



University of Kentucky
UKnowledge

Theses and Dissertations--Computer Science

Computer Science


2019

Learning to Map the Visual and Auditory World

Tawfiq Salem

University of Kentucky, tawfiq.salem@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0001-6232-0542>

Digital Object Identifier: <https://doi.org/10.13023/etd.2019.340>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Salem, Tawfiq, "Learning to Map the Visual and Auditory World" (2019). *Theses and Dissertations--Computer Science*. 86.

https://uknowledge.uky.edu/cs_etds/86

This Doctoral Dissertation is brought to you for free and open access by the Computer Science at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Computer Science by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

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.

Tawfiq Salem, Student

Dr. Nathan Jacobs, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

Learning to Map the Visual and Auditory World

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
Tawfiq Salem
Lexington, Kentucky

Director: Dr. Nathan Jacobs
Associate Professor of Computer Science
Lexington, Kentucky 2019

Copyright© Tawfiq Salem 2019

ABSTRACT OF DISSERTATION

Learning to Map the Visual and Auditory World

The appearance of the world varies dramatically not only from place to place but also from hour to hour and month to month. Billions of images that capture this complex relationship are uploaded to social-media websites every day and often are associated with precise time and location metadata. This rich source of data can be beneficial to improve our understanding of the globe. In this work, we propose a general framework that uses these publicly available images for constructing dense maps of different ground-level attributes from overhead imagery. In particular, we use well-defined probabilistic models and a weakly-supervised, multi-task training strategy to provide an estimate of the expected visual and auditory ground-level attributes consisting of the type of scenes, objects, and sounds a person can experience at a location. Through a large-scale evaluation on real data, we show that our learned models can be used for applications including mapping, image localization, image retrieval, and metadata verification.

KEYWORDS: computer vision, machine learning, deep neural networks, remote sensing, geospatial analysis, mapping

Author's signature: _____ Tawfiq Salem

Date: _____ July 31, 2019

Learning to Map the Visual and Auditory World

By
Tawfiq Salem

Director of Dissertation: Nathan Jacobs

Director of Graduate Studies: Mirosław Truszczyński

Date: July 31, 2019

This work is dedicated to the memory of my parents, Mousa and Huda Salem. I would have never become who I am now without their inspiration and limitless love and support. They taught me the importance of commitment, hard work, and going above and beyond to achieve my goals and dreams.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation and gratitude to my advisor, Dr. Nathan Jacobs, for his constant encouragement, support, and guidance throughout my doctoral program. Dr. Jacobs was always there to listen to me and to give me sincere advice. He showed me different ways to approach a research problem. More importantly, he inspired me to work hard by being a role model himself. I would have never completed my research work without his help, guidance, and support.

My sincere thanks go to the rest of my committee members, Dr. Seales, Dr. Yang, and Dr. Cheung for their helpful feedback and discussions which helped me curate this document and present my work in a meaningful way. Special thanks to Dr. Sama for agreeing to be my outside examiner in my defense. I am grateful to the graduate coordinator, Dr. Mirek Truszczynski, for his encouragement and support during my Ph.D. studies. I would also like to extend my thanks to all the faculty members of the computer science department for providing me with a solid foundation, not only in computer science but also in academia.

I will forever be thankful to my lab-mates, including Scott Workman, Menghua Zhai, Zach Bessinger, Connor Greenwell, Weilian Song, M. Usman Rafique, Hunter Blanton, and others. We have all been there for one another and have taught ourselves and each other many tools and skills. I know that I could always ask them for advice, opinions, and ideas on different research problems. Thank you all for the fun and support. I am eagerly looking forward to having all of you as colleagues in the years ahead.

Last but not least, I would like to thank my family, whose continuous love, support, and encouragement have been the light of my life. My brothers and sisters, thank you for your support, and special thanks goes to my brother, Saeed Salem, who was the first person to encourage me to pursue a graduate education. And, of course, thanks for the constant support through the ups and downs of my academic career. To my son Bilal and my daughters Jena and Huda, you have always been a source of joy, inspiration, and encouragement to work hard toward achieving my goals. To my wife, Wafaa, who inspired me all the time and for the endless love and support she provided during my journey.

* * *

Table of Contents

Acknowledgments	iii
Table of Contents	iv
List of Figures	vi
List of Tables	ix
Chapter 1 Introduction	1
1.1 Image Driven Mapping	2
1.2 Mapping Using Overhead Imagery	3
1.3 Contributions	4
1.4 Dissertation Outline	5
Chapter 2 Technical Background	7
2.1 Learning with Convolutional Neural Networks	7
2.2 Transfer Learning	8
Chapter 3 General Framework for Mapping Geospatial Attributes	12
3.1 Estimating Images Attributes	12
3.2 Mapping Time-Variant Image Attributes	13
3.3 General Framework for Mapping	14
Chapter 4 Learning Static Maps of Visual Appearance	16
4.1 Introduction	16
4.2 Related Work	17
4.3 Approach	17
4.4 Evaluation	20
4.5 Conclusion	23

Chapter 5	Learning Dynamic Maps of Visual Appearance	24
5.1	Introduction	24
5.2	Related Work	26
5.3	Problem Definition	28
5.4	Dynamic Visual Appearance Mapping	28
5.5	Evaluation	31
5.6	Discussion	39
5.7	Conclusion	40
Chapter 6	Learning to Map Soundscapes	41
6.1	Introduction	41
6.2	Cross-View Aural Mapping	42
6.3	Experiments	46
6.4	Conclusion	49
Chapter 7	Discussion	50
Bibliography		52
Vita		60

LIST OF FIGURES

1.1	The semantic content of the overhead image and the co-located ground level image are similar.	2
2.1	Convolutional neural network structure. [1]	7
2.2	The common way of extracting features from a model is by removing the last couple of layers and use the rest of the model to extract features for input data from the new task.	9
2.3	The general approach for fine-tuning a pre-trained model on a new related problem.	10
2.4	The general approach for model-to-model learning.	10
3.1	Different models trained for estimating ground-level attributes.	13
3.2	Transient attributes of a scene change over time.	13
3.3	Visual appearance changes dramatically due to differences in location and time. Our work takes advantage of sparsely distributed, ground-level image data, with associated location and time metadata, in conjunction with overhead imagery to construct dynamic maps of visual appearance attributes.	14
4.1	An overview of our network architecture.	18
4.2	For a given ground image, we show the top-3 overhead images that give the highest probability for the given image. The top row is based on Places, the second on Imagenet, and the last on Object counts.	20
4.3	Given a query ground-level image (left), we can construct a heatmap (right) that represents the score where the greener the dot on the map the more likely the image was taken in that location.	21
4.4	Localization accuracy of the different learned probabilistic models on the test-set of the ground-level imagery	22
4.5	Overhead images with the highest scores for the <i>car</i> label. The <i>park</i> label score is increased from left to right, transitioning the images from industrial to rural scenes while focusing on roads. Each column represents multiple images with similar scores for the query labels.	22

5.1	Visual appearance changes dramatically due to differences in location and time. Our work takes advantage of sparsely distributed, ground-level image data, with associated location and time metadata, in conjunction with overhead imagery to construct dynamic maps of visual appearance attributes.	25
5.2	An overview of our network architecture, which includes the network we train to predict visual attributes (left) and the networks we use to estimate visual attributes (right).	27
5.3	The spatial distribution of the dataset. The green (red) dots represent the training (testing) data.	30
5.4	The temporal distribution of the dataset.	31
5.5	Dynamic visual attribute maps for different transient attributes and months. In each, yellow (blue) corresponds to a higher (lower) value for the corresponding attribute. Each attribute exhibits unique spatial and temporal patterns, which closely match the authors' personal travel experiences.	33
5.6	Dynamic visual attribute maps for different methods on the transient attribute <i>sunny</i>	33
5.7	For a given location and corresponding overhead image, predictions of the <i>sunrise-sunset</i> attribute at different hours for two different months. This highlights that our model has learned that days are longer during the summer.	35
5.8	For each overhead image, we predict the visual attributes using our full model and compute the average distance between them and those of the ground-level images in the test set. (left) The overhead images of two query locations. The closest images when using August at 5pm as input (middle) and when using August at 2am (right).	35
5.9	Localization accuracy as a function of candidate images searched. Our approach, <i>image+time+loc</i> , outperforms all baselines.	36
5.10	Given a query ground-level image (top), we show localization results (bottom) for different scoring strategies, visualized as a heatmap. Red (blue) represents a higher (lower) likelihood that the image was captured at that location.	37

5.11	An example highlighting temporal patterns learned by our model. For each example, we show the original image and the overhead image of its location. For every possible hour and month, we use our full model to predict the visual attribute. The heatmap shows the distance between the true and predicted visual attributes, with dark green (white) representing smaller (larger) distances. In the top example, there are two narrow bands of small distances, centered around dawn and dusk. In the top example, we see small distances during the nighttime hours.	38
6.1	We propose a multimodal approach for relating overhead image appearance with sounds in order to map soundscapes. (left) Overhead image; (right) Similar ground-level images and sounds output by our method.	42
6.2	An overview of our network architecture.	43
6.3	The distribution of the collected audio files in our CVS dataset.	44
6.4	A word cloud for the tags associated with the sounds in the CVS dataset.	45
6.5	The model architecture for predicting a distribution over sound clusters from an overhead image.	46
6.6	Our work explores the relationship between overhead image appearance and sound. Given an overhead image (top), our model outputs a distribution over sound clusters (bottom).	48
6.7	Block-level audio mapping: (left) An overhead image of a small geographical region on Miami beach. (right) A per-pixel labeling of sound clusters.	48
6.8	City-level audio mapping: (left) An overhead image covering New York City. (right) A per-pixel labeling of sound clusters.	48
6.9	Country-level audio mapping: visualizing the sound clusters over USA. Gaps (white) are regions where the CVUSA dataset does not have imagery.	49

LIST OF TABLES

5.1	Comparing performance on the visual attribute prediction task.	32
6.1	Quantitative performance of different networks.	47

Chapter 1

Introduction

“Even before you understand them, your brain is drawn to maps.”

– Ken Jennings

Looking at the world from above, you will see a diverse range of scenery, from mountains to beaches, deserts to rainforests, and many other types of scenery. Furthermore, the appearance of the same scene may change over time. For example, if we were to look at a forest in summer, we will expect to see green trees and sunny weather, whereas if we look at it in winter, we will likely see leafless trees and snowy weather. Depending on geolocation and time, our world can look very different. One of the most efficient ways to represent these differences is by maps. Maps can represent our big world at a much smaller scale for people to explore and provide information that helps them decide where to live and how to travel from a location to another. Moreover, cities and countries often use maps for monitoring land-use and land-cover changes, especially those caused by human activities, and help them in making decisions toward planning and development.

There are many different types of maps based on the information they provide. For example, a street map shows you roads, their names, and different points of interests along these roads. A topographic map represents information about land elevations and features, a park map that shows you trails and playgrounds, and a city map shows the roads and locations of important buildings such as hospitals. Constructing such maps at a global scale can be very costly and time-consuming since there will be a need to collect information from all over the world.

Recently, the widespread use of social media and the increasing availability of smartphones have conveniently allowed people to capture and share most of their life activities in terms of images and videos. This data is often associated with precise time and location metadata. Moreover, advances in remote sensing and satellite technologies enabled

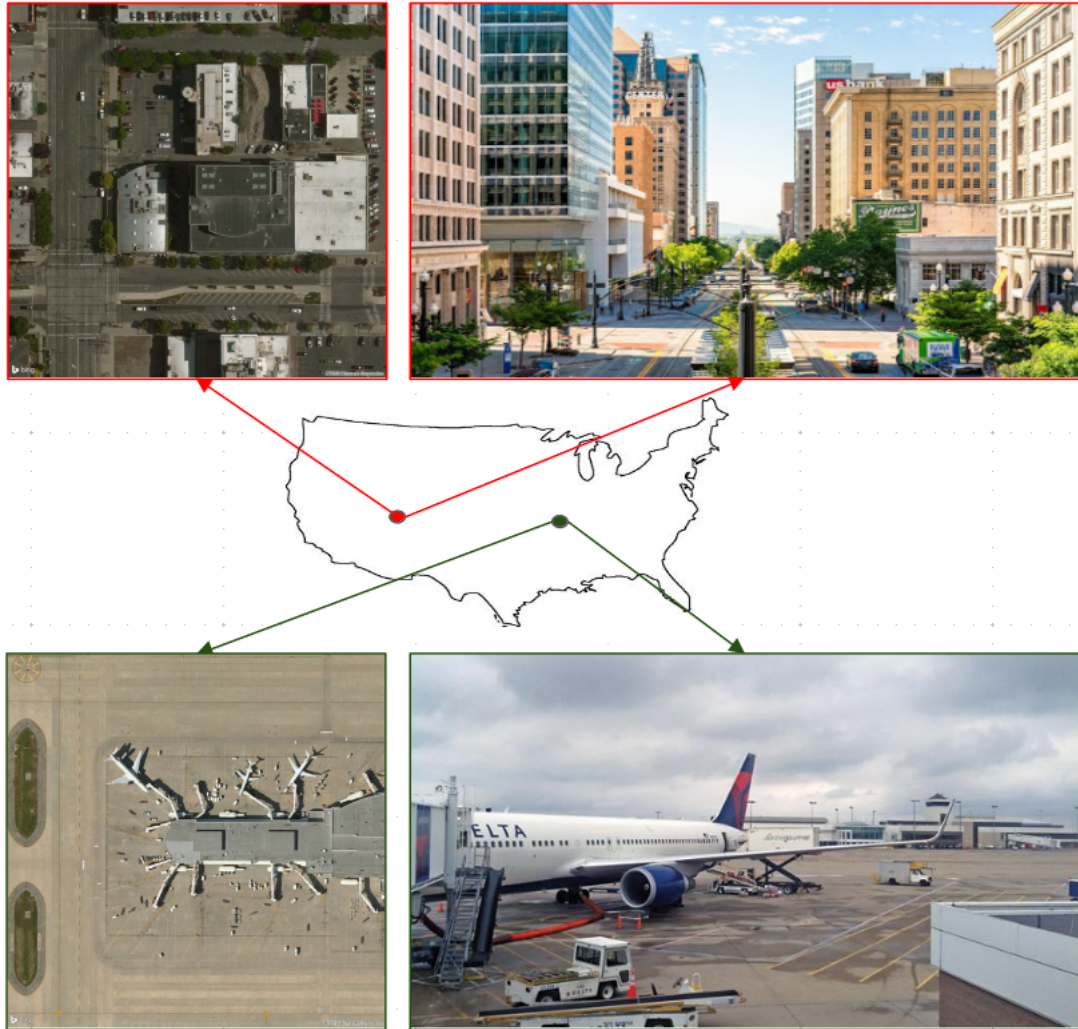


Figure 1.1: The semantic content of the overhead image and the co-located ground level image are similar.

the accumulation of overhead imagery that covers most of the globe. Currently, billions of geotagged images are publicly available on the Internet. This rich source of images, coming from all over the world, could be very beneficial to better understand our globe. This dissertation proposes a general framework for utilizing these publicly available images for mapping different ground level attributes.

1.1 Image Driven Mapping

When we, as humans, look at an image it does not take much effort for us to recognize a cat or a human face in the image. Although humans perform these tasks effortlessly, these are actually hard problems to solve with a computer since computers see an image as a

matrix of numbers. In computer vision, researchers focus on how computers can learn to gain high-level understanding from digital images or videos. Different works in computer vision have focused on different ground-level image understanding tasks, such as image classification [26, 56] and object recognition [60, 74]. Researchers have also proposed algorithms for estimating image attributes including scene classification [64, 80], weather estimation [7, 33], and image geolocalization [25, 68].

Recently, with the availability of large datasets of geotagged images, different works have proposed methods for mapping ground-level attributes based on ground-level image appearance. Gebru et al. [19] proposed a method for estimating and mapping socio-economic characteristics of different US cities using millions of Google street-view images. Another work by Arietta et al. [4] proposed a method for automatically identifying and mapping the relationships between the visual appearance of a city and its non-visual attributes including crime statistics, housing prices, and population density. Other proposed methods for mapping people’s visual appearance [10, 28]. Because of the sparsity of the existing ground-level imagery, methods proposed to map ground-level attributes based on ground-level images lack the ability to generate dense maps at a global scale. Furthermore, these methods are biased towards locations with large numbers of ground-level images.

1.2 Mapping Using Overhead Imagery

To tackle the problem of sparsely distributed ground-level imagery, this dissertation proposes a general framework for mapping ground-level attributes from overhead imagery. Although overhead imagery is uniformly distributed and available for almost every location on earth, the lack of annotated data for model training remains a challenge for learning from overhead imagery. However, pairing overhead and co-located ground images gives hope at tackling this challenge. If we were to look at the different pairs of overhead and the co-located ground-level images in Figure 1.1, we will notice the similarity in the semantic contents of the different views. The overhead image (left) in the top row shows an image covering a city with roads and buildings, and we can see similar content if one looks at the ground image (right) taken at the center of the overhead image. We can also notice the same observations in the bottom pair of images that cover an airport with runways, terminals, and airplanes.

