




2018

Modeling and Mapping Location-Dependent Human Appearance

Zachary Bessinger

University of Kentucky, zach.bessinger@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0001-8719-1249>

Digital Object Identifier: <https://doi.org/10.13023/etd.2018.469>

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

Recommended Citation

Bessinger, Zachary, "Modeling and Mapping Location-Dependent Human Appearance" (2018). *Theses and Dissertations--Computer Science*. 75.

https://uknowledge.uky.edu/cs_etds/75

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.

Zachary Bessinger, Student

Dr. Nathan Jacobs, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

Modeling and Mapping Location-Dependent Human Appearance

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
Zachary Bessinger
Lexington, Kentucky

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

Copyright© Zachary Bessinger 2018

ABSTRACT OF DISSERTATION

Modeling and Mapping Location-Dependent Human Appearance

Human appearance is highly variable and depends on individual preferences, such as fashion, facial expression, and makeup. These preferences depend on many factors including a person's sense of style, what they are doing, and the weather. These factors, in turn, are dependent upon geographic location and time. In our work, we build computational models to learn the relationship between human appearance, geographic location, and time. The primary contributions are a framework for collecting and processing geotagged imagery of people, a large dataset collected by our framework, and several generative and discriminative models that use our dataset to learn the relationship between human appearance, location, and time. Additionally, we build interactive maps that allow for inspection and demonstration of what our models have learned.

KEYWORDS: human appearance, neural networks, computer vision, social media, geographic location, mapping

Author's signature: Zachary Bessinger

Date: December 13, 2018

Modeling and Mapping Location-Dependent Human Appearance

By
Zachary Bessinger

Director of Dissertation: Nathan Jacobs

Director of Graduate Studies: Mirosław Truszczyński

Date: December 13, 2018

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to my advisor, Dr. Nathan Jacobs. I learned more than I ever expected under his guidance, even beyond computer vision and machine learning. He not only helped fund a large part of my research, but was a patient and strong mentor. He taught me the importance of being pragmatic and understanding problems in their simplest form. I would have never accomplished what I have without his assistance.

I would also like to thank my committee members who guided me to this point: Dr. Judy Goldsmith, Dr. Ruigang Yang, and Dr. Matthew Zook. They provided helpful feedback and discussions which helped me curate this document and present my work in a meaningful way.

I had the privilege of working with many intelligent and helpful peers, including Menghua Zhai, Connor Greenwell, and other members of my lab group. Their insightful discussions, research-related or otherwise, were helpful when distilling and creating new knowledge. This document would not exist were it not for all of these individuals, and I may have forgotten several people, but to everyone involved I am forever thankful.

Table of Contents

Acknowledgments	iii
Table of Contents	iv
List of Figures	vi
List of Tables	x
Chapter 1 Introduction	1
1.1 Understanding People using Computer Vision	1
1.2 Geo-dependent Human Image Analysis	2
1.3 Mapping Human Appearance	4
1.4 Contributions	5
Chapter 2 Technical Background	6
2.1 Traditional Methods for Image Analysis	6
2.2 Convolutional Neural Networks	7
2.3 Deep Generative Models	10
2.4 Geotagged Image Analysis	13
Chapter 3 WhoGoesThere? A Large-Scale Dataset of Geotagged Human Faces 14	
3.1 Introduction	14
3.2 Infrastructure	16
3.3 The WGT Dataset	17
3.4 Dataset Visualization Using Conditional Averaging	20
3.5 Interactive Maps of Facial Appearance	21
Chapter 4 Baseline Generative and Discriminative Models of Facial Appearance 27	
4.1 Introduction	27
4.2 Related Work	27

4.3	Predicting Facial Appearance	28
4.4	Multi-modal Distributions	31
4.5	Evaluation	36
4.6	Conclusions	37
Chapter 5 A Deep Generative Model of Facial Appearance		38
5.1	Introduction	38
5.2	Related Work	40
5.3	Approach	42
5.4	Evaluation	46
5.5	Conclusions	51
Chapter 6 A High Spatial Resolution Model of Clothing Style		52
6.1	Introduction	52
6.2	Related Work	55
6.3	The XViewClothing Dataset	57
6.4	Approach	58
6.5	Evaluation	60
6.6	Conclusions	66
Chapter 7 Conclusion		67
Bibliography		69
Vita		84

LIST OF FIGURES

1.1	Attribute maps of the past. (a) A map of the Americas by Sebastian Münster in 1572 and shows different known attributes of the time in the New World. (b) A contemporary example from a children’s book [99] showing geographic and cultural characteristics of Switzerland, such as the Swiss Alps, the character Heidi, and a man playing an alphorn.	3
1.2	Our framework for geo-dependent human image analysis. (a) We collect geo-tagged imagery of people, highlighting important areas of interest. (b) We learn high-level semantic features relating to their facial appearance, clothing choices, and background scene. (c) We develop web-based, interactive attribute maps for each proposed model.	4
2.1	The AlexNet [72] convolutional neural network architecture.	7
2.2	An abstract example of a GAN architecture trained using the MNIST digits dataset [78].	11
3.1	We use geotagged social-media images to learn how human facial appearance varies across the globe. This montage shows representative images for different clusters of people. See Section 4.4 for details.	15
3.2	Infrastructure of “Who Goes There?”	16
3.3	Word cloud of user tags in the WGT dataset.	18
3.4	Distributions of age and gender in the WGT dataset.	20
3.5	Top three rows: Average Indian females arranged from youngest to oldest averaging using similarity alignment (first row), perspective alignment (second row), and Collection Flow [65]. (third row). Bottom three rows: Average Indian males using the same approaches.	22
3.6	Multiscale visualization. Zooming in reveals finer details about world populations.	23
3.7	Web interface showing the capability to toggle between the average male (a) and female (b) appearance based on spatial boundaries.	24
3.8	Web interface showing the average male and female appearance in China.	25
4.1	Results of learning models conditioned individually on age, gender, location, and face shape and then conditioned on all four of these attributes. The predicted components are then used to reconstruct the original image.	29

4.2	Reconstructions from our linear model of facial appearance. The x -axis corresponds to age and the y -axis corresponds pose. We found that a linear model resulted in reasonable reconstructions for some parameters, but not the location parameter, which is highly non-linear.	30
4.3	These montages show exemplar faces for the eight clusters that are more likely to be from the corresponding location.	33
4.4	(left) An exemplar face for a given class, c_i . (right) The conditional distribution of that class for each location, $P(c = c_i \text{location})$	33
4.5	The automated geographical subpopulation factor analysis with five latent factors shows spatial regularities in location posterior (left) and the most differentiable appearances for each factor (right).	35
4.6	Quantifying appearance diversity using the fraction of variability explained by the top k PCA components of the $FC8$ identity features. For a given number of components, larger values imply less diversity, because more of the variability is explained by the top k components.	36
5.1	We propose a generative model that incorporates geospatial metadata, along with additional human-related attributes, and allows for synthesis of people within a given area. The color of the bounding box in the map corresponds to randomly generated women from their respective regions.	39
5.2	Samples from the WGT dataset [10] used in our work. Unlike face datasets that have been previously used to train generative models, such as CelebA [86], our dataset has not been manually filtered and contains a wide variety of image qualities.	43
5.3	Our proposed model, $GPS2Face$, has two components: a landmark prediction network, L , and an appearance generation network supported by the other sub-networks. Landmarks are used to guide synthesis and improve the quality of generated faces since identity is not used as a regularizer. L uses latent factors, c , to predict facial landmarks, s . Predicting landmarks allows us to model how facial structure changes with respect to latent factors and also serves to avoid manually specifying a large set of landmarks at test time.	44

5.4	Examples of encoding an input set of images (a) in randomly selected to have certain poses, and transforming them by manipulating the latent factors. (b) shows the reconstruction using ground truth labels. (c) shows changing the latent factors used to generate (a) into females, ages 25–32, frontalized, and each row is fixed to the following set of countries: United Kingdom, Germany, Italy, India, Taiwan, Ethiopia, Iran, Sudan. (d) shows changing the latent factors to be males, ages 38–43, pitch = -35° , yaw = 45° , and each row fixed to the same countries as used in (c).	47
5.5	Qualitative comparison of random samples from our method and a previous method from Bessinger <i>et al.</i> [10]. Input images (a) are encoded through our network to predict z , which is used as input to our generator to decode (b). Using the same conditioning terms, in (c) we change z to be a sample from the prior. (d) is generated using [10].	48
5.6	We observe that for a fixed sample, z , we can vary pose and preserve individual identity. We fix the age and gender to be a 25 year old female. We vary the pose to be $\pm 20^\circ$ yaw and $\pm 30^\circ$ pitch. We then sample locations from the countries shown in the captions above.	49
5.7	We highlight appearance diversity within each country by generating faces sampled from the prior. In each montage, age and gender are randomized, while pose and geographic location are fixed.	50
6.1	We propose a model that uses satellite imagery to predict a distribution over human attire within a given geographic region.	54
6.2	Three kinds of clothing style imagery showing high variability under different constraints. The first image shows traditional fashion imagery used by stores to sell items of clothing which have both constrained lighting and pose. The second image shows stock imagery in which the environment in which the image is captured has semi-constrained pose and lighting, as well as a relatively clean background. The third image shows images captured “in the wild,” which have both unconstrained lighting and pose and may suffer from external factors, such as occlusion.	56
6.3	Histogram of confidences for patches extracted with Faster-RCNN. The x -axis is confidence and the y -axis is the number of patches.	58

6.4	Our approach involves three steps. Given an input image, we first extract scene features from a scene classifier [155] and detect/crop all people found in the image using Faster-RCNN [115]. For each cropped person, we extract an associated clothing style feature and concatenate the scene feature to each clothing style feature. Next, we cluster the concatenated features and assign each person image to their centroid. Finally, we use a CNN that takes an aerial image as input and predicts a probability distribution over centroids.	59
6.5	Top- k accuracy of style cluster prediction on the test set for different settings of $k \in \{1, 3, 5, 10, 25, \text{ and } 50\}$	61
6.6	Given an aerial image from our test set (left), we show the true appearance (center) of people from the location and predicted appearance (right).	62
6.7	Appearance trends for each month of the year. Each row shows an aerial image over a geographic region (left), samples from a particular cluster (center), and the frequency in which those clusters appear conditioned on the month of the year (right).	63
6.8	High-resolution maps of clothing style. Each row is a location corresponding to London, Myrtle Beach (USA), Central Park (USA), Tokyo, and Dubai respectively. The first column of each row is an aerial image captured over a large geographic region and the remaining columns show the probability for a particular style of clothing to appear in that location ranging from low (green) to high (red) likelihood.	64

LIST OF TABLES

3.1	Comparison of existing large-scale face datasets with geotags.	18
4.1	Average RMSE of PCA reconstruction for each regression method conditioned on various types of input.	31
5.1	Quantitative evaluation of our proposed method.	51
6.1	Statistics for the XVC Great Britain subset.	57

Chapter 1

Introduction

Human appearance is highly variable and modulated by location and time. One example of this is occupational uniforms worn by public servants, such as firefighters, military, and police officers. In another example, if we were to imagine someone on the beach, we would likely picture a person in swim-wear and not a clown costume. People also get married on beaches, so it is reasonable to see someone in a suit on the beach depending on the location and time of year. There are specific modes of appearance that depend on the time at various temporal scales. We are more likely to see someone wearing a clown costume on Halloween than any other day of the year. Fashion trends are a prime example of temporally-dependent human appearance across years that is constantly evolving. Similarly, people also use their appearance as a form of expression. This can be for identifying with their friends and social groups by dressing similarly, or supporting their favorite sports teams on game day. There is a strong correlation between location, time, and human appearance. The complexity this correlation motivates us to understand the dependencies between each factor and their relationship to human appearance. In this work, we develop computer vision models to support a large-scale framework for understanding location- and time-dependent human appearance.

1.1 Understanding People using Computer Vision

Computer vision research has long-sought to comprehend *how* to view our world using both biological and engineering inspiration. Early vision researchers believed that if we want to understand the fundamental relationship between visual perception and human appearance, we could be inspired by how a child learns to represent human faces. The earliest of this research claimed that the mammalian visual cortex, the component of the brain that processes visual information, could be modeled by Gabor filters [22]. Vision researchers in

the 1990’s successfully discovered multiple ways to represent human faces [134, 5]. Since then, computer vision research on human appearance has moved beyond the single task of face representation and towards other ways to holistically represent people. These applications are typically used for the tasks of detection, surveillance, and re-identification. Other common human-appearance related tasks computer vision addresses are object recognition, pose estimation, and attribute detection.

Accomplishments in computer vision and machine learning in the past decade are due in part to new algorithms, faster hardware, open source code, and ease of sharing ideas. However, an often overlooked cause for rapid research advancement is due to the increase in large, quality benchmark datasets. Good datasets are imperative for all learning tasks, especially supervised learning which assumes human knowledge is provided in the form of a label to guide the learning process. One of the most well-known of benchmark datasets is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [118], which contains 14,197,122 annotated images collected from various search engines for the general purpose tasks of object detection, single-object localization, and image classification. If not for imagery available through Internet search engines, acquiring this amount of data would have been arduous.

Datasets used for understanding human appearance, such as those curated to support the tasks of face detection and recognition, have grown exponentially in size in the past two decades. Face datasets of the 1990’s [122, 38] contained hundreds of images with single labels, whereas contemporary face datasets contain hundreds of thousands to millions of images [53, 97, 42] and many labels. For complex tasks, such as per-pixel segmentation of clothing, the labeling process is expensive and the number of human annotated data per dataset is between hundreds and thousands [146]. Though the amount and quality of image data is becoming increasingly available, most current research still ignores two important human appearance related cues: location and time.

1.2 Geo-dependent Human Image Analysis

Only within the past several years have computer vision experts began exploring techniques for geo-dependent human imagery analysis. A large-scale dataset of geotagged faces, GeoFaces [56], has been introduced and shown to be useful for the tasks of city [57], country [56], and attribute [39] classification from faces captured “in the wild.” Recently, Wang et al. [138] showed that weather and time (temporal consistency) can be used to improve the task of facial attribute classification. These works have shown substantial promise in capturing the relationship between human appearance, location, and time. We believe lo-



(a)



(b)

Figure 1.1: Attribute maps of the past. (a) A map of the Americas by Sebastian Münster in 1572 and shows different known attributes of the time in the New World. (b) A contemporary example from a children’s book [99] showing geographic and cultural characteristics of Switzerland, such as the Swiss Alps, the character Heidi, and a man playing an alphorn.

cation and time have been ignored by vision researchers over the years for a number of possible reasons: 1) it was simply neglected information and 2) the kind and amount of data needed to incorporate location and time into human appearance algorithms simply did not exist. We argue that relationship between image, location, and time is a largely ignored, untapped cue that can be useful in improving our understanding of human appearance.

Location can refer to either an object’s geospatial coordinates in the world or the scene in which an object resides. Both definitions of location are important to distinguish if we are to relate human appearance and location. If one were asked to imagine how a person on Waikiki Beach looks during summer, they might have relied on geospatial location using their prior knowledge of being in that specific region or of other beaches during summer. If asked to picture the clothing worn in a boardroom, one might rely on location as a scene or use prior cues from settings in movies and television shows. Time also plays an important role in relation to human appearance. In fact, time is an important contextual cue for both of the previously mentioned definitions for location. When location is defined geospatially, time can be indicative of short-term events, such as the transition between day and night, or long-term events such as the change of seasons. When location is defined as a scene, time can be used to discriminate whether a restaurant might have people dressed in work-casual at lunch time or cocktail attire for dinner.

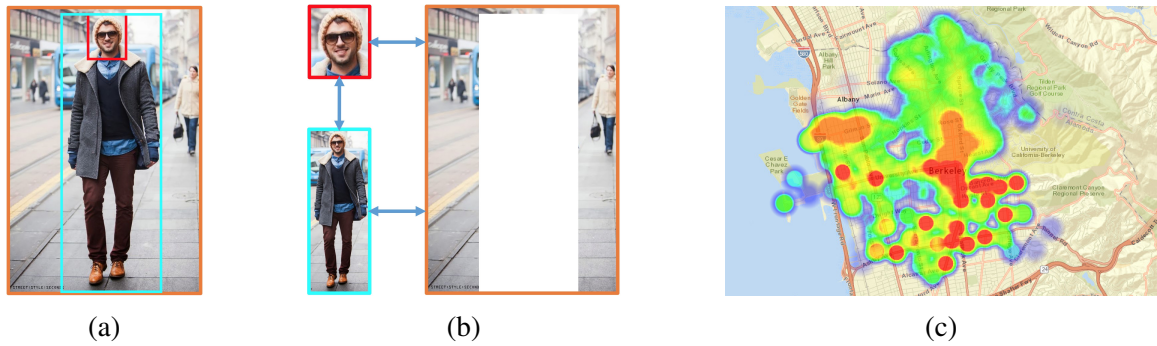


Figure 1.2: Our framework for geo-dependent human image analysis. (a) We collect geo-tagged imagery of people, highlighting important areas of interest. (b) We learn high-level semantic features relating to their facial appearance, clothing choices, and background scene. (c) We develop web-based, interactive attribute maps for each proposed model.

1.3 Mapping Human Appearance

By capturing the relationship between an image of a person and their geographic location, we can construct maps that visually bridge the two concepts. Cartographers of the past sought to understand the world in a similar manner by creating paper maps that showed various attributes about a region, such as those shown in Figure 1.1. Similar kinds of maps can be found in children’s books as a way to highlight cultural diversity. The artists would design these kinds of attribute maps that highlight specific cultural elements, like activities, architecture, or inventions. Today we have many different digital, web-based mapping libraries and applications, most notably Google and Bing maps. These tools are useful for answering quantitative questions about nearby restaurants and how to navigate from point A to point B. However, their capabilities are limited to answering only these quantitative questions and fail to answer qualitative questions, such as those related to human appearance. Our work is a step in the direction of modeling and creating interactive maps of human appearance.

Our work proposes a framework to answer the following research question: what novel computer vision and machine learning techniques can we develop to analyze the relationship between social media imagery of people, their geolocation, and the time in which they were captured? Figure 1.2 provides a three-part, visual description of our work. We want to tap into the vast quantity of available geotagged social media imagery and reason about an image by breaking it down into different parts that compose the image. When observing a full body portrait of someone, there are at least three different parts of image that may describe it: the person’s face, clothing, and scene. Visual information extracted from the face can give us estimates for age, gender, expression, hair styles, and any accessories they

might be wearing such as earrings or sunglasses. When we look at a person’s clothing, we can estimate their style, perhaps other attributes from style, including age and even the outdoor temperature. Scene-level visual cues can tell us if a user is outdoor or indoor, and from that we can estimate time of day, what the weather is like, or place of residence. Since we are focused on the location aspect, if we assume that location is given, we can make maps that show the spatial distribution of similar people, wearing similar outfits, at similar times of the day and weather conditions. Using computer vision and machine learning techniques to model the relationship between all three of these areas can provide us with an understanding to reason about our world and answer visual questions in ways that are currently impossible using today’s tools.

1.4 Contributions

The focus of this research is modeling the relationship between geographic location and human appearance using a large dataset of geotagged consumer photographs. We develop a framework for collecting, analyzing, modeling, and presenting digital maps of human attributes at a worldwide scale. The contributions of this dissertation are as follows:

- A framework for collection, analysis, modeling, and interactive attribute maps. This is reported in [10].
- The largest known, public dataset of geotagged facial imagery currently available.
- A generative model of facial appearance that allows for generating faces given any geographic location in the world. Additionally, the model allows for attribute transformation of input faces and flexible control of non-spatial related predictor variables, such as age and gender. This is reported in [9].
- A discriminative model that uses satellite imagery and image capture time to model the distribution of human appearance that can be used to predict the distribution of clothing and trends for any geographic location in the world.

Chapter 2 provides a technical background to understand the work in this dissertation. Chapter 3 introduces a large-scale dataset of faces. Chapter 4 describes preliminary research that has been done towards generative modeling of the human face. Chapter 5 improves upon the facial generative model introduced in Chapter 4 by introducing a factored, latent variable generative model. Chapter 6 introduces models developed to learn about the distribution of clothing styles using satellite images, their location, and image capture time.

Chapter 2

Technical Background

In this section, I review the background materials necessary to understand the compiled work in this dissertation.

2.1 Traditional Methods for Image Analysis

In order to understand and solve computer vision and machine learning problems, we must have some meaningful way of representing the data. These are often referred to as *features* and are often expressed with respect to a sample or set of samples of input data. In the case of images, we could use pixels as our feature representation to learn to reason about an image, however this particular representation is problematic. Pixel representations are high-dimensional and contain redundant information due to high spatial correlations. Since computational resources are finite, we want to learn on a representation that is both informative and compact.

