A Generative Model of Worldwide Facial Appearance

Zachary Bessinger Nathan Jacobs
Department of Computer Science
University of Kentucky
{zach, jacobs}@cs.uky.edu

Abstract

Human appearance depends on many proximate factors, including age, gender, ethnicity, and personal style choices. In this work, we model the relationship between human appearance and geographic location, which can impact these factors in complex ways. We propose GPS2Face*, a dual-component generative network architecture that enables flexible facial generation with fine-grained control of latent factors. We use facial landmarks as a guide to synthesize likely faces for locations around the world. We train our model on a large-scale dataset of geotagged faces and evaluate our proposed model, both qualitatively and quantitatively, against previous work.*

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 [7], the frequency of various facial expressions [25], and types of clothing [14].

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) [31], 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 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.

Figure 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.

Our approach significantly improves upon previous works that attempt to model the relationship between geographic location and facial appearance. Disciplines including anthropology [15] and evolutionary biology [28] 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 [6]. 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 [4], 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 1.

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

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 [8, 27] but these characteristics are often used implicitly to recognize individuals [34, 37, 40]. 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 [16, 39], 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 [43] and models that explicitly represented face pose [5]. More recent work has built upon the Generative Adversarial Network (GAN) framework proposed by Goodfellow *et al*. [10]. 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 [35], has been applied to a wide variety of image synthesis tasks, including facial synthesis. Recently, many approaches [2, 13, 3] 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 [1, 36] 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 [32], 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 [26, 33] 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 [23, 47] or transient attributes, such as facial hair and hair color [29, 45]. 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 [17, 46].

Similar to our model is the recent work by Tran *et al*. [42]. 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 [47], however we use LeakyReLU instead of ReLU activations in all networks, replace the transpose convolution layers with *PixelShuffle* [38] 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*. [18] 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 [11] rather than other common facial tasks such as recognition. Islam *et al*. [19] used image features extracted from a pre-trained CNN to predict in which of 50 cities a face image was captured. Wang *et al*. [44] 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*. [4] 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.

Figure 2: Samples from the WGT dataset [4] used in our work. Unlike face datasets that have been previously used to train generative models, such as CelebA [30], our dataset has not been manually filtered and contains a wide variety of image qualities.

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 2 shows samples from the dataset. A diagram of our network architecture is shown in Figure 3.

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 [4] for all experiments. Unlike other recently created large-scale face datasets, such as CelebA [30] and MegaFace [22], 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 [41], however it only includes faces from images that are geotagged. In total, it

