UNIVERSITY OF CALIFORNIA SAN DIEGO

Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Devendra Pratap Yadav

Committee in charge:

Professor Garrison W Cottrell, Chair
Professor Manmohan Chandraker
Professor Sanjoy Dasgupta

2020

The thesis of Devendra Pratap Yadav is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California San Diego

2020

TABLE OF CONTENTS

Signature Page . . . iii

Table of Contents . . . iv

List of Figures . . . vi

List of Tables . . . viii

Acknowledgements . . . ix

Vita . . . x

Abstract of the Thesis . . . xi

Chapter 1 Introduction . . . 1
1.1 Related Work . . . 4
1.1.1 Generative Adversarial Networks . . . 4
1.1.2 Face Attribute Modification . . . 4
1.1.3 Modifying the First Impressions Evoked by Faces . . . 5

Chapter 2 Datasets . . . 6
2.1 Social Traits Dataset . . . 6
2.1.1 Data Collection Procedure . . . 7
2.1.2 Rater Consistency for Social Traits . . . 9
2.2 Attribute Prediction Model: PredCNN . . . 10
2.3 Celeba-HQ-60k . . . 10
2.3.1 Generating Additional Attribute Labels . . . 12

Chapter 3 GSC-GAN: Gated Skip Connection Generative Adversarial Network . . . 14
3.1 Problem Formulation . . . 15
3.2 Gated Skip Connection (GSC) Module . . . 15
3.3 GSC-GAN Architecture . . . 18
3.4 Loss Functions and Objective . . . 19
3.4.1 Adversarial Loss . . . 19
3.4.2 Attribute Modification Loss . . . 20
3.4.3 Reconstruction loss . . . 21
3.4.4 Final Objective . . . 21

Chapter 4 Experiments . . . 22
4.1 Training Procedure . . . 22
4.2 Evaluation Metrics . . . 23
4.3 Continuous Attribute Modification . . . 24

4.3.1 Qualitative Evaluation 24
4.3.2 Quantitative Evaluation 25
4.4 Discrete Attribute Modification 28
4.4.1 Qualitative Evaluation 28
4.4.2 Quantitative Evaluation 29
4.5 Parameter Efficiency 31
4.6 Model Ablations 32
4.7 Visualizing Modified Features Using GSC-GAN 33

Chapter 5 Modifying Perceived Social Impressions of Faces 37
5.1 Correlated Visual Features for Social Traits 38
5.2 Human Validation Study 40

Chapter 6 Conclusion 42
6.1 Future Work 42
6.2 Broader Impact and Ethics 43

Appendix A High Resolution Face Modification Results 45
A.1 Discrete CelebA Attributes 45
A.2 Continuous Facial Action Units 45
A.3 Perceived Social Traits 45

Bibliography 50

LIST OF FIGURES

Figure 1.1: Continuous modification on a relative scale for a CelebA attribute (young), Action Unit (AU25) and social trait (kind) using our proposed model, GSC-GAN. Please zoom in for details. 3

Figure 2.1: The task page shown to human raters on Amazon Mechanical Turk to collect ratings for the Social Traits Dataset. 7

Figure 2.2: Morphed faces representing the 5 highest and 5 lowest rated faces per gender for selected social traits. 8

Figure 2.3: Intraclass Correlation Coefficient (ICC) for each trait in the Social Traits Dataset. 9

Figure 2.4: Architecture of our attribute prediction model: PredCNN, used to label CelebA-HQ-60k and evaluate accuracy of GAN-based attribute modifications. 11

Figure 2.5: CelebA-HQ-60k samples obtained after denoising and GAN based super-resolution. Zoom in for details. 12

Figure 3.1: Architecture of our Gated Skip Connection (GSC) module. $\circ$ denotes element-wise product. 16

Figure 3.2: Architecture of GSC-GAN, showing the encoder-decoder architecture using GSC modules. 17

Figure 4.1: Action Unit modification with GANimation, STGAN, and GSC-GAN at 256×256 resolution. The columns show modification to 5.0 absolute AU intensity. Please zoom in for details. 25

Figure 4.2: Interpolation along relative AU values using GSC-GAN at 256×256 resolution. Relative values of -1 and 1 denote absolute AU intensity of 0 and 5 respectively. 26

Figure 4.3: AU25 (Lips Part) modification to 5.0 absolute intensity with occluded faces for GANimation, STGAN, and GSC-GAN. 27

Figure 4.4: CelebA attribute modification results for GSC-GAN, STGAN, STGAN-BCE and GANimation. Each attribute is inverted and column labels show the target value. 29

Figure 4.5: A visualization of residual features learned by GSC-GAN for Action Unit attributes. We also show the pixel-wise difference between input and output image for comparison. 34

Figure 4.6: A visualization of residual features learned by GSC-GAN for CelebA attributes. We also show the pixel-wise difference between input and output image for comparison. 35

Figure 4.7: A visualization of residual features learned by GSC-GAN for CelebA attributes, for relative attribute values of -1.5 and 1.5. We also show the pixel-wise difference between input and output image for comparison. . . . 36

Figure 5.1: Social trait modification to low and high values using GSC-GAN. Top two rows show generated images, and the bottom two rows show the modified face features (by visualizing residual features). 38
Figure 5.2: Social trait modification to low and high values using GSC-GAN. Top two rows show generated images, and the bottom two rows show the modified face features (by visualizing residual features). 39
Figure 5.3: Human verification accuracy for social trait modification using GSC-GAN. ICC denotes Intraclass Correlation Coefficient which measures rater agreement when rating a social trait. 40

Figure A.1: CelebA attribute modification using GSC-GAN at 512×512 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details. 46
Figure A.2: Action Unit attribute modification using GSC-GAN at 512×512 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details. 47
Figure A.3: Social Trait attribute modification using GSC-GAN at 256×256 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details. 48
Figure A.4: Social Trait attribute modification using GSC-GAN at 256×256 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details. 49

LIST OF TABLES

Table 3.1: Specification of different components used in GSC-GAN's generator and discriminator. 19
Table 3.2: Architecture of GSC-GAN for 128 and 512 image size. See Table 3.1 for specification of components. N_a denotes number of attributes in the dataset. A column represents the components in a sequential order. 20

Table 4.1: Evaluation of GSC-GAN (ours), STGAN and GANimation for continuous Action Unit (AU) attributes. Training image size is denoted by subscripts. Note: RMSE uses AU intensity in [0,5]. 27
Table 4.2: Evaluation of GSC-GAN (ours), STGAN, STGAN-BCE and GANimation for CelebA discrete attributes. Models are trained at image size as denoted by subscripts. 30
Table 4.3: Number of generator (G) and discriminator (D) parameters for GSC-GAN and STGAN at different image sizes (denoted by subscript). 32
Table 4.4: Quantitative results of a model ablation study for discrete CelebA attributes at 128×128 resolution. GSC-GAN-gru replaces GSC module with a GRU based gating module. 33
Table 4.5: Quantitative results of a model ablation study for Action Unit attributes at 128×128 resolution. GSC-GAN-gru replaces GSC module with a GRU based gating module. 33

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Garrison Cottrell for his guidance as the chair of my committee. The thesis would not be possible without his support in developing the ideas. Apart from the vast subject knowledge, his humor was essential throughout my work at UC San Diego

The work done during this thesis is part of a larger inter-disciplinary project to study social first impressions of faces. I would like to acknowledge Professor Ed Vul, Amanda Song and Weifeng Hu for their work on the project which laid the foundation for my thesis.

I would also like to acknowledge members of The Cottrell Lab a.k.a. GURU: Gary's Unbelievable Research Unit. The frequent exchange of ideas, insights and discussions with members of the group greatly improved the quality of my work.

I acknowledge the Computer Science department at UC San Diego which provided me with ample computational and knowledge resources to conduct research to the best of my abilities.

All chapters, in part, have been submitted for publication of the material. Yadav, Devendra Pratap; Hu, Weifeng; Song, Amanda; Vul, Edward; Cottrell, Garrison. "Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification". The thesis author was the primary investigator and author of this paper.

VITA

2018	Bachelor of Technology in Computer Science and Engineering, Indian Institute of Technology Ropar
2020	Master of Science in Computer Science, University of California San Diego

ABSTRACT OF THE THESIS

Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification

by

Devendra Pratap Yadav

Master of Science in Computer Science

University of California San Diego, 2020

Professor Garrison W Cottrell, Chair

Face image modification is a variant of the image-to-image translation task where we modify features of a face image to evoke given target attributes, while preserving the identity of the pictured person. Generative Adversarial Networks (GANs) using an encoder-decoder architecture have been widely used to modify both discrete and continuous face attributes, with a few different architectures designed to address the challenge of preserving identity through the modification. We propose a novel GAN architecture that introduces gated skip connections in the generator's decoder for this task. Our model enables high fidelity (512×512) modification with minimal changes to irrelevant facial regions, while using fewer parameters than existing

approaches. We demonstrate the model on discrete CelebA attributes, continuous facial Action Unit labels, and perceived social impression traits such as “attractive”, “kind”, and “trustworthy”. Our model is also able to selectively visualize the modified face features, allowing us to extract plausible visual explanations for face attributes, including, for the first time, social impression traits. An experiment with human raters validates that our model can effectively alter a face’s social impressions.

Chapter 1

Introduction

Generative Adversarial Networks (GANs) have shown state-of-the-art performance on several image-to-image translation tasks such as image super-resolution [WYW+18, HSU18], synthetic-to-real images [HLBK18] and semantic maps to real images [WLZ+18, PLWZ19]. Face attribute modification is an image-to-image translation task where face images are transformed according to discrete or continuous attributes [CCK+18, HZK+19, PAM+18].

A critical feature of face modification is that it must preserve the person's identity, otherwise it reduces to a conditional GAN that generates arbitrary faces based on only attribute labels. Spatial attention based approaches [ZKSC18, PAM+18] achieve this by copying pixels from the input image, but this strategy only works well when modifications are region specific. For instance, if the face modification task aims to change the face pose, a large region of the face needs to be generated and the utility of input image pixels is limited. Instead of pixels, copying more abstract features of the input image though skip connections, in an encoder-decoder based generator, preserves details for arbitrary modifications [HZK+19, IZZE17]. However, as we add more skip connections in decoder layers, the attribute modification intensity decreases, when training with unpaired training samples [LDX+19]. Liu et al. [LDX+19] propose STGAN to address the decrease in modification capacity by selective transfer of encoder features through

skip connections. STGAN uses Selective Transfer Units (STUs), which are inspired by Gated Recurrent Units [CVMBB14], to select and transform encoder features. A STU generates a state which is passed as input to the next STU, and an output which is concatenated with decoder features. This approach generates features corresponding to the image output through STUs as well as the decoder layers. Compared to a basic U-net [RFB15] architecture, STGAN's approach adds an additional set of feature decoding layers for STUs. This leads to an increase in the number of generator parameters, making it inefficient as we scale to higher image resolutions.

We aim to perform selection of useful encoder features in a parameter efficient manner. To this end, we propose a novel gating module that selects which encoder features to use in skip connections to preserve image details. We use our Gated Skip Connection (GSC) module to replace decoder layers in a U-net [RFB15] architecture. For a given generator depth and number of feature channels per layer, our approach requires roughly half as many generator parameters as STGAN, while maintaining equivalent performance. We also compare our method with spatial attention based GANimation [PAM+18], for continuous modification of facial Action Units [FE78], and show significant improvement in image quality and identity preservation. Our efficient architecture scales from 128×128 to 512×512 resolution using only 43% more parameters, producing accurate attribute modifications with ~99% identity preservation accuracy. To the best of our knowledge, no existing work tackles unsupervised face attribute modification at 512×512 resolution.

Since our model, Gated Skip Connection GAN (GSC-GAN), has access to the original encoded features though skip connections, the decoder primarily generates face modifications; allowing us to explain its operation by visualizing the modified features. We use this to study a novel dataset of perceived social traits from face images. Our model extracts visual explanations for traits such as intelligent, kind and healthy. We verify our model's correctness through a human rater study and achieve ~81% accuracy. This shows that our model can perform realistic face modifications to alter the social perception of faces. To summarize, the key contributions of our

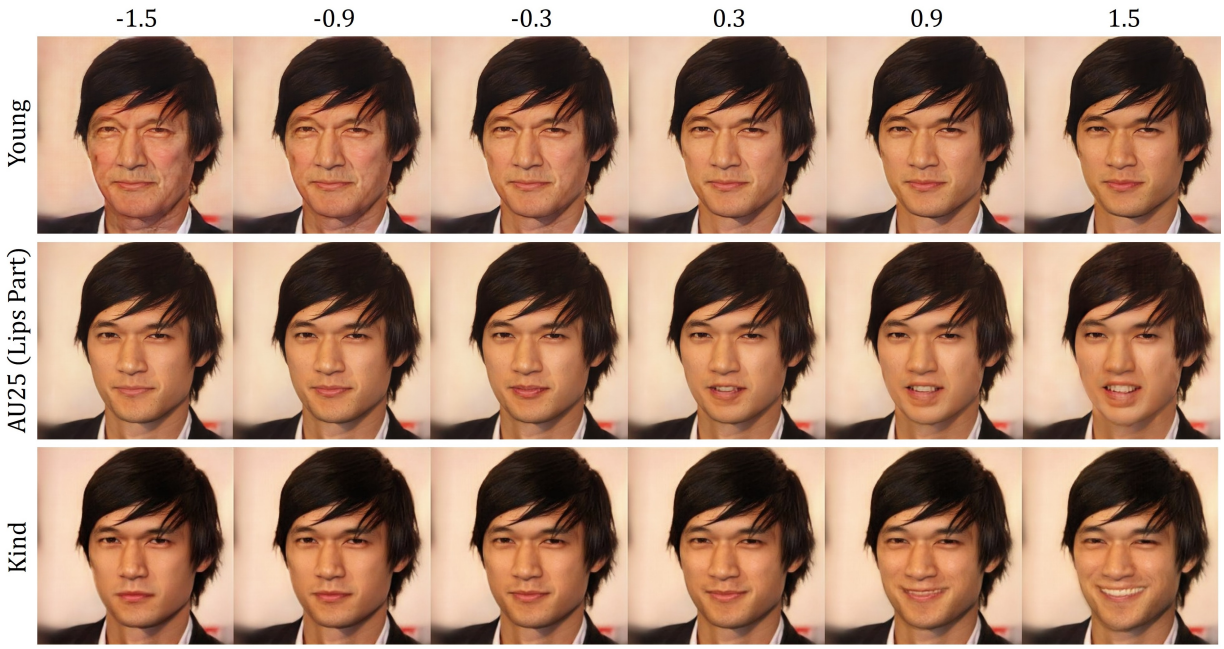

Figure 1.1: Continuous modification on a relative scale for a CelebA attribute (young), Action Unit (AU25) and social trait (kind) using our proposed model, GSC-GAN. Please zoom in for details.

work are:

- A Gated Skip Connection (GSC) module that selects useful features from skip connections in a U-net architecture. Our proposed model achieves equivalent performance in face editing to state-of-the-art STGAN, with roughly half the generator parameters.
- Our model performs high-fidelity, identity-preserving face attribute modification at 512×512 resolution. Additionally, for a given face image and target attribute, our model can visualize the feature modifications, providing insight into salient features for an attribute.
- We use our model to modify facial features driving subjective impressions of 10 social traits, achieving $81.3 \pm 12.1\%$ accuracy with human raters.

