Face Modeling and Editing with Deep Neural Networks

by

Mohammad Amin ALIARI

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS IN INFORMATION TECHNOLOGY ENGINEERING
M.A.Sc.

MONTREAL, AUGUST 9, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

M. Eric Paquette, Thesis supervisor
Department of Software Engineering and Information Technology, École de technologie supérieure

M. Adrien Gruson, Chair, Board of Examiners
Department of Software Engineering and Information Technology, École de technologie supérieure

M. Carlos Vázquez, External Examiner
Department of Software Engineering and Information Technology, École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON JULY 17, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to take this moment to thank everyone who helped me during my master's program, especially my supervisor, Prof. Eric Paquette. He shared his knowledge, wisdom, and experience with me. He trusted me and gave me the freedom and guidance to explore new ideas and directions. I am truly honored that I was one of his students. Next, I want to thank Prof. Tiberiu Popa for being a great source of inspiration and knowledge. He always provided me with constructive feedback and valuable suggestions that improved the quality of my work. I also want to express my gratitude to the jury members for taking their time and agreeing to evaluate my thesis. Moreover, I want to acknowledge the great support and collaboration of our industry partners: Ubisoft and Mitacs. They funded this research and provided me with the opportunity to work on real-world problems with state-of-the-art technologies, tools, and resources. I am immensely grateful to the research team at Ubisoft. Namingly, Olivier Pomarez, Andre Beauchamp, and Abdallah Dib. They were very helpful and friendly and shared their expertise and insights with me. They also created an enjoyable research environment where I learned a lot and had fun.

On a personal note, I want to thank my parents and two sisters for their unconditional love and support throughout my life. My mother deserves special praise for her constant encouragement to follow my passion. Last but not least, I want to mention my beautiful home country, Iran. The place where I was born and raised. The place that I will always be proud of and will always be in my heart.

Modélisation et édition de visages avec réseaux de neurones profonds

Mohammad Amin ALIARI

RÉSUMÉ

Nous proposons une approche basée sur les réseaux génératifs profonds pour l'édition interactive de visages 3D. La plupart des méthodes actuelles pour l'édition de visages se basent sur des méthodes linéaires et ne peuvent pas exprimer de déformations complexes et non linéaires. Par opposition aux modèles 3D déformables basés sur l'analyse en composantes principales (ACP), nous proposons une nouvelle architecture basée sur les autoencodeurs variationnels. Notre architecture a plusieurs encodeurs (un par partie du visage, comme le nez et la bouche) qui sont reliés à un seul décodeur. En conséquence, chaque sous-vecteur du vecteur latent représente une partie du visage. Nous entraînons notre modèle avec une nouvelle fonction de coût, qui désintrique l'espace latent selon les différentes parties du visage. La sortie du réseau est un nouveau visage. Ainsi, contrairement aux méthodes par partie basées sur l'ACP, notre modèle apprend intrinsèquement à regrouper les parties et ne requiert pas de processus additionnel pour joindre les parties du visage. Pour permettre l'édition interactive du visage, nous optimisons les variables latentes selon des contraintes positionnelles sur des sommets, qui sont fournies par l'utilisateur. Pour éviter les changements globaux ailleurs sur le visage, nous optimisons seulement le sous-ensemble du vecteur latent qui correspond à la partie du visage qui est modifiée. Notre optimisation d'édition converge en moins d'une seconde. Nos résultats montrent que l'approche proposée supporte un large éventail de contraintes d'édition et génère des visages 3D plus réalistes. Finalement, nous explorons l'idée d'ajouter des textures aux visages générés puisque ceci complémente notre modèle génératif et le rend plus versatile.

Mots-clés: infographie, modélisation de surface, maillages de polygones, réseaux de neurones profonds, synthèse de texture

Face Modeling and Editing with Deep Neural Networks

Mohammad Amin ALIARI

ABSTRACT

We propose an approach for interactive 3D face editing based on deep generative models. Most of the current face modeling methods rely on linear methods and cannot express complex and non-linear deformations. In contrast to 3D morphable face models based on Principal Component Analysis (PCA), we introduce a novel architecture based on variational autoencoders. Our architecture has multiple encoders (one for each part of the face, such as the nose and mouth) which feed a single decoder. As a result, each sub-vector of the latent vector represents one part. We train our model with a novel loss function that further disentangles the space based on different parts of the face. The output of the network is a whole 3D face. Hence, unlike part-based PCA methods, our model learns to merge the parts intrinsically and does not require an additional merging process. To achieve interactive face modeling, we optimize for the latent variables given vertex positional constraints provided by a user. To avoid unwanted global changes elsewhere on the face, we only optimize the subset of the latent vector that corresponds to the part of the face being modified. Our editing optimization converges in less than a second. Our results show that the proposed approach supports a broader range of editing constraints and generates more realistic 3D faces. Finally, we explore the idea of adding textures to the generated faces as it can complement our generative model and make it more useful.

Keywords: computer graphics, shape modeling, mesh geometry models, deep neural networks, texture synthesis

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

3DMM	3D Morphable Face Model
AE	Autoencoder
AI	Artificial Intelligence
DNN	Deep Neural Network
GAN	Generative Adversarial Network
LLM	Large Language Model
PCA	Principal Component Analysis
VAE	Variational Autoencoder
VRAM	Video Random Access Memory

INTRODUCTION

Face modeling and editing have been very active topics in computer vision and graphics. It has a wide range of applications in multiple contexts, such as video games, film, visual effects, and the metaverse. In the context of the metaverse and video game production, especially in open-world games, artists need to generate and edit new 3D meshed faces and textures at a large scale, and this with a simple and intuitive interface. Current commercial tools (e.g., ZBrush®) for manual 3D face modeling and editing require a lot of expertise to achieve the desired output.

Our work focuses on developing an approach for generating new faces at a large scale while providing an intuitive editing capacity for the artists. This will have a considerable impact on the time spent by artists on the task.