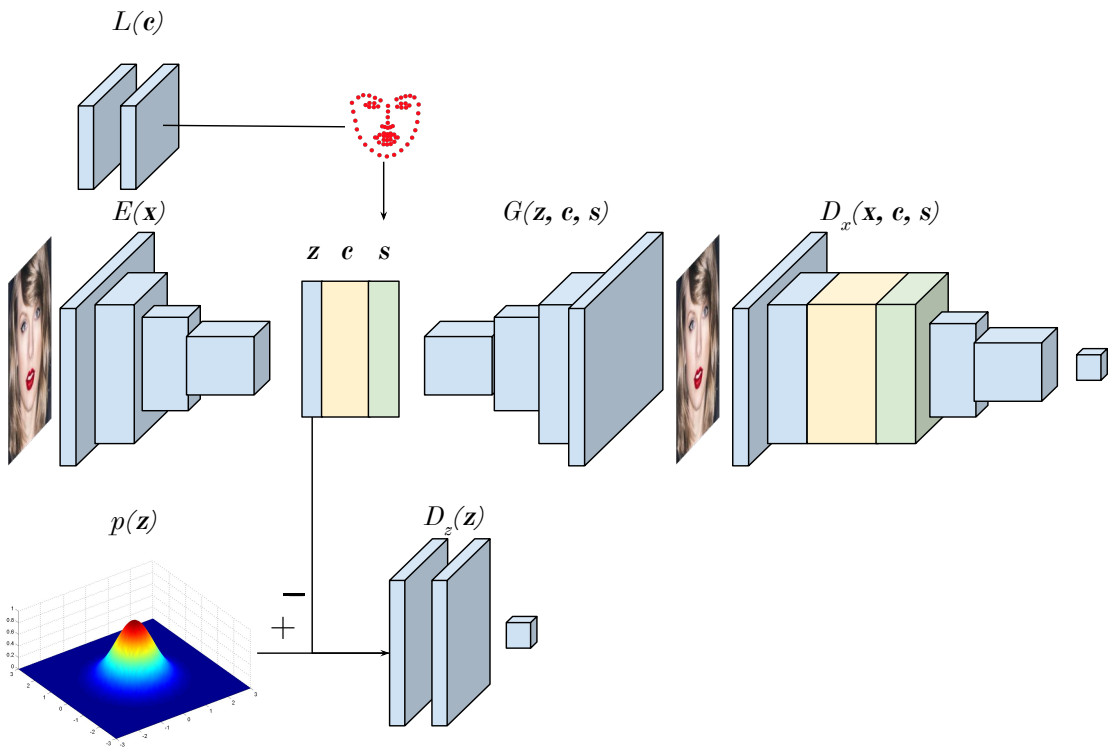


Figure 3: Our proposed model, *GPS2Face*, has two components: a landmark prediction network, L, and an appearance generation network supported by the other sub-networks. Landmarks are used to guide synthesis and improve the quality of generated faces since identity is not used as a regularizer. L uses 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.

contains 2.1 million geotagged face images, along with automatically estimated facial landmark locations [21] and age/gender [27]. We augment this dataset by estimating the pitch, yaw, and roll of each face using the provided landmarks and the perspective-n-point algorithm.

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 [9]:

$$\mathcal{L}_{\text{huber}}(x, y) = \frac{1}{n}\sum_i z_i$$
$$z_i = \begin{cases} 0.5(x_i - y_i)^2, & \text{if } |x_i - y_i| < 1 \\ |x_i - y_i| - 0.5, & \text{otherwise} \end{cases}, \quad (1)$$

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.

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 subnetworks: an encoder, a decoder, and discriminators for

both images and latent space. The encoder, E, takes as input a face patch, $\boldsymbol{x}$, 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 $\boldsymbol{x}$ to represent an image, $\boldsymbol{y}$ 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 $G(z)$:

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

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:

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

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:

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

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 [20] 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:

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

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 [12] format.

4. Evaluation

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

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, from each country to form our final training set. We trained *GPS2Face* using Adam [24] 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.

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 4, we manipulate faces from the testing set in a variety of ways. Figure 4a shows a montage of example images, organized by pitch (y-axis) and yaw (x-axis). Figure 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 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

(a) (b) (c) (d)

Figure 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).

(a) Input images (b) Ours (encoded) (c) Ours (identity) (d) Bessinger *et al*. [4]

Figure 5: Qualitative comparison of random samples from our method and a previous method from Bessinger *et al*. [4]. 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 [4].

shapes and lighting 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.

4.3. Qualitative Comparison with Previous Work

We qualitatively compare the results of *GPS2Face* against the previous work of Bessinger *et al*. [4]. 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 we provide a qualitative comparison of this method versus ours using faces and attributes from the held-out set. Figure 5a shows real images that will be encoded and whose latent factors are used to condition each model. Figure 5b shows faces generated with our method after encoding input images to the latent space, then reconstructing using the conditional encoded sample. Figure 5c shows faces generated with our method using conditioned samples from the prior. Figure 5d shows reconstructions from the predicted PCA coefficients. The reconstructions in Figure 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 4.4.

In Figure 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 each montage. We also fix age and gender to be a 25 year old female. We then vary facial pose pitch $\pm$ 30° and yaw $\pm$ 20°, 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.

Table 1: Quantitative evaluation of our proposed method.

	Inception Score	**PSNR**	**SSIM**
Bessinger *et al*. [4]	1.475 ± 0.004	13.068	0.339
Ours (encoded)	$\mathbf{1.7370 \pm 0.007}$	**19.131**	**0.513**
Ours (identity)	1.609 ± 0.004	–	–
Real data	3.483 ± 0.015	–	–

The most representative faces are those with neutral pose (in the center of each montage).