One way to obtain a compact representation of high dimensional data is to use principal component analysis (PCA). PCA is an unsupervised learning technique that projects data onto a new coordinate system using a linear transformation where each axis is orthogonal to all others and the set of all axes are ordered by most to least amount of variance. Assume we are given a data matrix, $\mathbf{X}^{n \times d}$, where n is the number of samples, d is the number of features, and the data matrix is mean-centered. PCA can be performed using the singular value decomposition (SVD),

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top, \tag{2.1}$$

where \mathbf{U} is the orthonormal basis of left singular vectors, \mathbf{S} are the singular values, and \mathbf{V} is the orthonormal basis of right singular vectors. The scores are found by the product, $\mathbf{U}\mathbf{S}$. The rank- k approximation of the original input, $\hat{\mathbf{X}}^{n \times d}$, can be reconstructed by projecting

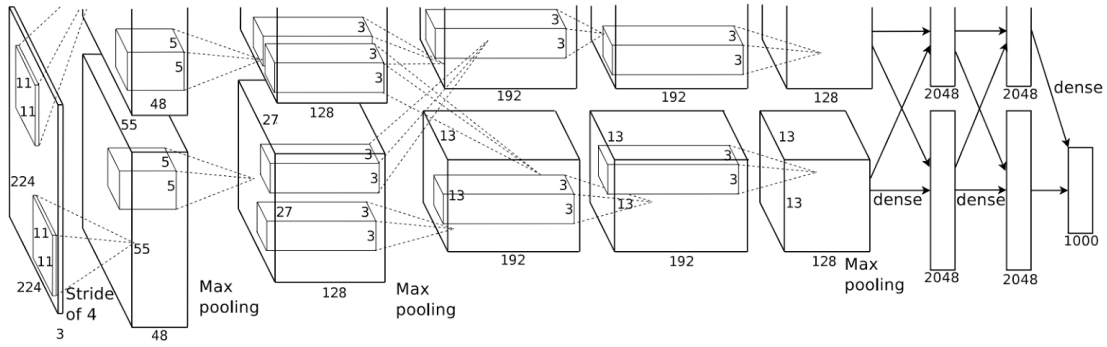


Figure 2.1: The AlexNet [72] convolutional neural network architecture.

$U_k S_k$ onto principal components V_k^T . The low-dimensional approximation of the original input is more computationally efficient for machine learning algorithms. Although counterintuitive, choosing a small k can increase performance in machine learning algorithms, despite information loss. One early technique for understanding human appearance that uses PCA, and has been successfully applied on the task of facial recognition, is *eigenfaces* [134].

There are many ways to reduce complexity of inputs while maintaining necessary information and reduce computational cost. Traditional methods relied on human expertise about image structure. These methods are referred to as *feature extractors* and include such techniques as the spatial envelope (GIST) [105] used for scene recognition, and more general purpose techniques scale-invariant feature transformation (SIFT) [88] and histograms of oriented gradients (HoG) [20]. Each feature extractor has its own specific set of hyperparameters and a good choice of these hyperparameters can change dramatically as the data source changes, even if they share similar semantics. Historically, computer vision research involved hyperparameter tuning of the feature extractor to manually find a good configuration for downstream tasks.

2.2 Convolutional Neural Networks

Instead of spending valuable time searching for the right feature extractor for a particular dataset, it would be better to learn the features from the data itself that are optimal for a particular task. One way this was done was using convolutional neural networks [78], which fell out of favor with researchers when support vector machines [11, 17] were developed, but have recently come back into academic interest. The revival of convolutional neural networks (CNNs) came due to the work of Krizhevsky *et al.* [72] who used CNNs in their

submission to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [118] 2012 classification task. Their network architecture, AlexNet, was able to substantially outperform other challengers of the time. Since then there has been major interest in developing new CNN architectures and training methods. The AlexNet architecture is shown in Figure 2.1. State-of-the-art performance on benchmark challenges [118, 157] in computer vision that include the tasks of object recognition, detection, localization, classification, scene recognition, and semantic segmentation has been achieved with CNNs. Many libraries have been developed to work with these neural networks including Caffe [61], Tensorflow [1], and PyTorch [108].

Convolutional neural networks are a special class of neural network. The *convolutional* part of their name comes from the process by which outputs are calculated. In a traditional neural network, the output layer $i + 1$ from an input layer i is calculated by computing the inner product of a set of weights, W_i and its current inputs x , then optionally adding a bias term, b_i . The size of these weights must match the size of the input, which is problematic in the case of images where the weight matrix of the input layer of the neural network can grow large. Convolutional neural networks address this by applying a shared set of weights *convolutionally* across inputs. Weight sharing significantly reduces the number of parameters when compared with a multilayer perceptron network. Other layer operations, such as the max pooling layer, allow for the network to learn a small amount of invariance to translations and rotations. Convolutional neural networks can be abstractly conceptualized as a composition of functions. A feature representation for some input is the output of any function within the composition.

A set of weights for the input layer of a convolutional neural network that operates on RGB imagery is a 4-dimensional tensor defined by h, w, c_{in}, c_{out} , which refer to height, width, number of input channels, and number of output channels, respectively. Similar to the functionality of neural networks, we compute the inner product of the weights and inputs, which yield a single output. These inner product are applied in a strided manner across inputs, yielding single numbers for c_{out} many layers. The dimensionality of the output layer is completely controlled by choice of stride and input/output padding.

We learn the weights of a CNN using back-propagation [117]. This process computes the gradient of the loss/objective function with respect to each input. Errors are sent backwards through the network, applied using the chain rule, and optimized using gradient descent. For ImageNet image classification, the activation function for the final layer is the softmax function:

$$\text{softmax} = \frac{\exp(x_i)}{\sum_i \exp(x_i)}. \quad (2.2)$$

This activation function exponentiates the output log probabilities (logits) of the final layer

and divides each element by their sum to ensure it is a valid probability distribution which sums to 1. After the softmax, we minimize the cross entropy loss:

$$- \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{train data}}} \log p_{\text{net}}(\mathbf{x}), \quad (2.3)$$

which optimizes for the distribution of the classifier, p_{net} , to be similar to the training data distribution, $\hat{p}_{\text{train data}}$, for a set of training examples, \mathbf{x} .

The weights of a network trained for ImageNet classification can be used as a feature extractor for other kinds of natural imagery. This is due in part to the large quantity and variety of images found in the ImageNet dataset. Since a set of weights has been learned over many diverse samples of natural imagery, we can use the learned set of weights from ImageNet classification for other tasks. This method of learning is referred to as *transfer learning* because we can transfer the knowledge we learn for one task to another due to similarities in the data or task. The lowest level layers of the network have already been learned for generic objects, therefore these layers do not need to be relearned. Instead, we can freeze these layers and apply additional layers to the top of the network and optimize for the specific task at hand. In the literature, this is referred to as *fine-tuning*. The approach is similar to what vision researchers did in the past, except treating the frozen neural network as a feature extractor. Many modern published papers rely on the learned weights from these widely-accepted architectures as generic feature extractors, just as vision researchers of the past used hand-crafted feature extractors. Notable CNN architectures include AlexNet [72], Zeiler-Fergus Net (ZFNet) [152], VGGNet [127], Inception [129], Residual Networks (ResNet) [48], and Squeeze-and-Excitation Networks (SENet) [51]. These architectures are notable in the computer vision community because they are the annual winners of the ILSVRC [118] for image classification from 2012 – 2017. While 2017 was the last year for the ILSVRC workshop, researchers still report state-of-the-art metrics on this dataset, including most recent notable examples DenseNet [52] and Network Architecture Search Network (NASNet) [160].

In addition to being useful for ImageNet classification, CNNs are a base framework that can be applied in different architectural forms. One particular way they are used, and applied our work, is in the form of autoencoders. An autoencoder is a model often used in representation learning to learn a compact representation of the data. They can be seen from the view of dimensionality reduction, similar to other methods such as principal component analysis and factor analysis. Autoencoder architectures are hourglass-shaped and are composed of two networks: an encoder and a decoder. The encoder takes some input image, I and encodes it to a latent space, z , where $|z| \ll |I|$. The decoder uses z as input and tries to reconstruct I from the latent space, modeling $p(I|z)$. The loss of an

autoencoder is the reconstruction error of the input and is typically the L_2 loss. Autoencoders with a reconstruction error are another means of dimensionality reduction and can be considered non-linear PCA because they optimize for the same objective function but apply non-linear activation functions in each layer. An autoencoder is core component of our work in Chapter 5.

2.3 Deep Generative Models

Broadly speaking, there are two primary classes of models used in computer vision: discriminative and generative. Let \mathbf{x} represent an image and \mathbf{y} represent a set of labels associated with the image. Discriminative models are those that represent the probability distribution $p(\mathbf{y}|\mathbf{x})$, essentially learning to predict the labels given an image. Examples of such models include scene classification, age estimation, and face detection. In contrast, generative models represent $p(\mathbf{x}|\mathbf{y})$, which means we can generate an image given the labels. In general, generative models are more difficult to train than discriminative models because the output space is more complicated, complete realistic images as opposed to, for example, simple discrete labels.

2.3.1 Generative Adversarial Networks

Generative adversarial networks (GANs) are new class of generative models introduced by Goodfellow *et al.* [37] and have been studied extensively [36, 139, 19, 74] in recent years. GANs are composed of two networks: a generator and a discriminator. The generator's objective is to take in some input from a prior distribution that can be sampled from and produce an output that matches the distribution of the true data. The discriminator then observes the real images and the generated (fake) images and makes a binary decision to say whether an image is real or not. The two play a minimax game in which the discriminator maximizes its ability to distinguish between real and fake images and the generator minimizes its error when creating fakes. Generative adversarial networks are optimized using the cross-entropy loss between samples drawn from the distribution of the empirical data and the generated data using the loss function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log (1 - D(G(\mathbf{z})))] \quad (2.4)$$

An example GAN learning to generate MNIST digits is shown in Figure 2.2.

When GANs are used for images, we can make the analogy of a forensic artist and a criminal witness. The forensic artist (generator) tries to create a sketch of the criminal and

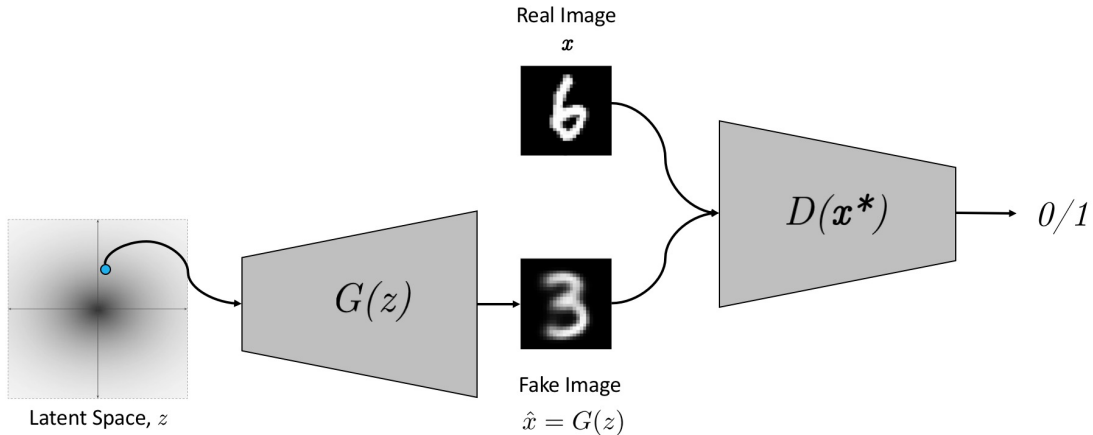


Figure 2.2: An abstract example of a GAN architecture trained using the MNIST digits dataset [78].

the witness looks at the created sketch. The witness then assesses which parts of the sketch looked like the criminal or which parts did not. For the artist to make a more accurate sketch, he must be able to perceive what the witness remembers. The forensic artist is able to make a perfect recreation when the witness sees the sketch and does not suggest any further modifications. This means that the generator has learned how to generate a sample that appears to the discriminator as though it were from the true data distribution (looks like the criminal).

Generative adversarial networks in their original formulation do not make it possible to have direct control over the generated output. One way to incorporate prior information into GANs are through conditional GANs (C-GANs) [98]. In this work, the authors have some class information corresponding to the data distribution and include that information as input to both the generator and the discriminator. The C-GAN objective modifies the original GAN objective as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x}|\mathbf{y})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log (1 - D(G(\mathbf{z}|\mathbf{y})))] \quad (2.5)$$

where \mathbf{y} is a set of any known factors about the sample that can be used to aid the likelihood functions of the generator and discriminator. Using our aforementioned example of a forensic sketch artist and a witness, if the witness were able to say with certainty that the criminal is male or female, the forensic sketch artist would be able to reduce their search space of all possible faces to those that are more shaped like the criminal's gender.

There are a variety of ways to introduce known information into a GAN. Some of these ways include directly concatenating information into the model [98, 154] or using additional classifiers [104, 103, 133]. Recent research into new C-GAN architectures involve

having two networks as generators: one that maps from latent space to an image and one that maps from image space to a latent vector [30, 27]. Conditional GANs have proven to be quite successful in generating images that match the additionally included information.

It has been suggested in recent works that GANs are able to generate sharp images because they are based on a perceptual loss. An example of this based recent work from in next-frame prediction [87], super-resolution [79], in-painting [109], and domain transfer [59, 69]. Other recent works have suggested that using an adversarial loss in combination with mean squared error can produce perceptually better images [62, 29, 109], and therefore an improved generative model that can better represent the true data distribution.

One of the key challenges for GANs is understanding the optimization dynamics by studying their behavior and improving their training stability. Though Goodfellow *et al.* [37] show in the original paper that there are theoretical guarantees for GAN convergence, there are practical issues that arise which make learning a GAN difficult. There have been numerous works which make suggestions on how to improve GAN training [112] including addressing issues with numerical stability [96, 95, 3] (or lack thereof) and how to best traverse the latent space [140] when interpolating between samples. Many approaches [4, 41, 7] have been recently proposed to simultaneously increase the stability and output resolutions of GANs. The Wasserstein GAN (WGAN) [4] uses the Earth Mover's distance applied to the distribution. An improvement upon WGAN was made by Gulrajani *et al.* [41] by adding an additional penalty to the gradient in the discriminator update. The motivation for this was to circumvent the need to Lipschitz continuity via gradient clipping needed for proving that WGANs can work. The current state-of-the-art of GANs use progressively growing architectures [63] and zero-based gradient penalties [94]. These recent works also show how high-resolution imagery, 1024×1024 , of celebrities can be generated using their methods.

Another important advancement with GANs has been the development of disentangled representations [102]. In the context of human appearance, suppose we are given an image of a face for whom we know the age and gender. If we were to train an autoencoder to map the image to latent space and reconstruct, the latent space would entangle the pixels of the image, the age of the person, and their gender into a compact, entangled representation that is most efficient for the network to reconstruct. This compact representation cannot be parsed in a meaningful way, so what we would like is disentangle the representation into components that can be manipulated. InfoGAN [14] is a notable work of unsupervised GANs by making a simple modification that encourages a latent variable to maximize the mutual information between when it was input to the generator and when it was output from the discriminator. Disentangled representations are a way to have control of the axes

in latent space so that we can manipulate them in a meaningful way to importance of each term.

2.4 Geotagged Image Analysis

Two common discriminative tasks in the area of geotagged image analysis are geolocalization and image-driven mapping. The task of geolocalization takes an image as input and estimates where it was captured. The seminal work of Hays *et al.* [47] involved collecting six million geotagged Flickr images and used a non-parametric approach to predict the location of a query image. This work has been recently revisited using a deep learning approach [137]. Works since then have proposed a variety of different approaches and using multiple sources of geotagged imagery, such as Google Street View [151] and combinations of aerial, ground-level, and landcover imagery [84, 141].

Image-driven mapping is the process of recognizing attributes from imagery and then conveying the learned attributes in the form of a map. Accurate image-driven maps of natural scenes are imperative in facilitating the development of geographical information systems (GIS). Not only are maps for natural scenes desirable, but as urban populations increase so does the need for urban maps. Crandall *et al.* [18] explore visual and textual characteristics of 35 million geotagged images from Flickr. Jacobs *et al.* [60] use geotagged webcam imagery to learn environment related attributes, such as weather and phenology. Xie *et al.* [142] use geotagged imagery to construct dense maps showing attributes such as scenicness.

Only recently has the relationship between the image of a human face and the location it was captured been explored in the area of computer vision. There are several works that have explored this domain and several large geo-facial datasets [97, 58] have been curated to explore the relationship between location and facial appearance. Greenwell *et al.* [39] develop a pipeline for processing geotagged imagery from Flickr and map several detected attributes. Our work is different in that we create maps of facial appearance not the distribution of the presence of facial attributes, such as age and gender. Islam *et al.* [56] provides a broad overview of problems in geo-facial image analysis. Wang *et al.* [138] show that using weather and location along with faces extracted from egocentric video improve face attribute classification.

Chapter 3

Who Goes There? A Large-Scale Dataset of Geotagged Human Faces

3.1 Introduction

According to the United States Census Bureau, the estimated world population as of January 2016 is 7.3 billion and rising. The increasing population density puts extreme pressure on resources and presents many opportunities for conflict. Understanding cultural and demographic trends and their spatial distribution is increasingly essential for individuals, corporations, and governments. Social scientists attempt to discover such trends, but the vast scale of this problem means that traditional approaches, which often involve manual data collection and scholarly dissemination, are insufficient. With the advent of social media, it has become quite easy to collect data that reflects these trends. However, novel methods for transforming this data into useful information are still needed.

We propose creating visualizations, based on geotagged social-media imagery, that enable novice users to understand world populations. Most social-media sites offer tools to visualize the appearance of people in different places. The most advanced tools allow users to issue a textual query, such as “person”, and then browse a map to see images of people in different locations. These visualizations are quite limited; they only show sparse samples of the underlying distribution of human appearance and do not make trends easy to discover. We address this problem with a combination of computational techniques and high-level, user-focused visualizations.

There are two main types of approaches we could consider for learning the relationship between human appearance and geographic location, discriminative and generative. Islam et al. [57] use a discriminative approach by addressing the face localization prob-



Figure 3.1: We use geotagged social-media images to learn how human facial appearance varies across the globe. This montage shows representative images for different clusters of people. See Section 4.4 for details.

lem. They propose using a deep convolutional neural network to estimate the city in which a given facial image was captured. This approach is appealing because it lends itself toward straightforward quantitative evaluation, however it does not directly support our goals of enabling user-focused visualizations. Therefore we take a generative approach; we attempt to construct a model that allows us to estimate the appearance of an individual for a given geographic location. We investigate a variety of different methods for incorporating geospatial priors into data-driven models to visualize human facial appearance.

We develop a collection of location-dependent human appearance models. We propose several strategies for representing the distribution of facial appearance for different geographic locations. We show how conditional averaging can be used to show how human

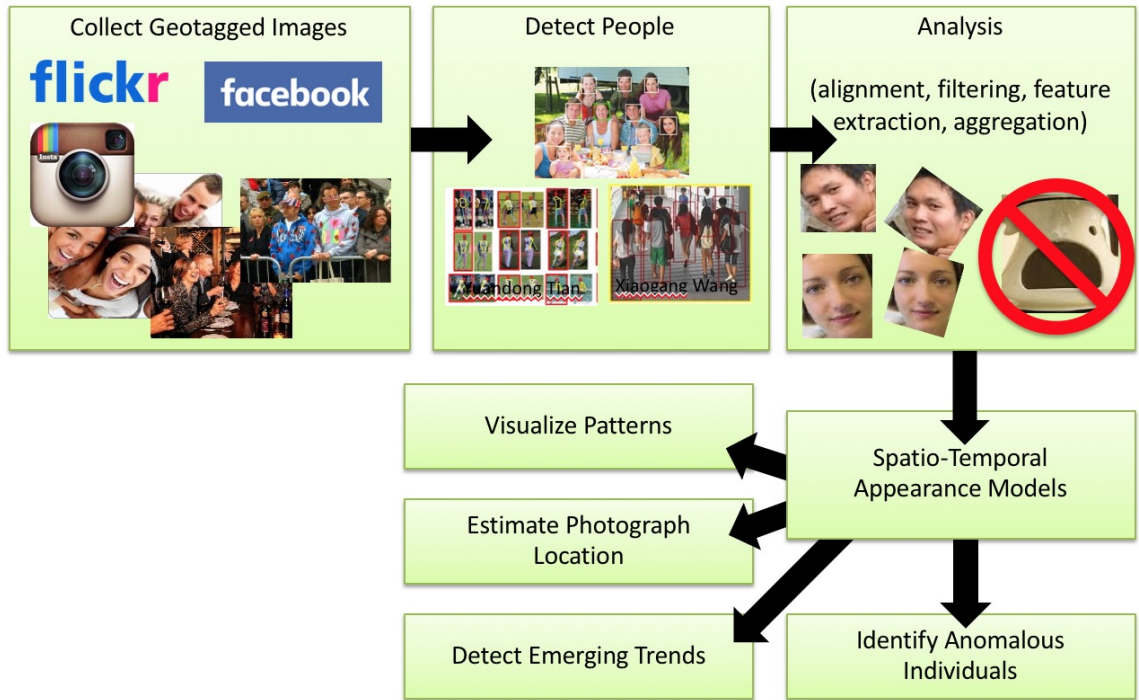


Figure 3.2: Infrastructure of “Who Goes There?”

appearance varies based on location and other factors. This approach works quite well, but is limited because it is not able to extrapolate or infer appearance where data is sparse. To address this, we propose learning-based approaches for modeling the distribution, $P(f|\ell)$, of facial appearance, f , for arbitrary geographic locations, ℓ . We show how to extend these models by conditioning on other attributes, such as gender, age, and face shape. We use the resulting models to support visualizations, web-based applications, and further analytical processing.

The main contributions of this work are: 1) A massive new dataset of geotagged face images, 2) A regression model that uses location, and potentially other attributes, to predict the facial appearance distribution for any location in the world, and 3) mapping applications that allow novice and expert users to explore our dataset and the learned models. We hope this work will serve as a foundation for a more ambitious platform for understanding human appearance.

3.2 Infrastructure

A complete overview of the system we have developed is shown in Figure 3.2. We leverage the availability of images with geotags from Flickr as a source to download imagery on the

fly. In theory, this could be any other social media website with imagery that were to appear in the future. We ingest URLs and their metadata into our system which operates using MapReduce [23]. MapReduce is a three step process, that involves mapping inputs using a key-value pair, sorting the keys lexicographically, then the final reduction stage which groups the sorted keys and combines the groups into a new set outputs. A series of chained maps and reduces are performed to complete our infrastructure. The cluster we operate this on is a 15-node cluster, each machine having 4 GB of RAM. We use Hadoop, the Apache implementation of MapReduce, along with the Avro binary format for serialization, and Python along with a development stack similar to that found in the Anaconda Python distribution (numpy, scipy, matplotlib). There are various architecture choices, such as the allocated memory of each container on a worker, which are important for maximum parallelization.

