# Modifying social dimensions of human faces with ModifAE

**Chad Atalla**[1]
Computer Science and Engineering
University of California, San Diego
chada@ucsd.edu

**Amanda Song**[1]
Cognitive Science
University of California, San Diego
mas065@ucsd.edu

Bartholomew Tam
Electrical and Computer Engineering
University of California, San Diego
b4tam@ucsd.edu

Asmitha Rathis
Computer Science and Engineering
University of California, San Diego
arathis@eng.ucsd.edu

Gary Cottrell
Computer Science and Engineering
University of California, San Diego
gary@eng.ucsd.edu

## Abstract

At first glance, humans extract social judgments from faces, including how trustworthy, attractive, and aggressive they look. These impressions have profound social, economic, and political consequences, as they subconsciously influence decisions like voting and criminal sentencing. Therefore, understanding human perception of these judgments is important for the social sciences. In this work, we present a modifying autoencoder (ModifAE, pronounced "modify") that can model and alter these facial impressions. We assemble a face impression dataset large enough for training a generative model by applying a state-of-the-art (SOTA) impression predictor to faces from CelebA. Then, we apply ModifAE to learn generalizable modifications of these continuous-valued traits in faces (e.g., make a face look slightly more intelligent or much less aggressive). ModifAE can modify face images to create controlled social science experimental datasets, and it can reveal dataset biases by creating direct visualizations of what makes a face salient in social dimensions. The ModifAE architecture is also smaller and faster than SOTA image-to-image translation models, while outperforming SOTA in quantitative evaluations.

**Keywords:** neural networks; generative models; face recognition; social perception; image modification

## Introduction and Related Work

Humans quickly form subjective impressions of faces, judging traits like facial attractiveness, trustworthiness, and aggressiveness (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Despite the continuous scale and subjective nature of these social judgments, there is often a consensus among humans in how traits are perceived (e.g., human raters agree that certain faces appear relatively more trustworthy) (Falvello, Vinson, Ferrari, & Todorov, 2015; Eisenthal, Dror, & Ruppin, 2006). Social judgments of faces have a significant impact on social outcomes, ranging from electoral success to sentencing decisions (Dumas & Testé, 2006; Oosterhof & Todorov, 2008). Modeling is one way to understand these critical split-second impressions. Another way is through explicit human-judged experiments, which require carefully controlled datasets (e.g., building a dataset of faces which vary in "trustworthiness" while remaining consistent across age, gender, and "attractiveness"). In this work, we develop a system to model these impressions, visualize human perceptual biases, and create isolated image modifications for experimental datasets.

Choosing a subset of social impressions for modeling, we look to the 10k US Adult Faces Database (Bainbridge, Isola, & Oliva, 2013a). Bainbridge et al. (2013a) investigated what social attributes influence the memorability of a face. They compiled a list of 20 spontaneous social judgments and the corresponding opposite traits. Then, they assembled a human-judged dataset of trait ratings on 2,222 faces from the 10k US Adult Faces Database. Among the 40 traits, "aggressive," "attractive," "intelligent," "emotional," and "trustworthy" were frequently used in human-written face descriptions, played a significant role in face memorability, and had high rating agreement levels between human judges. Therefore, we choose them as the subset of social impressions for modeling in this paper.

To create controlled face datasets and visualize perceptual biases, a generative model is needed. Recent generative image models have been successful in creating high-resolution, high fidelity, and diverse images (Brock, Donahue, & Simonyan, 2018; Karras, Aila, Laine, & Lehtinen, 2017; Choi et al., 2017). However, in the face space, most generative models have focused on editing or modifying categorical and objective attributes, such as expression, gender, hair color, and identity (Choi et al., 2017). These categorical changes are referred to as "image to image translation." Here, we focus on modifying continuous attributes of an image, which we refer to as "continuous image modification" (Isola, Zhu, Zhou, & Efros, 2016). Regarding continuous image modification, there has been work on modifying the memorability (Khosla, Bainbridge, Torralba, & Oliva, 2013), and attractiveness of a face (Leyvand, Cohen-Or, Dror, & Lischinski, 2008), but these models do not generalize to wider sets of social impressions. Also, some researchers have generated fake faces with particular social impressions, but these models cannot modify real face images (Vernon, Sutherland, Young, & Hartley, 2014; Oosterhof & Todorov, 2008). So, no prior work has attempted to automatically modify general continuous social impressions of real face photographs.

