Perceptually Validated Precise Local Editing for Facial Action Units with StyleGAN

by

Alara Zindancıoğlu

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Data Science

December 15, 2021

Perceptually Validated Precise Local Editing for Facial Action Units with StyleGAN

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Alara Zindancıoğlu

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. T. Metin Sezgin (Advisor)

_____

Prof. Peter Robinson

_____

Prof. Yücel Yemez

Date: _____

ABSTRACT

Perceptually Validated Precise Local Editing for Facial Action Units with StyleGAN
Alara Zindancıoğlu
Master of Science in Data Science
December 15, 2021

The ability to edit facial expressions has a wide range of applications in computer graphics. The ideal facial expression editing algorithm needs to satisfy two important criteria. First, it should allow precise and targeted editing of individual facial actions. Second, it should generate high fidelity outputs without artifacts. We build a solution based on StyleGAN [Karras et al., 2019], which has been used extensively for semantic manipulation of faces. As we do so, we add to our understanding of how various semantic attributes are encoded in StyleGAN. In particular, we show that a naive strategy to perform editing in the latent space results in undesired coupling between certain action units, even if they are conceptually distinct. For example, although brow lowerer and lip tightener are distinct action units, they appear correlated in the training data. Hence, StyleGAN has difficulty in disentangling them. We allow disentangled editing of such action units by computing detached regions of influence for each action unit and restrict editing to these regions. We validate the effectiveness of our local editing method through perception experiments conducted with a total of 43 subjects. The results show that our method provides higher control over local editing and produces images with superior fidelity compared to the state-of-the-art methods.

ÖZETÇE

StyleGAN kullanarak Yüz Eylem Üniteleri için Algısal Olarak Doğrulanmış Hassas Yerel Düzenleme

Alara Zindancıoğlu

Veri Bilimleri, Yüksek Lisans

15 Aralık 2021

Yüz ifadelerini düzenleme yeteneği, bilgisayar ara yüzlerinde çok çeşitli uygulamalara sahiptir. İdeal yüz ifadesi düzenleme algoritmasının iki önemli kriteri karşılaması gerekir. İlk olarak, bireysel yüz eylemlerinin kesin ve hedefli düzenlenmesine izin vermelidir. İkincisi, yapaylık olmadan yüksek kaliteli çıktılar üretmelidir. Biz yüz ifadesi düzenlemesi için, yüzlerin anlamsal manipülasyonu alanında yaygın olarak kullanılan StyleGAN [Karras et al., 2019]'a dayalı bir çözüm oluşturuyoruz. Bunu yaparken, StyleGAN'da çeşitli anlamsal niteliklerin nasıl kodlandığına dair anlayışımıza katkıda bulunuyoruz. Özellikle, gizli uzayda düzenleme yapmak için naif bir stratejinin belirli eylem birimleri arasında, birbirlerinden bağımsız olsalar bile, istenmeyen bağlantılarla sonuçlandığını gösteriyoruz. Örneğin, kaş indirme ve dudak gerdirme farklı eylem birimleri olmasına rağmen, eğitim verilerinde ilişkili görünüyorlar. Bu nedenle, StyleGAN onları çözmekte zorluk çekiyor. Her bir eylem birimi için ayrılmış etki bölgelerini hesaplayarak bu tür eylem birimlerinin bağımsız düzenlenmesine izin veriyoruz ve düzenlemeyi bu bölgelerle sınırlandırıyoruz. Toplam 43 denekle gerçekleştirilen algı deneyleri ile yerel surat ifadesi düzenleme yöntemimizin etkinliğini doğruluyoruz. Sonuçlar, yöntemimizin yerel düzenleme üzerinde daha yüksek kontrol sağladığını ve son teknoloji yöntemlerle karşılaştırıldığında üstün ve resmin aslına uygun görüntüler ürettiğini gösteriyor.

ACKNOWLEDGMENTS

I would like to thank my advisor Assoc. Prof. T. Metin Sezgin for his invaluable guidance and encouragement. His expertise in the field and personal support has always motivated and steered me throughout my research. I would also like to extend my sincere gratitude to our team at Intelligent User Interfaces (IUI) Lab, for their friendship and unconditional support. Special thanks to Ezgi Dede for her immense assistance in conducting the user studies, and Alpay Sabuncuoğlu for his valuable insights and allowing me to publish his edited images.

With a background in Physics and Philosophy, I had little experience in Computer Science. Thanks to my advisor and all my professors, I managed improve myself in the field. I feel very lucky for being their student, and I am extremely grateful for the education I have received at Koç University.

My deepest gratitude goes to my family and friends for standing by me every step of the way. Thank you to my parents for their guidance and for providing everything I need for my academic endeavours. Thank you to my childhood friends Duru, Ece, and Meltem for their emotional support and always being there for me. Thank you to Begüm, Beste, and Işılay for making my masters journey priceless. Lastly, thank you to Berke for always believing in me and pushing me to do better.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AU	Action Unit
FACS	Facial Action Coding System
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
SGD	Stochastic Gradient Descent
VAE	Variational Autoencoder

Chapter 1

INTRODUCTION

Facial expression editing aims to manipulate the expression of a given face image, while preserving the personal identity and other facial features. Due to the complex structure of human face, the task is especially challenging since it involves synthesizing unseen muscle movements realistically. Photorealistic facial expression editing has a wide range of applications in various fields such as human-computer interaction, entertainment industry, virtual reality, and photography.

The task of facial expression editing has been approached with traditional computer vision, as well as deep learning based methods. Traditional approaches mainly make use of mapping, warping, and morphing operations [Wang et al., 2003, Ma et al., 2008, Liu et al., 2001, Zhang et al., 2005, Yang et al., 2011]. They require excessive supervision of operations for extracting complex geometric information from the face, computing flow maps, or doing face decomposition. Approaches based on deep learning attempt to build end-to-end generative models to synthesize face images with multiple facial expressions [Ding et al., 2018, Choi et al., 2018, Song et al., 2018, Qiao et al., 2018, Pumarola et al., 2019]. With the advancements in variational autoencoders (VAEs) and generative adversarial networks (GANs), impressive results have been reported in face synthesis, facial expression editing, and transfer. The main procedure with generative models is to encode images into a semantically meaningful latent space and perform editing by manipulating the latent vector.

The field of facial expression editing is closely related to Psychology. Facial Action Coding System (FACS) developed by Paul Ekman [Friesen and Ekman, 1978] is widely used in the area to label facial actions. It describes emotions and related

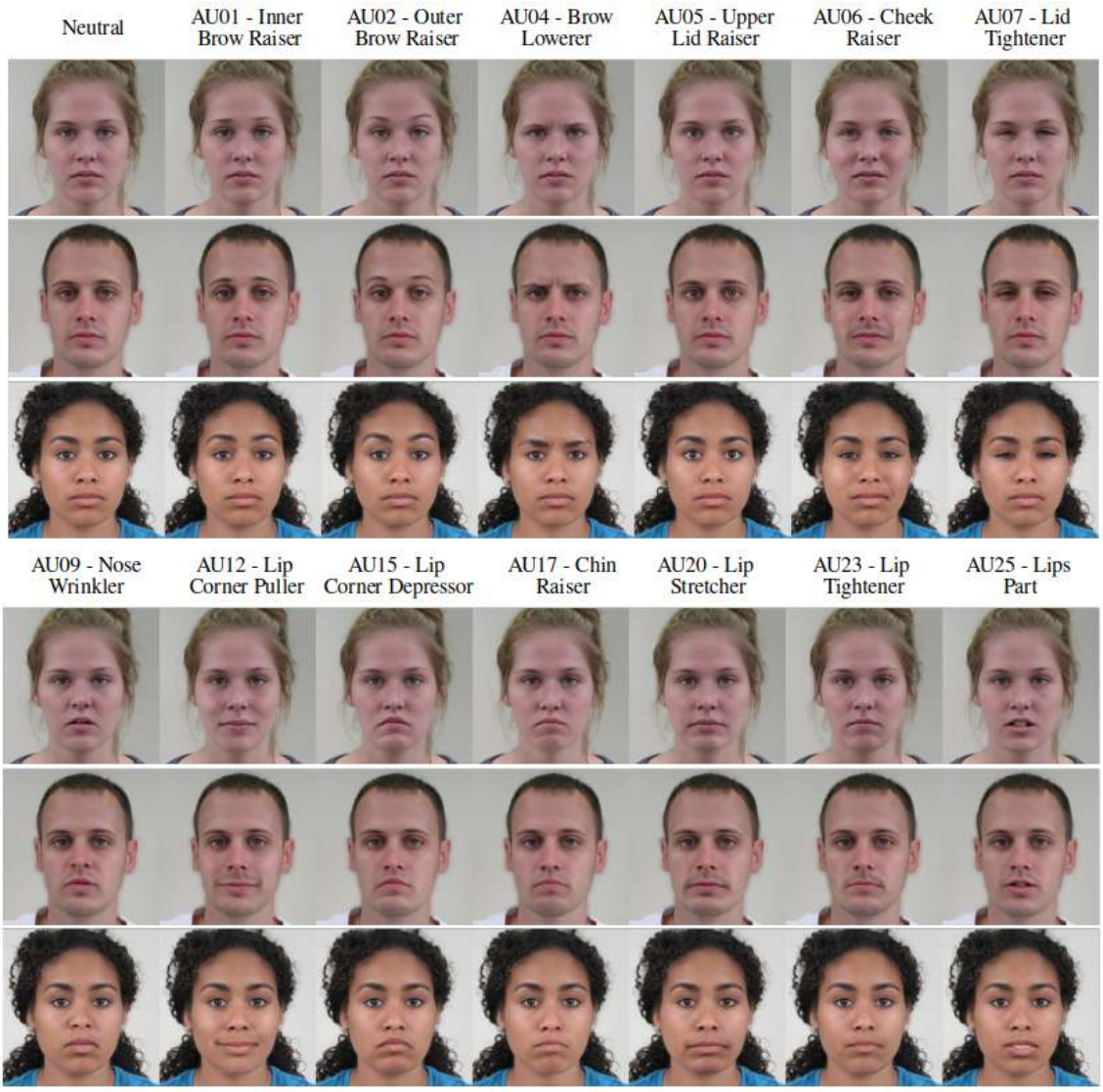


Figure 1.1: Sample editing results for individual action units.

facial movements in terms of Action Units (AUs). In some works, AU activations are used as labels or conditions for face editing with generative models [Susskind et al., 2008, Pumarola et al., 2019].