Several existing works have approached this problem. For example, Lee et al. [35] proposed an approach to estimate geo-informative attributes such as population density and demographic properties. Similar to our approach, the work in [70] proposed a method for mapping between ground-level and overhead image and utilized the proposed method

for the problem of image localization. The previous methods assumed that the predictions of the ground-level images were distributed in a Gaussian manner, and only learned to predict the mean of the distribution. This is problematic when a particular overhead image could have multiple possible scenes. For example, the overhead image (left) in the top row in Figure 1.1 covering parks, stadium, and parking lots.

To capture this multi-label feature of images, we propose to learn probabilistic models that capture the distribution of the different ground-level image labels. We also introduce a method for constructing dynamic maps at a global scale, which is a probability distribution conditioned not only on the geographic location but also on time. These maps capture the change in different ground level attributes over time. To the best of our knowledge, our work is the first to integrate dense overhead imagery with location and time metadata to better model the relation between image appearance, location, and time. We further propose a method for mapping soundscapes to show the type of sounds one can hear at a given geolocation.

1.3 Contributions

In this work, we propose a general framework for mapping ground-level attributes from overhead imagery. We propose different models for constructing static and dynamic maps for different types of attributes. We also conduct different experiments to show the different applications of our proposed methods. For the purpose of the research in this dissertation, a map can be thought of as a probability distribution, conditioned on geographic location, over a collection of attributes; typical attributes include land cover, land use, elevation, or place name.

Main Contributions: The main contributions of this dissertation are:

- A general framework based on deep convolutions neural networks for mapping ground-level attributes.
- A model for constructing a static map that captures the change in the type of scenes, objects, and object counts from overhead imagery.
- A model for constructing dynamic maps for different ground-level attributes that change over time and geolocation.
- A model for understanding the types of sounds that could be heard at a specific geographic location and mapping the soundscapes at a global scale.
- A detailed evaluation, both quantitative and qualitative, of the learned models for a variety of different settings.

- Present different applications for these models, including mapping, image localization, image retrieval, and metadata verification.

1.4 Dissertation Outline

The remainder of this document consists of the following chapters:

- **Chapter 2** provides a technical background that is necessary for understanding the work in this dissertation. We provide an overview of related research works in three areas: learning with convolutional neural networks, transfer learning, and image attributes.
- **Chapter 3** presents a general framework for mapping ground-level attributes from overhead imagery. A key element of our approach is that we use visual attributes present in ground-level images as a supervisory signal for model training, thus requiring no labeled overhead imagery.
- **Chapter 4** introduces an approach that makes it possible to draw a variety of conclusions from overhead imagery. In particular, we propose using well-defined probabilistic models and a weakly-supervised, multi-task training strategy to learn to predict properties and their uncertainties for a given location. We show that our learned models can be used directly for applications in mapping and image localization. The material in this chapter has been published [50].
- **Chapter 5** introduces an approach for constructing dynamic maps of visual appearance attributes that can provide an estimate of the expected appearance at any geographic location and time. Our approach integrates dense overhead imagery with location and time metadata. A key element of our method is that we use visual attributes present in ground-level images as a supervisory signal for model training, thus requiring no labeled overhead imagery. Through a large-scale evaluation on real data, we find that combining overhead imagery with metadata results in more accurate predictions and better performance on a variety of tasks.
- **Chapter 6** explores the problem of mapping soundscapes, that is, predicting the types of sounds that are likely to be heard at a given geographic location. Using a novel dataset, which includes geo-tagged audio and overhead imagery, we develop an approach for constructing an aural atlas, which captures the geospatial distribution of soundscapes. We build on previous work relating sound to ground-level imagery

but incorporate overhead imagery to overcome the limitations of sparsely distributed geo-tagged audio. In the end, all that we require to construct an aural atlas is overhead imagery of the region of interest. We show examples of aural atlases at multiple spatial scales, from block-level to country. The material in this chapter has been published [52].

- **Chapter 7** summarizes the contributions of this dissertation and our most important findings. In addition, we discuss possible future research directions that will lead to improved methods for geo-visual analysis and understanding.

Chapter 2

Technical Background

In this chapter, we provide an overview of related research works in two areas: learning with convolutional neural networks and present the different techniques for transfer learning.

2.1 Learning with Convolutional Neural Networks

A neural network is a computational model that is inspired by the way biological neural networks in the human brain process information. Convolutional Neural Networks (ConvNets or CNNs) is a category of neural networks that has gained attention since Alex Krizhevsky [32] developed the AlexNet structure and used it in 2012 to win the ImageNet competition (classifying 1.2 million images to 1.000×10^3 classes). AlexNet considerably outperformed the previous state-of-the-art methods, dropping the classification error from 26% to 15%. Since then CNNs has generated a lot of excitement in research and industry. Researchers in computer vision continue using CNNs and have shown great success on traditional computer vision problems: image classification [26, 56], object recognition [60, 74], and scene classification [64, 80]. Furthermore, different big companies

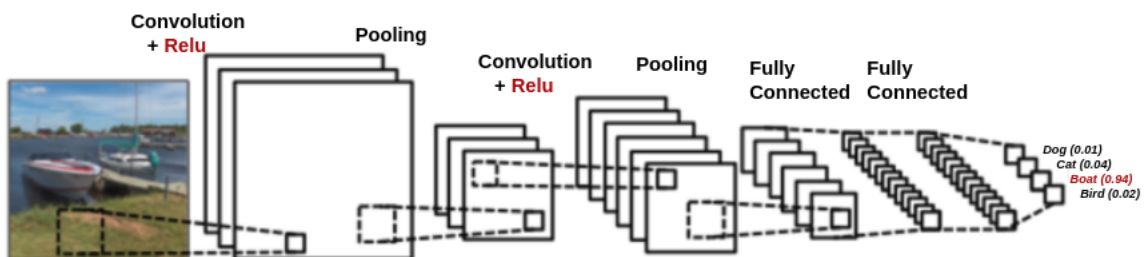


Figure 2.1: Convolutional neural network structure. [1]

started employing deep learning in their services, such as face recognition at Facebook, photo search at Google, product recommendation system at Amazon. One of the main reasons behind the great performance of CNNs is its ability to learn hierarchical features representation of the input images while traditional methods use hand-engineered features.

How CNNs learns the hierarchical features representation of an image? computers see an image as an array of numbers with size equals width \times height \times channels. For example, in image classification the input is an image, array of numbers, and the expected output is a probability distribution over the different classes that the input image belongs to. To perform image classification, CNNs looks for low-level features such as edges and curves in the training images and then construct abstract concepts through a series of linear and nonlinear operations achieved by a combination of layers.

The major components of a CNNs model is a number of convolutional and subsampling layers optionally followed by fully connected layers. The excellent performance of CNNs most of the time come on problems that involve learning discriminative models that usually map a high-dimensional data (e.g. image) to a class label as shown in Figure 2.1. This learning approach, known as supervised learning, for training convolutional neural networks depend on large amounts of labeled samples (training data). Different problems lack the amount of labeled data required for training CNNs. The common practice in deep learning for such problems is to use transfer learning approaches.

2.2 Transfer Learning

The traditional method for training CNNs on a discriminative problem depends on the availability of labeled data. For example, the performance improvement achieved in Krizhevsky et al. [32] was only possible due to the existence of massive training labeled records (1.2 Million training-set). Nowadays, researchers rarely train an entire CNNs model from scratch (with random initialization), because it is hard and costly to have enough training labeled data [76]. Instead, it is a common practice to use and adapt pre-trained models for the new related task of interest. Next, we will explain in more details the different ways to use pre-trained models on a new related task.

2.2.1 Fixed feature extractor

Different works have shown that the features extracted from a model trained on a task can be useful for other related tasks. Razavian et al. [55] have shown that features extracted from OverFeat [54] that has the same structure as AlexNet and trained on ImageNet, per-

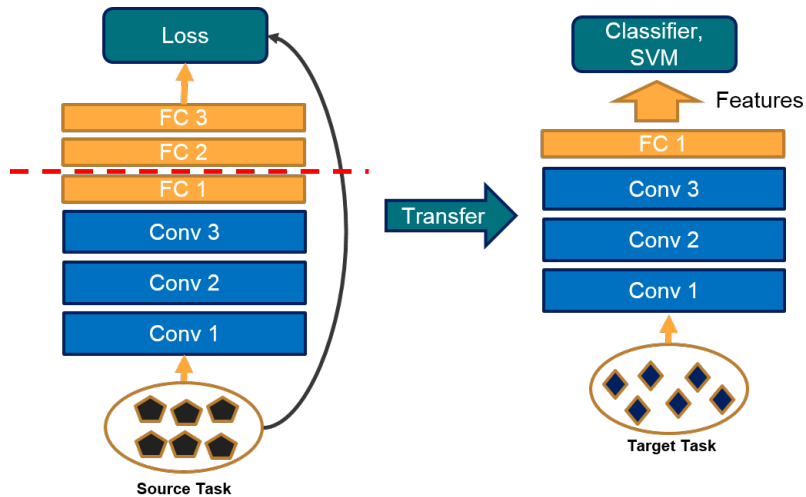


Figure 2.2: The common way of extracting features from a model is by removing the last couple of layers and use the rest of the model to extract features for input data from the new task.

formed reasonably well on a whole variety of tasks, including image classification, and image retrieval. A common way of extracting useful features from a well-trained model such as pre-trained AlexNet is by removing the last fully-connected layer(s) and using the rest of the trained-model as a fixed feature extractor for the new data on the new domain. Different classifiers, e.g. SVM, can be applied over the extracted features as shown in Figure 2.2. A recent work by Owens et al. [45] proposed a method for learning scenes and object detection by training a model for predicting the audio label from an image. Then, extracted the mid-level features from this model for thousands of images and found that different neurons are selective on specific objects and showed that these mid-level features can be very helpful for object and scene detection and classification.

2.2.2 Domain Adaptation

Would it be possible to do better than off-the-shelf features? the answer is yes in many of the cases by what is known as domain adaptation. In domain adaptation, we start with a well-trained model for a ‘nearby’ task, such as AlexNet that has been trained on 1.2 million images for image classification, and you cut off the top layer(s) and replace it with a new layer(s) suitable for the new related task. We fine-tune the new network by running iterations of stochastic gradient descent and back-propagation on the training data for the new problem with new loss function as can be seen in Figure 2.3. It is possible to fine-tune all the layers of the new model or keep some of the earlier layers fixed and only fine-tune some higher-level portions of the network.

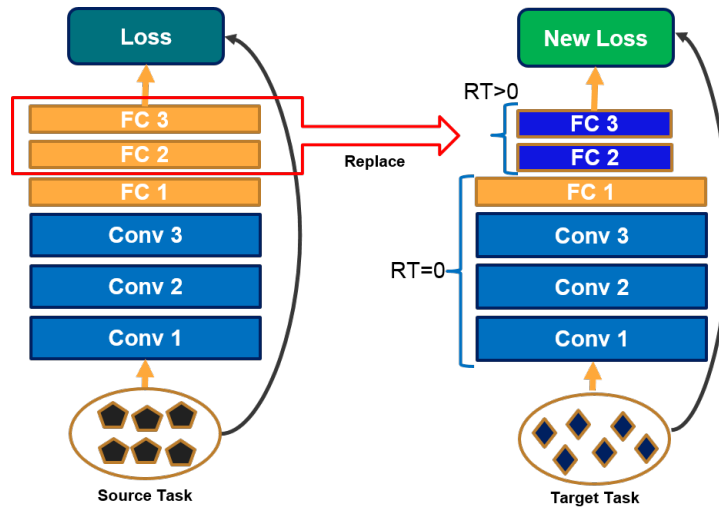


Figure 2.3: The general approach for fine-tuning a pre-trained model on a new related problem.

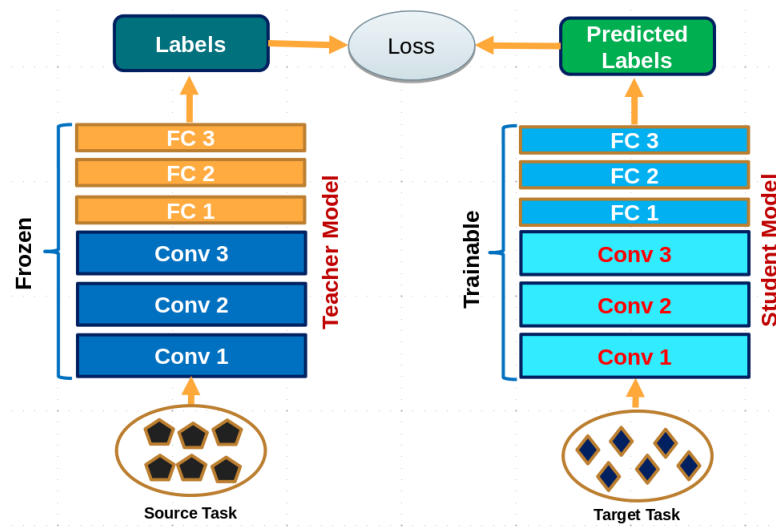


Figure 2.4: The general approach for model-to-model learning.

This is motivated by the observation that the earlier features of convolutional networks contain lower level features, e.g. edges and curves that should be useful to many different tasks, but the final layers of the convolutional networks learn high-level details of the classes contained in the original dataset. This approach usually results in faster training times than training new convolutional networks from scratch [16, 18, 22, 63].

2.2.3 Knowledge Distillation

Over the past few years, deep learning techniques have enabled rapid progress on different core problems in computer vision including, image classification and detection. One of the main challenges in training an efficient deep learning model on a new problem is the need for lots of data which is most of the time hard and costly to obtain. Recently, Ba et al. [6] proposed a training procedure to transfer knowledge from a previously trained model (teacher) that achieve high performance to a model on a related task (student) as shown in Figure 2.4. Different work recently used this approach, e.g., the work by Workman [70] proposed a cross-view training approach for learning to predict the scenes categories from overhead-imagery. Workman et al. [70] used the prediction of the last layer of the well-trained AlexNet-Places model that produces a categorical-distribution over 250 scenes categories for any given outdoor ground-level image as a weak signal during training. Another interesting work by Aytar et al. [5] proposed a method for transferring from vision to other modalities and built a model of learning sound features using two vision-based trained models, Places-CNN and ImageNet-CNN. Similar to this, in this work we use models trained on the ground-level to provide a supervisory signal to train models on co-located overhead imagery.

Chapter 3

General Framework for Mapping Geospatial Attributes

This dissertation proposes a general framework that exploits the publicly available data and the similarity between overhead and ground imagery to make it possible to draw wide-ranging conclusions from overhead imagery. In particular, we propose leveraging the recent advances in ground-level image understanding to learn and map different ground-level attributes from overhead-imagery without requiring annotated data. The proposed approaches utilize pre-existing CNNs to extract categorical distributions of co-located ground-level images to provide a weak signal for training models on overhead imagery. The unique advantages of our proposed methods are: 1) it can operate at a global scale and 2) they are not prone to the bias of sparse data since overhead imagery is uniformly distributed and available for almost every location on earth.

3.1 Estimating Images Attributes

The word *attribute* is a generic term and it can refer to any property that is associated with the image's appearance. It takes us, Humans, a single glance to make a higher-level judgment about a scene as a whole; whether it is a park, airport, or Museum as shown in Figure 3.1b. Different works in computer vision have focused on the problem of scene recognition. The work by Zhou et al. [80] proposed a model called *Places-CNN* using a neural network that has been trained on over two million images from 205 places categories, and then the same authors proposed a new model called *Places2* trained on over ten million images from 365 places categories. Other works have been proposed for estimating the age and gender based on the human appearance in an image [37, 75], the date of when

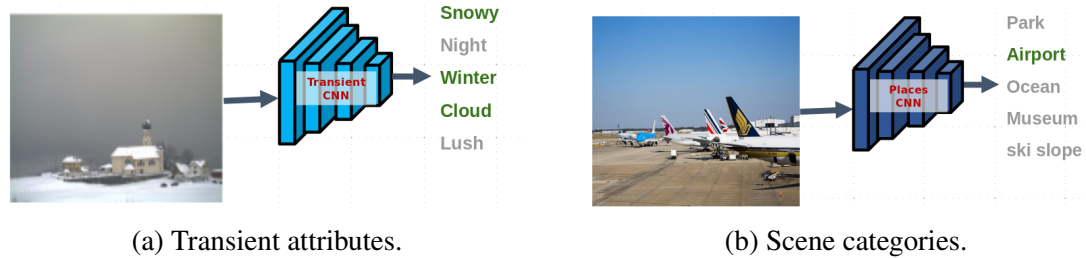


Figure 3.1: Different models trained for estimating ground-level attributes.