1.1 Related Work

1.1.1 Generative Adversarial Networks

Generative Adversarial Networks [GPAM+14] use an adversarial learning process to synthesize realistic images from random noise input. Conditional GANs [MO14] add a conditioning input to control the properties of generated images. WGAN variants [ACB17, GAA+17] use Wasserstein distance to improve adversarial training stability. Recent work has shown that GANs can generate realistic images at over 1024×1024 resolution [WLZ+18, KALL17, KLA19].

1.1.2 Face Attribute Modification

GANs have been widely used for image-to-image translation according to domain labels or specific image attributes. This can be achieved in a supervised setting by training with paired input/output image samples [IZZE17, WLZ+18, PLWZ19], but such paired samples are scarce. GANs can use labeled but unpaired images to learn image-to-image translations between domains [ZPIE17, CCK+18, HLBK18]. GANs with an encoder-decoder architecture can modify face images according to discrete facial attributes, with the best architectures using skip connections [HZK+19, LDX+19]. Continuous valued Facial Action Units [FE78] have been modified by GANimation [PAM+18] using a spatial attention based generator. While paired image translation has been done at high resolution [WLZ+18, PLWZ19], modifying unpaired faces at high resolution remains a challenge. With paired images, a loss using pixels [IZZE17] or features [WLZ+18] provides supervision for both realism and target attribute generation. In our experiments with unpaired images, we observe that the losses for realism and attribute generation compete with each other when optimized together in the generator. At higher resolutions, it's easier to discriminate real and fake images, causing the model to favor input image reconstruction over attribute generation.

1.1.3 Modifying the First Impressions Evoked by Faces

Upon seeing a face, people quickly form impressions of their social traits such as kindness, attractiveness, trustworthiness; these perceptions have profound influences on subsequent social judgments and behavior [TODMS15]. Although psychologists have explored how first impressions of faces vary across cultures, [OT08, TODMS15, SLZ+18], attempts to explain what facial features affect social perceptions have been relatively ad hoc. We know of only one paper that attempts to modify faces to alter specific social impressions [Ata19], but it uses an auto-encoder, resulting in very low resolution images that limit its usefulness and interpretability.

Acknowledgment Chapter 1, in part, has been submitted for publication of the material. Yadav, Devendra Pratap; Hu, Weifeng; Song, Amanda; Vul, Edward; Cottrell, Garrison. "Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification". The thesis author was the primary investigator and author of this paper.

Chapter 2

Datasets

2.1 Social Traits Dataset

Existing face image datasets such as CelebA [LLWT15] or EmotionNet [FBQSM16] contain labels for discrete face attributes (such as Beard, Eyeglasses, Smiling) or facial Action Units for which the salient facial features are known. We collected a new dataset to understand how facial features affect the perception of complex social traits from face images. Based on the key dimensions of facial impressions identified by [TODMS15, SLZ$^+$18], we used 10 social traits grouped into three categories: (1) Warmth traits: trustworthy, humble, kind; (2) Physical Appearance traits: attractive, masculine, healthy; and (3) Capability traits: intelligent, powerful, responsible, successful. We select 1611 Caucasian faces from the US 10K Adult Database [BIO13] and obtain ratings on all 10 social traits for each face using Amazon Mechanical Turk. Each social trait rating is a continuous value in $[1.0, 9.0]$. Note that these ratings reflect raters' perception of these traits from images, rather than the unknown latent traits of the people pictured in the images. We refer to this dataset as "Social Traits Dataset" in our work.

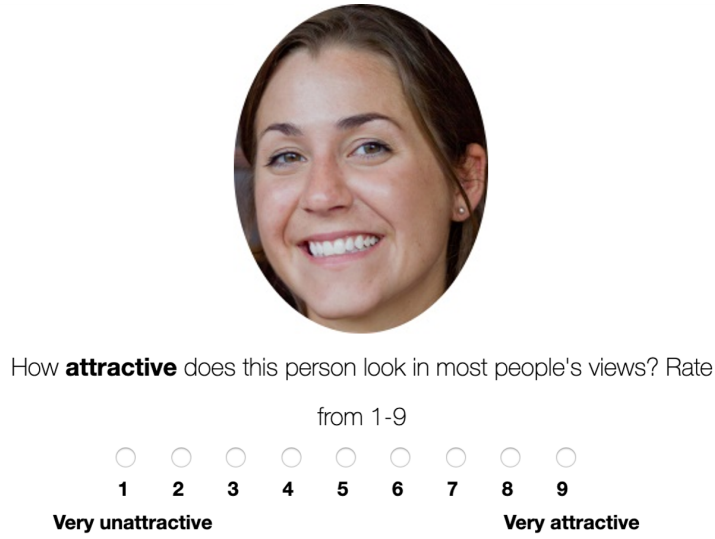

Figure 2.1: The task page shown to human raters on Amazon Mechanical Turk to collect ratings for the Social Traits Dataset.

2.1.1 Data Collection Procedure

We selected 1611 Caucasian faces from the US 10K Adult Database [BIO13], where faces are cropped to an oval shape, limiting the influence of image background when rating faces on social traits. In our online Amazon Mechanical Turk experiment, participants are asked to indicate their first impression of an image on a specific social trait by providing an integer rating on a scale of 1-9. An illustration of the rating task is shown in Figure 2.1. We then asked raters what they think others would rate the face. Using our unpublished data, we found that this reduces social desirability biases when offering potentially contentious opinions. Participants saw 100 faces in a sequence, and rated one face at a time for a specific social trait.

We recruit 428 Caucasian-American participants (254 female) using Amazon Mechanical Turk. The participants' median age range is 30-39 years old. Since rating social traits is a subjective task, we designed a screening mechanism to ensure participants were paying attention to the task. The screening consisted of 20 randomly selected unique faces and a randomly-selected social trait to rate. The 20 faces were presented, then they were shuffled and shown again, resulting in a 40-trial sequence. If a participant's reliability was significantly above zero

(as measured by Spearman's rank correlation of test/retest ratings), and they used at least three different scores from the 9 point scale, the participant passed the "reliability test". A reliable rater is shown 100 new faces and asked to rate them on a social trait. Unreliable raters are not used to collect ratings for our dataset. For our dataset with 1611 face images, we collected at least ten ratings per image-trait combination, which are averaged to get a continuous rating in [1.0,9.0] for each image-trait pair.

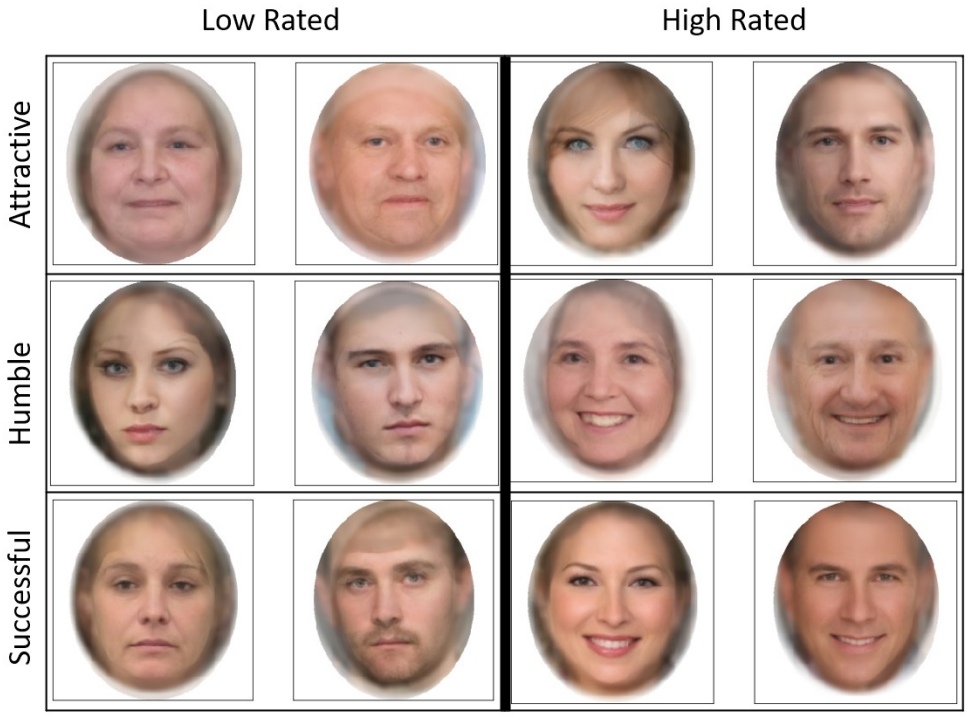

Figure 2.2: Morphed faces representing the 5 highest and 5 lowest rated faces per gender for selected social traits.

Using our dataset of perceived social trait ratings, we visualize an average high or low rated face for selected social traits in Figure 2.2. We average 5 faces to generate a morphed face as follows: 1) Detect 68 landmarks for all faces using Dlib [Kin09]. 2) Average the landmarks across all faces to get a 68 average landmarks. 3) Use Delaunay triangulation and affine transformation to morph each face to the 68 average landmarks. 4) Average all morphed faces to get the final output face image. We can observe salient facial features such as smile for humble and successful.

We also observe an effect of age and wrinkles for attractive, although these are averaged out when morphing faces. Later in our study of perceived social traits, we present a method to better visualize salient facial features for each social trait using a GAN.

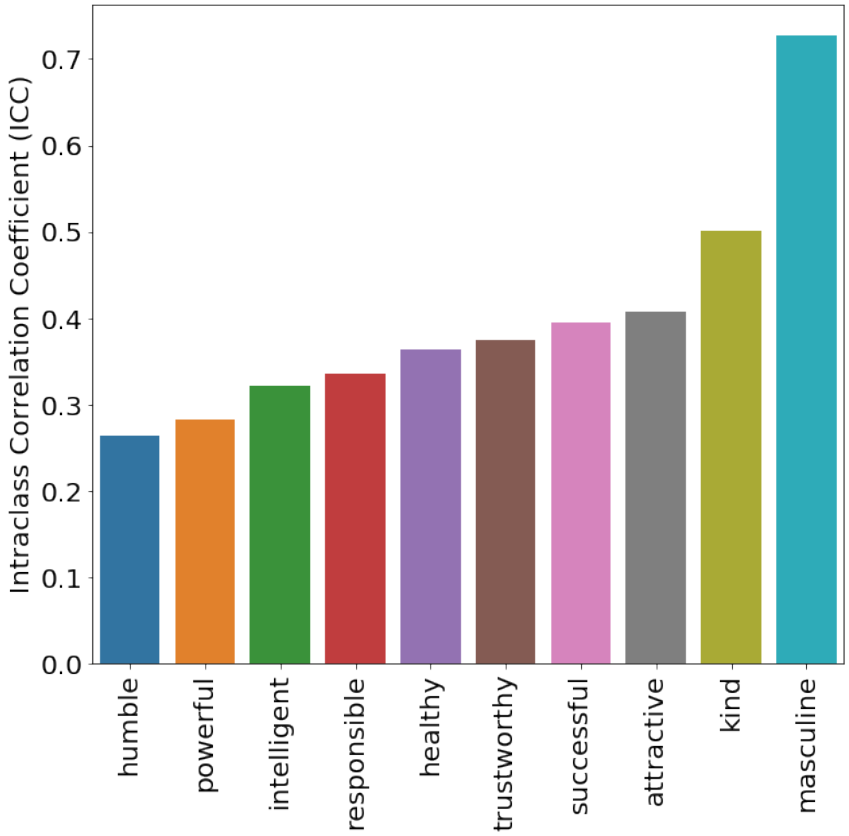

Figure 2.3: Intraclass Correlation Coefficient (ICC) for each trait in the Social Traits Dataset.

2.1.2 Rater Consistency for Social Traits

Due to the subjective nature of perceived social traits, rater may disagree with each other when rating a given face. We analyze the consistency among the raters in our dataset for each social trait. We used a one-way Intraclass Correlation Coefficient (ICC) [Bar66] to measure the agreement level by evaluating the ratio of the variance of item random effects to the overall rating variance. Figure 2.3 shows the ICCs of each social trait. Similar to previous research [HSFS17], we found that there is more agreement for traits representing warmth and

appearance-based appraisals (e.g., kind, attractive and masculine), than for competence-related traits (such as responsible, intelligent, and powerful). This is not surprising as attractiveness, youth, and propensity to smile are much more evident in a face image than traits like responsible or intelligent.

2.2 Attribute Prediction Model: PredCNN

We train a deep convolutional neural network model (referred to as PredCNN) to predict attributes for face images. PredCNN is used to augment the social trait ratings dataset, and evaluate attribute generation accuracy for images modified using the GAN. PredCNN uses ResNet-50 [HZRS16] pre-trained on the ImageNet classification task to extract image features, and fine-tunes it during training. A 512-dimensional feature output from ResNet-50 is used with 256-dimensional fully connected layers to predict attribute labels. We use the same architecture for both discrete and continuous attributes, using sigmoid activation with binary cross entropy loss and mean squared error loss respectively. The architecture of the model is shown in Figure 2.4. The model is trained using stochastic gradient descent, with a learning rate of 0.001 for 40 epochs. We use a train/val/test split of 0.9/0.05/0.05 for all samples in a given training dataset, using the validation set to select the best performing model.

2.3 Celeba-HQ-60k

CelebA-HQ [KALL17] is a dataset of 30k face images at 1024×1024 resolution. While CelebA contains image of size 218×178, CelebA-HQ uses the online image sources for CelebA dataset and refines them to obtain 30k images. They perform image denoising, artifact removal and super-resolution to obtain high quality faces. We use a similar methodology as CelebA-HQ and obtain an additional 30k faces at a smaller 512×512 resolution, annotated with the same

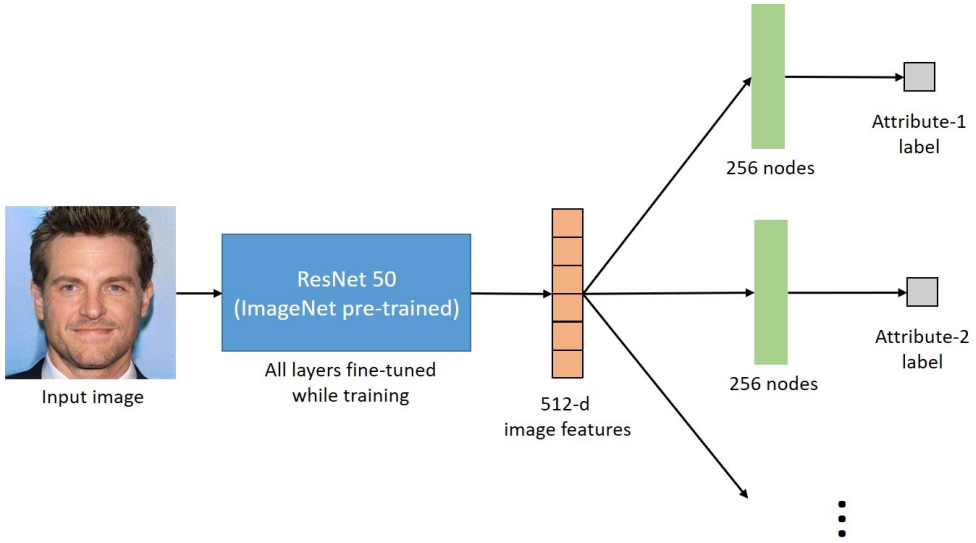

Figure 2.4: Architecture of our attribute prediction model: PredCNN, used to label CelebA-HQ-60k and evaluate accuracy of GAN-based attribute modifications.