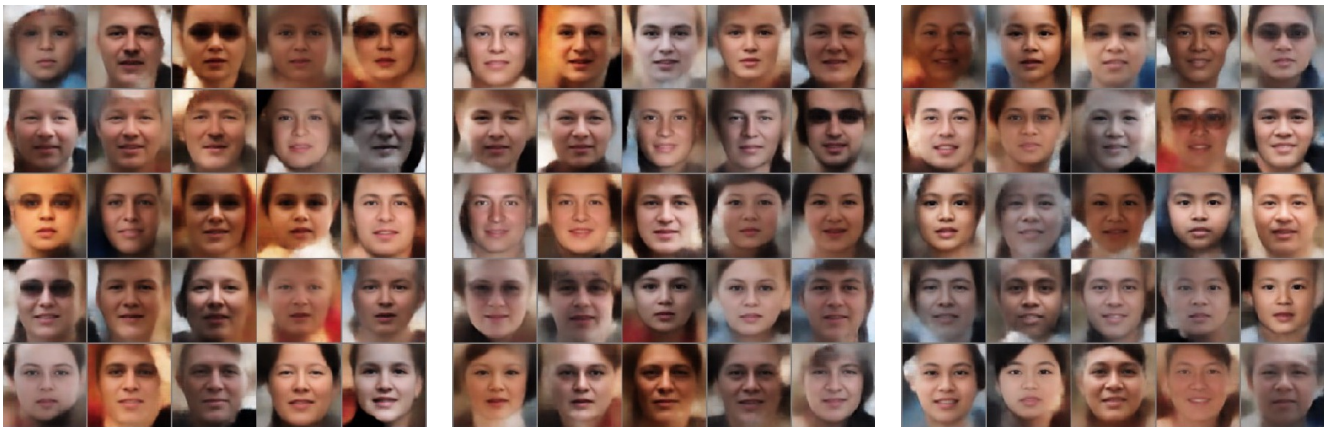
Figure 7 shows montages of synthesized faces from various locations around the world. For each montage, we draw 25 values of z from the prior. We select a configuration of latent factors where age and gender are randomized, and pose is frontal. In total, we show 25 faces in each country for three different countries.

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*. [36], 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 changes makes the task no longer an image-to-image translation, we do not

(a) Japan (b) United Kingdom (c) DR Congo

Figure 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° yaw and $\pm$ 30° pitch. We then sample locations from the countries shown in the captions above.



(a) United Kingdom (b) Russia (c) Philippines

Figure 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.

compute PSNR and SSIM on identity-modified images.

Our results are shown in Table 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 [4], 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. 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.

Acknowledgment

We gratefully acknowledge the financial support of NSF CAREER grant IIS-1553116 and computing resources provided by the Univ. of Kentucky Center for Computational Sciences, including a Power8 system donated by IBM.

References

[1] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[3] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[4] Z. Bessinger, C. Stauffer, and N. Jacobs. Who goes there?: approaches to mapping facial appearance diversity. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), 2001.

[6] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3), 2016.

[7] M. DeMello. *Faces around the world: a cultural encyclopedia of the human face*. ABC-CLIO, 2012.

[8] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12), 2014.

[9] R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, 2015.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[11] C. Greenwell, S. Spurlock, R. Souvenir, and N. Jacobs. GeoFaceExplorer: Exploring the Geo-Dependence of Facial Attributes. In *ACM SIGSPATAL International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD)*, 2014.

[12] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5), 2010.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, 2017.

[14] K. T. Hansen. The world in dress: Anthropological perspectives on clothing, fashion, and culture. *Annu. Rev. Anthropol.*, 33, 2004.

[15] W. A. Haviland, H. E. Prins, D. Walrath, and B. McBride. *Anthropology: The human challenge*. Cengage Learning, 2013.

[16] S. Hosoi, E. Takikawa, and M. Kawade. Ethnicity estimation with facial images. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[17] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017.

[18] M. T. Islam, C. Greenwell, R. Souvenir, and N. Jacobs. Large-Scale Geo-Facial Image Analysis. *EURASIP Journal on Image and Video Processing (JIVP)*, 2015(1), 2015.

[19] M. T. Islam, S. Workman, and N. Jacobs. Face2gps: Estimating geographic location from facial features. In *International Conference on Image Processing*, 2015.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[21] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[22] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz. Illumination-aware age progression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[24] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] K. Krys, C.-M. Vauclair, C. A. Capaldi, V. M.-C. Lun, M. H. Bond, A. Domínguez-Espinosa, C. Torres, O. V. Lipp, L. S. S. Manickam, C. 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.

[26] H. Kwak and B.-T. Zhang. Ways of conditioning generative adversarial networks. *arXiv preprint arXiv:1611.01455*, 2016.

[27] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *CVPR Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, 2015.

[28] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866), 2008.

[29] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015.

[31] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[32] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[33] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[35] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[36] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[38] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. 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.

[39] N. Srinivas, H. Atwal, D. C. Rose, G. Mahalingam, K. Ricanek, and D. 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] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[41] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[42] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[43] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1991.

[44] J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[45] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, 2016.

[46] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. *IEEE International Conference on Computer Vision*, 2017.

[47] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.