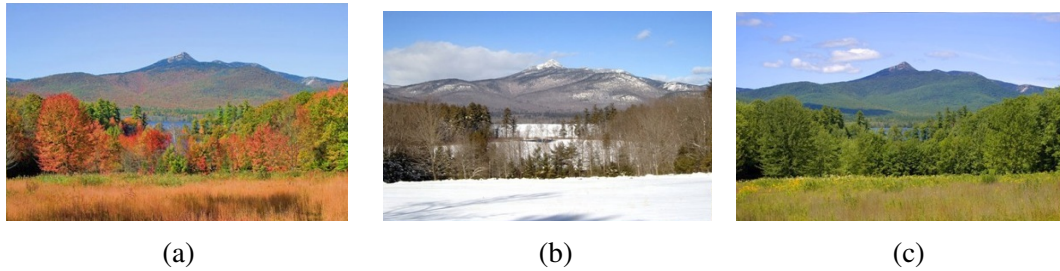


Figure 3.2: Transient attributes of a scene change over time.

an image was captured [17, 30, 34, 46, 51, 78], and the geolocation of where an image was taken [25, 70].

Another interesting ground-level attributes that got attention recently is the problem of estimating weather conditions that exist in the input image (Figure 3.1a). The work by Laffont et al. [33] proposed a method for predicting the presence of 40 transient attributes including winter and summer in an image and developed an interesting approach for outdoor image editing based on these transient attributes. Another work by Baltenberger et al. [7] proposed a fast method based on deep learning for estimating the same 40 different transient attributes for an image. Automatically estimating the properties associated with an image have a lot of potential applications, including high-level image understanding, image geolocalization, image retrieval, image editing, and environmental monitoring.

3.2 Mapping Time-Variant Image Attributes

The time when an image was taken has a direct impact on many image attributes, such as transient attributes. Looking at an image of the same scene at different times as in Figure 3.2, you can notice the obvious changes in the image appearance from summer to fall, and then to winter. Similarly, different other attributes changing over time, for example, people wear different types of clothes at different times of the year. In an image captured in summer, we will see people wearing shorts and t-shirts, whereas in winter they wear coats

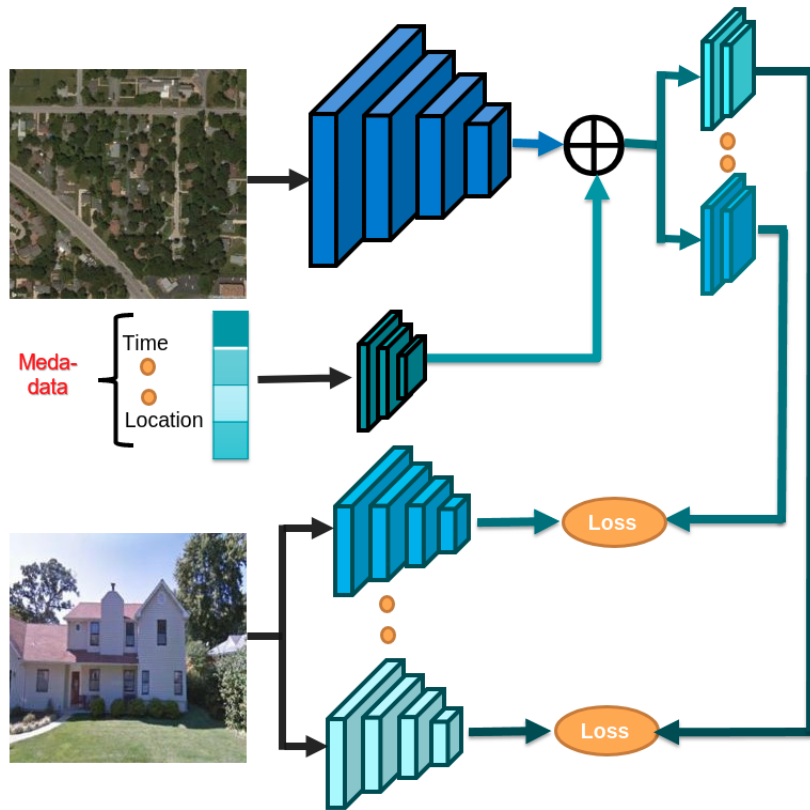


Figure 3.3: Visual appearance changes dramatically due to differences in location and time. Our work takes advantage of sparsely distributed, ground-level image data, with associated location and time metadata, in conjunction with overhead imagery to construct dynamic maps of visual appearance attributes.

and jackets. Another interesting nonvisual attribute of an image that got less attention is the sound that you could hear at an image’s environment. If you were asked to predict the type of sounds you could hear in the environment of an image of a city that was taken at daytime, you will probably predict noisy and crowded sounds. Whereas if you were asked to predict the sound for another image of the same city taken at night, you will probably predict less noisy sounds.

3.3 General Framework for Mapping

Different works in computer vision have proposed approaches for constructing maps that show the differences of objects with respect to the image’s geolocations. Hayes et al. [25] proposed a method for mapping images appearance to geolocation. Other proposed methods for mapping people’s visual appearance [10, 28]. Other works proposed methods for mapping different geospatial attributes from ground-level images including household in-

come, education level, crime statistics, housing prices, and population density [4, 19].

Recently, different works have been proposed for estimating ground-level attributes from overhead imagery. Lee et al. [35] proposed an approach to estimate geo-informative attributes such as population density and demographic properties. Another work by Song et al. [58] proposed a method to estimate a road segment’s free-flow speed from overhead imagery and road metadata. Similar to our approaches, the work in [70] proposed a method for mapping between ground-level and overhead image and utilized the proposed method for the problem of image localization. But these methods ignore the fact that many attributes change not only over geolocation but also continuously changing over time.

In this work, we propose a general framework for image-driven mapping. In particular, we propose a cross-modal distillation strategy to learn to predict the distribution of fine-grained properties from overhead imagery, without requiring any manual annotation of overhead imagery. With this strategy, we are able to estimate probability distributions over categories that would normally be considered too difficult for overhead imagery understanding, due to the lack of available training data. By using large numbers of GPS-tagged consumer photographs, we use off-the-shelf networks including image classification, scene classification, and weather condition estimation in ground-level images to build a sparse training dataset for overhead image understanding. Different ground-level attributes change for many reasons including the time and geolocation. Therefore, our framework has the option to integrate the time and geolocation for constructing dynamic maps.

Figure 3.3 shows an overview of our architecture. We first construct a feature embedding for each conditioning variable (time, location, overhead image) using a set of *context* neural networks (top left). We combine these context features to predict the visual attributes that are coming from the co-located ground-level images using *estimator* networks (top right). On the bottom left, we have a set of pre-trained networks that extract visual attributes from the ground-level images. These networks are only used for extracting visual attributes, which are used to train the context and estimator networks.

Our proposed approach has several advantages.

- it does not require any manually annotated training data.
- it can model spatiotemporal trends, without the need for overhead imagery at every time.
- it is extendable to a wide range of visual attributes.

In the next chapters, we demonstrate the effectiveness of our general framework on mapping different ground-level attributes including places, objects, weather conditions, and sounds.

Chapter 4

Learning Static Maps of Visual Appearance

4.1 Introduction

Traditional approaches to pixel-level labeling of remote sensed imagery rely on the manual specification of semantic categories. In our view, this limits the ability to predict categories for which a human annotator has low confidence. This means that we are unable to learn to make predictions about less certain categories. We propose to overcome this problem using a weakly-supervised learning strategy that uses manually specified labels in a domain for which humans are confident (ground-level imagery) but allows us to make less confident predictions for overhead imagery. With this strategy we are able to estimate probability distributions over categories that would normally be considered too difficult for overhead imagery understanding, due to the lack of available training data.

In particular, using large numbers of GPS-tagged consumer photographs, we use off-the-shelf networks for image classification, scene classification, and object detection in ground-level images to build a sparse training dataset for overhead image understanding. We extend the approach in [70] by modeling the distribution of labels. The previous work assumed that the predictions of the ground-level images were distributed in a Gaussian manner, and only learned to predict the mean of the distribution. This is problematic when a particular overhead image could have multiple possible interpretations from a ground-level perspective. For example, if the overhead image contains a beach and a parking lot, then the ground-level image may be of a beach or a parking lot, depending on the orientation.

To capture such uncertainty, we model the distribution of ground-level image labels

as samples from a Dirichlet distribution. We use a multi-task approach and predict the parameters of prior distributions over three label spaces: scene categorization [79], image classification [32], and object detection [49].

4.2 Related Work

Many recent works jointly reason about ground-level and overhead image viewpoints. Zhai et al. [77] incorporate a transformation between co-located ground-level and overhead images to learn semantic features for overhead imagery. Cross-view image geolocalization strategies [38, 39, 69, 70] learn a feature mapping between the two viewpoints in order to leverage densely available overhead imagery. Other works have used large-scale image collections to map properties of the world. For example, Lee et al. [35] estimate geoinformative attributes such as population density and demographic properties. Another work by Salem et al. [52] proposed an approach for constructing an aural atlas, which captures the geospatial distribution of soundscapes. Most similar to our work, Workman et al. [70] proposed an approach for cross-view training to learn similar feature representation for co-located ground and overhead images and use this for geolocalization. Greenwell et al. [23] proposed a similar cross-view learning approach to learn a model that is capable of predicting the type and count of objects that are likely to be seen from a ground-level perspective conditioned on the overhead image. The previous two methods work on a single ground-level distribution. We propose a general architecture that can learn all labels that we can get from the ground perspective.

4.3 Approach

We propose a cross-view training strategy (Figure 4.1) that uses pre-existing CNNs to extract categorical distributions of ground-level images to provide a weak signal for predicting the parameters of probabilistic models conditioned on co-located overhead imagery. We simultaneously learn three such probabilistic models that model separate ground-level distributions (Places categories, ImageNet categories, and MS-COCO object counts).

4.3.1 Dataset

In our work, we use the 5.51851×10^5 Flickr geotagged ground-level images contained in the Cross-View USA (CVUSA) dataset [70]. For each ground-level image, we extracted two categorical distributions: one over 365-Places categories using a VGG16-Places365

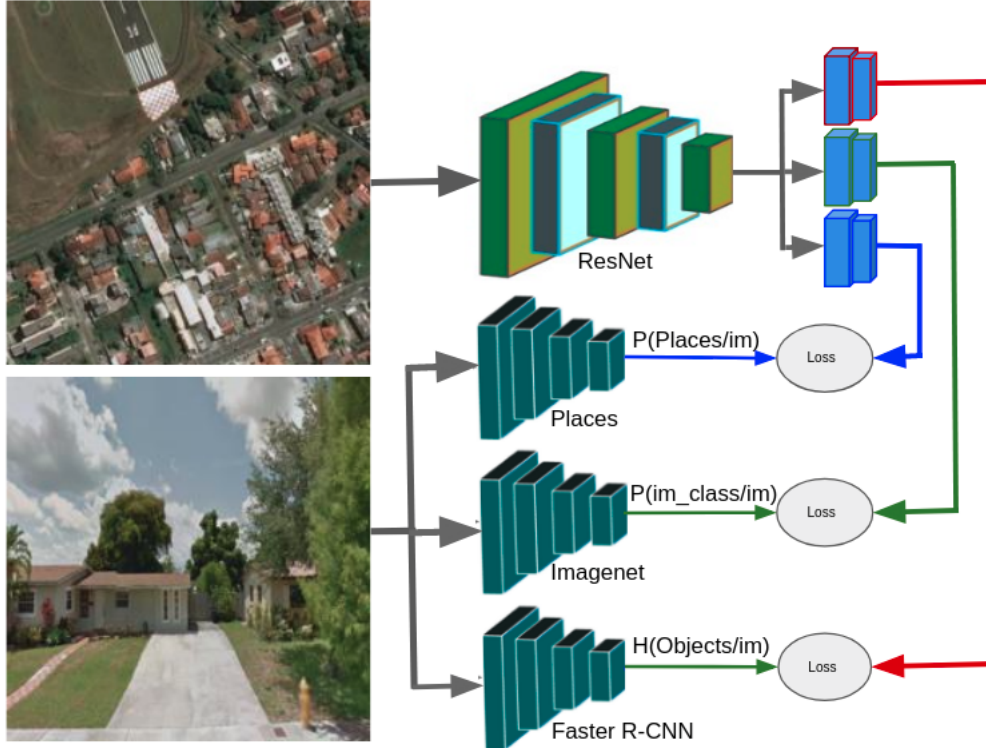


Figure 4.1: An overview of our network architecture.

scene recognition model trained on Places2 [79], and another over 1.000×10^3 -Objects categories using VGG16-Imagenet trained for the task of image classification [32]. We also use the constructed histogram describing the objects present in each image following the work of Greenwell et. al. [23] that uses the output of Faster R-CNN ResNet 101 [49] detector trained on the MS-COCO challenge dataset [40]. To train our model, we split our data into 93% training, 2% validation, and 5% testing.

4.3.2 Distribution Representation

We use two common distribution functions to model priors over ground-level image distributions: Dirichlet and Poisson. The Dirichlet distribution is the conjugate prior of the categorical distribution, meaning samples drawn from a Dirichlet distribution are themselves categorical distributions. Given parameters α_i , the probability density function is given by the following equation:

$$f(x_1, \dots, x_k; \alpha_1, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i - 1} \quad (4.1)$$

where $B(\alpha)$ is a normalizing constant. Using this we can model the one-to-many relationship between overhead imagery and potential ground-level scene and object probabilities

through a discrete set of parameters.

The Poisson distributions describe the likelihood of an event happening k times in some fixed interval. In our case, this will be the probability of k objects of a class being present in the spatial extent of the scene. For each object class, the probability of k objects of that class appearing is given by the following equation, where λ is the interval rate, which varies per class.

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (4.2)$$

Using a Poisson distribution, we can directly model probabilities of not only the types of objects expected in a ground-level scene, but also the number of expected occurrences.

4.3.3 Network Architecture

Our network (Figure 4.1) has two main components, the first is a collection of pre-trained models that we use for extracting ground-level predictions. The second is a shared CNN which takes an overhead image as input and produces a feature which is passed to 3 separate prediction heads. These heads separately predict 1) parameters of a Dirichlet distribution over Places scene categories, 2) parameters of a Dirichlet distribution over ImageNet classes, and 3) parameters of Poisson distributions over the histogram of objects in the image for each overhead image. Each head consist of two fully connected layers. The first layer of each head contains 1024 neurons. The second layer is different for each task, 365, 1.000×10^3 , and 91 respectively. During training we use three losses, one for each distribution, that minimize the mean negative log-likelihood of the resulting probability distributions as in the following equations:

$$L = \min \frac{1}{N} \sum_a -\log(p(g|a, w)) \quad (4.3)$$

where g represents the distribution coming from the ground-level image, a is the overhead imagery, and w the learned weights .

4.3.4 Implementation Details

Our model is implemented in TensorFlow. We trained ResNet-v2-50 [27] using the cross-view learning approach as in [52]. Specifically, the model is trained to predict distributions over ground-level scene and Imagenet categories from overhead imagery using the KL-divergence as a loss function. The trained ResNet is then frozen for subsequent training of the three prediction heads. The heads are initialized randomly using Xavier initialization.



Figure 4.2: For a given ground image, we show the top-3 overhead images that give the highest probability for the given image. The top row is based on Places, the second on Imagenet, and the last on Object counts.

We optimize each head simultaneously by minimizing the negative log-likelihood (Equation 4.3). For both ResNet pre-training and training of the prediction heads, we use the Adam optimizer with a learning rate of 0.001 and a weight decay factor of 0.0005 for 6 epochs with batch size 32. The input images are re-sized to 224×224 , scaled to $[-1, 1]$, and augmented by random horizontal and vertical flipping.

4.4 Evaluation

We evaluated our learned models quantitatively and qualitatively on the test-set defined in the Dataset section.

4.4.1 Cross-View Image Retrieval

For a given overhead image, we extract the output parameters to define the three distributions. The distributions are used to compute the log-probability for any given ground-level image to identify the top-3 overhead images with highest probability. Two qualitative examples are shown in Figure 4.2. The right image shows the ground-level image and on the left we show the top-3 overhead-images for each distribution. For example, in (a) the top row shows overhead images covering Baseball field and *highways* where in the middle and bottom covering *runways*.

4.4.2 Cross-View Localization

Our model can be used for ground-level image localization by defining the distributions for every overhead image in a reference dataset based on the predicted parameters. In Figure 4.3, we took 4.88224×10^5 overhead images from CVUSA. For each overhead image,



Figure 4.3: Given a query ground-level image (left), we can construct a heatmap (right) that represents the score where the greener the dot on the map the more likely the image was taken in that location.

we predict the Places-Dirichlet distributions and, for any ground-level image, get a score for each reference image. The heatmap represents the score, where the greener the dot, the more likely the image was taken in that location. In the middle row, our method clearly identifies the image as having been captured in the coast of USA. Here we show results based on the Places-Dirichlet distribution, but the other two distribution can be used in the same way.