With successful GAN architectures capable of synthesizing photorealistic face images, a recent research approach has been to utilize pre-trained GAN models for image editing. Along the same line, we build our solution for facial action unit editing on the state-of-the-art StyleGAN model. The mechanism through which the latent space of StyleGAN maps to various semantic attributes is not well understood, but there have been notable attempts to characterize it. Some authors have proposed finding linear directions corresponding to specific attributes within the StyleGAN latent space. For example, Shen et al. [Shen et al., 2020] show that there are linear directions within the latent space corresponding to certain attributes such as pose, age, gender, and the presence of a smile. Others have identified certain layers and channels within the StyleGAN architecture that control discrete semantic attributes [Collins et al., 2020]. Unfortunately, neither of these approaches successfully characterize variations in facial action units, which are localized, fine scale and delicate in nature. We add to our understanding of the latent space of StyleGAN by showing that fine scale variations in facial action units require manipulating the latent vector representation of a source content, as well as identifying and adjusting localized patches within the StyleGAN layers and activation maps. To achieve this we present two complementary methods: first is global editing, where latent encodings of images are manipulated in a way that does not necessarily focus on specific regions on the resulting image, and local editing, which refers to restricting the global edit only to a region of interest. Sample editing results are shown in Figure 1.1.

Our contributions are as follows:

- We present a simple yet effective way of facial action unit editing in the latent space of StyleGAN, producing high-fidelity editing results.

- We offer methods for both global and local editing of facial action units. Our approach allows adjusting the strength of the action units on a continuous

scale and supports joint editing of multiple action units.

- We add to our understanding of the latent and activation spaces of StyleGAN by analysing how semantics of facial action units are encoded.

- We show that combining editing techniques in latent and activation spaces improves the editing quality for facial action units.

Chapter 2

RELATED WORK

Facial expressions are essential in social communication since they convey important information about people's internal emotional states. Facial Action Coding System (FACS) represents facial muscle movements that are related to emotions through Action Units (AUs). There are a number of AUs defined and their combinations produce a wide range of facial expressions and emotions. AU recognition has long been studied, and it is usually intertwined with facial landmark and emotion recognition tasks [Baltrušaitis et al., 2015, Ruiz et al., 2015, Onal Ertugrul et al., 2019].

2.1 Deep generative models

Common deep generative models for image synthesis are variational autoencoders (VAEs) and generative adversarial networks (GANs). VAEs are probabilistic models with an encoder-decoder architecture. GAN models are composed of a generator and a discriminator trained jointly in a minimax game setup. With recent improvements in architectures and loss functions [Liu et al., 2017, Kingma and Welling, 2013, Larsen et al., 2016], we often see combined VAE-GAN architectures to boost performance and capabilities. In addition to realistic image synthesis, deep generative models are widely used for a wide range of tasks such as conditional image generation [Mirza and Osindero, 2014], image-to-image translation [Isola et al., 2017, Liu et al., 2017, Choi et al., 2018], super-resolution imaging [Ledig et al., 2017], and facial attribute editing [Perarnau et al., 2016, Shen et al., 2020, He et al., 2019, Liu et al., 2019].

2.2 Facial expression editing

Facial expression editing is an intricate task as it requires manipulation of various face regions simultaneously and realistically while preserving the identity of the face. Some generative models are conditioned on discrete emotion categories [Choi et al., 2018, Ding et al., 2018, Zhou and Shi, 2017]. StarGAN [Choi et al., 2018] uses a single model for multi-domain image translation, which can be utilized for facial expression transfer. Others use facial geometry in order to perform continuous editing of facial expressions on a continuous range of intensities [Qiao et al., 2018, Song et al., 2018, Pumarola et al., 2019]. Facial expression transfer is another problem that has been widely studied, where the aim is to transfer the expression of one face to another [Xu et al., 2017, Wiles et al., 2018].

There are also lines of work that utilize action unit labels to perform editing on the face [Tripathy et al., 2020, Pumarola et al., 2019, Wu et al., 2020]. All aformentioned methods, except [Pumarola et al., 2019, Wu et al., 2020] are methods that process the image as a whole. GANimation [Pumarola et al., 2019], on the other hand, incorporates an attention later that focuses the network on regions of the face that are most relevant for synthesizing the target expression. Another work that focuses on local regions is Cascade EF-GAN [Wu et al., 2020], which processes images via global and local branches that focus on selected expression-intensive regions - eyes, nose, and mouth.

2.3 Utilizing pre-trained models

Utilizing pre-trained GAN models for image editing has attracted attention recently. Since most successful GANs for image synthesis generate images based on a random latent code, real images first need to be encoded in the latent space of the specific GAN for subsequent use. There are several GAN-inversion techniques that allow encoding real images in the latent space. While some methods train and encoder neural network to find the latent representation [Chai et al., 2021, Kingma and Welling, 2013], others optimize the latent code such that the generated image matches the target image [Baylies, 2019, Creswell and Bharath, 2018, Abdal et al.,

2019].

Two main approaches to manipulate images with a pre-trained GAN are investigating the semantics in the latent space to perform edits on the latent vectors and manipulating the activation maps. Latent code based methods make use of intrinsic semantics of the latent space in order to learn a manifold or find directions through which latent vectors can be moved [Shen et al., 2020, Nitzan et al., 2020]. Activation-based methods focus on the semantics learned by feature maps of convolutional layers. While some manipulate spatial positions on the activation tensors of certain layers [Bau et al., 2018, Suzuki et al., 2018], others study if certain feature maps focus on specific objects and perform editing by only manipulating those channels [Collins et al., 2020].

Our work can be viewed as a combination of latent code and activation-based methods. To perform global edits on the input image, we use latent space manipulation and investigate the disentangled semantics learned in the StyleGAN latent space. We also present two techniques for local editing: 1) optimizing the latent encoding and 2) patching the activation maps. In addition to editing, we study and report insights on the semantics learned by the latent and activation spaces of StyleGAN.

Chapter 3

METHODOLOGY

Instead of building a generative adversarial network from scratch, we use the state-of-the-art GAN model StyleGAN to synthesize images, and perform editing in its latent and activation spaces. Given an input image $\mathbf{I}_{orig}$, we first encode it to the intermediate latent space $\mathcal{W}$ of StyleGAN using the StyleGAN-encoder [Baylies, 2019] to obtain $\mathbf{w}_{orig}$. Next, $\mathbf{w}_{orig}$ is fed to the *global editor network*, which is essentially an action unit detector but is utilized to edit latent vectors with respect to desired action unit intensities. The details of global editing are discussed in Section 3.1. As we elaborate later, our experiments show that some action units appear to be entangled with each other, i.e. changing certain action units causes changes at parts of the face that were not intended to change. To overcome this problem, we use *spherical k-means clustering* [Buchta et al., 2012, Collins et al., 2020] to find local regions of the face that correspond to each action unit. Together with these local regions, globally edited latent vector $\mathbf{w}_{global}$ is fed to the *local editor network*, which modifies only the desired local regions of the face. The overall model architecture is demonstrated in Figure 3.1.

We created an encoded dataset consisting of 10,323 images from the RaFD [Langner et al., 2010], Bosphorus [Savran et al., 2008], and CFEE [Du et al., 2014] datasets. For labeling, we use the OpenFace software [Baltrusaitis et al., 2018, Baltrušaitis et al., 2015] and get regression values of 16 action units for each image. This datasets is used to train the global editor network.

3.1 Global Editing

For global editing, we train an action unit detector that predicts action unit values from the latent encodings of images. The network takes as input the latent vectors $\mathbf{w}$

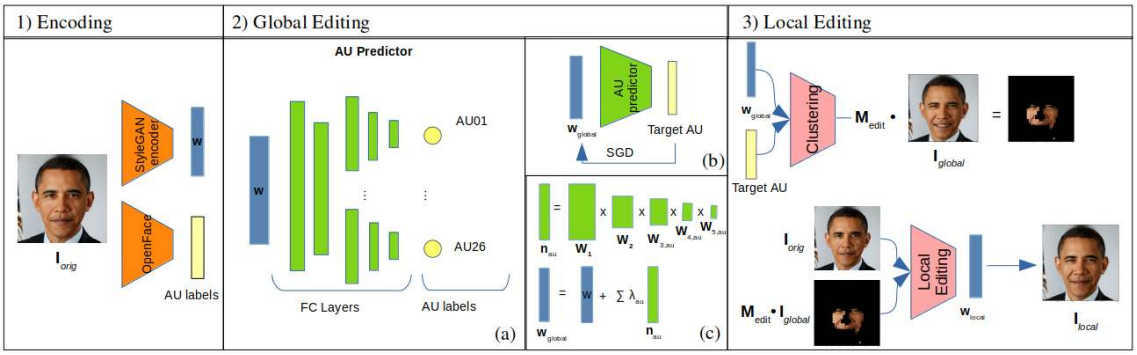


Figure 3.1: Our pipeline for image editing in the latent space of StyleGAN: 1) Images are first encoded in the latent space of StyleGAN and AU labels are prepared with OpenFace [Baltrusaitis et al., 2018]. 2) Latent vectors are fed to the global editor, which is essentially an AU predictor (a) utilized in two different ways for editing: (b) via optimization and (c) by extracting linear directions. 3) Finally local editing is performed on globally edited latent vector $\mathbf{w}_{global}$.

