LEARNING A FACE SPACE FOR EXPERIMENTS ON HUMAN IDENTITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models of human identity and appearance have broad applicability to behavioral science and technology, but the exquisite sensitivity of human face perception means that their utility hinges on alignment of the latent representation to human psychological representations and the photorealism of the generated images. Meeting these requirements is an exacting task, and existing models of human identity and appearance are often unworkably abstract, artificial, uncanny, or heavily biased. Here, we use a variational autoencoder with an autoregressive decoder to learn a latent face space from a uniquely diverse dataset of portraits that control much of the variation irrelevant to human identity and appearance. Our method generates photorealistic portraits of fictive identities with a smooth, navigable latent space. We validate our model's alignment with human sensitivities by introducing a psychophysical Turing test for images, which humans mostly fail, a rare occurrence with any interesting generative image model. Lastly, we demonstrate an initial application of our model to the problem of fast search in mental space to obtain detailed police sketches in a small number of trials.

1 INTRODUCTION

Face processing tasks are popular in computer vision and machine learning, yet they often focus on specific goals such as detecting or recognizing faces in diverse pose and lighting conditions, a useful aspiration for social media and security applications. Meanwhile, psychologists and neuroscientists study the way that human process faces, but deal with simpler models, and often restrict their stimulus sets to small collections of real photographs, photographic averages, or uncanny syntheses from 3D face parameterization software (Inversions, 2008). Ideally, one type of good latent face space is one that aligns with human processing sensitivity, and such a space could also double as a candidate stimulus space to both study human face representation and assist in human-centric tasks. To our knowledge, such a space does not yet exist.

Figure 1: Random fictive identity from our model (512×512 px)

In the context of human interaction, decoding latent representations into visually coherent sets of pixels is just as important as the face space itself, as machine feature representations are not perceptible to humans. The problem aligns with image generation and feature inversion tasks in machine learning. Few machine learning methods are successful in generating realistic images for complex domains, and there is little pressure to attain perfection given that human-interpretability is not often a primary goal. Further, using such spaces to study humans places strong constraints on how biased the dataset should be (e.g., redundant identities of attractive celebrities carves out only one small portion of human mental face space). The perfect generative face model should aim to capture a broad range of human identity, produce realistic image renderings to represent the underlying latent face space, and be fast enough for human interaction. This could better assist current experimental methods for studying humans (Sanborn & Griffiths, 2007; Greene et al., 2014;

Vondrick et al., 2015), or applications that allow humans to communicate detailed information from their mental representations in the absence of being an expert visual artist.

In what follows, we offer the following contributions:

- We present a novel face dataset with well-controlled variation, aimed at allowing us to learn useful generative models of human identity with highly realistic reconstructions and samples.
- We outline a visual Turing test for assessing the quality of such models for use with humans, and show that our best model nearly masters this test.
- We show that samples from our model can be enlarged and drastically improved while preserving identity.
- Lastly, we demonstrate an initial application of our model to quickly extracting information from human mental representations, yielding expertless "police sketching", improving on previous methods by a factor of 10.

2 LEARNING A LATENT FACE SPACE FOR HUMAN EXPERIMENTS

The face space learned by a network reflects in part the images used to train it. Many image sets of faces used for training deep neural networks are confounded by variation unrelated to identity, while lacking representative variation essential to it. Confounded properties include those of the photographer and equipment (photographer, camera model, lens model, focal length, focal point, aperture, exposure time, distance to subject, camera placement, light design), the subject's ephemeral state (facial expression, pose, arousal), the environment (background imagery, ambient lighting), and post-processing (resolution, digital format, color grading, white balance, gamma correction, compression quality, watermarking, and digital alteration of skin complexion, hair color, and face shape), among others. Though learning a representation of identity that is invariant with respect to these properties requires a training set that varies along them, holding these properties constant greatly simplifies the learning problem in contexts where such invariance is irrelevant (Zhang & Gao, 2009). And though there are public datasets of human faces that control much of this variation, they tend to have too few unique individuals, precluding the possibility of learning a very universal face space.