4.4.3 Localization Accuracy

We evaluated the accuracy of our learned probabilistic models on the task of localization. In Figure 4.4 the y-axis represents the accuracy and the x-axis is k , where top- k represents that the correct geolocation for the ground-level image appears in $k\%$ of the overhead images. We can see that using the distribution based on Places gives the highest accuracy and the worst performance (but still better than random) is based on the Poisson distribution.

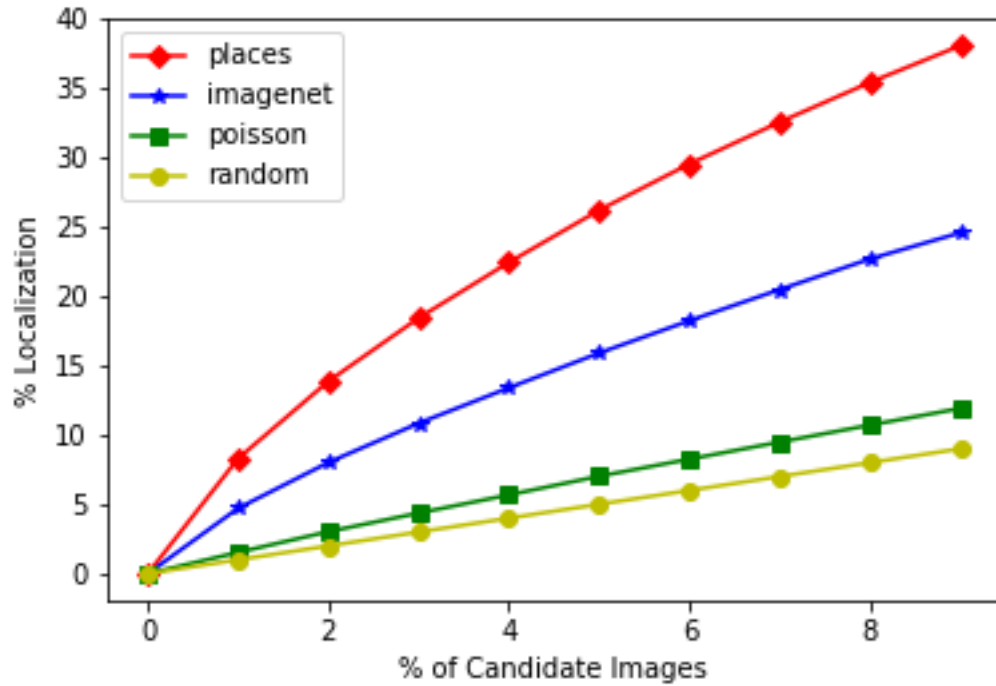


Figure 4.4: Localization accuracy of the different learned probabilistic models on the test-set of the ground-level imagery



Figure 4.5: Overhead images with the highest scores for the *car* label. The *park* label score is increased from left to right, transitioning the images from industrial to rural scenes while focusing on roads. Each column represents multiple images with similar scores for the query labels.

We suspect this is due to the fact that most of the 91 labels used to train the predicted parameters for the Poisson distribution are difficult to learn from the overhead imagery, such as the label *stop sign*.

4.4.4 Multi-Attribute Region Search

We find the geolocation based on the output images that give a high score for a search query attribute. For example, In Figure 4.5, images with the highest score for *car* (an ImageNet label) are given, with increasing score for *park* (a Places label) from left to right. While each image is centered around a large road, they transition from industrial to rural as the *park* score increases.

4.5 Conclusion

We created a location-dependent model of geo-place understanding conditioned on overhead imagery. We show how our model can be used to generate maps at varying spatial scales. In the future, we will extend this work to include time, because what you can expect to see and experience at a location are highly dependent on time.

Chapter 5

Learning Dynamic Maps of Visual Appearance

5.1 Introduction

Through experience, humans develop an understanding of the relationship between time, location, and the visual appearance of a scene. We learn to know, for example, when to expect it to be dark, whether a scene will likely contain snow, and where to stand for the best sunset. While this *common sense*, geo-visual understanding has many practical uses, relatively little research in the computer vision community has explored the problem. In this work, we present an important step towards the construction of a computational model of geo-visual understanding. Potential uses of our framework include verifying the integrity of image metadata, geolocating images, providing advice to photographers in search of a beautiful sunset, and enabling further studies of the relationship between the visual environment and health [53].

The field of computer vision has traditionally focused on the problem of extracting visual attributes of scenes captured from a human perspective. For example, categorizing the scene type [79], estimating weather conditions [33], or predicting demographic properties [19]. Recently, there has been a significant interest in the problem of image geolocalization, i.e., estimating the geographic location of the camera, or an object in the scene, given visual attributes extracted from the image. Solving this problem requires two elements: the ability to extract the visual attributes and an understanding of the geospatial distribution of these visual attributes. This naturally leads to the problem of image-driven mapping, in which we extract visual attributes from images with known geolocation and use these to construct a map of visual appearance.

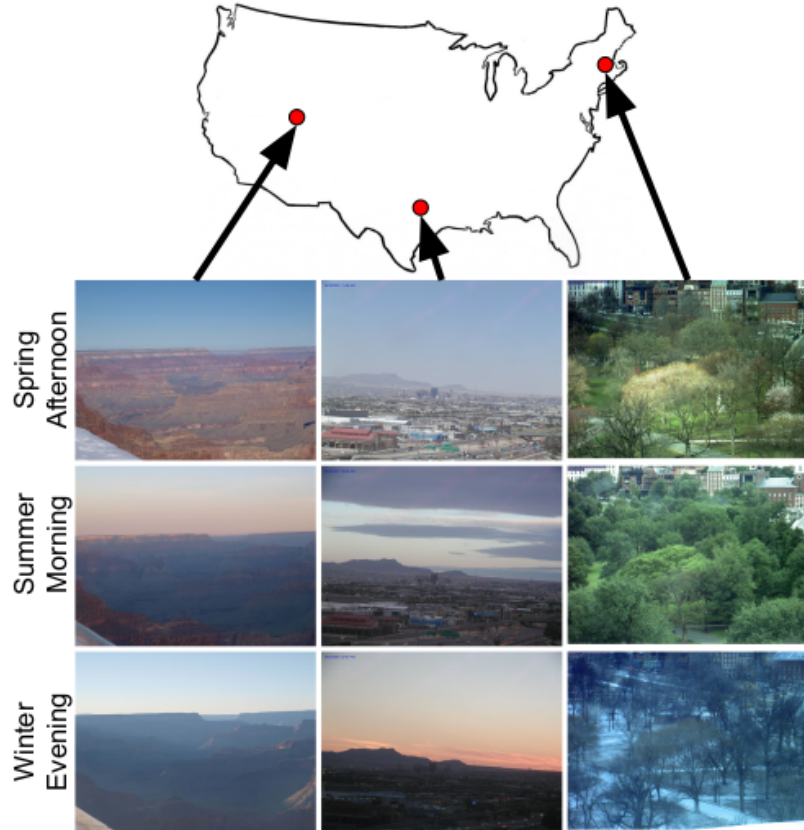


Figure 5.1: Visual appearance changes dramatically due to differences in location and time. Our work takes advantage of sparsely distributed, ground-level image data, with associated location and time metadata, in conjunction with overhead imagery to construct dynamic maps of visual appearance attributes.

A map can be thought of as a probability distribution, conditioned on geographic location, over an attribute; typical attributes include land cover, land use, elevation, or place name. In this work, we propose to construct a map of visual attributes of the world. These attributes change for many reasons, including seasonal changes in plants and illumination changes due to the diurnal cycle. Therefore, we propose to build a dynamic map, which is a probability distribution conditioned on geographic location and time. The visual attributes of a ground-level image, such as those shown in Figure 5.1, are samples from this distribution.

We collect a large number of GPS-tagged and time-stamped images, extract visual features, and then learn to predict distributions over these features based on when and where the picture was captured. Directly predicting these features from geographic location alone is difficult because of the complexity of the distribution. We find that including the overhead image as a conditioning variable results in significantly better predictions. This is a promising approach because many features that relate the appearance of a place are

visible from above and high-resolution overhead imagery is available across the globe and is updated at increasing frequencies, some providers even promise daily updates [47].

In our work, we focus primarily on two visual attributes: the scene category [79], such as whether the image views an attic or a zoo, and transient attributes [33], which consists of time-varying properties such as sunny and foggy. We selected these because they are well known, easy to understand, and have very different spatiotemporal characteristics. The former is relatively stable over time, but can change rapidly with respect to location, especially in urban areas. The latter has regular, dramatic changes throughout the day and with respect to the season.

Contribution The key contribution of this work is a novel approach for image-driven mapping. Our approach has several advantages: it does not require any manually annotated training data; it can model differences in visual attributes at large and small spatial scales; it captures spatiotemporal trends, but does not require overhead imagery at every time; and is extendable to a wide range of visual attributes. We demonstrate the effectiveness of our approach through evaluation on large datasets for several different tasks. In each case, our model, which combines overhead imagery and metadata, is superior.

5.2 Related Work

A significant amount of research has sought to extract meaningful visual attributes from ground-level imagery. Dubey et al. [14] describe methods for quantifying the urban perception of safety, liveliness, wealth, and more. Laffont et al. [33] estimate attributes describing scene appearance, such as if it is sunny or foggy. Other examples include interpreting weather conditions [41], estimating the local temperature [20], and predicting demographic properties [35]. When additional context about the imagery is known, such as the location or time of capture, these methods allow for characterizing properties of the underlying physical world.

Alternatively, studies have shown how additional image context can aid visual understanding. Tang et al. [61] integrate the location an image was captured as an input to a framework for image classification and demonstrate improved results. Zhai et al. [78] describe methods for learning geo-temporal features from location and time metadata. Wang et al. [65] use location information along with weather conditions to learn a feature representation for facial attribute classification. Unique from these works, we explore integrating location and time metadata with overhead imagery to produce high resolution dynamic maps of visual attributes.

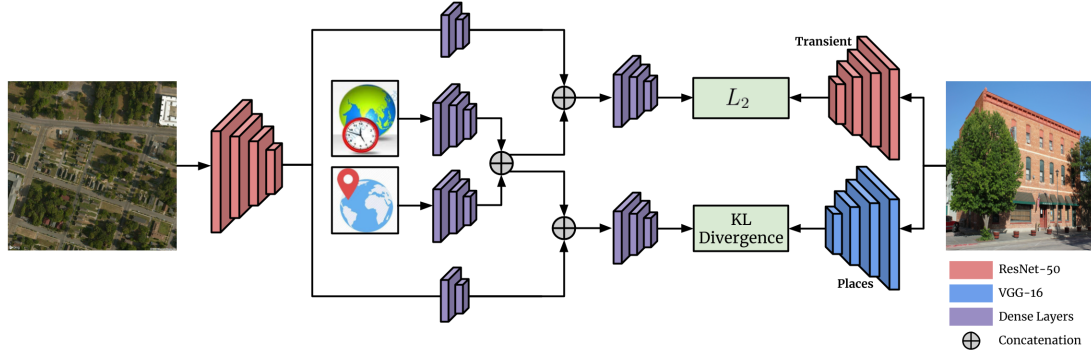


Figure 5.2: An overview of our network architecture, which includes the network we train to predict visual attributes (left) and the networks we use to estimate visual attributes (right).

Typically image-based methods for generating maps of visual attributes start by extracting information from large-scale geotagged image collections and then apply a form of smoothing, such as locally weighted averaging, to produce a map. For example, Lee et al. [35] estimate geo-informative attributes such as population density and elevation. Wang et al. [67] create maps of snowfall by automatically recognizing snowy scenes. Bessinger et al. [10] map the visual appearance of people across the globe using a large corpus of geotagged face imagery. In our work, we integrate location and time as additional context in the learning process and take advantage of overhead imagery to produce higher resolution maps. There are numerous other papers that use similar approaches [36, 66, 73].

Imagery captured from an overhead perspective has been shown to aid ground-level image understanding. Luo et al. [42] use overhead imagery as additional context to improve event recognition in ground-level photos. Overhead imagery has also been used to synthesize images from a ground-level perspective [13, 48, 77]. Another research area has explored using ground-level imagery as a form of weak supervision for learning about overhead imagery. For example, Zhai et al. [77] introduce a method for labelling overhead imagery by learning a semantic transformation between co-located ground-level and overhead images. Other studies have explored the joint understanding of the two modalities. For instance, learning a feature mapping between ground-level and overhead image viewpoints enables image localization in regions without nearby ground-level images [38, 39, 69, 70]. More recently, Workman et al. [71, 72] integrate nearby ground-level images to improve the prediction of semantic properties.

Most relevant to our work, overhead imagery has been used to improve image-driven mapping by enabling fine-grained higher resolution maps. Workman et al. [71] combine overhead and ground-level imagery to map the scenicness of a region. Salem et al. [52]

proposed a method for constructing maps of soundscapes by relating overhead image appearance to geolocated sounds. Our work extends this area by integrating location and time metadata as input to our framework.

5.3 Problem Definition

Our objective is to construct a map that represents the expected appearance at any geographic location and time. The expected appearance is defined using a set of visual attributes, which could be low level, such as a color histogram, or higher-level, such as the scene category. For a given visual attribute, a , such a map can be modeled as a conditional probability distribution, $P(a|t, l)$, given the time, t , and location, l , of the viewer.

We assume we are given a set of images, $\{I_i\}$, each with associated capture time, $\{t_i\}$, and geo-location metadata, $\{l_i\}$. We assume these images are captured from a ground-level viewpoint. Furthermore, we assume we have the ability to calculate, or estimate with sufficient accuracy, each visual attribute from all images. The computed visual attributes, $\{a_i\}$, can be considered samples from the probability distribution, $P(a|t, l)$, and used for model fitting.

5.4 Dynamic Visual Appearance Mapping

We present a general approach to constructing a visual appearance map that works well across a broad range of tasks. We refine the distribution described in the previous section by adding an additional conditioning variable. In particular, we define a conditional probability distribution, $P(a|t, l, I(l))$, where $I(l)$ is an overhead image of the geographic location, l . Critically, the overhead image is not dependent on the time. This means that an overhead image is not required for every timestamp, t , of interest. An overhead image *is* required for each location, but this is not a significant limitation given the wide availability of high-resolution satellite and aerial imagery.

5.4.1 Network Architecture Overview

Our model uses a mixture of convolutional and fully-connected neural networks to map from the conditioning variables to the parameters of distributions over a visual attribute, $P(a|F(t, l, I(l); \Theta))$, where Θ are the parameters of all neural networks. See Figure 5.2 for an overview of our complete architecture, which simultaneously predicts two visual attributes. From the left, we first construct a feature embedding for each conditioning vari-

able using a set of *context* neural networks. We combine these context features to predict the visual attributes using a per-attribute, *estimator* network. From the right, a set of pre-trained networks extract visual attributes from the ground-level images. These networks are only used for extracting visual attributes, which are used to train the context and estimator networks. The attribute estimation networks are not trained in our framework.

5.4.2 Network Architecture Details

We have defined a macro-architecture for training a dynamic visual appearance map. In the remainder of this section, we define the specific neural network architectures and hyper-parameters we used.

Visual Attributes We focus on two visual attributes: *Places* [79], which is a categorical distribution over 3.65×10^2 scene categories, and *Transient* [33], which is a multi-label attribute with 4.0×10^1 values that each reflect the degree of presence of different time-varying attributes, such as sunny, cloudy, or gloomy. To extract the *Places* attributes, we use the pre-trained VGG-16 [57] network. To extract the transient attributes, we use a ResNet-50 [27] model that we trained using the original Transient Attributes Dataset [33].

Context Networks The context networks encode every conditioning variable, i.e., time, geographic location, and overhead image, to a 1.28×10^2 -dimensional feature vector. For the time and geolocation inputs, we use two similar encoding networks, consisting of three fully connected layers each. The first layer contains 2.56×10^2 neurons, the second has 5.12×10^2 neurons, and the third 1.28×10^2 neurons. The geographic location is represented in earth-centered earth-fixed coordinates, scaled to the range $[-1, 1]$. The time is factored into two components: the month of the year and the hour of the day. Each is scaled to the range $[-1, 1]$. For the overhead image, we use a ResNet-50 model to extract the 2.048×10^3 -dimensional feature vector from the last global average pooling layer. This feature is passed to a per-attribute head. Each head consists of two fully connected layers that are randomly initialized using the Xavier scheme [21]. The layers of each head have 2.56×10^2 and 1.28×10^2 neurons, respectively.

Estimator Networks For each visual attribute, there is a separate estimator network that directly predicts the visual attribute. These consist of fully-connected layers, each with a ReLU activation. The input for these is the concatenation of the outputs of the encoding networks. For each, the first two layers contain 256 and 512 neurons, respectively. The third layer represents the output, with the number of neurons depending on the visual attribute. In this case, there are 365 output neurons for the *Places* estimator, with a *softmax* activation, and 40 for the *Transient* estimator, with a *sigmoid* activation.