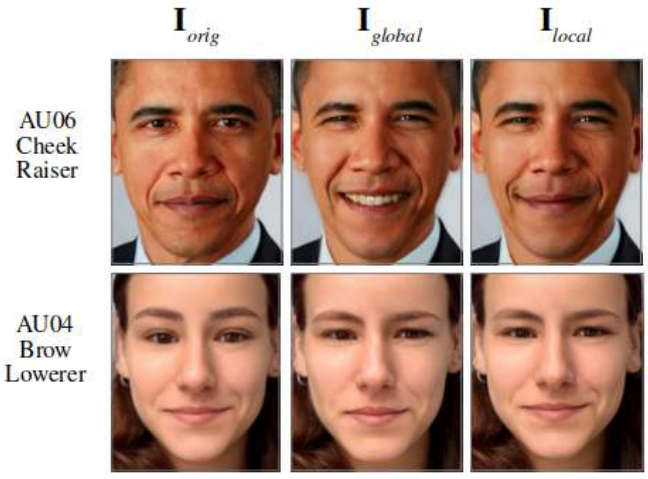


Figure 3.2: With global editing of some action units, changes at unrelated regions of the face can be observed ($\mathbf{I}_{global}$). Local editing restricts manipulations to corresponding regions ($\mathbf{I}_{local}$).

and outputs a vector of regression values for 16 action units **AU**. The AU detector is a feedforward network with two common fully connected layers followed by 16 AU branches, each with three fully connected layers. ReLU activation function is used at each layer except the output nodes, where linear activation function is used. The network is trained with MSE loss and SGD optimizer.

There are two ways in which we use this network to edit action units in the latent space. The first one is an optimization based approach, where we update the $\mathbf{w}$ vector via gradient descent by setting the target action unit values as the output of the AU detector. This approach allows precise editing by only changing specific action unit values and keeping others the same. It can also be used to transfer the target expression from one image to another. But this approach is slow due to its iterative nature.

A faster approach is to get weights of the network after training, and take the product of the weights to get a linear direction in the latent space for each action unit. Let $\mathbf{W}_1$ and $\mathbf{W}_2$ refer to weight matrices of first two layers of the global editor network, and $\mathbf{W}_{3,au}$, $\mathbf{W}_{4,au}$, and $\mathbf{W}_{5,au}$ refer to the third, forth, and fifth layers of network branches that correspond to the action unit $au \in \mathbf{AU}$. The linear direction $\mathbf{n}_{au}$ in the latent space for editing action unit au is then given by:

$$\mathbf{n}_{au} = \mathbf{W}_1 \cdot \mathbf{W}_2 \cdot \mathbf{W}_{3,au} \cdot \mathbf{W}_{4,au} \cdot \mathbf{W}_{5,au} \tag{3.1}$$

To edit multiple action units at once, we take linear combinations of them with varying intensities. Let λ_{au} be the intensity of au and $\mathbf{AU}_{\text{edit}}$ be the list of action units to be edited. Combined direction $\mathbf{n}_{\text{edit}}$ is found by:

$$\mathbf{n}_{edit} = \mathbf{W}_1 \cdot \mathbf{W}_2 \cdot \lambda_{au} \sum_{\mathbf{AU}_{edit}} \mathbf{W}_{3,au} \cdot \mathbf{W}_{4,au} \cdot \mathbf{W}_{5,au} \tag{3.2}$$

Adding the resulting vector to $\mathbf{w}_{orig}$ moves it in a linear direction that enhances the desired action unit value. In both approaches, optimization and moving in linear directions, multiple action units can be modified with varying intensities.

With global editing, especially in the fast approach, entanglement between some action units that often occur together is observed. Figure 3.2 shows the result of entanglement as a result of modifying certain action units. This can be explained

by dataset or latent space bias. To suppress editing in regions where changes were not intended, we further optimize the **w** vector with local editing.

3.2 Local Editing

For local editing, we first identify semantic regions on the face by applying spherical k-means clustering on the activation maps. We then propose two ways of performing local edits on the face: first is optimizing the latent encoding, and second is multiplexing activation maps.

3.2.1 Spherical K-means Clustering

Following Collins *et al.* [Collins et al., 2020], we apply spherical k-means clustering to find local regions learned in the StyleGAN activation space. Clustering is applied to the activation maps $\mathbf{A} \in \mathbb{R}^{N \times C \times H \times W}$ where N is the number of images, C is the number of channels, and H and W are height and width. With K clusters, we obtain a cluster catalog which outputs a tensor of cluster memberships, $\mathbf{U} \in \{0,1\}^{N \times K \times H \times W}$. From K clusters, only those that fall within the face region are used. Using encoded vectors of images from RaFD [Langner et al., 2010], Bosphorus [Savran et al., 2008] and CFEE [Du et al., 2014] datasets, we train the clustering network and determine regions on the face that are most affected by global editing of individual action units. Details of this analysis are given in section 4.2. With both visual inspection of facial regions and the results of this study, we construct a dictionary of action unit – cluster relationships. Thus for each action unit, we have a list of associated clusters.

3.2.2 Optimizing the Latent Encoding

Let $\mathbf{w}_{global}$ be the latent vector of the globally edited image, and $\mathbf{AU}_{\text{edit}}$ be the list of action units to be edited. The latent vector is first fed to the StyleGAN generator to compute the 7th layer[1] activation maps $\mathbf{A} \in \mathbb{R}^{512 \times 32 \times 32}$ and the output

[1]Based on our observation and findings by Collins *et al.* [Collins et al., 2020], layer 7 appears to be the most semantically meaningful layer for facial parts.

image $\mathbf{I}_{global} \in \mathbb{R}^{1024\times1024\times3}$. Then, spherical k-means clustering is applied on the activations $\mathbf{A}$ to obtain the local clusters $\mathbf{U} \in \{0,1\}^{K\times32\times32}$. The local clusters are resized to the output image size and only k clusters that are related to $\mathbf{AU}_{\text{edit}}$ are selected to obtain the tensor $\mathbf{U} \in \{0,1\}^{k\times1024\times1024}$. Note that clusters are selected from a pre-computed dictionary of AU-cluster relationships (Section 4.2). The mask for local editing $\mathbf{M}_{edit}$ is calculated by,

$$\mathbf{M}_{edit} = \sum_{k} \mathbf{U}_{k,1024,1024} \tag{3.3}$$

To locally edit the image, we update the latent vector $\mathbf{w}_{global}$ by minimizing the L2 loss between the output image $\mathbf{I}_{local}$ and globally edited image $\mathbf{I}_{global}$ within the editing mask region. Editing loss is defined as:

$$\mathcal{L}_{edit} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{M}_{edit} \odot (\mathbf{I}_{local} - \mathbf{I}_{global}))^2 \tag{3.4}$$

where $\odot$ denotes elementwise multiplication.

Outside the editing mask, we minimize the L2 loss between the output image $\mathbf{I}_{local}$ and the original image $\mathbf{I}_{orig}$. Regions outside editing mask are defined as $1 - \mathbf{M}_{edit}$. Consistency loss is defined as:

$$\mathcal{L}_{const} = \frac{1}{N}\sum_{n=1}^{N}((1 - \mathbf{M}_{edit}) \odot (\mathbf{I}_{local} - \mathbf{I}_{orig}))^2 \tag{3.5}$$

The total loss is given by,

$$\mathcal{L} = \mathcal{L}_{edit} + \mathcal{L}_{cons} \tag{3.6}$$

After the optimization, the locally edited latent vector $\mathbf{w}_{local}$ is obtained. Finally the locally edited image $\mathbf{I}_{local}$ can be generated by feeding the vector to StyleGAN.

3.2.3 Multiplexing Activation Maps

Another way of performing local editing is multiplexing activation maps, which is faster but does not yield the locally edited latent vector $\mathbf{w}_{local}$. As in optimization based local editing, first $\mathbf{M}_{edit}$ is obtained by applying spherical k-means clustering on the activation maps of globally edited latent vector $\mathbf{w}_{global}$. This time the mask

is not resized to the output image size, as it will be applied on he activation maps of size (512, 32, 32). As the original latent vector $\mathbf{w}_{orig}$ is sent through the network, the patch of 7th layer activation maps $\mathbf{A}_{orig}$ that corresponds to the $\mathbf{AU}_{\text{edit}}$ is replaced by the patch of activation maps from the globally edited image $\mathbf{A}_{global}$.

$$\mathbf{A}_{local} = (1 - \mathbf{M}_{edit}) \odot \mathbf{A}_{orig} + \mathbf{M}_{edit} \odot \mathbf{A}_{global} \tag{3.7}$$

With the locally edited activation map $\mathbf{A}_{local}$, StyleGAN generator successfully incorporates the patch and outputs the locally edited image. Resulting image does not contain artifacts from the patching, since the operation is performed in the intermediate layers and the generator blends the patch as it passes through the subsequent layers.