We also use this infrastructure to build our maps. We use Google Maps TMS (tile manager system) to show facial averages conditioned on their geographic location relative to the particular tile. Our method for tile generation uses an agglomerative (bottom-up) approach. The function is given both the face and its respective geographic location in latitude/longitude format. We then convert the latitude/longitude coordinates into TMS tiles and find the parent of the particular cell for a given zoom level. The mapper emits a key that is the x, y, z tile components. The output key/value pairs are sorted and grouped and sent to their respective reducers. While the inputs are of the same x, y time coordinates we compute a streaming average by summing all of the input faces for the x, y tile up until the tile coordinates change. Upon their change, we then divide by the number of faces seen. These outputs are stored to the disk and then passed back through the same algorithm, however processing for zoom level $z - 1$. This approach continues with increasing speed each iteration, due to reduced input, all the way to the most coarse zoom level, zero. We describe our mapping approach mathematically in Section 3.5.

3.3 The WGT Dataset

We have curated a large dataset of geotagged face images, referred to as the WhoGoes-There? (WGT) dataset, to support algorithm development and evaluation. We processed all geotagged images in the Yahoo Flickr Creative Commons 100M (YFCC100M) dataset [132], which contains 100 million images and their associated metadata. The metadata consists of 34 attributes including the date uploaded, user and machine tags. Approximately 49 million are geotagged.

We acknowledge two similar datasets, MegaFace [97] and GeoFaces [56], and provide

Table 3.1: Comparison of existing large-scale face datasets with geotags.

	Geofaces [58]	MegaFace [97]	Ours (WGT)
# of face patches	248 000	1 000 000	2 106 468
# of unique geotagged images	< 248 000	84 614	1 703 749
Alignment	Similarity	None	Similarity, perspective
Detector	Commercial (Omron)	HeadHunter [92]	HoG pyramid [64]
Fiducial markers	12	49 [143]	68 [64]
Addl. Metadata	Bounding boxes	Bounding boxes	Bounding boxes, city, country, age, gender



Figure 3.3: Word cloud of user tags in the WGT dataset.

a comparison to our dataset in Table 3.1. The MegaFace dataset, like our dataset, is constructed from images in YFCC100M, however MegaFace uses additional sources of data and was developed to support evaluation of face recognition systems. We downloaded MegaFace and checked for the number of unique geotagged YFCC100M images, finding this to be 84 614, versus 1 703 479 in our dataset. The GeoFaces dataset is quite similar to WGT, with a few important differences. WGT is built using a publicly available open dataset, whereas GeoFaces was crawled from Flickr using face-related search terms. Our dataset contains significantly more fiducial keypoints, detected using a state-of-the-art algorithm [64], and nearly an order of magnitude more face patches. In addition, we will be releasing extra image and metadata features, which we detail in Section 3.3.2.

3.3.1 Face Detection and Filtering

Using all geotagged images, we detect a bounding box for the face, extract the face patch, and align it using the detected landmark points. Facial bounding boxes and their landmark points are extracted using the method of Kazemi et al. [64] which is known to have an extremely low false positive rate. This process resulted in 2 106 468 geotagged face patches, each containing 68 fiducial landmarks.

Once the fiducial landmarks have been detected for a face, we extract four different face patches. The first patch we extract is an area that is 40% larger than the detected bounding box and is scaled to 256×256 . The second patch is the tight cropped version of this patch, which has dimension 153×153 . Both face patches are aligned to a canonical position using a similarity transformation between the centers of each eye and a set of reference eye centers. We also align using a perspective transformation, conditioned on gender, between the entire set of detected landmarks and a gender-specific reference set of landmarks. We do gender-specific alignment because of the subtle differences in male and female facial structure.

3.3.2 Feature Extraction

We extract three features for each of the face patches: PCA (appearance), VGG Face [107] FC8 (identity), and additional metadata. We randomly sample 200 000 faces from our dataset and learn a PCA basis using the similarity aligned, tight cropped patches as input. The remaining approximately two million images are reserved for experiments and evaluation. The tight crops of faces are used in order to reduce the variance that would otherwise be present by considering the 40% expanded region patches. In our experiments, we find that 200 000 is a sufficient number of training images to learn a basis that captures facial appearance.

Identity features are extracted in a similar manner. We use the VGG Face network and extract features from the network’s FC8 layer which correspond to the semantic labels. This network is trained on a dataset of 2 622 different celebrity faces for identity recognition. We find that these identity features are more invariant to lighting and pose than using the PCA appearance features.

We are also interested in a data-driven approach to learning demographics at a world-wide scale. To support this, we provide age and gender estimates using the convolution neural networks of Levi et al. [81] and reverse geocodings that include country and local administrative regions. Visualizations of the distributions of age and gender in our dataset are shown in Figure 3.4. The WGT dataset, including the extracted face patches, appear-

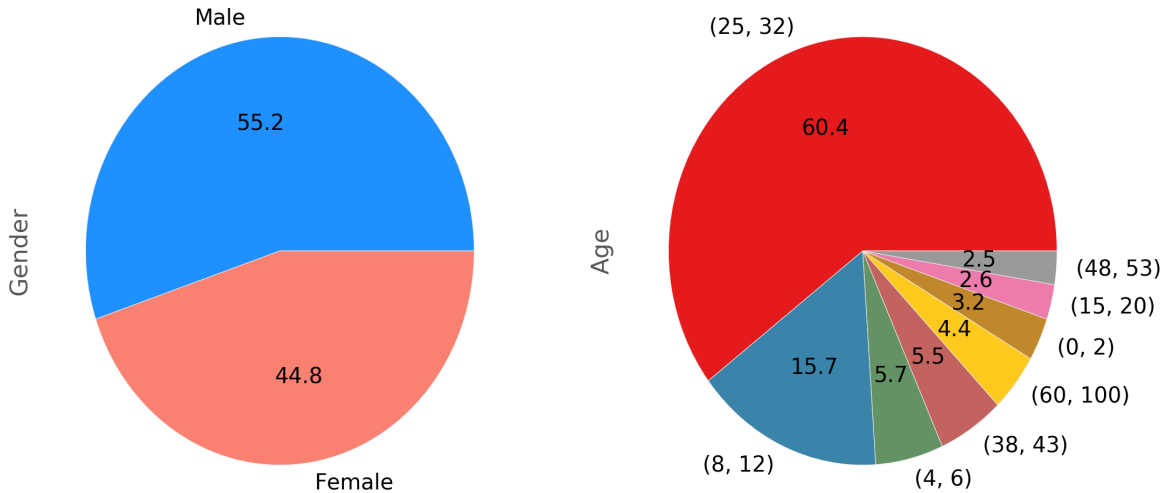


Figure 3.4: Distributions of age and gender in the WGT dataset.

ance and identity image features, detected fiducial points, and age/gender estimates will be publicly available pending publication. The following sections discuss our different approaches for visualizing worldwide appearance.

3.4 Dataset Visualization Using Conditional Averaging

The proposed dataset is a resource for understanding the geospatial distribution of human appearance. In this section, we consider simple visualizations constructed by averaging subsets of the dataset, in other words constructing conditional average images. In addition, this will serve as a baseline for comparison to more sophisticated techniques we explore in subsequent sections. Zhu et al. [158] show that using the technique of image averaging can reveal impressive visualizations of certain classes of scenes, objects, and animals. In a similar manner, we show that the average appearance of people is affected by their spatial distribution.

We highlight the visual differences of image averaging using different alignment strategies in Figure 3.5. We begin by selecting individuals from India and aggregating them by their age and gender. We sample 200 faces from each age/gender group. The top three rows of Figure 3.5 are females sorted by age along the columns. Each row within in the top three rows uses a different alignment strategy, where the first row is the similarity aligned average, the second row is the perspective aligned average, and the third row uses the iterative image refinement method of Kemelmacher et al. [65], referred to as Collection Flow. The average of similarity aligned faces are sharp near the eyes, but are blurry elsewhere. We can improve the average face appearance by using perspective aligned faces instead.

Some attributes we observe are that females appear to smile more, from the shape of their dimples and lips. We can also see the development in mustaches in men, and in both genders we can see how hair turns gray with age. We use the perspective aligned faces in two web-based visualizations described in Section 3.5.

Average images are a quickly computed, good way to visualize a set of images. However, observing the average image alone is limiting. It does not allow us to extrapolate and gain insight from a variety of appearance-related factors. If we want to conditionally average on additional attributes, information will inherently become lost. This motivates us to consider regression methods to predict appearance.

3.5 Interactive Maps of Facial Appearance

One primary goal of our work is to develop practical applications to visualize worldwide appearance diversity. We have designed two web applications to qualitatively demonstrate both novelty and practicality of our models. These applications will allow any naïve user to explore worldwide facial diversity. In this section, we describe implementation details and show screen shots from each application.

Conditioning on Geopolitical Boundaries In Figure 3.8, we condition on geopolitical boundaries. That is, we discretize latitude and longitude into their respective countries. We then compute the average perspective aligned face for each country. These averages show sharp boundaries in appearance as we move from country to country. To better understand how the distribution of average appearance changes at a finer scale, we design a multiscale, purely spatial model.

Conditioning on Multiple Spatial Scales In this section, we describe a method for multiscale visualization of human appearance. We begin by discretizing the world into a set of spatial bins. The bounding box, b , is defined as the set of all intersecting bins with the bounding box. With our discrete set of bounding boxes, we can solve for $P(f|b)$, where f is the facial appearance within b . This results in millions of bins that may contain a sparse number of faces. Since the WGT dataset contains millions of images and the world contains millions of bins, it is imperative that our method be able to scale to handle large bounding box queries and the addition of new images. To handle the addition of new images, we apply an on-line method by maintaining the sufficient statistics of a Gaussian distribution for each bin. In each bin, we maintain the count of images in the bin, c , and the running sum of the images, $I_{sum} = \sum_k^c I_k$. These two values allow us to generate the mean

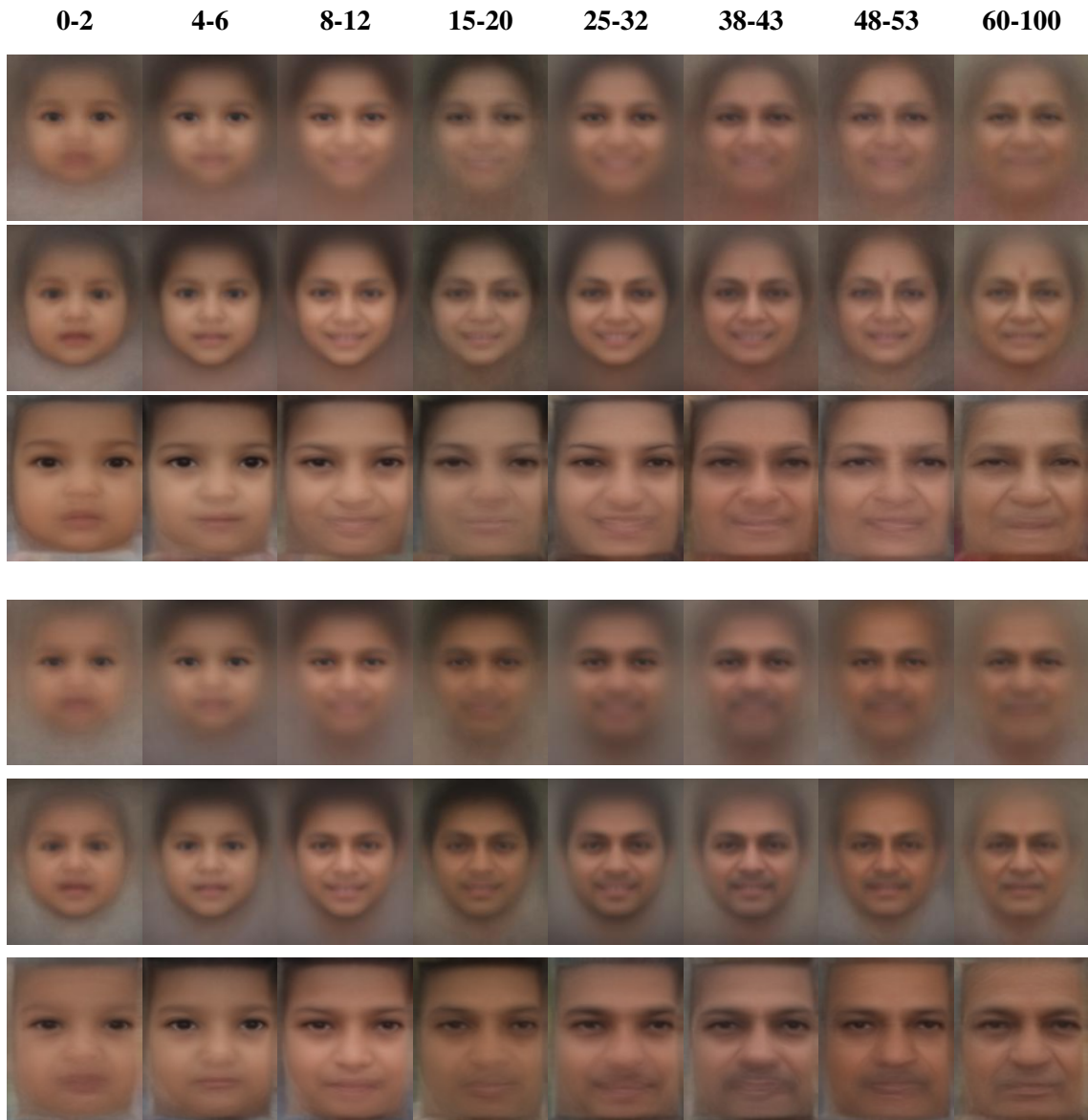
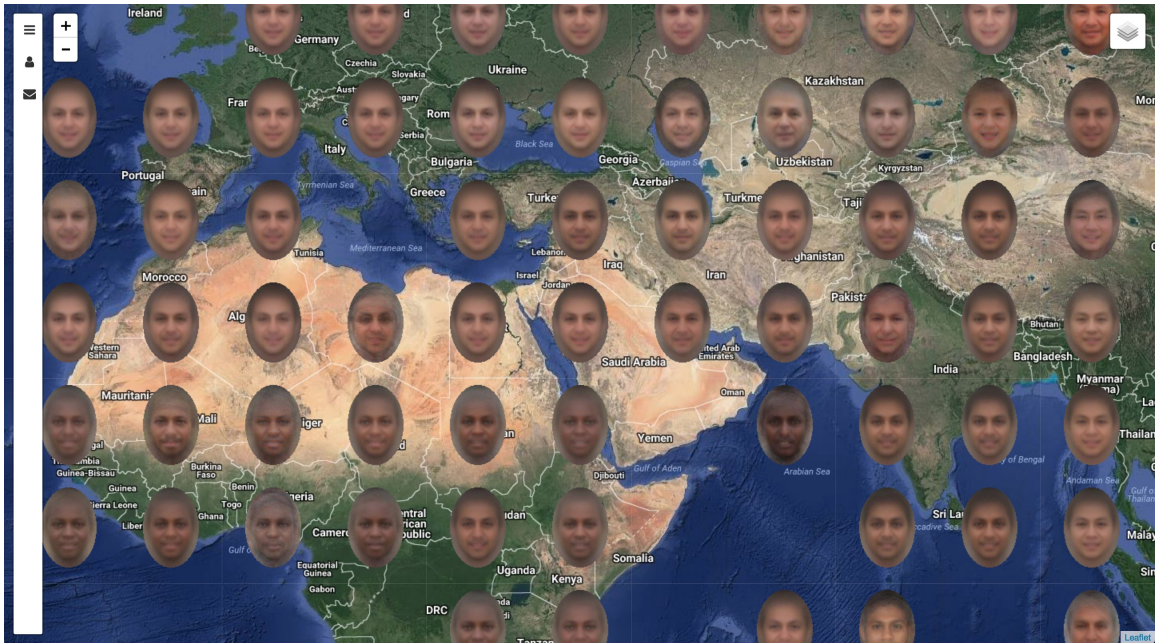


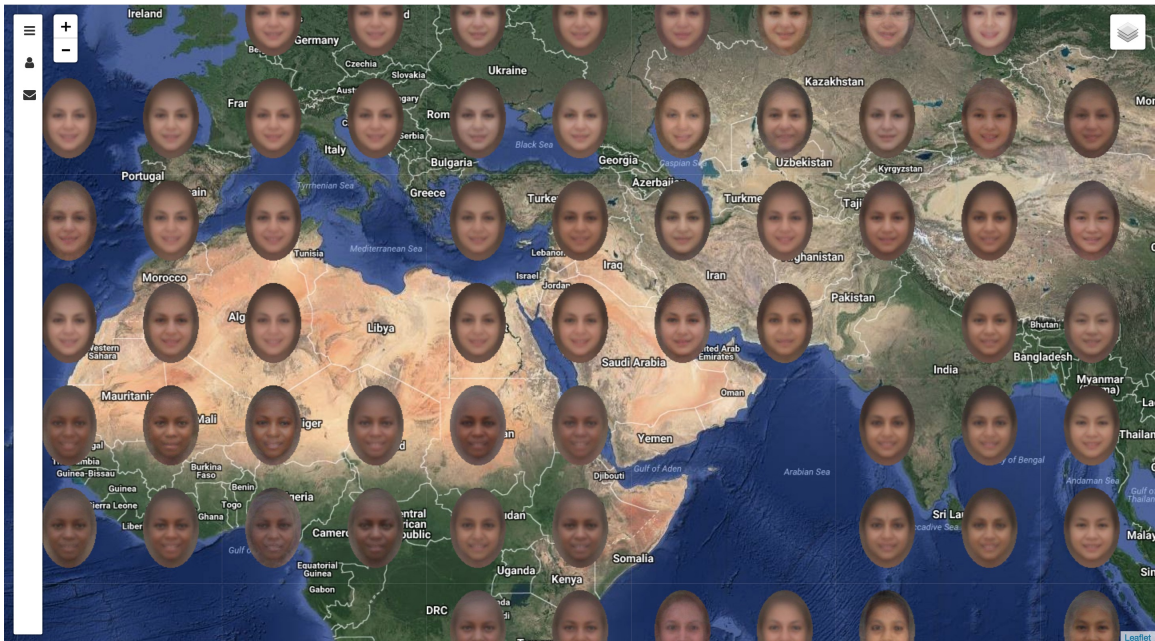
Figure 3.5: Top three rows: Average Indian females arranged from youngest to oldest averaging using similarity alignment (first row), perspective alignment (second row), and Collection Flow [65]. (third row). Bottom three rows: Average Indian males using the same approaches.



Figure 3.6: Multiscale visualization. Zooming in reveals finer details about world populations.



(a)



(b)

Figure 3.7: Web interface showing the capability to toggle between the average male (a) and female (b) appearance based on spatial boundaries.



Figure 3.8: Web interface showing the average male and female appearance in China.

and covariance in an efficient manner for any given bounding box, enabling us to visualize the average facial distribution of any queried region.

Figure 3.6 shows a screenshot from the multiscale web application. The top image is at a higher zoom level, meaning that facial appearance from Africa, Europe, Asia, and Australia all pool together to form the average face. The bottom image shows that, at a lower zoom level, over the continent of Africa we find finer-grained appearance. The user is also able to toggle the age and gender representation, be it male, female or a gender neutral representation.

Figure 3.7, shows the average male (top image) and female (bottom image) appearance. Immediately, we can see the smooth transitions in appearance between males and females when panning and zooming across the map, as opposed to the country-level conditioning which shows rigid transitions when moving from country to country. Together, these web visualizations highlight both the subtle and profound distinctions in population appearances.

3.5.1 Conclusion

In this chapter, we described our new dataset of faces, compared with other datasets, and described the weak labels that were added to support further research. We described a simple way to visualize the distribution of faces using aligned, conditional averages of faces. We incorporated the conditional averages into an interactive, web-based mapping tool to visualize the distribution of appearance at multiple scales.

Chapter 4

Baseline Generative and Discriminative Models of Facial Appearance

4.1 Introduction

The conditional averages shown in the previous chapter are a nice way to visualize the dataset and are quick to compute. However, this visualization comes at the harsh limitation that it is unable to extrapolate, and are therefore incapable of gaining insight about the individual effects of each factor of variation. In this section, we describe a regression method that allows us to control these factors and visualize the effects of manipulation.

4.2 Related Work

4.2.1 Modeling Facial Appearance

Understanding human faces has been a long-standing problem in computer vision and much research on understanding faces has been done, especially related to detection [64, 92, 113], recognition [107, 34], pose estimation [35, 45, 159], and attribute estimation [50, 86, 135]. Many issues can arise when observing facial images, including unsuitable lighting conditions and challenging camera angles. One of the most fundamental problems in constructing facial feature representations is to achieve invariance to these issues. Holistic and parts-based engineered feature representations require tremendous amounts of preprocessing and complex learning methods to achieve invariance and preserve semantics. Recent advances in convolutional neural networks have led to learned feature representations that are both compact and maintain semantics of identity, while being invariant to lighting and pose. DeepFace by Taigman et al. [131] uses a Siamese network that optimizes the

L_1 distance between positive and negative exemplar faces. Schroff *et al.* [123] proposed FaceNet, which uses the GoogLeNet architecture and minimizes the triplet loss between the query image, an image of the same person, and an image of a different person. Parkhi *et al.* [107] propose VGG Face, which is similar to FaceNet [123], however they train on a large dataset of celebrity faces and use the VGG architecture. Many works on generative adversarial networks have been applied to faces, including both unconditional and conditional GAN variants [37, 14, 112, 28, 83, 77, 130]. These works, while successful, neglect location and time information that would be useful in generating samples of human faces.