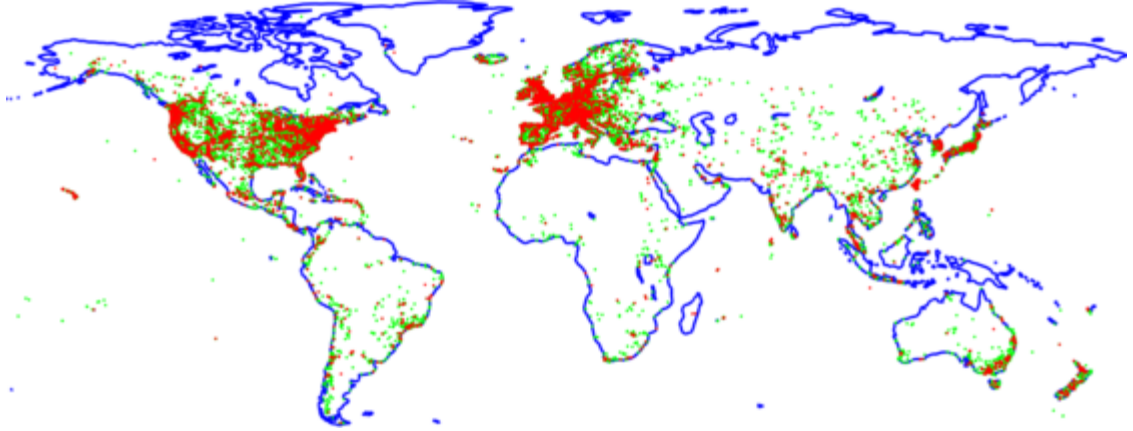


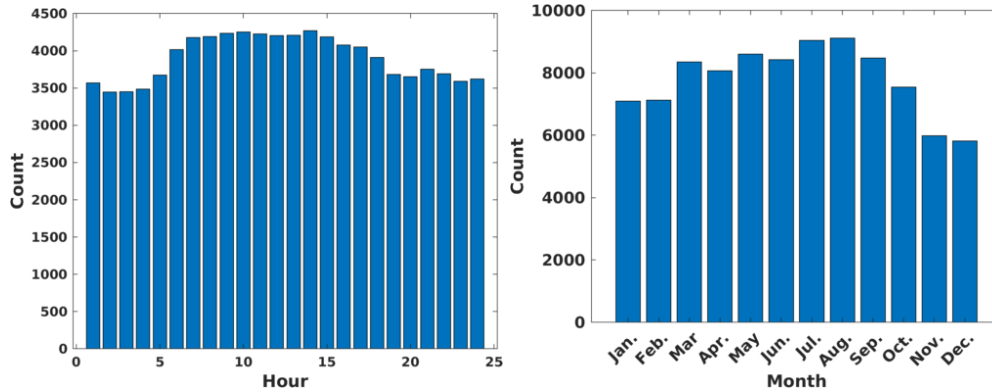
Figure 5.3: The spatial distribution of the dataset. The green (red) dots represent the training (testing) data.

5.4.3 Implementation Details

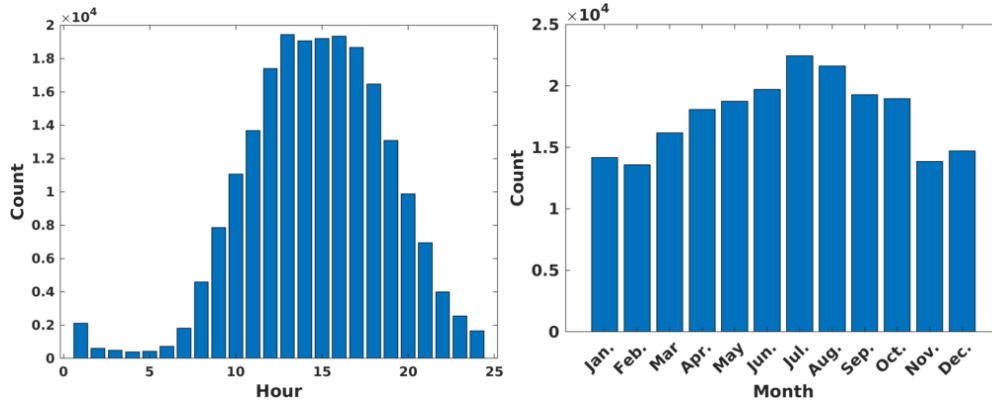
We jointly optimize all estimator and context networks with losses that reflect the quality of our prediction of the visual attributes extracted from ground-level images, $\{I_i\}$. For the *Places* estimator, the loss function is the KL divergence between attributes estimated from the ground-level image and the network output. For the *Transient* head, the loss function is the L_2 distance between the attribute and the network output. These losses are optimized separately using Adam [31] with mini-batches of size 3.2×10^1 . We applied L_2 regularization with scale 5×10^{-4} and trained all models for 1.0×10^1 epochs with learning rate 1×10^{-3} .

All networks were implemented using TensorFlow [2] and will be shared with the community. Input images are resized to 224×224 and scaled to $[-1, 1]$. We pre-trained the overhead context network to directly predict *Places* and *Imagenet* categories of co-located ground-level images, minimizing the KL divergence for each attribute. The weights are then frozen and only the added attribute-specific heads are trainable.

For extracting transient attributes from the ground-level images, we use the dataset introduced by the authors [33] and train a ResNet-50 network, minimizing the L_2 distance. The weights were initialized randomly using the Xavier scheme. We trained the network using Adam [31] until convergence with learning rate 0.001 and batch size 64. The resulting model achieves a mean squared error (MSE) of 3.04%. This improves upon the approach from the original authors, which used hand-engineered features and achieved an MSE of 4.2%.



(a) Webcam Images



(b) Cell Phone Images

Figure 5.4: The temporal distribution of the dataset.

5.5 Evaluation

We evaluated our approach, both quantitatively and qualitatively, on a variety of tasks.

5.5.1 Dataset

We built an evaluation dataset by combining images from two sources. The first source is a set of images from the Archive of Many Outdoor Scenes (AMOS) [29], a collection of over a billion images captured from public outdoor webcams around the world. The subset [44] includes images captured between the years 2.013×10^3 and 2.014×10^3 , from 5.0×10^1 webcams, totaling 9.8633×10^4 images. Each image is associated with the location of the webcam and a timestamp (UTC) indicating when the image was captured. The second source is a subset of the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) [62]. This subset [78] contains only geotagged outdoor images, with time stamps, from smart phones. We combined both datasets to form a hybrid dataset containing

Table 5.1: Comparing performance on the visual attribute prediction task.

Context	Transient (MSE)	Places (KL)
<i>loc</i>	1.468	4.160
<i>time</i>	1.486	4.550
<i>image</i>	1.398	3.652
<i>time+loc</i>	1.239	3.897
<i>image+loc</i>	1.375	3.527
<i>image+time</i>	1.194	3.402
<i>image+time+loc</i>	1.159	3.300

3.05011×10^5 images, 2.5000×10^4 of which are held out for testing. For each image, we also downloaded an orthorectified overhead image from Bing Maps (800×800 , 0.60 meters/pixel), centered on the geographic location. Figure 5.3 shows the spatial distribution of the training (green dots) and testing images (red dots). Visual analysis of the distribution reveals that the images are captured from all over the world, with more images from Europe and the United States. Further, examining the capture time associated with each image shows that the images cover a wide range of times. In Figure 5.4 (top), the distribution over month and hour for webcam images is essentially uniform, which is as anticipated because webcams are always on, capturing imagery at a standard interval. Figure 5.4 (bottom) shows the same visualization for cell phone images.

5.5.2 Baselines

For comparison, we trained several variants of our full model, *image+time+loc*. For each, we omit either one or two of the conditioning variables but keep all other aspects the same. We use the same training data, training approach, and micro-architectures. In total, we trained six baseline models: *loc*, *image*, *time*, *time+loc*, *image+loc*, and *image+time*.

5.5.3 Prediction Accuracy

Using the testing set, we evaluate the prediction quality of our method and all baseline methods. Table 5.1 shows the errors for all approaches on both visual attributes. We find that our method has lower average error. However, the ranking of baseline models changes depending on the visual attribute. For example, the predictions for the *image+loc* model are relatively worse for the *Transient* attribute than the *Places* attribute. This makes sense because the former is highly dependent on when an image was captured and the latter is more stable over time. We also note the significant improvement, for both attributes,

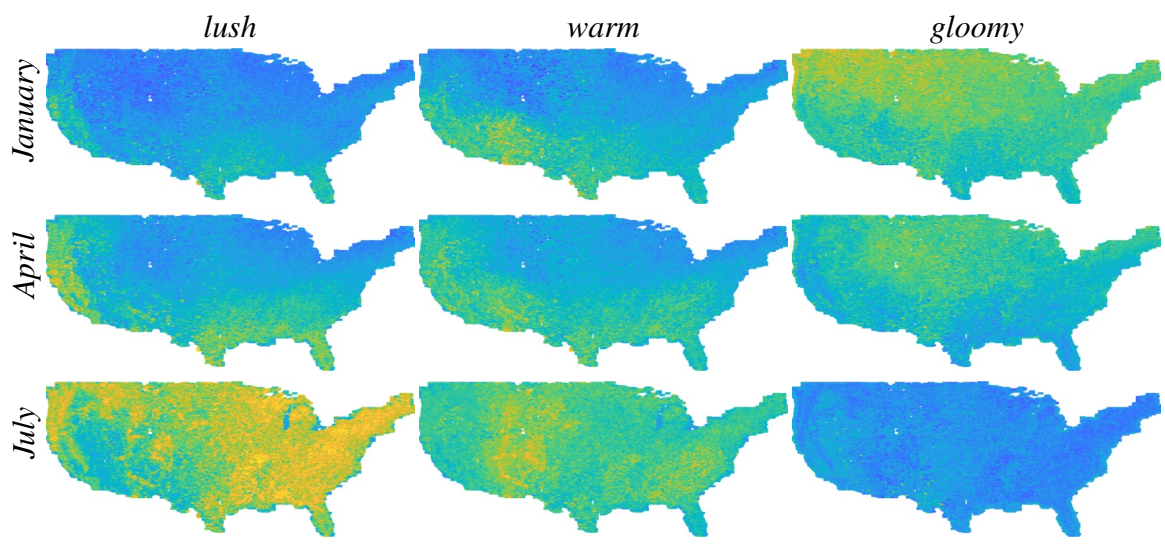


Figure 5.5: Dynamic visual attribute maps for different transient attributes and months. In each, yellow (blue) corresponds to a higher (lower) value for the corresponding attribute. Each attribute exhibits unique spatial and temporal patterns, which closely match the authors' personal travel experiences.

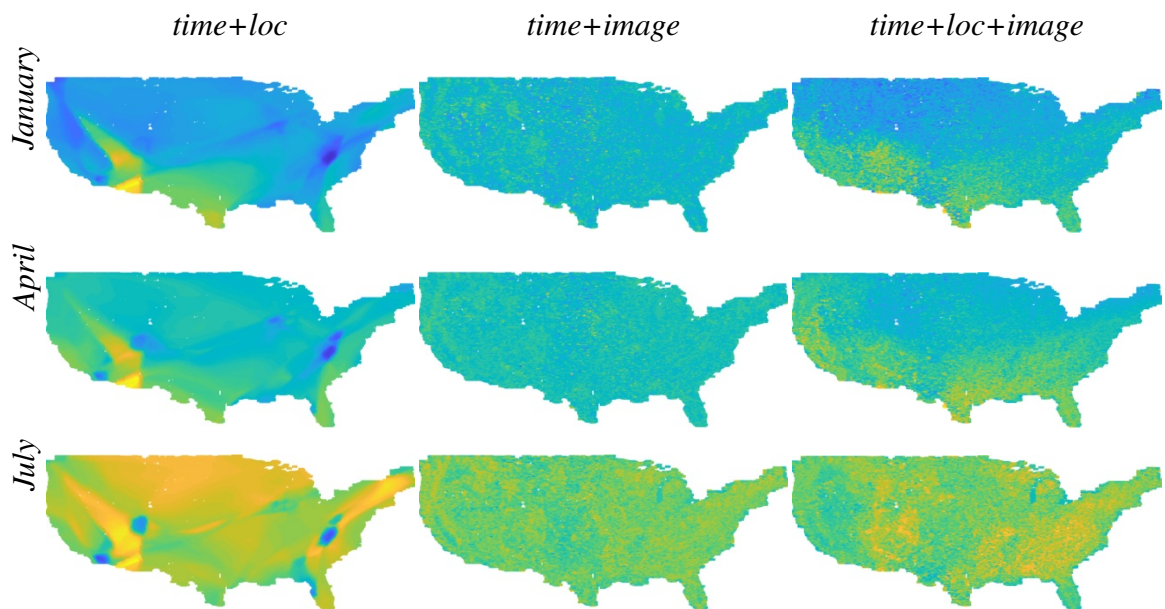


Figure 5.6: Dynamic visual attribute maps for different methods on the transient attribute *sunny*.

obtained by including overhead imagery in the model. See, for example, the *time+loc* model is significantly worse than our full model.

5.5.4 Examples of Visual Attribute Maps

Figure 5.5 shows several example attribute maps rendered from our model. To construct these maps, we use overhead images from the CVUSA dataset [70] which contains overhead imagery across the continental United States. Specifically, we use a subset of 4.88243×10^5 overhead images associated with the Flickr images in the dataset. For each overhead image, we compute visual attributes using our full model, *image+time+loc*. We specify the time of day as 4pm, and vary the month.

The trends we observe are in line with our expectations. For example, for the transient attribute *lush*, which refers to vegetation growing, January has low values (blue) in the northernmost regions. However, the highest estimates (yellow) include regions like Florida and California. The lushness estimate progressively increases from January through April, achieving its highest value in July. Similarly, the *warm* attribute is highest in the southwest during both winter and spring, but reaches higher overall values in the summer months. Meanwhile, the *gloomy* attribute is highest during winter, with a bias towards the Pacific Northwest, and decreases during the summer.

We show additional dynamic attribute maps rendered from our model versus several baselines. For this experiment, we only compared against baselines that incorporate time as an input. For each, we specified the time of day as 4pm, and varied the month. We show example maps for the attribute *sunny* in Figure 5.6. The first column shows the *time+loc* model, the middle column the *time+image* model, and the last column our *time+loc+image* model.

Figure 5.7 shows how the visual attribute varies over the day for different attributes, locations, and months. The first example, which is located in Florida, shows that our model has captured the difference in day length between winter and summer.

5.5.5 Application: Image Retrieval

In Figure 5.8, we show how our model can be used to identify a set of ground-level images that would be likely to be observed at a given location and time. Our process is as follows: for a given overhead image, location, and input time, we first extract the *Places* and *Transient* attributes. Then we compare these attributes against visual attributes extracted from a set of ground level images using KL divergence for the *Places* attribute and L_2 distance for the *Transient* attribute. We produce a final scoring by adding the individual

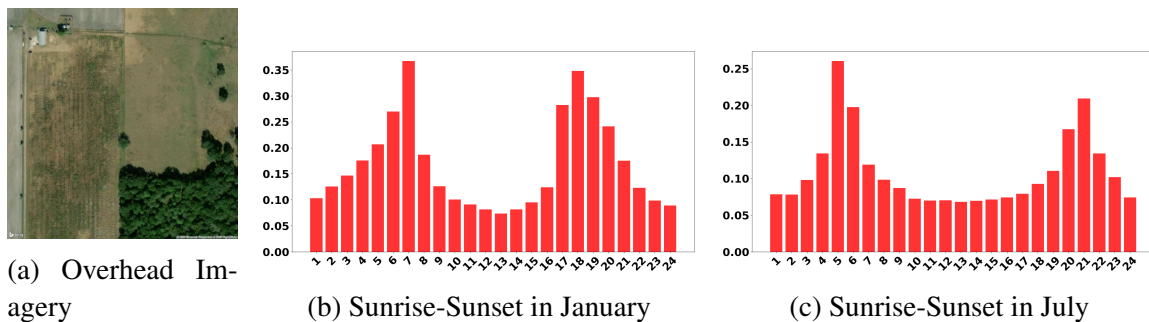


Figure 5.7: For a given location and corresponding overhead image, predictions of the *sunrise-sunset* attribute at different hours for two different months. This highlights that our model has learned that days are longer during the summer.