Our goal is to produce a latent space of human identity useful for empirical experiments with humans. For this reason, it is just as important that this representation be decodeable into images that are free of distortions and artifacts, because human judgments of such images may often mistake them for features, or otherwise remove the natural context of the visual stimuli that humans often reason within. This is a difficult problem in machine learning, and will likely require a dataset with well-controlled variation so it can be learned using current methods, yet it also requires just the right diversity, avoiding heavy bias towards attractive celebrities and certain ethnicities.

2.1 THE HUMANÆDATASET

To assist in the effective training of a representative latent face space for human identity, we use a novel dataset based on Humanæ, an artistic work by Angèlica Dass that explores human diversity by mapping the gamut of skin tone across over $3,300$ people, with minimal constraints on participant selection. Of particular note is that the images are so nearly unvaried in their production that variation across the images focuses solely on the diversity of human visual identity. The abundance of controlled variation relevant to identity makes the dataset useful despite its relatively small size when compared to other face datasets used to train deep neural networks.

The dataset consists of $3,353$ front-facing portraits scraped from the project's website (http://humanae.tumblr.com/) with permission. Portraits were first resized by Lanczos resampling to 1024×1024 pixels. Next, the portraits were aligned through Procrustes superimposition of facial landmarks, which rotates, scales, and translates each portrait to minimize the Procrustes distance between the detected landmarks and a target, in this case a composite that places each landmark in its average position across all portraits. Facial landmarks were detected using a pre-trained ensemble-of-regression-trees detector (Gerbrands, 1981). After cropping a 640×640 square from the center of each portrait, they were then downsampled to the training size of 512×512 pixels by Lanczos resampling.

Figure 2: Sample images from the aligned/cropped Humanæ dataset.

We note that, although this dataset is small by the standards of modern deep learning applications, it is nevertheless effective given its well-controlled production by the artist. Eighty images from this dataset can be seen in Figure 2. The background of each image was matched to the skin tone of each face by the artist, a property we left intact.

2.2 Generative Models

To assess candidate latent spaces, we employed two different families of generative models, and one hybrid. Generative Adversarial Networks (GANs; Goodfellow et al., 2014) are often known to have the best sample quality, an important requirement for creating human-viewable images, and although they can also suffer from mode collapse, affecting sampling, they often learn reasonably good latent spaces that might be well-suited to our dataset. In particular, we use WGAN-GP (Gulrajani et al., 2017) because it showed slight improvements over other GANs we trained. Samples from the best performing model are shown in Figure 3. As expected, the quality is high, despite numerous artifacts.

Figure 3: Samples from WGAN-GP trained on humanæ.

The second type of model we explore are variational autoencoders (VAEs; Kingma & Welling, 2013). Most VAEs produce samples and reconstructions that are too smooth for our purposes, but the added help of a perceptual loss may actually be competitive given the high degree of alignment in our dataset. To this end, we opted to train a deep feature-consistent variational autoencoder (DFC-VAE; Hou et al., 2017). Surprisingly, samples from this model are indeed competitive with WGAN-GP (Figure 4), although the model has trouble rendering hair.

Figure 4: Samples from DFC-VAE trained on humanæ.

Lastly, we opted not to train autoregressive models, which often obtain the most competitive likelihoods, because they do not learn an easily traversable latent space. Instead, we compromised by training a hybrid model, PixelVAE (Gulrajani et al., 2016), which captures high-level variation with a simple latent variable, and uses a partially autoregressive decoder to handle finer visual detail. This

is a particularly good match for capturing human identity from portraits because it allows for an accessible latent space, while encoding fine image detail elsewhere. Indeed, we find that this approach is effective for our data. The samples in Figure 5 appear perfect in many respects, including striking detail, variation, and little to no artifacts. While it seems clear to us that this model is superior (and is a good candidate for human experiments), each model has its advantages. In the next section, we turn to a more concrete assessment of sample quality.

Figure 5: Samples from PixelVAE trained on humanæ.