4.3 Predicting Facial Appearance

The appearance of a face is dependent on many factors, including the individual’s age, gender, face shape, and pose. In addition to these proximate factors, the appearance is also dependent, albeit indirectly, on the geographic location where the image was captured. The location impacts various elements of appearance, including fashion choices, ethnicity, and typical facial expressions. We propose using regression to predict facial appearance from varying combinations of these factors to better understand the relationship between facial appearance and geographic location. In this section, we focus on unimodal regression methods and minimize the L_2 loss function for all models. In some sense the generated faces can be considered conditional averages. The key difference from the approach in the previous section is that we now have a natural means to interpolate. In Section 4.4 we consider multimodal models.

We propose models of how age, gender, facial shape, and location affect the expected value of appearance given these factors. We begin by preprocessing the data. Age and gender features, the categorical age and gender labels, assigned by the convolutional neural networks in [81], are one-hot encoded. Location is represented as a 3D vector in ECEF coordinates and shape is a 2D vector in pixel coordinates. Location and shape are normalized to the range, $[0, 1]$. Using these input features, we fit a linear regression model to predict facial appearance, by predicting the top 2048 PCA coefficients, for different subsets of predictor variables. Samples from this linear regression model for different values of age and pose are shown in Figure 4.2. The result is a montage that clearly reflects changes in these two parameters. This approach did not work well for other the location variables.

Using a 80/20 training and testing dataset split, we fit a random forest of ten trees by taking as input our objective attributes, age, gender, facial shape, and location, to predict the same PCA appearance features as defined above. We trained a model for each individual attribute and one that combined features from all attributes.

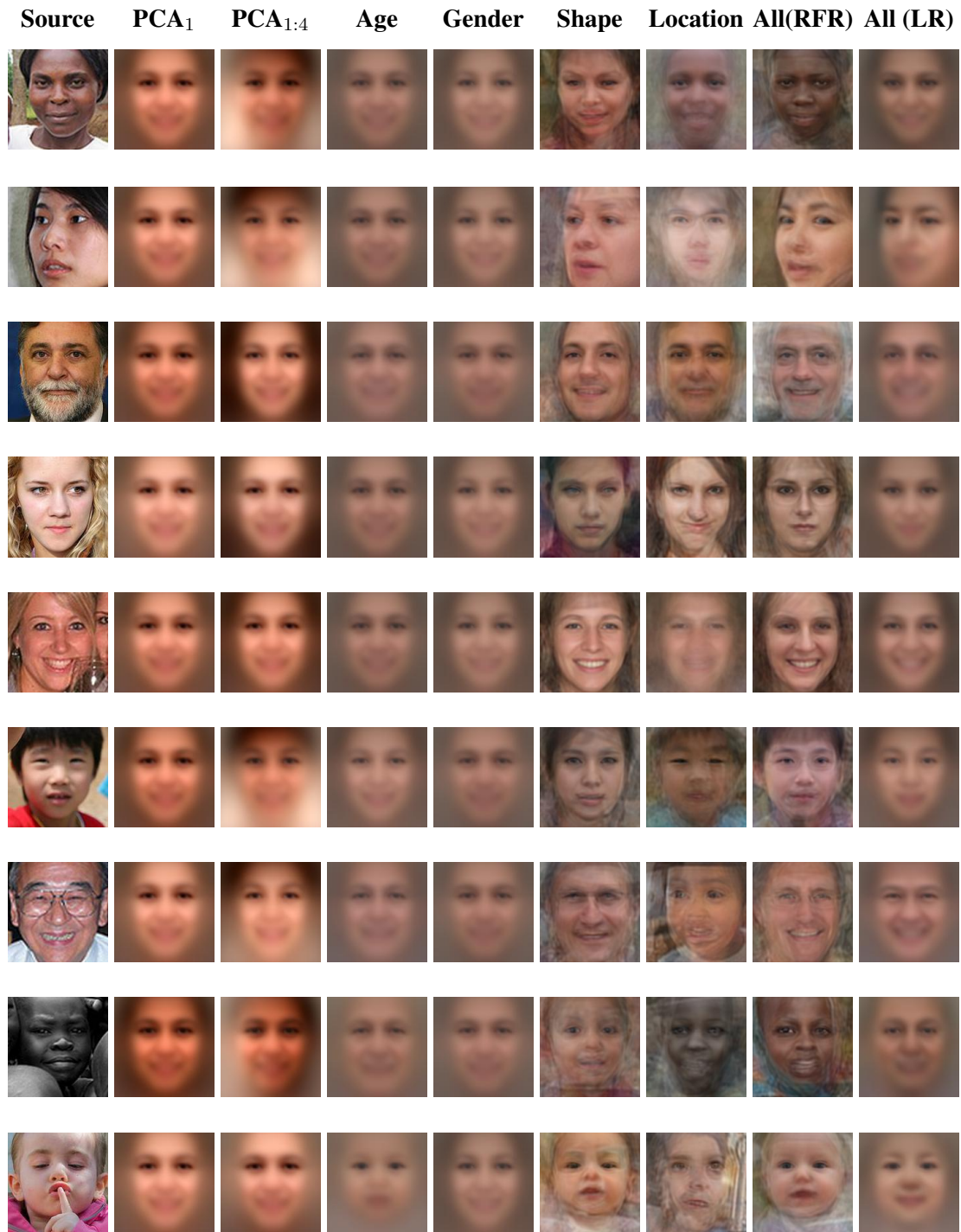


Figure 4.1: Results of learning models conditioned individually on age, gender, location, and face shape and then conditioned on all four of these attributes. The predicted components are then used to reconstruct the original image.



Figure 4.2: Reconstructions from our linear model of facial appearance. The x -axis corresponds to age and the y -axis corresponds pose. We found that a linear model resulted in reasonable reconstructions for some parameters, but not the location parameter, which is highly non-linear.

Table 4.1: Average RMSE of PCA reconstruction for each regression method conditioned on various types of input.

	Linear Regression	Random Forest
Age	10.328	10.328
Gender	10.322	10.324
Shape	10.320	10.298
Location	10.331	10.245
All	10.313	10.281

Figure 4.1 shows faces generated from our random forest model (RF) based on the attributes of a given real face. The first column shows the source patch and compares the effects of reconstruction using a single PCA component and the top four PCA components in the second and third columns. The second column captures illumination factors and the third column captures lighting from various angles. The following columns show the predicted reconstruction based on each objective attribute. Examining the second to last column of Figure 4.1, we observe that conditioning on age results in an average baby and conditioning on gender shows an average female. However, conditioning on shape reveals significantly more information than either age or gender. Location tends to generate features that are indicative of ethnicity. Finally, by conditioning on all objective attributes, we can see a baby whose appearance is qualitatively most similar to the source face patch.

We report the average root mean reconstruction error from the test set in Table 4.1. Both models minimize the reconstruction error to a relatively similar number, however the random forest model preserves significantly more facial structure. We find that conditioning on location achieves the lowest RMSE quantitatively. We observe a stark difference, from a qualitative standpoint, between the results from our random forest model and our linear regression model. The linear regression model conditioned on all objective attributes (right most column) tends to reconstruct an image that is highly similar to the mean, however the random forest model that conditions on the same set of objective attributes reveals significantly more discriminative features such as pose, lighting, age, and gender.

4.4 Multi-modal Distributions

In locations with diverse populations, a conditional average face, or any individual exemplar image, is likely insufficient to accurately reflect the diversity of facial appearance. To overcome this, we propose learning a conditional multi-modal distribution. The key idea is to cluster faces in image feature space, assign each face to a cluster, and then learn to predict cluster membership for a given geographic location. Once this model is trained, we

can provide a location and obtain a distribution over the types of faces we can expect to find. We are essentially fitting a mixture model, $P(f|\ell) = \sum P(f|c)P(c|\ell)$, where f is a facial feature, c is a cluster, and ℓ is a location. In the remainder of this section, we describe our approach for obtaining clusters, fitting the conditional distribution, and one analytical application of the model.

4.4.1 Clustering Faces

Our goal is to cluster faces into groups such that members of a given group have similar facial appearance. At one extreme, we could group all faces into one cluster, this is essentially the approach used in the previous section. This approach does not allow us to model the multi-modal nature of human appearance. At the other extreme, we could attempt to make each cluster only contain images from a single individual. This approach would make learning a conditional distribution difficult because there would likely be few samples per label. Experimentally, we found that clustering into $k = 250$ groups was a good compromise. Initially, we also found that using appearance features (PCA) resulted in clusters mostly grouped faces by pose and lighting conditions. We later discovered that clustering the identity features (VGG Face FC8), which were discriminatively trained for face recognition (and hence invariant to pose and lighting), resulted in more semantically meaningful clusters.

For each similarity aligned face patch in our dataset, we extract its identity feature, $f_i \in \mathbb{R}^{2622}$. We then cluster the identity features using iterative k -means clustering on a subset of our dataset. Initially we sampled faces uniformly at random from the training set, but the resulting clusters did not contain any clusters that were mostly of African descent. This is not surprising since relatively few images were captured in Africa. To minimize the impact of this dataset bias, we use a stratified sampling approach. We first discretize the world into a 10×10 grid, linearly sampling in latitude and longitude. To form our stratified training set we randomly sample faces separately from each bin, ensuring that no more than 500 faces are sampled from each bin.

The final step of the clustering process is to construct an exemplar face. Given the cluster assignments, for each cluster we select the 5000 faces nearest to their cluster center. We then compute the average landmarks for the 5000 closest faces and preserve top 800 faces whose landmarks are nearest to the average landmarks to ensure the face is mostly front-facing. We then take these 800 faces and apply Collection Flow [65] to structurally refine the exemplar face. The resulting image is then assigned as the exemplar image for each individual cluster. A subset of exemplar faces we found are shown in Figure 4.4.



Figure 4.3: These montages show exemplar faces for the eight clusters that are more likely to be from the corresponding location.

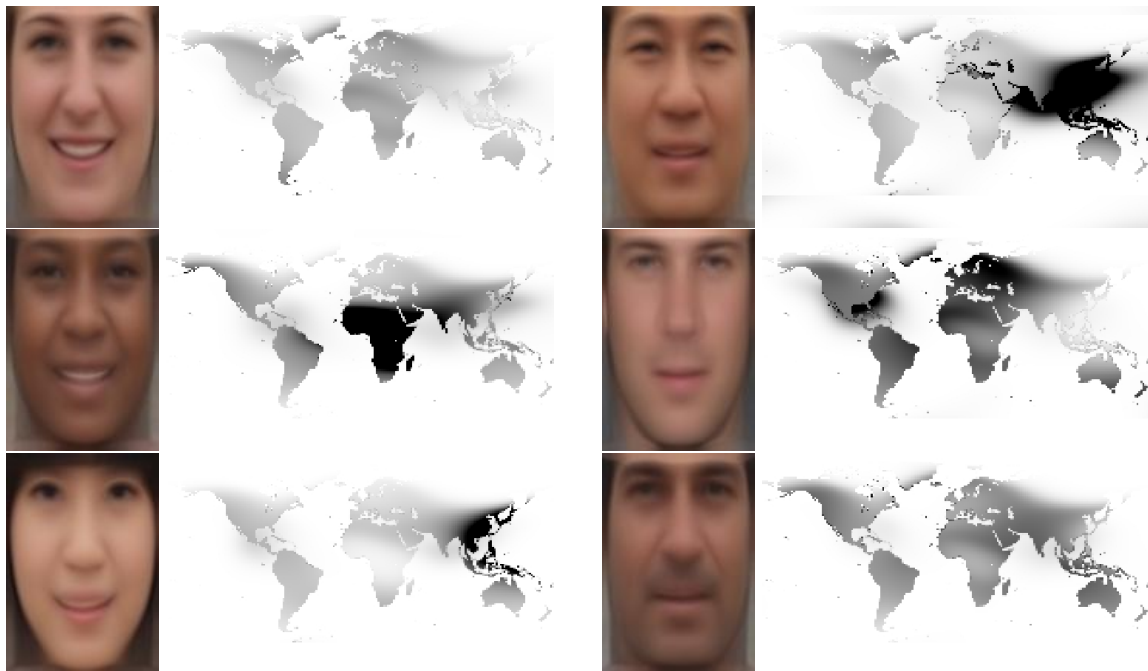


Figure 4.4: (left) An exemplar face for a given class, c_i . (right) The conditional distribution of that class for each location, $P(c = c_i | \text{location})$.

4.4.2 Conditional Distribution of Clusters

With our faces now clustered in feature space, we use a neural network to represent the distribution over cluster assignments. Our network takes geographic location and landcover class as input and outputs the conditional probability of the cluster assignment. We found that including the landcover class improved the accuracy of our model significantly and made training converge more reliably. The network is feed-forward with three hidden layers, with 100, 100, and 50 nodes respectively. All activations are hyperbolic tangent and L_2 regularization ($\lambda = 1e^{-5}$) is used. We train the network using stochastic gradient descent (batch size = 10 000) with a cross-entropy loss function.

Locations from around the world are sampled and we visualize the top eight clusters with the highest membership likelihood. Figure 4.3 shows, for six locations, the eight face clusters that are most common in that region, based on the output of our neural network, $P(c|\ell)$. From this we can see clear trends in facial appearance. For example, the most common clusters from the location sampled in Asia appear to be of Asian descent. Figure 4.4 visualizes $P(\ell|c = c_i)$, which we estimate using Bayes rule, for six different clusters. This distribution reflects where you would be most likely to find a face belonging to the given cluster center, c_i . Specifically, for each map we sample from $P(\ell|c = c_i)$ for a particular cluster, c_i , at a dense grid of geographic locations. The darker the location on the map the more likely it is that a face seen at that location will be from the cluster. Similar to Figure 4.3, these maps highlight that our model has learned distributions that reflect the appearance of individuals we can expect to see in different locations. In the following section, we show how we can group these clusters into higher-level categories.

4.4.3 Subpopulation Factor Analysis

The population at any point on the earth is composed of a mixture of underlying subpopulations that exhibit variability in appearance factors such as age and sex and regularities in appearance factors such as ethnicity. Almost every location has some children, adults, and elderly, while some locations can be either ethnically and culturally homogeneous or heterogeneous. This subsection describes our approach to estimating these subpopulation factors as a distribution over face clusters.

We represent each face as a pair, $f_i = l_i, c_i$, with location, l_i , and cluster assignment, c_i . We sampled 10 000 locations on the earth randomly and computed the histogram of the 200 nearest faces, $p(c|l)$. Using pLSA [49], we estimate a mixture of k latent subpopulation models that minimize the Kullback-Leibler divergence between the observed population



Figure 4.5: The automated geographical subpopulation factor analysis with five latent factors shows spatial regularities in location posterior (left) and the most differentiable appearances for each factor (right).

distributions (counts) and our latent estimate:

$$p(s_i, l_j) = \sum_{f=1}^k p(f)p(c_i|f)p(l_j|f) \quad (4.1)$$

where $p(f)$ is the weight of factor f , $p(l_j|f)$ is likelihood of the location given each factor, and $p(c_i|f)$ is the likelihood over appearances given each factor. This is analogous to factor analysis on documents, where documents are location samples and words are visual appearances.

Figure 4.5 shows a coarse set of five latent factors estimated from 10 000 location samples with our dictionary of 250 cluster centers. On the left are maps showing the posterior likelihood of subpopulation factor given location using 5×5 degree samples, $p(f|l)$. The right shows the appearance models ordered by how discriminative they are for each factor, more exactly, the ratio of posterior likelihood to the second most likely factor. Though there is significant noise in individual appearance model scores, the factors appear to encode to major ethnic groups: African, Asian, Indian/South American, Asian, and two Indo-European clusters.

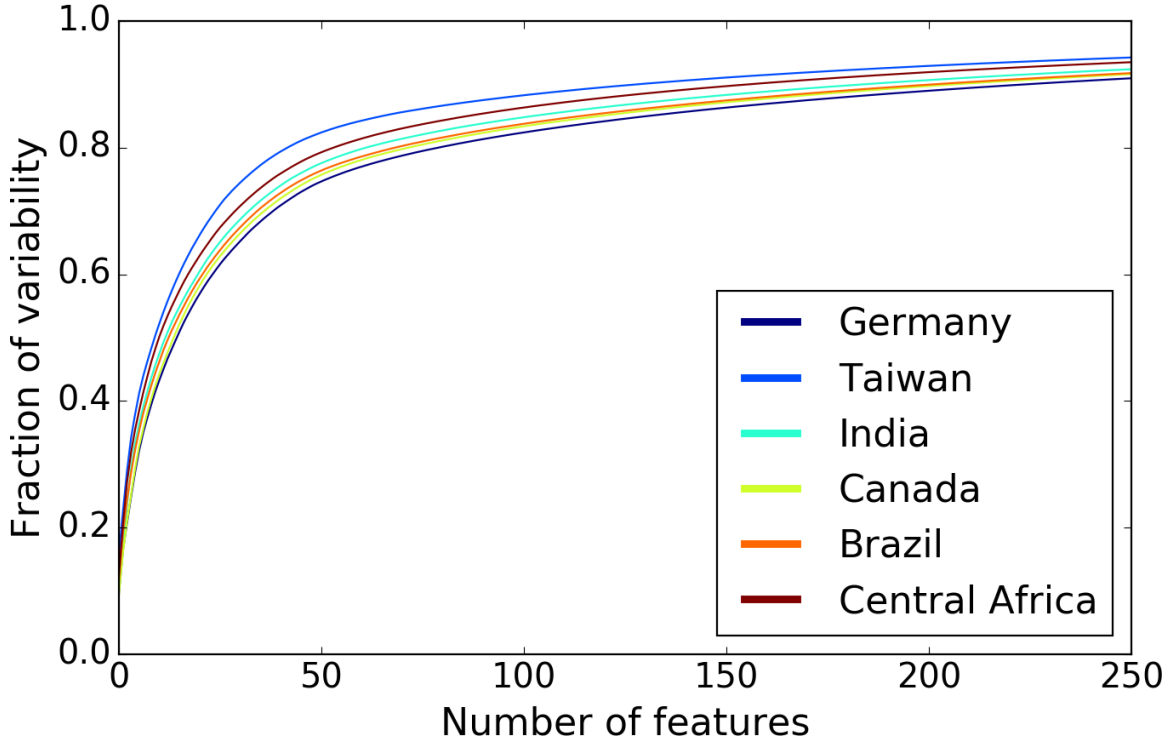


Figure 4.6: Quantifying appearance diversity using the fraction of variability explained by the top k PCA components of the *FC8* identity features. For a given number of components, larger values imply less diversity, because more of the variability is explained by the top k components.

4.5 Evaluation

In this section, we provide a quantitative evaluation of our work by measuring country-level appearance diversity.

4.5.1 Quantifying Appearance Diversity

Intuitively, appearance diversity around the world differs as move from one location to another. Using the *FC8* image identity features, we quantitatively measure diversity of a population by their fraction of variability. We begin by querying a set of largely populated countries scattered throughout the world. Since Africa has a relatively sparse number of images, we select several countries from Central Africa to compare diversity. For each country, we compute the covariance of the identity features and apply SVD to the covariance matrix. The fraction of variability is defined as:

$$\lambda_n = \frac{\sum_{i=1}^n \lambda_i}{\sum_{i=1}^N \lambda_i}, \quad (4.2)$$

where λ are the set of eigenvalues calculated by SVD, n is the i th λ , and $N = |\lambda|$. This metric allows us to examine multivariate variability. In our case, this implies that the more appearance diverse a region is, the lower the fraction of variability will be. Conversely, less diverse regions will have a higher fraction of variability. Figure 4.6 shows the fraction of variability for the selected countries. If we compare Taiwan with Germany using the top 50 eigenvalues, their fractions of variability are 0.744 and 0.822, respectively. These values tell us that Germany is 9.49% more diverse in appearance than Taiwan when considering how much diversity is captured within 50 dimensions.

4.6 Conclusions

Overall, we have curated a large-scale dataset of geotagged faces and shown many ways that we can both qualitatively and quantitatively model worldwide appearance and diversity. We have demonstrated there are a variety of ways this data can be visualized and shown several pragmatic applications of our work. We will be releasing our dataset and web-based visualizations for public use to spur interest in developing and exploring practical applications using our dataset. Our hope is that this work will serve as a multidisciplinary foundation towards furthering our understanding of human appearance diversity.

Chapter 5

A Deep Generative Model of Facial Appearance

5.1 Introduction

Differences in human phenotypes, the amalgam of observable characteristics, are dependent on many factors. These factors may be biological, such as gender, age, and ethnicity, or more ephemeral, such as personal style and mood. Together, the biological and ephemeral factors both depend on geographic location, time of day, and current/forecasted weather conditions. This dependence has been demonstrated for make-up and facial hair choices [24], the frequency of various facial expressions [73], and types of clothing [44].

This motivates us to consider a model that explicitly captures geographic location and its relationship to human appearance. To build this model on a global scale, we propose using a large dataset of geotagged images collected from a popular photo sharing website. While the model we learn will inherit the biases inherent in the underlying data source, it is sufficiently diverse to enable us to highlight the capabilities of our model and learn the latent structure present in the data.

We propose a novel generative model, *GPS2Face*, that captures the complex relationship between human appearance and geographic location. We utilize adversarial autoencoders (AAEs) [91], which have shown great promise in providing a way to generate samples from complex distributions, such as natural images of faces, by using a distribution from which it is easy to sample. The distribution of face images is complex due to drastic differences in pose, illumination, expression, and occlusion, especially when considering faces that are captured in unconstrained settings. Capturing the complete relationship between an image and other proximal factors, such as age, gender, and location, enables us to



Figure 5.1: We propose a generative model that incorporates geospatial metadata, along with additional human-related attributes, and allows for synthesis of people within a given area. The color of the bounding box in the map corresponds to randomly generated women from their respective regions.

sample faces from anywhere on Earth. A geographically conditioned generative model has many potential uses, including discovering emerging trends in facial appearance (which could be due to mass migration), providing an interactive visualization for educational purposes, or the natural evolution of style.