Conditional generative adversarial networks (GANs) (Goodfellow et al., 2014) have become the most popular tool for the image to image translation task, so we compare against a recent GAN as a state-of-the-art (SOTA) reference point (Isola et al., 2016; Mirza & Osindero, 2014; Lee & Seok, 2017). StarGAN (Choi et al., 2017) is a SOTA conditional GAN that can modify multiple binary categorical traits at once, maintaining identifying traits of the original image using "cycle consistency" (Zhu, Park, Isola, & Efros, 2017). StarGAN consists of two networks: a generator and discriminator. The generator takes an image and a set of desired categorical traits, producing a modified image. The discriminator takes an image and makes a prediction about its realism and categorical traits. By comparing the fake images to genuine images, the discriminator gives feedback to the generator about how to make the image and desired traits appear more realistic.

Despite the success of GANs in categorical image-to-image translation, they cannot perform continuous image modification without binarizing the task and have architectural downsides. GANs typically have many parameters and long training times. They are also sensitive to hyper-parameter selection and the delicate balance between generator and discriminator training. Therefore, they can be difficult to train compared to a single-network model. Finally, they suffer from a lack of interpretability, offering no means of visualizing or understanding why the model makes the modifications it does.

In this work, we address these architectural concerns while designing a neural network to model and automatically modify continuous-scale face traits (rated from 1 to 9) in real face images. We create a sufficiently large dataset for training a generative model by combining CelebA images with a SOTA face impression predictive model (Liu, Luo, Wang, & Tang, 2015). Enabling interpretable bias visualization and controlled dataset creation for human face impressions, we introduce ModifAE. ModifAE (pronounced "modify") is a single-network image modification autoencoder.

## Subjective Judgment Face Dataset

### Building a Large Scale Facial Impression Dataset

To train a generative model on continuous face traits, we need a large and diverse dataset. We start with images from the CelebA dataset (Liu et al., 2015), which are annotated with binary categorical labels such as "wearing a hat" but lack continuous ratings of social impressions.

To generate continuous social impression ratings of these faces, we use our previous social impression predictive model (Song, Li, Atalla, & Cottrell, 2017). The model was trained on a smaller dataset (2,222 faces from the MIT 10k US faces dataset (Bainbridge, Isola, & Oliva, 2013b)) that had been annotated with ratings of 40 social traits on a scale from 1 to 9 by 15 raters for each face. Now, we focus on the subset of traits with the highest correlation between human judges: emotional, aggressive, trustworthy, responsible, attractive and intelligent. We apply this predictive model to about 190,000 faces from the CelebA dataset. Example faces and their predicted ratings are shown in Figure 1. Note that 6-8 are high ratings, and 2-4 are low ratings.

### Validating the Algorithm-Augmented Dataset

Evaluating the effectiveness of this algorithm-augmented dataset, we collect human judgments of the model's predictions in two ways: pairwise comparison and single image ratings. All participants were recruited from Amazon Mechanical Turk (AMT).

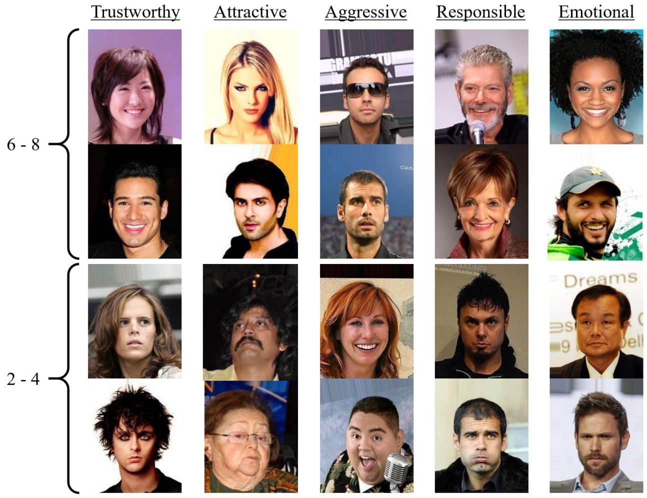

Figure 1: CelebA faces and their predicted traits.