Over the past 20 years, linear generative models (Blanz & Vetter, 1999) have been the dominant technology for face generation because of their simplicity and efficiency. With such generative models, a new face is generated by a linear combination of an orthogonal basis. This basis is obtained by applying a statistical analysis (Principal Component Analysis – PCA) on a set of face scans. PCA-based linear models suffer from multiple drawbacks. Their control mechanism is not intuitive. Also, the generated faces are generally bound by the statistical prior space; If the parameters are varied a little, the generated faces are believable but rather similar to the ones in the dataset, and if the parameters are varied a lot, we get novel faces, but they are not necessarily realistic. In other words, these methods tend not to generalize very well. Furthermore, local face editing is not possible with most PCA-based methods because the orthogonal basis does not provide a semantically meaningful separation of different parts of the face. More recently, non-linear generative models (Ranjan *et al.*, 2018) based on Deep Neural Networks (DNN) have emerged and achieved a better generalization capacity than PCA-based methods. However, it is challenging to balance the realism of the face with the requirement of allowing the generation of

a large variety of faces. Moreover, providing an intuitive user interface for local face editing remains a big challenge.

In this work, we propose a novel approach for face modeling that enables users to generate a wide variety of realistic faces using a simple and intuitive user interface. We frame the problem as an optimization problem over the latent space of a graph-based variational autoencoder (VAE) (Kingma & Welling, 2014) that incorporates both a set of part-based encoders as well as a global face decoder. This combination is key to allowing a large variety of faces as well as maintaining local user control. The part-based encoders allow for more flexibility in the generated result as well as local user control, while the global decoder ensures the realism of the result. This formulation also allows the user to perform direct vertex manipulation as an editing paradigm. Furthermore, our method runs at interactive rates, taking under one second to generate a new face after user interaction. Our novel contributions are summarized as:

- A graph convolutional VAE architecture with a contrastive loss function designed to disentangle the latent space into local parts;
- An approach to editing the face through an optimization of the latent variables enforcing the locality of the edit.

Compared to state-of-the-art methods, our approach has an improved generalization in contrast to PCA-based methods and a better locality of the editing compared to DNN methods.

As an additional step, we explore the idea of adding textures to the generated faces as it can complement our generative model and make it more useful for content creators. More specifically, we use a Style-GAN network (Karras, Laine & Aila, 2018) to generate albedo maps.

In Chapter 1, we review the current 3D Face modeling methods literature. Chapter 2 describes our novel deep generative model and then explains our method for face modeling. It also contains the details of our texture generation workflow. Chapter 3 presents our datasets' details, followed

by results, ablation studies, and comparisons to a number of new and classic methods. In the final section of this chapter (Sec. 3.5), we run an experiment to see how our randomly generated faces would look when textured with albedo maps. Chapter 4 discusses our method's limitations and the areas where it needs improvements. Finally, we summarize the content presented in this thesis and make several suggestions for possible future work.

I would like to express my sincere appreciation and gratitude to our industry partners: Ubisoft and Mitacs, who kindly supported and enabled this work. I am also thankful to the research team members at Ubisoft, namely Olivier Pomarez, Andre Beauchamp, Abdallah Dib, and others. Their expertise and assistance were essential for the success of this work.

A version of this thesis, excluding Sec. 2.8 and Sec. 3.5, was published in Computer Graphics Forum (Aliari, Beauchamp, Popa & Paquette, 2023). It was also presented at the Eurographics 2023 conference. I was the first author of that paper. My key contributions were writing the manuscript, designing the neural network, preprocessing the datasets, training the models, and developing a method for neural face editing. I also carried out the ablation studies, comparisons, and evaluations. While working on the method, I discussed my ideas with Prof. Eric Paquette, Prof. Tiberiu Popa, and Andre Beauchamp from Ubisoft. I used their supervision and technical knowledge to explore new research topics and methods to improve my model. After I wrote the initial version of the paper, I further improved the manuscript by following Prof. Eric Paquette and Prof. Tiberiu Popa's comments and feedback.

CHAPTER 1

LITERATURE REVIEW

This chapter provides an overview of various methods and techniques used for 3D shape representation, 3D face modeling, and editing. In addition, we introduce some of the neural network-based approaches that are commonly used for texture generation. This field is also related to our work since we use texture generation in Sec. 2.8 and Sec. 3.5 to color 3D faces generated by our model.

1.1 Image Synthesis with Generative Artificial intelligence

Image synthesis is a widely explored topic in the field of generative AI. Surveys such as the work of Zhan, Yu, Wu, Zhang & Lu (2021) explain most of the commonly used methods in detail. In short, many of the models rely on 2D convolution; the kernel weights are the training parameters, and the local features of the image are learned by applying several cascading 2D convolutions. For example, autoencoders (Rumelhart, Hinton & Williams, 1986) use this method to compress the input image to a lower-level representation called the latent vector. Then they decompress it to reconstruct the initial image. However, since the loss function is focused on reconstruction, the model is less capable of generative tasks. Variational autoencoders (VAEs) (Kingma & Welling, 2014) address this issue by adding a prior to the training loss. The prior forces the latent space distribution to be closer to the normal distribution. Thus, it makes the latent space more meaningful and makes it easier to generate new images. Despite this, VAEs tend to produce blurry images and do not work well on higher-resolution images. Solutions based on generative adversarial networks (GANs) (Goodfellow *et al.*, 2014), such as StyleGAN (Karras *et al.*, 2020), are a newer take on this problem. They use a generator-discriminator setup where the generator tries to make plausible images that would convince the discriminator that they come from the original dataset. This architecture has proven to be a robust solution for image synthesis. It has been widely used on many categories of images to produce high-quality and high-resolution results. Diffusion model (Ho, Jain & Abbeel, 2020) is another recent advancement in the field of generative AI. It is a probabilistic model for high-quality image synthesis with performance

comparable to GAN-based models. The model is trained by adding noise to the input image and learning how to recover the initial input from noise. Stable diffusion (Rombach, Blattmann, Lorenz, Esser & Ommer, 2021) is a famous open-source model based on diffusion models. It is a text-to-image generative model, so its key feature is that it takes text prompts as the context for image generation. At inference, the input text prompt is encoded by a language model and is mixed with a noise patch. Then, this encoded mix is given to the model decoder to create the final image. Nevertheless, the model is primarily designed to be used with text descriptions and becomes less suitable for unconditional image generation. On top of that, the training process requires a large dataset and can become resource-demanding, and the inference is also slow. Therefore, while diffusion models are becoming more powerful and are starting to suppress state-of-art GAN-based models (Dhariwal & Nichol, 2021), they still have the mentioned issues. Figure 1.1 shows a summary about the discussed methods.