Our approach significantly improves upon previous works that attempt to model the relationship between geographic location and facial appearance. Disciplines including anthropology [46] and evolutionary biology [82] have historically relied on manual methods of field research to acquire human phenotype data. These datasets are often small, expensive to collect, and prone to human bias. Our work is novel in that it uses a significantly larger sample size and relies less on human biases and predispositions. In principle, this means it has the potential to overcome some of the pitfalls of previous works if we are able to train our model using a truly unbiased dataset.

There is a long tradition in using discriminative computational approaches to understand human phenotype from imagery. This work has largely been conducted in the surveillance and biometrics community [21]. While these approaches are interesting and relatively easy to evaluate, they are limited in that they do not provide a generative process that makes it possible to understand what the model has captured about the human phenotype distribution. Our proposed model is generative and, when compared with previous data-driven approaches [10], produces more realistic faces, enables a variety of facial manipulations, and provides an explicit method for sampling facial appearance for different attribute settings. Furthermore, our model is fast so it can be used to directly support other applications such as interactive visualization, as shown in Figure 5.1.

Our work makes the following contributions: 1) Significantly improved image quality compared to previous geolocation-conditioned generative models of human facial appear-

ance, 2) a novel pose representation that enables continuous pose manipulation, compared to discrete poses used in previous generative models of human facial appearance, 3) a factored latent variable model that makes it simple to manipulate and constrain semantically meaningful facial attributes, such as pose, age, and gender, and 4) an extensive evaluation highlighting the capabilities of our model. Our results are comparable with other recent generative models, which were trained on hand-curated datasets, despite being trained on unfiltered social media imagery.

5.2 Related Work

Soft Biometrics In the context of computer vision, soft biometrics are roughly defined as observable characteristics, such as facial geometry, eye color, and gait, that are easy for humans to perceive without special equipment. In some applications, it is desirable to estimate these characteristics directly [31, 81] but these characteristics are often used implicitly to recognize individuals [107, 123, 131]. A goal for such approaches is often to achieve invariance to unimportant factors for the given application. For example, ideally a model for predicting age of an individual will work equally well regardless of their gender and eye color. Similarly, when predicting the ethnicity of an individual the pose and lighting conditions should not affect the result. While achieving invariance is a useful goal for such discriminative tasks, it makes it difficult to visualize the relationship between human appearance and the latent factors. In our work, we use a generative model which makes visualizing this relationship relatively easy.

Soft biometrics approaches have typically ignored the geographic location at which a photograph was captured; it is assumed that a model should be invariant to the geographic location. However, there have been attempts to estimate the race/ethnicity of an individual [50, 128], which is correlated with geographic location. Such approaches discretize the space of ethnicity into a small number of disjoint categories. For our purposes, this representation is problematic because it oversimplifies a complex attribute and would therefore limit the expressiveness of our generative model. In our work we do not explicitly define ethnicity nor limit it to a fixed number of categories. Rather, we learn about the relationship between appearance and geographic location, which implicitly includes a variety of factors, ranging from ethnicity to local fashionability.

Facial Synthesis The goal of facial synthesis is to generate realistic looking faces based on an easy-to-specify, typically low dimensional, representation. Early work on this task proposed subspace models [134] and models that explicitly represented face pose [16].

More recent work has built upon the Generative Adversarial Network (GAN) framework proposed by Goodfellow *et al.* [37]. In this framework, two networks are trained: a discriminator and a generator. In the context of image inputs, the discriminator’s goal is to distinguish between real and synthesized images. The generator’s goal is to synthesize images that fool the discriminator into believing they are real. The networks are trained *adversarially* where each is attempting to defeat the other, ideally achieving an equilibrium condition in which the generator synthesizes realistic images. Through this process, the generator learns to map random samples from a low-dimensional, known prior distribution into realistic images. One architecture in particular, DCGAN [112], has been applied to a wide variety of image synthesis tasks, including facial synthesis. Recently, many approaches [4, 41, 7] have been proposed to simultaneously increase the stability and output resolutions of GANs. The stability of GAN training is an active area of research and some recent works have provided general techniques [3, 121] for doing such.

Our goal is to be able to generate faces based on a variety of latent factors. Unfortunately, in the basic formulation, GANs do not allow for explicit control of the output, thereby limiting their usefulness. A variant, called conditional GANs [98], offers a solution. This is done by including categorical or numeric metadata as an input to the generator, in addition to the random sample from the prior. There are many ways [75, 104] to incorporate this metadata and to train conditional GANs.

A key requirement of many facial synthesis tasks is that the identity of the synthesized image appears similar to the input image. One example of this is the attribute transfer task, where the goal might be to change a person’s hair color or expression. An approach that made Brad Pitt look like Donald Trump wouldn’t be of much use. Recently, several methods have been proposed for transferring fine-grained attributes such as age [67, 154] or transient attributes, such as facial hair and hair color [83, 149]. Another task in which identity preservation is imperative is facial frontalization. Given a face that is captured at an extreme pose, the task is to normalize the pose. Some recent approaches have used facial symmetry as a way to synthesize the missing part, while most recently others have used GANs [55, 150].

Similar to our model is the recent work by Tran *et al.* [133]. However, their focus is on discriminative as opposed to generative tasks which leads to different model design choices. For example, they choose to represent face pose as a single variable by discretizing only yaw. Our model uses continuous pitch, yaw, and roll angles instead, enabling finer-grained control of the synthesized images. Additionally, their model disentangles factors of variation by using a fixed set of identities in their discriminator. Our proposed method uses geographic location and operates at a worldwide scale, so relying on a fixed set of identities

would be intractable. Instead, our model learns a soft identity representation conditioned on the contextual factors of age, gender, facial morphology, and location. We use a generative architecture similar to [154], however we use LeakyReLU instead of ReLU activations in all networks, replace the transpose convolution layers with *PixelShuffle* [124] convolution layers, and add an additional component to emphasize facial morphology and pose. We found that these design changes were necessary and result in a network that is noticeably more stable during training, converges to similar quality images in a fraction of the time, and allows for increased control of factors of variation.

Geospatial Analysis of Facial Appearance Work in this area has sought to use large datasets of geotagged face images to better understand human facial appearance. Islam *et al.* [56] provides a broad overview of tasks and challenges in the geospatial analysis of facial appearance, which they called *geofacial analysis*. Work in this area, which can be seen as a sub-domain of soft biometrics, has typically taken a discriminative approach but usually focuses on attribute prediction [39] rather than other common facial tasks such as recognition. Islam *et al.* [57] used image features extracted from a pre-trained CNN to predict in which of 50 cities a face image was captured. Wang *et al.* [138] used egocentric geotagged videos with scene related characteristics, such as weather, to learn facial attributes. Most early work in geofacial analysis has focused on discriminative tasks. In a notable exception, Bessinger *et al.* [10] proposed a method for location-based face synthesis using a simple subspace representation. However, this approach generates images that lack realistic details, is unable to represent multiple modes of appearance in regions with diverse populations, and provides significantly less control over the synthesized images than our approach. Our work is the first to propose modeling the relationship between geographic location and appearance using a generative-adversarial approach.

5.3 Approach

We propose *GPS2Face*, a framework that is capable of representing the relationship between latent factors of human appearance and geographic location and allows for conditioned facial image generation. Our neural network consists of two primary components: one that predicts facial landmarks and a second that generates facial appearance. We train *GPS2Face* on a large-scale dataset of geotagged social media images of faces. Figure 5.2 shows samples from the dataset. A diagram of our network architecture is shown in Figure 5.3.

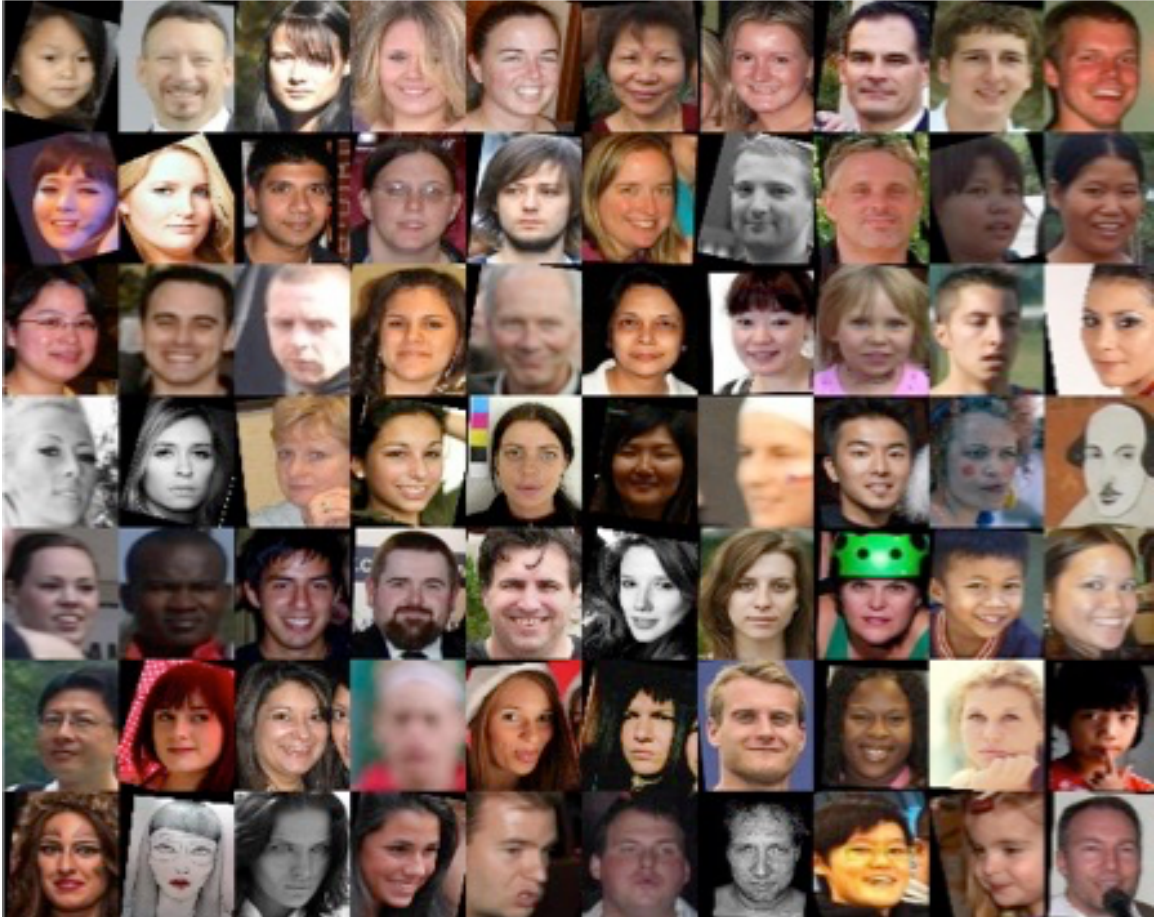


Figure 5.2: Samples from the WGT dataset [10] used in our work. Unlike face datasets that have been previously used to train generative models, such as CelebA [86], our dataset has not been manually filtered and contains a wide variety of image qualities.

5.3.1 Dataset

Since our model is a data-driven approach to understand how human appearance varies around the world, we need data that can appropriately model the problem in scale, distribution, and appearance diversity. We use the WhoGoesThere? (WGT) dataset [10] for all experiments. Unlike other recently created large-scale face datasets, such as CelebA [86] and MegaFace [66], the WGT dataset includes the geolocation data we need to train our model. CelebA is commonly used to evaluate the performance of generative models, however it is of higher image quality and captured under more controlled settings (good lighting, solid backgrounds) than the raw, social media imagery found in the WGT dataset. Similarly to MegaFace, the WGT dataset is a subset of the Yahoo Flickr Creative Commons 100 million (YFCC100m) image dataset [132], however it only includes faces from images that are geotagged. In total, it contains 2.1 million geotagged face images, along with auto-

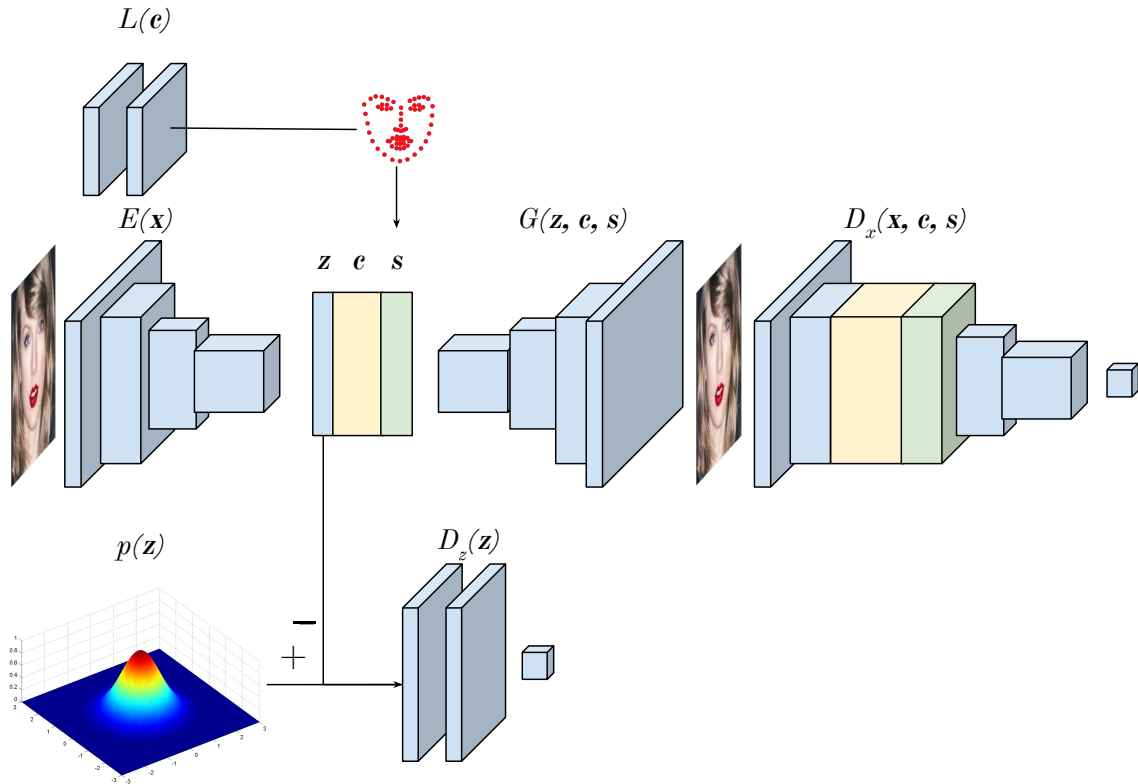


Figure 5.3: Our proposed model, *GPS2Face*, has two components: a landmark prediction network, L , and an appearance generation network supported by the other sub-networks. Landmarks are used to guide synthesis and improve the quality of generated faces since identity is not used as a regularizer. L uses latent factors, c , to predict facial landmarks, s . Predicting landmarks allows us to model how facial structure changes with respect to latent factors and also serves to avoid manually specifying a large set of landmarks at test time.

matically estimated facial landmark locations [64] and age/gender [81]. We augment this dataset by estimating the pitch, yaw, and roll of each face using the provided landmarks and the perspective-n-point algorithm.

5.3.2 Landmark Regression

The shape of one’s facial features are dependent upon many biological factors, including age, gender, and ethnicity. For example, the roundness of a child’s face is due to the lack of age-induced bone development. On average, adult men and women tend to have slightly different facial shapes. These shape differences are subtle, however we are attuned to both recognizing and differentiating them when observing the appearance of other people. Therefore, to capture the conditional dependency of face shape on these biological factors, we leverage the large quantity of images in our dataset to regress facial landmarks using

a neural network, L . The input to L is latent factors, c , consisting of age, gender, and location. The output is the predicted shape of the face, s . The landmark regression network is trained by minimizing the Huber loss [33]:

$$\mathcal{L}_{\text{huber}}(x, y) = \frac{1}{n} \sum_i z_i \quad (5.1)$$

$$z_i = \begin{cases} 0.5(x_i - y_i)^2, & \text{if } |x_i - y_i| < 1 \\ |x_i - y_i| - 0.5, & \text{otherwise} \end{cases},$$

where x and y are vectors of target and predicted landmarks. We choose this loss over L_2 for improved training stability.

We then use the predicted landmarks to guide our generative model on where to draw specific facial parts. We use the landmark locations generated by this network as input to an appearance generation network. By using face landmarks as inputs, we guide the appearance generation network to synthesize particular facial features, such as the eyes, mouth, and chin.

5.3.3 Generating Facial Appearance

Given the facial landmark locations, the next component in our network renders the image. The appearance generation component of *GPS2Face* is composed of four sub-networks: an encoder, a decoder, and discriminators for both images and latent space. The encoder, E , takes as input a face patch, \mathbf{x} , to produce a latent vector, z . This latent vector is used as input to a decoder/generator, G , to produce a synthetic image. The first discriminator, D_x , is for images and its purpose is to force the generator to produce realistic facial images. The second discriminator, D_z , is for the latent space, z . The goal of D_z is to force the encoder to map z to look like a sample drawn from the prior distribution, p_z . This constraint on z allows us to readily generate samples from p_z that are distributed in the same way as our training dataset. Our prior distribution is assumed to be uniform, $\mathcal{U}(-1, 1)$. Details of the network architecture are provided in the supplemental materials.

We denote \mathbf{x} to represent an image, \mathbf{y} are the set of latent factors, c , and landmarks s , associated with the image, z is a low-dimensional sample drawn from the prior distribution, and λ_* are parameters controlling the weight of the losses. Each iteration of our procedure for optimizing *GPS2Face* consists of four phases. In the first phase we only optimize the parameters of the image discriminator, D_x , using a true image, x , and a fake image $G(z)$:

$$\mathcal{L}_1 = \lambda_1 \cdot \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D_x(\mathbf{x}, \mathbf{y})] + \lambda_1 \cdot \mathbb{E}_{z \sim p_z(z)} [\log (1 - D_x(G(z), \mathbf{y}))].$$

This loss encourages the image discriminator to tell the difference between real and fake images. In the second phase, we optimize for the latent space discriminator, D_z :

$$\mathcal{L}_2 = \lambda_2 \cdot \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log D_z(\mathbf{z})] + \lambda_2 \cdot \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log (1 - D_z(E(\mathbf{x})))] .$$

This ensures that samples encoded by the generator appear like they are from the prior distribution so we can effectively sample. In the third phase, we optimize for the reconstruction error between a real image and a generated image:

$$\mathcal{L}_3 = \lambda_3 \cdot \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\|\mathbf{x} - G(E(\mathbf{x}), \mathbf{y})\|_1] .$$

Minimizing the reconstruction error makes sure that the colors of pixels in our generated image appear similar to the encoded image. The reconstruction loss of autoencoders is often the L_2 loss, however we choose to minimize the L_1 loss based on results from various works [59] showing that generated images using L_1 loss are less blurry and more realistic than their L_2 loss counterparts. In the fourth phase we update G and E with the adversarial penalty:

$$\mathcal{L}_4 = \lambda_4 \cdot \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log G(E(\mathbf{x}), \mathbf{y})] .$$

We use the following conditioning variables in our network: age, gender, latitude/longitude location, country code, pose, and landmarks. We found that the number of conditioning terms and their dimensionality made training the model difficult. Additionally, it was important to weigh the discriminator updates to avoid large spikes in gradient and preserve model stability. λ_1 , λ_2 , and λ_4 are each set to 0.01 and λ_3 is set to 1.0 empirically. All discrete variables (age, gender, and country code) are represented in a one-hot encoding. Pose is represented as Euler angles in degrees, and landmarks are represented by 68 keypoints in Multi-PIE [40] format.

5.4 Evaluation

We qualitatively and quantitatively evaluated *GPS2Face* using a large dataset of facial images captured in the wild by many different photographers.

5.4.1 Implementation Details

We randomly split the WGT dataset into training (80%), testing (10%), and held-out (10%) sets, stratifying by country to ensure representation for each country in all sets. To reduce the bias toward more populous countries, we sampled 50 000 faces, with replacement,



Figure 5.4: Examples of encoding an input set of images (a) in randomly selected to have certain poses, and transforming them by manipulating the latent factors. (b) shows the reconstruction using ground truth labels. (c) shows changing the latent factors used to generate (a) into females, ages 25–32, frontalized, and each row is fixed to the following set of countries: United Kingdom, Germany, Italy, India, Taiwan, Ethiopia, Iran, Sudan. (d) shows changing the latent factors to be males, ages 38–43, pitch = -35° , yaw = 45° , and each row fixed to the same countries as used in (c).