40 face attributes as CelebA-HQ. CelebA provides the original uncropped source images used to create the dataset, which range from 100 to 4000 pixels for the shorter side. We perform face detection on all source images and select the ones with face width greater than 256 pixels. Note that we do not consider the 30k face images already present in CelebA-HQ dataset. The faces are cropped, aligned, denoised and resized to a size of 512x512. For faces smaller than 512x512, we use a super-resolution GAN based on ESRGAN [WYW$^+$18], which is trained for the super-resolution task on the CelebA-HQ dataset. Finally, we use a deep image quality assessment network [BMM$^+$17] to evaluate and select the top 30,000 images. Combining these new 30k images with the existing 30k images in CelebA-HQ (resized to 512×512), we get CelebA-HQ-60k containing 60k images at 512×512 resolution.

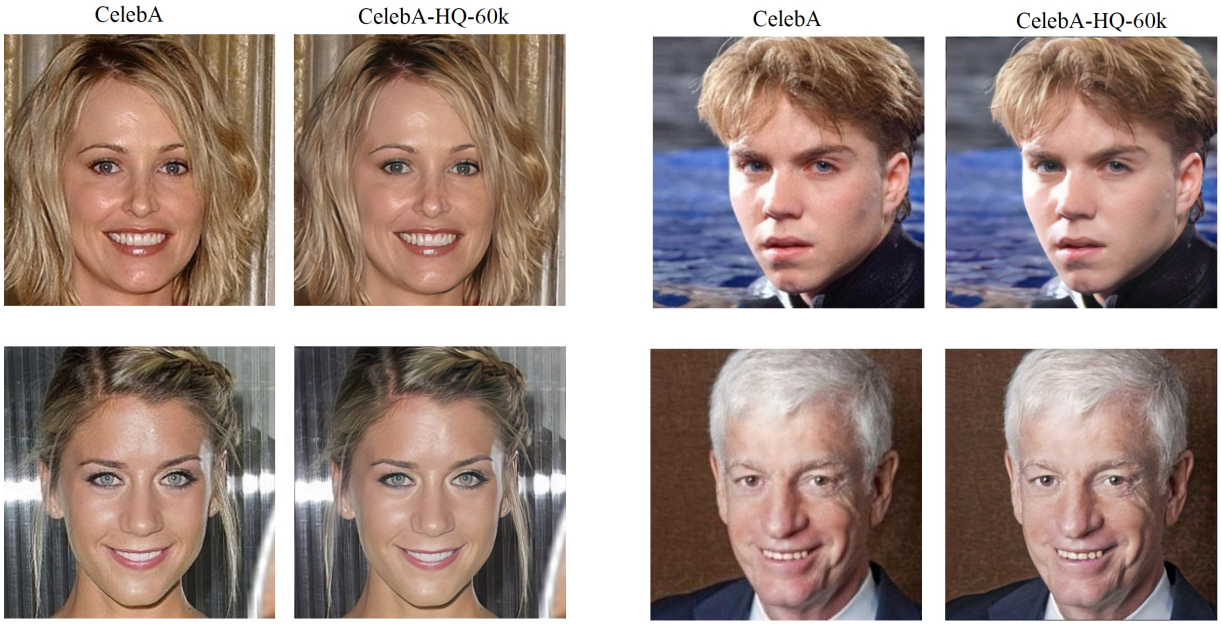

Figure 2.5: CelebA-HQ-60k samples obtained after denoising and GAN based super-resolution. Zoom in for details.

2.3.1 Generating Additional Attribute Labels

Continuous CelebA Attribute Labels

While CelebA-HQ-60k contains binary attribute labels, we aim to convert them to probability based-continuous labels to evaluate GAN based attribute modification methods. Our experiments show that these continuous labels improve face attribute modification performance. We train PredCNN from Section 2.2 on 8 selected binary attributes in the CelebA-HQ-60k dataset: *Bald, Bangs, Black_Hair, Blond_Hair, Eyeglasses, Heavy_Makeup, Male, No_Beard, Smiling, Young*. The model achieves a classification accuracy of 92.6% on the test set. For each attribute, the predicted probability is linearly transformed to get a continuous label in $[-1.0, 1.0]$, where -1.0 denotes absence of the face attribute.

Continuous Facial Action Unit Labels

We label the CelebA-HQ-60k faces for 17 facial Action Unit (AU) attributes, which denote specific facial muscle movements based on the Facial Action Coding System [FE78]. We consider the following Action Units: {*AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45*}. We use OpenFace 2.0 [BZLM18] to predict a continuous AU intensity in [0.0, 5.0] as well as a binary label denoting presence of AU. The absolute AU intensity values in [0,5] are linearly transformed to relative values in [-1.0, 1.0].

Perceived Social Trait Labels

We train PredCNN on the Social Traits Dataset described in Section 2.1. This model achieves Root Mean Square Error (RMSE) $= 0.78$ on the test set, where the ratings are in [1.0,9.0]. We use the model to annotate the images in CelebA-HQ-60k with social traits. The predicted ratings in $[1.0, 9.0]$ are z-scored and linearly transformed to lie in $[-1.0, 1.0]$. We z-score the ratings so that all social traits have losses on a similar scale while training our models.

Our final CelebA-HQ-60k dataset contains 60k faces at 512×512 resolution, labeled on three attribute categories: (1) 8 CelebA face attributes, (2) 17 facial Action Units and, (3) 10 perceived social traits.

Acknowledgment Chapter 2, in part, has been submitted for publication of the material. Yadav, Devendra Pratap; Hu, Weifeng; Song, Amanda; Vul, Edward; Cottrell, Garrison. "Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification". The thesis author was the primary investigator and author of this paper.

Chapter 3

GSC-GAN: Gated Skip Connection Generative Adversarial Network

Generative Adversarial Networks (GANs) with an encoder-decoder architecture have been used to modify face attributes in [HZK+19], [LDX+19] and [PAM+18]. The generator takes an input image and target attribute values to create a modified image. An encoder-decoder architecture reduces the spatial resolution of an image to extract a high-level feature encoding, which can lead to loss of fine image details. To retain these details, AttGAN [HZK+19] employs a skip connection in the decoder. While this improves image quality, [LDX+19] shows that adding more skip connections decreases the attribute modification accuracy significantly. STGAN [LDX+19] proposes Selective Transfer Units (STUs), inspired by Gated Recurrent Units (GRUs) [CVMBB14], which selectively transfer the encoder features to the decoder layers, preventing the reconstruction of face regions that should be modified. While this improves modification accuracy, it adds an additional set of feature decoding layers for STUs, significantly increasing the number of parameters in the generator (68% increase compared to a corresponding U-net [RFB15]). We aim to make the selection of useful encoder features more efficient, with minimal loss in performance.

We propose a novel gating module inspired by Minimal Gated Units (MGUs) [ZWZZ16] that selects parts of the encoder features before combining them with the decoder features. Starting with a U-net architecture [RFB15], we use Gated Skip Connection (GSC) modules to replace the layers in the generator's decoder. In the following sections we describe the architecture of the proposed GSC module, the generator and the discriminator. We detail the loss functions and the objective used to train the model.

3.1 Problem Formulation

We are given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ labeled with a N dimensional attribute vector $\mathbf{v}_{orig} = (v_1, ..., v_N)^\top$, where v_k represent the value of the k^{th} attribute. Let $\mathbf{v}_{tar}$ be a desired attribute vector. We aim to learn a mapping $\mathcal{F} : (\mathbf{x}, \mathbf{v}_{tar}) \rightarrow \mathbf{y}$ where $\mathbf{y}$ is an image representing the attributes $\mathbf{v}_{tar}$. A supervised method would require tuples of $(\mathbf{x}, \mathbf{v}_{tar}, \mathbf{y})$, but instead, we employ an unsupervised learning approach using tuples $(\mathbf{x}, \mathbf{v}_{orig}, \mathbf{v}_{tar})$. A dataset containing tuples $(\mathbf{x}, \mathbf{v}_{orig})$ is used with randomly sampled $\mathbf{v}_{tar}$ (from the dataset) to learn $\mathcal{F}$. Recent work [LDX+19] on this problem shows that using an attribute difference vector $\mathbf{v}_{diff} = \mathbf{v}_{tar} - \mathbf{v}_{orig}$ improves learning for GAN-based methods. Hence, we reformulate the mapping as $\mathcal{F} : (\mathbf{x}, \mathbf{v}_{diff}) \rightarrow \mathbf{y}$.

3.2 Gated Skip Connection (GSC) Module

Consider a GAN generator with a U-net architecture, where the decoder layers are replaced with a module. We have N encoder layers $\{\mathbf{E}_1, ..., \mathbf{E}_N\}$ and N decoder modules $\{\mathbf{M}_N, ..., \mathbf{M}_1\}$, where $\mathbf{M}_1$ produces the final image output. Now we define the operations for the k^{th} decoder module (or Gated Skip Connection module) in the generator. Let $\mathbf{v}_{diff}$ be a vector representing the target attribute modifications. Let $\mathbf{f}_{enc}^{k-1}$ denote the features from the $k-1^{th}$ encoder layer, and $\mathbf{f}_{dec}^{k+1}$ denote the output features from the $k+1^{th}$ decoder module. The k^{th} decoder module uses

$\mathbf{f}_{enc}^{k-1}$, $\mathbf{f}_{dec}^{k+1}$ and $\mathbf{v}_{diff}$ as input to produce output features $\mathbf{f}_{dec}^{k}$ as,

$$\hat{\mathbf{f}}_1 = \mathbf{W}_1 *_T \mathbf{f}_{dec}^{k+1}, \tag{3.1}$$

$$\mathbf{g} = \sigma(\mathbf{W}_2 * [\ \mathbf{f}_{enc}^{k-1},\ \hat{\mathbf{f}}_1,\ \mathbf{v}_{diff}\ ]), \tag{3.2}$$

$$\hat{\mathbf{f}}_2 = \mathbf{g} \circ \hat{\mathbf{f}}_1 + (1 - \mathbf{g}) \circ \mathbf{f}_{enc}^{k-1}, \tag{3.3}$$

$$\mathbf{f}_{dec}^{k} = tanh(\mathbf{W}_3 * [\ \hat{\mathbf{f}}_1,\ \hat{\mathbf{f}}_2,\ \mathbf{v}_{diff}\ ]), \tag{3.4}$$

where $[\cdot,\cdot]$ denotes concatenation, $*_T$ denotes transposed convolution, $*$ denotes convolution, $\sigma(\cdot)$ denotes sigmoid activation and $\circ$ denotes element-wise multiplication. $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ denote learned weight matrices. $\mathbf{v}_{diff}$ is replicated spatially before concatenation.

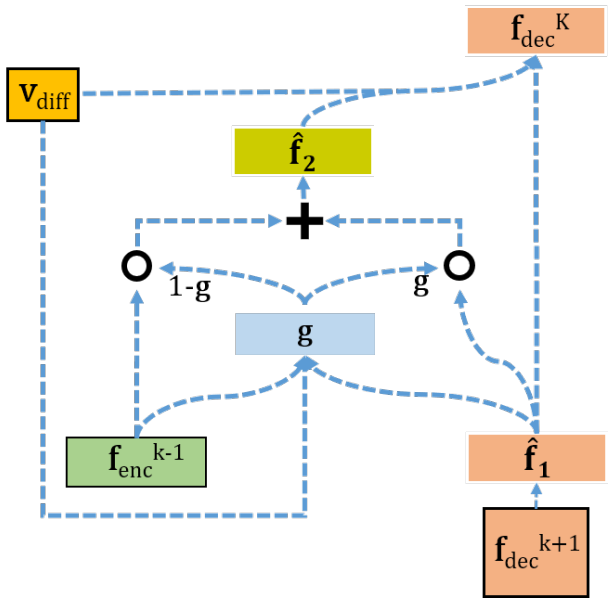

Figure 3.1: Architecture of our Gated Skip Connection (GSC) module. $\circ$ denotes element-wise product.

We show the architecture of the Gated Skip Connection (GSC) module in Figure 3.1. The previous decoder features $\mathbf{f}_{dec}^{k+1}$ are upsampled using a transposed convolution to get $\hat{\mathbf{f}}_1$,which has the same shape as the encoder features $\mathbf{f}_{enc}^{k-1}$. $\mathbf{v}_{diff}$ and two feature maps are used to predict a gating

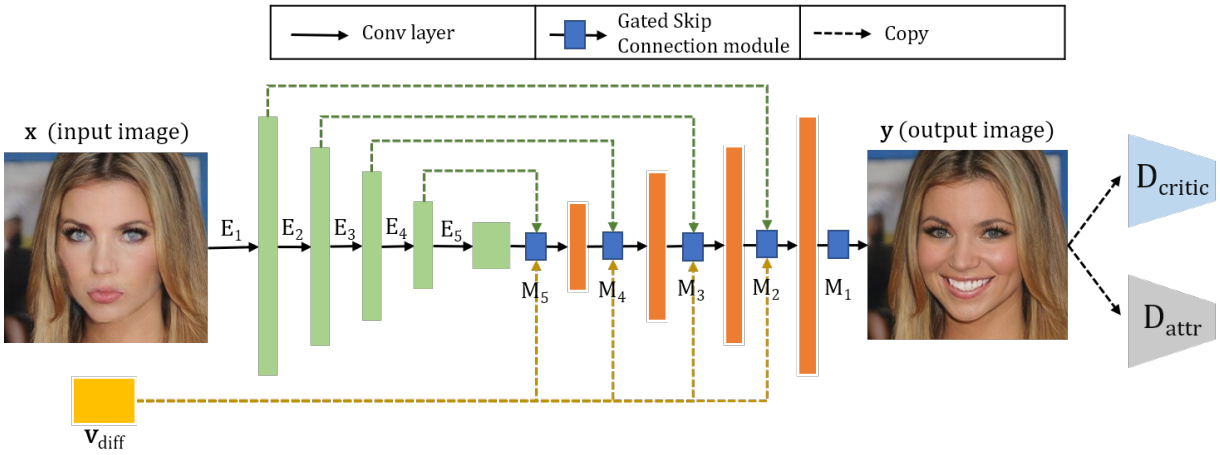

Figure 3.2: Architecture of GSC-GAN, showing the encoder-decoder architecture using GSC modules.

mask $\mathbf{g}$ through a convolutional layer. We use a sigmoid activation to generate values in [0,1] for the gating mask. The gating mask $\mathbf{g}$ performs an element-wise linear combination/selection of features from the decoder ($\hat{\mathbf{f}}_1$) and the skip connections ($\mathbf{f}_{enc}^{k-1}$) to yield $\hat{\mathbf{f}}_2$. The selected encoder features can be used to improve the modified image features as well as reconstruct parts of the input image.