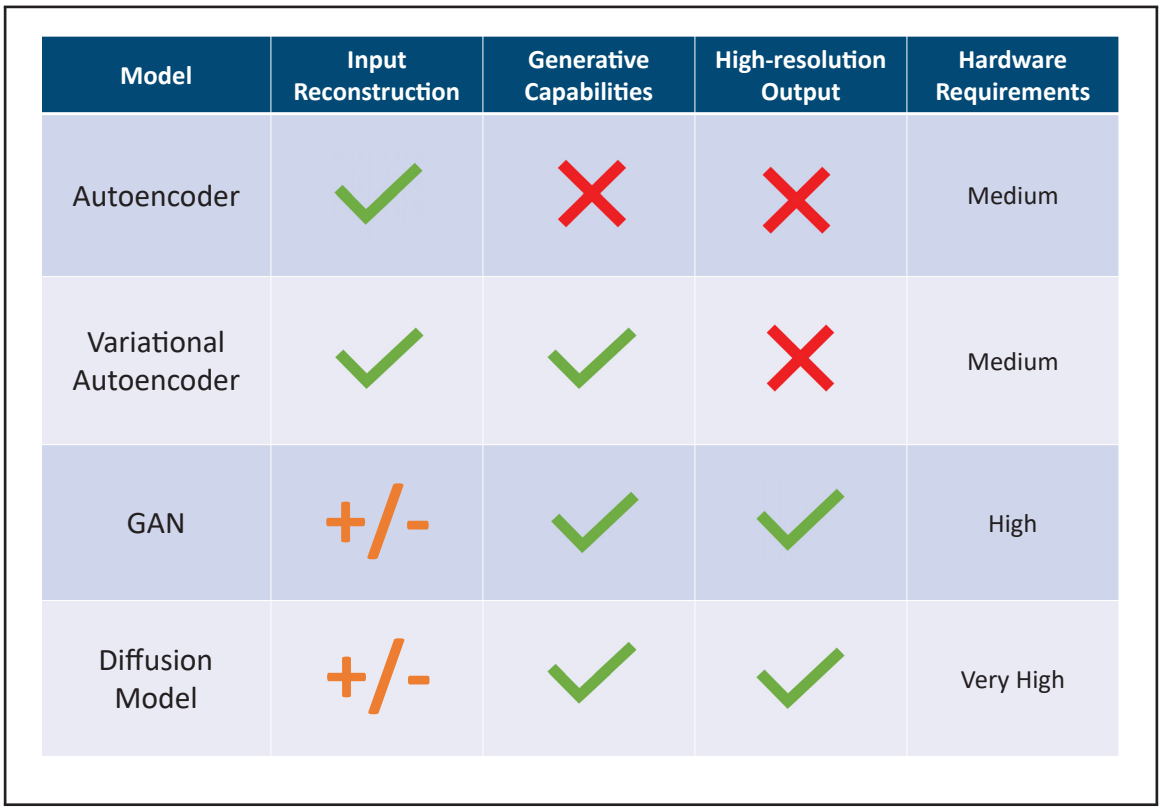

Figure 1.1 Summary of image synthesis methods

1.2 3D Shape Modeling and Editing

Classical 3D morphable face model (3DMM) methods as well as many recent ones (Cao, Weng, Zhou, Tong & Zhou, 2014; Booth, Roussos, Zafeiriou, Ponniah & Dunaway, 2016; Li, Bolkart, Black, Li & Romero, 2017; Booth, Roussos, Ponniah, Dunaway & Zafeiriou, 2018; Ploumpis, Wang, Pears, Smith & Zafeiriou, 2019; Ploumpis *et al.*, 2021; Ghafourzadeh *et al.*, 2020; Egger *et al.*, 2020) use PCA to sample the distribution of the face models. The popularity of PCA-based methods is due to their simplicity and efficiency. User control is achieved by optimizing for the eigenvector weights given vertex-based constraints, typically resulting in a linear system of equations. In many cases this is done globally on the entire face (Cao *et al.*, 2014; Booth *et al.*, 2016; Li *et al.*, 2017; Booth *et al.*, 2018) resulting in a lack of local control. Local control is less important when the application is global face reconstruction, but it is a critical limitation when the application is face editing. To address this limitation, Ghafourzadeh *et al.* (2020) propose a method that further decomposes the face into semantic parts (like in Figure 1.2) allowing independent generation for each part followed by an ARAP-inspired reassembling of the parts into one coherent face mesh.

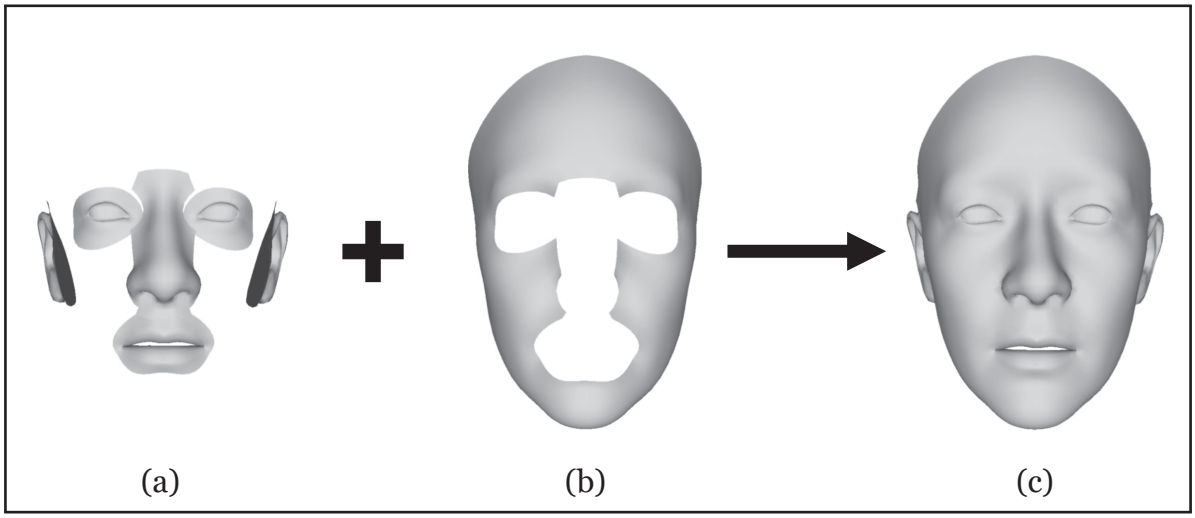

Figure 1.2 (a), (b): Face segmentation. (c): Blending the segments.
Taken from Ghafourzadeh *et al.* (2021)