3 Visual Turing Tests

While a number of quantitative heuristic standards exist for measuring sample quality, we are interested in learning a space that humans can interact with, allowing experimenters to probe human mental representations, among other applications. Faces have a particularly strong precedent here, because synthesized faces are notoriously susceptible to the uncanny valley effect (Mori, 1970), a product of the extremely sensitive nature of human face perception machinery. This allows to answer the question of which models can successfully fool humans and how this ability interacts with image size. The latter is also important because successful generative image models are often trained and tested with very small images (e.g., 32×32 px), yet is is unclear how many pixels are required for humans to make sensible judgments.

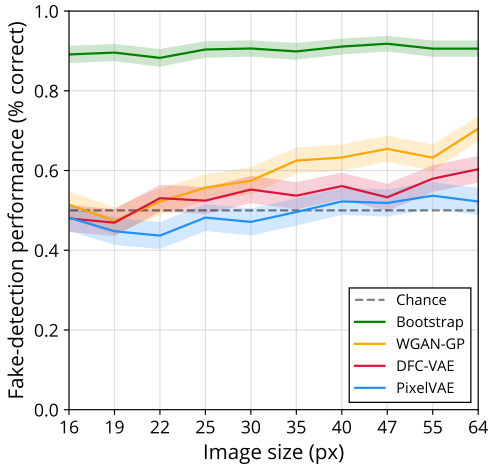

Figure 6: Psychophysical detection curves for each model. PixelVAE performs consistently better, and straddles the dotted line representing chance performance.

To evaluate the degree to which we can traverse this valley, we use a visual Turing test for synthesized images that draws from psychophysics to measure people's ability to distinguish real from synthesized photographs of faces, where the synthesized photographs are samples from each of the candidate models. This stands in contrast to two other forms of visual Turing test, one in which a human adjudicator determines which of two images was generated by human behavior (Lake et al., 2015), the other in which a visual recognition task is used to determine whether the observer performing the task is a human (Von Ahn et al., 2003; Geman et al., 2015).

We collected such judgments from 250 participants recruited from Amazon Mechanical Turk, each of whom performed 40 trials (10 trials for each log-spaced size from 16×16 to 64×64, the largest size and source output sample dimension from our models). One each trial, one image was an image from the training set and the other was a sample from one of the three candidate models. We also included a fourth control model that participants can easily distinguish from photographs: images composed of bootstrap-resampled pixels from portraits in the Humanæ dataset. The results are plotted in Figure 6. WGAN-GP does worst, likely due to diverse artifacts, whereas PixelVAE consistently outperforms all other models in fooling humans. In fact, it stays near or below the line of chance for all image sizes, effectively traversing the uncanny valley for this domain. It is notable that near-perfect samples are rarely obtained in generative image models, outside of small domains such as MNIST (LeCun et al., 2010) and SVHN (Netzer et al., 2011).

4 Improving image quality and exploring the latent space

While our visual Turing tests show that our samples can fool human subjects, they are still relatively small compared to the average comfortable stimulus sizes used in human experiments. This is an unfortunate feature of most modern generative image networks, which struggle to learn distributions over large sets of pixels. However, we note that the 64×64-pixel portraits generated by the PixelVAE carry enough information to convey identity (Bachmann, 1991). This suggests our samples can be improved by a super-resolution network that enlarges the image, inventing plausible fine details while preserving identity.

To do this, we use a generative adversarial network with an added content loss as our super-resolution upsampler network (Ledig et al., 2016). We train this network to enlarge 64×64-pixel portrait inputs to 512×512-pixel outputs. Panel A in Figure 7 shows enlarged PixelVAE samples. To our knowledge, these are among the highest quality synthesized facial identities produced by a deep neural network. Panel B demonstrates through enlargements of linear interpolations between random samples that identities and their composites are preserved, and only the quality and size of the image are improved.

We also tested for overfitting by randomly selecting 8 samples and finding their nearest neighbor in the training set, defined by pixelwise linear correlation. As evident in 9, samples and their nearest neighbors in the training set depict different identities.