Chapter 4

EXPERIMENTS

In this section, we first give the implementation details of the model. We then analyze the latent and activation spaces of StyleGAN, and report insights on the semantics of the activation maps.

4.1 Implementation Details

AU detector network

Our action unit detector network, detailed in Section 3.1 of the main paper, has two common fully connected layers, followed by three fully connected layers for each of the 16 action unit branches. The input to the network is the flattened **w** vector of size 9216 (18×512). Sizes of the layers are as follows:

FC-1: 3072

FC-2: 1024

FC-3: 256

FC-4: 32

FC-5: 1

Each fully connected layer is followed by ReLU, and the last layers have a linear activation function. The model is trained for 50 epochs with MSE loss and SGD optimizer, with learning rate of 0.01.

Other parameters

In Section 3.1 of the main paper, we show how linear directions are obtained from weights of the AU detector network. The λ_{au} parameter is used to control the

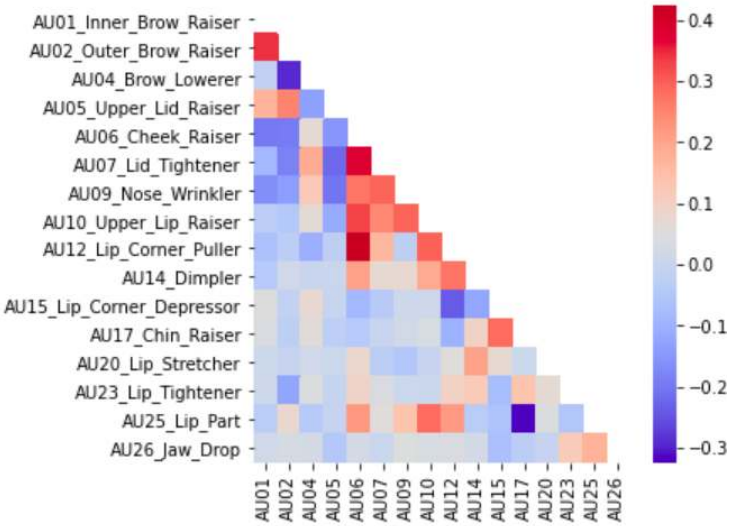


Figure 4.1: Correlation between action units.

intensity of each action unit. In our experiments, we found [-0.5, 0.5] to be the best range for the value of λ_{au}.

For spherical k-means clustering (Sections 3.2.1 and 3.2.2 of the main paper), we choose the initial K value to be 37 via visual inspection of resulting regions. From 37 clusters, we select only the clusters that fall within the face region, which reduces them to k = 22. We find these number of clusters to be sufficient to represent action unit regions. Nevertheless, these values are not sentitive, so experimenting with values for K within the range of 30 to 50 can give equally well results.

4.2 Linear Directions and Entanglement

There are multiple ways of finding linear directions in the latent space of StyleGAN. For example Shen *et al.* train independent SVMs on pose, smile, age, gender, and eyeglasses attributes[Shen et al., 2020] to find linear hyperplanes that separate the latent space such that samples that possess an attribute are placed on the same side. Instead of training independent SVMs for each action unit, we train one network with multiple branches and shared weights, which we refer to as the *global editor network*. We obtain linear directions by multiplying the weights of this network (Section 3.1). Moving the latent vector in the positive direction that corresponds

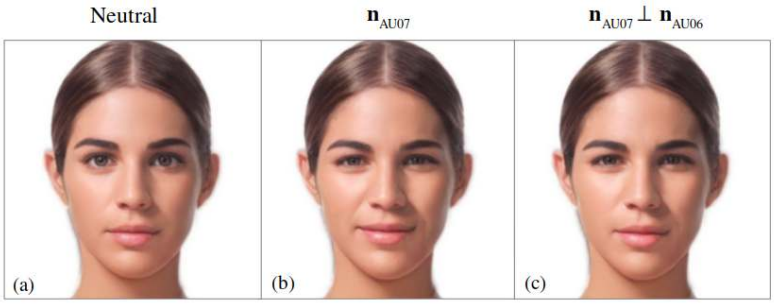


Figure 4.2: Moving in orthogonal directions. Linear direction vector $\mathbf{n}_{\text{AU07}}$ for AU07 - Lid Tightener is highly correlated with direction vector $\mathbf{n}_{\text{AU06}}$ for AU06 - Cheek Raiser (b). Moving in the direction $\mathbf{n}_{\text{AU07}} \perp \mathbf{n}_{\text{AU06}}$ results in only editing AU07 without effecting AU06 (c).

to an action unit enhances the action unit value. Similarly, moving in the negative direction reduces the action unit intensity.

Although action units are conceptually distinct, some are activated together in most existing datasets. For example, brow lowerer and lip tightener, or cheek raiser and lips part action units are distinct and can be performed individually, but they are commonly observed together in emotion related datasets. Thus some of the linear directions that are computed are not totally disentangled with the others. To demonstrate this, we report the correlation of each action unit with respect to the others by calculating the size of projected vector. The correlation graph is presented in Figure 4.1. Orthogonal portions of linear direction vectors can be extracted to decouple editing between correlated action units. Figure 4.2 shows the result of this operation.

4.3 Insights on the Activation Maps

For local editing, we need to form a dictionary of AU-cluster relationships so that editing is limited only to regions that are relevant. To do so, we compare channel activations of neutral and edited images at the clustering regions. This also gives us an insight on how well the global editor network localizes.

After visual inspection and consulting previous research [Collins et al., 2020, Karras et al., 2019], we decide that 7th layer with 32×32 resolution is the best

layer for semantic facial part learning. Using the global editor network, we first edit 180 neutral images from the RaFD dataset for each action unit. Then we feed both edited and neutral vectors to StyleGAN in order to get 7th layer activation maps. We obtain the resulting tensor $\mathbf{A} \in \mathbb{R}^{AU \times N \times C \times H \times W}$, where AU denotes action unit index, N is number of neutral images, C is number of channels, and H and W are spatial dimensions of the layer. In our experiments, $(AU, N, C, H, W) = (17, 180, 512, 32, 32)$, where $AU = 0$ denotes the neutral image and $AU \in [0, 16]$ are action unit indices.

For each neutral and edited image, we also the compute local clusters with spherical k-means clustering to obtain the tensor $\mathbf{U} \in \{0, 1\}^{AU \times N \times K \times H \times W}$, where K denotes the number of clusters.

Using these matrices, our aim is to find: 1) regions that are most affected by global editing of action units and 2) channels that are most activated by certain local regions on the face. With the following operation, we obtain the AU-cluster-channel relationship matrix $\mathbf{R} \in \mathbb{R}^{AU,K,C}$:

$$\mathbf{R}_{au,k,c} = \sum_{n} \left(\frac{\sum_{h,w} (\mathbf{U}_{au,n,k,h,w} \odot (\mathbf{A}_{au,n,c,h,w} - \mathbf{A}_{0,n,c,h,w})^2)}{\sum_{h,w} \mathbf{U}_{au,n,k,h,w}} \right) \tag{4.1}$$

The AU-cluster relationship is computed by:

$$\mathbf{R}_{au,k} = \sum_{c} \mathbf{R}_{au,k,c} \tag{4.2}$$

Table 4.1 presents most affected facial regions by each action unit via global editing. The table shows that global editing affects irrelevant parts of the face for some action units. For example, editing Inner Brow Raiser affects the nose area (cluster 20) the most, which is unexpected. The reason for this may be the bias in the datasets used for training the global editor network, or latent space bias in StyleGAN.

Furthermore, we inspect the facial region and channel relationship in the StyleGAN activation space, independent of action units. Previous research by Collins *et al.* [Collins et al., 2020] suggests that certain channels of the convolutional layers of StyleGAN focus on specific regions. Following their approach we compute facial

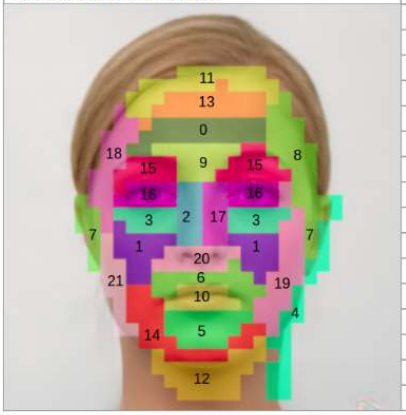
Labeled Face Clusters	Action Units	Affected Clusters
	AU01: Inner Brow Raiser	20, 15, 2, 9
	AU02: Outer Brow Raiser	15, 16, 9, 0
	AU04: Brow Lowerer	15, 16, 9, 20
	AU05: Upper Lid Raiser	16, 3, 15, 20
	AU06: Cheek Raiser	10, 6, 20, 5
	AU07: Lid Tightener	16, 20, 10, 3
	AU09: Nose Wrinkler	20, 6, 10, 2
	AU10: Upper Lip Raiser	10, 20, 6, 5
	AU12: Lip Corner Puller	10, 6, 20, 5
	AU14: Dimpler	20, 10, 6, 2
	AU15: Lip Corner Depressor	10, 20, 5, 6
	AU17: Chin Raiser	10, 5, 20, 6
	AU20: Lip Stretcher	10, 6, 20, 5
	AU23: Lip Tightener	10, 6, 5, 20
	AU25: Lips Part	10, 6, 20, 5
	AU26: Jaw Drop	20, 10, 5, 6

Table 4.1: Face clusters obtained by spherical k-means clustering. Affected clusters column presents top four relevant clusters for each action unit computed with Equation 4.2.

region - channel relationship with a slight modification on Equation 4.1,

$$\mathbf{R}_{k,c} = \sum_{n} \left(\frac{\sum_{h,w} (\mathbf{U}_{n,k,h,w} \odot (\mathbf{A}_{n,c,h,w})^2)}{\sum_{h,w} \mathbf{U}_{n,k,h,w}} \right) \tag{4.3}$$

which outputs the contribution of each channel towards each semantic cluster.