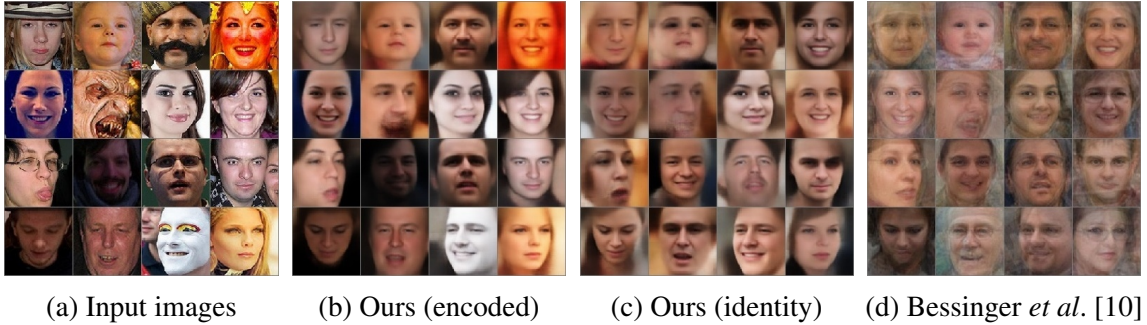


Figure 5.5: Qualitative comparison of random samples from our method and a previous method from Bessinger *et al.* [10]. Input images (a) are encoded through our network to predict z , which is used as input to our generator to decode (b). Using the same conditioning terms, in (c) we change z to be a sample from the prior. (d) is generated using [10].

from each country to form our final training set. We trained *GPS2Face* using Adam [71] for 100 000 minibatches using a learning rate of 0.0001 ($\beta_1 = 0.5$, $\beta_2 = 0.999$). Each minibatch contained 64 face patches that were resized to 128×128 and whose intensity values were scaled to the range $[-1.0, 1.0]$. We implemented our neural network models in PyTorch.

5.4.2 Attribute Manipulation

We highlight the effectiveness of our model by transforming images using various combinations of latent factors. We show several different applications of our architecture including identity-preserving pose deformations and changes in other latent factors. In Figure 5.4, we manipulate faces from the testing set in a variety of ways. Figure 5.4a shows a montage of example images, organized by pitch (y-axis) and yaw (x-axis). Figure 5.4b shows the reconstruction of the example images using our model. Each image was encoded and reconstructed using *GPS2Face* with the ground-truth latent factors. These images lose some details, such as the microphone in the upper left image, but show that our model can represent a diverse set of faces while preserving important characteristics.

Figure 5.4c shows how we can manipulate the latent variables to achieve different effects. In this montage, we changed the gender to be all female, constrained age to be in the range 25–32, and frontalized pose (setting pitch and yaw to 0°). Finally, as a test of our ability to encode for geolocation, we change the latitude/longitude location of each row to be the locations of capitol cities from the following countries (from top to bottom): United Kingdom, Germany, Italy, India, Taiwan, Ethiopia, Iran, and Sudan. Focusing on the fifth row of each montage, we observe several important aspects. Females in this row, samples 1, 2, and 5, do not have their gender changed and their respective hairstyle shapes and light-



Figure 5.6: We observe that for a fixed sample, z , we can vary pose and preserve individual identity. We fix the age and gender to be a 25 year old female. We vary the pose to be $\pm 20^\circ$ yaw and $\pm 30^\circ$ pitch. We then sample locations from the countries shown in the captions above.

ing are preserved. In addition, males in this row, samples 3, 5, and 8, have lost their facial hair and appear more feminine. These results highlight that *GPS2Face* can represent many complex aspects of appearance and its relationship to latent factors, including geographic location.

5.4.3 Qualitative Comparison with Previous Work

We qualitatively compare the results of *GPS2Face* against the previous work of Bessinger *et al.* [10]. In their work, the authors propose using latent factors to predict the PCA components and then use those predicted components to generate a face. Note that their method does not allow for a principled approach to sampling faces, whereas our method forces the latent space to obey a prior distribution with a known probability density function.

In Figure 5.5 we provide a qualitative comparison of this method versus ours using faces and attributes from the held-out set. Figure 5.5a shows real images that will be encoded and whose latent factors are used to condition each model. Figure 5.5b shows faces generated with our method after encoding input images to the latent space, then reconstructing using the conditional encoded sample. Figure 5.5c shows faces generated with our method using conditioned samples from the prior. Figure 5.5d shows reconstructions from the predicted PCA coefficients. The reconstructions in Figure 5.5d are lower-quality samples than ones we have generated due to a significant amount of artifacts and color reproduction. We quantify these claims in Section 5.4.4.

In Figure 5.6 we evaluate the effect of geographic location on facial appearance. We draw a single z from our prior, the uniform distribution, $\mathcal{U}(-1, 1)$, and leave it fixed for



Figure 5.7: We highlight appearance diversity within each country by generating faces sampled from the prior. In each montage, age and gender are randomized, while pose and geographic location are fixed.

each montage. We also fix age and gender to be a 25 year old female. We then vary facial pose pitch $\pm 30^\circ$ and yaw $\pm 20^\circ$, left to right, and vary the country in each montage. The effects of changing the country are noticeable, yet subtle, as both skin tone and facial morphology changes with location. The most representative faces are those with neutral pose (in the center of each montage).

Figure 5.7 shows montages of synthesized faces from various locations around the world. For each montage, we draw 25 values of z from the prior. We select a configuration of latent factors where age and gender are randomized, and pose is frontal. In total, we show 25 faces in each country for three different countries. One observation we can make from the synthesized faces is that our model does not handle the presence of sunglasses very well. We believe a reason for this is the inclusion of landmarks in the network, which direct the generator to produce eyes in an area that should be occluded.

5.4.4 Quantitative Evaluation

We quantitatively evaluate our method using several metrics that have been used to measure performance in many recent works of generative models. The first of these is the inception score proposed by Salimans *et al.* [121], which measures how similar a generated sample is to its predicted class and according to the authors correlates well with human judgment. In image-to-image translation works, two other fidelity metrics are also measured: the peak signal-to-noise ratio (PSNR) and structured similarity metric (SSIM). PSNR will assess how much noise is present in the generated samples, relative to the real data. SSIM compares two images and produces is a value ranging from $[0, 1]$, where 1 is the result of comparing the structure of an image with itself. Since changing the identity of a person

Table 5.1: Quantitative evaluation of our proposed method.

	Inception Score	PSNR	SSIM
Bessinger <i>et al.</i> [10]	1.475 ± 0.004	13.068	0.339
Ours (encoded)	1.7370 ± 0.007	19.131	0.513
Ours (identity)	1.609 ± 0.004	–	–
Real data	3.483 ± 0.015	–	–

changes makes the task no longer an image-to-image translation, we do not compute PSNR and SSIM on identity-modified images.

Our results are shown in Table 5.1. For inception score, the objective is to attain a score that is as high as the distribution of the real data allows. Not only does our encoded image model outperform [10], but our identity-manipulated model does as well. This metric implies that the faces our model can generate, from both autoencoded samples and random samples from the prior, are more realistic and diverse than samples generated in previous work.

5.5 Conclusions

Advances in mapping technology have made it possible to quickly see what a street corner looks like in most major cities of the world. In this work, we presented *GPS2Face*, which is a first step towards making it possible to see what people might look like on those street corners. We demonstrated that *GPS2Face* can learn the complex relationship between geographic location and various facial attributes despite the noisy nature of our dataset. The resulting model is fast to sample from at test time, enables fine-grained control over facial appearance, and generates realistic looking, and novel faces.

Chapter 6

A High Spatial Resolution Model of Clothing Style

6.1 Introduction

Retailers, marketers, and fashion designers have for decades been trying to find answers to a short, yet complex question: what factors influence how people choose to dress? In recent years, computer vision researchers have also been trying to provide answers to this question. We know that people’s clothing choices are often determined by several factors, including their friends [76], environment, and activity. These works have either neglected or overlooked the factor of geographic location. Our work seeks to answer the question of how geographic location influences people’s choice in attire.

Modeling the relationship between choices in human attire and geographic location is a challenging task for several reasons: First, there exists no public dataset of high-quality fashion imagery with precise geotags. Second, the available geotagged imagery is often lower quality and captured in real-world scenarios. This presents further issues that must be handled, such as unconstrained lighting and pose. Third, the geographic distribution of geotagged human imagery is often concentrated in cities and is biased towards countries that upload to the image source. These challenges motivate us to model the relationship between attire and geographic location with computer vision using a data-driven approach.

There are several ways to learn a distribution of human attire choices conditioned on geographic location. One way is to use professionally captured fashion photography. It is often captured under moderately to highly constrained settings where the explicit focus is on the clothing items. This kind of imagery is not only limited in availability, but also does not reflect the real-world scenarios in which the kinds of clothing may be worn. Another

way to learn such a model is to use social media websites where people share images and descriptions of garments and outfits. This approach is similar to the one proposed by Simo-Serra *et al.* [126]. The images uploaded to these social media services are captured in semi-constrained settings where lighting and pose are mostly controlled, but the background is usually noisy and may distract from the clothing items. These kinds of images are more available than professional fashion photography, and websites hosting this kind of imagery occasionally provide a crude location for the outfit, such as city and country, however they do not provide fine-grained geotags. Rather than relying on social media imagery to collect fashion data, one could use a self-collection method. The person collecting data would walk around a region capturing images, videos, and additional metadata about people and their surroundings. In this way, the data collector can acquire fine-grained geotags for every image and collect many examples for a particular individual under various poses and lighting conditions. This is similar to the approach proposed by Wang *et al.* [138]. A model trained using this self-collected data can perform well on small, local scales, such as at the city-level, and has the advantage that any additional metadata can be collected. However, a disadvantage is that difficult to extrapolate and use for learning global trends. Additionally, it requires significant efforts from those collecting and organizing the data. Our goal is to model worldwide of appearance distributions, which is made difficult by ground-level image sparsity. There are a number of reasons why ground-level images might be sparse, including selection bias or simply a lack of data that has been collected from a particular location.

Once the data is collected, you could simply train a simple classifier/regressor whose input is geographic location in latitude/longitude coordinates and learn a distribution over clothing styles. The problem with this approach is that it leads to sparse solutions imposed by the sparse distribution of geotags in the clothing imagery. While latitude/longitude can describe a precise location on the Earth, it fails to capture other potentially useful features about the environment, such as geological and man-made structures. Instead of using latitude/longitude coordinates as inputs to the classifier, we propose to use aerial imagery as a proxy. It addresses the concern of capturing man-made structures and provides a smoother estimate of the distribution of clothing choices.

Our proposed method uses a combination geotagged social media and aerial imagery from web-based mapping services to understand the relationship between geographic location and clothing style choices. An overview of our approach is visualized in Figure 6.1. Specifically, we propose using aerial imagery, which inherently encodes geographic location, as a predictor for human appearance. Using aerial imagery has several advantages over an approach that uses only ground-level imagery. A notable advantage is that aerial

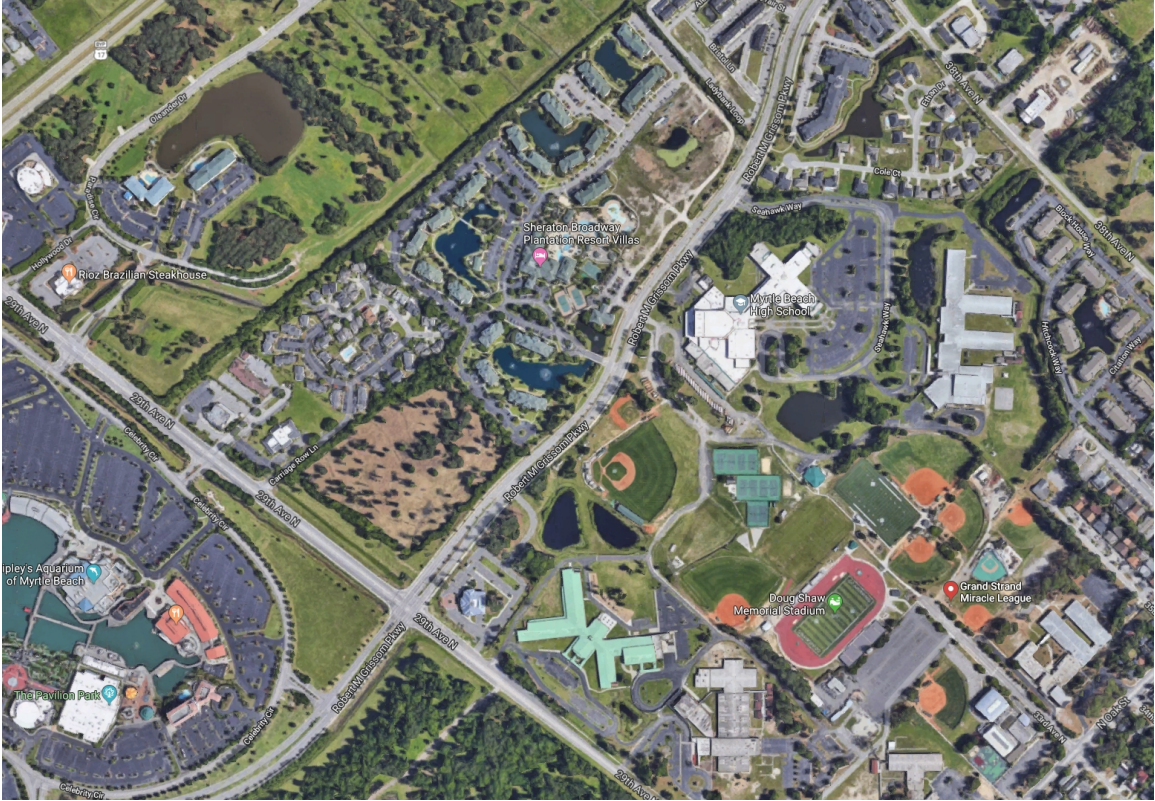


Figure 6.1: We propose a model that uses satellite imagery to predict a distribution over human attire within a given geographic region.

imagery has dense coverage compared to sparsely distributed ground-level imagery. Additionally, the natural and man-made structures people interact with can be used as additional supervision to increase performance over a variety of tasks. We use a convolutional neural network whose input is aerial imagery to predict the distribution over groups of clothing styles. As a side effect, the use of satellite imagery allows us to construct fine-grained, high-resolution maps of likely choices in attire at a worldwide scale.

Our work makes the following contributions:

- a new clothing dataset, XViewClothing, that contains over 8 million geotagged person images and co-located aerial images,
- a convolutional neural network that models the relationship between location (in the form of aerial imagery) and choices in human attire, the first of its kind to the authors' knowledge, and
- high spatial resolution maps showing the distribution of human attire choices for any geographic location in the world.

6.2 Related Work

6.2.1 Understanding Cities

Understanding the environment in which a person is present is critical towards learning about their culture and lifestyle. We can also observe objects, such as buildings, automobiles, and street signs may be useful cues for a variety of tasks. Recently there has been an increased interest in applying computer vision techniques to learn attributes for understanding urban areas. One of the most actively researched attributes in this domain is automatically assessing the safety of a region [2, 101, 106]. Other recent works have learned how certain aspects of cities, such as their architecture [26] and architectural evolution [80], can be used as cues for city recognition. Zhou *et al.* [156] train CNNs to estimate a set of city attributes that are used to distinguish cities. Porzi *et al.* [110] also train CNNs to perform safety ranking in an end-to-end manner from Google Street View images.

Most work on this task has focused on scene-level appearance attributes, while relatively little has explored object-level appearance attributes. Bessinger *et al.* [8] focused on learning the relationship between visual elements of a house and its value. Salesses *et al.* [120] and Quercia *et al.* [111] have successfully estimated attributes of wealth and beauty of a city based on images obtained from Google Street View and social media. Our work is similar in that we want to relate object-level appearance attributes of people to satellite imagery.

6.2.2 Cross-view Learning

The approach of learning a model using co-located aerial and ground-level imagery is referred to in the literature as *cross-view*. This work was proposed by Lin *et al.* [84] and later improved by Workman *et al.* [141] for the task of geolocalization where the goal is to return the geographic location of an image in the world. More recent works have used aerial imagery for the prediction of attributes such as home prices [6], object counts [15], sound [119], and time of day [153]. Most of these methods can learn a “soft-matching” between non-transient structures located in both images, such as buildings and road markers, which help to relate the different views. In our case, we cannot use soft-matching, as we need to learn higher level concepts that relate structure to scene and scene to human appearance. Some recent works have used satellite imagery as a predictor for human attributes, such as obesity [90], and our work is similar in that we use aerial imagery as a predictor for clothing styles.



Figure 6.2: Three kinds of clothing style imagery showing high variability under different constraints. The first image shows traditional fashion imagery used by stores to sell items of clothing which have both constrained lighting and pose. The second image shows stock imagery in which the environment in which the image is captured has semi-constrained pose and lighting, as well as a relatively clean background. The third image shows images captured “in the wild,” which have both unconstrained lighting and pose and may suffer from external factors, such as occlusion.

6.2.3 Computer Vision for Clothing Styles

Clothing understanding is a difficult problem due to the variability that can be found in images of people. Figure 6.2 highlights three different kinds of imagery with increasing complexity that are typically used: stock, staged, and in the wild. Learning the relationship between social groups and their choice of attire is an important step in performing large-scale analysis of the effect of clothing and location. Early computer vision research on understanding clothing involved semantic segmentation [146, 125, 136], and how to use clothing as a cue for social group analysis [100, 76, 25]. This problem has gained even more interest in recent years due to the effects of social media imagery, globalization, and learning trends in fashion. Problems in understanding clothing style choices include style classification [12, 68, 89], popularity analysis [68, 144], fashion image retrieval [145, 147], and clothing article recognition [116, 148, 85]. In the literature, this is often referred to as understanding *fashion*, however a more specific term for this research is understanding *clothing styles* since the term *fashion* is often associated with specific designers and brands.

One of the more interesting problems in understanding clothing styles is cross-domain clothing matching. The problem states that given an image of a particular item of clothing worn by someone in the street, can we match it to the a similar or exactly the same item in the shop (and vice-versa)? In this case, the domains we observe are multiple scenarios where similar apparel are worn. This problem is of interest to companies and advertisers who want to see the global outreach of their products, who is actually wearing them, and the demographics, such as age, location, etc. That said, the difficulty of this problem shares similar reasons as understanding faces. In the street domain, images are captured in highly

Table 6.1: Statistics for the XVC Great Britain subset.

	Confidence	Width	Height	Aspect Ratio
mean	0.970	152	200	0.774
std	0.037	96	113	0.246
min	0.850	40	15	0.500
25%	0.957	68	96	0.592
50%	0.987	130	185	0.710
75%	0.996	215	286	0.887
max	1.000	499	499	5.689

unconstrained settings and are subject to the same pose and lighting issues that affect face image understanding. Recent approaches to understanding cross-domain clothing style recognition include new ranking losses [54] and performing an exact match from street image to shop [43]. Our work falls under the task of cross-domain clothing matching. Our approach is the first of its kind to learn the relationship between aerial imagery and the distribution of clothing.

6.3 The XViewClothing Dataset

To support algorithm development, we curated a large-scale dataset of geotagged human and co-located aerial imagery called the XViewClothing (XVC) dataset. XVC is a set of 8 million images of people and their co-located aerial imagery. For each person patch, we download their associated aerial imagery at a particular zoom level from Microsoft Bing Maps. Note that since some photographs may have more than one person, each person in that photograph will share the same aerial image.

Our data source for collecting geotagged human imagery is the Yahoo Flickr Creative Commons 100M (YFCC100M), which contains roughly 100 million images in a variety of different settings. We downloaded only the images which contain geolocation metadata. Once downloaded, we used Faster-RCNN [115] trained on Pascal VOC 2012 [32] to extract bounding boxes and crops for every object classified as “person.” We used the default settings for Faster-RCNN by setting the minimum confidence threshold at 0.85 and the non-maximum suppression threshold at 0.30. For this work, we only operate on a subset of the XVC dataset for the country of Great Britain.

In Table 6.1 we provide some statistics about the human imagery on a subset of our dataset from Great Britain. The total number of person image collected is 1235446. We can see the average aspect ratio is approximately similar to that used in fashion image

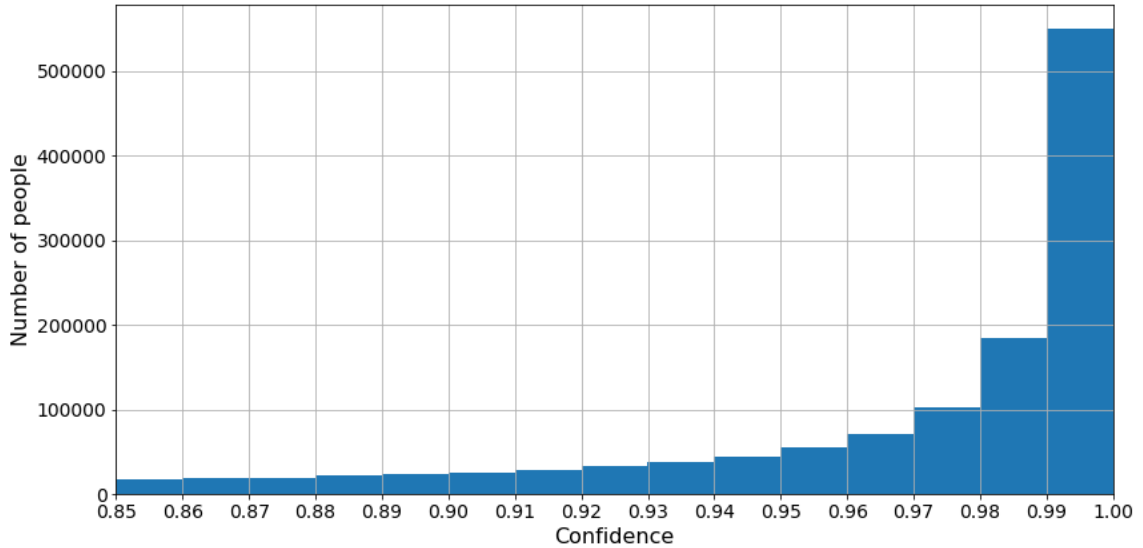


Figure 6.3: Histogram of confidences for patches extracted with Faster-RCNN. The x -axis is confidence and the y -axis is the number of patches.

photography, a ratio of 4:5 width to height. Figure 6.3 compares the prediction confidence for each patch versus the number of patches. We can see that more than half of the person patches in the Great Britain subset are predicted with a confidence greater than 0.98. The Great Britain subset represents approximately 15% of the entire dataset.

