Places model, computed for each image in the overhead reference database of both the Charleston and San Francisco datasets. Starting from these features, we apply principal component analysis (PCA) to generate a synthetic overhead image for each region which we then visualize and compare versus a ground-truth land cover map. The results of this process are shown in Figure 2.6.



(a) Charleston (b) San Francisco

Figure 2.6: Synthetic overhead images (right), constructed by performing PCA analysis on Places *fc8* features from small overhead images, highlights different types of land cover (left). For example, regions that are over water (pink), forest (yellow), and urban (green) areas are all clearly visible as unique colors.

To generate the synthetic overhead image we use the top three principal components. For each map location we have a 3D PCA coefficient; we use the first, second, and third coefficient as the red, green, and blue color channels, scaling each color channel to $[0, 1]$ and using natural neighbors interpolation. The result is an image that encodes the dominant feature appearance variations as different colors. Upon closer inspection of the synthetic overhead images and corresponding land cover maps, the three PCA coefficients of the CNN feature vectors of overhead imagery are very closely related to land cover. For example, pink corresponds to areas containing water, and green corresponds to urban areas.

To explore this relationship further, we augmented the overhead images in the reference map database for San Francisco with a ground-truth land cover label obtained from the National Land Cover Database [36] (NLCD). NLCD uses a 16 class land cover classification scheme which we aggregate into higher level groups resulting in 8 classes: water, developed, barren, forest, shrubland, herbaceous, planted/cultivated and wetlands. Figure 2.7 shows the embedding computed using t-Distributed Stochastic Neighbor Embedding [120] (t-SNE), a popular nonlinear dimensionality reduction technique, visualized against the land cover labels. Despite not being trained on overhead images, or for the task of land cover estimation, the features are highly geo-informative and closely related to land cover classes. The embedding clearly groups together overhead images with the same ground-truth land cover class.

2.4.3 Analyzing Semantic Co-occurrence Between Viewpoints

Given our findings that deep image representations are useful for describing both ground-level and overhead imagery, we now investigate whether such representations are useful as a joint feature representation by analyzing the co-occurrence of feature activations for

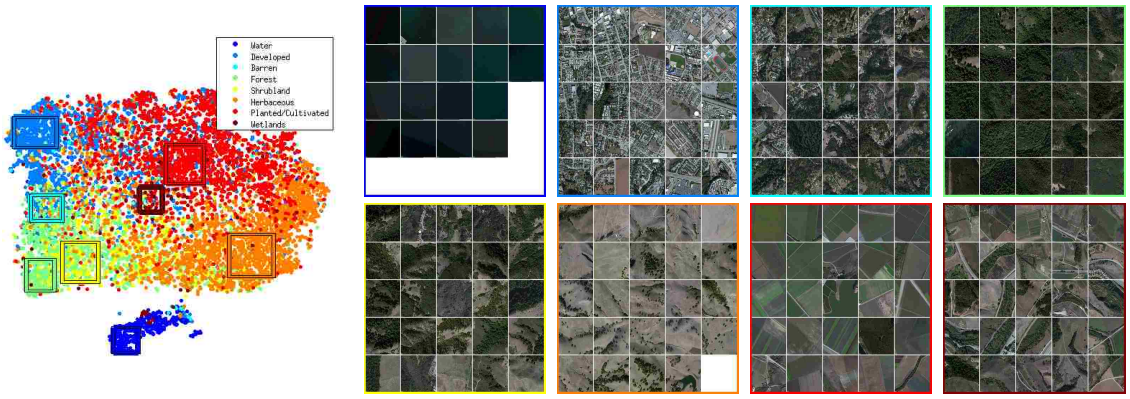


Figure 2.7: Visualizing land cover in overhead imagery using t-SNE [120], a non-linear unsupervised dimensionality reduction technique, to embed Places *fc8* features. The embedding produces well defined clusters in relation to the ground-truth land cover classes and shows separation in the high-dimensional feature space.

ground-level and overhead images captured at the same location. Our hypothesis is that deep image representations from models trained solely on ground-based imagery are positively correlated for co-located overhead imagery, despite drastic differences in viewpoint.

For this experiment, we take advantage of the known semantic meaning which can be inferred from the last fully connected layer, *fc8*, in the AlexNet architecture. Applying the softmax function, $\sigma(x_j) = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}}$, to each element of the K-dimensional output vector $\mathbf{x}$, results in a categorical probability distribution, $\sigma(\mathbf{x})$, over K classes, since $\sum_{i=1}^{K} \sigma(x_i) = 1$. For the Places model, this results in a distribution over 205 scene classes.

We start with the ground-level and overhead image pairs in the San Francisco dataset. For each image, we compute the *fc8* feature from Places and convert it to the corresponding categorical probability distribution. To examine the relationship between the distributions of co-located images, we compute the Pearson correlation coefficient between each pair of a subset of the 205 scene classes. To do this, we stack the distributions together to form two matrices, one for overhead and one for ground-level images, of size 174,217 $\times$ 205, with dimensions corresponding to number of images and number of scene classes, respectively. We then compute the Pearson's correlation coefficient between each pair of columns in these two matrices, producing a correlation matrix of size 205×205.

To highlight semantic groupings, we sort the correlation matrix using a Ward-linkage based hierarchical clustering algorithm applied to each row. This results in a semantically meaningful ordering of the scene categories. Figure 2.8 visualizes the result of this experiment. We observe that the correlation matrix has a visible block diagonal structure. The main block in the upper left contains mostly rural classes and the larger block in the lower right is mostly urban classes. Examples of classes for which the ground-level and overhead

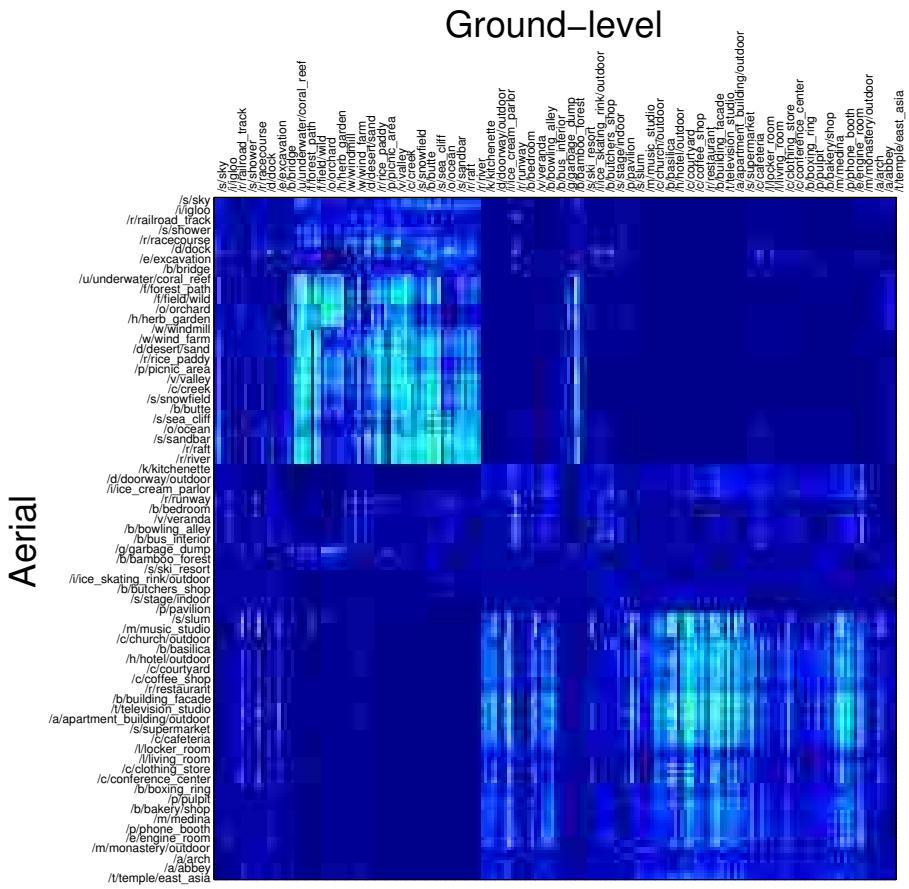


Figure 2.8: Analyzing the semantic co-occurrence of features extracted from co-located ground-level and overhead imagery in the San Francisco dataset. See Section 2.4.3 for algorithm details. Due to space constraints only every third label is shown.

image features are highly correlated include "desert" and "river". These results support our hypothesis that deep image representations are useful as a joint feature representation for ground-level and overhead images.

2.4.4 Improving Geospatial Modeling

Given the encouraging results from the previous sections, we now investigate the extent to which overhead imagery can be used to directly make predictions and improve geospatial modeling. Our hypothesis is that overhead imagery can be leveraged to augment ground-level imagery in order to create fine-grained geospatial models.

We test our hypothesis using the following strategy. Similar to Section 2.4.3, we compute the categorical probability distribution from the Places model for each ground-level image in the San Francisco dataset. We then select the entries corresponding to three scene

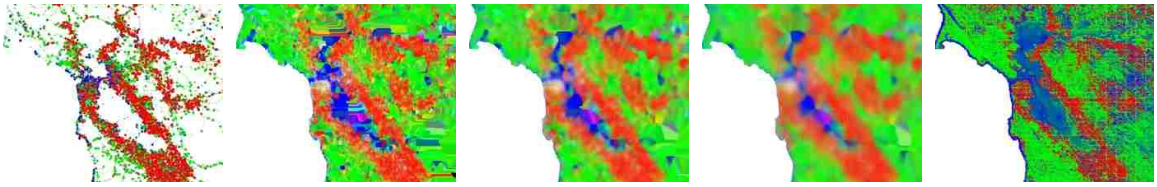


Figure 2.9: Leveraging overhead imagery to improve geospatial modeling (see Section 2.4.4 for details). The results shown correspond to three scene categories (urban="parking lot", rural="field/wild", and water="ocean") for the San Francisco dataset. The images above represent false-color distributions (red=urban, green=rural and blue=water) represented by: (left) a scatter plot of ground-level images, (middle) Nadaraya-Watson kernel regression with three different bandwidths on the sparse samples, and (right) using dense overhead imagery instead.

classes that are highly correlated between ground-level and overhead viewpoints: "parking lot", "field/wild", and "ocean". This results in a 3×1 vector of class probabilities for each image. We then generate a scatter plot of the ground-level image locations, using the class probabilities as a false-color image (the red channel corresponding to "parking lot", green as "field/wild", and blue as "ocean"). Figure 2.9 (left) visualizes this intermediate result. We then interpolate this sparse set of samples using Nadaraya-Watson kernel regression, with the latitude/longitude location of each ground-level image as the input features. In Figure 2.9 (middle), the results show that no choice of kernel bandwidth is free of noticeable artifacts. The result is either too smooth or too noisy.

To overcome the artifacts introduced by interpolating sparse ground-level imagery, we apply the same strategy to the densely sampled overhead imagery in the San Francisco map database instead. In Figure 2.9 (right), the results show this method is able to capture high resolution local structure with minimal artifacts. As an example, the coastline of San Francisco is clearly visible only when using dense overhead imagery, as opposed to the sparse ground-level samples, or kernel regression techniques. These results support our hypothesis that overhead imagery is useful for improving geospatial modeling. This suggests that cross-view, i.e., ground-level and overhead, image analysis is a powerful tool for capturing geospatial distributions. The unprecedented density, both spatial and temporal, of overhead imagery has the potential to make maintaining up-to-date high-resolution geospatial models much more cost-effective.

2.4.5 Image-Based Search

The richness of these deep image representations applied to overhead imagery suggests a novel user-focused application in image-based search. Consider the following scenario: a

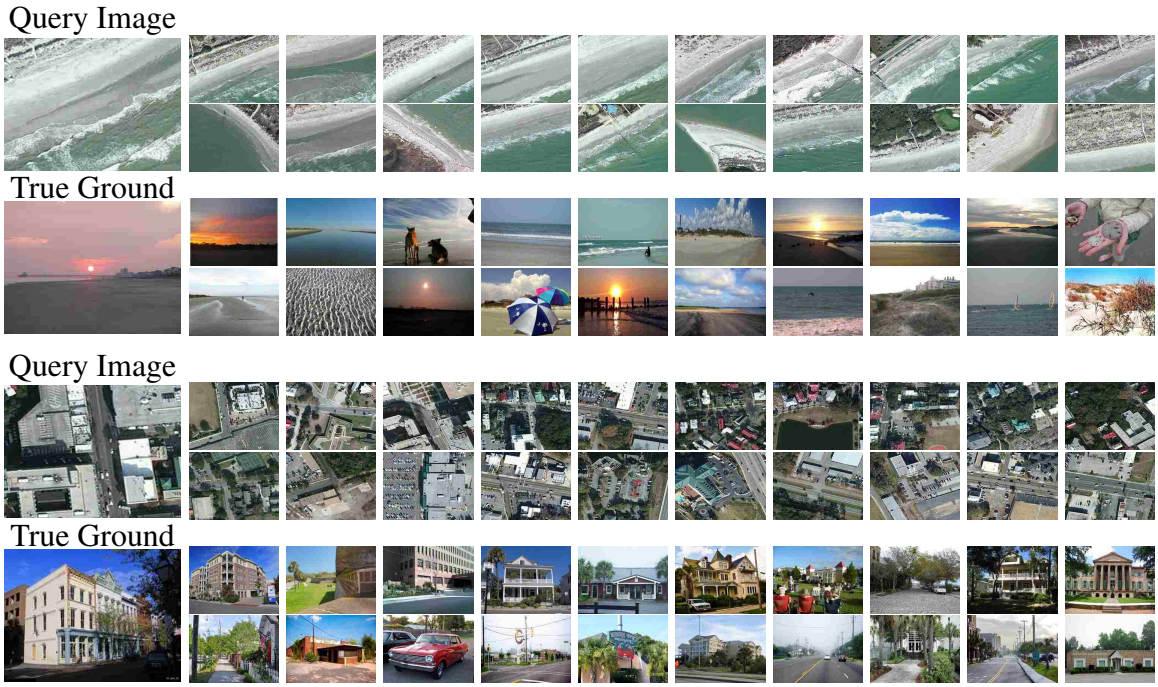


Figure 2.10: Overhead image-based search for characterizing unimaged ground-level locations. Given a query overhead image (top, left), we find the most similar overhead images (top, right) in the map database, and infer hypothetical ground-level images (bottom, right). The results are realistic when compared to the true ground-level image (bottom, left).

person is browsing a map and is curious about the ground-level appearance of a particular location, but no ground-level image is available at that particular location.

We propose a method to search for ground-level images using only the current overhead image and a reference dataset of overhead and ground-level image pairs. Our approach consists of three steps: (1) compute CNN features on the current map location, (2) compute the Euclidean distance between this feature and all map images in the reference database, and (3) present the user with the ground-level images that had the most similar overhead images.

Several results of this approach, shown in Figure 2.10, demonstrate that we are able to retrieve a realistic set of ground-level images (as compared to the true ground-level image) by querying on the appearance of the overhead view. Our approach clearly finds images that would not have been found by matching on the ground-level view. For instance, in Figure 2.10 (bottom), querying using the overhead image results in images that do not contain a building, contrary to what one would expect if the query were the ground-level image (which contains a building).

2.5 Cross-View Image Matching

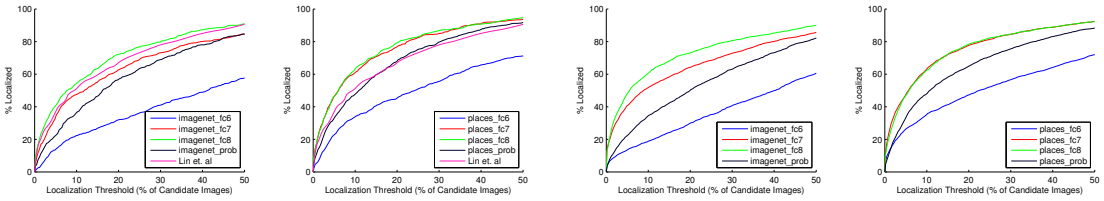
The canonical computer vision task in this domain is image localization. Given an image, where was it captured? While some images provide strong localization cues and are easily found, such as a view of the Statue of Liberty from Ellis Island or the Coliseum in Rome, others only provide weak evidence of their geographic location. For such images, it may only be possible to guess the region in which the image was taken. A wide variety of approaches have been proposed for the former problem, while the latter problem has only received significant attention recently.

Data-driven image localization is often reformulated as an image retrieval problem, often called visual place recognition. Standard approaches use machine learning techniques to find visually overlapping images from a reference set of ground-level images with known geographic location. These methods generally fall into two categories, matching using local features [4, 13, 15, 100, 107] or global image features [34, 43]. Many other cues for localization have been explored which take advantage of photometric and geometric properties [41, 42, 128].

When no nearby ground-level imagery is available, existing methods that localize via direct visual similarity [34] are not applicable. To address this, the problem of cross-view image localization [66, 67, 127, 129] has recently been investigated. In this scenario, ground-level imagery is matched to overhead imagery instead. The underlying premise is that overhead imagery, which is available practically everywhere compared to the relatively sparse coverage of geotagged ground-level images, can be exploited to produce dense localization estimates.

The cross-view localization problem is inherently more difficult than the single-view problem, due to the dramatic differences in viewpoint of the two image sets. Lin et al. [66] explored several strategies for characterizing this relationship. Their methods build on hand-engineered global image descriptors, such as GIST, and combine them with land cover attributes in an attempt to learn a feature translation between the two viewpoints. Their most successful method combines a feature averaging strategy with a supervised learning technique.

Given our findings that deep image representations are highly location-dependent, even for overhead imagery, we analyze their performance for this task. Our strategy is similar to Lin et al. [66]: given a query image, we first find the closest 30 ground-level images in the training set by comparing their associated feature vector. For this set of neighbors, we average the features of their corresponding overhead images and use this as our query to search the overhead image reference database. In both cases, we use Euclidean distance as



(a) Charleston (b) San Francisco

Figure 2.11: Accuracy of localization as a function of retrieved candidate locations. Our method, using Places *fc8* features, significantly outperforms Lin et al. [66], the previous best method on the Charleston dataset.

our distance metric. We convert the distance, d, between two feature vectors to a similarity score using: $s = \frac{1}{1+d}$. This results in a score for every map location.

We evaluate this technique on both the Charleston and San Francisco datasets, using the isolated set of images (images for which no nearby ground-level images exist). The performance metric used is the same as described by Lin et al. [66]; given the scores for each location, we compute the rank of the ground-truth location in the sorted list. Figure 2.11 visualizes our results as a cumulative distribution function of the fraction of query images correctly localized versus the percentage of candidate images retrieved.

In Figure 2.11 (left), we compare the results of our method on the Charleston dataset, using several model and feature combinations (here *prob* corresponds to the categorical probability distribution produced by applying a softmax to the *fc8* feature). Our approach is highly effective, outperforming Lin et al. [66], the previous state-of-the-art method of those using hand-engineered features, by a large margin without requiring any land cover imagery, manual selection of features, or learning. In terms of top 1% accuracy, our best result, using Places *fc8* features, correctly localizes 18.45% of query images versus the 17.37% reported by Lin et al. [66], a 1.08% increase and a relative improvement of 6.22%. This trend continues as the localization threshold, the percentage of candidate images retrieved, is increased. Figure 2.11 (right) shows similar results on the San Francisco dataset. In both cases, we find Places *fc8* features to be the best for this task.