Collins *et al.* [Collins et al., 2020] perform local editing by interpolating activation maps of the source and target images at a certain layer with slope of interpolation defined by the tensor of region - channel relationships $\mathbf{R}_{k,c}$. Assigning more weight to the relevant channels that affect a region results in transferring local attributes from source image to the target image [Collins et al., 2020]. To demonstrate localization of convolutional channels, we visualize activation maps that are most relevant to three regions: mouth, eyes and nose in Figure 4.3. Just by visual inspection, it can be seen that even the most responsible channels for a region have high activation on parts of the face that are not relevant. Thus performing local editing by interpolating the activation maps as a whole results in either editing parts of the face that were not intended, or very weak editing results where the change is hard to observe. Editing results by GAN Local Editing are shown in Figure 4.4, where we can observe the loss in the target subject's identity because of unlocalized attribute transfer.

When we perform local editing via manipulation of activation maps, we do not

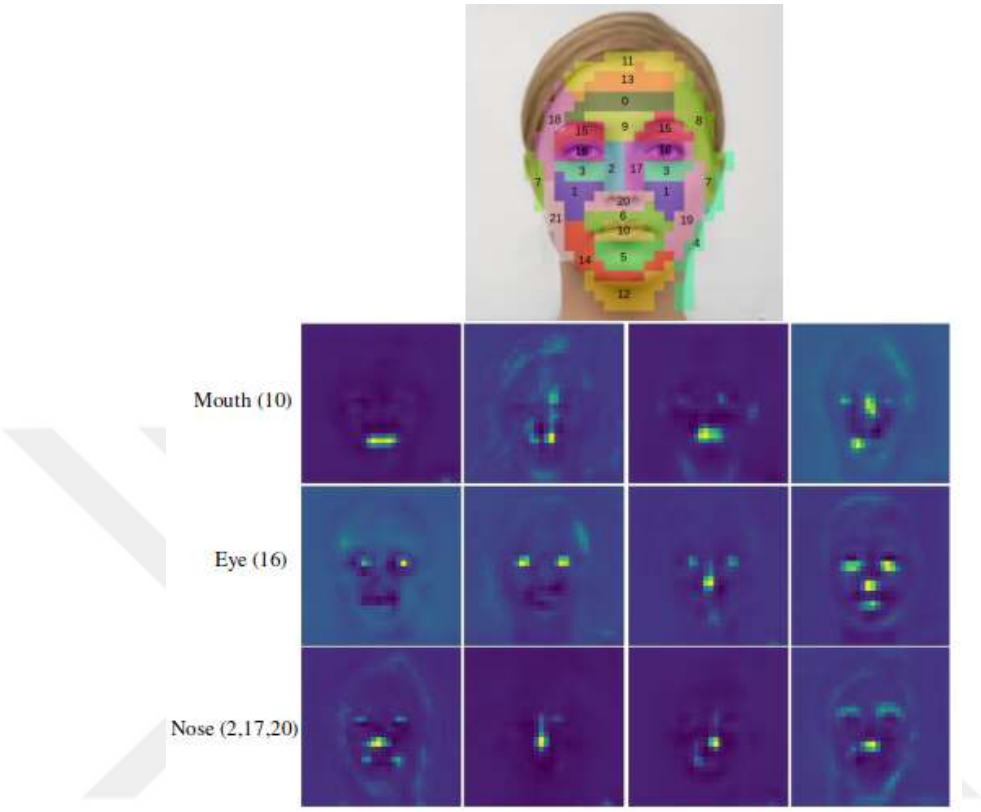


Figure 4.3: Four most relevant channel outputs for the mouth, eye, and nose regions.

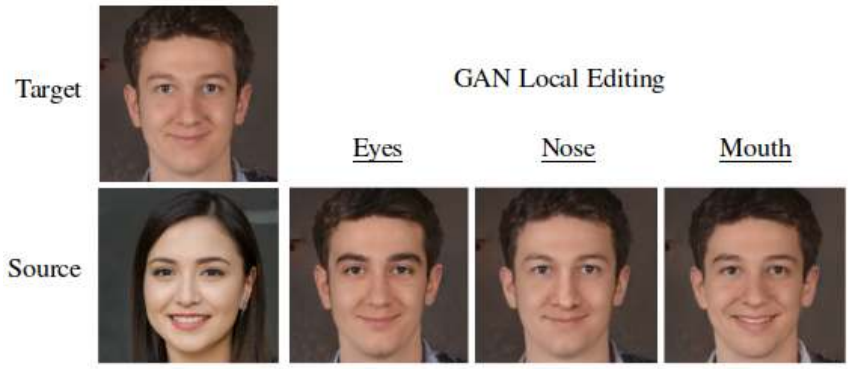


Figure 4.4: Results by GAN Local Editing [Collins et al., 2020]

interpolate activation maps as a whole, but transfer the patch of corresponding region at every channel in order to avoid transfer of attributes outside the region of interest or loss of intensity of attributes at a local region. The difference between two methods are presented at Chapter 5.

Chapter 5

COMPARISONS AND PERCEPTUAL USER STUDIES

We compare our editing results with the state-of-the-art facial expression editing models for three main tasks: local editing, facial expression transfer and emotion editing. Additionally, we perform a perceptual study to see the difference between evaluation of editing results on familiar and unfamiliar faces. Results show that people are more sensitive to editing quality on familiar faces, which affects the model's overall assessment.

5.1 Local Editing

For local editing in the activation space, we compare our results with GAN Local Editing by Collins *et al.* [Collins et al., 2020]. We use the same technique to find local regions on the face by applying spherical k-means clustering on the activation maps of a certain layer. After determining the regions, they compute the contribution of each channel to the region of interest and interpolate activation maps by giving more weight to the relevant ones. As discussed in Section 4.2, we show that most activation maps do not focus on specific semantic regions, but influence multiple locations on the face. It can be seen in Figure 4.3 that even the most relevant channels for a region affect irrelevant parts of the face. So transferring attributes by selecting activation maps based on their contribution to a region and discarding seemingly unrelated activation maps result in either transferring attributes from outside the region of interest or losing the intensity of attributes in the region. Based on our analysis of the activation space, we suggest that channels need to be considered cumulatively in StyleGAN. Thus we transfer local patches cut out from all activation maps of a layer from the source image to the target image. Comparison results are presented in Figure 5.1. It should be noted that unlike GAN Local Editing, our purpose is

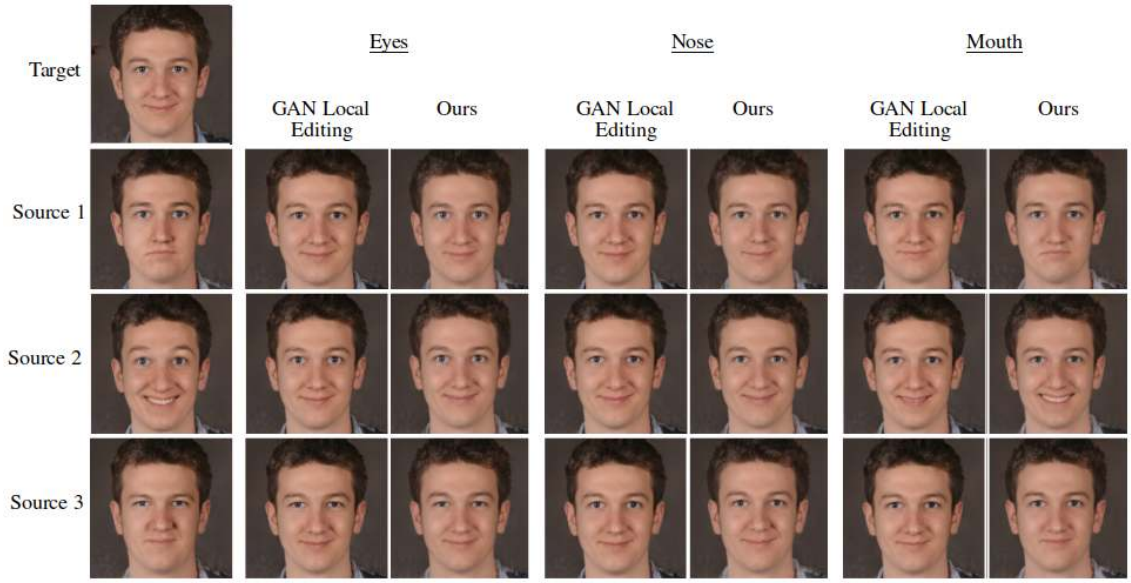


Figure 5.1: Comparison of local editing results. We transfer attributes from the local regions eyes, nose and mouth of source images to the target image using GAN Local Editing [Collins et al., 2020] and our local editing method. Especially in nose and mouth regions, it can be seen that GAN Local Editing cannot fully transfer the attributes. For example, the mouth region of Source 2 is poorly transferred with GAN Local Editing, while our approach transfers the same mouth to the target without affecting other regions on the face.

to transfer local attributes between original and edited images of the same person. Using our method for transferring local attributes between different identities would produce artifacts because of differences such as dimensions of the face or skin color. In order to transfer attributes between different identities, we would suggest using a combination of both approaches, i.e. interpolating activation maps and patching the region of interest, so that undesired changes do not appear at irrelevant parts of the face as in Figure 4.4.