6.4 Approach

Our proposed approach, shown Figure 6.4, is a three step process of feature extraction, style clustering, and style prediction from aerial imagery. We begin by describing the process for acquiring features from ground-level human imagery. We use features from both clothing style and scene estimation networks to learn how to predict clothing style from satellite imagery.

For clothing style features, we train a neural network on the StreetStyle dataset [93]. Our network is a pre-trained Resnet-50 initialized using ImageNet weights. We remove the final classification layer and add 12 task-specific heads, each representing a classifier for a different style attribute including predictions of color, style, and accessories. Each classifier head is composed of three fully-connected layers of respective sizes [256, 256, and $n_classes$]. We train the clothing feature network end-to-end using stochastic gradient descent with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. Class weighting is applied on each task head of the network using the balanced approach proposed in [70]. After the network has been trained, we extract clothing style features for

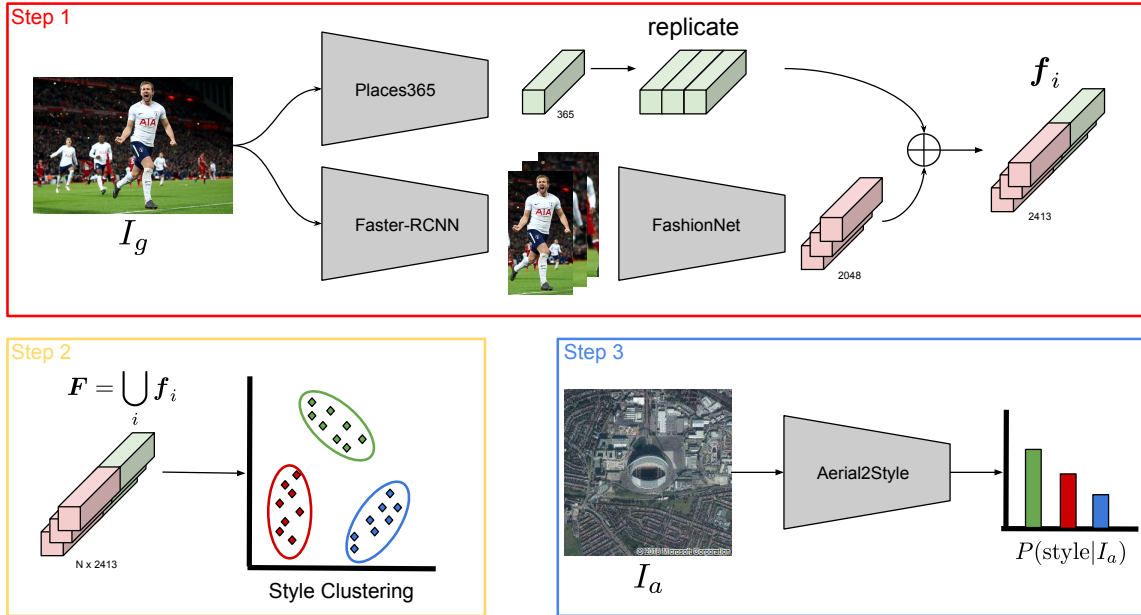


Figure 6.4: Our approach involves three steps. Given an input image, we first extract scene features from a scene classifier [155] and detect/crop all people found in the image using Faster-RCNN [115]. For each cropped person, we extract an associated clothing style feature and concatenate the scene feature to each clothing style feature. Next, we cluster the concatenated features and assign each person image to their centroid. Finally, we use a CNN that takes an aerial image as input and predicts a probability distribution over centroids.

each person in the XVC dataset from the shared final pooling layer with dimensionality 2048.

For scene features, we use a Resnet-50 initialized using weights trained on the Places365 database [155]. The extracted scene feature is from the final layer which is 365 dimensional. Since a person patch is cropped using the tightest possible bounding box, we cannot use the person patch as input to the scene network. A negative side effect of using the person patch as input is that the scene prediction for multiple individuals in a single image could be drastically different. Therefore, we use the entire image as input to the Places365 trained network. If an image contains more than one individual, the same scene feature is assigned to each individual in the photograph. Each sample, \mathbf{x}_i , in our dataset now has a corresponding clothing style-related features $c(\mathbf{x}_i)$, and scene feature, $s(\mathbf{x}_i)$.

We then concatenate the clothing style and scene features to form one holistic feature, f_i , that represents both scene and clothing. We apply k-means clustering on the features using cluster size $k = 200$, which was empirically selected as a compromise between intra- and inter-cluster variance. In our initial experiments, the clusters generated from human appearance features alone were insufficient for aerial image to appearance classification.

We incorporated ground-level scene information along with our human appearance features into the model allowed it to better reason about the objects in the aerial imagery and the appearance of people near them.

We then assign the predicted cluster center for each person as the label for each co-located aerial image. Finally, we train an Resnet-50, initialized using pretrained ImageNet weights, whose input is an aerial image and output is a probability distribution over each cluster center. Our network is optimized using the cross-entropy loss.

6.5 Evaluation

We create a training and testing set using an approximately 90/10 percent split of the dataset. A random split of the data might result in people from the same image being in both the training and test sets. We remedy this by constructing splits on the unique image identifier, so that all people in a particular image are either in the training or test sets. We also create a validation set using 10% of the training data to select model hyperparameters. During training we apply stratified sampling over style clusters to ensure the network learns an unbiased distribution from aerial imagery to clothing styles. We pre-process our inputs using data augmentation on the aerial imagery at training time by applying uniform random rotations between $[0, 360]$ degrees to make the network robust to changes in orientation. Each aerial image is then center cropped to 224×224 pixels, as resizing would affect the spatial resolution (zoom level). We train our networks on aerial imagery at zoom level 14.

Our model is optimized using the AMSGrad variant [114] of Adam with a learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We regularize the network using weight decay at a rate of 0.0001. We trained the model for 15 epochs and evaluate our approach both quantitatively and qualitatively.

6.5.1 Quantitative Metrics

We report top- k classification accuracy on our test set for various assignments of k as shown in Figure 6.5. The top- k accuracy is a particularly useful metric in our case when visually similar clusters may receive different class labels. We compare accuracy using our trained network to predict clothing styles for an aerial image against random selection. For all choices of k , we significantly outperform random chance. It quantifies that we have learned a useful feature representation for predicting clothing styles from aerial imagery.

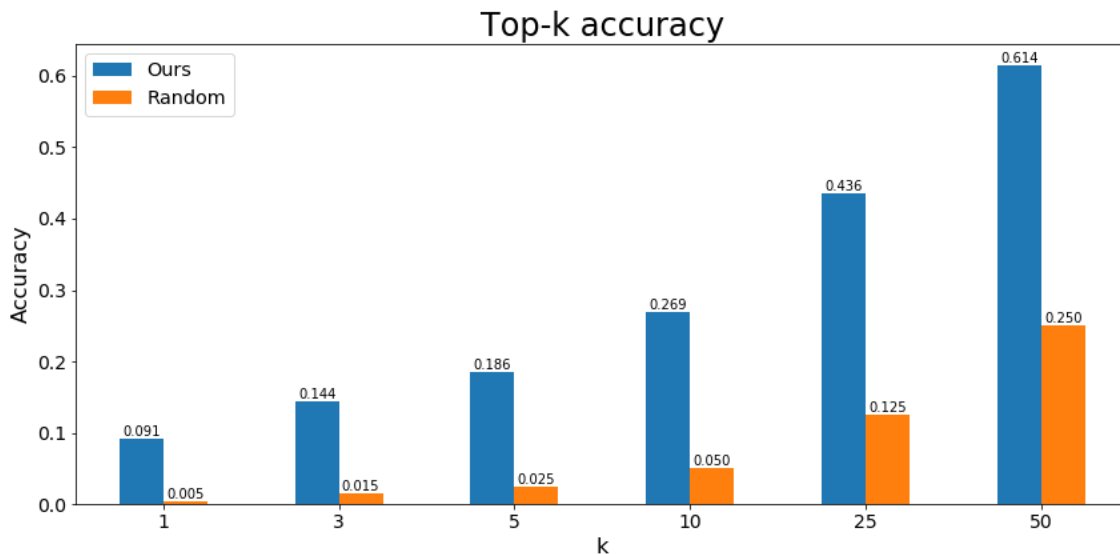


Figure 6.5: Top- k accuracy of style cluster prediction on the test set for different settings of $k \in \{1, 3, 5, 10, 25, \text{ and } 50\}$.

We observe that our model is able to predict the correct style cluster with 9.1% accuracy compared to random chance at 0.5%.

6.5.2 Style Predictions

In this section, we show our model’s ability to predict style clusters from aerial imagery and show our results in Figure 6.6. We use our network to classify an aerial image into a distribution over possible clothing styles. We select four diverse areas with varying structures from London to highlight differences in appearance environmentally in the aerial image and stylistic in person images. Each row shows an input aerial image from the test set, a sample of people from the ground truth class, and samples from the top-5 most likely predicted clusters. We show predictions from the top-5 clusters because some clusters may be visually similar to other clusters, however have different label assignments. The first two rows shows our network learns distinct region styles, such as stadiums and beaches, and more challenging urban areas in the last two rows.

6.5.3 Learning Style Trends

In this section, we demonstrate how our model can be used to show how clothing style choices change throughout the year. To enable this, we train our network with an additional embedding layer where each embedding corresponds to a month of the year. The embedding is concatenated to the output of the final pooling layer.

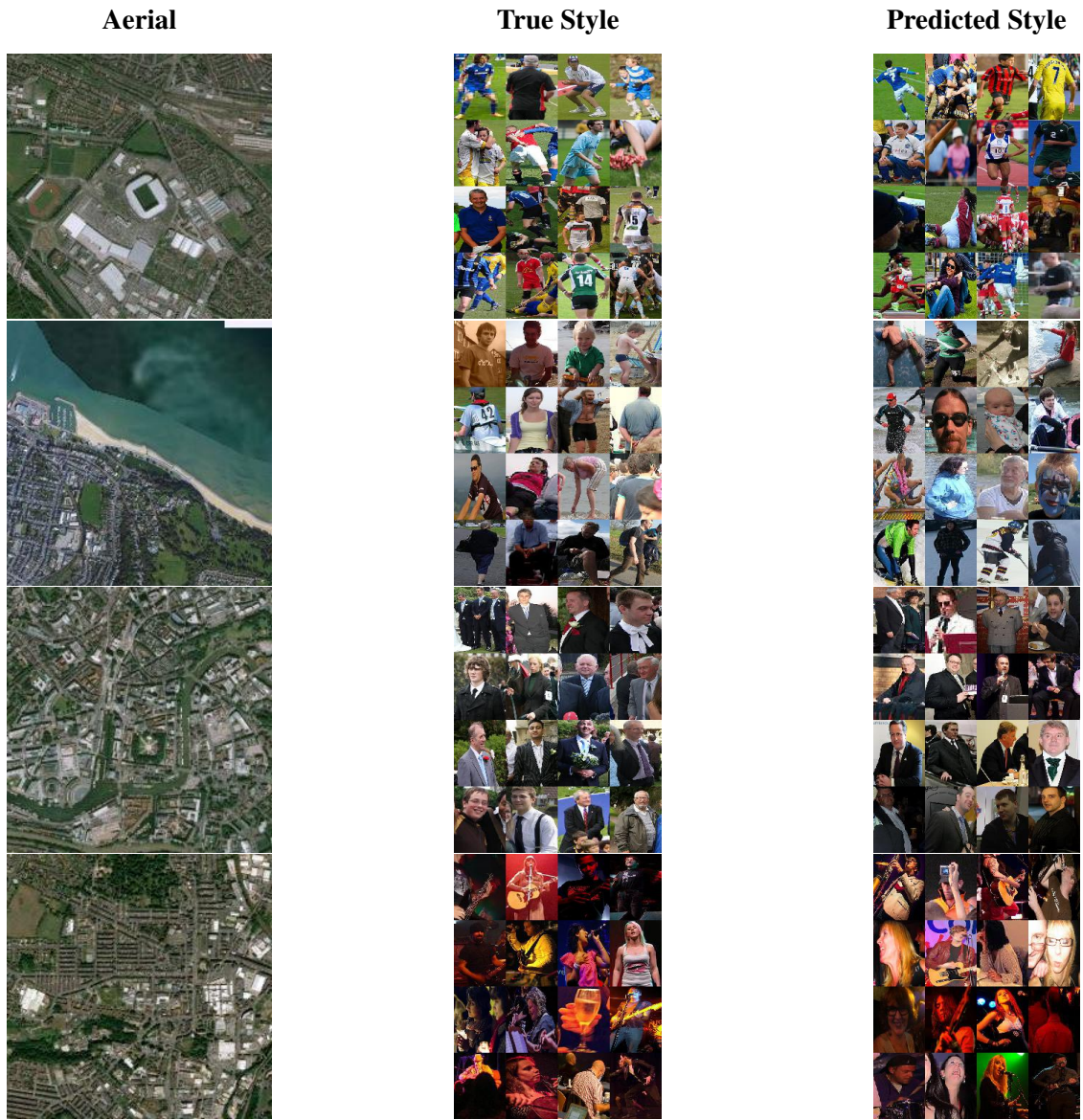


Figure 6.6: Given an aerial image from our test set (left), we show the true appearance (center) of people from the location and predicted appearance (right).

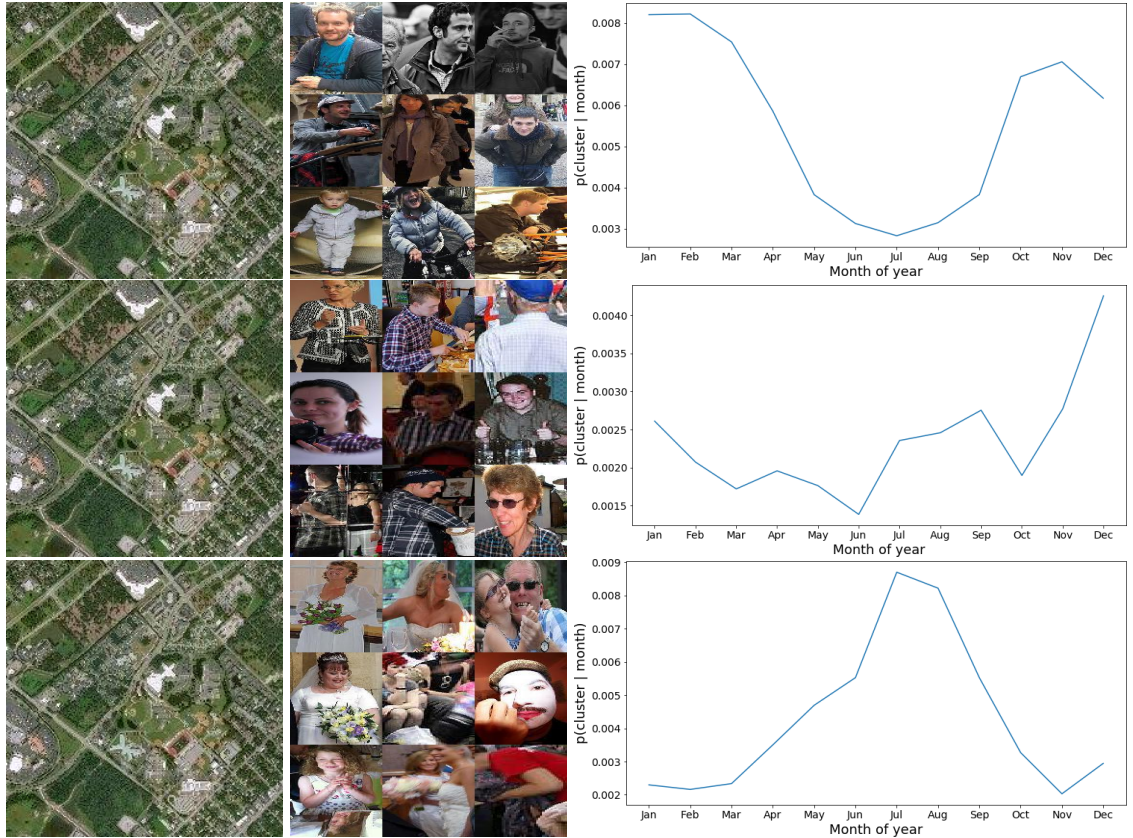


Figure 6.7: Appearance trends for each month of the year. Each row shows an aerial image over a geographic region (left), samples from a particular cluster (center), and the frequency in which those clusters appear conditioned on the month of the year (right).

Our results are shown in Figure 6.7. We select a region outside of the training set that has a diverse set of geographic and man-made structures. The first row shows how the likelihood of people wearing jackets in this region decreases during warmer months and increases during colder months. The second row shows how clothing patterns, specifically plaid, increase in appearance frequency around the fall and winter seasons. The third row shows the frequency of wedding dresses, indicating that most wedding events occur in summer months. In addition to learning clothing style trends, our model can be used to learn the frequency of special events, such as weddings. Our model can be conditioned in different ways that are most suitable to individual applications, such as analyzing the time of year a particular item, brand, or style of clothing is being worn in a geographic region.

6.5.4 High-Resolution Style Maps

The previous section, Section 6.5.2, showed results on aerial imagery from the test set and samples from the predicted cluster. It might be the case that some examples from our test

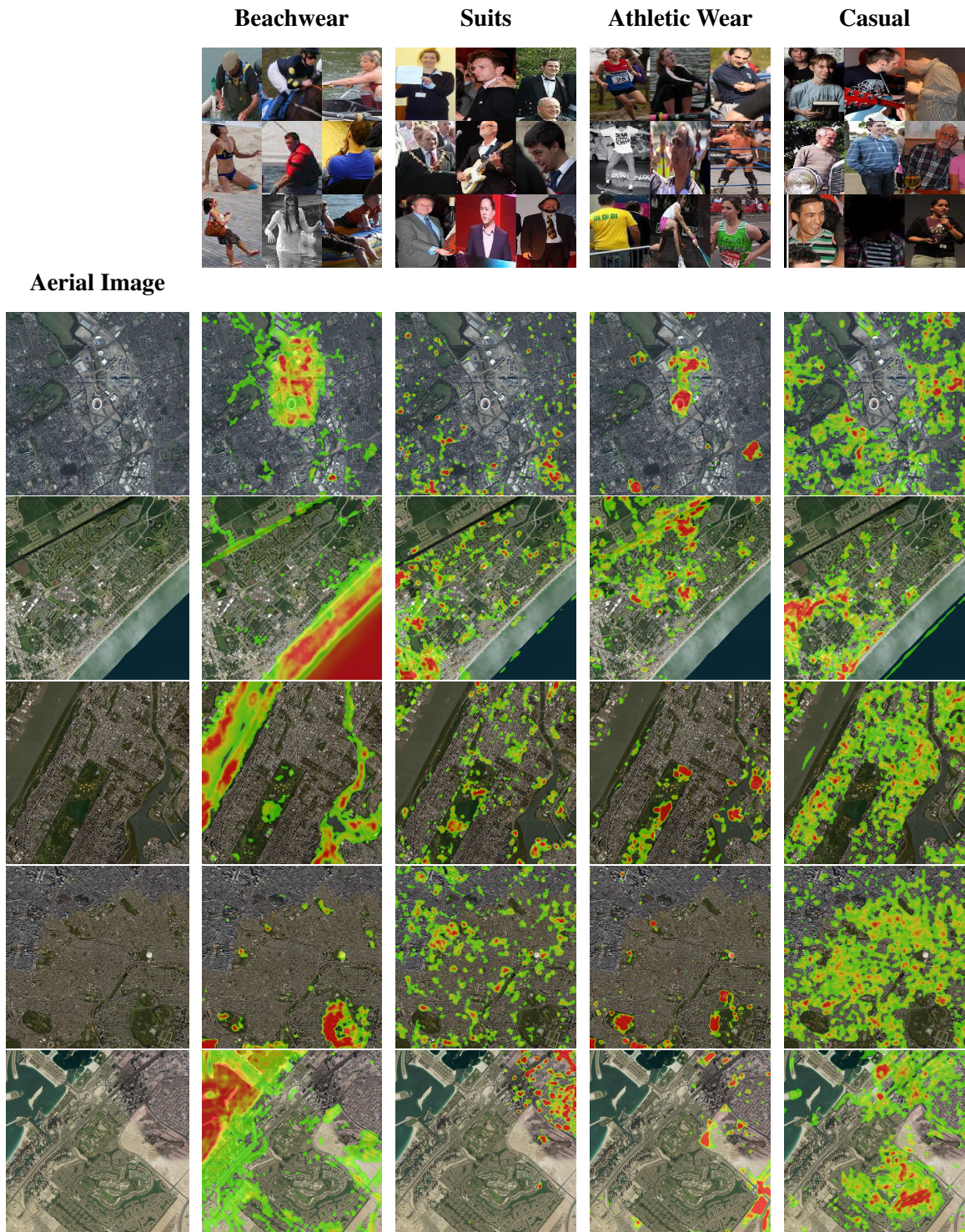


Figure 6.8: High-resolution maps of clothing style. Each row is a location corresponding to London, Myrtle Beach (USA), Central Park (USA), Tokyo, and Dubai respectively. The first column of each row is an aerial image captured over a large geographic region and the remaining columns show the probability for a particular style of clothing to appear in that location ranging from low (green) to high (red) likelihood.

set share geospatially overlapping regions. To address this concern, we look at how style can be predicted in regions that are completely outside of the training set of aerial imagery.