In Figure 2.12, we visualize the localization result for three example query images from each dataset using several different features extracted from the Places model. To generate these results, we visualize the similarity scores for each map location as a heatmap, where red indicates a higher likelihood that the image was captured at that location. Similar to the quantitative results in Figure 2.11, we observe qualitatively that *fc8* features outperform the features from other layers.

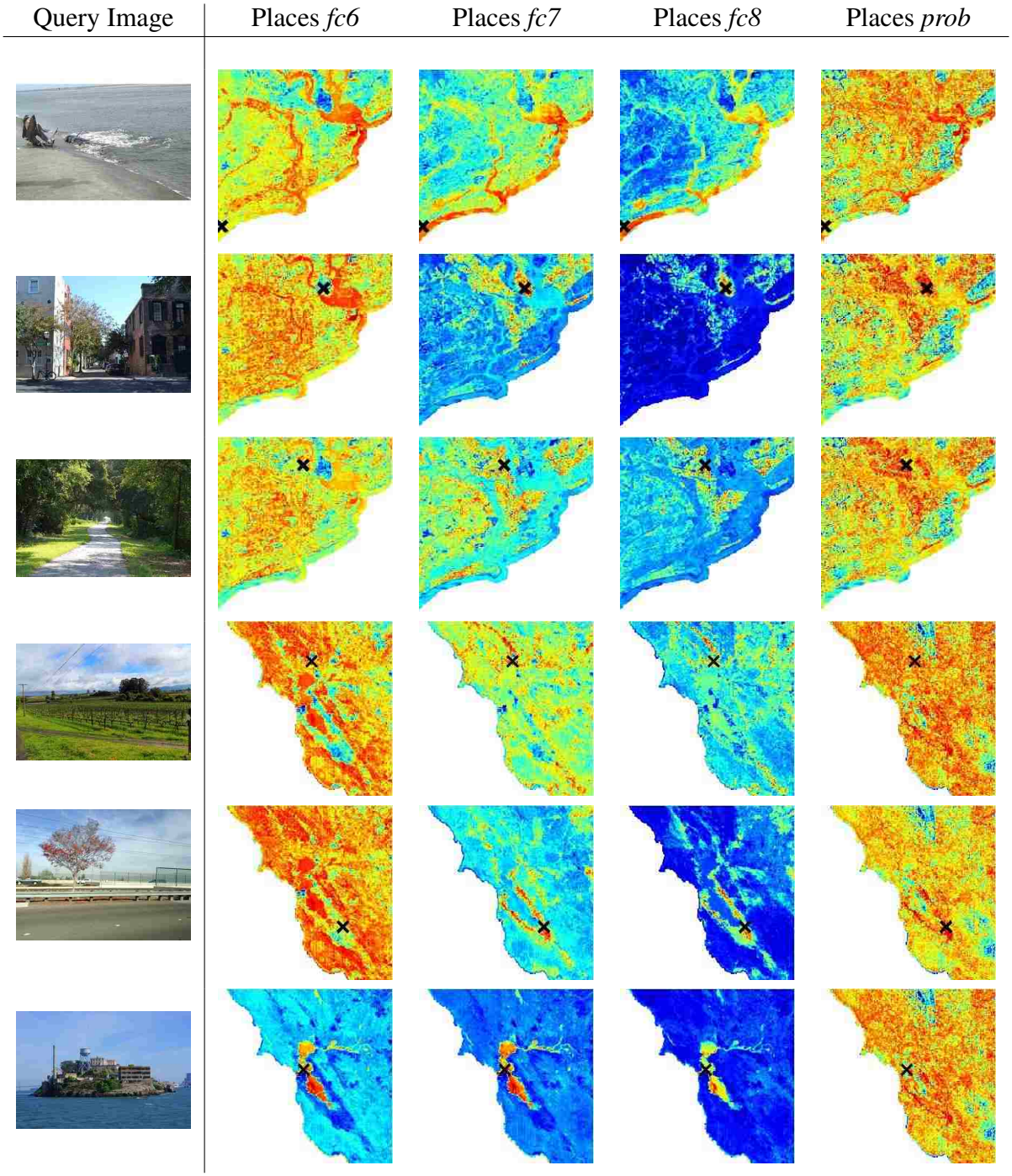


Figure 2.12: False-color images that represent the likelihood that an image is at a particular location. In each, red represents high likelihood, blue represents low, and the 'x' marks the true location. See Section 2.5 for an algorithm description.

2.6 Conclusion

This chapter attempted to answer the question, "Are deep image representations geo-informative?" Through experiments on a wide variety of geolocation-related computer vision tasks, we found that deep image representations are significantly more powerful than hand-engineered features.

Our experiments showed that for distinguishing the region of ground-level images, deep image representations outperformed a commonly used off-the-shelf feature descriptor and also provide a method to identify images that capture the relative appearance of two places. In addition, we found that CNN features give state-of-the-art results on the challenging problem of cross-view image geolocalization, when compared with methods that use hand-engineered features.

Interestingly, features from the Places model, which was trained for scene classification, outperformed features from the ImageNet model, which was trained for object recognition, on all geospatial problems we explored. This is in line with recent findings in a study of transferability of features by Yosinski et al. [139]; when transferring features, performance is related to the specificity of the task. In other words, features suited for scene classification appear to be more geo-informative than those for object detection.

Most notably, we found that both the ImageNet and Places models extract strongly location-related features on overhead imagery, and demonstrated that these features can be exploited to relate co-located ground-level and overhead images. This is surprising because these models were trained on imagery from a vastly different viewpoint. This points to a promising direction for future research in building deep-learning based models that are directly targeted at problems of localization and location-related feature extraction from ground-level and overhead imagery.

Chapter 3

Wide-Area Image Geolocalization with Overhead Reference Imagery

3.1 Introduction

In this chapter, we address the problem of cross-view image geolocalization, which aims to localize ground-level query images by matching against a database of overhead images (Figure 3.1). 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 geotagged 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 [15].

Cross-view image geolocalization is motivated by the observation that the distribution of geotagged ground-level imagery is relatively sparse in comparison to the abundance of high-resolution overhead imagery. The underlying idea is to learn a mapping between ground-level and overhead image viewpoints, such that a ground-level query image can be directly matched against an overhead image reference database. In contrast to previous work [66] 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 [54, 146].

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 overhead image pairs of the same location to learn to extract semantic, geo-informative

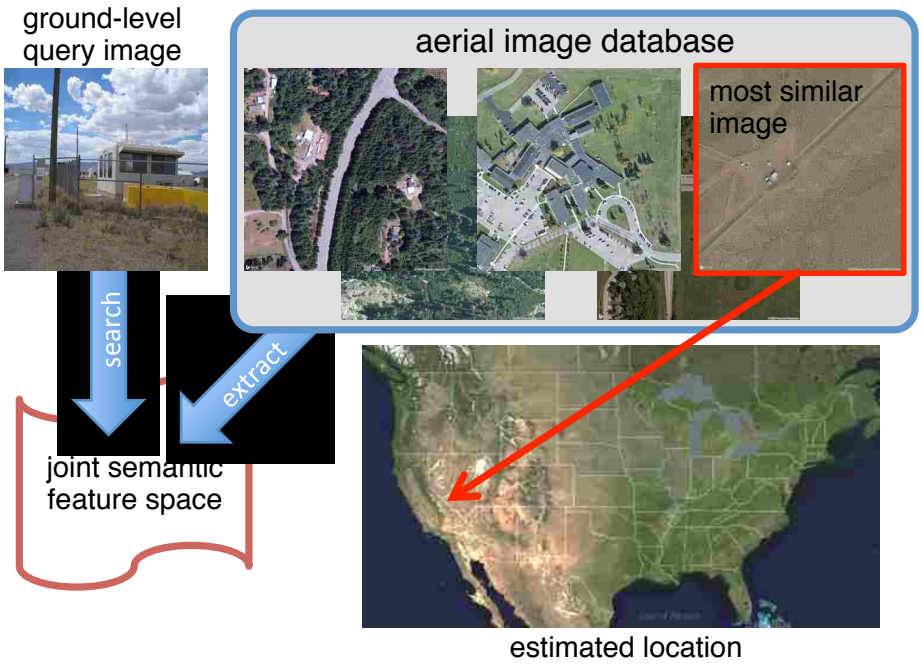


Figure 3.1: We learn a joint semantic feature representation for overhead and ground-level imagery and apply this representation to the problem of cross-view image geolocalization.

features from overhead 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 overhead imagery; (4) state-of-the-art performance on two smaller-scale evaluation benchmarks for cross-view geolocalization; and (5) extensive qualitative evaluation, including visualizations, which highlights the utility of cross-view training.

3.2 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 [4,19,34,53,61,149]. A wide variety of visual cues have been investigated, including photometric and geometric properties such as sun position [14, 56, 128], shadows [48, 99, 133], and weather [41, 42, 111].

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 geotagged 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 [4, 13, 15, 100, 107, 116, 140]. The second category of methods match using global image features [34,43,149]. 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 [19,23,39,89]. 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 [54], object detection [30], and scene classification [146]. Razavian et al. [91] further show that these feature hierarchies are useful as generic descriptors. Lee et al. [59] estimate geo-informative attributes from an image using convolutional neural network classifiers.

Only recently has overhead imagery been discovered as a valuable resource for ground-level image understanding [7, 70]. Shan et al. [104] geo-register ground-level multi-view stereo models using ground-to-aerial image matching. Viswanathan et al. [121] 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. [66]. Workman et al. [127] show that features extracted from convolutional neural networks are useful for problems in geospatial image analysis. Most akin to our work, Lin et al. [67] apply a siamese CNN architecture for learning a joint feature representation between ground-level images and 45° oblique overhead imagery. Our approach is more general; we operate on orthorectified overhead imagery, do not require scale and depth metadata for each query, and our joint feature representation is semantic.

3.3 Cross-View Training for Overhead Image Feature Extraction

We propose a cross-view training strategy that uses deep convolutional neural networks to extract features from overhead imagery. The key idea is to use pre-existing CNNs for extracting ground-level image features and then learn to predict these features from overhead

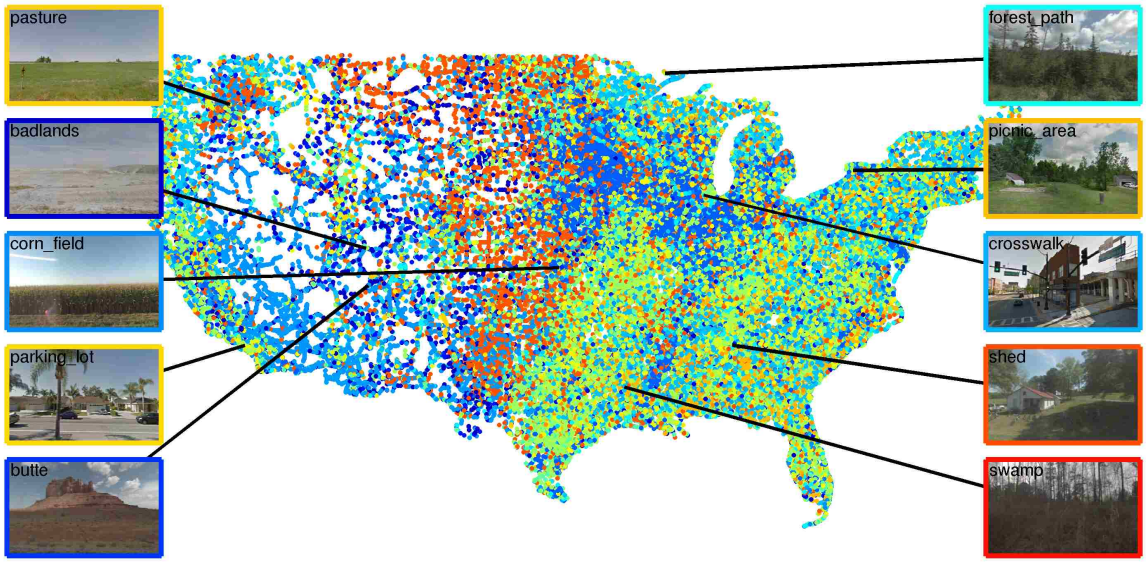


Figure 3.2: Existing CNNs trained on ground-level imagery provide high-level semantic representations which can be location dependent. Each point represents a geotagged image extracted from a Google Street View panorama, colored according to the predicted scene category from the Places [146] network.

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 [16], where the source domain is the ground-level view and the target domain is overhead imagery. The end result of cross-view training is a CNN that is able to extract semantically meaningful features from overhead images without manually specifying semantic labels.

3.3.1 Cross-View Feature Representations

We assume the existence of two functions: $f_a(l;\Theta_a)$, which extracts features from the overhead 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 overhead images. This is possible if the location of the ground-level imagery is known. For example, in Figure 3.2, we show the estimated label from the Places [146] 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 overhead 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 3.4 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 overhead imagery.

3.3.2 Cross-View Training a Single-Scale Model

Given a semantically meaningful feature representation for ground imagery, we propose to extract features from overhead 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:

$$J(\Theta_a) = \sum_i \|f_a(l_i; \Theta_a) - f_g(I_i; \Theta_g)\|_2. \tag{3.1}$$

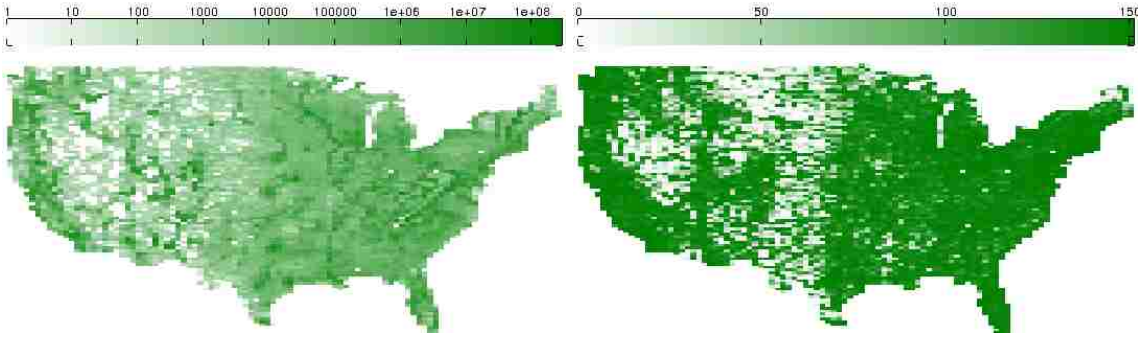
Intuitively, the objective is to learn to extract features from the overhead imagery that match those from a corresponding ground-level image.

3.3.3 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 overhead imagery. To address this issue, we extend our overhead image feature function, f_a, to support extracting features at multiple spatial scales. Rather than mapping a single ground-level image to a single overhead image, the multi-scale approach allows for a ground-level image to be matched to overhead images at multiple scales. In support of multi-scale, cross-view training, we introduce a large dataset of ground-level and overhead image pairs.

3.3.4 A Large Cross-View Training Dataset

Previous cross-view datasets have been limited in spatial scale and number of training images. The largest dataset [127] contains 174,217 training image pairs sampled from a $200km \times 200km$ area around San Francisco. Features learned using such a dataset are



(a) Google Street View (b) Flickr

Figure 3.3: The distribution of ground-level images in the CVUSA dataset.

Figure 3.4: Example matched ground-level and overhead images from the CVUSA dataset.

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 overhead images from across the United States, called the Cross-View USA (CVUSA) dataset.

Geotagged ground-level images were collected from both Google Street View and Flickr. For Google Street 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 [146] 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 3.3 visualizes the relative density of each set of images. For each ground-level image, we downloaded an 800×800 overhead 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 overhead image locations and a total of 1,588,655 million geotagged, image matched pairs. Figure 3.4 shows several example matched ground-level and

overhead images from our dataset.

3.4 Application to Cross-View Localization

We focus on the problem of cross-view image geolocalization [66] in which the goal is to use a database of overhead 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 overhead viewpoints.

3.4.1 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. [66] 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 overhead image and land-cover attribute map centered at its location. The overhead image reference database contains 182,988 images. The second benchmark dataset, San Francisco, is introduced by Workman et al. [127] 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 overhead image centered at the ground-level image location. The overhead 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.

3.4.2 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 overhead 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. [66], 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 overhead 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.

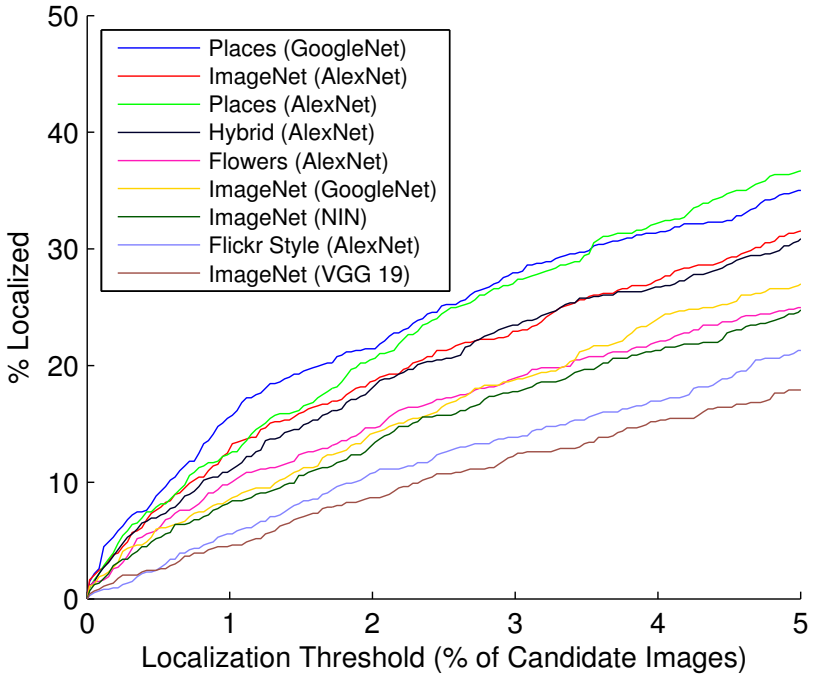
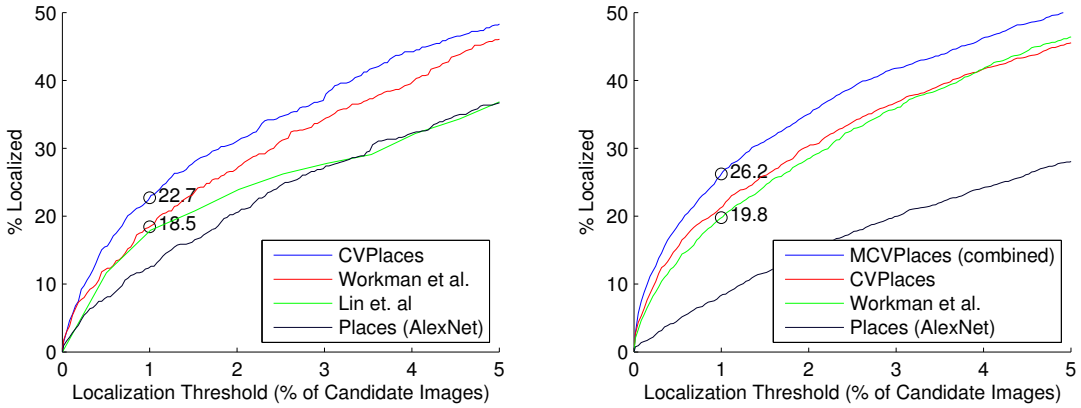


Figure 3.5: Comparison of several off-the-shelf CNN features in terms of localization accuracy on the Charleston dataset.