5.2 Facial Expression Transfer

For facial expression transfer, we first predict action unit values of the target image using the action unit predictor. Then we update the latent vector corresponding to the input expression via gradient descent to match the target expression. Results can be seen in Figure 5.2. We performed a user study with 23 participants to compare

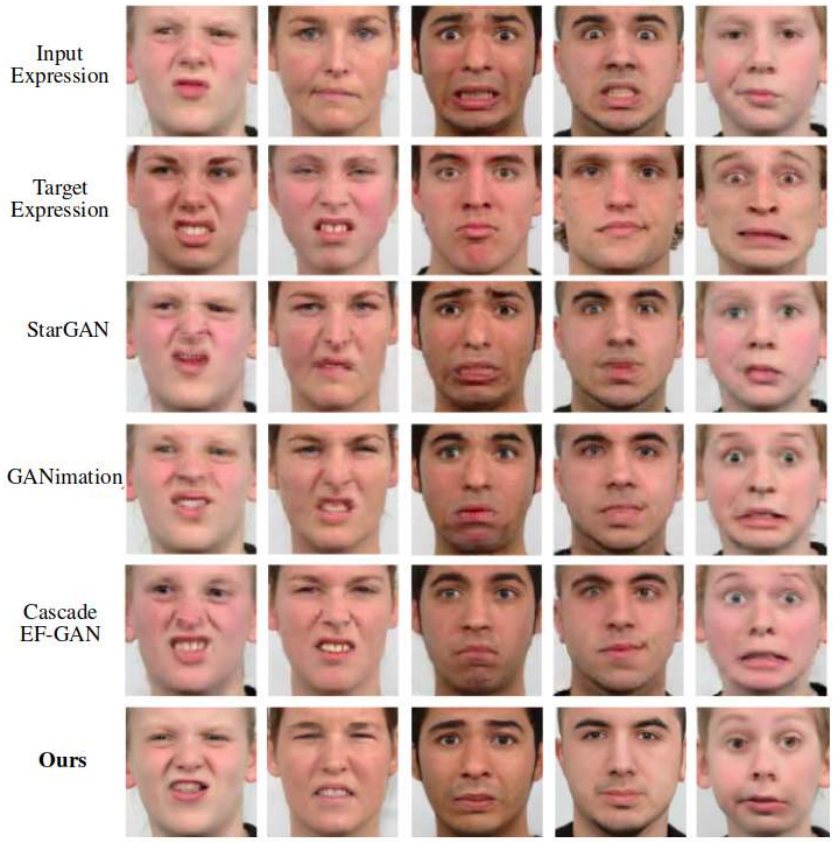


Figure 5.2: Comparison of facial expression transfer results. According to the user study (Figure 5.3), our method produces much less artifacts and preserves the identity of the input image better. Our method also outperforms others, except Cascade EF-GAN, in expression transfer.

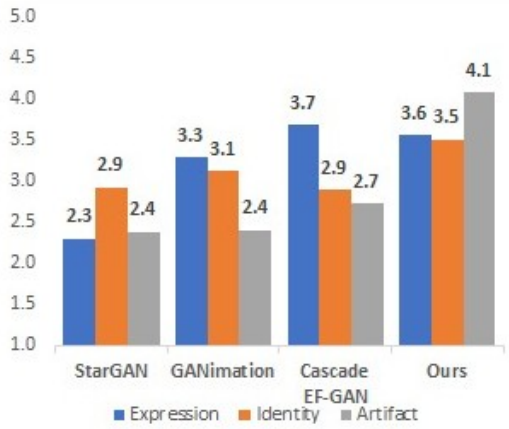


Figure 5.3: Perceived artifacts are significantly less in our model compared to others ($p<0.001$). Identity preservation is also rated better ($p<0.05$). For expression transfer, our model performs significantly better than StarGAN ($p<0.01$) and GANimation ($p=0.11$). Although Cascade EF-GAN performs better in expression transfer, the difference is not significant ($p=0.43$).

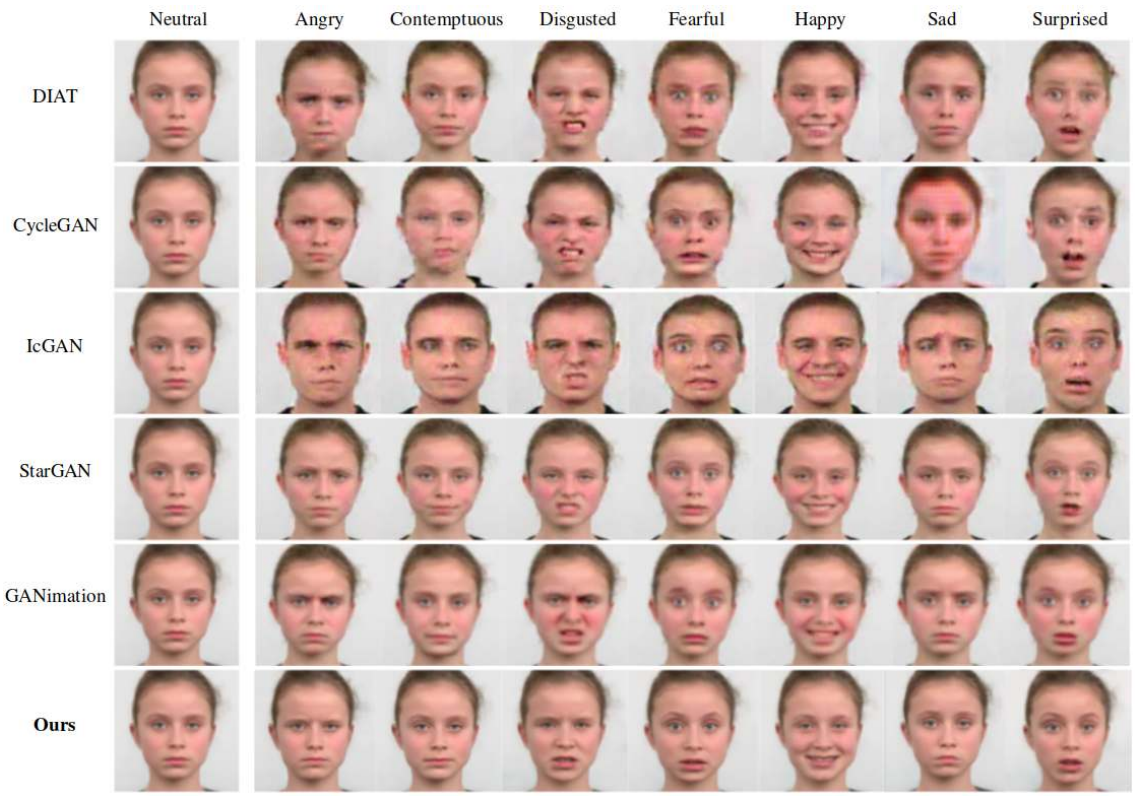


Figure 5.4: Comparison results on emotion editing. Our method produces much less artifacts than others and successfully synthesizes the correct emotion while preserving the identity.

our editing results with StarGAN [Choi et al., 2018], GANimation [Pumarola et al., 2019] and Cascade EF-GAN [Wu et al., 2020] models on facial expression transfer. We asked participants to evaluate edited images on three conditions: success in facial expression transfer, preservation of identity, and lack of artifacts. Resulting ratings on Likert scale are provided in Figure 5.3. The results show that our method outperforms others significantly for absence of artifacts on the edited image and preservation of identity from the input image. For expression transfer, our method performs significantly better than StarGAN and GANimation.

5.3 Emotion Editing

To reproduce the emotions, we refer to studies on FACS [Ekman, 1997] to understand which action units should be activated for each emotion. We then perform combined

editing of action units by boosting the intensity of the relevant ones. Comparison results with the models StarGAN [Choi et al., 2018] and GANimation [Pumarola et al., 2019] are shown in Figure 5.4. It should be noted that although our method is not originally designed for expression transfer or emotion editing, it allows us to manipulate multiple action units at once with varying intensities to perform high-fidelity expression editing.

5.4 Perceptual User Study

Existing works on expression editing primarily resort to two means of providing evidence in support of the success of their methods: 1) they present edited versions of faces from arbitrary people, and appeal to the subjective opinion of the reader, 2) they report metrics based on automated expression recognition accuracies, PSNR, FID scores, etc [Ding et al., 2018, Wu et al., 2020, Song et al., 2018].