Vertex-based editing was added to this part-based method (Ghafourzadeh *et al.*, 2021). Similarly to our approach, the user edits the face by moving mesh vertices. Alas, linear models such as PCA tend to perform poorly in the presence of large complex changes, and blending together local parts may result in uncanny effects, as shown in Figure 1.3.

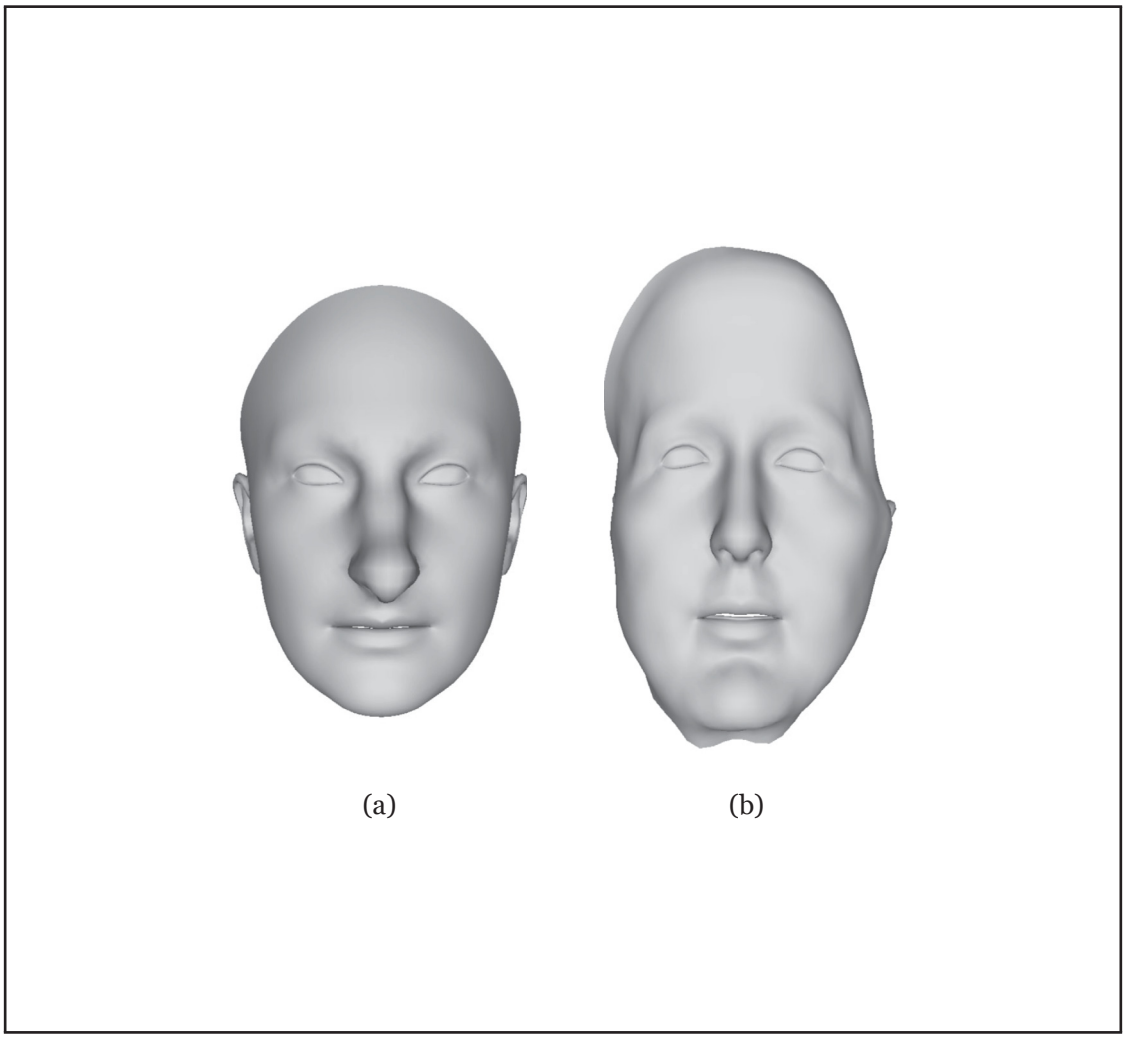

Figure 1.3 Examples of faces becoming odd-looking after editing their parts in isolation and blending them back together. Taken from Ghafourzadeh *et al.* (2021)

More recent neural methods leverage the generative power of deep learning by replacing the PCA with either autoencoders (Tan, Gao, Lai & Xia, 2018; Bagautdinov, Wu, Saragih, Fua & Sheikh, 2018; Bouritsas, Bokhnyak, Ploumpis, Bronstein & Zafeiriou, 2019) or Generative Adversarial

Networks (GAN) (Bouritsas *et al.*, 2019; Cheng *et al.*, 2019). Fernandez-Abrevaya (2020) uses a fully connected network along with UV representation of the vertex positions. Image representation of the geometry is particularly useful as many state-of-the-art convolutional neural network (CNN) models can be used with this type of data. However, reconstructing the final mesh from a UV mesh has its own problems. There are areas around the lips or the eyes where there is no data. As a result, those vertices would collapse to (0, 0, 0) coordinates. That is why Fernandez-Abrevaya (2020) uses the flattened representation as input of the GAN model while training the discriminator with the geometry map. Li *et al.* (2020) also use a UV representation and a hybrid method to address reconstruction issues. They use a combined linear and non-linear method. The face area, where there are reliable pixel values, is sampled directly, while the rest of the geometry is morphed using linear 3DMM deformation modes.