3.4.3 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 overhead and ground-level query image using a variety of network architectures trained for different target label spaces. The network architectures used included GoogleNet [112], AlexNet [54], NIN [65], and VGG 19 [12]. Training databases included Places [146], ImageNet [98], Hybrid [146], Oxford Flowers [81], and Flickr Style [49]. We evaluated multiple such configurations, all publicly available as Caffe [47] model files.

Our findings from this experiment are visualized in Figure 3.5. The top two performing configurations in terms of top 5% accuracy are trained for the task of scene classification on the Places [146] database, which contains over two million images labeled from 205 different categories. These two networks vastly outperform the next best network, 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 overhead imagery. For the rest of the experiments, we apply cross-view training to learn an overhead image feature extractor for Places features using the *AlexNet* architecture [146], which we refer to as *Places*.



(a) Charleston (b) San Francisco

Figure 3.6: Accuracy of localization as a function of retrieved candidate locations on two benchmark datasets.

3.4.4 Localization using Cross-View Features

The *AlexNet* architecture [54] 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 1,000 possible categorical labels).

Given the architecture and weights, Θ_g, of *Places*, we apply the cross-view training approach described in Section 3.3 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 overhead images in our CVUSA dataset (zoom level 18). We refer to this model as *CVPlaces*. Figure 3.6 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 1,000 matched pairs of images from each benchmarks training set as a validation set for model selection. Our models are implemented using the Caffe toolbox [47] and trained using stochastic gradient descent with a Euclidean loss for parameter fitting to reflect (3.1).

3.4.5 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 overhead 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 overhead imagery (zoom levels 16 and 14). Figure 3.6 shows a comparison of our multi-scale approach versus our single-scale approach and a recent 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%.

3.5 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.

3.5.1 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/overhead images from CVUSA and recorded the activations for each. Figure 3.7 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 overhead 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 overhead images where particular scene categories are likely to be observed from a ground-level viewpoint.

3.5.2 Geospatial Visualization of Overhead Image Features

We visualize the geospatial distribution of high-level features extracted from the high-resolution overhead reference imagery from the Charleston dataset [66]. The result is a

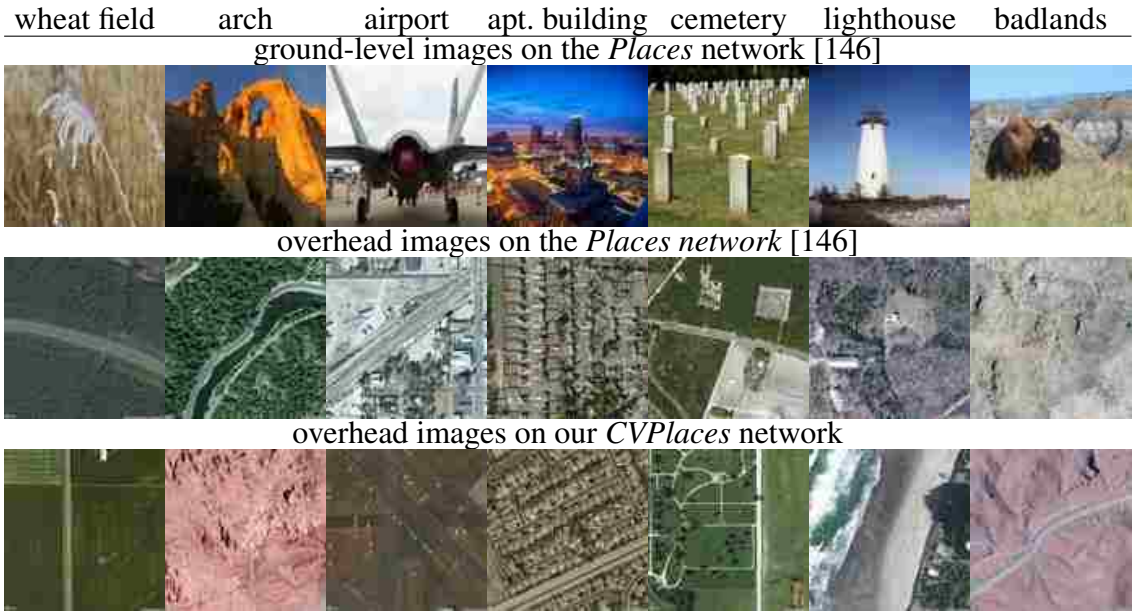


Figure 3.7: Images that result in high activations for particular scene categories. (top) The high-activation ground-level images are exemplars for the corresponding semantic class. (middle) The high-activation overhead images for the network trained on ground-level images are, not surprisingly, less 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 overhead domain, the high-activation images are a better match to the respective categories.

coarse-resolution false-color image that summarizes the semantic information extracted by a particular CNN from the overhead 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 overhead 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 overhead image (Figure 3.8) 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 overhead 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.

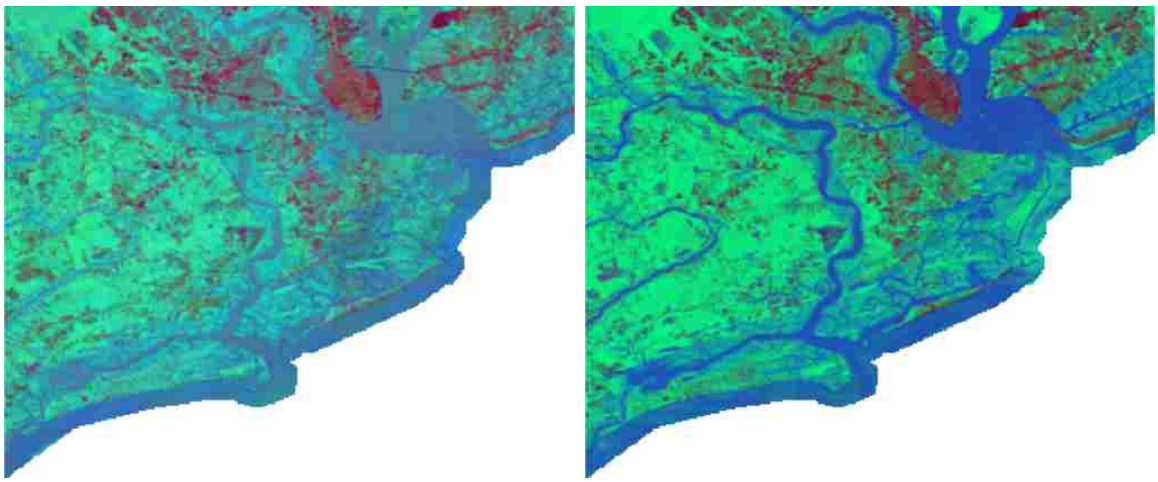


Figure 3.8: (left) A false-color image generated by applying the *Places* network to overhead imagery. In both images the colors are semantically meaningful (red=urban, green=rural, blue=water-related). (right) The same as (left) but with our *CVPlaces* network (trained on the entire USA dataset, with no Charleston-specific fine tuning).

3.5.3 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 overhead images in the dataset. Figure 3.9 shows qualitative results as a heatmap that represents the distance between the query and corresponding overhead 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 3.10 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 between the *Places fc8* feature vector of the query image and the corresponding *CVPlaces* feature of the sub-window of the overhead 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

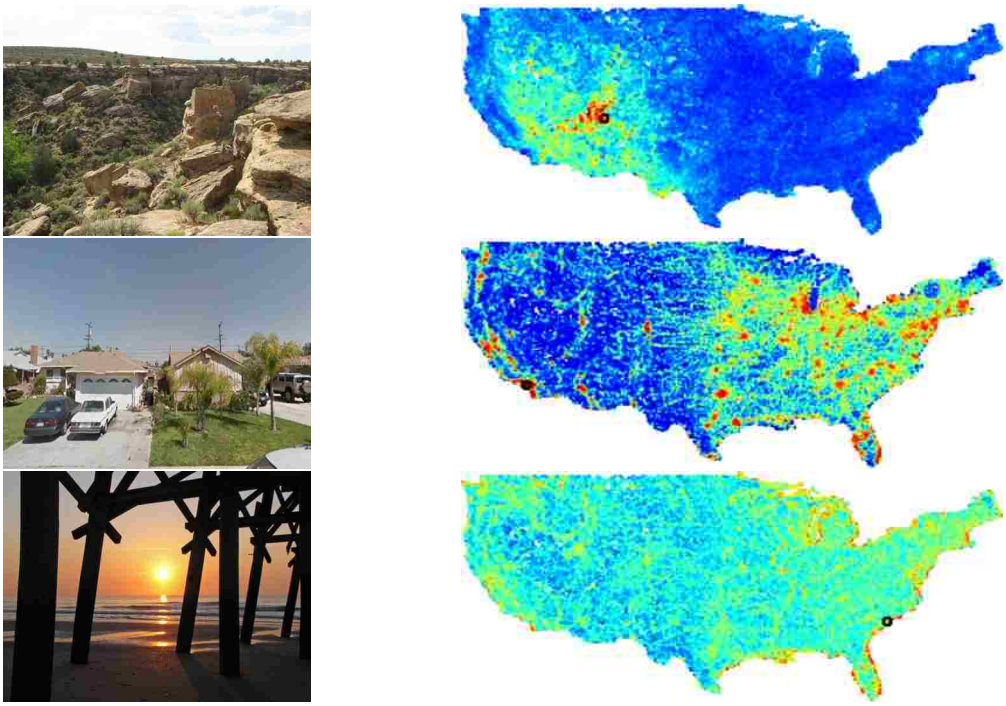


Figure 3.9: Localization examples at a continental scale. (left) A ground-level query image. (right) A heatmap of the distance between the *Places fc8* feature of the query image and the corresponding *CVPlaces* feature of an overhead image at that location (red: more likely location, blue: less likely location). The black circle marks the true location of the camera.

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.

3.6 Conclusion

We proposed a cross-view training approach, in which we learn to predict features extracted from ground-level imagery from overhead imagery of the same location. We introduced a massive dataset of such pairs and proposed single and multi-scale networks for extracting overhead 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 overhead imagery. We tried fixing the overhead parameters, Θ_a, using pre-existing networks, and optimizing over Θ_g, but the performance was poor. We also attempted jointly optimiz-

Figure 3.10: Examples of localization at finer spatial scales. (top) The ground-level query image. (middle) An overhead image centered at the ground location. (bottom) An overlay showing the distance between the ground-level image feature and the overhead image features at each location, computed using a sliding window approach (red: more likely, blue: less likely).

ing 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 overhead 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.

Chapter 4

Understanding and Mapping Natural Beauty

4.1 Introduction

Recent advances in learning with large-scale image collections have led to methods that go beyond identifying objects and their interactions toward quantifying seemingly subjective high-level properties of the scene. For example, Isola et al. [40] explore image memorability, finding that memorability is a stable property of images that can be predicted based on the image attributes and features. Other similar high-level image properties include photographic style [110], virality [18], specificity [45], and humor [11]. Quantifying such properties facilitates new applications in image understanding.

In this chapter we consider "scenicness", or the natural beauty of outdoor scenes. Despite the popularity of the saying "beauty lies in the eye of the beholder," research shows that beauty is not purely subjective [57]. For example, consider the images in Figure 4.1; mountainous landscapes captured from an elevated position are consistently rated as more beautiful by humans than images of power transmission towers.

Understanding the perception of landscapes has been an active research area (see [151] for a comprehensive review) with real-world importance. For example, McGranahan [75] derives a natural amenities index and shows that rural population change is strongly related to the attractiveness of a place to live, as well as an area's popularity for retirement or recreation. Seresinhe et al. [102] show that inhabitants of more beautiful environments report better overall health. Runge et al. [97] characterize locations by their visual attributes and describe a system for scenic route planning. Lu et al. [69] recover cues from millions of geotagged photos to suggest customized travel routes.

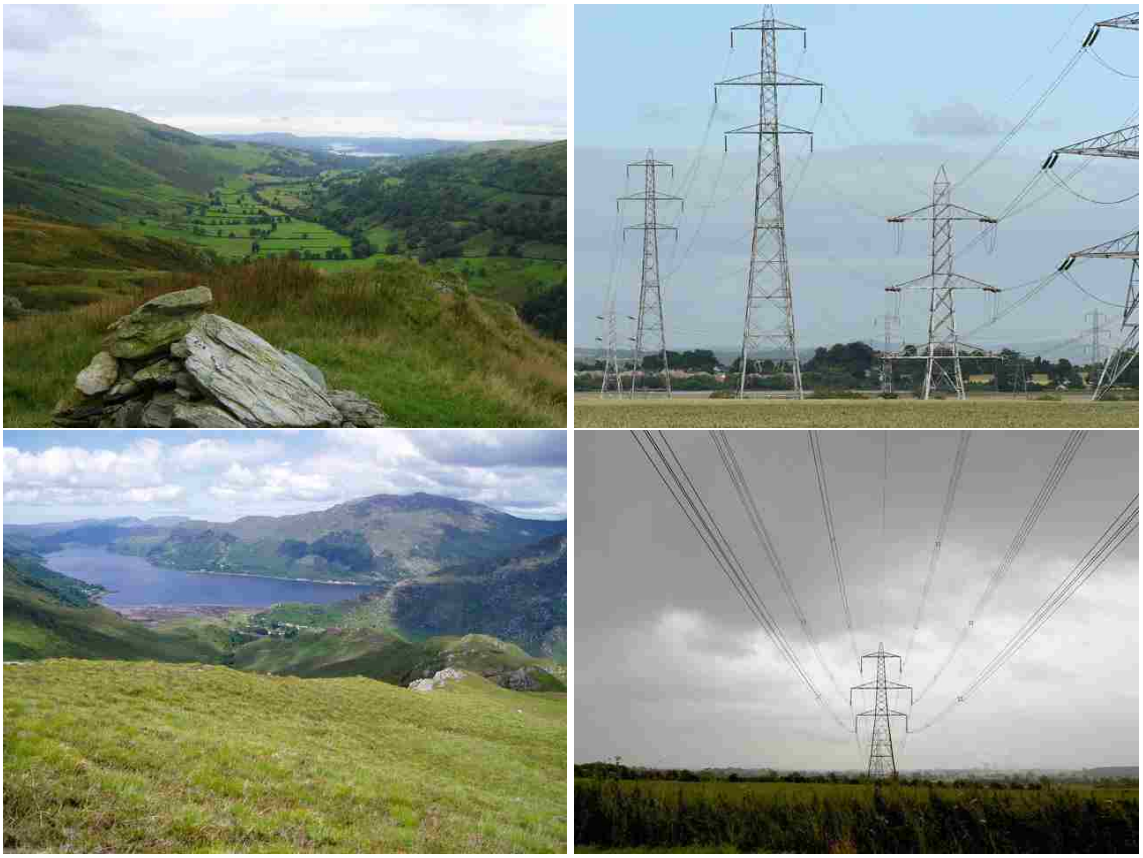
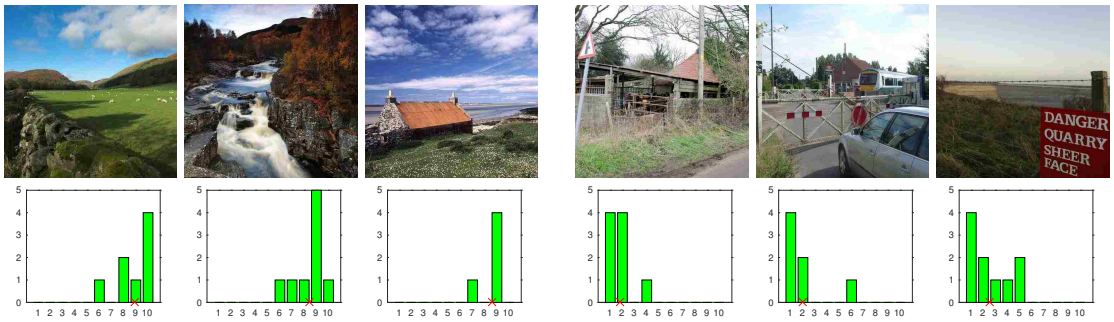


Figure 4.1: Most observers agree that images of mountains are more scenic than power lines. Our work seeks to automatically quantify "scenicness" and demonstrate applications in image understanding and mapping.

Recently, a number of algorithms have been developed to automatically interpret high-level properties of images. Laffont et al. [55] introduce a set of transient scene attributes and train regressors for estimating them in novel images. Lorenzo et al. [87] use a convolutional neural network to estimate urban perception from a single image. Deza and Parikh [18] study the phenomenon of image virality. Similarly, a significant amount of work has sought to understand the relationship between images and their aesthetics [50, 71, 138]. Karayev et al. [49] recognize photographic style. Su et al. [110] propose a method for scenic photo quality assessment using hand-engineered features. Developed independently from our work, Seresinhe et al. [101, 103] explore models for quantifying scenicness. Lu et al. [68] apply deep learning to rate images as high or low aesthetic quality.

In this work, we start with a large-scale dataset containing hundreds of thousands of images, individually rated by humans, to quantify and predict image scenicness. Our main contributions are:

- an analysis of outdoor images to identify semantic concepts correlated with scenic-



(a) Scenic (b) Non-Scenic

Figure 4.2: Example images (and human-provided scenicness ratings) from the ScenicOrNot (SoN) dataset: (a) "scenic" images (average rating above 7.0) and (b) "non-scenic" images (average rating below 3.0).

ness;

- a method for estimating the scenicness of an image which accounts for variance in the ratings and human perception of scenicness;
- a new dataset of ground-level and overhead images with crowdsourced scenicness scores; and
- a novel cross-view mapping approach, which incorporates both ground-level and overhead imagery to address the spatial sparsity of ground-level images, to provide country-scale predictions of scenicness.

4.2 Exploring Image Scenicness

Our work builds on a publicly-available crowd-sourced database collected as part of an online game, ScenicOrNot,[1] which contains images captured throughout Great Britain. As part of the game, users are presented a series of images from around the island of Great Britain and invited to rate them according to their scenicness, or natural beauty, on a scale from 1-10. From a user standpoint, in addition to being exposed to the diverse environments of England, Scotland and Wales, the purpose of the game is to compare aesthetic judgments against those of other users.

We apply our work to a database of 185,548 images and associated natural beauty rating histograms. Each image in the dataset was rated at least five times. We refer to this set of images as the ScenicOrNot (SoN) dataset. In addition to retaining the rating distribution and average rating, we partition the set of images into "scenic" (average rating above 7.0)

[1] ScenicOrNot (`http://scenicornot.datasciencelab.co.uk/`) is built on top of Geograph (`http://www.geograph.org.uk/`), an online community and photo-sharing website.