Source	Dependent Var.	df	F	p	$\eta^2_{partial}$
Subject	Expression	3	4.281	0.006	0.053
	Identity	3	4.524	0.004	0.056
	Artifact	3	8.894	<0.001	0.105
	Overall	3	6.422	<0.001	0.078
Editing	Expression	2	0.485	0.616	0.004
	Identity	2	6.384	0.002	0.053
	Artifact	2	1.925	0.148	0.017
	Overall	2	4.381	0.014	0.037

Table 5.1: 4x3 Two-Way Repeated MANOVA - Multivariate Tests

However, proper assessment of how well a generated facial display portrays a particular expression is inherently a perceptual task that requires human appraisal. Unfortunately, there are no established guidelines, or protocols on how this should be done within the literature on facial expression editing. In this section we fill an important gap in the literature through a perception-based user study. Our study

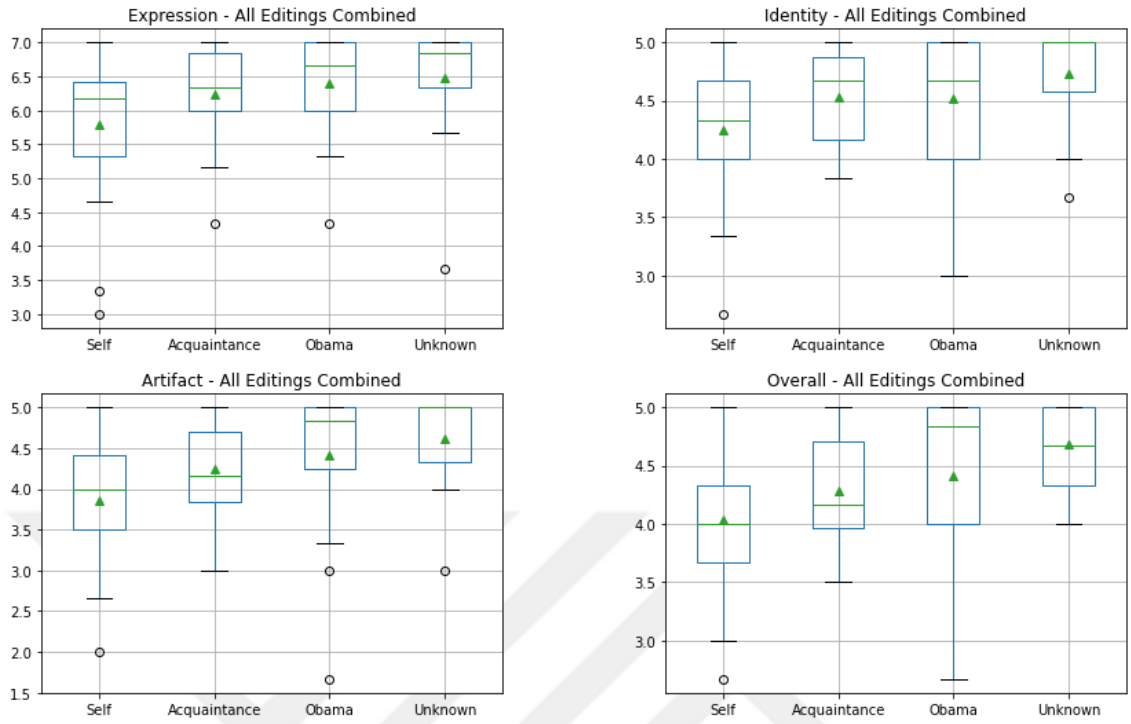


Figure 5.5: Distribution of participant ratings for each question. T-tests show that there is no significant difference between ratings for the acquaintances, thus the analyses for acquaintances are combined.

moves us closer to formalizing what perceptual aspects of model outputs should be assessed, and how they should be assessed.

A successful editing operation should portray the target expression well, while preserving the identity of the subject. It should also be free of editing artifacts. Therefore, the perceptual assessment protocol that we prescribe resorts to human appraisals for quantifying success along the criteria listed above. Furthermore, we hypothesize that sensitivity of a rater's appraisal depends strongly on the level of rater's familiarity with the person whose facial image is being edited. Therefore, unlike the existing approaches, which resort to reader's assessment of results on arbitrarily selected faces, we collect subjective appraisals over a range of images whose familiarity to the rater ranges from "unknown to the rater" to "the rater himself/herself." This allows us to assess the practical utility of facial editing algorithms for a wide range of application scenarios, where the manipulated face may be familiar to the user.

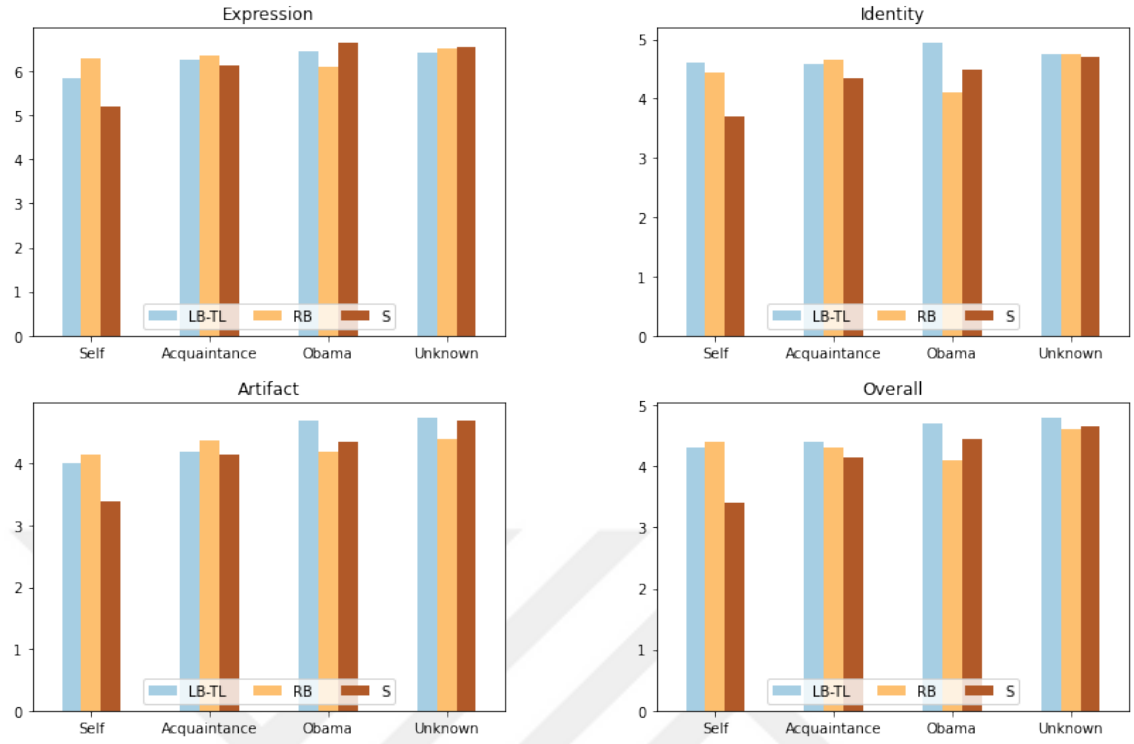


Figure 5.6: Mean of responses for each editing result. LB-TL: Lowered Brows and Tightened Lips, RB: Raised Brows, S: Smile.

Our aim is to show that people are more attentive and sensitive in evaluating the quality of facial expression editing on their own images or on images of people they are acquainted with. We performed a perceptual study with 20 participants in order to assess the success of our model, and to measure the effects of familiarity on subject sensitivity to editing artifacts, shift in identity, and portrayal of the intended expression. Participants were asked to submit a image of their own and two of their acquaintances with a neutral expression. We selected two additional subjects to present each participant: Obama and an unknown female subject. Thus each participant was provided with a custom survey on images of 5 subjects: an image of their own, two images of their acquaintances, an image of someone they are familiar with from the media but have not met in person, and an image of an unknown subject. Three distinct editing operations were performed on each image: Lowered brows and tightened lips (LB-TL), raised brows (RB), and smiling (S). Smiling expression is a combination of cheek raiser, lip corner puller, and dimpler.

For each editing result, participants were asked to answer four questions with the following response scales:

- How satisfied are you with the portrayal of the intended facial expression in the edited image?
  7. Very satisfied
  6. Satisfied
  5. Slightly satisfied
  4. Neither satisfied nor dissatisfied
  3. Slightly dissatisfied
  2. Dissatisfied
  1. Very dissatisfied

- How well is the identity preserved in the edited image?
  5. Very well
  4. Well
  3. Moderately well
  2. Slightly well
  1. Not well at all

- What is the amount of artifacts in the edited image?
  5. None at all
  4. A little
  3. A moderate amount
  2. A lot
  1. A great deal

- Overall, how well is the editing result? (*Linear scale*)
  5. Extremely well
  1. Not well at all

First 4x3 Two-Way Repeated MANOVA test is performed on the results. 4 subjects (Self, Acquaintance, Obama, Unknown) and 3 types of editing (LB-TL,

Subjects		p-values			
i	j	Expression	Identity	Artifact	Overall
Self	Acq.	**0.04**	**0.022**	**0.019**	**0.027**
	Obama	**0.001**	0.061	**0.002**	**0.015**
	Unknown	**0.01**	**0.002**	**<0.001**	**<0.001**
Acq.	Self	**0.04**	**0.022**	**0.019**	**0.027**
	Obama	0.257	0.95	0.21	0.308
	Unknown	0.105	**0.037**	**0.005**	**0.003**
Obama	Self	**0.001**	0.061	**0.002**	**0.015**
	Acq.	0.257	0.95	0.21	0.308
	Unknown	0.621	0.114	0.104	0.096
Unknown	Self	**0.01**	**0.002**	**<0.001**	**<0.001**
	Acq.	0.105	**0.037**	**0.005**	**0.003**
	Obama	0.621	0.114	0.104	0.096

Table 5.2: Pairwise comparisons for subjects. Mean difference is significant at p=0.05 level.

Editing		p-values	
i	j	Identity	Overall
LB-TL	RB	0.252	0.339
	S	**0.008**	**0.02**
RB	LB-TL	0.252	0.339
	S	0.242	0.317
S	LB-TL	**0.008**	**0.02**
	RB	0.242	0.317

Table 5.3: Pairwise comparisons for editing types. Mean difference is significant at p=0.05 level. Results for the expression and artifact questions are not significant, thus are not reported. LB-TL: Lowered Brows and Tightened Lips, RB: Raised Brows, S: Smile.