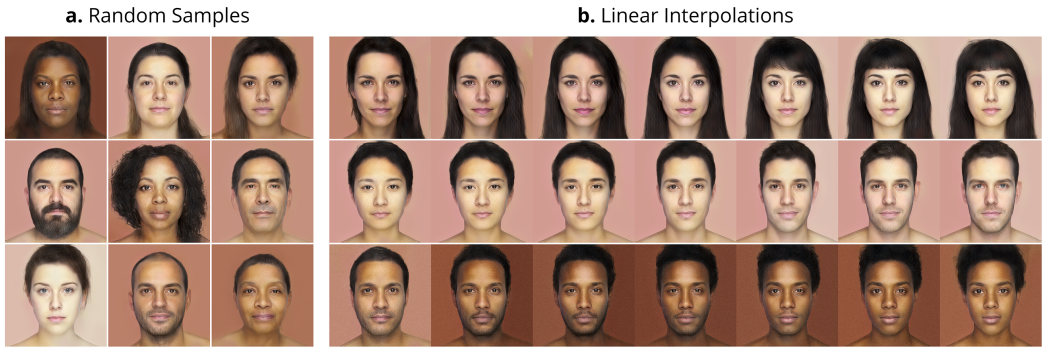

Figure 7: **a.** Decoded random samples drawn from the prior. **b.** Three decoded 7-point linear interpolations between two random samples drawn from the prior.

A useful machine representation to us is one that can be explored locally by humans, and yield interesting variations on a concept. Figure 8 shows four levels of noise added to a base sample. This simple procedure allows us to propose small adjustments to facial features or hair, as well as dramatically alter identity along several meaningful dimensions. In the next section, we will use these properties to interact with human explorers.

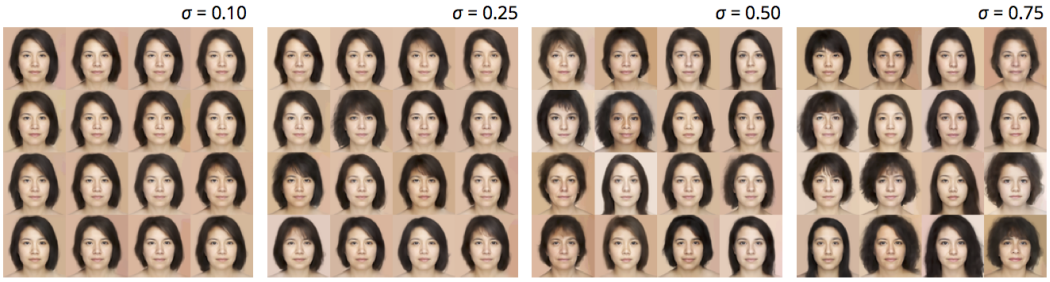

Figure 8: Latent noise perturbations at four levels of intensity.

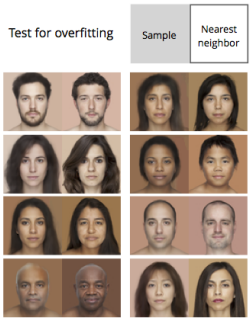

Figure 9: Samples and their nearest neighbors in the training set.

5 COMPOSITE SKETCHES VIA INTERACTIVE EVOLUTIONARY COMPUTATION

A key challenge in face modeling is the revelation of identities and appearances that are available only in a viewer's memory. For example, in police composite sketching, a trained artist with expertise in portraiture aims to translate witness testimony into a workable sketch that will enable law enforcement officials to recognize the perpetrator of a crime on sight (Mancusi, 2010). Or when creating cover art for a book, an illustrator may wish to render a character's likeness using vague suggestions from the text.

Searching the latent space of a generative model of human identity provides a way to render these obscured identities. For example, software-based facial-composite tools used by some law enforcement agencies guide a witness through a series of sequential decisions about facial features to arrive at a well-formed composite sketch. When that generative model has been learned directly from a corpus of human identity, the model's latent space may reflect the underlying variation in a way that benefits the search.