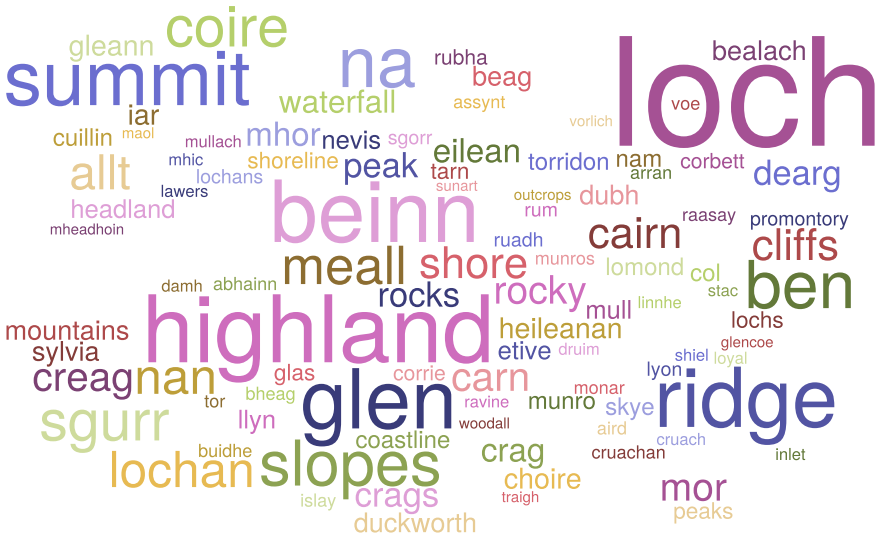


Figure 4.3: The word cloud depicts the relative frequency of title and caption terms found in scenic images from the SoN dataset.

and "non-scenic" (average rating below 3.0) subgroups. Figure 4.2 shows sample images from the dataset. In the remainder of this section, we explore image properties that may be associated with scenicness, including: text annotations, color statistics, and semantic image attributes.

4.2.1 Image Captions

Like most images hosted on image sharing sites, the SoN images have associated metadata, including a title and caption. For example, the image in Figure 4.1 (top, left) is titled *From Troutbeck Tongue* and has the following caption: "Looking over the cairn down Trout Beck. Windermere and the sea in the distance". For all of the images in the SoN dataset, we analyzed the title and captions to explore whether these associated text annotations are correlated with scenicness.

Using the scenic and non-scenic subsets, we compute the relative term frequency for each of the extracted words. Figure 4.3 shows a word cloud of the most frequent 100 extracted terms from scenic images, where the size of the word represents the relative frequency. While some of the terms (e.g., "ridge", "cliffs", "summit") may universally correlate with scenicness, other terms, such as "loch", "na", and "beinn" reflect the fact the data originates from Great Britain. Conversely, example terms that are negatively correlated with scenicness include "road", "lane", "house", and "railway".

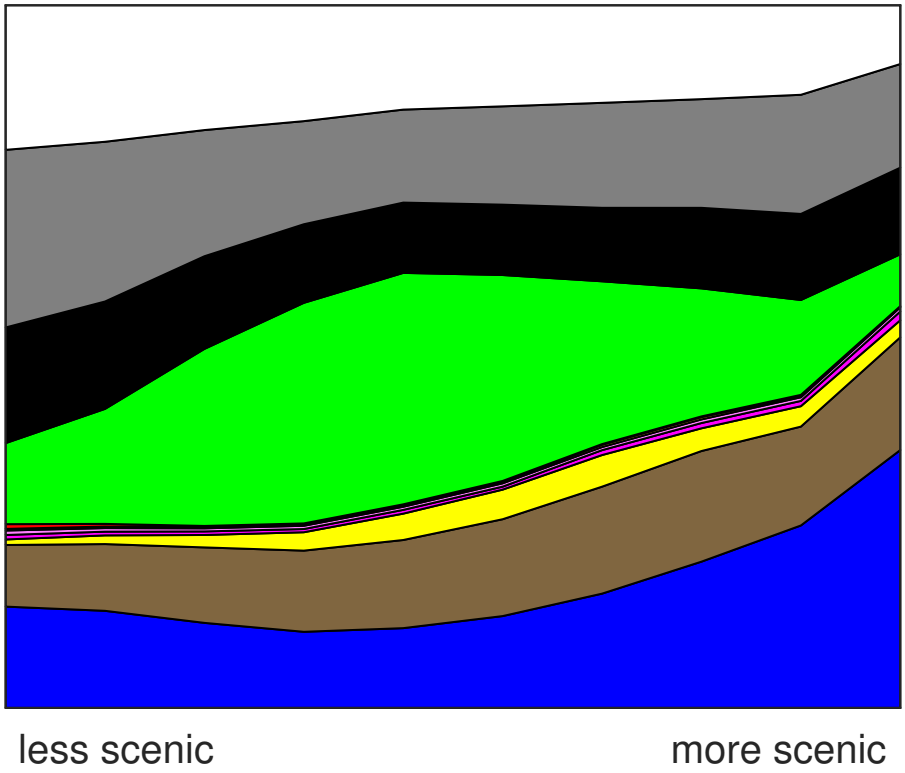


Figure 4.4: Distribution of color with respect to the average scenicness rating of the SoN image set.

4.2.2 Color Distributions

The images in Figure 4.2 and terms in Figure 4.3 suggest that images with blue skies, green fields, water, and other natural features tend to be rated as more scenic. For this analysis, we computed the distribution of quantized color values, using the approach of Van De Weijer et al. [119], as a function of the average scenicness rating of the SoN image set. Figure 4.4 shows the distribution, where we see blue overrepresented in scenic images and, conversely, black and gray overrepresented in non-scenic images.

4.2.3 Scene Semantics

For each image, we compute SUN attributes [84], a set of 102 discriminative scene attributes spanning several types (e.g., function, materials). Figure 4.5 shows an occurrence matrix for a subset of attributes correlated with image scenicness. Attributes such as "asphalt", "man-made", and "transporting things or people" occur often in less scenic images, suggesting that urban environments are more typical of images with low scenicness. In contrast, attributes such as "ocean", "climbing", and "sailing/boating" occur more often in

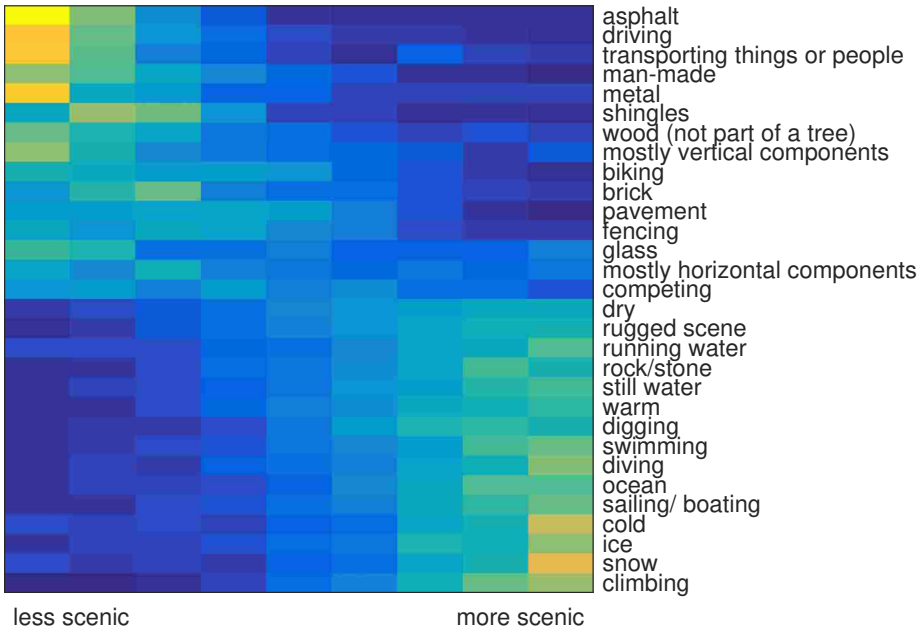


Figure 4.5: Distribution of the frequency of SUN attributes [84] in "scenic" versus "not scenic" images. Warm colors indicate higher frequency.

the most scenic images.

Similarly, we compared scenicness to the scene categorizations generated by the Places [146] convolutional neural network. Of the 205 Places scene classes (e.g., "airplane cabin", "hotel room", "shed"), 135 describe outdoor categories. We aggregate the outdoor classes into seven higher-level scene categories (similar to Runge et al. [97]), such as "buildings and roads", "nature and woods", and "hills and mountains". Each image is classified using Places into one of these high-level categories. Figure 4.6 shows the frequency of each category as a function of the average user-provided rating of SoN images. The trend follows previously observed patterns; on the whole, images containing natural features, such as hills, mountains, and water, are rated as more scenic than images containing buildings, roads, and other man-made constructs.

4.2.4 Summary

This analysis shows that scenicness is related to both low-level image characteristics, such as color, and semantic properties, such as extracted attributes and scene categories. This suggests that it is possible to estimate scenicness from images. In the following section, we propose a method for directly estimating image scenicness from raw pixel values.

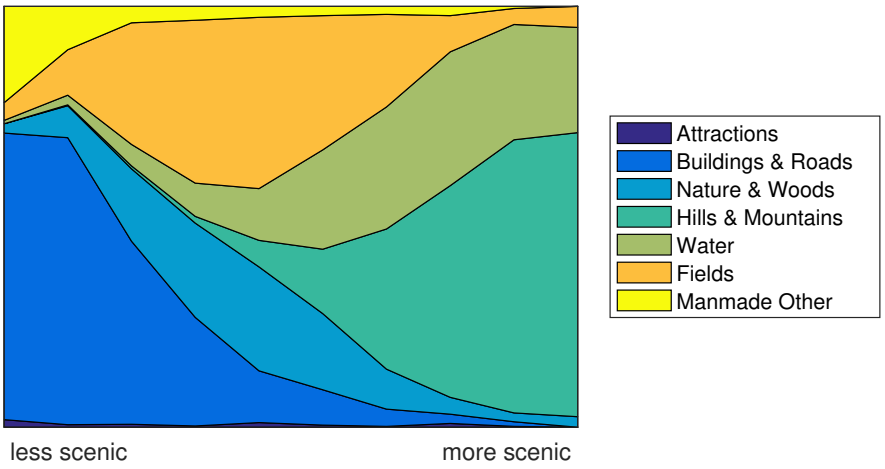


Figure 4.6: Distribution of high-level categories for the images in the SoN dataset.

4.3 Predicting Image Scenicness

We use a deep convolutional neural network (CNN) to address the task of automatically estimating the scenicness of an image. Following other approaches (e.g., [126, 132]), we partition the output space and treat this prediction as a discrete labeling task where the output layer corresponds to the integer ratings (i.e., $1, 2, \ldots, 10$) of scenicness. We represent our CNN as a function, $G(I; \Theta_g)$, where I is an image and the output is a probability distribution over the 10 scenicness levels. We consider multiple loss functions during training to best capture the distribution in human ratings of scenicness for a given image.

The baseline approach follows recent work (e.g., [136]), which trains a model to predict a single value. For this variant, each image is associated with the label corresponding to the mean human rating, rounded to the nearest integer value, $\bar{r}$. Training involves minimizing the typical cross-entropy loss:

$$E = -\frac{1}{N}\sum_{n=1}^{N} \log(G(I_n; \Theta_g)(\bar{r}_n)), \tag{4.1}$$

where N is the number of training examples.

The baseline approach assumes a single underlying value for scenicness. However, as shown in Figure 4.2, for many images, there may be high variability in the ratings. In these cases, the mean scenicness may not serve as a representative value. So, instead of directly predicting the mean scenicness, we train the model to predict the human rating distribution for a particular image. For this variant, we treat the normalized human ratings as a target distribution and train the model to predict this distribution directly, by minimizing the

cross-entropy loss:

$$E = -\frac{1}{N}\sum_{n=1}^{N}\sum_{r=1}^{10} p_{nr}\log(G(I_n;\Theta_g)(r)), \tag{4.2}$$

where p_{nr} is the proportion of r ratings for image n.

However, the previous formulation assumes a large number of ratings so that p_n approaches the true distribution. In our case, this assumption does not hold. As an alternative to predicting the mean scenicness or the empirical scenicness distribution, we model the set of ratings for an image as a sample from a multinomial distribution. Each training example is associated with a set of (potentially noisy) labels $\{(I_1, \{v_{1i}\}), \ldots, (I_N, \{v_{Ni}\})\}$, where $\{v_{ji}\}$ is the set of ratings for image I_j. This results in the following loss:

$$E = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{V_n} p_{ni}\log(G(I_n;\Theta_g)(v_{ni})), \tag{4.3}$$

where V_n is the total number of ratings for image n.

4.3.1 Comparison with Human Ratings

We evaluate our scenicness predictions using the SoN dataset. We reserved 1,413 images that have at least ten ratings as test cases for evaluation, with the remaining data used for training and validation. For predicting scenicness, we modify the GoogleNet architecture [112] with weights initialized from the *Places* network [146]. We selected this CNN because it had been trained for the related task of outdoor scene classification; however, our methods could be applied to other related architectures or trained from scratch with sufficient data. Our implementation uses the Caffe [47] deep learning toolbox. For training, we randomly initialize the last layer weights and optimize parameters using stochastic gradient descent with a base learning rate of 10^{-4} and a mini-batch size of 40 images. Roughly 10% of the training data is reserved for validation.

We refer to the three models as: (1) AVERAGE, the baseline approach that predicts the mean scenicness (Equation 4.1); (2) DISTRIBUTION, the model that minimizes cross-entropy loss to the normalized distribution of human ratings (Equation 4.2); and (3) MULTINOMIAL, which maximizes the multinomial log-likelihood (Equation 4.3). We compare performance on two tasks: (1) predicting the average human rating and (2) predicting the distribution of ratings for a given image.

The output of each network is a posterior probability for each integer rating for a given input image. To evaluate the average user predictions, we consider the order of the predictions, ranked by posterior probability and use the information retrieval metric, *Normalized*

Table 4.1: Quantitative results comparing models with different loss functions. For each metric, higher is better.

Loss	Metric	
	nDCG	K-S
AVERAGE	.9780	14.8%
DISTRIBUTION	.9678	50.0%
MULTINOMIAL	.9745	58.4%

Discounted Cumulative Gain (nDCG), which penalizes "out of order" posterior probabilities, given the ground-truth rating. The second column of Table 4.1 shows the nDCG scores for each of the three models. Overall, the models trained using different loss functions performed similarly well under this evaluation metric.

For the task of predicting the distribution of ratings for a given image, the performance of the models diverged. We take a hypothesis testing approach and consider whether or not the set of human ratings could be drawn from the distribution represented by the output probabilities of the CNN. For this, we applied the one-sample *Kolmogorov-Smirnov (K-S)* test with a non-parametric distribution and computed the proportion of testing images for which the human ratings come from the posterior distribution at the 5% significance level. The last column of Table 4.1 shows the percentage of testing images that matched the predicted distribution. The models trained using distribution of ratings, DISTRIBUTION and MULTINOMIAL, significantly outperform the model trained on average rating, with MULTINOMIAL showing the best performance.

Figure 4.7 visualizes these results qualitatively. Several example images are shown alongside the distribution of human ratings (green) and predictions from the three models. In general, the results follow the quantitative analysis. The MULTINOMIAL method better captures human uncertainty as compared to the other methods. For example, in Figure 4.7 (row 1), the baseline approach, AVERAGE, provides a much higher posterior probability for a rating of 2 than the distribution of humans ratings. Comparatively, MULTINOMIAL is more consistent with human ratings and closer to the average user predictions. For the remaining experiments, the MULTINOMIAL model is used unless otherwise specified.

4.3.2 Receptive Fields of Natural Beauty

For additional insight into our scenicness predictions, following Zhou et al. [147], we apply receptive field analysis to highlight the regions of the image that are most salient in generating the output distribution. Briefly, the approach computes the differences in output predictions for a given image with a small (i.e., 7×7) mask applied. Using a sliding

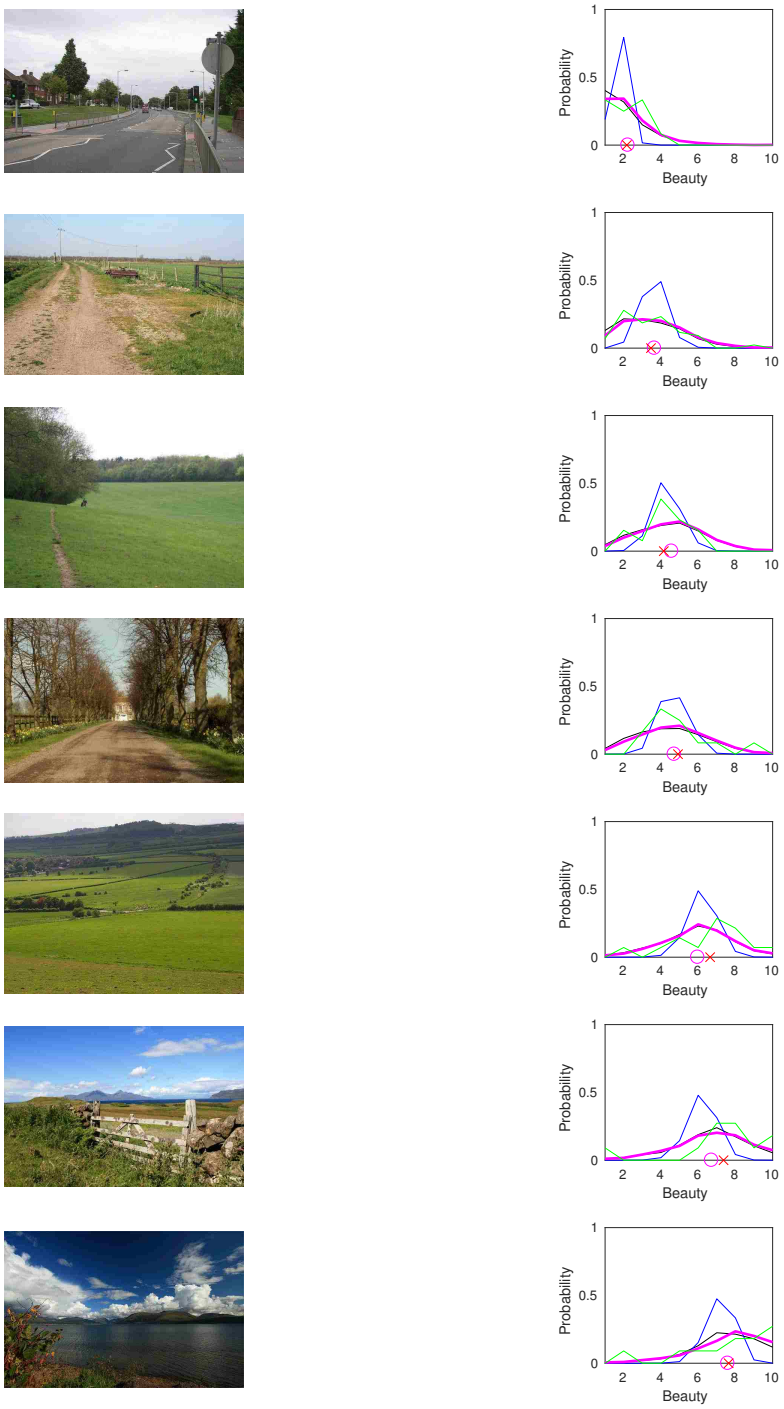


Figure 4.7: Example images alongside the distribution of human ratings (green), and the outputs of AVERAGE (blue), DISTRIBUTION (black), and MULTINOMIAL (magenta). The red $\times$ corresponds to the mean rating and the magenta $\circ$ the weighted average of the MULTINOMIAL prediction.

window approach, the prediction differences (compared to the unmasked image) are computed on a grid across the image. A large difference signifies the masked region plays a significant role in the output prediction. This process leads to a saliency map over the input image. For visualization purposes, we represent the map as a binary mask (thresholded at 0.6). Figure 4.8 shows several examples of this analysis. Each pair of images shows the input and the image regions with the most contribution to the (high or low) scenicness score. In most cases, the receptive fields match the intuition and semantic analysis of scenicness. Regions containing water, trees, and horizons contribute to scenicness, while man-made objects, such as buildings and cars, contribute to non-scenicness.