One useful feature of our approach is that it can be used to generate high-resolution maps of human appearance. We begin by observing clusters, c_i , predicted by the model and group them based on similar appearances into higher-order clusters, C' . We then download a 1024×1024 image for a particular region, shown in the first entry of the row. We select different kinds of environments, including a coastal city, an urban center, and a region with sparse man-made structures. We apply our network in a sliding window fashion over each regional image with a stride of 2 pixels. This results in 160000 predictions over the 200 clusters in our dataset. We select several groups from C' and color each map by the probability distribution for $P(c = c_i \in C' | I)$. The presence of color in each map shown in Figure 6.8 indicates a greater than random chance of a particular clothing style appearing in a given location. Green represents a small increase and red represents a large increase in the likelihood of a style cluster given a location. Note that objects in Figure 6.8 appear smaller than those in Figure 6.6 due to resizing of a larger spatial extent for visualization purposes.

We created high-resolution maps for five cities: London, Myrtle Beach (USA), New York City, Tokyo, and Dubai. These cities were selected for their dispersion around the world and environmental diversity. Our network was trained on images from Great Britain, so we can expect good performance for regions in London (first row). However, our network has never seen the other four cities, so maps of these regions are a test of our network’s ability to transfer knowledge of geographic location and structural elements to human appearance. We can see from these plots that the network has learned to relate geospatial structures to what people are likely to wear in that location. Beach regions are highlighted when people wearing beach attire and urban areas are highlighted when people are wearing business suits. People who wear athletic wear tend to be around park areas and stadiums. The “casual” clusters, however, exhibit an interesting behavior. If we focus on the “casual” predictions for New York City (third row, fifth column), we could imagine people dressed in this clothing style to appear almost anywhere. Our map of appearance likelihood reflects this assumption is true in nearly all areas except the Hudson river and Central Park.

These predictions could be the result of a dataset bias, such as a photographer standing in a park capturing only marathon runners, or a cultural trend, such as Americans wearing shorts when traveling in Europe. If we trained the model on a different location, for example New York City (NYC), our model would capture trends in NYC and the maps we generate would all be dependent on NYC styles. Tourists may also have an effect on the style of a city if they are heavily populated or capture many more photos of themselves

dressed in a different way than the general population. Unfortunately, using our dataset, we are unable to tell if someone is a tourist or not to determine if the styles we have learned are styles of residents or an amalgam of tourists and residents. Additionally, if our clothing style clusters were more specific, such as all people wearing a distinctive fashion brand design or logo, it might provide a stronger discernment between dataset bias and cultural trend.

6.6 Conclusions

We presented a model that is given an aerial image and predicts a probability distribution over a set of clothing styles. We quantitatively and qualitatively demonstrated our model learns features useful for accurately predicting a distribution over clothing styles. We also showed how our model can be used to construct high-resolution maps using aerial imagery and showed that our model picks up on cues from natural and man-made structures found in aerial imagery to assist in learning about how people dress in different settings. These maps also demonstrated the model’s ability to transfer knowledge it learned about structures from one geospatial region to another, suggesting its use for understanding the distribution of personal style choices at a global scale.

A limitation in our approach is that it relies on a fixed feature representation of the input. That is, the features that were used to generate our style clusters were learned from pre-trained networks. In future work, we plan to address this using an end-to-end learned representation that is able to better represent the relationship between geographic location and clothing style choices. Another avenue for future work involves providing additional context to the network using co-located street-view imagery. The context provided by street-view imagery could help our model know if a person standing outside a building in a city were in front of a restaurant or a store. We demonstrated our approach on over one million images from Great Britain and plan to learn a new model with full global coverage for all eight million images in the XVC dataset.

Chapter 7

Conclusion

This thesis presented a computational, data-driven approach to geo-ethnography. Our generative and discriminative models were designed in such a way that they enable high-resolution predictions for facial and clothing appearance, despite having sparsely distributed ground-level imagery. These maps provide a visual presentation of our models and show how our models can be practically applied in real-world applications and scenarios.

In Chapter 3 we introduced a large-scale geotagged dataset of human faces, WGT, with age and gender estimates. This is the largest publicly available dataset of geotagged facial images. We compared WGT with existing geotagged facial image datasets, highlighting number of samples, demographic labels, and additional metadata. We presented a way to visualize our dataset using a bottom-up approach using conditional averages based on age, gender, and country. We incorporated the conditional averages into an interactive, web-based mapping tool which allowed us to view worldwide, smooth appearance trends at multiple spatial scales.

In Chapter 4 we developed baseline generative and discriminative models of our dataset. We evaluated a linear regression model and a random forest model whose inputs are the age, gender, pose, and geographic location and outputs are PCA components. The predicted PCA components are then projected onto a basis formed on a held out set of faces to conditionally generate faces. While this method can synthesize face-like images, its disadvantages include less sample diversity in the generated samples and a very large model size caused by the random forest.

In Chapter 5, we introduced a factored, latent variable generative model that used age, gender, pose, and geographic location to conditionally generate faces. It allowed us to significantly improve upon sample generation introduced in Chapter 4 with increased sample diversity and higher fidelity. We demonstrated this with both qualitative and quantitative

results. Despite the noisy nature of our dataset, our model can learn the complex relationship between geographic location and various facial attributes. The model requires two orders of magnitude less storage space than our model in Chapter 4, can operate on GPUs, and allows us to quickly generate realistic, diverse sets of faces.

Finally, in Chapter 6 we introduced a discriminative model to learn the relationship between satellite images, location, and time to clothing choices. We quantitatively and qualitatively demonstrated that our model learns features that are useful for accurately predicting a distribution of clothing styles. We showed how the model can be used to learn how fashion trends appear and disappear over months of the year. We also showed how our model can be used to construct high spatial resolution maps using aerial imagery and showed that our model picks up on cues from natural and man-made structures found in aerial imagery to assist in learning about how people dress in different settings. These maps also demonstrated the model’s ability to transfer knowledge it learned about structures from one geospatial region to another, suggesting its use for understanding the distribution of personal style choices at a global scale.

There are several avenues of work for future studies on this topic. In particular, two important issues that should be considered when working with human image datasets are human-introduced biases and privacy. Most face recognition datasets are gender and ethnically biased due to using a single data source or selection bias. This problem can be somewhat mitigated using methods such as stratified sampling and is actively being addressed in some recent work [13] which proposed a new, balanced dataset. Our generative model of faces could be used to address location distribution biases by synthesizing new samples from sparsely populated areas, however it is not a panacea for lack of real data. Additionally, one could address tourist bias using a form of anomaly detection. Identifying and counteracting data biases is essential as machine learning algorithms are increasingly used in important decision making. Another important issue is the implication on privacy. Our generative model introduced in Chapter 5 could be used by someone to generate samples from a geographic region and use them as a visual guide to blend in or disguise themselves with people native to the region. Our work developed in this thesis serves as a test-bed for future research into high-spatial-resolution, location-dependent human appearance.

Bibliography

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, 2016. 8
- [2] Sean M Arietta, Alexei Efros, Ravi Ramamoorthi, Maneesh Agrawala, et al. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2014. 55
- [3] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017. 12, 41
- [4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017. 12, 41
- [5] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 1997. 2
- [6] Archith J Bency, Swati Rallapalli, Raghu K Ganti, Mudhakar Srivatsa, and BS Manjunath. Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. *arXiv preprint arXiv:1610.04805*, 2016. 55
- [7] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 12, 41
- [8] Zachary Bessinger and Nathan Jacobs. Quantifying curb appeal. In *International Conference on Image Processing*, 2016. 55
- [9] Zachary Bessinger and Nathan Jacobs. A generative model of worldwide facial appearance. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 5

- [10] Zachary Bessinger, Chris Stauffer, and Nathan Jacobs. Who goes there?: approaches to mapping facial appearance diversity. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016. vii, viii, 5, 39, 42, 43, 48, 49, 51
- [11] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 1992. 7
- [12] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. Apparel classification with style. In *Asian Conference on Computer Vision*, 2012. 56
- [13] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 2018. 68
- [14] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016. 12, 28
- [15] Nathan Jacobs Connor Greenwell, Scott Workman. What goes where: Predicting object distributions from above. In *IEEE Geoscience and Remote Sensing Society*, 2018. 55
- [16] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), 2001. 40
- [17] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3), 1995. 7
- [18] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world’s photos. In *International World Wide Web Conference*, 2009. 13
- [19] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1), 2018. 10

- [20] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2005. HOG. 7
- [21] Antitza Dantcheva, Petros Elia, and Arun Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3), 2016. 39
- [22] John G Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision research*, 20(10), 1980. 1
- [23] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 2008. 17
- [24] Margo DeMello. *Faces around the world: a cultural encyclopedia of the human face*. ABC-CLIO, 2012. 38
- [25] Abhinav Dhall, Jyoti Joshi, Karan Sikka, Roland Goecke, and Nicu Sebe. The more the merrier: Analysing the affect of a group of people in images. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, 2015. 56
- [26] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 31(4), 2012. 55
- [27] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 12
- [28] Hao Dong, Paarth Neekhara, Chao Wu, and Yike Guo. Unsupervised image-to-image translation with generative adversarial networks. *arXiv preprint arXiv:1701.02676*, 2017. 28
- [29] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv:1602.02644*, 2016. 12
- [30] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 12
- [31] Eran Eidinger, Roe Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12), 2014. 40

- [32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>. 57
- [33] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, 2015. 45
- [34] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 27
- [35] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. R-CNNs for pose estimation and action detection. *arXiv preprint arXiv:1406.5212*, 2014. 27
- [36] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 10
- [37] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 10, 12, 28, 41
- [38] Daniel B Graham and Nigel M Allinson. Characterising virtual eigensignatures for general purpose face recognition. In *Face Recognition*. Springer, 1998. UMIST. 2
- [39] Connor Greenwell, Scott Spurlock, Richard Souvenir, and Nathan Jacobs. GeoFace-Explorer: Exploring the Geo-Dependence of Facial Attributes. In *ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD)*, 2014. 2, 13, 42
- [40] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5), 2010. 46
- [41] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, 2017. 12, 41
- [42] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016. 2

- [43] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *IEEE International Conference on Computer Vision*, 2015. 57
- [44] Karen Tranberg Hansen. The world in dress: Anthropological perspectives on clothing, fashion, and culture. *Annu. Rev. Anthropol.*, 33, 2004. 38
- [45] Tal Hassner, Shai Harel, Eran Paz, and Roei Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 27
- [46] William A Haviland, Harald EL Prins, Dana Walrath, and Bunny McBride. *Anthropology: The human challenge*. Cengage Learning, 2013. 39
- [47] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 13
- [48] 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. 9
- [49] Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems*, 2000. 34
- [50] Satoshi Hosoi, Erina Takikawa, and Masato Kawade. Ethnicity estimation with facial images. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004. 27, 40
- [51] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 9
- [52] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 9
- [53] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008. LFW. 2

- [54] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *IEEE International Conference on Computer Vision*, 2015. 57
- [55] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017. 41
- [56] Mohammad T Islam, Connor Greenwell, Richard Souvenir, and Nathan Jacobs. Large-Scale Geo-Facial Image Analysis. *EURASIP Journal on Image and Video Processing (JIVP)*, 2015(1), 2015. 2, 13, 17, 42
- [57] Mohammad T Islam, Scott Workman, and Nathan Jacobs. Face2gps: Estimating geographic location from facial features. In *International Conference on Image Processing*, 2015. 2, 14, 42
- [58] Mohammad T. Islam, Scott Workman, Hui Wu, Richard Souvenir, and Nathan Jacobs. Exploring the Geo-Dependence of Human Face Appearance. In *IEEE Winter Conference on Applications of Computer Vision*, 2014. 13, 18
- [59] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 12, 46
- [60] Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kyla Miskell, Bobby H. Braswell, Andrew D. Richardson, and Robert Pless. The Global Network of Outdoor Webcams: Properties and Applications. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2009. 13
- [61] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 8
- [62] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *arXiv preprint arXiv:1603.08155*, 2016. 12
- [63] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 12

- [64] Vahdat Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 18, 19, 27, 44
- [65] Ira Kemelmacher-Shlizerman and Steven M Seitz. Collection flow. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. vi, 20, 22, 32
- [66] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 43
- [67] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M Seitz. Illumination-aware age progression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 41
- [68] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Hipster wars: Discovering elements of fashion styles. In *European Conference on Computer Vision*, 2014. 56
- [69] Taeksoo Kim, Moon-su Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 12
- [70] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2), 2001. 58
- [71] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 48
- [72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. vi, 7, 9
- [73] Kuba Krys, C-Melanie Vauclair, Colin A Capaldi, Vivian Miu-Chi Lun, Michael Harris Bond, Alejandra Domínguez-Espinosa, Claudio Torres, Ottmar V Lipp, L Sam S Manickam, Cai Xing, et al. Be careful where you smile: culture shapes judgments of intelligence and honesty of smiling individuals. *Journal of nonverbal behavior*, 40(2), 2016. 38
- [74] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. The gan landscape: Losses, architectures, regularization, and normalization. *arXiv preprint arXiv:1807.04720*, 2018. 10

- [75] Hanock Kwak and Byoung-Tak Zhang. Ways of conditioning generative adversarial networks. *arXiv preprint arXiv:1611.01455*, 2016. 41
- [76] Iljung S Kwak, Ana Cristina Murillo, Peter N Belhumeur, David J Kriegman, and Serge J Belongie. From bikers to surfers: Visual recognition of urban tribes. In *British Machine Vision Conference*, 2013. 52, 56
- [77] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 28
- [78] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998. vi, 7, 11
- [79] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 12
- [80] Stefan Lee, Nicolas Maisonneuve, David J. Crandall, Alexei A. Efros, and Josef Sivic. Linking past to present: Discovering style in two centuries of architecture. In *IEEE International Conference on Computational Photography, ICCP*, 2015. 55
- [81] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *CVPR Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, 2015. 19, 28, 40, 44
- [82] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866), 2008. 39
- [83] Mu Li, Wangmeng Zuo, and David Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016. 28, 41
- [84] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 13, 55

- [85] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 56
- [86] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015. vii, 27, 43
- [87] William Lotter, Gabriel Kreiman, and David Cox. Unsupervised learning of visual structure using predictive generative networks. *arXiv preprint arXiv:1511.06380*, 2015. 12
- [88] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004. SIFT. 7
- [89] Yihui Ma, Jia Jia, Suping Zhou, Jingtian Fu, Yejun Liu, and Zijian Tong. Learning to appreciate the aesthetic effects of clothing. In *AAAI Conference on Artificial Intelligence*, 2017. 56
- [90] Adyasha Maharana and Elaine Okanyene Nsoesie. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Network Open*, 1(4), 2018. 55
- [91] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 38
- [92] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, 2014. 18, 27
- [93] Kevin Matzen, Kavita Bala, and Noah Snavely. Streetstyle: Exploring world-wide clothing styles from millions of photos. *arXiv preprint arXiv:1706.01869*, 2017. 58
- [94] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, 2018. 12
- [95] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, 2017. 12
- [96] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016. 12

- [97] Daniel Miller, Evan Brossard, Steven M Seitz, and Ira Kemelmacher-Shlizerman. MegaFace: A million faces for recognition at scale. *arXiv preprint arXiv:1505.02108*, 2015. 2, 13, 17, 18
- [98] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 11, 41
- [99] Aleksandra Mizielińska and Daniel Mizieliński. *Maps*. Big Picture Press, 1 edition, 10 2013. vi, 3
- [100] Ana C Murillo, Iljung S Kwak, Lubomir Bourdev, David Kriegman, and Serge Belongie. Urban tribes: Analyzing group photos from a social perspective. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012. 56
- [101] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo. Streetscore—predicting the perceived safety of one million streetscapes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014. 55
- [102] Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, 2017. 12
- [103] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016. 11
- [104] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 11, 41
- [105] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), 2001. 7
- [106] Vicente Ordonez and Tamara L Berg. Learning high-level judgments of urban perception. In *European Conference on Computer Vision*, 2014. 55
- [107] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 19, 27, 28, 40
- [108] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 8

- [109] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. *arXiv preprint arXiv:1604.07379*, 2016. 12
- [110] Lorenzo Porzi, Samuel Rota Bulò, Bruno Lepri, and Elisa Ricci. Predicting and understanding urban perception with convolutional neural networks. In *ACM Conference on Multimedia*, 2015. 55
- [111] Daniele Quercia, Neil Keith O’Hare, and Henriette Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2014. 55
- [112] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 12, 28, 41
- [113] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 27
- [114] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Machine Learning*, 2018. 60
- [115] 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. ix, 57, 59
- [116] Rasmus Rothe, Marko Ristin, Matthias Dantone, and Luc Van Gool. Discriminative learning of apparel features. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, 2015. 56
- [117] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088), 1986. 8
- [118] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3), 2015. 2, 8, 9

- [119] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. A multimodal approach to mapping soundscapes. In *IEEE Geoscience and Remote Sensing Society*, 2018. 55
- [120] Philip Salesses, Katja Schechtner, and César A Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7), 2013. 55
- [121] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016. 41, 50
- [122] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision*, 1994. Olvetti faces. 2
- [123] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 28, 40
- [124] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 42
- [125] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. A high performance crf model for clothes parsing. In *Asian Conference on Computer Vision*, 2014. 56
- [126] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 53
- [127] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 9
- [128] Nisha Srinivas, Harleen Atwal, Derek C Rose, Gayathri Mahalingam, Karl Ricanek, and David S Bolme. Age, gender, and fine-grained ethnicity prediction using convolutional neural networks for the east asian face dataset. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017. 40

- [129] 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. 9
- [130] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 28
- [131] Yaniv Taigman, Ming Yang, Marc’ Aurelio Ranzato, and Lars Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 27, 40
- [132] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. 17, 43
- [133] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 11, 41
- [134] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1991. 2, 7, 40
- [135] Carles Ventura, David Masip, and Agata Lapedriza. Interpreting cnn models for apparent personality trait regression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 27
- [136] Sirion Vittayakorn, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Runway to realway: Visual analysis of fashion. In *IEEE Winter Conference on Applications of Computer Vision*, 2015. 56
- [137] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *IEEE International Conference on Computer Vision*, 2017. 13
- [138] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 13, 42, 53
- [139] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, and Fei-Yue Wang. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4), 2017. 10

- [140] Tom White. Sampling generative networks: Notes on a few effective techniques. *CoRR*, abs/1609.04468, 2016. 12
- [141] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, 2015. 13, 55
- [142] Ling Xie and Shawn Newsam. Im2map: deriving maps from georeferenced community contributed photo collections. In *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media*, 2011. 13
- [143] Xuehan Xiong and Fernando Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 18
- [144] Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. Chic or social: Visual popularity analysis in online fashion networks. In *ACM Conference on Multimedia*, 2014. 56
- [145] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *IEEE International Conference on Computer Vision*, 2013. 56
- [146] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2, 56
- [147] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Retrieving similar styles to parse clothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5), 2015. 56
- [148] Kota Yamaguchi, Takayuki Okatani, Kyoko Sudo, Kazuhiko Murasaki, and Yuki-nobu Taniguchi. Mix and match: joint model for clothing and attribute recognition. In *British Machine Vision Conference*, 2015. 56
- [149] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, 2016. 41
- [150] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. *IEEE International Conference on Computer Vision*, 2017. 41

- [151] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, 2010. 13
- [152] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014. 9
- [153] Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. Learning geo-temporal image features. In *British Machine Vision Conference*, 2018. 55
- [154] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 11, 41, 42
- [155] 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. ix, 59
- [156] Bolei Zhou, Liu Liu, Aude Oliva, and Antonio Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *European Conference on Computer Vision*, 2014. 55
- [157] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016. 8
- [158] Jun-Yan Zhu, Yong Jae Lee, and Alexei A Efros. Averageexplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics*, 33(4), 2014. 20
- [159] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 27
- [160] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 9

Vita

Zachary Bessinger

Education

2014–Present	Ph.D., Computer Science	University of Kentucky <i>Lexington, KY</i>
2012–2013	M.S., Computer Science	Western Kentucky University <i>Bowling Green, KY</i>
2007–2011	B.S., Computer Science	Western Kentucky University <i>Bowling Green, KY</i>

Appointments

Data Scientist Intern Oct. 2017 – Present	Zillow Group <i>Seattle, WA</i>
Teaching/Research Assistant Jun. 2014 – Present	University of Kentucky <i>Lexington, KY</i>
Software Developer Jul. 2013 – Jun. 2014	Publishers Printing Co. <i>Shepherdsville, KY</i>
Teaching/Research Assistant Jun. 2011 – May 2013	Western Kentucky University <i>Bowling Green, KY</i>

Publications

Refereed Conference Publications

- [1] Zachary Bessinger and Nathan Jacobs. A generative model of worldwide facial appearance. In *IEEE Winter Conference on Applications of Computer Vision*, 2019.
- [2] 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.
- [3] Zachary Bessinger, Chris Stauffer, and Nathan Jacobs. Who goes there? approaches to mapping facial appearance diversity. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016.
- [4] Zachary Bessinger and Nathan Jacobs. Quantifying curb appeal. In *International Conference on Image Processing*, 2016.
- [5] Radu Mihail, Scott Workman, Zachary Bessinger, and Nathan Jacobs. Sky segmentation in the wild: An empirical study. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.
- [6] Zachary Bessinger, Guangming Xing, and Qi Li. Localization of drosophila embryos using connected components in scale space. In *International Conference on Image Processing*, 2012.
- [7] Qi Li and Zachary Bessinger. Learning scale ranges for the extraction of regions of interest. In *International Conference on Image Processing*, 2012.

Honors and Awards

- 2014–2016 – University of KY Teaching Assistantship
- 2017 – CVPR Student Volunteer
- 2013 – Ogden College Outstanding Graduate Student Award in Computer Science
- 2013 – Ogden College Outstanding Graduate Research Student Award (honorable mention)

Service

- 2017 – Reviewer for ACM Transactions on Multimedia Computing Communications and Applications (TOMM)