Table 1: Validation of the impression prediction model

| Attribute | Accuracy | Attribute | Correlation |
|---|---|---|---|
| Aggressive | 0.95 | Aggressive | $0.76^{***}$ |
| Emotional | 0.92 | Attractive | $0.90^{***}$ |
| Trustworthy | 0.88 | Trustworthy | $0.73^{***}$ |
| Responsible | 0.78 | Intelligent | $0.62^{***}$ |

For pairwise comparison, we test four attributes: aggressive, responsible, trustworthy and emotional. For each trait, we compose 40 pairs of images. Within each pair, one is from the 40 faces of highest scores, and the other is from the 40 faces of lowest scores, as predicted by the model. We then ask human participants which face better exemplifies the predicted trait. Each trait's 40 pairs are evaluated by 30 AMT workers. We then calculate the overall likelihood that the face of higher predicted score is chosen, which we call "accuracy." The results are shown in the left side of Table 1. The attributes predicted by the model align well with human judgments.

For the single-rating experiment, we examined four traits: attractive, aggressive, trustworthy and intelligent. For each trait, we chose 80 faces whose predicted scores are evenly spread across a range of predictions (i.e., from 2 to 8). Each participant is presented with a random sequence of 80 faces, and is asked to give each face a rating on a 1-9 scale for the specified trait. Every face is rated by 15 subjects, and we compute the average. Lastly, we compute the Spearman rank correlation between the average human ratings and the model's predictions of the same set of faces for each trait. For all four traits, human average ratings are significantly correlated with model predictions ($^{***}$ indicates $p < 0.001$), as seen in the right side of Table 1.

Given the pairwise and single image rating results, we consider the predicted scores as roughly equivalent to human judgments. Hence, in the next section, we train our face modification model with these ratings.

## ModifAE

ModifAE is a single network autoencoder which implicitly learns to modify continuous face impression traits in images (illustrated in Figure 2). Here, we elaborate on the architecture, training procedure, and mechanism of the ModifAE model.

### Model Architecture

The ModifAE architecture consists of a single autoencoder with two (image and trait) sets of inputs which pass through an encoding stage, are fused (by averaging) in the middle of the network, and are then fed into an image decoder.

The image encoder and decoder are identical to the encode and decode portions of the StarGAN generator network, scaled to fewer channels (Choi et al., 2017). More specifically, the network has two downsampling convolutional layers with stride two, four residual blocks, a bottleneck with 16 channels, four more residual blocks, then two upsampling transposed convolutional layers with stride two (Choi et al., 2017). All layers have ReLU activation. We use the first half of this network (including the bottleneck) as the image encoder. We use the remainder of the network as the image decoder. Theoretically, this portion could consist of the encode and decode halves of any image autoencoder; we chose the architecture from StarGAN for the sake of comparability.

The trait encoder takes a 1-dimensional set of traits, feeds these into a single dense layer with Leaky ReLU activation, and reshapes the output to create a vector of the identical shape as the image encoder output. The outputs of the trait and image encoders are then combined into a single latent representation by averaging.

In order to encourage the model to encode the trait information, which is otherwise unnecessary to reproduce the image, 50% dropout is applied to the values from the image encoder. This is then averaged with the trait encoder output to arrive at the combined latent representation. The image decoder projects the representation back into image space, creating the single output image. The architecture is depicted in Figure 2, where “convs” refers to residual convolutional blocks from StarGAN.