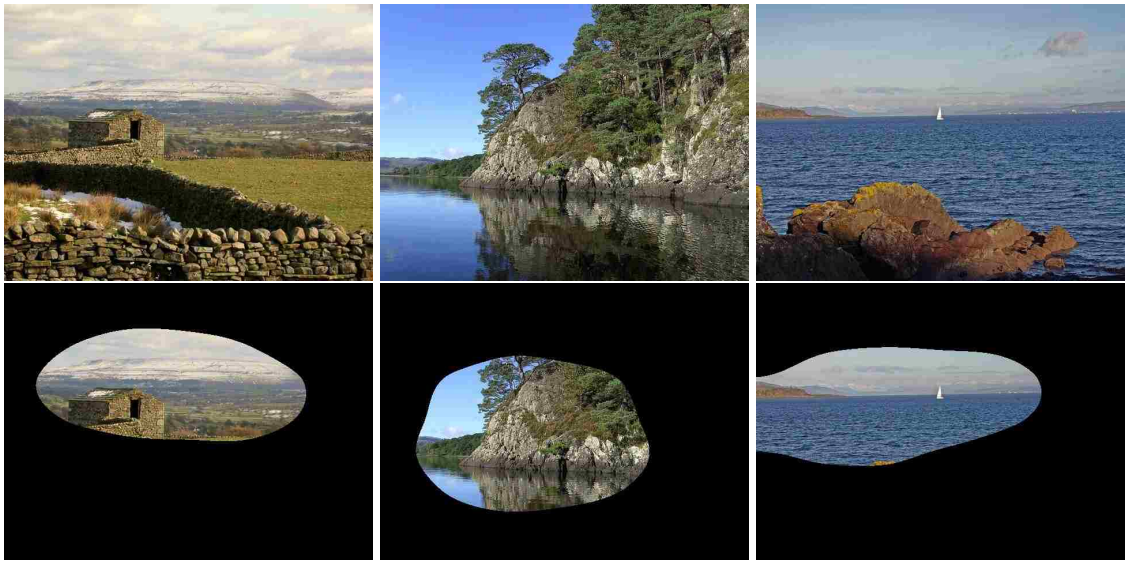
4.3.3 Scenicness-Aware Image Cropping

The previous experiment shows that components within a given image contribute differently to the overall scenicness. For this experiment, we solve for the image crop that maximizes scenicness. This approach follows the style of previous methods for content-aware image processing (e.g., seam carving for image resizing [3]). We used constrained Bayesian optimization [27] to solve for the position and size of the maximally scenic image crop, where scenicness is measured as the weighted average prediction from the MULTINOMIAL network. Figure 4.9 shows representative examples. In some cases, cropping improved the scenicness scores greatly. For example, in the top image in Figure 4.9, cropping out the vehicles increased the predicted scenicness from 5.0 to 7.3.

4.4 Mapping Image Scenicness

The previous sections considered scenicness as a property of an image. Here, we consider scenicness as a property of geographic locations and propose a novel approach for estimating scenicness over a large spatial region. We extend our approach for single-image estimation to incorporate overhead imagery. The result is a dense, high-resolution map that reflects the scenicness for every location in a region of interest. Such a map could, for example, be used to provide driving directions optimized for "sight seeing" [90, 97] or suggest places to go for a walk [88].

We consider geotagged images as noisy samples of the underlying geospatial scenicness function. The challenge is that ground-level imagery is sparsely distributed, especially away from major urban areas and tourist attractions. This means that methods which estimate maps using only ground-level imagery [2, 87, 136] typically generate either low-resolution or noisy maps.



(a) Scenic

(b) Non-Scenic

Figure 4.8: Network receptive field analysis. Given an input image (top), the output mask (bottom) highlights the region(s) that most significantly impact the maximal label assigned by our network.

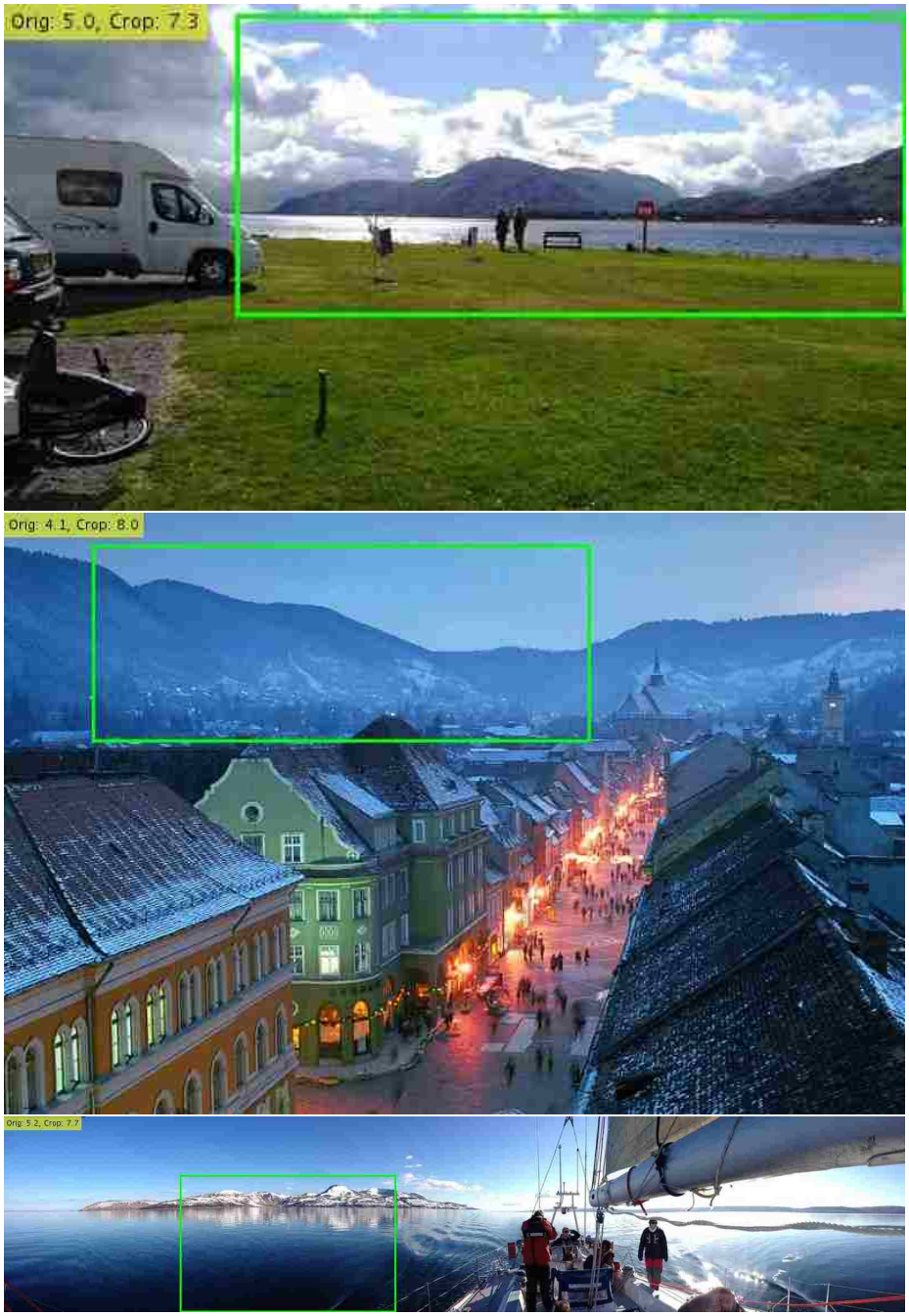


Figure 4.9: For each image, the green bounding box shows the image crop that maximizes scenicness. The predicted scenicness scores for both the entire image and the cropped region are shown in the inset.

Figure 4.10: Examples of the co-located ground-level (top) and overhead (bottom) image pairs contained in the Cross-View ScenicOrNot (CVSoN) dataset.

To deal with the problem of interpolating sparse examples over large spatial regions, we apply a cross-view training and mapping approach. Cross-view methods [67, 129, 141] incorporate both ground-level and overhead viewpoints and take advantage of the fact that, while ground-level images are spatially sparse, overhead imagery is available at a high-resolution in most locations. Jointly reasoning about ground-level and overhead imagery has become popular in recent years. Luo et al. [70] use overhead imagery to perform better event recognition by fusing complementary views. Lin et al. [66, 67] introduce the problem of cross-view geolocalization, where an overhead image reference database is used to support ground-level image localization by learning a feature mapping between the two viewpoints. Workman et al. [127, 129] study the geo-dependence of image features and propose a cross-view training approach.

To support these efforts, we extend the ScenicOrNot (SoN) dataset to incorporate overhead images. Specifically, for each geotagged, ground-level SoN image, we obtained a 256×256 orthorectified overhead image centered at that location from Bing Maps (zoom level 16, which is $\sim$2.4 meters/pixel). Figure 4.10 shows co-located pairs of ground-level and overhead images from the Cross-View ScenicOrNot (CVSoN) dataset.

4.4.1 Cross-View Mapping

To predict the scenicness of an overhead image even though labeled overhead images are not available, we apply a cross-view training strategy; instead of predicting the scenicness of the overhead image, we predict the scenicness of a ground-level image captured at the same location. We use the same network architecture and training methods as with the ground-level network, with two changes: (1) overhead (instead of ground-level) images are used as input and (2) the weights are initialized with those learned from the ground-level

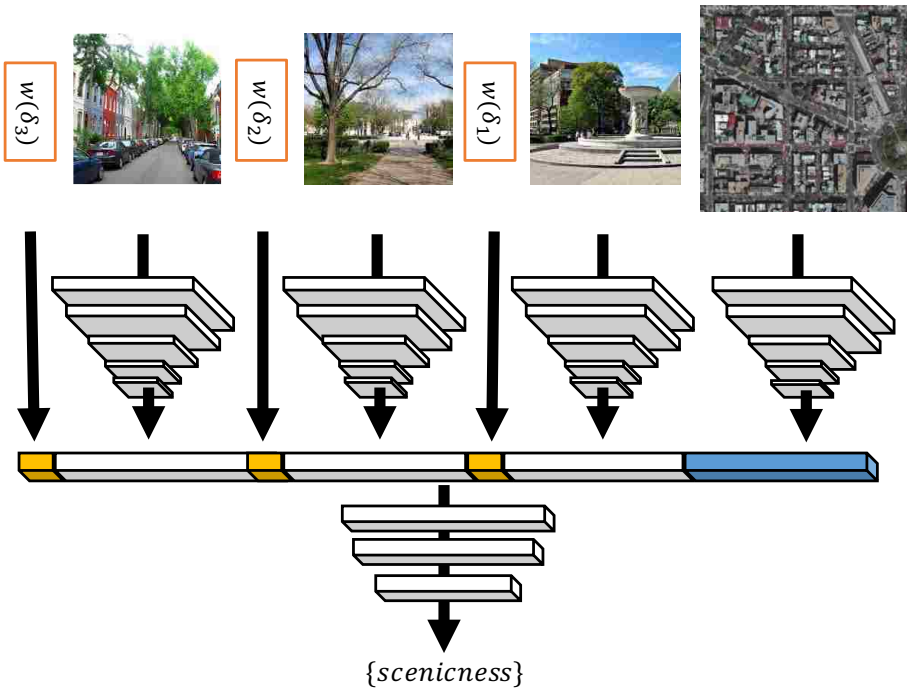


Figure 4.11: The architecture for our hybrid approach to cross-view mapping.

network. Similar to our ground-level network, after training, the output of this overhead image network is a distribution over scenicness ratings.

While using overhead images as input may address the issue of sparse spatial coverage of ground-level imagery, an overhead-only network may miss, for example, scenic views hidden amongst dense urban areas. To address this issue, we introduce a novel variant to the cross-view approach for combining ground-level and overhead imagery to estimate the scenicness of a query location. This is similar to our framework for estimating geospatial functions [131].

Figure 4.11 shows an overview of our hybrid cross-view approach. For a given query location, q, consider the co-located overhead image, set of the k closest ground-level images, and the distances of the ground-level images to the query location, $\{\delta_1, \delta_2, \ldots, \delta_k\}$. For the images, we can compute scenicness features using the existing ground and overhead networks. For the hybrid approach, we learn and predict scenicness from the fused features (overhead image features, ground-level features, weighted distances) using a small feed-forward network, with three hidden layers containing 100, 50, and 25 neurons, respectively. The activation function on the internal nodes is the hyperbolic tangent sigmoid. The network weights are regularized using an L_2 loss with a weight of 0.5. The output is the predicted distribution of ratings for a ground-level image taken at the input location. We refer to this as the *Cross-View Hybrid (CVH)* network.

Table 4.2: Comparison of mapping strategies.

Method	1NN	LWA	CVH
AUC	64.38%	66.86%	68.51%

4.4.2 Mapping the Scenicness of Great Britain

To evaluate CVH, the CVSoN dataset is divided as before, with the same 1,413 ground-level images (with at least 10 ratings) held out for testing. For CVH, the test input includes the co-located overhead image. We compare against two baseline methods for constructing dense maps of visual properties:

- *1NN*: return the prediction from the ground-level image closest to the query location; and
- *LWA*: return the locally weighted average prediction of neighboring ground-level images with a Gaussian kernel ($\sigma = 0.01$ degrees).

To compare our methods, we formulate a binary classification task to determine if a given test image is above or below a scenicness rating of 7. Table 4.2 shows the results for each method as the area under the curve (AUC) of the ROC curve computed from the output distributions. The results show that including orthographic overhead imagery improves the resulting predictions.

These results are supported qualitatively in Figure 4.12, which shows scenicness maps for several regions around Great Britain. We observe that by including overhead imagery we are able to construct a significantly more accurate map than purely interpolating scenicness estimates obtained from ground-level images alone. The maps created using only ground-level images (e.g., 1NN, LWA) are susceptible to both underprediction (e.g., no nearby scenic ground-level images) and overprediction (e.g., a single nearby scenic image with a narrow field of view). On the other hand, the cross-view approach can be more robust against these types of mispredictions due to effectively averaging across many images (by marginalizing through the overhead imagery), not just those in the nearby area.

4.5 Conclusion

We explored the concept of natural beauty as it pertains to outdoor imagery. Using a dataset containing hundreds of thousands of ground-level images rated by humans, we showed it is possible to quantify scenicness, from both ground-level and overhead viewpoints. To our knowledge, this is the first time a combination of overhead and geotagged ground-level

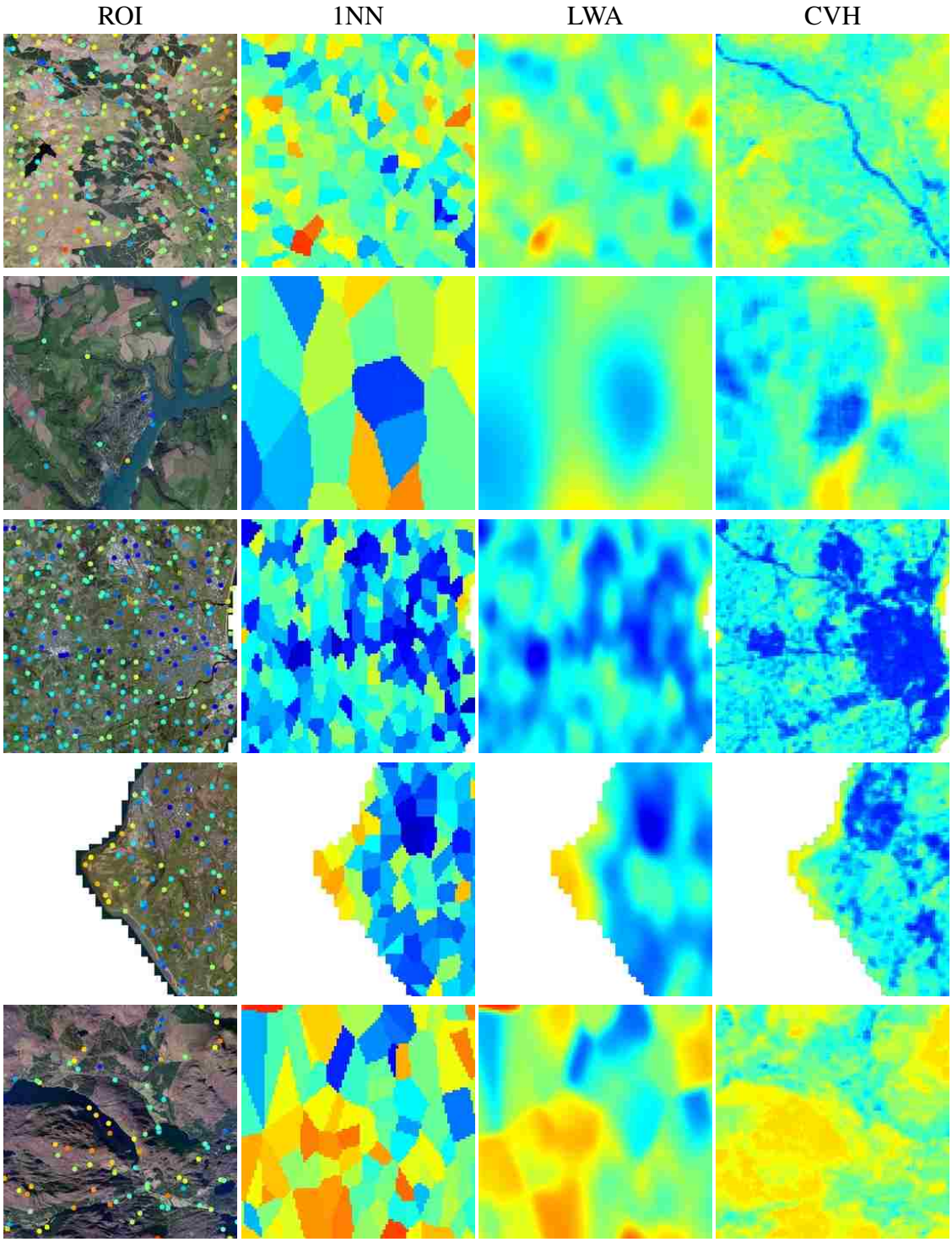


Figure 4.12: Scenicness maps. The first column shows an overhead image where dots correspond to geotagged ground-level imagery, colored by average scenicness rating (warmer colors correspond to more scenic images). The remaining columns show false-color images that reflect the average scenicness predicted by each method.

imagery has been used to map the scenicness of a region. The resulting maps are higher-resolution than those constructed by previous approaches and can be quickly computed. Such methods have significant practical importance to many areas, including: tourism, transportation routing, and environmental monitoring.

Chapter 5

A Unified Model for Near and Remote Sensing

5.1 Introduction

From predicting the weather to planning the future of our cities to recovering from natural disasters, accurately monitoring widespread areas of the Earth's surface is essential to many scientific fields and to society in general. These observations have traditionally been collected through remote sensing from satellites, aerial imaging, and distributed observing stations and sensors. These approaches can observe certain properties like land cover and land use accurately and at a high resolution, but unfortunately, not everything can be seen from overhead imagery. For example, Wang et al. [123] evaluate approaches for urban zoning and building height estimation from overhead imagery, and conclude that urban zoning segmentation "is an extremely hard task from aerial views," that building height estimation is "either too hard, or more sophisticated methods are needed," and that "utilizing ground imagery seems a logical first step."