Vesdapunt, Rundle, Wu & Wang (2020) use modeling bones to represent the face. This representation benefits from the lower count of parameters due to the compactness of bone data. Their model learns the skinning weights to reconstruct the 3D faces. The predicted modeling bones can be used for face modeling. However, this approach needs a base model with modeling bones and initial hand-painted skinning. Also, the bones are directly transformed to deform the face, and neither higher-level control nor non-linear transformations exist.

Ranjan *et al.* (2018) use graph convolution to build their CoMA model. A novel mesh sampling operation is also used to capture details of the face at different levels. It also helps reduce the convolution input dimension so the model can converge more easily. Figure 1.4 shows how this is done in CoMA. As a result, CoMA and other works inspired by CoMA (Tan *et al.*, 2018; Li, Liu, Lai & Yang, 2019; Yuan, Li, Lai, Liu & Yang, 2019) can outperform linear models in face reconstruction tasks. Bagautdinov *et al.* (2018) use a variational autoencoder (VAE) architecture to model faces, but there are several fundamental differences, both conceptual and technical, between their work and ours. Their goal is high-fidelity 3D face reconstruction from images and video. As such, their VAE design matches this goal by globally reconstructing the faces using several latent spaces corresponding to a range from coarse to dense levels of detail. In contrast,

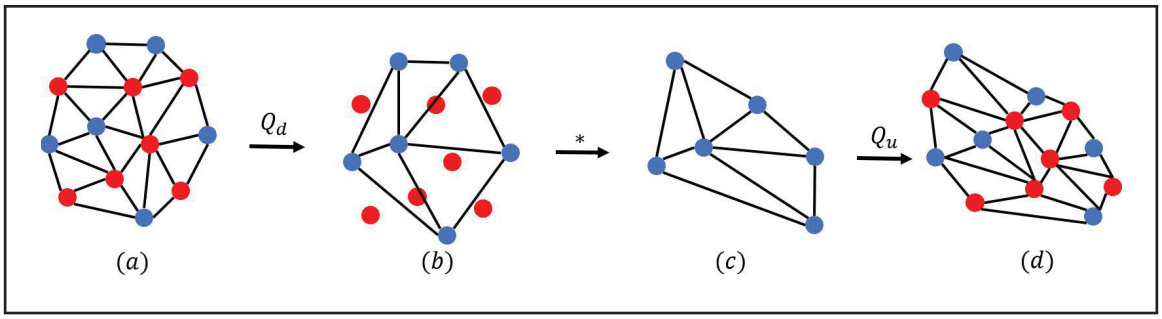

Figure 1.4 CoMA mesh sampling operation: Red vertices are removed from the mesh (a) to down-sample it. The vertices are selected in way that quadric error is minimized. The result is mesh (b) that is transformed by convolution to mesh (c). At (d), the removed vertices are added back and the mesh is up-sampled. Taken from Ranjan *et al.* (2018)

our goal is novel face synthesis through local editing, and, as such, we employ an architecture where we segment the face and use part-based encoders and a global decoder.

While the methods presented so far use triangular meshes as a surface representation, alternative representations like voxel based exist but are primarily suitable for lower-level detail or a specific category of 3D meshes. For instance, Guan, Jahan & van Kaick (2020) use voxelization to reconstruct and generate 3D desk and chair shapes.

Most of the work in DNN for faces has been about encoding the identity and expression of the face rather than enabling editing operators. Tan *et al.* (2018) use a set of expression labels with a conditional-VAE to achieve control. However, describing face modeling with labels is not feasible as we would need a much lower level of control over the face. Yang *et al.* (2020) propose a method that can reconstruct a rigged model from a single photograph, while Wang *et al.* (2022) and Bao *et al.* (2021) propose methods that can reconstruct a face model from RGB-D data. Very recently, Jung *et al.* (2022) proposed a DNN model to generate 3D caricatures. While this model works remarkably well for caricatures, it exhibits a global deformation behavior during editing that makes it difficult to control.

Another network architecture that has been very successful and widely used in the field of large language models (LLMs) is Transformers (Vaswani *et al.*, 2017). Due to the success of this

network architecture, its usage has been expanded to other fields as well. Namingly, Chandran, Zoss, Gross, Gotardo & Bradley (2022) have used it in the realm of 3D shape models. The authors present a non-linear parametric 3D shape model called Shape Transformer. Its most important features are not relying on a fixed mesh topology, supporting different mesh representations such as triangular and point clouds, and being capable of outputting high-resolution meshes. However, this model does not offer a way to edit the 3D shapes, let alone edit them locally. It is also very memory-consuming and slow to be used in an interactive application. Therefore, while the model suggests several new and interesting solutions in the field of mesh completion (filling the missing parts of the mesh) and marker-based facial performance capture, it has not been designed for applications where precise control over the generation process is important.

We propose an approach that outperforms PCA-based 3DMM methods in terms of generalization from the faces in the dataset. Furthermore, our approach outperforms current DNN methods in terms of application to face editing, in particular providing a local editing paradigm. Figure 1.5 shows a summary of the discussed methods with respect to the important criteria for our work.