### Training Procedure

ModifAE is exclusively trained on an autoencoding task. We train ModifAE using the Adam optimizer (Kingma & Ba, 2014) and train for 100 epochs on CelebA images (Liu et al., 2015). The objective is to optimize a single loss function based on two terms. We use the $L_1$ loss on the image autoencoder. We also optimize the $L_1$ loss between the trait encoder and image encoder. The total loss is:

$$L = \frac{1}{N}\sum_{p=1}^{N} |x_p - AE(x_p)| + |E(x_p) - E(y_p)| \qquad (1)$$

where $x_p$ is the $p^{th}$ image example, $y_p$ is its trait vector, $E(\cdot)$ is the result of the trait or image encoder, and $AE(\cdot)$ is the output of the full-architecture autoencoder. The second term in this loss function encourages the network to have a similar representation between the trait and the image encodings. The trait encoder obviously does not “know” what the image is, but this constrains the image encoding to include information about the trait.

### Why the Model Learns Implicitly to Modify Images

Each image is encoded along with its predicted traits. The image encoder compresses the image down to a bottlenecked latent space, where higher level features about the image are encoded. Simultaneously, the trait encoder projects the given traits to the same latent space, creating an average face representation with those ratings.

Because dropout is applied to the face encoding, the decoder has to use the trait information to “fill in the gaps” in the face representation. Therefore, at training time, faithfully reconstructing the image is reliant on information coming from the trait encoder, and the trait encoder learns to mimic average latent distributions of images with the provided ratings.

At test time, an image can be passed in with any desired traits. The trait encoder estimates the latent space for images with those traits, and the decoder responds by altering the face image towards the encoded trait. Hence, the output image resembles the original but is changed according to the provided traits.

## Experiments and Results

In this section, we provide examples of ModifAE’s modifications and interpretable transformation maps. We also report an experiment which quantitatively compares the effectiveness of ModifAE and StarGAN with a user study, and we numerically compare the ModifAE architecture with other relevant systems.

### Qualitative Evaluation

**Multi-Trait Traversals** Here, we show that ModifAE is capable of making continuous modifications on multiple traits with a single model (see Figure 3 and Figure 4). This enables ModifAE to modify some traits while holding others traits constant, which can be applied to creating datasets with controlled and isolated modifications for social psychology experiments.

For Figure 3, we trained ModifAE on two traits: “attractive” and “aggressive.” The picture in the upper left corner is the original. At the (0,0) point in Figure 3 (unattractive and not aggressive) the man’s mouth is fairly neutral, and his features are not very pronounced. As attractiveness and aggressiveness increase, the angles of the face become sharper, there is more definition of features like eyes and eyebrows, and the smile shrinks.

Figure 4 shows interpolations generated by two models. Each was trained on a social trait and a gender category from CelebA. Then, each trait was interpolated while holding the gender bias constant. The resulting figure shows how perception of “aggressiveness” may vary across genders. Likewise,

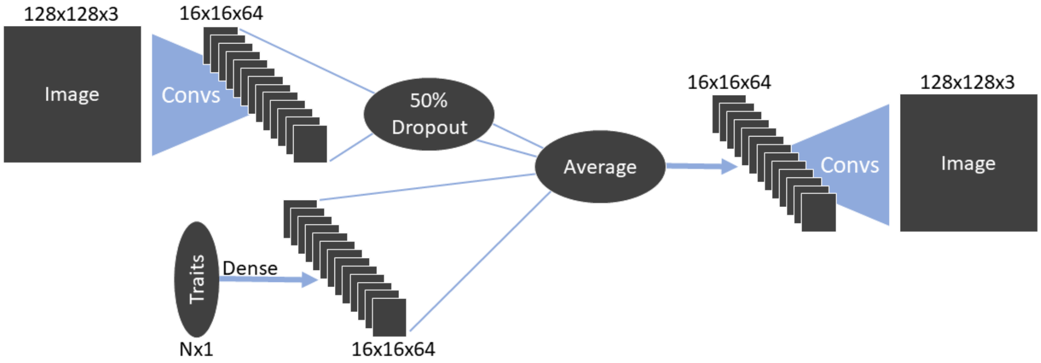

Figure 2: General illustration of ModifAE architecture.

this method can show how other traits may be less correlated with gender perception.