More recently, the explosive popularity of geotagged social media has raised the possibility of using online user-generated content as a source of geospatial information, sometimes called *image-driven mapping* or *proximate sensing*. For example, online images from social network and photo sharing websites have been used to estimate land cover for large geographic regions [60, 150], to observe the state of the natural world by recreating maps of snowfall [122], and to quantify perception of urban environments [21]. Despite differing applications, these works all wish to estimate some unobservable *geospatial function*, and view each social media artifact (e.g., geotagged ground-level image) as an observation of this function at a particular geographic location.

Figure 5.1: We use overhead imagery and geotagged ground-level imagery as input to an end-to-end deep network that estimates the values of a geospatial function by performing fine-grained pixel-level labeling on the overhead image.

The typical approach [2, 136] is to (1) collect a large number of samples, (2) use an automated approach to estimate the value of the geospatial function for each sample, and (3) use some form of locally weighted averaging to interpolate the sparse samples into a dense, coherent estimate of the underlying geospatial function. This estimation is complicated by the fact that observations are noisy; state-of-the-art recognition algorithms are imperfect, some images are inherently confusing or ambiguous, and the observations are distributed sparsely and non-uniformly. This means that in order to estimate geospatial functions with reasonable accuracy, most techniques use a kernel with a large bandwidth to smooth out the noise, which yields coarse, low-resolution outputs. Despite this limitation, the proximate sensing approach can work well if ground-level imagery is plentiful, the property is easily estimated from the imagery, and the geospatial function is smoothly varying.

Figure 5.2: What type of building is shown in the overhead view (left)? Identifying and mapping building function is a challenging task that becomes considerably easier when taking into context nearby ground-level imagery (right).

In this chapter, we propose a novel neural network architecture that combines the strengths of these two approaches (Figure 5.1). Our approach uses deep convolutional neural networks (CNNs) to extract features from both overhead and ground-level imagery. For the ground-level images, we use kernel regression and density estimation to convert the sparsely distributed feature samples into a dense feature map spatially consistent with the overhead image. This differs from the proximate sensing approach, which uses kernel regression to directly estimate the geospatial function. Then, we fuse the ground-level feature map with a hidden layer of the overhead image CNN. To extend our methods to pixel-level labeling, we extract multi-scale features in the form of a hypercolumn and use a small neural network to estimate the geospatial function of interest. A novel element of our approach is the use of a spatially varying kernel that depends on features extracted from the overhead imagery.

Our network is trained end-to-end, so that all free parameters, including kernel bandwidths and low-level image features, are automatically tuned to minimize our loss function. In addition, our architecture is very general because it could be used with most state-of-the-art CNNs, and could be easily adapted to use any sparsely distributed media, including geotagged audio, video, and text (e.g., tweets). We evaluate our approach with a large real-world dataset, consisting of most of two major boroughs of New York City (Brooklyn and Queens), on estimating three challenging labels (building age, building function, and land use), all of which are notoriously challenging tasks in remote sensing (Figure 5.2). The results show that our technique for fusing overhead and ground-level imagery is more accurate than either the remote or proximate sensing approach alone, and that our automatically-estimated spatially-varying kernel improves accuracy compared to one that is uniform.

5.2 Related Work

Many recent studies have explored analyzing large-scale image collections as a means of characterizing properties of the physical world. A number of works have tried to estimate properties of weather from geotagged and time-stamped ground-level imagery. For example, Murdock et al. [79, 80] and Jacobs et al. [44] use webcams to infer cloud cover maps, Li et al. [63] use ground-level photos to estimate smog conditions, Glasner et al. [31] estimate temperature, Zhou et al. [149] and Lee et al. [59] estimate demographic properties, Fedorov et al. [24, 25] and Wang et al. [122] infer snow cover, Khosla et al. [51] and Porzi et al. [87] measure perceived crime levels, Leung and Newsam [60] estimate land use, and so on.

Many of these works' contribution is exploring a novel application, as opposed to proposing novel techniques. They mostly follow a very similar recipe in which standard recognition techniques are applied to individual images, and then spatial smoothing and other noise reduction techniques are used to create an estimate of the geospatial function across the world. Meanwhile, remote sensing has long used computer vision to estimate properties of the Earth from satellite images. Of course, overhead imaging is quite different from ground-level imaging, and so remote sensing techniques have largely been developed independently and in task-specific ways [95].

We know of relatively little work that has proposed general frameworks for estimating geospatial functions from imagery, or in integrating visual evidence from both ground-level and overhead image viewpoints. Tang et al. [114] show how location context can improve image classification, but they do not use overhead imagery and their goal is not to estimate geospatial functions. Luo et al. [70] use overhead imagery to give context for event recognition in ground-level photos by combining hand-crafted features for each modality. Xie et al. [137] use transfer learning to extract socioeconomic indicators from overhead imagery. Most similar is our work on mapping the subjective attribute of natural beauty [130] where we propose to use a multilayer perceptron to combine high-level semantic features. Recent work in image geolocalization has matched ground-level photos taken at unknown locations to georegistered overhead views [66, 67, 127, 129], but this goal is significantly different from inferring geospatial functions of the world.

Several recent works jointly reason about co-located ground-level and overhead image pairs. Máttyus et al. [73] perform joint inference over both monocular aerial and ground-level images from a stereo camera for fine-grained road segmentation, while Wegner et al. [124] detect and classify trees using features extracted from overhead and ground-level images. Ghouaiel and Lefèvre [29] transform ground-level panoramas to an overhead per-

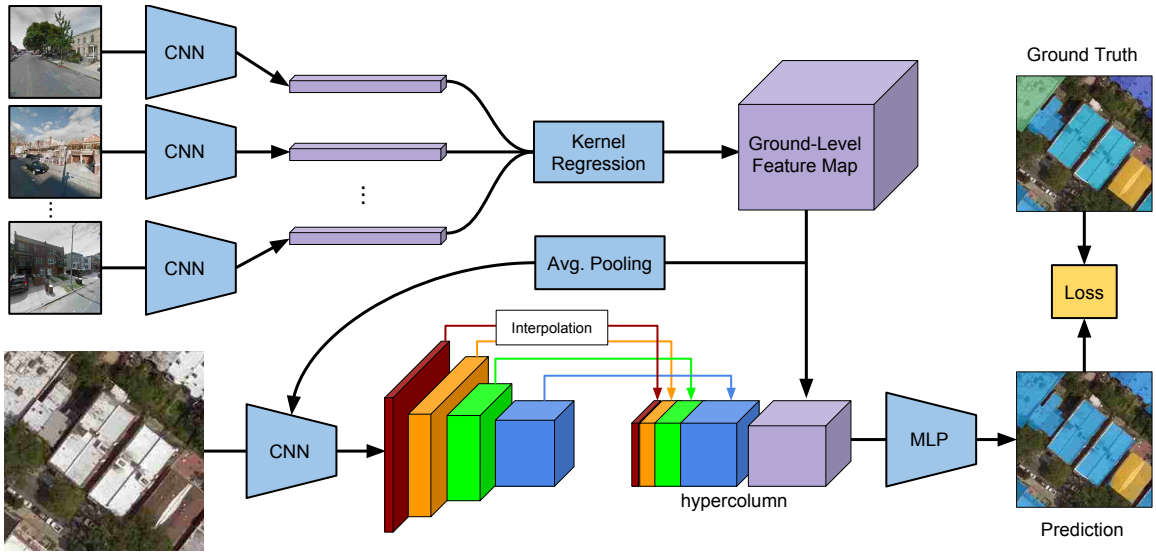


Figure 5.3: An overview of our network architecture.

spective for change detection. Zhai et al. [141] propose a transformation to extract meaningful features from overhead imagery.

In contrast with the above work, our goal is to produce a general framework for learning that can estimate any given geospatial function of the world. We integrate data from both ground-level imagery, which often contains visual evidence that is not visible from the air, and overhead imagery, which is typically much denser. We demonstrate how our models learn in an end-to-end way, avoiding the need for task-specific or hand-engineered features.

5.3 Problem Statement

We address the problem of estimating a spatially varying property of the physical world, which we model as an unobservable mathematical function that maps latitude-longitude coordinates to possible values of the property, $F : \mathbb{R}^2 \rightarrow \mathcal{Y}$. The range $\mathcal{Y}$ of this function depends on the attribute to be estimated, and might be categorical (e.g., a discrete set of elements for land use classification — golf course, residential, agricultural, etc.) or continuous (e.g., population density). We wish to estimate this function based on the available observable evidence, including data sampled both densely (such as overhead imagery) and sparsely (such as geotagged ground-level images). From a probabilistic perspective, we can think of our task as learning a conditional probability distribution $P(F(l) = y | S_l, \mathbf{G}(l))$, where l is a latitude-longitude coordinate, S_l is an overhead image centered at that location, and $\mathbf{G}(l)$ is a set of nearby ground-level images.

5.4 Network Architecture

We propose a novel convolutional neural network (CNN) that fuses high-resolution overhead imagery and nearby ground-level imagery to estimate the value of a geospatial function at a target location. While we focus on images, our overall architecture could be used with many sources of dense and sparse data. Our network can be trained in an end-to-end manner, which enables it to learn to optimally extract features from both the dense and sparse data sources.

5.4.1 Architecture Overview

The overall architecture of our network (Figure 5.3) consists of three main components, the details of which we describe in the next several sections: (1) constructing a spatially dense feature map using features extracted from the ground-level images (Section 5.4.2), (2) extracting features from the overhead image, incorporating the ground-level image feature map (Section 5.4.3), and (3) predicting the geospatial function value based on a hypercolumn of features (Section 5.4.4). A novel element of our proposed approach is the use of an adaptive, spatially varying interpolation method for constructing the ground-level image feature map based on features extracted from the overhead image (Section 5.4.5).

5.4.2 Ground-Level Feature Map Construction

The goal of this component is to convert a sparsely sampled set of ground-level images into a dense feature map. For a given geographic location l, let $\mathbf{G}(l) = \{(G_i, l_i)\}$ be a set of N elements corresponding to the closest ground-level images, where each (G_i, l_i) is an image and its respective geographic location. We use a CNN to extract features, $f_g(G_i)$, from each image and interpolate using Nadaraya–Watson kernel regression,

$$f_G(l) = \frac{\sum w_i f_g(G_i)}{\sum w_i}, \tag{5.1}$$

where $w_i = exp(-d(l, l_i; \Sigma)^2)$ is a Gaussian kernel function where a diagonal covariance matrix Σ controls the kernel bandwidth and $d(l, l_i; \Sigma)$ is the Mahalanobis distance from l to l_i. We perform this interpolation for every pixel location in the overhead image. The result is a feature map of size $H \times W \times m$, where H and W are the height and width of the overhead image in pixels, and m is the output dimensionality of our ground-level image CNN.

The diagonal elements of the covariance matrix are represented by a pair of trainable weights, which pass through a *softplus* function (i.e., $f(x) = ln(1 + e^x)$) to ensure they are

positive. Here, the value of Σ does not depend on geographic location, a strategy we call *uniform*. In Section 5.4.5, we propose an approach in which Σ is spatially varying.

In our experiments, the ground-level images, $\mathbf{G}(l)$, are actually geo-oriented street-level panoramas. To form a feature representation for each panorama, G_i, we first extract perspective images in the cardinal directions, resulting in four ground-level images per location. We replicate the ground-level image CNN, $f_g(G_i)$, four times, feed each image through separately, and concatenate the individual outputs. We then add a final 1×1 convolution to reduce the feature dimensionality. For our experiments, we use the VGG-16 architecture [106], initialized with weights for Place categorization [148] ($m = 205$, layer name *fc8*). The result is an 820 dimensional feature vector for each location, which is further reduced to 50 dimensions.

It is possible that the nearest ground-level image may be far away, which could lead to later processing stages incorrectly interpreting the feature map. To overcome this, we concatenate a kernel density estimate, using the kernel defined in equation (5.1), of the ground-level image locations to the ground-level image feature map. The result is an $H \times W \times 51$ feature map that captures appearance and distributional information of the ground-level images.

5.4.3 Overhead Feature Map Construction

This section describes the CNN we use to extract features from the overhead image and how we integrate the ground-level feature map. The CNN is based on the VGG-16 architecture [106], which has 13 convolutional layers, each using 3×3 convolutions, and three fully connected layers. We only use the convolutional layers, typically referred to as conv-$\{1_{1-2}, 2_{1-2}, 3_{1-3}, 4_{1-3}, 5_{1-3}\}$. In addition, we reduce the dimensionality of the feature maps that are output by each layer. These layers have output dimensionality of $\{32, 64, 128, 256, 512\}$ channels, respectively. Each intermediate layer uses a leaky ReLU activation function ($\alpha = 0.2$).

To fuse the ground-level feature map with the overhead imagery, we apply average pooling with a kernel size of 6×6 and a stride of 2. Given an input overhead image with $H = W = 256$, this reduces the ground-level feature map to $32 \times 32 \times 51$. We then concatenate it, in the channels dimension, with the overhead image feature map at the seventh convolutional layer, 3_3. The input to convolutional layer 4_1 is then $32 \times 32 \times 179$. We experimented with including the ground-level feature map earlier and later in the network and found this to be a good tradeoff between computational cost and expressiveness.

5.4.4 Geospatial Function Prediction

Given an overhead image, S_l, we use the ground-level and overhead feature maps defined above as input to the final component of our system to estimate the value of the geospatial function, $F(l(p)) \in 1 \dots K$, where $l(p)$ is the location of a pixel p. This pixel might be the center of the image for the image classification setting or any arbitrary pixel in the pixel-level labeling setting. To accomplish this we adapt ideas from the PixelNet architecture [6], due to its strong performance and ability to train using sparse inputs. However, our approach for incorporating sparsely distributed inputs could be adapted to other semantic labeling architectures.

We first resize each feature map to be $H \times W$ using bilinear interpolation. We then extract a *hypercolumn* [33] consisting of a set of features centered around p, $h_p(S) = [c_1(S,p), c_2(S,p), \dots, ...c_M(S,p)]$, where c_i is the feature map of the i-th layer. For this work, we extract hypercolumn features from conv-$\{1_2, 2_2, 3_3, 4_3, 5_3\}$ and the ground-level feature map. The resulting hypercolumn feature has length 1,043. Note that resizing all intermediate feature maps to be the size of the image is quite memory intensive. Following Bansal et al. [6], we subsample pixels during training to increase the number (and therefore diversity) of images per mini-batch. At testing time, we can either compute the hypercolumn for all pixels to create a dense semantic labeling or a subset to label particular locations.

This hypercolumn feature is then passed to a small multilayer perceptron (MLP) that provides the estimate of the geospatial function. The MLP has three layers of size 512, 512, and K (the task dependent number of outputs). Each intermediate layer uses a leaky ReLU activation function.

5.4.5 Adaptive Kernel Bandwidth Estimation

In addition to the *uniform* kernel described above for forming the ground-level image feature map (Section 5.4.2), we propose an *adaptive* strategy that predicts the optimal kernel bandwidth parameters for each location in the feature map. We estimate these bandwidth parameters using a CNN applied to the overhead image. This network shares the first three groups of convolutional layers, conv-$\{1_1, \dots, 3_3\}$, with the overhead image CNN defined in Section 5.4.3. The output of these convolutions is passed to a sequence of three convolutional transpose layers, each with filter size 3×3 and a stride of 2. These layers have output dimensionality of 32, 16, and 2, respectively. The final layer has an output size of $H \times W \times 2$, which represents the diagonal entries of the kernel bandwidth matrix, Σ, for each pixel location. Similar to the *uniform* approach, we apply a *softplus* activation

on the output (initialized with a small constant bias) to ensure positive kernel bandwidth. When using the *adaptive* strategy, these bandwidth parameters are used to construct the ground-level feature map ($H \times W \times 51$).

5.5 Experiments

We evaluated the performance of our approach on a challenging real-world dataset, which includes overhead imagery, ground-level imagery, and several fine-grained pixel-level labels. We proposed two variants of our approach: *unified (uniform)*, which uses a single kernel bandwidth for the entire region, and *unified (adaptive)*, which uses a location-dependent kernel that is conditioned on the overhead image.

5.5.1 Baseline Methods

In order to evaluate the proposed macro-architecture, we use several baseline methods that share many low-level components with our proposed methods.

- *random* represents random sampling from the prior distribution of the training dataset.
- *remote* represents the traditional remote sensing approach, in which only overhead imagery is used. We use the *unified (uniform)* architecture, but do not incorporate the ground-level feature map in the overhead image CNN or the hypercolumn.
- *proximate* represents the proximate sensing approach in which only ground-level imagery is used. We start from the *unified (uniform)* architecture but only include the ground-level image feature map (minus the kernel density estimate) in the hypercolumn.
- *grid* is similar to the *proximate* method. Starting from *unified (uniform)*, we omit all layers from the overhead image CNN prior to concatenating in the ground-level feature map from the hypercolumn. The motivation for this method is that the additional convolutional layers are able to capture spatial patterns which the final MLP cannot, because it operates on individual hypercolumns.

5.5.2 Implementation Details

All methods were implemented using Google's TensorFlow framework [1] and optimized using Adam [52] with default training parameters, except for an initial learning rate of 10^{-3}

Figure 5.4: Sample overhead imagery and nearby street-level panoramas included in the Brooklyn and Queens dataset.

(decreasing by 0.5 every 7,500 mini-batches) and weight decay of 5×10^{-4}. During training, we randomly sampled 2,000 pixels per image per mini-batch. The ground-level CNNs have shared weights. All other network weights were randomly initialized and allowed to vary freely. We applied batch normalization [38] (decay $= 0.99$) in all convolutional and fully connected layers (except for output layers). For our experiments, we minimize a cross-entropy loss function and consider the nearest 20 street-level panoramas. Each network was trained for 25 epochs with a batch size of 32 on an NVIDIA Tesla P100.

5.5.3 Brooklyn and Queens Dataset

We introduce a new dataset containing ground-level and overhead images from Brooklyn and Queens, two boroughs of New York City (Figure 5.4). It consists of non-overlapping overhead images downloaded from Bing Maps (zoom level 19, approximately $30cm$ per pixel) and street-level panoramas from Google Street View. From Brooklyn, we collected imagery for the entirety of King's County. This resulted in 73,921 overhead images and 139,327 panoramas. A significant number (30,316) of the overhead images are over water; we discard these and only consider those which contain buildings. We hold out 4,361 overhead images for testing. For Queens, we selected a held out region solely for evaluation and used the same process to collect imagery. This resulted in a dataset with 10,044 overhead images and 38,603 panoramas.

Using data made publicly available by NYC Open Data,[1] we constructed a per-pixel labeling of each overhead image for the following set of labels.

Building Function. We used 206 building classes, as outlined by the New York City Department of City Planning (NYCDCP) in the Primary Land Use Tax Lot Output (PLUTO) dataset, to categorize each building in a given overhead image. PLUTO contains detailed geographic data at the tax lot level (property boundary) for every piece of land in New York City. Example labels include: Multi-Story Department Stores, Funeral Home, and Church.

[1] `https://data.cityofnewyork.us/`

To this set we add two classes, background (non-building, such as roads and water) and unknown, as there are several thousand unlabeled tax lots. To form our final labeling, we intersected the tax lot data with building footprints obtained from the NYC Planimetric Database. For reference, there are approximately 331,000 buildings in Brooklyn.

Land Use. From PLUTO, we generated a per-pixel label image with each contained tax lot labeled according to its primary land use category. The land use categories were specified by the New York City Department of City Planning. In total, there are 11 land use categories. Example land use categories include: One and Two Family Buildings, Commercial and Office Buildings, and Open Space and Outdoor Recreation. Similar to building function, we add two classes, background (e.g., roads) and unknown.