Finally, we allow the module to combine $\hat{\mathbf{f}}_2$ with the upsampled decoder features $\hat{\mathbf{f}}_1$ to produce the output decoder features $\mathbf{f}_{dec}^{k}$. This step has a similar function as the update gate in GRUs, where we can choose to use the information from the skip connection-based features $\hat{\mathbf{f}}_2$. While GRUs and MGUs use gates to forget and update the hidden state, our module instead forgets parts of the input ($\mathbf{f}_{enc}^{k-1}$), and our ablation studies show that this change improves performance over GRU based gating. Since the GSC module uses a single gate, compared to two gates in STGAN's STU [LDX+19], it's a parameter efficient feature gating method for skip connections. We provide $\mathbf{v}_{diff}$ as input to guide feature combination, and we find that removing concatenation of $\mathbf{v}_{diff}$ in Equation 3.4 reduces modification accuracy (2% decrease for Action Unit attributes).

3.3 GSC-GAN Architecture

GSC-GAN consists of an encoder-decoder based generator G and a discriminator D. The generator's encoder consists of N convolutional layers $\{\mathbf{E}_1, ..., \mathbf{E}_N\}$, with corresponding outputs $\{\mathbf{f}^1_{enc}, ..., \mathbf{f}^N_{enc}\}$. The generator's decoder consists of N decoder (GSC) modules $\{\mathbf{M}_N, ..., \mathbf{M}_1\}$ with corresponding outputs $\{\mathbf{f}^N_{dec}, ..., \mathbf{f}^1_{dec}\}$. Figure 3.2 shows the architecture of our model's generator with $N = 5$. The GSC module $\mathbf{M}_k$ computes its output ($\mathbf{f}^k_{dec}$) by combining the previous decoder representation ($\mathbf{f}^{k+1}_{dec}$) with the corresponding encoder representation ($\mathbf{f}^{k-1}_{enc}$). At the first decoder module $\mathbf{M}_N$, we assign $\mathbf{f}^{N+1}_{dec} = \mathbf{f}^N_{enc}$, and at the last decoder module $\mathbf{M}_1$ (where input $\mathbf{f}^0_{enc}$ is unavailable), we simply use a transposed convolution of the decoder feature $\mathbf{f}^2_{dec}$ followed by a convolutional layer to produce $\mathbf{f}^1_{dec}$ – the modified image output. $\mathbf{E}_1$ extracts features with C_{init} channels from the input, and the number of channels is doubled by every layer up to a maximum of 1024. Specifically, an encoder layer doubles the number of channels only if it is less than 1024, otherwise it keeps 1024 channels in the output features.

The discriminator D consists of two components: the adversarial critic D_{critic} and the attribute predictor D_{attr}. Both components share a series of convolutional layers, denoted by D_{conv}, to extract image features. The image features are passed to two Fully Connected (FC) layers to predict an output. D_{attr} and D_{critic} do not share the FC layers. D_{attr} predicts attribute labels and D_{critic} predicts a real value which is used to distinguish real and generated samples. The number of channels in D_{conv} are set in a manner similar to G's encoder, starting with C_{init} and doubled by every layer upto a maximum of 1024.

We present the different components used in the generator and discriminator architecture for GSC-GAN in Table 3.1. Table 3.2 uses these components to describe the architecture of generator and discriminator. Note that we show the architecture for 128 and 512 image size. The architecture used for 256 image size is obtained by using generator from the 128 size model, and discriminator from the 512 size model.

Table 3.1: Specification of different components used in GSC-GAN's generator and discriminator.

Component Name	**Specification**
BN/IN	Batch/Instance Normalization
ReLU / LReLU	ReLU and Leaky ReLU activations
TrConv/Conv (c, k, s)	Transposed Convolutional / Convolutional layer with c output channels, kernel size k and stride s
FC (d)	Fully connected layer with d nodes
$\mathbf{E}_i(c)$	Conv(c , 4, 2) , BN, LReLU
$\mathbf{M}_i(c), i > 1$	W1 = TrConv(c , 4, 2)
	W2 = Conv(c , 3, 1), BN, Sigmoid
	W3 = Conv(c , 3, 1), BN, tanh
$\mathbf{M}_1(c)$	TrConv(c , 4, 2) , BN, ReLU, Conv(3 , 3, 1), tanh
$\mathbf{D}_{conv_i}(c)$	Conv(c , 4, 2) , IN, LReLU

3.4 Loss Functions and Objective

Following our problem formulation in Section 3.1, for a given input image $\mathbf{x}$, let $\mathbf{v}_{orig}$ denote the true attribute labels. We wish to modify $\mathbf{x}$ according to target attributes $\mathbf{v}_{tar}$. We sample $\mathbf{v}_{tar}$ at random from the dataset, thus ensuring that we respect the correlation structure of dataset labels in the target vectors. The attribute difference vector $\mathbf{v}_{diff} = \mathbf{v}_{tar} - \mathbf{v}_{orig}$ is used to inform the generator about the magnitude and direction of attribute modifications. The generator outputs a modified image $\mathbf{y} = G(\mathbf{x}, \mathbf{v}_{diff})$. Now we describe the losses used to train GSC-GAN.

3.4.1 Adversarial Loss

We train the generator G to output realistic images using the adversarial learning formulation proposed by Wasserstein GAN [ACB17] along with the improvements proposed by WGAN-GP [GAA+17]. The adversarial loss for the generator ($\mathcal{L}_{G_{adv}}$) and the discriminator ($\mathcal{L}_{D_{adv}}$) are,

$$\min_G \mathcal{L}_{G_{adv}} = -\mathbb{E}_{\mathbf{x},\mathbf{v}_{diff}} D_{critic}(G(\mathbf{x}, \mathbf{v}_{diff})), \tag{3.5}$$

$$\min_{D_{critic}} \mathcal{L}_{D_{adv}} = \mathbb{E}_{\mathbf{y}} D_{critic}(\mathbf{y}) - \mathbb{E}_{\mathbf{x}} D_{critic}(\mathbf{x}) + \lambda_{gp} \mathbb{E}_{\tilde{\mathbf{x}}}[(\|\nabla_{\tilde{\mathbf{x}}} D_{critic}(\tilde{\mathbf{x}})\| - 1)^2], \tag{3.6}$$

Table 3.2: Architecture of GSC-GAN for 128 and 512 image size. See Table 3.1 for specification of components. N_a denotes number of attributes in the dataset. A column represents the components in a sequential order.

Image size	Generator		Discriminator	
	Encoder	**Decoder**	$\mathbf{D}_{critic}$	$\mathbf{D}_{attr}$
128	$\mathbf{E}_1(64)$	$\mathbf{M}_5(512)$	$\mathbf{D}_{conv_1}(64)$	
	$\mathbf{E}_2(128)$	$\mathbf{M}_4(256)$	$\mathbf{D}_{conv_2}(128)$	
	$\mathbf{E}_3(256)$	$\mathbf{M}_3(128)$	$\mathbf{D}_{conv_3}(256)$	
	$\mathbf{E}_4(512)$	$\mathbf{M}_2(64)$	$\mathbf{D}_{conv_4}(512)$	
	$\mathbf{E}_5(1024)$	$\mathbf{M}_1(32)$	$\mathbf{D}_{conv_5}(1024)$	
			FC(1024), LReLU	FC(1024), LReLU
			FC(1)	FC(N_a)
512	$\mathbf{E}_1(32)$	$\mathbf{M}_6(512)$	$\mathbf{D}_{conv_1}(32)$	
	$\mathbf{E}_2(64)$	$\mathbf{M}_5(256)$	$\mathbf{D}_{conv_2}(64)$	
	$\mathbf{E}_3(128)$	$\mathbf{M}_4(128)$	$\mathbf{D}_{conv_3}(128)$	
	$\mathbf{E}_4(256)$	$\mathbf{M}_3(64)$	$\mathbf{D}_{conv_4}(256)$	
	$\mathbf{E}_5(512)$	$\mathbf{M}_2(32)$	$\mathbf{D}_{conv_5}(512)$	
	$\mathbf{E}_6(1024)$	$\mathbf{M}_1(16)$	$\mathbf{D}_{conv_6}(1024)$	
			FC(512), LReLU	FC(512), LReLU
			FC(1)	FC(N_a)

$\mathcal{L}_{G_{adv}}$ guides the generator G to output realistic images. $\mathcal{L}_{D_{adv}}$ guides the discriminator network D_{critic} to distinguish between real and fake images. The gradient penalty [GAA+17], having a coefficient λ_{gp}, is calculated using samples $\tilde{\mathbf{x}}$ obtained by random interpolation between real and fake image pairs. Following WGAN-GP [GAA+17], we set $\lambda_{gp} = 10$ and use a 5:1 update rate for D and G respectively.

3.4.2 Attribute Modification Loss

While we don't have the ground truth output image, we use the known target attributes to constrain the output image. The regression loss $\mathcal{L}_{D_{reg}}$ trains D_{critic} to accurately predict $\mathbf{v}_{orig}$ given $\mathbf{x}$. Additionally, we want the generated image $\mathbf{y}$ to represent the target attribute values $\mathbf{v}_{tar}$. This loss is denoted by $\mathcal{L}_{G_{reg}}$,

$$\min_{D_{attr}} \mathcal{L}_{D_{reg}} = \mathbb{E}_{\mathbf{x},\mathbf{v}_{orig}} \left\| \mathbf{v}_{orig} - D_{attr}(\mathbf{x}) \right\|_2^2, \tag{3.7}$$

$$\min_{G} \mathcal{L}_{G_{reg}} = \mathbb{E}_{\mathbf{y},\mathbf{v}_{tar}} \left\| \mathbf{v}_{tar} - D_{attr}(\mathbf{y}) \right\|_2^2, \tag{3.8}$$

3.4.3 Reconstruction loss

While modifying images, we require the generator to reconstruct parts of the image which are not relevant to the target modifications. Hence, if $\mathbf{v}_{diff} = \mathbf{v}_{zero} = \vec{0}$ we want the generated image $\mathbf{y} = \mathbf{x}$. This is enforced using ℓ_1 loss which retains fine details better than ℓ_2 loss. The reconstruction loss $\mathcal{L}_{rec}$ is given by,

$$\min_{G} \mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}} \left\| \mathbf{x} - G(\mathbf{x}, \mathbf{v}_{zero}) \right\|_1, \tag{3.9}$$

3.4.4 Final Objective

The final objective combines the aforementioned losses and alternatively optimizes the generator and discriminator. The discriminator loss $\mathcal{L}_{dis}$ and generator loss $\mathcal{L}_{gen}$, which use hyperparameters $\lambda_{D_{reg}}$, $\lambda_{G_{reg}}$ and λ_{rec} to balance different loss components, are given by,

$$\min_{G} \mathcal{L}_{dis} = \mathcal{L}_{D_{adv}} + \lambda_{D_{reg}} \mathcal{L}_{D_{reg}}, \tag{3.10}$$

$$\min_{D} \mathcal{L}_{gen} = \mathcal{L}_{G_{adv}} + \lambda_{G_{reg}} \mathcal{L}_{G_{reg}} + \lambda_{rec} \mathcal{L}_{rec}. \tag{3.11}$$

Acknowledgment Chapter 3, in part, has been submitted for publication of the material. Yadav, Devendra Pratap; Hu, Weifeng; Song, Amanda; Vul, Edward; Cottrell, Garrison. "Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification". The thesis author was the primary investigator and author of this paper.

Chapter 4

Experiments

4.1 Training Procedure

We train our GSC-GAN model on the CelebA-HQ-60k dataset for the three types of attributes it is labeled on. We use $\lambda_{D_{reg}} = 1$, $\lambda_{G_{reg}} = 10$, $\lambda_{rec} = 40$ for all models, except $\lambda_{G_{reg}} = 15$ for CelebA attributes. Our experiments use a train/test split of 59k/1k images on the CelebA-HQ-60k dataset.

The model architecture is varied based on the training image size as shown in Table 3.2. For 128×128 image size, we use 5 layers in generator's encoder, 5 GSC modules in the generator's decoder, and 5 convolutional layers in D_{conv}, with $C_{init} = 64$. At 256×256, our default architecture increases layers in D_{conv} to 6, and reduces nodes in fully connected layers for D_{critic} and D_{attr} to 512. Note that for a fair comparison with STGAN at 256×256 resolution, we instead use the same $C_{init} = 48$ and D_{conv} depth (5 layers) as STGAN's architecture in our Action Unit attribute experiments. At 512×512, we further set $C_{init} = 32$ and increase encoder layers and GSC modules in G to 6. All models are trained using the ADAM [KB14] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We use a learning rate of 10^{-4} and train for 100 epochs with a batch size of 32 (reduced to 16 for 512 image size). We compare our method with state-

of-the-art models for discrete and continuous face attribute modification: STGAN [LDX+19] and GANimation [PAM+18] respectively. We use the authors' code for GANimation[1] and a Pytorch re-implementation for STGAN. We train both models for 100 epochs using their default hyperparameters and architecture.

4.2 Evaluation Metrics

Attribute Modification Accuracy We evaluate if the GANs modify attributes accurately using a predictive model. We generate images by modifying one attribute at a time, and predict their attribute values. The generated and target attribute values are used to measure accuracy or root mean squared error (RMSE). For discrete CelebA attributes, we use the PredCNN model described in Section 2.2 to predict binary labels. For continuous facial Action Unit (AU) attributes, we use OpenFace 2.0 [BZLM18] to predict a continuous AU intensity as well as a binary label which denotes presence/absence of an Action Unit.

Fréchet Inception Distance (FID score) Fréchet Inception Distance or (FID score) [HRU+17] is used to evaluate image quality for GAN generated images [KLA19] [KALL17]. This measure embeds the real and generated images into the feature space of a specific layer from Inception Net [SLJ+15]. The embedding layer is assumed to be a continuous multivariate Gaussian. The Fréchet distance or the Wasserstein-2 distance between these two Gaussians (real and fake) is used to calculate the FID score. A lower FID corresponds to more realistic generated images.

Identity Preservation Accuracy While modifying faces, we prefer to preserve the identity of the person. This is desirable for attributes such as facial Action Units or CelebA face attributes. To measure identity preservation, we use a state-of-the-art face recognition model based on SENet-50 [CSX+18, HSS18] trained on the VGGFace2 dataset [CSX+18]. We use the model to

[1] https://github.com/albertpumarola/GANimation

extract features for a face image pair (original and modified image), and use a threshold of 0.4 on the cosine distance between them to get identity preservation accuracy.

4.3 Continuous Attribute Modification

Facial Action Units (AUs) denote independent facial muscle movements based on the Facial Action Coding System [FE78]. These AUs combine to generate various facial expressions. We study the task of modifying these AU attributes for face images in a controllable, continuous manner. We use OpenFace 2.0 [BZLM18] to label 17 AUs for faces in the CelebA-HQ-60k dataset as described in Section 2.3.1. We adapt STGAN for continuous AUs by replacing the binary cross entropy loss using mean squared error loss in the attribute discriminator. While AUs are annotated using continuous intensity values in [0,5], we normalize and linearly transform the values to lie in a relative range of [-1, 1] during training. We note that for a fair comparison with STGAN at 256×256 resolution, we use GSC-GAN with similar model capacity as STGAN's architecture ($C_{init} = 48$ and D_{conv} depth of 5 layers).