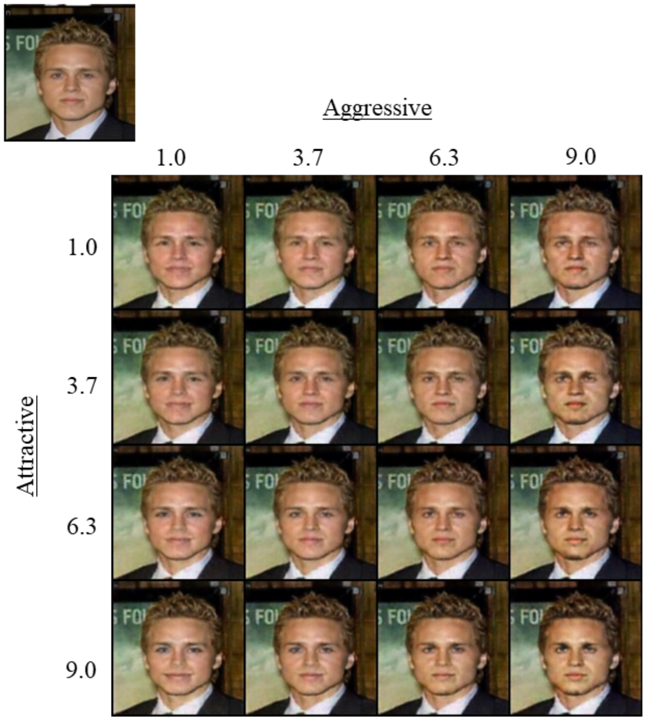

Figure 3: Continuous value, multi-trait image modification by ModifAE.

**Qualitative Comparison to StarGAN** Comparing our model to StarGAN (Choi et al., 2017), we binarize the continuous traits by doing a median split on the continuous-valued traits and train StarGAN on these two groups (low and high). This is necessary because StarGAN inherently only makes binary changes. The results are shown in Figure 5. While StarGAN produces high-resolution image reconstructions, they occasionally suffer from color distortions or lack of apparent changes. ModifAE makes subtle and reliable modifications to the original images, changing the way the social traits are perceived. In the images produced by ModifAE, more trustworthy faces smile more, and appear to have eyes set farther apart. The ModifAE attractive faces appear to smile more and notably have more well-defined eyes.

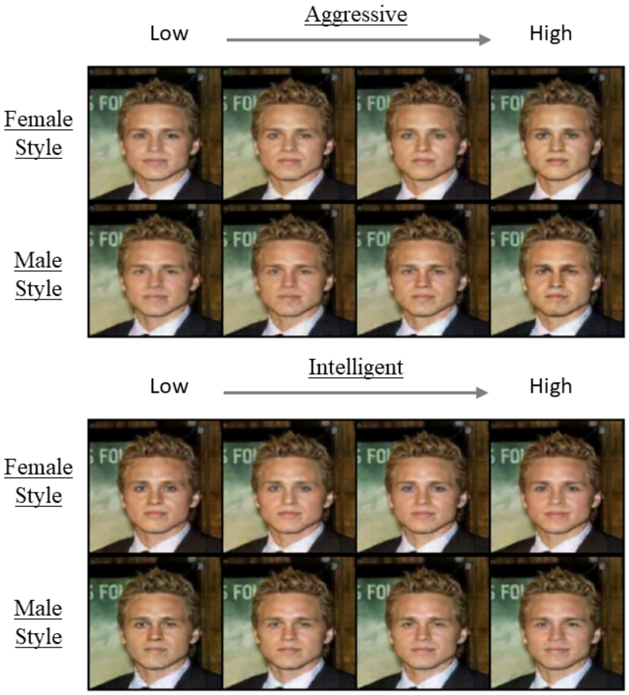

Figure 4: Continuous changes of a face while holding gender bias constant.