RB, and S) are set as independent variables, and the questions (Expression, Identity, Artifact, Overall) are set as dependent variables. Results indicate a significant effect of subjects on each question. Editing types have significant effect on the identity and overall questions (see Table 5.1).

Figure 5.5 shows distribution of responses given for each question. It can be seen that participants mainly rated their own images lower than others. Also, the unknown subject is generally rated higher. Pairwise comparisons (Table 5.2) show that the differences between participant responses to their own images and to others' are significant for each question (except for Self-Obama pair for the identity question). Results also indicate that the ratings for the unknown subject are significantly lower than ratings for self and acquaintances. On the other hand, responses to Obama's image have no significant difference from responses to acquaintances or the unknown subject. From these results, we can infer that people are more sensitive in evaluating their own images and images of their acquaintances than images of unknown subjects.

We further investigate whether editing types have an impact on the ratings. Figure 5.6 shows mean of responses for each question under each editing type. It can be seen that images with a smiling expression are generally rated lower than those with lowered brows tightened lips and raised brows. Pairwise comparisons for editings (Table 5.3) show that there is significant difference between ratings to lowered brows and tightened lips and smile in identity and overall questions. There are no significant differences to ratings of editing types in other questions.

Chapter 6

LIMITATIONS AND SOCIETAL IMPACT

Although using the pre-trained StyleGAN model has many advantages, such as access to a rich latent space and impressive generative power, it has limitations when working with real images. First, in order to work with real images their latent representations must be obtained. The best methods for latent encoding are iterative optimization-based approaches [Baylies, 2019], which are very time consuming. Second, facial images produced by the encoded representations preserve identity but can have a smoother skin, fewer wrinkles and facial marks. In our research, we paid special attention to achieve identity preservation while image encoding and action unit transfer, but preservation of finer details of the original face image is still a challenge.

At the time of development, StyleGAN2 was released. However, our assessment of the encoder (projector) for StyleGAN2 [Karras et al., 2020] found it to be particularly unsatisfactory in preserving subject identity. Hence, we continued to work with the initial StyleGAN and its encoder [Baylies, 2019]. Since its fundamentals are similar to the original StyleGAN, we would expect our approach to generalize to StyleGAN2. However, StyleGAN3 [Karras et al., 2021] takes a substantially different strategy for modeling latent and activation spaces, therefore efficacy of a solution based on it is an open research question.

The ability to control action units with the high degree of fidelity and precision afforded by our algorithm can have malicious applications in the form of deep fakes. However, it also opens up avenues for remarkable applications. In fact, the original motivator for our entire work has been the prospects of using it for rehabilitation of individuals suffering from Autism Spectrum Conditions, particularly those who have difficulty in producing facial expressions corresponding to specific emotional displays. Formative feedback in the form of an ASC sufferer's own face correctly

displaying emotions pushes the boundaries of formative feedback-based rehabilitation systems beyond the current state of the art [Schuller et al., 2015].

Chapter 7

CONCLUSION

In this paper we present methods for both global and local editing of facial images by utilizing the state-of-the-art StyleGAN model. The idea can be extended to any domain of editing or GAN architecture, with suitable labeling of datasets. Leveraging the latent properties and generative power of a pre-trained GAN architecture allows us to produce high-fidelity editing results without the burden of training a GAN from scratch. Our analysis of the latent and activation spaces of StyleGAN also provides valuable insights that add to our understanding of how action unit semantics are encoded. As a future work, we aim to utilize our approach for building rehabilitation systems for ASC sufferers to provide formative feedback in performing certain facial expressions correctly.

BIBLIOGRAPHY

[Abdal et al., 2019] Abdal, R., Qin, Y., and Wonka, P. (2019). Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441.

[Baltrušaitis et al., 2015] Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE.

[Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

[Baltrušaitis et al., 2015] Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6.

[Bau et al., 2018] Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

[Baylies, 2019] Baylies, P. (2019). stylegan-encoder. `https://github.com/pbaylies/stylegan-encoder`. Accessed: May 2020.

[Buchta et al., 2012] Buchta, C., Kober, M., Feinerer, I., and Hornik, K. (2012). Spherical k-means clustering. *Journal of statistical software*, 50(10):1–22.

[Chai et al., 2021] Chai, L., Wulff, J., and Isola, P. (2021). Using latent space regression to analyze and leverage compositionality in gans. *arXiv preprint arXiv:2103.10426*.

[Choi et al., 2018] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). 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.

[Collins et al., 2020] Collins, E., Bala, R., Price, B., and Susstrunk, S. (2020). Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780.

[Creswell and Bharath, 2018] Creswell, A. and Bharath, A. A. (2018). Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974.

[Ding et al., 2018] Ding, H., Sricharan, K., and Chellappa, R. (2018). Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[Du et al., 2014] Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462.

[Ekman, 1997] Ekman, R. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[Friesen and Ekman, 1978] Friesen, E. and Ekman, P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5.

[He et al., 2019] He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

[Karras et al., 2021] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. In *Proc. NeurIPS*.

[Karras et al., 2019] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405.

[Karras et al., 2020] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Langner et al., 2010] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388.

[Larsen et al., 2016] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR.

[Ledig et al., 2017] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-

realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

[Liu et al., 2019] Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., and Wen, S. (2019). Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3673–3682.

[Liu et al., 2017] Liu, M.-Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*.

[Liu et al., 2001] Liu, Z., Shan, Y., and Zhang, Z. (2001). Expressive expression mapping with ratio images. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 271–276.

[Ma et al., 2008] Ma, W.-C., Jones, A., Chiang, J.-Y., Hawkins, T., Frederiksen, S., Peers, P., Vukovic, M., Ouhyoung, M., and Debevec, P. (2008). Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Transactions on Graphics (TOG)*, 27(5):1–10.

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

[Nitzan et al., 2020] Nitzan, Y., Bermano, A., Li, Y., and Cohen-Or, D. (2020). Disentangling in latent space by harnessing a pretrained generator. *arXiv preprint arXiv:2005.07728*.

[Onal Ertugrul et al., 2019] Onal Ertugrul, I., Yang, L., Jeni, L. A., and Cohn, J. F. (2019). D-pattnet: Dynamic patch-attentive deep network for action unit detection. *Frontiers in computer science*, 1:11.

[Perarnau et al., 2016] Perarnau, G., Van De Weijer, J., Raducanu, B., and Álvarez,

J. M. (2016). Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*.

[Pumarola et al., 2019] Pumarola, A., Agudo, A., Martinez, A., Sanfeliu, A., and Moreno-Noguer, F. (2019). Ganimation: One-shot anatomically consistent facial animation.

[Qiao et al., 2018] Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., and Wang, H. (2018). Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*.

[Ruiz et al., 2015] Ruiz, A., Van de Weijer, J., and Binefa, X. (2015). From emotions to action units with hidden and semi-hidden-task learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3703–3711.

[Savran et al., 2008] Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Bosphorus database for 3d face analysis. In *European workshop on biometrics and identity management*, pages 47–56. Springer.

[Schuller et al., 2015] Schuller, B., Marchi, E., Baron-Cohen, S., Lassalle, A., O'Reilly, H., Pigat, D., Robinson, P., Davies, I., Baltrusaitis, T., Mahmoud, M., et al. (2015). Recent developments and results of asc-inclusion: An integrated internet-based environment for social inclusion of children with autism spectrum conditions. In *Proceedings of the of the 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2015) as part of the 20th ACM International Conference on Intelligent User Interfaces, IUI 2015*, pages 9–pages.

[Shen et al., 2020] Shen, Y., Gu, J., Tang, X., and Zhou, B. (2020). Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252.

[Song et al., 2018] Song, L., Lu, Z., He, R., Sun, Z., and Tan, T. (2018). Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 627–635.

[Susskind et al., 2008] Susskind, J. M., Hinton, G. E., Movellan, J. R., and Anderson, A. K. (2008). Generating facial expressions with deep belief nets. *Affective Computing, Emotion Modelling, Synthesis and Recognition*, pages 421–440.

[Suzuki et al., 2018] Suzuki, R., Koyama, M., Miyato, T., Yonetsuji, T., and Zhu, H. (2018). Spatially controllable image synthesis with internal representation collaging. *arXiv preprint arXiv:1811.10153*.

[Tripathy et al., 2020] Tripathy, S., Kannala, J., and Rahtu, E. (2020). Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3385–3394.

[Wang et al., 2003] Wang, H. et al. (2003). Facial expression decomposition. In *Proceedings ninth IEEE international conference on computer vision*, pages 958–965. IEEE.

[Wiles et al., 2018] Wiles, O., Koepke, A., and Zisserman, A. (2018). X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686.

[Wu et al., 2020] Wu, R., Zhang, G., Lu, S., and Chen, T. (2020). Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5021–5030.

[Xu et al., 2017] Xu, R., Zhou, Z., Zhang, W., and Yu, Y. (2017). Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*.

[Yang et al., 2011] Yang, F., Wang, J., Shechtman, E., Bourdev, L., and Metaxas, D. (2011). Expression flow for 3d-aware face component transfer. In *ACM SIGGRAPH 2011 papers*, pages 1–10.

[Zhang et al., 2005] Zhang, Q., Liu, Z., Quo, G., Terzopoulos, D., and Shum, H.-Y. (2005). Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on visualization and computer graphics*, 12(1):48–60.

[Zhou and Shi, 2017] Zhou, Y. and Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)*, pages 370–376. IEEE.