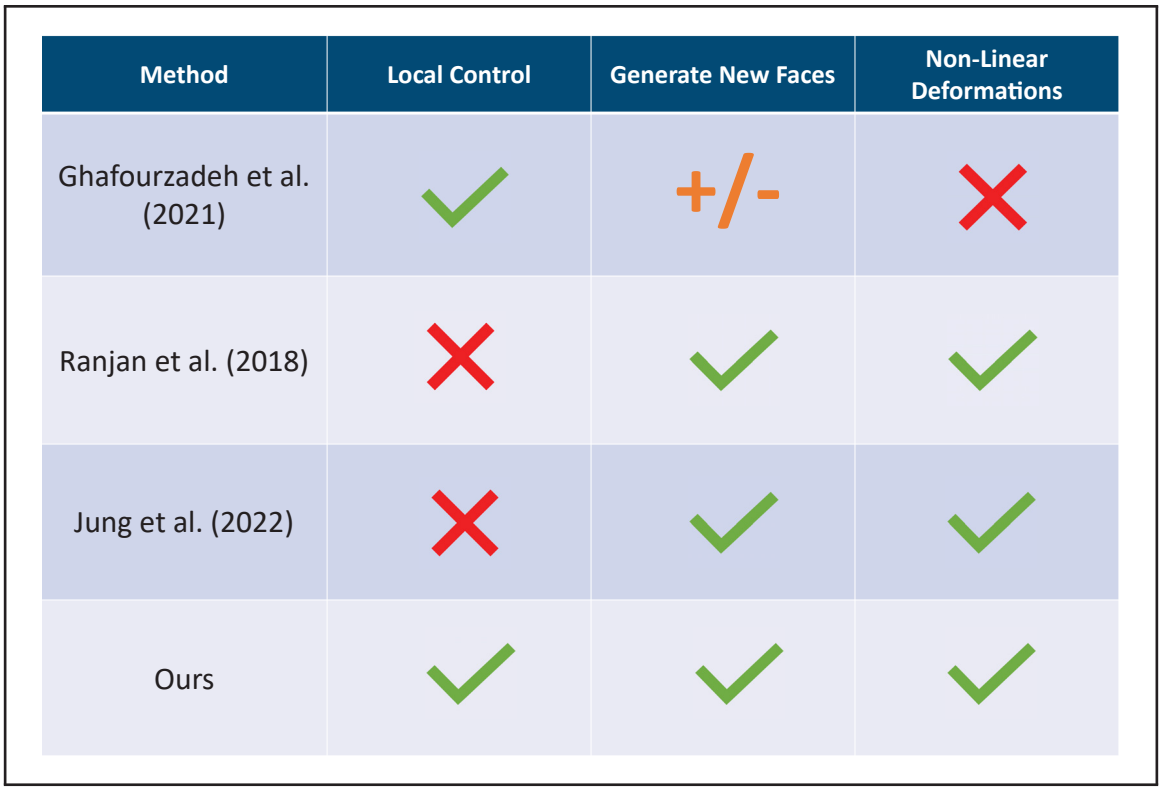

Figure 1.5 Summary of 3D shape modeling and editing methods

CHAPTER 2

METHOD

In this chapter, we present a DNN architecture to learn a compact latent space representation for 3D meshes of faces. The two main design objectives of our new architecture are to be able to disentangle different parts of the face and to allow local editing of the face mesh. For this, our key idea is to decompose the faces into different semantic parts and employ an autoencoder architecture where we feed each part to a separate encoder that produces a semantic latent vector for that part of the face. A single decoder is then used to aggregate all the latent vectors and produce the final mesh. The main intuition of having a separate embedding for each part of the face is to enforce local face editing by design. The use of a single decoder, that has access to all the embedding vectors, produces a realistic and consistent mesh. Our novel loss function enforces the disentanglement of the latent variables so that each has an influence on a localized region of the face. After training on a dataset of faces, we get a tailored latent space that enables multiple applications. It can be used to easily generate random faces with meaningful variations of facial characteristics, enabling character designers to generate a large set of facial assets. We also demonstrate that the latent space is very powerful in enabling the editing of the face through an optimization of the latent variables from user constraints in terms of dragging and dropping vertices. This provides a very intuitive interaction paradigm, and our latent space optimization outputs meaningful deformations of the face. This chapter introduces the face generator network, its training procedure, and our neural face editing approach (Figure 2.1).

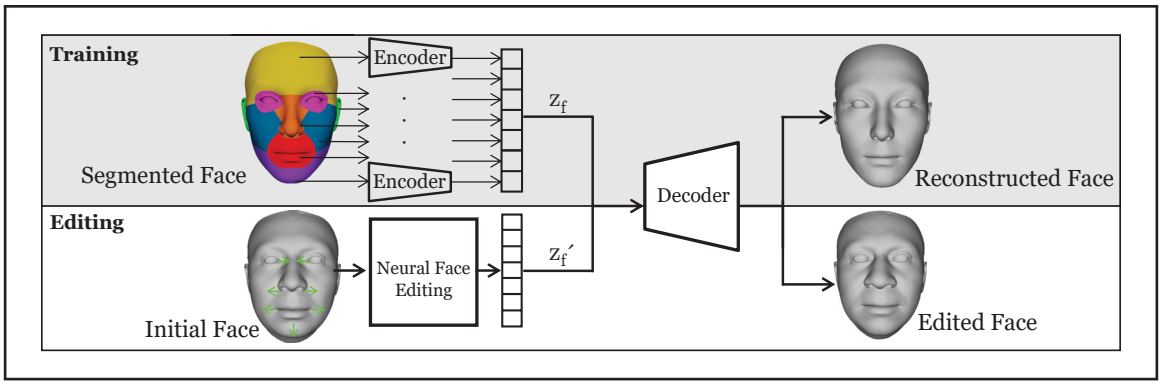

Figure 2.1 Our 3D face model. Training: We feed each segment of the face to its part encoder. Then, we merge and pass the encoded representation to a decoder that reconstructs the face. Editing: Our approach modifies the latent vector of the face based on user constraints

2.1 3D Face Generator

The first role of the generator is to learn a latent representation of the face that leverages the generalization capacity of the generative model to create new faces. For this goal, we introduce a network that encodes the face into a low-dimensional data representation. The second role of the generator is to learn a disentangled latent space where each group of latent variables is related to one specific part of the face. This encourages local and independent changes when the latent vector is modified.

2.2 Input and Output Data

The input of our network is a 3D face model represented as a 3D mesh. As we primarily use graph convolution operators, we represent the input as the canonical graph induced by the 3D mesh. For this reason, all of the faces in the dataset follow graph representation: $\mathcal{F} = (\mathcal{V}, \mathbf{A})$ with matrix $\mathcal{V} \in \mathbb{R}^{n\times 3}$, n vertices, and an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ showing the edge connections. Similarly to other methods (Blanz & Vetter, 1999; Ranjan *et al.*, 2018; Li *et al.*, 2019; Fernandez-Abrevaya, 2020; Ghafourzadeh *et al.*, 2020), we require that all faces be wrapped with one base head; they share an identical mesh topology (same triangles and vertex

connectivity). In addition, each face is segmented into seven different parts: forehead, eyes, ears, nose, cheeks, mouth, and chin. Each part $\mathcal{P}_i$ has n_i vertices $\mathcal{V}_i \in \mathbb{R}^{n_i \times 3}$. The segmentation is user-provided and shared by all faces. The output is the generated face with the same mesh topology and dimensions as the input face.