**Interpretable Transformation Maps** As mentioned above, ModifAE addresses the issue of interpretability in generative models. We provide a window into the model's representation of the traits by decoding the representation generated by the trait encoder without giving any actual image input. Figure 6 shows a traversal of the learned "trait faces" or "transformation maps" of attractiveness and intelligence. In this case, we trained the model on a combination

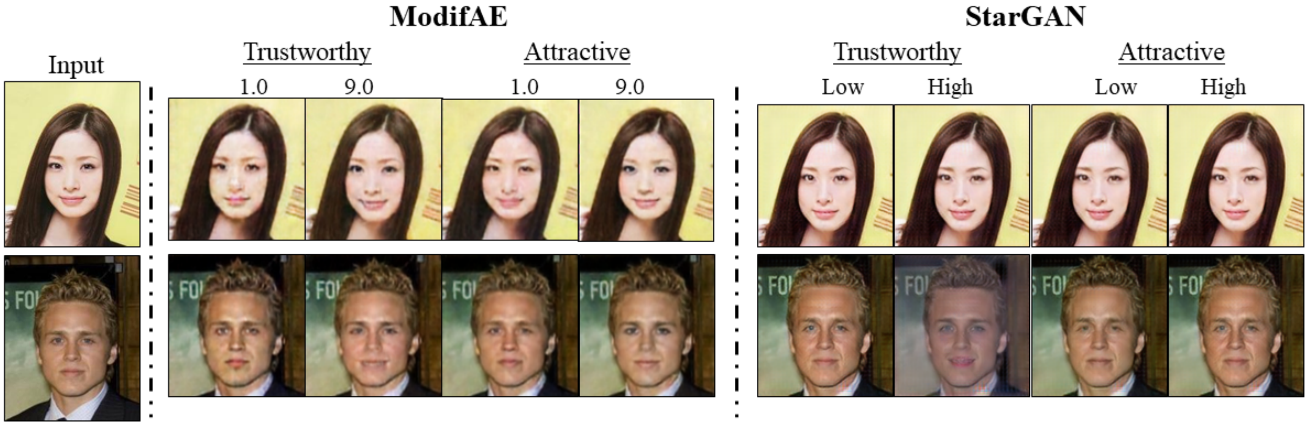

Figure 5: Comparison of ModifAE and StarGAN modifications.

of gender and the given trait, so we show a traversal of the model's representations for male and female faces separately.

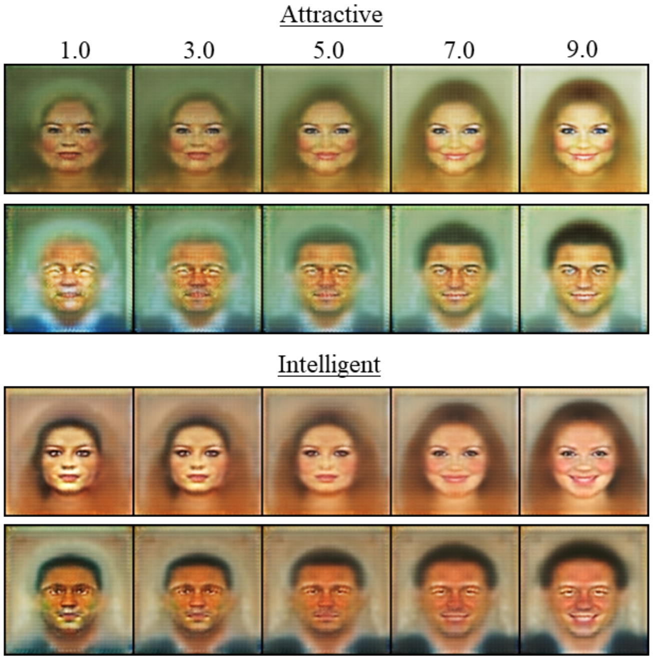

Figure 6: Visualization of model's internal perception of traits. Each is a traversal of a trait (increasing left to right) while gender is held constant.

## Quantitative Evaluation

**Quantitative Comparison with StarGAN** To evaluate the quality of ModifAE's continuous subjective trait modifications, we perform Amazon Mechanical Turk (AMT) experiments on four traits: aggressive, attractive, trustworthy and intelligent. For each trait, we created 90 image pairs, of which 80 are the same identity modified to be at high and low values of each trait. For StarGAN, we used a median split of low and high rated traits to train the model. ModifAE was trained as previously described. For each model, then, faces were modified to be low or high on each trait. Subjects judged which face had more of the particular trait. 10 pairs were repeats in order to judge subject consistency, and 10 pairs were unmodified CelebA faces with high and low ratings. These latter we called "ground truth" pairs to test whether subjects were paying attention. Subjects whose ratings on these pairs were at chance or below were rejected.