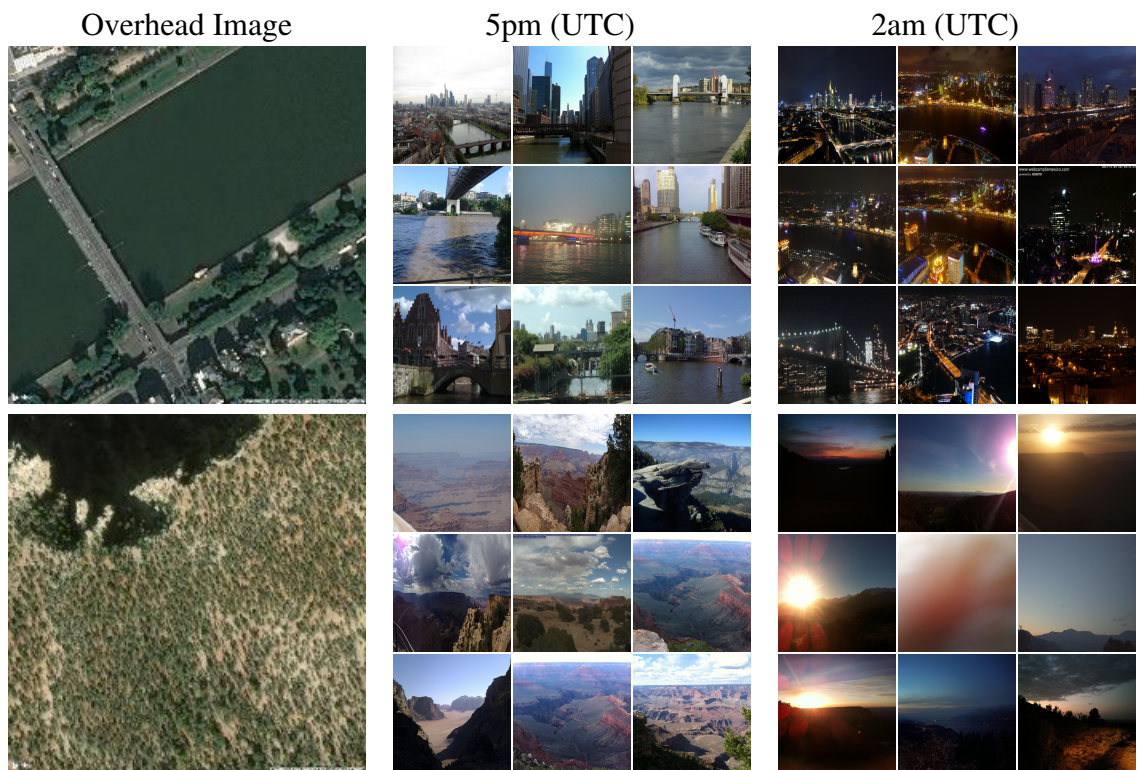


Figure 5.8: For each overhead image, we predict the visual attributes using our full model and compute the average distance between them and those of the ground-level images in the test set. (left) The overhead images of two query locations. The closest images when using August at 5pm as input (middle) and when using August at 2am (right).

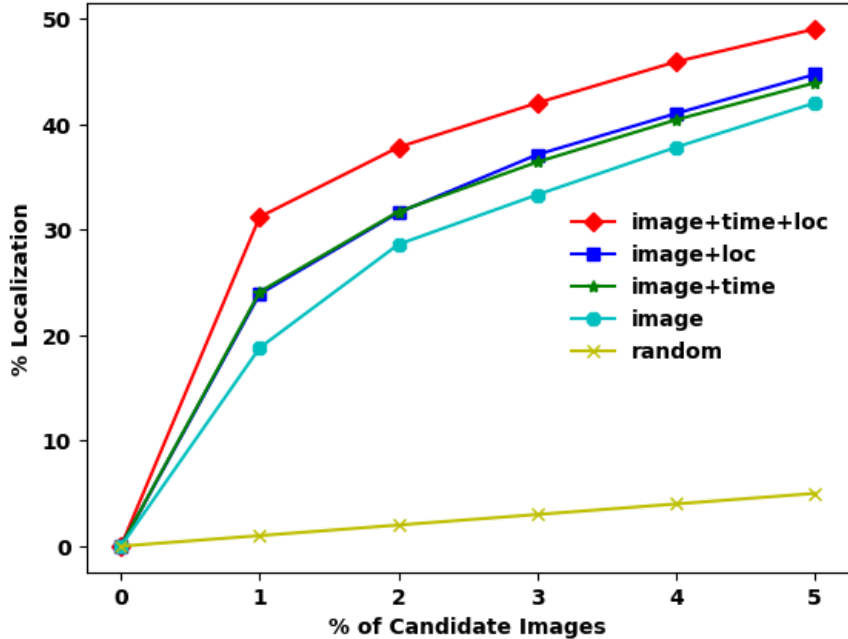


Figure 5.9: Localization accuracy as a function of candidate images searched. Our approach, *image+time+loc*, outperforms all baselines.

scores together. Based on the overall scoring, we identify the most similar ground-level images. We observe that the ground-level images contain both the expected scene type as well as represent the appropriate time of day. For example, the top left overhead image contains a bridge and the closest ground-level images are visually consistent at both input timestamps.

5.5.6 Application: Image Localization

We evaluated the accuracy of our models on the task of image geolocation, using a set of 1.000×10^3 ground-level query images randomly sampled from the test set. To localize an image, we first extract its visual attributes. Then, we predict the visual attributes for all 1.000×10^3 overhead images (using the associated location and the capture time of the ground-level image). Similar to our image retrieval experiment, we compute a score between the ground-level and overhead *Places* attributes using KL divergence. Finally, we compute the rank of the correct location and repeat this process for all query images.

Figure 5.9 visualizes the results. The x -axis represents the percentage of candidate locations with lower distances than the correct location and the y -axis represents the cumulative distribution. For a given threshold, a higher percentage localized is better. This experiment shows that our full model results in better localization accuracy and that using

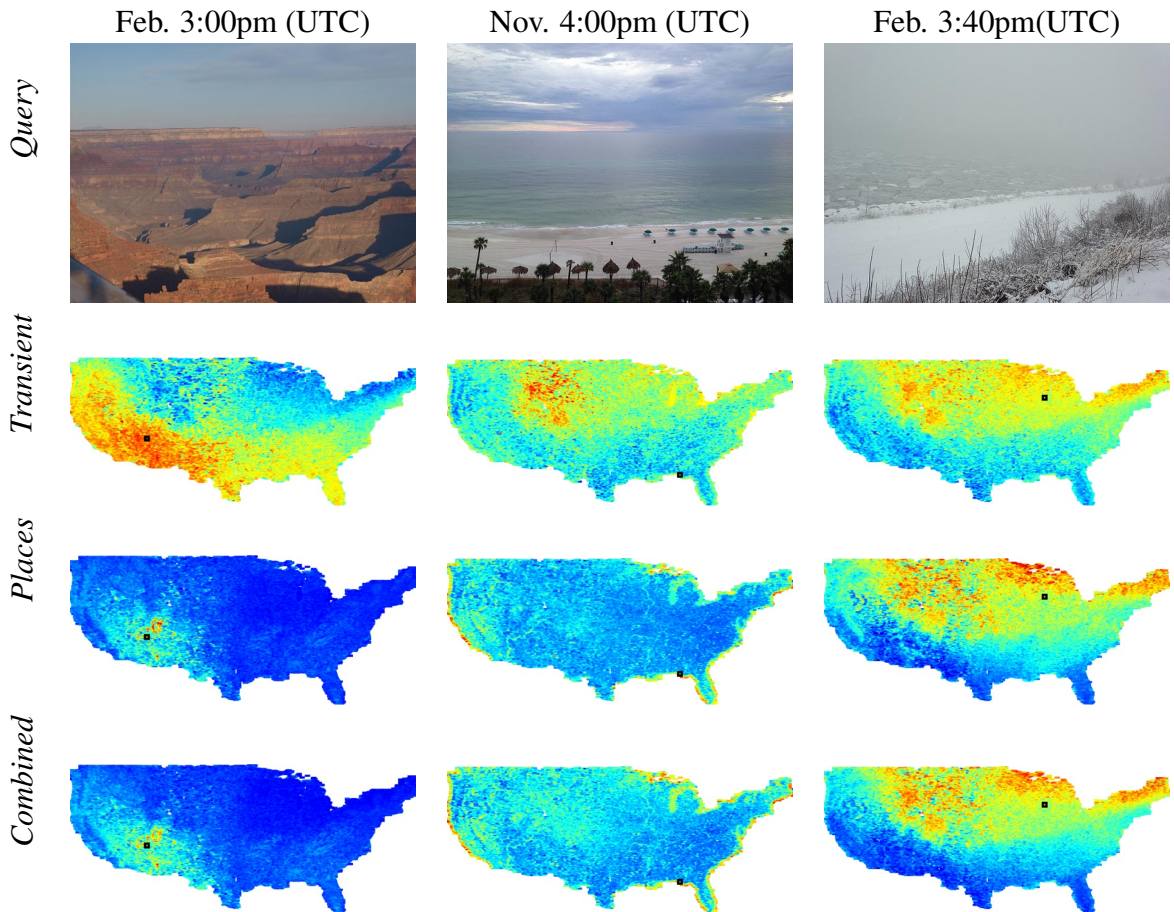


Figure 5.10: Given a query ground-level image (top), we show localization results (bottom) for different scoring strategies, visualized as a heatmap. Red (blue) represents a higher (lower) likelihood that the image was captured at that location.

only the imagery gives the worst performance. This highlights that our model is better able to capture the dynamic distribution of visual attributes.

In Figure 5.10 we show some qualitative results from our model for the localization task, using the method described in above . To summarize, we extracted the visual attributes of a query image and compared them against the visual attributes of an overhead image reference database, computed using the timestamp of the query image. For this experiment, we used 4.88224×10^5 overhead images from CVUSA as our reference database. The heatmap represents the likelihood that an image was captured at a specific location, where red (blue) is more (less) likely. Additionally, we compare the different scoring strategies on each row. Similar to our quantitative results, the *Places* attribute performs best for this task. This make sense as the *Places* attribute describes static scene elements that are more location dependent.

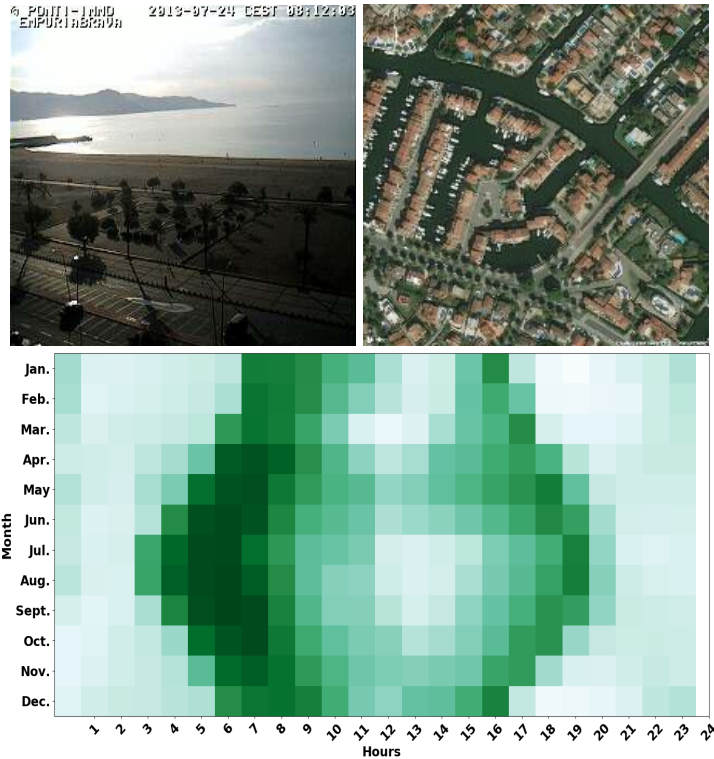


Figure 5.11: An example highlighting temporal patterns learned by our model. For each example, we show the original image and the overhead image of its location. For every possible hour and month, we use our full model to predict the visual attribute. The heatmap shows the distance between the true and predicted visual attributes, with dark green (white) representing smaller (larger) distances. In the top example, there are two narrow bands of small distances, centered around dawn and dusk. In the top example, we see small distances during the nighttime hours.

5.5.7 Application: Metadata Verification

Recent concern about “fake news” has led to a significant interest in verifying that imagery is real and un-manipulated. Early work on this problem focused on low-level image statistics [12, 15], but this approach is unable to detect the falsification of image metadata. Matzen and Snavely [43] introduce an approach for finding anomalous timestamps, but their method is based on visual correspondences and requires overlapping imagery. Recent work has begun to look at this problem more thoroughly, with new datasets [24] and proposals for comprehensive systems [11]. However, no work provides a complete model for mapping dynamic visual attributes which is necessary for detecting time/location metadata falsification.

We demonstrate that our model has the potential to be useful for this problem. Our approach is to take the purported metadata and make visual attribute predictions using our

full model. We can then compute the distance between the predictions and the actual visual attributes to see if they seem plausible. Figure 5.11 shows example heatmap of distances, using the *Transient* attribute, for a wide range of possible times. These highlight that many capture times aren't plausible. While this approach does not solve the problem of detecting metadata falsification, it demonstrates that our model has learned the essential elements that could be used to build such a system.

5.6 Discussion

Our model combines overhead imagery, time, and geographic location to predict visual attributes. We have demonstrated the superiority of this combination, but think there are several questions that naturally arise when considering our model. Here we provide answers, which we think are supported by the evaluation. (1) *Why do we need overhead imagery when it just depends on the location?* If our model was only dependent on geographic location, then we would need to learn a mapping between geographic location and the visual attribute. Consider something as simple as, “does this geographic location contain a road?”. This would be a very complicated function to approximate using a neural network and we have seen that it does not work well. In contrast, it is relatively easy to estimate this type of information from the overhead imagery. (2) *Why do we need to include geographic location if we have overhead imagery?* We think it makes it easier to learn larger scale trends, especially those that relate to time. For example, the relationship between day length and latitude. If we didn't include latitude we would have to estimate it from the overhead imagery, which would likely be highly uncertain. (3) *Why don't we need an overhead image for each time?* The overhead image provides information about the type of place. This is unlike a satellite weather map, which would tell us what the conditions are at a particular time. While we do lose some information, this is accounted for by including geographic location and time as additional context. In practice it is best if the overhead image is captured relatively close in time (within a few years) to account for major land use and land cover changes.

One of the limitations of this study is the reliance on social media imagery. This means that our visual appearance maps will exhibit biases about when people prefer to take pictures, or are willing to share pictures. For example, we are likely under-sampling cold and stormy weather conditions and oversampling sunsets. This is part of the motivation for incorporating imagery from the AMOS dataset. This, at least, doesn't have the same temporal bias because the webcams collect images on a regular basis, regardless of conditions. However, these are sparsely distributed spatially and, at least in our dataset, outnumbered

by the social media imagery. Despite this limitation we were still able to demonstrate effective learning and could overcome this problem as more data becomes available.

Another limitation is that our current approach cannot model longer-term, year-over-year trends in visual attributes. This results because our representation of time only reflects the month and time of day, not the year. Such a model, which we intend to explore in future work, could model changes in climate and land use.

5.7 Conclusion

We introduced a novel method for constructing dynamic visual attribute maps. In several large scale experiments, we demonstrated the practical usefulness of the model and highlighted the importance of including time, location, and an overhead image of the location as conditioning variables. Such a model has many potential uses, including image localization, trip planning, and media forensics. In future work, we plan to explore various low-level architectural choices, such as the model for extracting overhead image features, incorporate additional visual attributes, and train with a larger dataset.

Chapter 6

Learning to Map Soundscapes

6.1 Introduction

The visual appearance of a place and its soundscape, the totality of sounds one hears in a location, are inextricably linked. For example, in an urban environment, such as on a busy street corner, you can expect to hear honking, people talking, and, potentially, a siren. In contrast, in a rural environment, such as a forest, you could expect to hear animals chattering, leaves rustling, and perhaps the sound of rushing water. Given a photograph, humans have the ability to imagine the sounds they might hear in that moment.

Studies have shown that environmental noise affects social behavior [59], among other things. Basner et al. [8] summarize research related to noise exposure, including auditory and non-auditory health effects such as reduced cognitive performance and sleep disturbance. Models capturing the relationship between sound and specific locations could be used, for example, to help people decide where to live, or where to place sound barriers.

The objective of our work is to develop methods for understanding the types of sounds that could be heard at a specific geographic location. Several recent works have taken advantage of the synergy between sound and visual appearance to learn better representations. Aytar et al. [5] leverage two million unlabeled videos to learn a state-of-the-art sound representation for acoustic classification. Owens et al. [45] incorporate ambient sounds as a supervisory signal in order to learn visual representations. Most similar to our work, Aiello et al. [3] proposed a method for constructing sound maps by using sound-related image tags on a large set of geo-referenced ground-level imagery. This method requires high-quality image tags, which aren't always available, and performs poorly when ground-level imagery is sparsely distributed, such as away from major tourist landmarks.