Building Age. Again using PLUTO in conjunction with the NYC Planimetric Database, we generated a per-pixel label image with each building labeled according to the year that construction of the building was completed. Brooklyn and Queens have a lengthy history, with the oldest building on record dating to the mid-1600s. We quantize time by decades, with a bin for all buildings constructed before 1900. This resulted in 13 bins, to which we added a bin for background (non-building), as well as unknown for a small number of buildings without a documented construction year.

5.5.4 Semantic Segmentation

We report results using pixel accuracy and region intersection over union averaged over classes (mIOU), two standard metrics for the semantic segmentation task. In both cases, higher is better. When computing these metrics, we ignore any ground-truth pixel labeled as unknown. In addition, for the tasks of building function and age estimation, we ignore background pixels.

Classifying Land Use. We consider the task of identifying a parcel of land's primary land use. This task is considered especially challenging from an overhead only perspective, with recent work simplifying the task by considering only three classes [123]. We report top-1 accuracy for land use classification using the Brooklyn test set in Table 5.1 and on Queens in Table 5.3. Similarly we report mIOU for Brooklyn and Queens in Table 5.2 and Table 5.4, respectively. Our results support the notion that this task is extremely difficult. However, our approach, *unified (adaptive)*, is significantly better than all baselines, including an overhead image only approach (*remote*). Qualitative results for this task are shown in Figure 5.5.

Table 5.1: Brooklyn evaluation results (top-1 accuracy).

	Age	Function	Land Use
random	6.82%	0.49%	8.55%
proximate	35.90%	27.14%	44.66%
grid	38.68%	33.84%	71.64%
remote	37.18%	34.64%	69.63%
unified (uniform)	**44.08%**	43.88%	76.14%
unified (adaptive)	43.85%	**44.88%**	**77.40%**

Table 5.2: Brooklyn evaluation results (mIOU).

	Age	Function	Land Use
random	2.76%	0.11%	3.21%
proximate	11.77%	5.46%	18.04%
grid	16.98%	9.37%	37.76%
remote	15.11%	4.67%	31.70%
unified (uniform)	20.88%	13.66%	43.53%
unified (adaptive)	**23.13%**	**14.59%**	**45.54%**

Table 5.3: Queens evaluation results (top-1 accuracy).

	Age	Function	Land Use
random	6.80%	0.49%	8.41%
proximate	25.27%	22.50%	47.40%
grid	27.47%	26.62%	67.51%
remote	26.06%	29.85%	69.27%
unified (uniform)	29.68%	33.64%	68.08%
unified (adaptive)	**29.76%**	**34.13%**	**70.55%**

Table 5.4: Queens evaluation results (mIOU).

	Age	Function	Land Use
random	2.58%	0.09%	3.05%
proximate	5.08%	1.57%	15.04%
grid	7.31%	2.30%	28.02%
remote	7.78%	2.67%	28.46%
unified (uniform)	8.95%	3.71%	31.03%
unified (adaptive)	**9.53%**	**3.73%**	**33.48%**

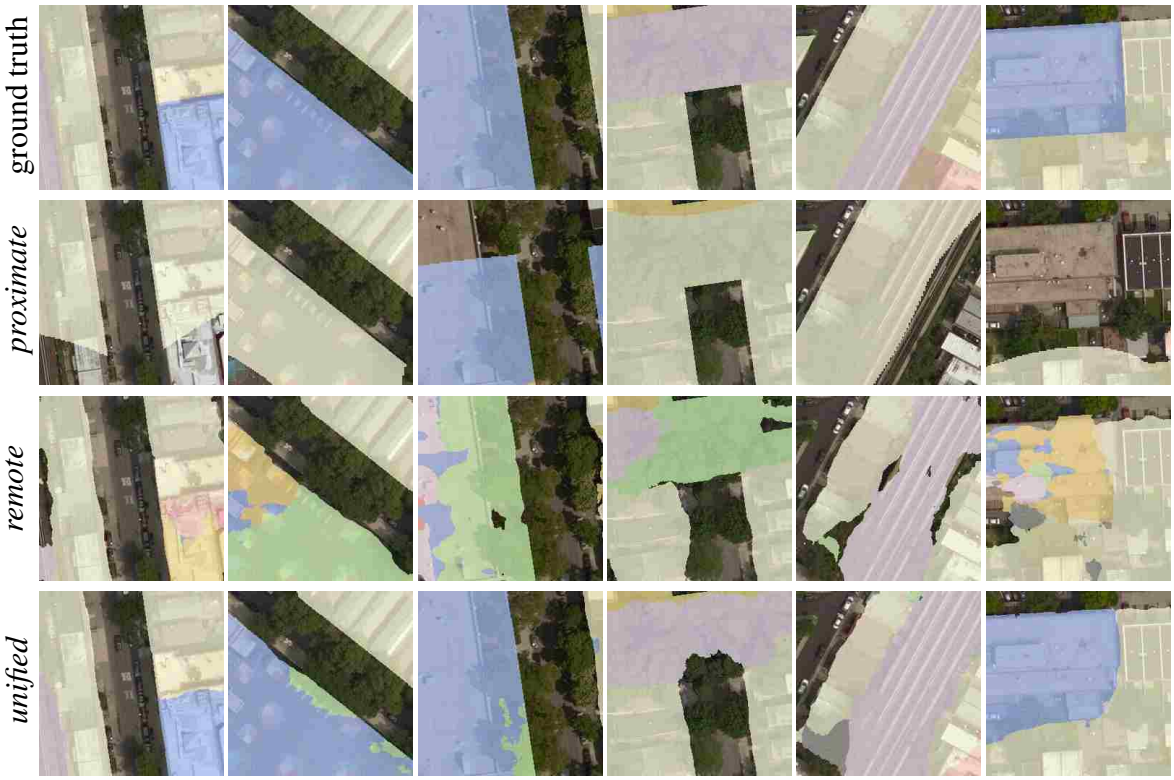


Figure 5.5: Sample results for classifying land use: (top–bottom) ground truth, *proximate*, *remote*, and *unified (adaptive)*.

Identifying Building Function. We consider the task of making a functional map of buildings. To our knowledge, our work is the first to explore this. For example, in Figure 5.2, it becomes considerably easier to identify that the building in the overhead image is a fire station when shown two nearby ground-level images. We report performance metrics for this task in Table 5.1 and Table 5.3 for accuracy, and Table 5.2 and Table 5.4 for mIOU. Qualitative results are shown in Figure 5.6. Given the challenging nature of this task, we visualize results as a top-k image, where each pixel is colored from green (best) to red, by the rank of the correct class in the posterior distribution. Our approach produces labelings much more consistent with the ground truth.

Estimating Building Age. Finally, we consider the task of estimating the year a building was constructed. Intuitively, this is an extremely difficult task from an overhead image only viewpoint, but is also non-trivial from a ground-level view. We report accuracy and mIOU metrics for this experiment in Table 5.1 and Table 5.2 for the Brooklyn region and Table 5.3 and in Table 5.4 for Queens. Our approach significantly outperforms the baselines. Example qualitative results are shown in Figure 5.7.

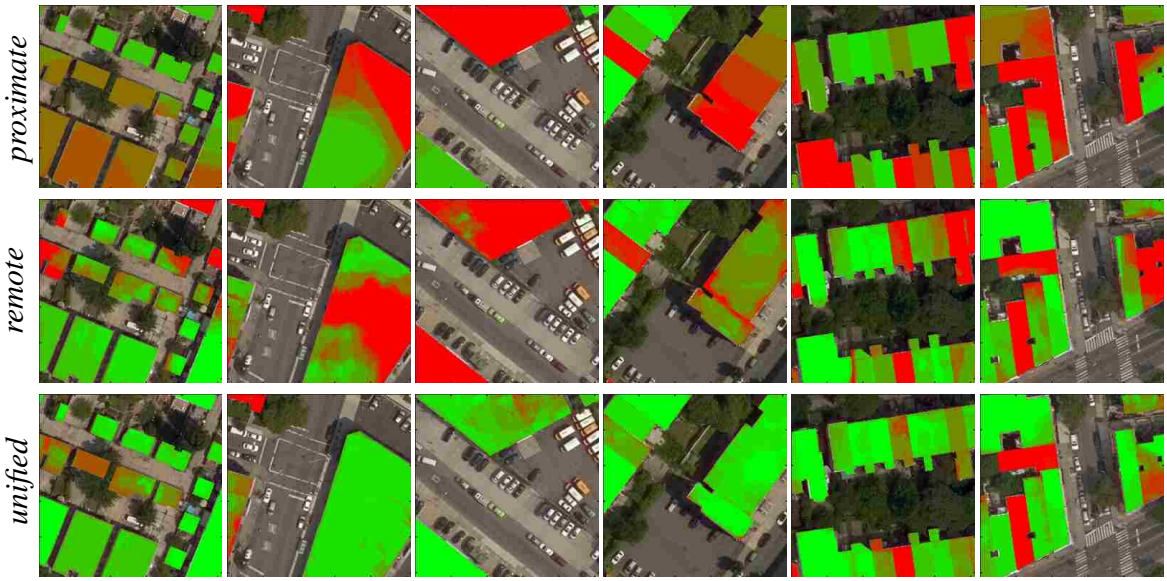


Figure 5.6: Sample results for identifying building function. From top to bottom, we visualize top-k images for the *proximate*, *remote*, and *unified (adaptive)* methods, respectively. Each pixel is color coded on a scale from green to red by the rank of the correct class in the posterior distribution, where bright green is the best (rank one).

Figure 5.7: Sample results for estimating building age: (top) ground truth and (bottom) *unified (adaptive)*.

5.5.5 Does Known Orientation Help?

In the evaluation above, we constructed the ground-level feature map (Section 5.4.2) using features from geo-oriented panorama cutouts. The cutout images were extracted in the cardinal directions and their features stacked in a fixed order. To better understand the value of the ground-level feature map, we investigated how knowing the orientation of the ground-level images affects accuracy. We repeated the land use classification experiment on Brooklyn using our *uniform (adaptive)* approach (retraining the network), but randomly circular-shifted the set of images prior to feature extraction. Note that orientation is not completely random, because doing so would have required regenerating cutouts. We observe a significant performance drop from 77.40% to 72.61% in top-1 accuracy, about 3% higher than using the overhead image only method. This experiment shows that knowing the orientation of the ground-level images is critical for achieving the best performance, but that including the ground-level images without knowing the orientation can still be useful.

5.6 Conclusion

We proposed a novel neural network architecture for estimating geospatial functions and evaluated it in the context of fine-grained understanding of an urban area. Our network fuses overhead and ground-level images and gives more accurate predictions than if either modality had been used in isolation. Specifically, our approach is better at resolving spatial boundaries than if only ground-level images were used and is better at estimating features that are difficult to determine from a purely overhead perspective. A key feature of our architecture is that it is end-to-end trainable, meaning that it can learn to extract the optimal features, for any appropriate loss function, from the raw pixels of all images, as well as parameters used to control the fusion process. While we demonstrated its use with ground-level images, our architecture is general and could be used with a wide variety of sparsely distributed measurements, including geotagged tweets, video, and audio.

Chapter 6

Discussion

Our thesis focused on combining ground-level and overhead image viewpoints to improve geospatial modeling. Specifically, we showed how to leverage overhead imagery, in addition to ground-level imagery, for tasks in localization, mapping, and understanding. Our work encompassed three primary research areas – learning a joint feature representation between ground-level and overhead imagery, inferring labels for overhead imagery from nearby ground-level images, and fusing ground-level imagery with overhead imagery.

In Chapter 2 we analyzed the discriminative ability of deep image representations, extracted from convolutional networks previously trained on traditional vision tasks, for several problems in geospatial image analysis. In addition to showing that deep image representations capture location-related information for ground-level imagery, our results show that such representations are useful for interpreting and understanding overhead images, despite the original networks being trained on images from a ground-level perspective. Next, we analyzed the co-occurrence of feature activations for ground-level and overhead images captured at the same location, finding that the representations are positively correlated. Motivated by these findings, we demonstrated how overhead imagery could be leveraged to improve geospatial modeling based on ground-level imagery alone. Finally, we highlighted potential applications in image-based search and cross-view image matching. Our results suggest the potential of building deep-learning based models that are directly targeted at problems of localization and location-related feature extraction from ground-level and overhead imagery.

In Chapter 3 we focused on the image geolocalization task, and proposed a method for learning a joint feature representation between ground-level and overhead images that enables fine-grained geolocalization results at varying spatial scales. The underlying idea was to learn a mapping between ground-level and overhead image viewpoints, such that a ground-level query image can be directly matched against an overhead image reference

database. We introduced a cross-view training approach that takes advantage of existing state-of-the-art feature representations for ground-level images in order to extract features for overhead imagery. Specifically, we used pre-existing CNNs for extracting ground-level image features and then learned to predict these features from overhead images of the same location. To support these efforts, we introduced a large dataset containing geotagged ground-level images and multi-scale overhead imagery. Using this dataset, we proposed single and multi-scale networks for extracting overhead image features, obtaining state-of-the-art results for cross-view localization on two benchmark datasets.

In Chapter 4 we used unlabeled overhead imagery to improve image-driven mapping. The attribute we sought to quantify was the scenicness, or natural beauty of an outdoor scene. We began by proposing a method to estimate scenicness from a single ground-level image that accounts for variance in the ratings and human perception of scenicness. We demonstrated quantitatively that our approach better captures human uncertainty compared to baseline methods. Then, we extended our approach to consider scenicness as a property of geographic locations. To deal with the sparsity of ground-level images, we applied a cross-view training approach to learn how to predict scenicness from unlabeled overhead imagery. To accomplish this, we infer target labels from nearby ground-level imagery. Specifically, we predict the scenicness of a ground-level image captured at the same location. Finally, we proposed a hybrid approach which combines ground-level and overhead imagery to estimate the scenicness of a query location. This approach considers the corresponding overhead image and a set of the closest ground-level images with their distances to the query location. Our results demonstrated that by including overhead imagery we are able to construct a significantly more accurate map than purely interpolating scenicness estimates obtained from ground-level images alone.

In Chapter 5 we proposed a general framework for fusing visual information from ground-level and overhead imagery and demonstrated its application to fine-grained urban understanding. Our approach combines the strengths of proximate and remote sensing in the form of an end-to-end trainable neural network, which uses kernel regression and density estimation to convert features extracted from the ground-level images into a dense feature map. This ground-level feature map has the same spatial coverage as the overhead image, and is fused internally at a hidden layer of the overhead image network. A key element of our approach is a spatially-varying kernel, conditioned on the overhead image, that improves accuracy compared to a uniform kernel. The final output of our network is a dense estimate of the geospatial function in the form of a pixel-level labeling of the overhead image. To evaluate our approach, we created a large dataset of overhead and ground-level images from a major urban area with three sets of labels: land use, building

function, and building age. Our results showed that our approach is more accurate, for all tasks, than if either modality had been used in isolation. A unique feature of our architecture is that it can be trained end-to-end, such that it can learn to extract the optimal features, for any appropriate loss function, from the raw images of each viewpoint, and the parameters used to control the fusion process. Our experiments demonstrated the application of our framework with ground-level images, but the architecture is general and could be used with a wide variety of sparsely distributed measurements, including geotagged tweets, video, and audio.

As part of our research efforts we constructed several large datasets, in some cases containing millions of images. A common feature of these datasets is that they often contain pairs of ground-level and overhead images captured at the same location. For the task of cross-view image geolocalization, we introduced several large cross-view datasets to support training models and benchmarking performance. Similarly, we introduced a cross-view dataset to support image-driven mapping of the subjective property of natural beauty. Finally, we introduced a dataset to support fine-grained understanding of an urban area. In all cases these datasets have been made available to the community, along with trained models and example code. It is our belief that the availability of large datasets such as these will stimulate research in localization, mapping, and understanding.

This thesis proposed several methods for jointly understanding ground-level and overhead imagery. Now widely available, overhead imagery offers a potential alternative to augment methods which rely solely on sparsely available ground-level images. Our work culminated in a general framework for estimating geospatial functions that integrates visual evidence from multiple viewpoints. There are several possible future research directions for extending this work, including: applying our framework to other tasks and sources of sparse measurements, integrating multi-scale overhead imagery, incorporating ground-level image attention, and exploring other architectures. Overall, we hope that our work will inspire the vision community to leverage overhead imagery as an additional source of context when searching for solutions to problems in geospatial modeling.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 68

[2] Sean M Arietta, Alexei A Efros, Ravi Ramamoorthi, and Maneesh Agrawala. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2624–2633, 2014. 52, 61

[3] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Transactions on Graphics*, 26(3):10, 2007. 52

[4] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *European Conference on Computer Vision*, 2012. 24, 29, 30

[5] Ruzena Bajcsy and Mohamad Tavakoli. Computer recognition of roads from satellite pictures. *IEEE Transactions on Systems, Man and Cybernetics*, 6(9):623–637, 1976. 4

[6] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta Ramanan, et al. Pixelnet: Representation of the pixels, by the pixels, and for the pixels. *arXiv preprint arXiv:1702.06506*, 2017. 67

[7] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with aerial image databases. In *ACM International Conference on Multimedia*, 2011. 30

[8] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*, 2014. 3

[9] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for scientific computing conference*, volume 4, page 3, 2010. 13

[10] Margaret Ann Boden. *Mind as machine: A history of cognitive science*. Clarendon Press, 2006. 1

[11] Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. We are humor beings: Understanding and predicting visual humor. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 42

[12] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014. 35

[13] David M Chen, Georges Baatz, K Koser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvanainen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 24, 30

[14] Fabio Cozman and Eric Krotkov. Robot localization using a computer vision sextant. In *International Conference on Robotics and Automation*, 1995. 29

[15] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *International World Wide Web Conference*, 2009. 3, 24, 28, 30

[16] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006. 31

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 13

[18] Arturo Deza and Devi Parikh. Understanding image virality. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 42, 43

[19] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 31(4), 2012. 9, 29, 30

[20] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, 2014. 9

[21] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*, 2016. 60

[22] Anne Eisenberg. Microsatellites: what big eyes they have. *The New York Times*, 2013. 4

[23] Quan Fang, Jitao Sang, and Changsheng Xu. Discovering geo-informative attributes for location recognition and exploration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1s):19, 2014. 9, 30

[24] Roman Fedorov, Piero Fraternali, Chiara Pasini, and Marco Tagliasacchi. Snowwatch: Snow monitoring through acquisition and analysis of user-generated content. In *IEEE International Conference on Multimedia and Expo*, 2015. 63

[25] Roman Fedorov, Piero Fraternali, and Marco Tagliasacchi. Snow phenomena modeling through online public media. In *IEEE International Conference on Image Processing*, 2014. 63

[26] Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*, 2014. 10

[27] Jacob Gardner, Matt Kusner, Kilian Q. Weinberger, John Cunningham, and Zhixiang Xu. Bayesian optimization with inequality constraints. In *International Conference on Machine Learning*, 2014. 52

[28] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[29] Nehla Ghouaiel and Sébastien Lefèvre. Coupling ground-level panoramas and aerial imagery for change detection. *Geo-spatial Information Science*, 19(3):222–232, 2016. 63

[30] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 9, 30

[31] Daniel Glasner, Pascal Fua, Todd Zickler, and Lihi Zelnik-Manor. Hot or not: Exploring correlations between appearance and temperature. In *IEEE International Conference on Computer Vision*, 2015. 63