4.3.1 Qualitative Evaluation

We train GSC-GAN, STGAN and GANimation at 128 and 256 image sizes. Figure 4.1 shows samples of AU modification at 256 size. Our model and STGAN accurately modify the AUs with constrained changes that preserve the face identity. GANimation generates higher AU intensity but does not preserve identity as well. We observe many artifacts in the images generated by GANimation. The quality of the learned image-level attention map in GANimation greatly affects the modified image regions, which leads it to modify irrelevant face regions.

Since the Action Unit attributes are continuous, we show that our model smoothly interpolates between low and high intensities in Figure 4.2. The model gradually creates modified features which are different from simply fading in two sets of features. We go beyond the -1

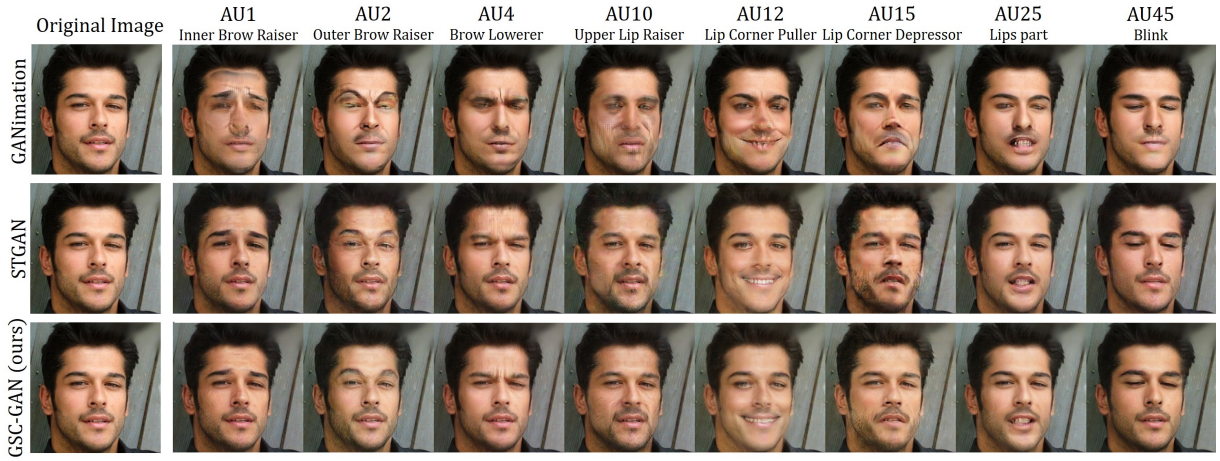

Figure 4.1: Action Unit modification with GANimation, STGAN, and GSC-GAN at 256×256 resolution. The columns show modification to 5.0 absolute AU intensity. Please zoom in for details.

to 1 relative AU values used during training to exaggerate the modified features. GANimation applies a spatial attention map at an image level, while our GSC module combines encoder and decoder features. As shown in Figure 4.3, the image level attention in GANimation doesn't handle occlusions well. STGAN, which also combines features instead of pixels, also handles occlusion better than GANimation. We also train GSC-GAN at 512×512, and observe high fidelity Action Unit modification with excellent identity preservation. We show the results of interpolation for selected Action Units in Appendix A.2.

4.3.2 Quantitative Evaluation

Table 4.1 shows results for quantitative evaluation of all methods using the metrics described in Section 4.2. For each test image, we modify AUs one at a time to 0 and 5 absolute intensity values, generating 34000 images for evaluation. We calculate AU generation accuracy and RMSE for each of the 17 Action Units, and use the 17 values to get the mean and standard deviation. We observe that GANimation achieves better modification accuracy and RMSE but a poor FID score and identity preservation. GSC-GAN and STGAN achieve much better identity

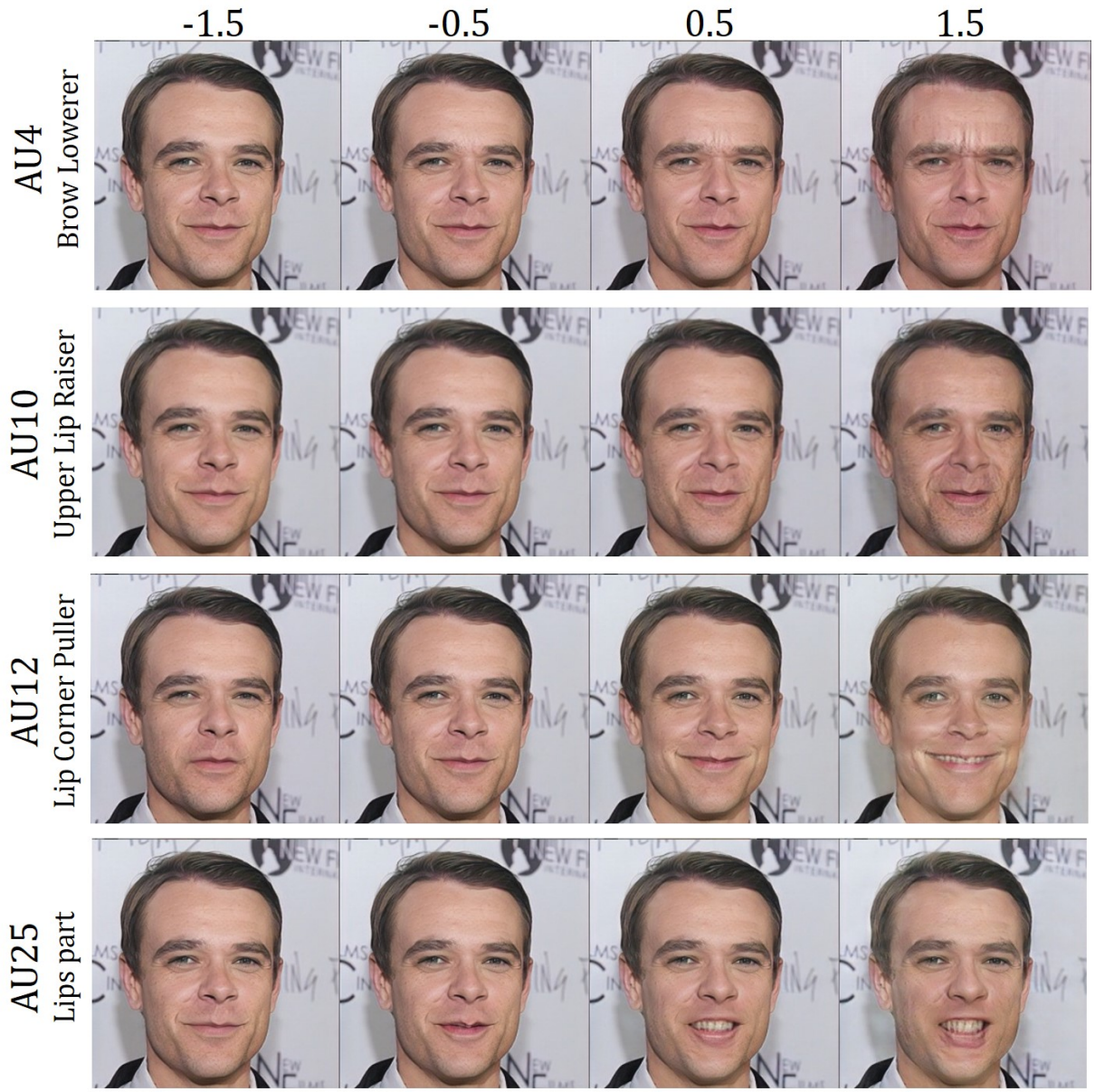

Figure 4.2: Interpolation along relative AU values using GSC-GAN at 256×256 resolution. Relative values of -1 and 1 denote absolute AU intensity of 0 and 5 respectively.

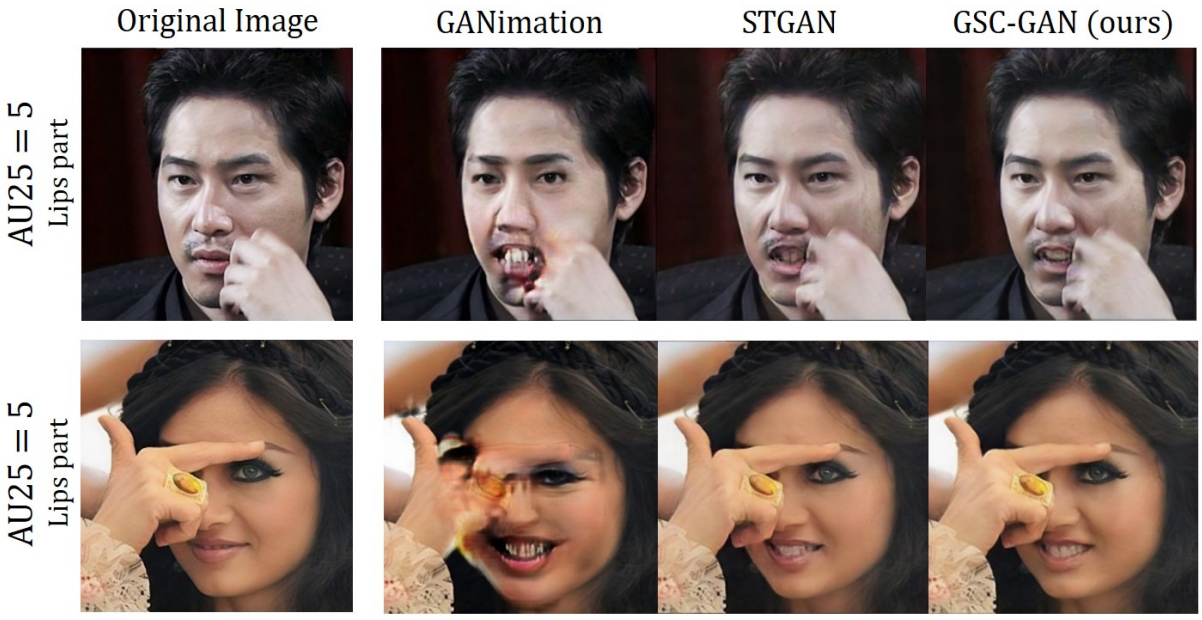

Figure 4.3: AU25 (Lips Part) modification to 5.0 absolute intensity with occluded faces for GANimation, STGAN, and GSC-GAN.

Table 4.1: Evaluation of GSC-GAN (ours), STGAN and GANimation for continuous Action Unit (AU) attributes. Training image size is denoted by subscripts. Note: RMSE uses AU intensity in [0,5].

Method$_{size}$	**AU generation accuracy**$(\%)\uparrow$	**RMSE**$\downarrow$	**FID**$\downarrow$	**Identity preservation accuracy**$(\%)\uparrow$
GANimation$_{128}$	**91.4 ± 4.9**	**1.34 ± 0.51**	17.02	20.6
STGAN$_{128}$	80.0 ± 8.0	2.09 ± 0.33	12.97	96.3
GSC-GAN$_{128}$	80.6 ± 6.7	2.04 ± 0.33	**12.43**	**99.4**
GANimation$_{256}$	**89.2 ± 5.9**	**1.42 ± 0.55**	14.03	48.2
STGAN$_{256}$	81.5 ± 5.9	1.81 ± 0.33	**11.15**	98.6
GSC-GAN$_{256}$	82.1 ± 6.1	1.71 ±0.35	11.41	**99.3**
GSC-GAN$_{512}$	80.4 ± 6.9	1.78 ±0.39	11.81	99.1

preservation and FID scores while achieving acceptable modification accuracy. GSC-GAN has a slight edge over STGAN on all metrics at a 128 image size. We see a similar trend at 256 size, except STGAN achieves better FID. Our results show that GSC-GAN scales to 512×512 resolution as well, with high identity preservation and modification accuracy. While STGAN has similar performance as GSC-GAN, the latter uses fewer parameters.

4.4 Discrete Attribute Modification

We use the CelebA-HQ-60k dataset to evaluate model performance on modifying discrete attributes. We select 8 face attributes from the CelebA dataset: *Bald, Bangs, Eyeglasses, Heavy_Makeup, Male, No_Beard, Smiling, Young*. These attributes are chosen since they modify different regions of the face, and include both local and global face modifications. Since GSC-GAN and GANimation are originally designed for modifying continuous attributes, we convert the binary labels into continuous probability-based labels in $[-1.0, 1.0]$ as described in Section 2.3. In order to utilize these continuous labels with STGAN, we replace the binary cross entropy loss in attribute discriminator with mean square error loss. We also train the original STGAN model with binary attribute labels and binary cross entropy loss, denoted by STGAN-BCE in our discussion, but find that it doesn't perform as well.

4.4.1 Qualitative Evaluation

We use low and high attribute values of -1.0 and 1.0 to generate modified face images at 128×128 image resolution. Figure 4.4 shows the qualitative results for our model, STGAN, STGAN-BCE and GANimation. We observe that both our model and STGAN, which use encoder features though skip connections, preserve identity well, while modifying attributes accurately at the same time. GANimation generated images show slightly higher attribute modification intensity but they contain unrealistic artifacts which result in poor identity preservation. Note

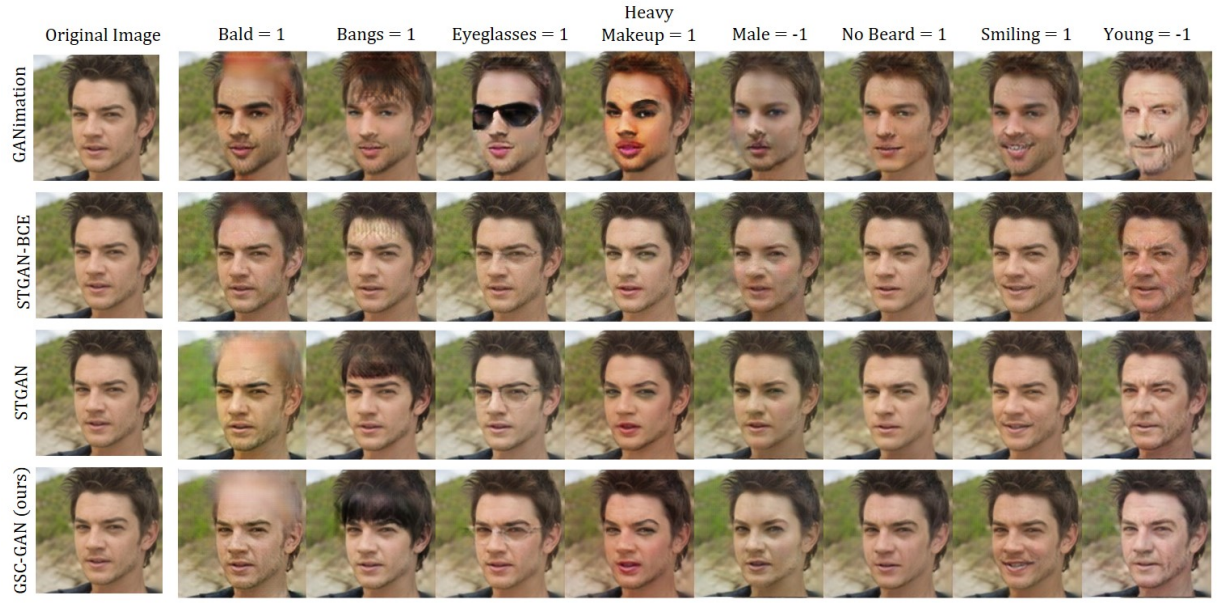

Figure 4.4: CelebA attribute modification results for GSC-GAN, STGAN, STGAN-BCE and GANimation. Each attribute is inverted and column labels show the target value.

that our model doesn't enforce identity preservation during training, but the skip connections implicitly guide the model to use the original image features to generate the output modified image. We observe that STGAN-BCE, which is trained with binary labels and cross entropy loss, generates much lower intensity attribute modifications. We also show that our model scales to 512×512 resolution, while maintaining accurate identity preservation and modification quality in Appendix A.1. We show interpolation along relative attribute values and observe that our model learns a continuous modification for most of the CelebA attributes, despite the original labels being binary.