We take a different approach and explore the problem of generating a location-dependent sound model. Our approach builds upon recent advances in both ground-level

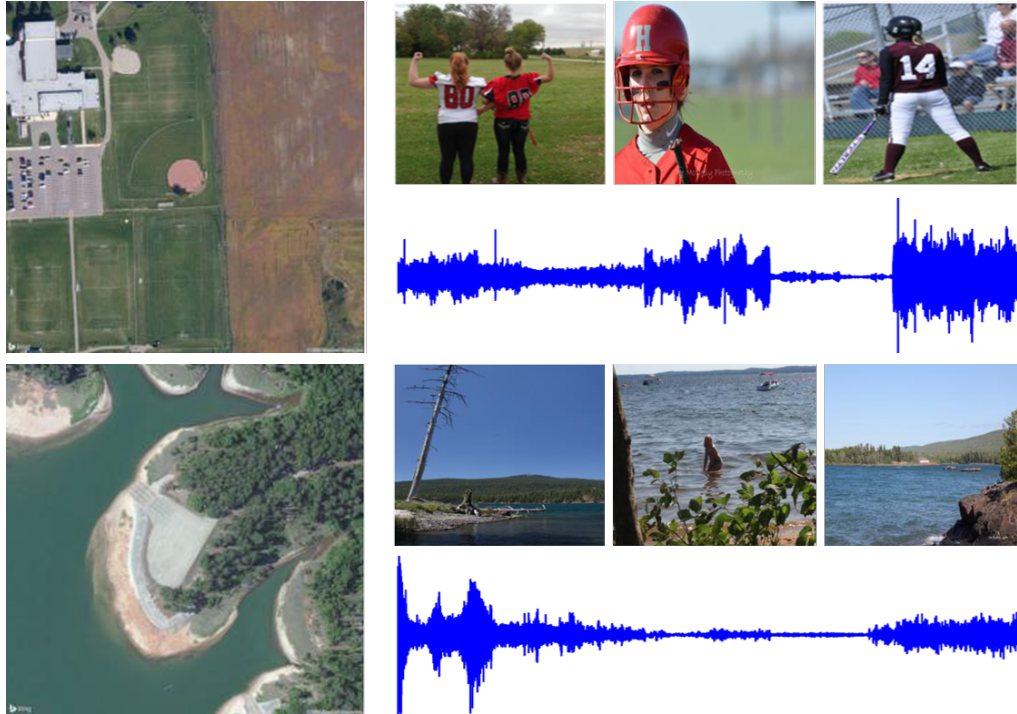


Figure 6.1: We propose a multimodal approach for relating overhead image appearance with sounds in order to map soundscapes. (left) Overhead image; (right) Similar ground-level images and sounds output by our method.

and overhead image understanding. A key element of our approach is that we learn a joint feature representation between sound, ground-level, and overhead image appearance (Figure 6.1). A unique advantage of our approach is that it enables us to generate a location-dependent sound map (or an aural atlas) using only overhead imagery, which is available at most locations.

6.2 Cross-View Aural Mapping

The objective of our work is to construct a map of the aural environment, which we represent as a conditional probability distribution, $P(s|l)$, where l is the geographic location and s represents the sound. In this work, we explore a novel approach, conditioning our aural map on the overhead imagery of a location, $P(s|I(l))$, where $I(l)$ is an overhead image of location l . This is a promising approach because many visual features that relate sound to location are visible from above. Furthermore, high-resolution overhead imagery is available across the globe and is updated frequently. We further factorize the distribution as:

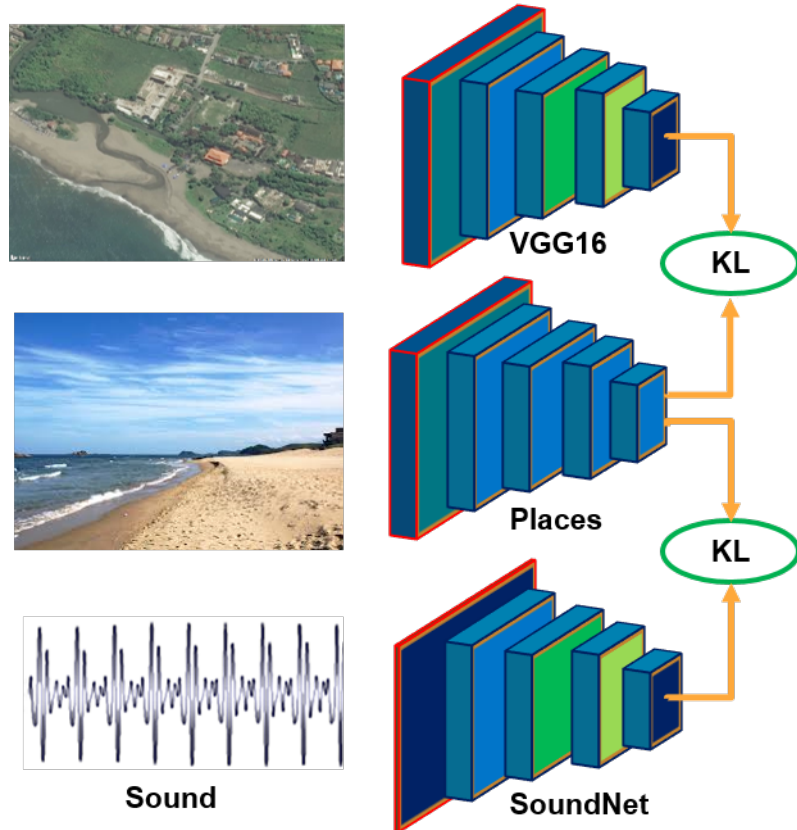


Figure 6.2: An overview of our network architecture.

$$P(s|I(l)) = \sum_c P(s|c)P(c|I(l)),$$

where c represents a cluster of related sounds.

Our approach consists of three phases: learning a suitable feature space, clustering sounds, and learning to predict a distribution over sound clusters from the overhead imagery using a convolutional neural network.

6.2.1 Cross-View Sound Dataset

To support our work, we constructed a dataset of geo-tagged sounds and co-located overhead images, which we refer to as the Cross-View Sound (CVS) dataset. We collected 23,308 geo-tagged audio files from FreeSound¹, a popular crowd-sourced repository. For each audio file, we downloaded the corresponding overhead image from Bing Maps (scale 0.60 m/pixel). Analysis of the geolocation associated with the sound files reveals that the sounds are recorded from around the world, with more sounds recorded in Europe and U.S

¹<https://freesound.org>



Figure 6.3: The distribution of the collected audio files in our CVS dataset.

than other parts of the world as can be seen in Figure 6.3. Further, examining the tags associated with the sound files shows that the sounds cover a wide range of human and natural aspects (Figure 6.4). For our experiments, we filtered out sounds that were shorter than 2 seconds and sounds for which there was no overhead imagery available at the selected scale. This results in 15,773 sounds and their corresponding overhead images.

6.2.2 Learning a Shared Feature Space

In this phase, our goal is to learn a shared feature representation that is suitable for our task. Specifically, we want a feature representation that can jointly describe audio and overhead imagery. To do this, we propose a convolutional neural network (CNN) architecture that relates sounds with co-located overhead images. Our approach builds on recent work that targets these two subproblems individually. An overview of our architecture is shown in Figure 6.2.

To extract audio features, we use SoundNet [5], a deep convolutional architecture for sound recognition, trained by transferring knowledge from existing visual recognition networks. To train SoundNet, images from unlabeled videos are passed through the *Places* network [79] (while different *Places* models are available, we always refer to the one used to train SoundNet), and the output distributions are used as the target label for a network that takes as input the corresponding audio file. Given an audio file, the output of Sound-



Figure 6.4: A word cloud for the tags associated with the sounds in the CVS dataset.

Net is a distribution over 401 visual scene categories. The resulting network performs remarkably well, despite being trained without any manually annotated audio files.

To learn the overhead image feature representation, we use a multimodal training approach similar to SoundNet and Workman et al. [70]. Specifically, we learn to predict a distribution over ground-level scene categories from overhead imagery. Each ground-level image is labeled using the *Places* network [79], generating a distribution over 401 scene categories. We then train a VGG-16 [56] network to predict these distributions using only the overhead image, minimizing the KL-divergence. We trained the network on the CVUSA dataset [70], which contains approximately 1.5 million geo-tagged pairs of overhead and ground-level images. The network is initialized to the weights of the *Places* network, and optimized using *Adam* with learning rate of 0.001 for 5 epochs.

This process results in a shared feature representation that allows the direct comparison of three different modalities: audio, ground-level imagery, and overhead imagery. We could, for example, use an image retrieval approach to identify sounds related to an overhead image. The problem with this approach is that the sounds close to the overhead image in the feature space will all be similar, and therefore potentially not representative of the

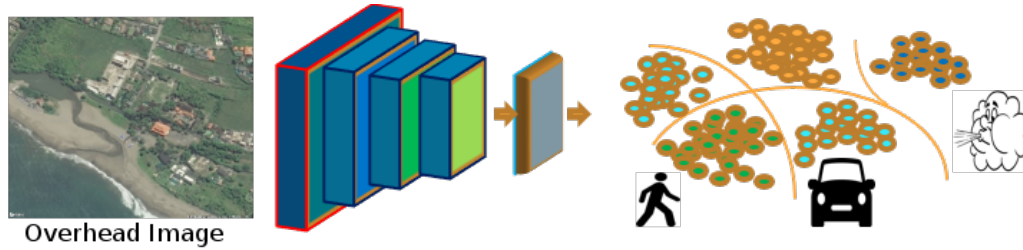


Figure 6.5: The model architecture for predicting a distribution over sound clusters from an overhead image.

diversity of sounds one could hear in a particular area. To overcome this, we introduce a clustering approach to group sounds, which we then use to map soundscapes.

6.2.3 Clustering Sounds

We group the sounds into a discrete set of clusters using hierarchical clustering [9]. For a given image-sound pair, we extract the predicted distributions over the 401 scene categories for each modality and concatenate them to form an 802-dimensional vector. Then, this concatenated representations is used as input for clustering. The result of this process is a set of clusters $C = c_1, \dots, c_k$. Finally, we filter out small clusters (less than 500 sounds), leaving 10 clusters. In the following section, we describe our process for estimating the conditional distribution over sound clusters for a given location.

6.2.4 Predicting Sound Clusters from Overhead Imagery

We assign each sound to a unique cluster and treat the cluster assignment, c_i , as the label of a given location, l_i . For each location, we obtain the co-located overhead image, $I(l_i)$, and train a CNN to predict the sound cluster, c_i , from the image. We fine-tune the network described in Section 6.2.2, adding a fully connected layer at the end with ten outputs (Figure 6.5). We minimize the cross-entropy loss using *Adam* with a learning rate of 0.001 for 20 epochs. We now have all of the components of our model and can use $P(c|I(l))$ to visualize soundscapes.

6.3 Experiments

We evaluated our approach both quantitatively and qualitatively using a TensorFlow [2] implementation. We begin with an analysis of the shared feature space.

Table 6.1: Quantitative performance of different networks.

Network	Precision	Recall	F_1 -score
<i>sound</i>	0.24	0.19	0.19
<i>joint</i>	0.51	0.34	0.36

6.3.1 How good is our feature space?

For a given overhead image, we extract the output distribution over scene categories and identify the closest sounds in CVS and the closest ground-level images in CVUSA, using KL-divergence. Several qualitative examples are shown in Figure 6.1. The leftmost column shows the overhead image and the right columns show the top three ground-level images above the top three sounds. For example, in the bottom row, the overhead image is of a lake, and the three closest ground-level images appear to be captured on or near a lake. The results are similar when listening to the closest sounds; in Figure 6.1 (top) the most similar sound contains people cheering. The predicted sounds, dataset, and more results will be made available online at <http://cs.uky.edu/~salem/audio-mapping/>.

6.3.2 What is the best way to cluster?

We compare our approach for clustering the sounds against a baseline approach using only sound features. As described in Section 6.2.4, we train two models on the two different clustering approaches. For evaluation, we split the CVS dataset into 90% training and 10% testing. The resulting test set contains 1,578 sounds and corresponding overhead images.

For a given overhead image, each model outputs a probability distribution over the sound clusters. The precision, recall and F_1 -score for these two models are shown in Table 6.1. The model that was trained on clusters generated from joint features achieved better performance compared to the model trained on sound features. The superior performance can be attributed to the fact that the clustering approach based on joint features takes into account the semantic relation between overhead images and the corresponding sounds. Figure 6.6 shows the output distributions over the 10 clusters for two test images.

6.3.3 Visualizing An Aural Atlas

Using the trained CNN model to predict a distribution over sound clusters from an overhead image enables us to construct sound maps at various spatial scales: block level, city level, and country level.

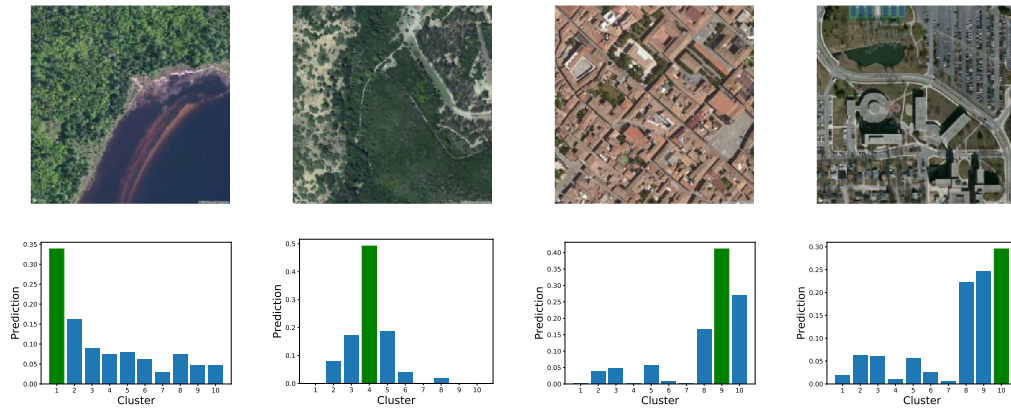


Figure 6.6: Our work explores the relationship between overhead image appearance and sound. Given an overhead image (top), our model outputs a distribution over sound clusters (bottom).

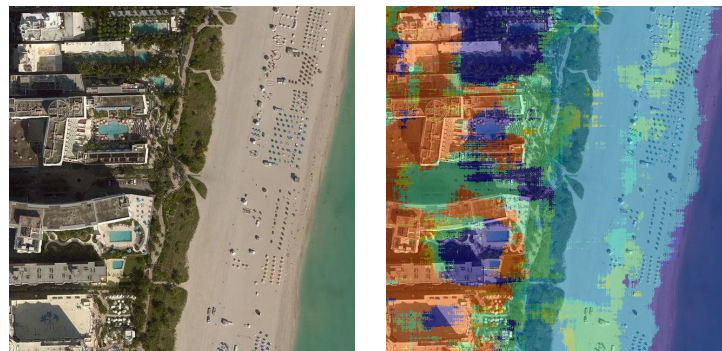


Figure 6.7: Block-level audio mapping: (left) An overhead image of a small geographical region on Miami beach. (right) A per-pixel labeling of sound clusters.

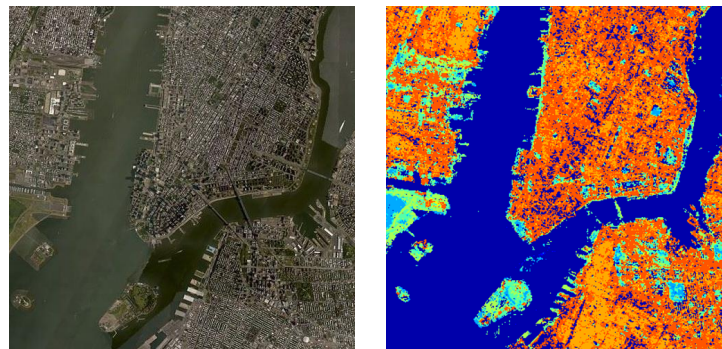


Figure 6.8: City-level audio mapping: (left) An overhead image covering New York City. (right) A per-pixel labeling of sound clusters.

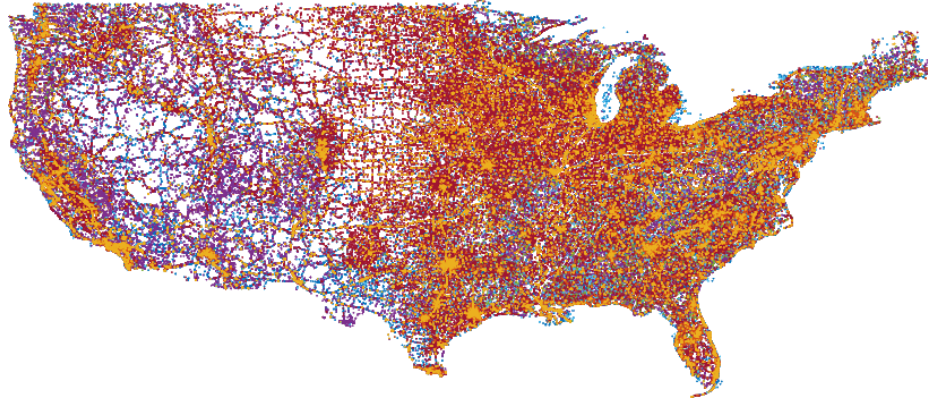


Figure 6.9: Country-level audio mapping: visualizing the sound clusters over USA. Gaps (white) are regions where the CVUSA dataset does not have imagery.