[32] Jorge Guillén, Antonio García-Olivares, Elena Ojeda, Andres Osorio, Oscar Chic, and Raul González. Long-term quantification of beach users using video monitoring. *Journal of Coastal Research*, 2008. 2

[33] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 67

[34] James Hays and Alexei A Efros. Im2gps: Estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 24, 29, 30

[35] Rob Holman, John Stanley, and Tuba Ozkan-Haller. Applying video sensor networks to nearshore environment monitoring. *Pervasive Computing, IEEE*, 2003. 2

[36] Collin Homer, Jon Dewitz, Limin Yang, Suming Jin, Patrick Danielson, George Xian, John Coulston, Nathaniel Herold, James Wickham, and Kevin Megown. Completion of the 2011 national land cover database for the conterminous united states-representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing*, 81(5):345–354, 2015. 19

[37] JM Idelsohn. A learning system for terrain recognition. *Pattern Recognition*, 2(4):293–301, 1970. 4

[38] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015. 69

[39] Mohammad T Islam, Scott Workman, Hui Wu, Nathan Jacobs, and Richard Souvenir. Exploring the geo-dependence of human face appearance. In *IEEE Winter Conference on Applications of Computer Vision*, 2014. 9, 30

[40] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 42

[41] Nathan Jacobs, Kylia Miskell, and Robert Pless. Webcam geo-localization using aggregate light levels. In *IEEE Workshop on Applications of Computer Vision*, 2011. 24, 29

[42] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Toward fully automatic geo-location and geo-orientation of static outdoor cameras. In *IEEE Workshop on Applications of Computer Vision*, 2008. 24, 29

[43] Nathan Jacobs, Scott Satkin, Nathaniel Roman, Richard Speyer, and Robert Pless. Geolocating static cameras. In *IEEE International Conference on Computer Vision*, 2007. 24, 30

[44] Nathan Jacobs, Scott Workman, and Richard Souvenir. Cloudmaps from static ground-view video. *Image and Vision Computing*, 52:154–166, 2016. 63

[45] Mainak Jas and Devi Parikh. Image specificity. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 42

[46] John R Jensen and Dave C Cowen. Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric engineering and remote sensing*, 65:611–622, 1999. 4

[47] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 13, 35, 36, 49

[48] Imran N Junejo and Hassan Foroosh. Gps coordinates estimation and camera calibration from solar shadows. *Computer Vision and Image Understanding*, 2010. 29

[49] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *British Machine Vision Conference*, 2014. 35, 43

[50] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 43

[51] Aditya Khosla, Byoungkwon An, Jasmine J Lim, and Antonio Torralba. Looking beyond the visible scene. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 63

[52] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 68

[53] Jan Knopp, Josef Sivic, and Tomas Pajdla. Avoiding confusing features in place recognition. In *European Conference on Computer Vision*, 2010. 29

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 12, 28, 30, 35, 36

[55] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (SIGGRAPH)*, 33(4), 2014. 9, 43

[56] Jean-François Lalonde, Srinivasa G Narasimhan, and Alexei A Efros. What do the sun and the sky tell us about the camera? *International Journal of Computer Vision*, 2010. 29

[57] Judith H Langlois, Lisa Kalakanis, Adam J Rubenstein, Andrea Larson, Monica Hallam, and Monica Smoot. Maxims or myths of beauty? a meta-analytic and theoretical review. *Psychological bulletin*, 126(3):390, 2000. 42

[58] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989. 12

[59] Stefan Lee, Haipeng Zhang, and David Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2015. 9, 10, 30, 63

[60] Daniel Leung and Shawn Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 3, 60, 63

[61] Y. Li, D.J. Crandall, and D.P. Huttenlocher. Landmark classification in large-scale image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 29

[62] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Mid-level deep pattern mining. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 13

[63] Yuncheng Li, Jifei Huang, and Jiebo Luo. Using user generated online photos to estimate and monitor air pollution in major cities. In *ACM International Conference on Internet Multimedia Computing and Service*, 2015. 63

[64] Liang Liang, Mark D Schwartz, and Songlin Fei. Validating satellite phenology through intensive ground observation and landscape scaling in a mixed seasonal forest. *Remote Sensing of Environment*, 115(1):143–157, 2011. 3

[65] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations*, 2014. 35

[66] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. vii, 13, 24, 25, 28, 30, 34, 37, 55, 63

[67] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 24, 30, 55, 63

[68] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rapid: Rating pictorial aesthetics using deep learning. In *ACM International Conference on Multimedia*, 2014. 43

[69] Xin Lu, Changhu Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. Photo2trip: generating travel routes from geo-tagged photos for trip planning. In *ACM International Conference on Multimedia*, 2010. 42

[70] Jiebo Luo, Jie Yu, Dhiraj Joshi, and Wei Hao. Event recognition: viewing the world with a third eye. In *ACM International Conference on Multimedia*, 2008. 30, 55, 63

[71] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, 2008. 43

[72] Larry Matthies, Mark Maimone, Andrew Johnson, Yang Cheng, Reg Willson, Carlos Villalpando, Steve Goldberg, Andres Huertas, Andrew Stein, and Anelia Angelova. Computer vision on mars. *International Journal of Computer Vision*, 75(1):67–92, 2007. 2

[73] Gellért Máttyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 63

[74] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 1943. 11

[75] David A McGranahan. Natural amenities drive rural population change. Technical report, United States Department of Agriculture, Economic Research Service, 1999. 42

[76] Marvin Minsky and Papert Seymour. *Perceptrons: an introduction to computational geometry*. MIT press, 1969. 11

[77] Jeffrey T Morisette, Andrew D Richardson, Alan K Knapp, Jeremy I Fisher, Eric A Graham, John Abatzoglou, Bruce E Wilson, David D Breshears, Geoffrey M Henebry, Jonathan M Hanes, et al. Tracking the rhythm of the seasons in the face of global change: phenological research in the 21st century. *Frontiers in Ecology and the Environment*, 2008. 2

[78] David J Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4):358–371, 2013. 4

[79] Calvin Murdock, Nathan Jacobs, and Robert Pless. Webcam2satellite: Estimating cloud maps from webcam imagery. In *IEEE Workshop on Applications of Computer Vision*, 2013. 63

[80] Calvin Murdock, Nathan Jacobs, and Robert Pless. Building dynamic cloud maps from the ground up. In *IEEE International Conference on Computer Vision*, 2015. 63

[81] M-E Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 35

[82] Aude Oliva. Gist of the scene. In *Neurobiology of attention*, pages 251–256. Elsevier, 2005. 1

[83] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 15

[84] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. ix, 3, 9, 46, 47

[85] Otávio AB Penatti, Keiller Nogueira, and Jefersson A dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *IEEE/ISPRS Workshop: Looking From Above: When Earth Observation Meets Vision*, 2015. 10

[86] Harold Everett Porter. *Aerial observation: the airplane observer, the balloon observer, and the army corps pilot*. Harper & Brothers, 1921. 4

[87] Lorenzo Porzi, Samuel Rota Bulò, Bruno Lepri, and Elisa Ricci. Predicting and understanding urban perception with convolutional neural networks. In *ACM International Conference on Multimedia*, 2015. 43, 52, 63

[88] Daniele Quercia, Luca Maria Aiello, Rossano Schifanella, and Adam Davies. The digital life of walkable streets. In *International World Wide Web Conference*, 2015. 52

[89] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2014. 9, 30

[90] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *ACM Conference on Hypertext and Social Media*, 2014. 52

[91] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014. 10, 13, 30

[92] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 3

[93] Andrew D Richardson, Bobby H Braswell, David Y Hollinger, Julian P Jenkins, and Scott V Ollinger. Near-surface remote sensing of spatial and temporal variation in canopy phenology. *Ecological Applications*, 2009. 2

[94] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 1958. 11

[95] Offer Rozenstein and Arnon Karnieli. Comparison of methods for land-use classification incorporating remote sensing and gis inputs. *Applied Geography*, 31(2):533–544, 2011. 4, 63

[96] Dominic Rüfenacht, Matthew Brown, Jan Beutel, and Sabine Süsstrunk. Temporally consistent snow cover estimation from noisy, irregularly sampled measurements. In *International Conference on Computer Vision Theory and Applications*, 2014. 2

[97] Nina Runge, Pavel Samsonov, Donald Degraen, and Johannes Schöning. No more autobahn!: Scenic route generation using googles street view. In *International Conference on Intelligent User Interfaces*, 2016. 42, 47, 52

[98] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 12, 35

[99] Frode Eika Sandnes. Determining the geographical location of image scenes based on object shadow lengths. *Journal of Signal Processing Systems*, 2011. 29

[100] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 24, 30

[101] Chanuki Illushka Seresinhe, Helen Susannah Moat, and Tobias Preis. Quantifying scenic areas using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*, 2017. 43

[102] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the impact of scenic environments on health. *Scientific Reports*, 5, 2015. 42

[103] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Using deep learning to quantify the beauty of outdoor places. *Royal Society Open Science*, 4(7), 2017. 43

[104] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Accurate geo-registration by ground-to-aerial image matching. In *International Conference on 3D Vision*, 2014. 30

[105] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 9

[106] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 66

[107] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Transactions on Graphics*, 25(3):835–846, 2006. 3, 24, 30

[108] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 12

[109] Birger Stichelbaut, Jean Bourgeois, and Nicholas Saunders. *Images of conflict: Military aerial photography and archaeology*. Cambridge Scholars Publishing, 2008. 4

[110] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H Hsu, and Shao-Yi Chien. Scenic photo quality assessment with bag of aesthetics-preserving features. In *ACM International Conference on Multimedia*, 2011. 42, 43

[111] 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. 29

[112] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 35, 49

[113] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3, 9

[114] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *IEEE International Conference on Computer Vision*, 2015. 63

[115] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[116] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 30

[117] John Townshend, Christopher Justice, Wei Li, Charlotte Gurney, and Jim McManus. Global land cover classification by remote sensing: present capabilities and future possibilities. *Remote Sensing of Environment*, 35(2-3):243–255, 1991. 4

[118] David M Tralli, Ronald G Blom, Victor Zlotnicki, Andrea Donnellan, and Diane L Evans. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(4):185–198, 2005. 4

[119] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, 2009. 46

[120] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008. vi, 19, 20

[121] Anirudh Viswanathan, Bernardo R Pires, and Daniel Huber. Vision based robot localization by ground to satellite matching in gps-denied situations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014. 30

[122] Jingya Wang, Mohammed Korayem, Saul Blanco, and David Crandall. Tracking natural events through social media and computer vision. In *ACM International Conference on Multimedia*, 2016. 60, 63

[123] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. In *IEEE International Conference on Computer Vision*, 2017. 60, 70

[124] Jan D Wegner, Steven Branson, David Hall, Konrad Schindler, and Pietro Perona. Cataloging public objects using aerial and street-level images-urban trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 63

[125] Paul Werbos. *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University, 1974. 12

[126] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, 2016. 3, 48

[127] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *IEEE/ISPRS Workshop: Looking From Above: When Earth Observation Meets Vision*, 2015. 7, 24, 30, 32, 34, 55, 63

[128] Scott Workman, R. Paul Mihail, and Nathan Jacobs. A pot of gold: Rainbows as a calibration cue. In *European Conference on Computer Vision*, 2014. 24, 29

[129] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, 2015. 7, 24, 55, 63

[130] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and mapping natural beauty. In *IEEE International Conference on Computer Vision*, 2017. 7, 63

[131] Scott Workman, Menghua Zhai, David J. Crandall, and Nathan Jacobs. A unified model for near and remote sensing. In *IEEE International Conference on Computer Vision*, 2017. 8, 56

[132] Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wild. In *British Machine Vision Conference*, 2016. 48

[133] Lin Wu and Xiaochun Cao. Geo-location estimation from two shadow trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 29

[134] Michael A Wulder and Nicholas C Coops. Make earth observations open access: freely available satellite imagery will improve science and environmental-monitoring products. *Nature*, 513(7516):30–32, 2014. 4

[135] Ling Xie, Alex Chiu, and Shawn Newsam. Estimating atmospheric visibility using general-purpose cameras. In *Advances in Visual Computing*. Springer, 2008. 2

[136] Ling Xie and Shawn Newsam. Im2map: deriving maps from georeferenced community contributed photo collections. In *ACM SIGMM International Workshop on Social media*, 2011. 48, 52, 61

[137] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *AAAI Conference on Artificial Intelligence*, 2015. 63

[138] Che-Hua Yeh, Yuan-Chen Ho, Brian A Barsky, and Ming Ouhyoung. Personalized photograph ranking and selection system. In *ACM International Conference on Multimedia*, 2010. 43

[139] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, 2014. 27

[140] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, 2010. 30

[141] Menghua Zhai, Zach Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 55, 64

[142] Haipeng Zhang, Mohammed Korayem, David J Crandall, and Gretchen LeBuhn. Mining photo-sharing websites to study ecological phenomena. In *International World Wide Web Conference*, 2012. 3

[143] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, 2016. 3

[144] Xiaoyang Zhang, Mark A Friedl, Crystal B Schaaf, Alan H Strahler, John CF Hodges, Feng Gao, Bradley C Reed, and Alfredo Huete. Monitoring vegetation phenology using modis. *Remote sensing of environment*, 84(3):471–475, 2003. 4

[145] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000. 3

[146] 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. vii, 9, 13, 17, 18, 28, 30, 31, 33, 35, 38, 47, 49

[147] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *International Conference on Learning Representations*, 2014. 50

[148] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2017. 66

[149] Bolei Zhou, Liu Liu, Aude Oliva, and Antonio Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *European Conference on Computer Vision*, 2014. 9, 15, 29, 30, 63

[150] Yi Zhu and Shawn Newsam. Land use classification using convolutional neural networks applied to ground-level images. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015. 60

[151] Ervin H Zube, James L Sell, and Jonathan G Taylor. Landscape perception: research, application and theory. *Landscape planning*, 9(1):1–33, 1982. 42

Vita

Scott Workman

Education

2006–2010	B.S. in Computer Science *Magna Cum Laude* with Honors	University of Kentucky

Appointments

Research Assistant 2011–Present	Center for Vis. & Virtual Envs., University of Kentucky *Lexington, KY*
Research Assistant 2010–2011	CEPIS, University of Kentucky *Lexington, KY*
Software Engineer Intern Summer 2010	CorrectCare - Integrated Health *Lexington, KY*
Network Engineer Intern 2008–2010	NEMOC, University of Kentucky *Lexington, KY*
Help Desk Technician 2007–2010	ECS, University of Kentucky *Lexington, KY*
Family Business 1997–Present	Deer Run Stables & Horseman's Retreat *Richmond, KY*

Publications

Refereed Journal Publications

[1] Nathan Jacobs, Scott Workman, and Richard Souvenir. Cloudmaps from Static Ground-View Video. *Image and Vision Computing (IVC)*, 52:154–166, August 2016.

[2] Scott Workman, Richard Souvenir, and Nathan Jacobs. Scene Shape Estimation from Multiple Partly Cloudy Days. *Computer Vision and Image Understanding (CVIU)*, 134:116–129, May 2015.

Refereed Conference Publications

[1] Connor Greenwell, Scott Workman, and Nathan Jacobs. What Goes Where: Predicting Object Distributions from Above. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018.

[2] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. A Multimodal Approach to Mapping Soundscapes. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018.

[3] Weilian Song, Scott Workman, Armin Hadzic, Xu Zhang, Eric Green, Mei Chen, Reginald Souleyrette, and Nathan Jacobs. FARSA: Fully Automated Roadway Safety Assessment. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[4] Scott Workman, Menghua Zhai, David J. Crandall, and Nathan Jacobs. A Unified Model for Near and Remote Sensing. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[5] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and Mapping Natural Beauty. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] Menghua Zhai, Zach Bessinger, Scott Workman, and Nathan Jacobs. Predicting Ground-Level Scene Layout from Aerial Imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon Lines in the Wild. In *British Machine Vision Conference (BMVC)*, 2016.

[8] Menghua Zhai, Scott Workman, and Nathan Jacobs. Camera Geo-Calibration using an MCMC Approach. In *International Conference on Image Processing (ICIP)*, 2016.

[9] Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting Vanishing Points using Global Image Context in a Non-Manhattan World. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] R. Paul Mihail, Scott Workman, Zach Bessinger, and Nathan Jacobs. Sky Segmentation in the Wild: An Empirical Study. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[11] Ryan Baltenberger, Menghua Zhai, Connor Greenwell, Scott Workman, and Nathan Jacobs. A Fast Method for Estimating Transient Scene Attributes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[12] Tawfiq Salem, Scott Workman, Menghua Zhai, and Nathan Jacobs. Analyzing Human Appearance as a Cue for Dating Images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[13] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-Area Image Geolocalization with Aerial Reference Imagery. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[14] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. DeepFocal: A Method for Direct Focal Length Estimation. In *International Conference on Image Processing (ICIP)*, 2015.

[15] Mohammad T. Islam, Scott Workman, and Nathan Jacobs. Face2GPS: Estimating Geographic Location from Facial Features. In *International Conference on Image Processing (ICIP)*, 2015.

[16] Scott Workman, R. Paul Mihail, and Nathan Jacobs. A Pot of Gold: Rainbows as a Calibration Cue. In *European Conference on Computer Vision (ECCV)*, 2014.

[17] Mohammad T. Islam, Scott Workman, Hui Wu, Richard Souvenir, and Nathan Jacobs. Exploring the Geo-Dependence of Human Face Appearance. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[18] Nathan Jacobs, Scott Workman, and Richard Souvenir. Scene Geometry from Several Partly Cloudy Days. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2013.

[19] Nathan Jacobs, Mohammad Islam, and Scott Workman. Cloud Motion as a Calibration Cue. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Refereed Workshop Publications

[1] Nathan Jacobs, Scott Workman, and Menghua Zhai. Cross-view Convolutional Networks. In *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2016.

[2] Scott Workman and Nathan Jacobs. On the Location Dependence of Convolutional Neural Network Features. In *IEEE/ISPRS Workshop: Looking from above: When Earth observation meets vision (EARTHVISION)*, 2015.

Abstracts

[1] Scott Workman and Nathan Jacobs. Scene Understanding using Clouds. In *International Computer Vision Summer School (ICVSS)*, 2014.

[2] J. David Smith, Ryan Baltenberger, Scott Workman, and Nathan Jacobs. User-in-the-Loop Calibration and Mensuration. In *National Conference on Undergraduate Research (NCUR)*, 2014.

[3] Ryan Baltenberger, James Knochelmann, Scott Workman, Mohammad Islam, Nathan Jacobs, and James Griffioen. Constructing a High-Resolution Mosaic of Kentucky Lake. In *Kentucky GIS Conference*, 2013.

[4] Xuzi Zhou, Scott Workman, Mohammad Islam, Nathan Jacobs, and James Griffioen. Cyber Infrastructure for the VOEIS Project. In *Symposium in the Mathematical, Statistical and Computer Sciences*, 2013.

[5] Scott Workman, James Knochelmann, Nathan Jacobs, David S. White, and Richard Hauer. Registration and Visualization of Scientific Aerial Imagery at Kentucky Lake. In *Kentucky EPSCoR Conference*, 2012.

Honors and Awards

- 5th Heidelberg Laureate Forum, 2017
- Outstanding Ph.D. Student in Computer Science, University of Kentucky, 2017
- Burton E. Heard Graduate Fellowship, 2016-2017
- NVIDIA Academic Hardware Grant (Tesla K40), 2015
- Dean's List, University of Kentucky, 2008-2010
- Alltel/Windstream Scholarship, University of Kentucky, 2008-2010