4.4.2 Quantitative Evaluation

We quantitatively evaluate our model, STGAN, STGAN-BCE and GANimation using the metrics described in Section 4.2. Each model is used to modify CelebA attribute one at a time to the opposite value as its original binary label. We do this for each of the 1000 images in our

Table 4.2: Evaluation of GSC-GAN (ours), STGAN, STGAN-BCE and GANimation for CelebA discrete attributes. Models are trained at image size as denoted by subscripts.

Method$_{size}$	**Attribute generation accuracy** $(\%) \uparrow$	**FID** $\downarrow$	**Identity preservation accuracy** $(\%) \uparrow$
GANimation$_{128}$	$\mathbf{65.1 \pm 30.6}$	22.90	30.1
STGAN-BCE$_{128}$	52.2 ± 33.9	21.63	97.1
STGAN$_{128}$	60.9 ± 29.4	**18.45**	**99.0**
GSC-GAN$_{128}$	60.3 ± 26.8	19.63	98.8
GSC-GAN$_{512}$	66.6 ± 28.9	18.60	98.8

CelebA-HQ-60k test set, which results in 8000 generated images. We use PredCNN from Section 2.2 trained on CelebA-HQ-60k (test set accuracy of 92%) to predict binary attribute labels for the generated images. An accuracy score is calculated based on the predicted and target attribute values. Note that since we used PredCNN based on ResNet-50 to generate the probability based continuous labels, we use a different ResNet-101 based PredCNN model to evaluate accuracy. We calculate attribute generation accuracy for each of the 8 CelebA attributes, and use the 8 values to get the mean and standard deviation. We also calculate FID score and identity preservation accuracy using the 8000 generated images.

Quantitative evaluation results are shown in Table 4.2. All models obtain good attribute modification accuracy but the quality of the generated samples varies across models. GANimation achieves high accuracy, but does poorly on image quality metrics. STGAN achieves slightly better FID score and identity preservation, compared to GSC-GAN. Our model, using about half the generator parameters as STGAN, achieves similar performance. We observe a high standard deviation for attribute generation accuracy since some attributes are modified accurately (Smiling:98%, Bangs:94%) while others show low modification accuracy (Bald:15%, Male:39%) for GSC-GAN. We observe a similar difference in accuracy for GANimation and STGAN as well. Our model, GSC-GAN, scales to 512×512 resolution, with improved accuracy and FID score. The results show that our GSC modules effectively utilize skip connections to generate high quality face modifications that preserve identity.

4.5 Parameter Efficiency

Our experiments on discrete and continuous attributes show that GSC-GAN provides similar performance to STGAN in both qualitative and quantitative comparisons. Now, we compare the total number of parameters for both models at various resolutions. Our model is efficient in both the generator and the discriminator architecture. In the generator, the efficiency is obtained by maintaining a single set of decoder layers, while STGAN contains two sets of feature decoding layers. In the discriminator, we find that adding additional layers as we increase the input image size, reduces the total parameters. The reduction is observed due to a decrease in the spatial size of the features extracted in D_{conv} before the fully connected layers. For instance, a feature with dimensions $512 \times 16 \times 16$, when flattened and passed to a 512-d fully connected layer, would require 67M parameters. If we add another convolutional layer, doubling the channels, but reducing the feature's spatial dimensions to $1024 \times 8 \times 8$, we only need 33M parameters. Moreover, since we limit the number of channels to 1024, the parameter reduction is larger for deeper networks.

The primary hyperparameters affecting the parameter count in the generator are the depth of the encoder-decoder component, and the feature channels being extracted at each layer. Both STGAN and GSC-GAN extract C_{init} number of channels from the input image and double it every layer in the encoder. For deeper architectures, GSC-GAN limits the number of channels to a maximum of 1024. For the same generator depth, and feature channels C_{init}, GSC-GAN contains only 47% of STGAN's generator parameters.

Table 4.3 shows the parameters for STGAN and GSC-GAN at 128, 256 and 512 image size. At 128×128, both models use same generator depth and C_{init}, hence a direct performance comparison is possible. As shown in our experiments, GSC-GAN achieves equivalent performance with half the generator parameters. At 256×256, we use a GSC-GAN architecture that uses the same C_{init} (48) and D_{conv} depth (5 layers) as STGAN. Hence, we can make a fair

Table 4.3: Number of generator (G) and discriminator (D) parameters for GSC-GAN and STGAN at different image sizes (denoted by subscript).

Method$_{size}$	**G parameters (million)**	**D parameters (million)**	**Total parameters (million)**
STGAN$_{128}$	74.8	44.7	119.5
STGAN$_{256}$	42.2	56.6	98.8
STGAN$_{512}$	42.2	207.6	249.8
GSC-GAN$_{128}$	35.3	44.7	80.0
GSC-GAN$_{256}$	19.8	56.6	76.4
GSC-GAN$_{512}$	35.3	78.3	113.6

parameter and performance comparison. Again, GSC-GAN uses half the generator parameters and achieves equivalent performance as shown in our experiments. Although STGAN authors do not evaluate an architecture for 512×512 resolution, we extend their 384×384 architecture to higher resolution based on their code[2]. At 512×512, STGAN uses 250M parameters versus GSC-GAN's 114M. The reduction in G parameters is smaller as we use 1 layer deeper G than STGAN. A deeper D reduces the total parameters significantly. As shown in our experiments, GSC-GAN achieves good performance at 512×512, while only using only 42% more parameters than a 128×128 resolution model.

4.6 Model Ablations

Our Gated Skip Connection (GSC) module proposes an efficient gating architecture inspired by Minimal Gated Units (MGUs), which uses a single gate. Here, we evaluate if our MGU based gating performs better than a Gated Recurrent Unit (GRU) based gating architecture. STGAN uses a GRU inspired gating architecture called Selective Transfer Unit (STU). STU uses the previous state, encoder features from a skip connection and the target attribute vector to generate two outputs: a state and a feature output. We directly adapt STU's gating operations (as described in STGAN [LDX+19]) into the GSC module, by discarding the state and using

[2]https://github.com/csmliu/STGAN

the feature output. We denote this model, that used GRU based gates, as GSC-GAN-gru. We train GSC-GAN-gru on CelebA attributes and facial Action Unit attributes, using the same hyperparameters as GSC-GAN at a 128×128 resolution.

Table 4.4: Quantitative results of a model ablation study for discrete CelebA attributes at 128×128 resolution. GSC-GAN-gru replaces GSC module with a GRU based gating module.

Method	**Attribute generation accuracy** $(\%) \uparrow$	**FID** $\downarrow$	**Identity preservation accuracy** $(\%) \uparrow$
GSC-GAN-gru	51.2 ± 28.3	**17.79**	98.7
GSC-GAN	$\mathbf{60.3 \pm 26.8}$	19.63	**98.8**

Table 4.5: Quantitative results of a model ablation study for Action Unit attributes at 128×128 resolution. GSC-GAN-gru replaces GSC module with a GRU based gating module.

Method	**Attribute generation accuracy** $(\%) \uparrow$	**RMSE** $\downarrow$	**FID** $\downarrow$	**Identity preservation accuracy** $(\%) \uparrow$
GSC-GAN-gru	75.7 ± 7.8	2.43 ± 0.28	12.46	99.1
GSC-GAN	$\mathbf{80.6 \pm 6.7}$	$\mathbf{2.04 \pm 0.33}$	**12.43**	**99.4**

Table 4.4 and Table 4.5 show the quantitative results for CelebA and Action Unit attributes respectively. We observe that GSC-GAN-gru achieves similar FID and identity preservation. However, the attribute generation accuracy decreases significantly. This ablation study shows that our gating architecture, which uses a gate to forget parts of the input instead of the state, improves attribute generation.

4.7 Visualizing Modified Features Using GSC-GAN

Our previous experiments show that GSC-GAN can generate accurate modifications while reconstructing irrelevant regions of the image using information from the skip connections. Since the decoder modules have access to encoder features at multiple spatial resolutions, its task is simplified to generating the modified features instead of generating features for the entire image. This form of residual feature learning enables us to visualize the modifications made by the model

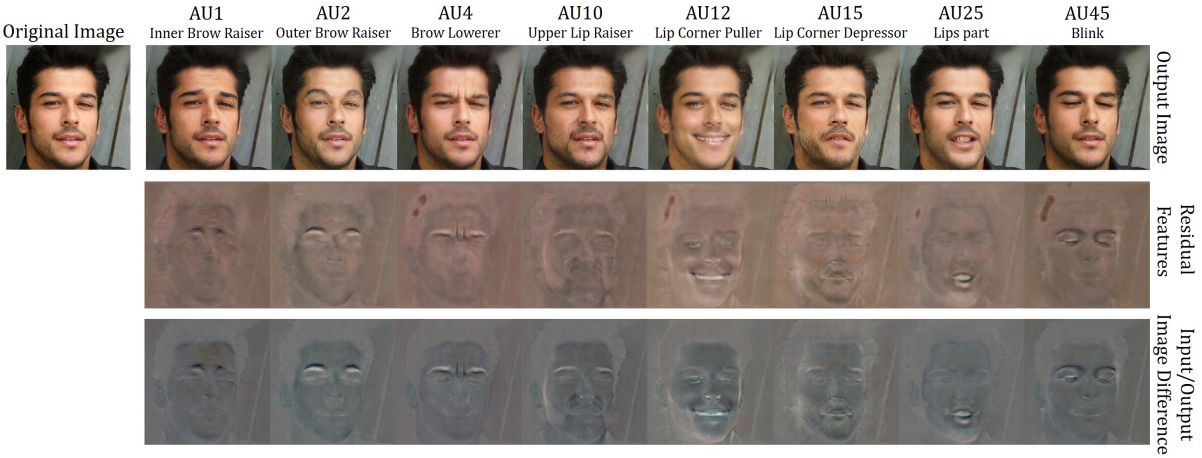

Figure 4.5: A visualization of residual features learned by GSC-GAN for Action Unit attributes. We also show the pixel-wise difference between input and output image for comparison.

by removing certain skip connections at inference time, and generating the output image. For our models, we find that removing the skip connection for decoder module $\mathbf{M}_2$ (we set $\mathbf{f}^1_{enc} = 0$) provides us with an output image that highlights the modified features.

Since our model reconstructs image regions with high quality, we can also use the pixel-wise difference between input and output image to visualize the feature modifications made for a given attribute. However, we need to scale the difference values to visualize them as an image, which can generate color and brightness noise. We compare our method of visualizing residual features with image difference for CelebA and Action Unit attributes. We visualize the residual features learned by the model for Action Unit attributes in Figure 4.5. We show the input image, the output image for each attribute, the corresponding residual features and the input/output image difference. We observe that both visualization methods produce similar images highlighting the localized feature modification for Action Units. This shows that GSC-GAN selectively learns the modified features, and copies encoder features from the skip connection to reconstruct irrelevant image regions. By visualizing the residual features, we add explainability to the model's operations. This method can also help us identify artifacts in generated images, such as the droplet artifacts seen in dark hair regions of residual features in Figure 4.5. Moreover,

Figure 4.6: A visualization of residual features learned by GSC-GAN for CelebA attributes. We also show the pixel-wise difference between input and output image for comparison.

we can also spot erroneous modifications made by the model such as the jaw brightness change for AU12 and AU15.

We visualize the residual features learned by the model for CelebA attributes in Figure 4.6. Again, we observe that the residual features help us visualize the feature modifications, although there is some reconstruction of irrelevant face features from previous skip connections. The direct input/output image difference produces brightness and color noise for Male and Young attributes, while the residual features provide a noise-free and color-accurate visualization. We show that our visualization method can be used to understand salient face features for an attribute in Figure 4.7. We extrapolate the attribute values outside the training range to -1.5 and 1.5, and generate the modified images. By comparing residual features for low and high attribute values, we can predict the salient face features for an attribute. In comparison, we note that direct input/output image difference produces color noise in some cases. Moreover, a direct pixel-wise difference does not highlight changes well for the Bangs attribute, where the modified feature has a color similar to the forehead region.

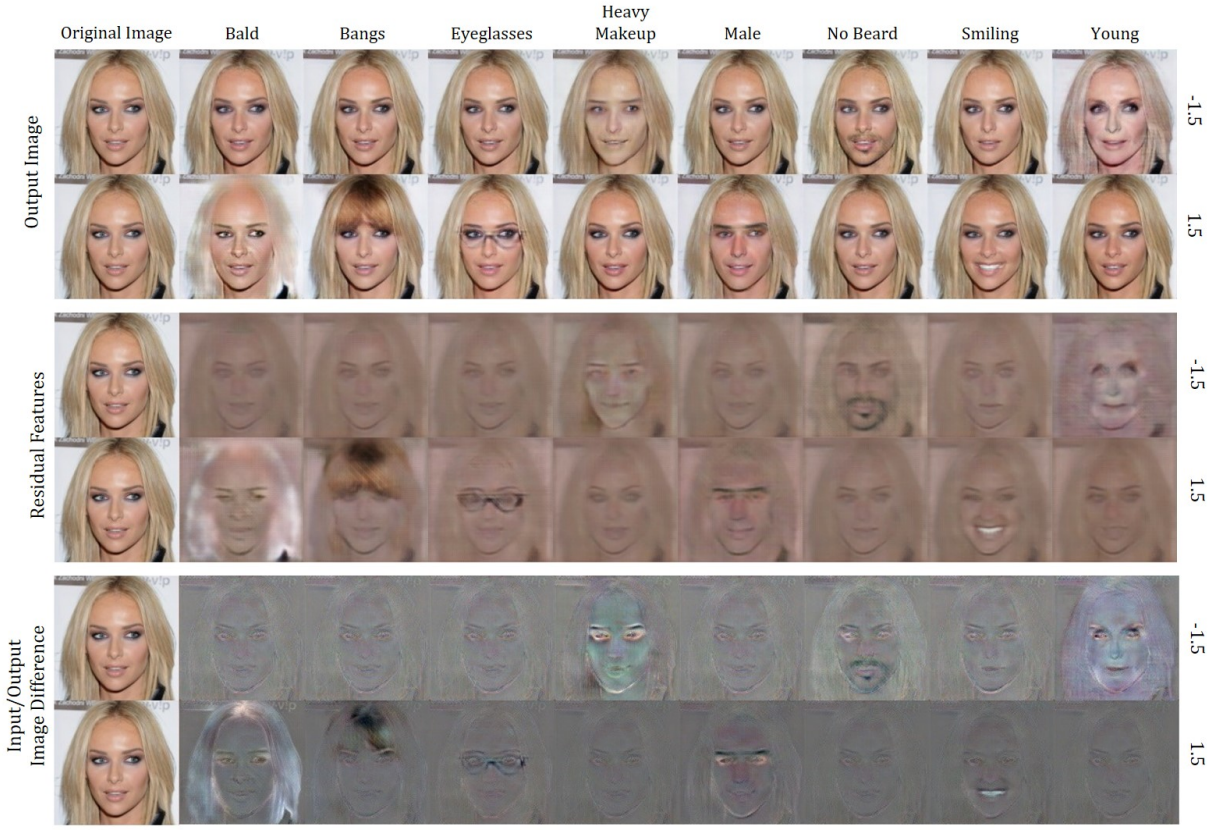

Figure 4.7: A visualization of residual features learned by GSC-GAN for CelebA attributes, for relative attribute values of -1.5 and 1.5. We also show the pixel-wise difference between input and output image for comparison.