2.3 Network Architecture

We choose variational autoencoders (Kingma & Welling, 2014) as our generative model. As in the work of Ranjan *et al.* (2018), we use fast spectral convolutions (Defferrard, Bresson & Vandergheynst, 2016) along with their mesh sampling operation. The mentioned graph convolution uses kernel g_θ, which is parameterized with Chebyshev polynomials

$$g_\theta(L) = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}), \tag{2.1}$$

where K is the order of the polynomial, $\tilde{L}$ is the scaled Laplacian, $\theta \in \mathbb{R}^K$ is Chebyshev coefficients, and $T_k \in \mathbb{R}^{n \times n}$ is the Chebyshev polynomial order k that can be obtained recursively. The definition of spectral convolution then becomes:

$$y_j = \sum_{i=0}^{F_{in}} g_{\theta_{i,j}}(L) x_i \in \mathbb{R}^n, \tag{2.2}$$

where $x_i \in \mathbb{R}^{n \times F_{in}}$ is the input feature map, $F_{in} = 3$ is the number of input feature assuming it represents 3D vertex positions, and y_j is each of $y \in \mathbb{R}^{n \times F_{out}}$ features. For more detail, we can refer to the spectral convolution papers (Defferrard *et al.*, 2016; Ranjan *et al.*, 2018).

To achieve latent space disentanglement, we design an asymmetric model where each face part is fed to a separate encoder called *part encoder*. Specifically, each encoder uses two Chebyshev convolutional filters with $K = 6$ polynomials and dimensions of 16 and 32, respectively. K is the Chebyshev filter size, determining the number of hops from the central vertex. It can be seen as similar to the filter size in 2D image convolutions. Therefore, as we increase this radius, we add neighboring vertices from further distances to be covered by the kernel. In practice, a large value

of K can overly smooth the final output and produce a low-detail mesh. Then, we apply the ELU activation function (Clevert, Unterthiner & Hochreiter, 2016) to each filter output. Furthermore, we place a down-sampling layer of factor two between each filter. Next, similar to traditional VAEs, we flatten the output and use two fully connected layers with 8 neurons to transform the output to two 8-dimensional vectors $\boldsymbol{\mu}_i$, and $\boldsymbol{\sigma}_i$, which are the mean and standard deviation of the part $\mathcal{P}_i$. The details of each part encoder are shown in Figure 2.2. The initial vertex count of each part affects the dimensions of the following layers, and this is the difference between each part encoder. Finally, we sample the latent vector $\mathbf{z}_i \in \mathbb{R}^8$ from $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$. As for the last step to encode the face, we concatenate all the latent vectors into the final latent vector $\mathbf{z}_f \in \mathbb{R}^{56}$. We feed $\mathbf{z}_f$ to our decoder block to reconstruct the face. This block is built with a similar set of components. It starts with a fully connected layer that maps $\mathbf{z}_f$ to the appropriate dimension for the convolutional filters. We use three filters of dimensions 32, 16, and 3. Also, we use ELU and up-sampling layers of factor two between the filters. The decoder output $D(\mathbf{z}_f) \in \mathbb{R}^{n\times 3}$ is the final face with the same dimension as the number of vertices in the initial face. Consequently, it will learn to merge the parts into a whole face. The decoder architecture is also shown in Figure 2.2.

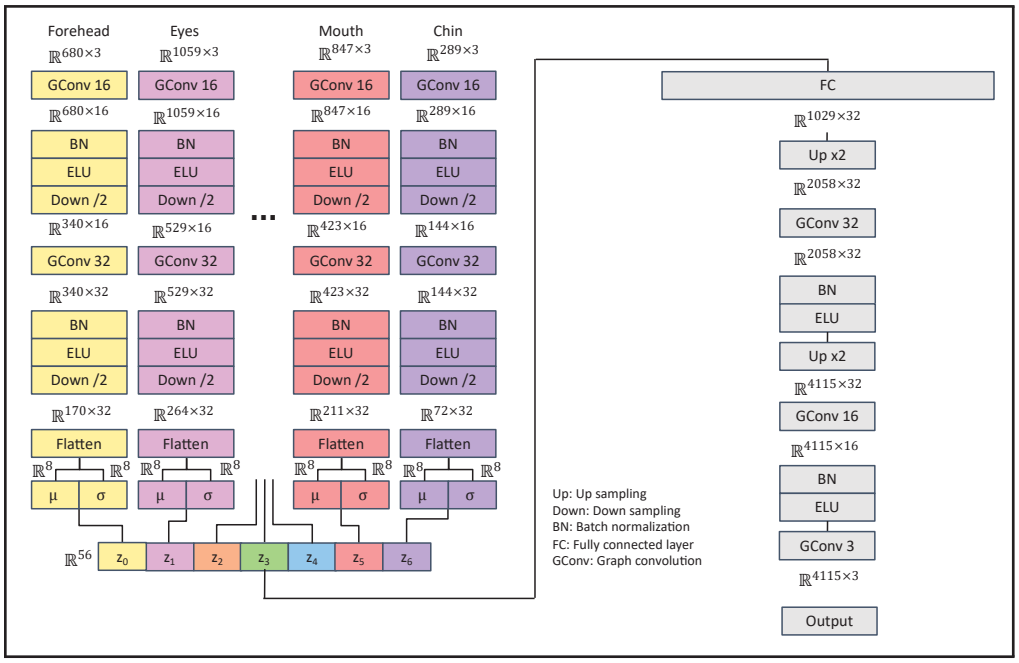

Figure 2.2 3D Face Generator: Network architecture

2.4 Loss Function

Our loss function is composed of three terms: a reconstruction loss, a KullbackLeibler (KL) divergence loss (Kingma & Welling, 2014), and a contrastive loss:

$$l = l_{Rec} + w_{kld} l_{KL} + w_{con} l_C. \tag{2.3}$$

We use an L_1 distance between the vertex positions of the ground-truth and decoded meshes.