Various search and optimization algorithms are available, with some having proven particularly useful for human-in-the-loop search. Interactive evolutionary computation, a technique that inserts humans into various stages of an evolutionary process (e.g., as the fitness function), has been successfully used for human-in-the-loop search over digital designs and in other contexts (Takagi, 2001).

Here, we used a crowdsourced form of interactive evolutionary computation that leverages the human capacity to compare the relative resemblance of displayed portraits to a target identity. The particular evolutionary algorithm that we used comes from a family of so-called "Natural Evolution Strategies" that approximates stochastic gradient descent, making using of gradient information to efficiently find high-scoring pockets of the latent space (Spall, 1992). The search begins by selecting a seed, a point θ_t in the latent space, typically the origin. On each round of the search, a lineup of portraits is generated by adding spherical Gaussian feature noise to the seed. A set of participants then ranks the portraits according to their resemblance to the target identity (e.g. "a young boy with

red hair", "Barack Obama"). Each portrait is assigned a score $F(\theta)$ equal to its average rank across all the participants. The next seed, θ_{t+1} is set to $\theta_t + \alpha \frac{1}{n\sigma} \sum_{i=1}^{n} F_i \epsilon_i$. Thus, the next seed reflects the differences between high- and low-ranked examples to the extent that there is agreement between participants about the resemblance rankings.

Ten participants completed each round of the search. We found that interactive evolutionary computation over the latent space of our learned PixelVAE model efficiently found regions of the face space corresponding to target identities. Figure 8 shows the seeds and example images for 10 rounds of evolutionary search for three target categories.

Figure 10: Three 10-round evolutionary searches through the latent space. The final seed for each set of trials represents a collective mental template for the verbal description below.

6 Discussion

What we have presented here is a finely-tuned coupling between just the right variation, and powerful modern generative models. The samples produced by our models are interesting in their own right and rival the image quality of many state-of-the-art model–dataset pairs. However, we find such models most interesting in terms of what they tell us about and do for humans. Our hope is that the Humanæ dataset will inspire similar approaches, and that our models like the ones we trained can be used as both empirical and practical tools. In the future, it will be worth exploring larger datasets, fine-tuning current models, testing new models, and finding new applications for interacting with humans, and their own fascinating stimulus representations. Other questions remain as well. Can we approximately quantify the extent to which such spaces are universal? That is, do that contain nearly every identity? To what extent do these space already align well with human face representations? What is the scope of information that can be extracted from human minds using such methods? To answer these questions, a collaborative human–machine research methodology will be required, a trend we intend hope to establish here.

References

Talis Bachmann. Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology*, 3(1):87–103, 1991.

Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015.

Jan J Gerbrands. On the relationships between svd, klt and pca. *Pattern recognition*, 14(1-6): 375–381, 1981.

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*, pp. 2672–2680, 2014.

Michelle R Greene, Abraham P Botros, Diane M Beck, and Li Fei-Fei. Visual noise from natural scene statistics reveals human scene category representations. *arXiv preprint arXiv:1411.5331*, 2014.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pp. 1133–1141. IEEE, 2017.

Singular Inversions. Facegen modeller (version 3.3)[computer software]. *Toronto, ON: Singular Inversions*, 2008.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Yann LeCun, Corinna Cortes, and Christopher JC Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

Stephen Mancusi. *The Police Composite Sketch*. Springer Science & Business Media, 2010.

Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

Adam Sanborn and Thomas L Griffiths. Markov Chain Monte Carlo with people. In *NIPS*, pp. 1265–1272, 2007.

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992.

Hideyuki Takagi. Interactive evolutionary computation: Fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE*, 89(9):1275–1296, 2001.

Luis Von Ahn, Manuel Blum, Nicholas J Hopper, and John Langford. Captcha: Using hard ai problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 294–311. Springer, 2003.

Carl Vondrick, Hamed Pirsiavash, Aude Oliva, and Antonio Torralba. Learning visual biases from human imagination. In *Advances in neural information processing systems*, pp. 289–297, 2015.

Xiaozheng Zhang and Yongsheng Gao. Face recognition across pose: A review. *Pattern Recognition*, 42(11):2876–2896, 2009.