Block level: Consider the overhead image on the left of Figure 6.7 which contains beach, water, roads, and buildings. Clearly the sounds at these places would be different. For every pixel in the image, we downloaded the corresponding overhead image and used our network (Section 6.2.4) to predict the distribution over sound clusters. We show the results of our approach in Figure 6.7 (right), as a per-pixel labeling where the color represents the most likely sound cluster (e.g., blue = water-related sounds and orange = traffic sounds). The color coding is the same for the next two spatial scales.

City level: Here we apply the same technique to a larger geographic area. Figure 6.8 shows the aural atlas for a portion of New York City. Note how the majority of the urban areas are colored orange and the water areas are dark blue.

Country level: Finally, we demonstrate the results of our method at the country level. We used 500,000 overhead images randomly sampled from the CVUSA dataset and extracted the sound cluster prediction with our trained model. Figure 6.9 shows the results. Note the orange regions covering the major metropolitan areas.

6.4 Conclusion

We created a location-dependent model of sound conditioned on overhead imagery. We showed how our model could be used for sampling a set of sounds that you would hear at a given location and to generate maps of soundscapes at varying spatial scales. To the best of our knowledge, our work is the first to model the relationship between overhead imagery and sound. In the future, we will extend our work to include time, as the sounds you might hear at a location are highly time dependent.

Chapter 7

Discussion

The research in this dissertation investigated the effectiveness of utilizing the publicly available images for learning and mapping different ground level attributes from overhead imagery. The major challenge for such learning tasks is the lack of annotated overhead imagery for model training. We presented an important step towards the construction of a computational model of geo-visual understanding using overhead imagery. In particular, we proposed to learn and map different ground-level attributes from overhead-imagery without requiring any annotated data. Instead, we used pre-existing CNNs to extract categorical distributions of co-located ground-level images to provide a weak signal for model training. Based on a quantitative and qualitative analysis of our research, we note that overhead imagery can be very beneficial to improve the understanding and capturing the different changes over geolocation and time.

In Chapter 3 we introduced a general framework for mapping ground-level attributes from overhead imagery. The proposed approach is based on deep learning techniques and for training, there is no need for any annotated data. We achieve this by using well performed trained models on ground-level images to provide a weak signal for the model to train on overhead imagery. Our approach has the option to integrate metadata with overhead imagery to improve the modeling of the changes in dynamic attributes that change over geolocation and time.

In Chapter 4 we created a location-dependent model of geo-visual understanding conditioned on overhead imagery. Our proposed approach captures the uncertainty in the ground-level labels by modeling the distribution of ground-level image labels as samples from a Dirichlet distribution. We used a multi-task approach and predicted the parameters of prior distributions over three label spaces: scene categorization, image classification, and object detection. We also showed how our model can be used to generate maps at varying spatial scales.

In Chapter 5 we proposed a novel method to integrate the time and geolocation with the overhead imagery to capture the changes of different ground level attributes over geolocation and time. To the best of our knowledge, our work is the first to integrate dense overhead imagery with location and time metadata into a general framework for image-driven mapping. A key element of our method is that we use visual attributes that are present in ground-level images as a supervisory signal for model training, thus requiring no labeled overhead imagery. Through a large-scale evaluation on real data, we demonstrated the practical applications of the model and highlighted the importance of including time, location, and an overhead image of the location as conditioning variables. We found that combining overhead imagery with metadata results in more accurate predictions and better performance on a variety of tasks.

In Chapter 6 we studied the relation between the visual appearance of a place and its soundscape, and we proposed a model for predicting the types of sounds that are likely to be heard at a given geographic location. We showed how our model can be used for constructing an aural atlas, which captures the spatial distribution of soundscapes. In our approach, we built on previous work relating sound to ground-level imagery. However, we incorporated overhead imagery to overcome the limitations of sparsely distributed geo-tagged audio. Our trained model requires only the overhead imagery to construct an aural atlas of the region of interest. In this work, we constructed a dataset of geo-tagged sounds and co-located overhead images (CVS). This dataset is considered the first of its kind and it was made publicly available to the research community. To the best of our knowledge, our work is the first to model the relationship between overhead imagery and sound.

This dissertation proposed different approaches for mapping ground-level attributes from overhead imagery. We showed how our models can be used to generate maps at varying spatial scales. There are several possible future research directions for extending our work, including exploring various low-level architectural choices, such as the model for extracting overhead image features, incorporating additional visual attributes, exploring other probabilistic models to capture the uncertainty in the ground-level labels, and training with a larger dataset.

Finally, we hope that our work will inspire other researchers to continue this direction of research for better geo-visual understanding. We believe there is still room for improvement and future work that can be done to improve the modeling and mapping of our visual and auditory world.

Bibliography

- [1] <https://www.clarifai.com/technology>. vi, 7
- [2] 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. 30, 46
- [3] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. Chatty maps: constructing sound maps of urban areas from social media data. *Royal Society open science*, 3(3):150690, 2016. 41
- [4] 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. 3, 15
- [5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016. 11, 41, 44
- [6] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, 2014. 11
- [7] 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*, 2016. 3, 13
- [8] Mathias Basner, Wolfgang Babisch, Adrian Davis, Mark Brink, Charlotte Clark, Sabine Janssen, and Stephen Stansfeld. Auditory and non-auditory effects of noise on health. *The Lancet*, 383(9925):1325–1332, 2014. 41

- [9] Doug Beeferman and Adam Berger. Agglomerative clustering of a search engine query log. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000. 46
- [10] Zachary Bessinger, Chris Stauffer, and Nathan Jacobs. Who goes there?: approaches to mapping facial appearance diversity. In *24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016. 3, 14, 27
- [11] Aparna Bharati, Daniel Moreira, Joel Brogan, Patricia Hale, Kevin Bowyer, Patrick Flynn, Anderson Rocha, and Walter Scheirer. Beyond pixels: Image provenance analysis leveraging metadata. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 38
- [12] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017, 2012. 38
- [13] Xueqing Deng, Yi Zhu, and Shawn Newsam. What is it like down there?: generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018. 27
- [14] 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. 26
- [15] Hany Farid. Image forgery detection. *IEEE Signal processing magazine*, 26(2):16–25, 2009. 38
- [16] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE international conference on computer vision*, 2013. 10
- [17] Basura Fernando, Damien Muselet, Rahat Khan, and Tinne Tuytelaars. Color features for dating historical color images. In *IEEE International Conference on Image Processing*, 2014. 13
- [18] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *32nd International Conference on Machine Learning*, 2015. 10

- [19] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017. 3, 15, 24
- [20] 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. 26
- [21] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 29
- [22] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 10
- [23] Connor Greenwell, Scott Workman, and Nathan Jacobs. What goes where: Predicting object distributions from above. In *IEEE International Geoscience and Remote Sensing Symposium*, 2018. 17, 18
- [24] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *IEEE Winter Applications of Computer Vision Workshops*, 2019. 38
- [25] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 3, 13, 14
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 7
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016. 19, 29
- [28] 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. 3, 14

- [29] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 31
- [30] Yong Jae Lee, Alexei A Efros, and Martial Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *IEEE international conference on computer vision*, 2013. 13
- [31] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 30
- [32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 7, 8, 17, 18
- [33] 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*, 33(4):149, 2014. 3, 13, 24, 26, 29, 30
- [34] Stefan Lee, Nicolas Maisonneuve, David Crandall, Josef Sivic, and Alexei A. Efros. Linking past to present: Discovering style in two centuries of architecture. *IEEE International Conference on Computational Photography*, 2015. 13
- [35] Stefan Lee, Haipeng Zhang, and David J Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2015. 3, 15, 17, 26, 27
- [36] 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. 27
- [37] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *IEEE conference on computer vision and pattern recognition workshops*, 2015. 12
- [38] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 17, 27
- [39] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocation. In *IEEE conference on Computer Vision and Pattern Recognition*, 2015. 17, 27

- [40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014. 18
- [41] Cewu Lu, Di Lin, Jiaya Jia, and Chi-Keung Tang. Two-class weather classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2510–2524, 2017. 26
- [42] 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. 27
- [43] Kevin Matzen and Noah Snavely. Scene chronology. In *European Conference on Computer Vision*, 2014. 38
- [44] Radu P 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*, 2016. 31
- [45] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, 2016. 9, 41
- [46] Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *European Conference on Computer Vision*, 2012. 13
- [47] <http://www.planet.com/>. 26
- [48] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 27
- [49] 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. 17, 18
- [50] Tawfiq Salem, Connor Greenwell, Hunter Blanton, and Nathan Jacobs. Learning to map nearly anything. In *IEEE International Geoscience and Remote Sensing Symposium*, 2019. 5
- [51] 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*, 2016. 13

- [52] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. A multimodal approach to mapping soundscapes. In *IEEE International Geoscience and Remote Sensing Symposium*, 2018. 6, 17, 19, 27
- [53] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the impact of scenic environments on health. *Scientific reports*, 5:16899, 2015. 24
- [54] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 8
- [55] 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. 8
- [56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 7, 45
- [57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 29
- [58] Weilian Song, Tawfiq Salem, Hunter Blanton, and Nathan Jacobs. Remote estimation of free-flow speeds. In *IEEE International Geoscience and Remote Sensing Symposium*, 2019. 15
- [59] Stephen Stansfeld, Mary Haines, and Bernadette Brown. Noise and health in the urban environment. *Reviews on Environmental Health*, 15(1-2):43–82, 2000. 41
- [60] 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. 3, 7
- [61] 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. 26
- [62] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015. 31

- [63] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 10
- [64] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. Modality and component aware feature fusion for rgb-d scene classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 7
- [65] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *IEEE International Conference on Computer Vision*, 2016. 26
- [66] Jingya Wang, Mohammed Korayem, Saul Blanco, and David J Crandall. Tracking natural events through social media and computer vision. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1097–1101. ACM, 2016. 27
- [67] Jingya Wang, Mohammed Korayem, and David Crandall. Observing the natural world with flickr. In *ICCV Workshop on Computer Vision for Converging Perspectives*, 2013. 27
- [68] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, 2016. 3
- [69] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *IEEE/ISPRS Workshop: EARTHVISION: Looking From Above: When Earth Observation Meets Vision*, 2015. 17, 27
- [70] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, 2015. 3, 11, 13, 15, 16, 17, 27, 34, 45
- [71] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and mapping natural beauty. In *IEEE International Conference on Computer Vision*, 2017. 27
- [72] Scott Workman, Menghua Zhai, David Crandall, and Nathan Jacobs. A unified model for near/remote sensing. In *IEEE International Conference on Computer Vision*, 2017. 27
- [73] Ling Xie and Shawn Newsam. Im2map: deriving maps from georeferenced community contributed photo collections. In *ACM SIGMM International Workshop on Social media*, 2011. 27

- [74] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 7
- [75] Dong Yi, Zhen Lei, and Stan Z Li. Age estimation by multi-scale convolutional network. In *Asian Conference on Computer Vision*, 2014. 12
- [76] 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. 8
- [77] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 17, 27
- [78] Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. Learning geo-temporal image features. In *British Machine Vision Conference*, 2018. 13, 26, 31
- [79] 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*, 2017. 17, 18, 24, 26, 29, 44, 45
- [80] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 2014. 3, 7, 12

Vita

TAWFIQ SALEM

Contact

Email	tawfiq.salem@uky.edu
Phone	(859) 257-3961
Address	329 Rose St., Lexington, KY, 40503
Website	http://cs.uky.edu/~salem/

Education

2014–2019	University of Kentucky Lexington, KY Ph.D. in Computer Science (GPA: 4.0), Adviser: Dr. Nathan Jacobs.
2010–2012	Purdue University Indianapolis, IN Master of Science in Computer Science (GPA: 3.73).
2002–2006	Islamic University Gaza, Palestine Bachelor of Science in Computer Science (GPA: 3.8, 1 st rank).

Experience

- 2015–2019 **University of Kentucky**, Computer Science Department – Lexington, KY
Research Assistant:
In Dr. Nathan Jacobs research group, working in the area of Computer Vision.
- 2014–2017 **University of Kentucky**, Computer Science Department – Lexington, KY
Teaching Assistant:
Teaching assistant for CS 221 (MATLAB), CS 215 (C++), and CS 115 (Python).
- Summer **University of Kentucky**, Computer Science Department – Lexington, KY
Summer Instructor:
Summer 2016- Teaching CS 115: Introductions to Computer Programming (Python)
Summer 2014- Teaching CS 221: First Course in Computer Science for Engineers (Matlab)
- 2010–2012 **Purdue University**, Computer Science Department – Indianapolis, IN
Tutor:
Assisted undergraduate students in programming problems (C,C++, Java, Python, R).
Helped undergraduate students in database (Oracle, MySQL).
- 2008–2010 **University College of Applied Science** – Gaza, Palestine
Instructor:
Teaching Introductions to Computer Programming(C++)
- 2006–2010 **Islamic University** – Gaza, Palestine
Teaching Assistant and Lab Technician:
Teaching assistant for computer science courses.
Maintained the computer labs: installing operating systems and software packages.

Publications

Submitted

- 2019 **Tawfiq Salem**, Scott Workman, Nathan Jacobs. *Learning a Dynamic Map of Visual Appearance*. [Submitted]
- 2019 Gongbo Liang, Hunter Blanton, **Tawfiq Salem**, Xin Xing, Nathan Jacobs, Xiaoqin Wang. *Joint 2D-3D Breast Cancer Classification*. [Submitted]

Published/Accepted

- 2019 **Tawfiq Salem**, Connor Greenwell, Hunter Blanton, Nathan Jacobs. *Learning to Map Nearly Anything*. In IEEE International Geoscience and Remote Sensing Symposium (IGARSS). [Oral]
- 2019 Weilian Song, **Tawfiq Salem**, Hunter Blanton, Nathan Jacobs. *Remote Estimation of Free-Flow Speeds*. In IEEE International Geoscience and Remote Sensing Symposium (IGARSS). [Oral]
- 2018 **Tawfiq Salem**, Menghua Zhai, Scott Workman, Nathan Jacobs. *A Multimodal Approach to Mapping Soundscapes*. In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR). [Sight and Sound Workshop]
- 2018 Menghua Zhai, **Tawfiq Salem**, Connor Greenwell, Scott Workman, Robert Pless, Nathan Jacobs. *Learning Geo-Temporal Image Features*. In British Machine Vision Conference (BMVC).
- 2018 **Tawfiq Salem**, Menghua Zhai, Scott Workman, Nathan Jacobs. *A Multimodal Approach to Mapping Soundscapes*. In IEEE International Geoscience and Remote Sensing Symposium (IGARSS).
- 2017 William Song, **Tawfiq Salem**, Nathan Jacobs, Michael Johnson. *Detecting the Presence of Bird Vocalizations in Audio Segments Using a Convolutional Neural Network Architecture*. In International Symposium on Acoustic Communication by Animals. [Abstract]
- 2016 **Tawfiq Salem**, Scott Workman, Menghua Zhai, Nathan Jacobs. *Analyzing Human Appearance as a Cue for Dating Images*. In IEEE Winter Conference on Applications of Computer Vision (WACV).

Honors and Awards

- | | |
|-----------|--|
| 2017 | ACM UK Award for Outstanding Teaching Assistant in CS, University of Kentucky. |
| 2016 | CS Department Award for Outstanding Teaching Assistant, University of Kentucky. |
| 2016 | \$1,000 Computer Science Department Travel Grant, University of Kentucky. |
| 2016 | \$400 Graduate School Travel Grant, University of Kentucky. |
| 2010-2012 | Fulbright scholarship, U.S. Department of State. |
| 2010-2011 | \$5,000 grant from computer science department, Purdue University, Indianapolis. |
| 2006 | Graduated with the highest grade point average (3.8, 1 st rank) among the students of the Faculty of Information Technology, Islamic University - Gaza (Class of 2006). |

Talks and Presentations

- | | |
|------|---|
| 2018 | "A Multimodal Approach to Mapping Soundscapes", June. 2018, Sight and Sound Workshop at IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah. |
| 2015 | "Computer Vision Applications of Deep Convolutional Neural Networks", Nov. 2015, Keeping Current Seminar, Computer Science Department, University of Kentucky. |
| 2016 | "Analyzing Human Appearance as a Cue for Dating Images", Mar. 2016, IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY. |

Services and Affiliations

- | | |
|------------------|---|
| Volunteer | <ul style="list-style-type: none">– Serving as Microteach Leader, Graduate School, University of Kentucky (2018).– Engineers Day (E-Day), University of Kentucky (2016).– Sharek Youth Forum, Gaza, Palestine (2008-2010).– Beit Lahia Development Association, Gaza, Palestine (2007-2009). |
| Member | <ul style="list-style-type: none">– Institute of Electrical and Electronics Engineers (IEEE).– Association for Computing Machinery (ACM). |