Hence, for each trait, we present participants with a sequence of 100 image pairs, and participants are asked to pick which image most exemplifies the trait in each pair.[1] Each pair was evaluated by 15 subjects.

We calculate the fraction of pairs in which subjects chose the image with the higher modified trait across all participants and all pairs. If they choose the face that was modified to be higher in the trait, then they agree with the model's modifications. The results are shown in Table 2. We perform a binomial test to determine whether each trait's accuracy is significantly below or above chance ($^{***}p < 0.001$). Note that the fourth column "Ground Truth" indicates the overall accuracy of the unmodified "ground truth" pairs. Given the variance in human impression judgments, these numbers serve as a reference ceiling for how well the models can perform.

**Evaluating ModifAE's Continuity** Since ModifAE is able to generate continuous modifications, we evaluated this property by creating two more same-face pairs: Ones modified to have low values and middle values, and ones modified to have middle values and high values. We obtain human agreement (accuracy) over the Low-Mid and Mid-High pairs for each of the four traits. The results are shown in Table 3.

## Model Size and Training Time

In contrast with GANs, ModifAE requires fewer parameters and less time to train. StarGAN takes about 24 hours to train on CelebA (Choi et al., 2017); ModifAE takes less than 11

[1] In a pilot experiment, we asked subjects to rate faces with different identities generated in a fine continuum, but found significant variance with no correlation to the intended scores, presumably because the images were not differentiable at that fine a grain.

Table 2: Comparison of ModifAE with StarGAN

| Attribute | **ModifAE** | StarGAN | "Ground Truth" |
|---|---|---|---|
| Aggressive | $0.68^{***}$ | $0.72^{***}$ | $0.90^{***}$ |
| Attractive | $0.68^{***}$ | 0.51 | $0.94$ $^{***}$ |
| Trustworthy | $0.63^{***}$ | 0.40 | $0.87^{***}$ |
| Intelligent | $0.68^{***}$ | $0.58^{***}$ | $0.81^{***}$ |

Table 3: ModifAE Low-Mid-High Level Self-comparison

| Attribute | Low-Mid | Mid-High | Low-High |
|---|---|---|---|
| Aggressive | $0.60^{***}$ | 0.52 | $0.68^{***}$ |
| Attractive | $0.59^{***}$ | 0.52 | $0.68^{***}$ |
| Trustworthy | $0.61^{***}$ | $0.53^{*}$ | $0.63^{***}$ |
| Intelligent | $0.60^{***}$ | 0.50 | $0.68^{***}$ |

Table 4: Model size for learning seven traits

| Model | CycleGAN | ICGAN | StarGAN | **ModifAE** |
|---|---|---|---|---|
| Parameters | 736M | 68M | 53M | **1M** |

hours. Table 4 shows the number of parameters required by different models trained on the CelebA dataset. The listed values are as reported in the original papers (Perarnau, van de Weijer, Raducanu, & Álvarez, 2016; Zhu et al., 2017) and in the parameter comparisons of Choi et al. (2017).

Note that the majority (over 40M) of StarGAN's parameters are in the discriminator network, and ModifAE uses a smaller version of the StarGAN generator. Also, ModifAE's relatively small trait encoder is the only part of the model which scales with supervising additional traits, so learning more traits with a single model is cheaper with ModifAE. Together, these properties mean that ModifAE takes over fifty times fewer parameters than any of the competing models.

# Discussion

## Quantitative Experiment Discussion

From Table 2, we can see that for all four traits, ModifAE produces pairs that yield above chance level human agreement. In three out of the four traits, ModifAE significantly outperforms StarGAN; whereas for the aggressive trait, StarGAN performs only slightly better than ModifAE. StarGAN is good at creating discrete changes in facial expressions, which accounts for this advantage.