Acknowledgment Chapter 4, in part, has been submitted for publication of the material. Yadav, Devendra Pratap; Hu, Weifeng; Song, Amanda; Vul, Edward; Cottrell, Garrison. "Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification". The thesis author was the primary investigator and author of this paper.

Chapter 5

Modifying Perceived Social Impressions of Faces

CelebA face attributes and Action Units correspond to facial features which can be easily visualized. When we modify these features using a GAN, we know what features should be generated for an accurate modification. We consider a new set of face attributes related to social traits, for which we cannot intuit the correlated face features easily. When we see a person's face, we quickly form impressions of them along social dimensions such as kindness, trustworthiness and attractiveness. Several studies [OT08, TODMS15, SLZ+18] have analyzed how these perceptions vary across individuals and cultures. However, attempts to explain possible factors such as facial features that affect social perception have been scarce.

In previous sections, we have shown that GSC-GAN can accurately modify attributes with high identity preservation. We use GSC-GAN with the Social Traits Dataset from Section 2.1 to study 10 perceived social traits: *trustworthy, humble, kind, attractive, masculine, healthy, intelligent, powerful, responsible, successful*. Since the Social Traits Dataset contains only 1611 images, we generate continuous social trait ratings in $[-1.0, 1.0]$ for CelebA-HQ-60k images as described in Section 2.3.1 for training GSC-GAN.

5.1 Correlated Visual Features for Social Traits

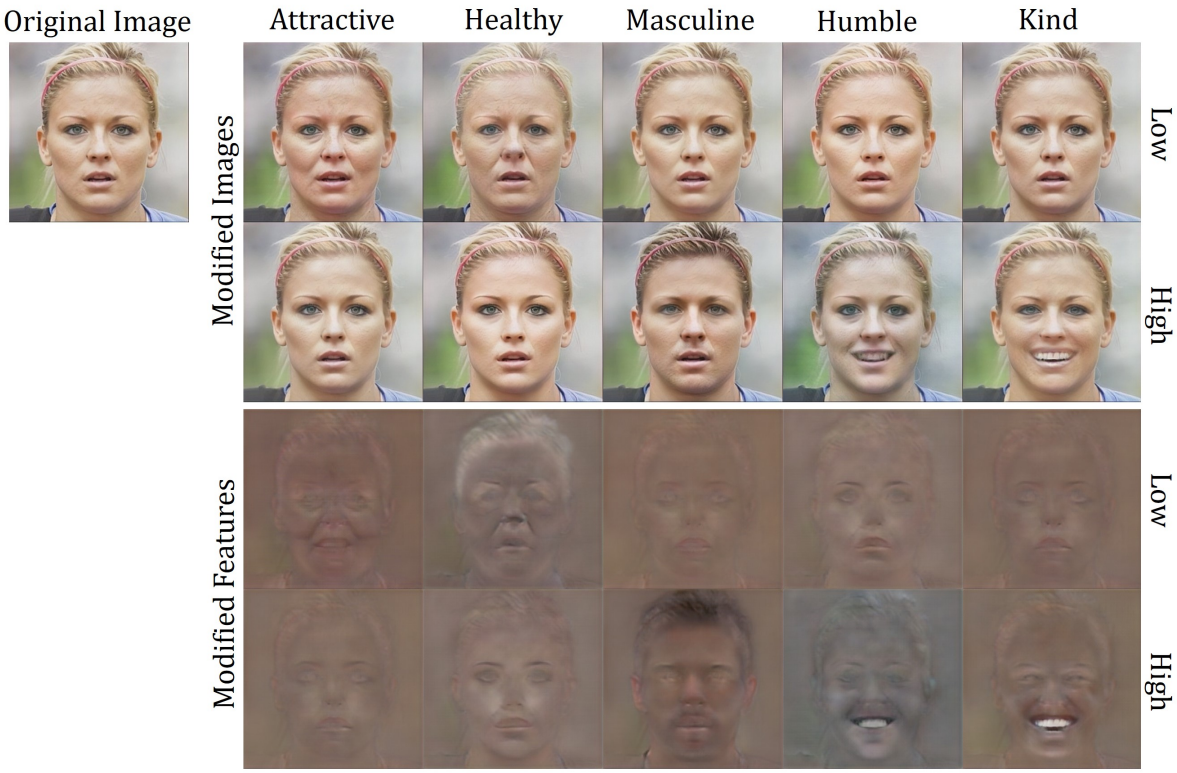

Figure 5.1: Social trait modification to low and high values using GSC-GAN. Top two rows show generated images, and the bottom two rows show the modified face features (by visualizing residual features).

We train GSC-GAN on CelebA-HQ-60k to modify 10 continuous social traits. We use our default GSC-GAN architecture for 256×256 resolution images. The model achieves a FID score 12.18 and 99.6% identity preservation accuracy on the CelebA-HQ-60k test set. Figure 5.1 and Figure 5.2 show a face modified for each social trait to low and high values, which correspond to 2.5 and 97.5 percentile rating values respectively. We observe selective transformation of facial features and good identity preservation across all traits. Attractive and healthy show a noticeable change in skin smoothness and age. Warmth related traits (kind, humble, trustworthy) show variation in smile intensity. For capability related traits (intelligent, powerful, and successful), the model slightly changes smiles, age, makeup and overall image colors. While we can visually

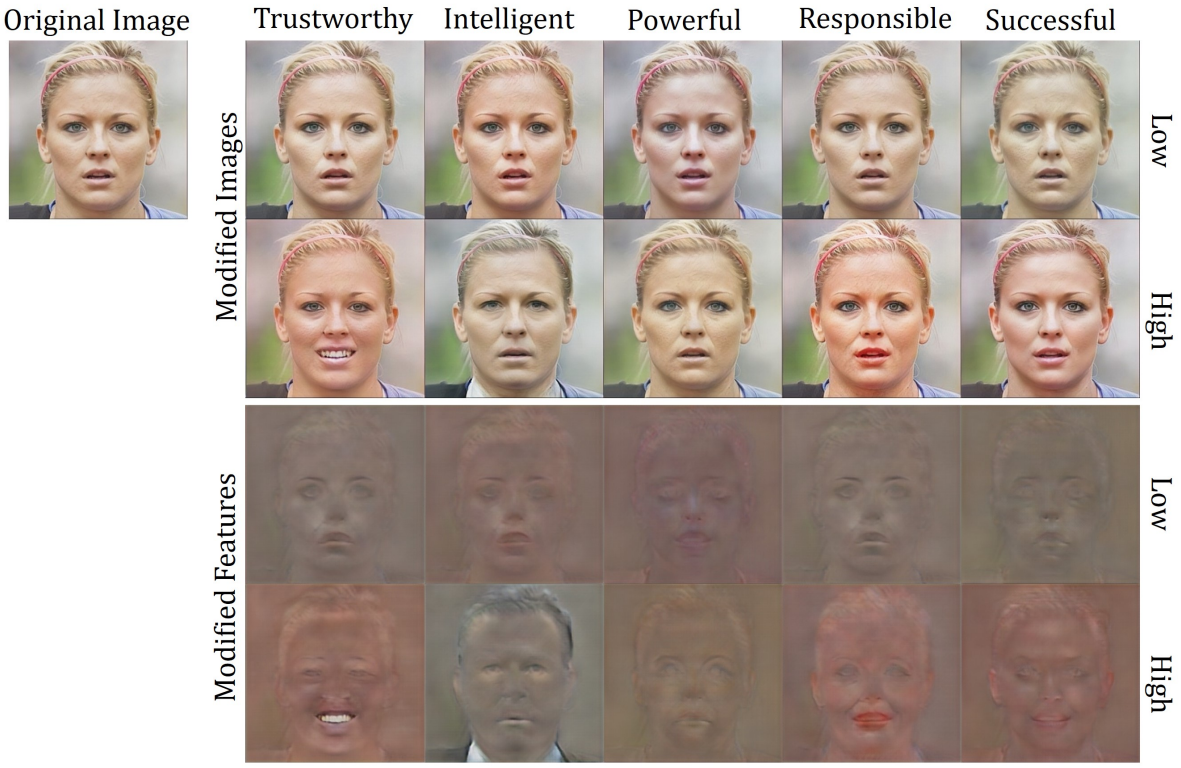

Figure 5.2: Social trait modification to low and high values using GSC-GAN. Top two rows show generated images, and the bottom two rows show the modified face features (by visualizing residual features).

compare low and high rated images, our model-based visualization of residual features better highlights the salient face features. This visualization method can also surface dataset biases. For example, the masculine face shows generation of darker hair. Analyzing the original ratings on 1611 images from Section 2.1, we find that only 1.4% of male images had blond hair compared to 26.8% of female images. We also show that GSC-GAN can perform continuous modification of these social traits in Appendix A.3. We use relative social trait ratings and interpolate from low to high values, showing generation of salient face features for each attribute.

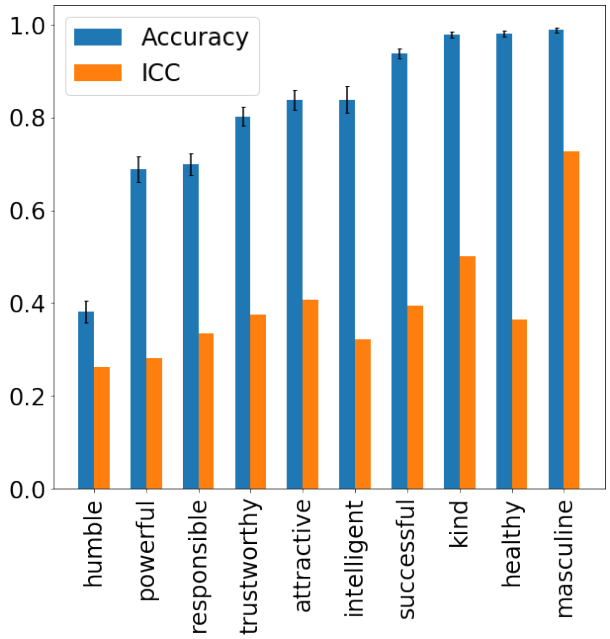

Figure 5.3: Human verification accuracy for social trait modification using GSC-GAN. ICC denotes Intraclass Correlation Coefficient which measures rater agreement when rating a social trait.

5.2 Human Validation Study

While our model helps visualize relevant face features for subjective social traits, we don't know if these face features actually affect human social perception. Hence, we conduct a human rating experiment to verify the modifications learned by our model. While we used CelebA-HQ-60k dataset for training GSC-GAN, we test on the originally rated 1611 faces from the Social Traits Dataset (Section 2.1). We select 50 images (25 male, 25 female) that are closest to the mean ratings for all traits, so that any GAN generated artifacts do not asymmetrically affect perception.

We conduct a human validation study on Amazon Mechanical Turk (AMT) by showing two images of a face modified to low and high social trait values, and ask raters to select the one they perceive to have a higher rating. We recruited 100 raters and obtained at least 10 ratings on each image pair. Figure 5.3 shows accuracy (how often people chose the image modified to a

high value of the trait) for each social trait (chance is 50%). Across all 10 social traits, we obtain an average accuracy of $81.3 \pm 12.1\%$.

Furthermore, we compare modification accuracy with human raters' agreement on each trait by calculating the Intraclass Correlation Coefficient (ICC) [Bar66] of human ratings on the 1611 faces. Traits with salient facial features such as masculine, kind, attractive show a high level of agreement, while humble, powerful and intelligent are more ambiguous and have a low ICC. Higher rater agreement translates to better training signal, leading to higher model accuracy for social traits with high ICC in general. Humble, which has the lowest rater agreement, has an accuracy of 38%, indicating that our model modified humbleness in the wrong direction. Overall, our experiment shows that GSC-GAN is able to learn face modifications corresponding to social traits, and is able to modify face images to change their social perception.

Acknowledgment Chapter 5, in part, has been submitted for publication of the material. Yadav, Devendra Pratap; Hu, Weifeng; Song, Amanda; Vul, Edward; Cottrell, Garrison. "Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification". The thesis author was the primary investigator and author of this paper.

Chapter 6

Conclusion

We study the task of face image modification and propose a novel GAN architecture that selectively uses skip connections to generate high quality outputs. Our experiments show that GSC-GAN achieves significantly better image quality and identity preservation than GANimation on modifying continuous attributes. GSC-GAN achieves performance similar to STGAN with a parameter efficient architecture. Using our high resolution CelebA-HQ-60k face dataset, we show that GSC-GAN scales up to 512x512 resolution, while retaining its modification accuracy and image quality. Our network learns to generate residual image features, which can be visualized to predict salient features for an attribute. Finally, we use GSC-GAN to modify faces along the dimension of perceived social traits. We show that our model can modify face features to change the perception of social traits, obtaining 81.3% accuracy on a human rating trial. Using a similar experimental framework, we can study arbitrary attributes that are labeled on images.

6.1 Future Work

While our model has good performance on identity preservation and image quality, the magnitude of changes is lower than GANimation. However, GANimation's modifications come at the cost of many image artifacts. We want to investigate improvements to our architecture

to make the modifications more accurate. We show that our model works for up to 512x512 resolution images. Evaluating the model at 1024x1024 resolution and generating high quality modified features is an area of interest. Additionally, we can incorporate multi-scale training for generator and discriminators to improve image quality. Multi-scale or progressive GAN training has achieved state-of-the-art results with StyleGAN [KLA19] and Progressive GAN [KALL17]. We also aim to use our model on non-face datasets and evaluate its image-to-image translation performance.

6.2 Broader Impact and Ethics

Our work paves the way for controllable, high resolution, identity preserving face modifications. The model can be used as a face editing tool that performs complex modifications along specific attribute dimensions, making it possible to create stimuli for social psychology experiments. We present an example in our paper where we generate images of the same person interpolated along a social trait. Similarly, we can produce high quality modifications of Facial Action Units to study the effects of facial expressions beyond the six basic emotions.

However, this work can also be used for nefarious purposes. For example, someone could make themselves look more attractive in photos to attract potential dates. Politicians could modify images of their opponents to make them look less intelligent, more aggressive and less trustworthy. Research has shown that facial impressions can affect election outcomes, so this is not an idle concern [TODMS15]. Similarly, hiring decisions can be affected by facial impressions [TODMS15]. In a remote hiring situation, a person could make themselves look more trustworthy.

On the other hand, it is important for social robots to understand facial impressions. If a robot can perceive that someone is not attractive or looks untrustworthy, they will be able to understand how that person might be treated by conspecifics, and attempt to mitigate these human biases. Again, these social impressions are subjective, not veridical, so they can lead to unfair

treatment simply due to how a person looks.