From Table 3, we find that all the low-mid pairs yield significantly above chance accuracy, yet for mid-high level, only trustworthy pairs have accuracy slightly above chance ($p < 0.05^{*}$). This suggests that human psychological face space is nonlinear and has more differentiation towards the low- to mid-range of social dimensions. Another possibility is that when our model generates faces that are of more extreme scores (e.g. 8 or 9), the model is extrapolating, and produces artifacts that lead to that face being rejected. This speculation requires further analysis to be confirmed.

## Interpreting Transformation Maps

The interpretability of the model may be useful in the field of social psychology, giving researchers new suggestions about what features of a face are most important for perceiving a given trait. It can also elegantly summarize the average opinions and biases of a group of raters who have created a dataset, or serve as a visual heuristic for understanding which traits are most similar to each other in human perception.

The "intelligent" transformation map appears to show that bigger heads are rated as more intelligent (at least, pictures in which the head appears larger or closer). This suggests a bias that to our knowledge, has not been previously observed. Of course, in this case, it is simply faces that subtend a larger visual angle, rather than real-world head size. In further experiments, the head size should be normalized across images to avoid this potential bias. In addition, experiments could be run where image head size is systematically manipulated with the same face (judged by different subjects), to verify the bias.

The "intelligent" transformation map appears to show that bigger heads are rated as more intelligent (at least, pictures in which the head appears larger or closer). This suggests a bias that to our knowledge, has not been previously observed. Of course, in this case, it is simply faces that subtend a larger visual angle, rather than real-world head size. In further experiments, the head size should be normalized across images to avoid this potential bias. In addition, experiments where humans rate images with systematically manipulated head size could be run to verify the bias.

# Conclusion

In this paper, we propose ModifAE: a single network autoencoder, which performs continuous image modification on subjective face traits in an interpretable manner. ModifAE does not require training multiple networks or designing hand-tailored features for image modification. Instead, a single network is trained to autoencode an image and its traits through the same latent space, implicitly learning to make meaningful changes to images based on trait values. Our experiments show that ModifAE requires fewer parameters and takes less training time than existing general methods. It also provides interpretable transformation maps of traits which demonstrably highlight biases in datasets and salient features in human perception of traits. Additionally, in this work, we compute and verify novel continuous subjective trait ratings for CelebA faces. Finally, we demonstrate that ModifAE makes more meaningful continuous image traversals than an equivalent SOTA method (Choi et al., 2017) and examine human agreement with ModifAE modifications in the subjective face trait space.

# References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013a). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013b). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., & Choo, J. (2017). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, *abs/1711.09020*.

Dumas, R., & Testé, B. (2006). The influence of criminal facial stereotypes on juridic judgments. *Swiss Journal of Psychology*, *65*(4), 237–244.

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*(1), 119–142.

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. *Social Cognition*, *33*(5), 368.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014, december 8-13 2014, montreal, quebec, canada* (pp. 2672–2680).

Isola, P., Zhu, J., Zhou, T., & Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *CoRR*, *abs/1611.07004*.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2013). Modifying the memorability of face photographs. In *International Conference on Computer Vision (ICCV-2013)* (pp. 3200–3207).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Lee, M., & Seok, J. (2017). Controllable generative adversarial network. *CoRR*, *abs/1708.00598*.

Leyvand, T., Cohen-Or, D., Dror, G., & Lischinski, D. (2008). Data-driven enhancement of facial attractiveness. *ACM Transactions on Graphics (TOG)*, *27*(3), 38.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (iccv)*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *CoRR*, *abs/1411.1784*.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Perarnau, G., van de Weijer, J., Raducanu, B., & Álvarez, J. M. (2016). Invertible conditional gans for image editing. *CoRR*, *abs/1611.06355*.

Song, A., Li, L., Atalla, C., & Cottrell, G. (2017). Learning to see people like people: Predicting social perceptions of faces. In *Proceedings of the 39th annual meeting of the cognitive science society, cogsci 2017, london, uk, 16-29 july 2017*.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Reviews of Psychology*, *66*(1), 519.

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361.

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision, ICCV 2017, venice, italy, october 22-29, 2017* (pp. 2242–2251). doi: 10.1109/ICCV.2017.244