Using a framework of learning residual features, our model's generator provides a way to visualize the changes it makes to the input image. This can be used to study how the feature modifications vary across input images and attribute values. Additionally, it can reveal the salient features for the attribute being modified. If the correct feature modifications are known for an attribute, we can use the modified feature visualization to detect biases in the model/dataset, as well as any erroneous changes made by the model. For example, although the intelligence ratings are very similar for men and women in our dataset, men are rated slightly higher on this trait. Figure 5.2 shows that the modification for increasing the perception of intelligence looks like a male in a white collar. This is a bias that has been amplified by the model, making it more salient.

Acknowledgment Chapter 6, in part, has been submitted for publication of the material. Yadav, Devendra Pratap; Hu, Weifeng; Song, Amanda; Vul, Edward; Cottrell, Garrison. "Gated Skip Connections for High Fidelity, Identity-Preserving, Continuous Face Modification". The thesis author was the primary investigator and author of this paper.

Appendix A

High Resolution Face Modification Results

A.1 Discrete CelebA Attributes

Figure A.1 presents interpolation along CelebA attributes: *Bald, Bands, Eyeglasses, Heavy Makeup, Male, No Beard, Smiling, Young* with relative values from -1.5 to 1.5. Note that GSC-GAN is trained with probability based continuous labels for the CelebA attributes, which leads to smooth generation of face features during interpolation.

A.2 Continuous Facial Action Units

Figure A.2 presents interpolation along Action Unit attributes: *AU7, AU9, AU10, AU12, AU20, AU25, AU26, AU45* with relative values from -1.5 to 1.5.

A.3 Perceived Social Traits

Figure A.3 presents interpolation along perceived social traits: *Successful, Attractive, Kind, Masculine, Powerful* with relative values from -1.5 to 1.5. Similarly, Figure A.4 presents interpolation along perceived social traits: *Healthy, Intelligent, Humble, Trustworthy, Responsible*.

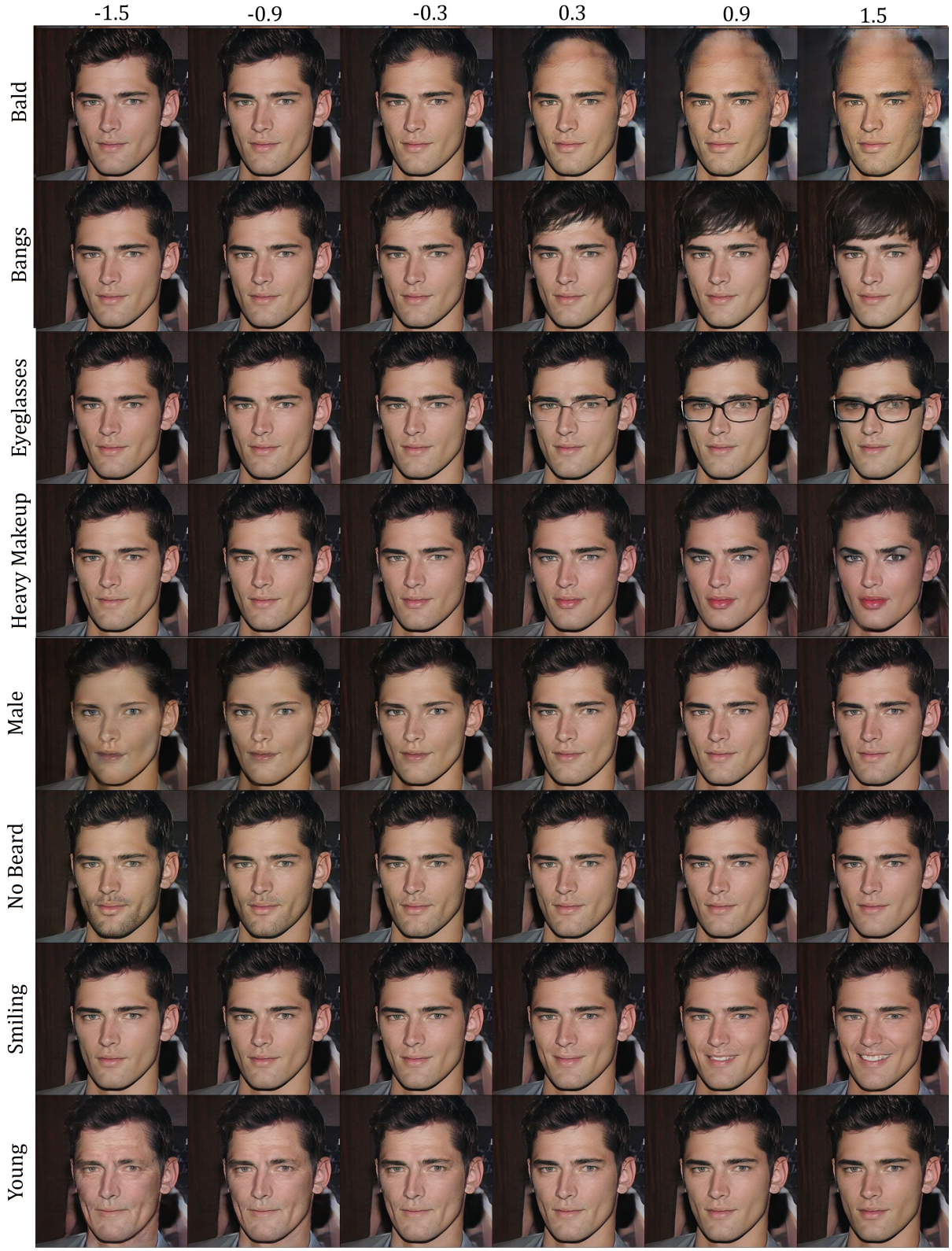

Figure A.1: CelebA attribute modification using GSC-GAN at 512×512 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details.

Figure A.2: Action Unit attribute modification using GSC-GAN at 512×512 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details.

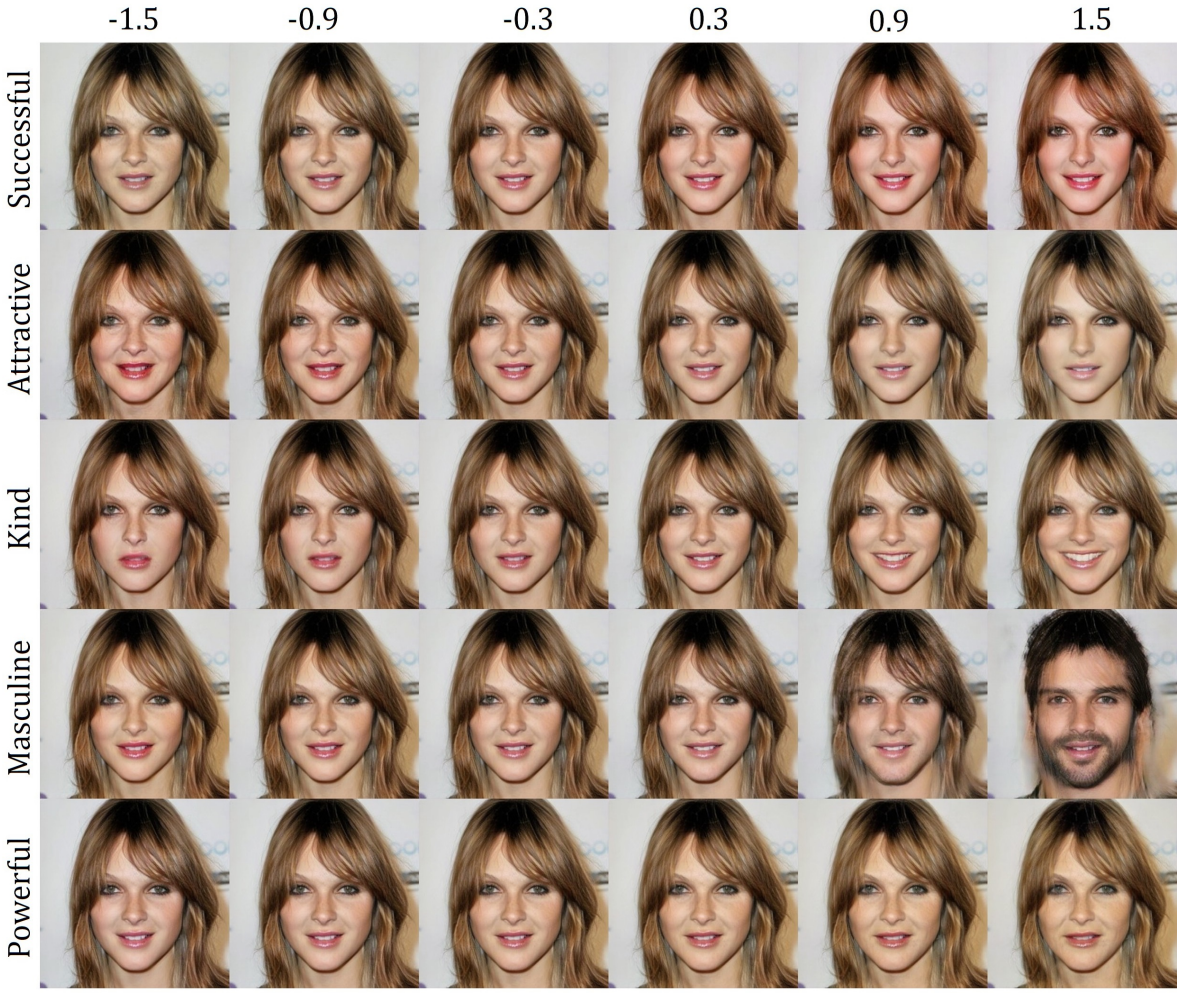

Figure A.3: Social Trait attribute modification using GSC-GAN at 256×256 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details.

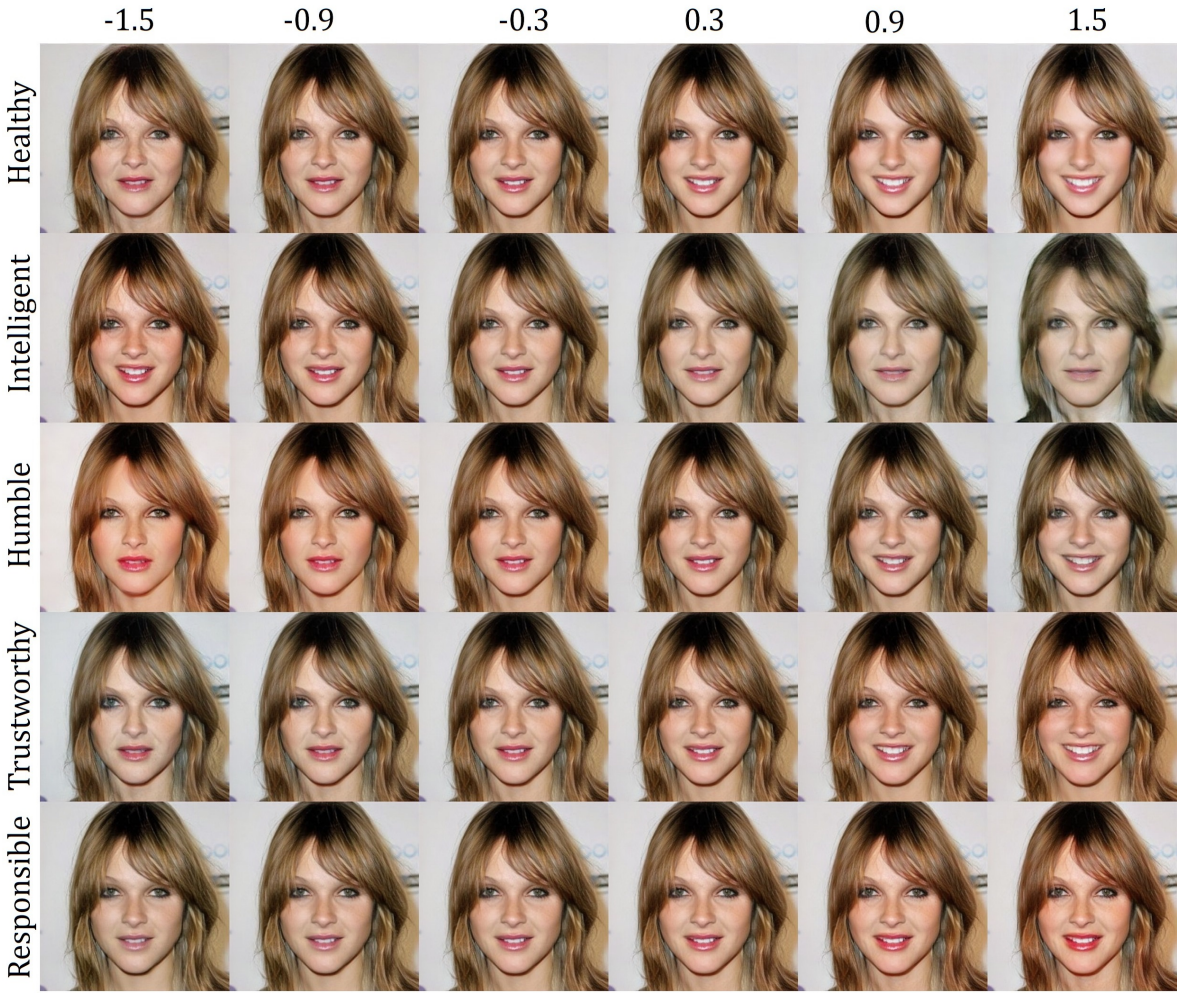

Figure A.4: Social Trait attribute modification using GSC-GAN at 256×256 resolution. We interpolate along relative attribute values from -1.5 to 1.5. Zoom in for details.

Bibliography

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[Ata19] Chad E Atalla. Modifying social dimensions of human faces with modifae. Master's thesis, UC San Diego, 2019.

[Bar66] John J Bartko. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19(1):3–11, 1966.

[BIO13] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323, 2013.

[BMM+17] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2017.

[BZLM18] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[CCK+18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[CSX+18] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[CVMBB14] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[FBQSM16] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[FE78] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978.

[GAA+17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[GPAM+14] 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*, pages 2672–2680, 2014.

[HLBK18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[HRU+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[HSFS17] Eric Hehman, Clare AM Sutherland, Jessica K Flake, and Michael L Slepian. The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, 113(4):513, 2017.

[HSS18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[HSU18] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[HZK+19] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[KALL17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kin09] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[KLA19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[LDX+19] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3673–3682, 2019.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[MO14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[OT08] Nikolaas N Oosterhof and Alexander Todorov. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087–11092, 2008.

[PAM+18] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[PLWZ19] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[SLZ+18] Clare AM Sutherland, Xizi Liu, Lingshan Zhang, Yingtung Chu, Julian A Oldmeadow, and Andrew W Young. Facial first impressions across culture: Data-driven modeling of chinese and british perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin*, 44(4):521–537, 2018.

[TODMS15] Alexander Todorov, Christopher Y Olivola, Ron Dotsch, and Peter Mende-Siedlecki. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, 66:519–545, 2015.

[WLZ+18] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[WYW+18] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[ZKSC18] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–432, 2018.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[ZWZZ16] Guo-Bing Zhou, Jianxin Wu, Chen-Lin Zhang, and Zhi-Hua Zhou. Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing*, 13(3):226–234, 2016.