$$l_{Rec} = \left\| \mathcal{F} - D(\mathbf{z}_f) \right\|_1 \tag{2.4}$$

The KL loss enforces a normal-like distribution for the latent vector $Q(z_i|\mathcal{F})$. The KL loss weight w_{kld} is 1e-3, and the number of face parts is $N_p = 7$.

$$l_{KL} = \sum_{i=0}^{N_p} KL\left(\mathcal{N}(0,1) \| Q(z_i|\mathcal{F})\right) \tag{2.5}$$

Inspired by the work of Deng, Yang, Chen, Wen & Tong (2020) in disentangling the latent space for human-face image generation, our third term reinforces latent space disentanglement and ensures that each part of the latent vector only affects the assigned part of the face:

$$l_C = \sum_{i=0}^{N_p} \left\| \left(\mathcal{F}_i' - \mathcal{F}\right) \odot \delta_{\notin \mathcal{P}_i} \right\|_1 . \tag{2.6}$$

For one part $\mathcal{P}_i$, we start with latent vector $\mathbf{z}_f$ and replace the part $\mathbf{z}_i$ with a randomly sampled $\mathbf{z}_i'$, while leaving the rest of $\mathbf{z}_f$ as before. We randomly sample with a uniform distribution $\mathcal{U}(-10, 10)$ to ensure we "push and pull" $\mathbf{z}_f$ far enough so that even large deformations of the face remain local. The result is a new final latent vector $\mathbf{z}_f{}'$ that only differs in $\mathbf{z}_i$. We calculate the L_1 distance between the vertex positions of the new face $\mathcal{F}_i'$ generated by decoding $\mathbf{z}_f{}'$ and the vertex positions of the face $\mathcal{F}$ decoded from the original $\mathbf{z}_f$. However, we ignore the current part vertices $\mathcal{V}_i$ to only penalize changes outside of part $\mathcal{P}_i$ by multiplying (Hadamard product) with $\delta_{\notin \mathcal{P}_i}$ defined that entries matching the part are zero and others are one. We sum up this

L_1 distance for each of the N_p parts. In addition, we set the w_{con} to 1e-4 at the beginning and gradually increase it by 10% ($w_{con} = 1.1w_{con}$) at the end of every epoch.

2.5 Training Procedure

We train the model for 70 epochs using the Adam optimizer (Kingma & Ba, 2015) and a batch size of 16. We tested with other numbers of epochs (up to 100), and 70 provided a good compromise between generalization and overfitting. Like CoMA (Ranjan *et al.*, 2018), we set the learning rate to 8e-4 and decay that rate by 0.99 every epoch. We set w_{kld} to 1e-3 to have a normal-like distribution (Equation 2.5), as mentioned before. Also, we gradually increase w_{con} to fortify the latent space disentanglement (Equation 2.6). Not starting with large w_{con} values helps the training process in the early stages since it enables the optimizer to converge to a good point in terms of reconstruction quality (Equation 2.4) first. We describe the details of our datasets in Sec. 3.1.

2.6 Random Face Generation

The trained network can be used as a 3D face generator by sampling from the learned latent space of the VAE. Our model consistently outputs plausible faces where the latent space is sampled from a normal distribution $\mathcal{N}(0, 1)$. This is in contrast to methods like the one of Ghafourzadeh *et al.* (2021) that might generate unacceptable faces and need an additional verification step to decline them.

2.7 Neural Face Editing

We can now use the trained face generator to create new faces. The user may directly manipulate each latent variable and observe the changes in the output of the network. However, operating directly on the latent space is often unintuitive as there is no clear semantic meaning behind each latent variable. For a more intuitive editing, we propose a flexible workflow where the

user controls vertex positions, and latent variables are automatically adjusted by our approach to reach the desired vertex movement (Figure 2.3).

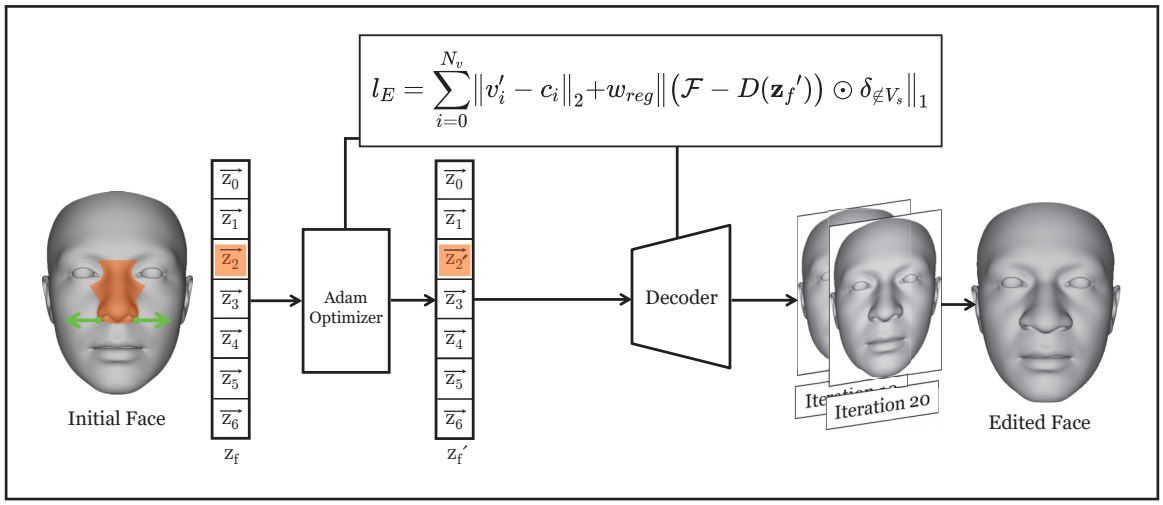

Figure 2.3 Neural face editing workflow example. The user input can be the movement of the selected vertices. For example, moving the green spheres in the direction of green arrows. Then to apply the change, we start from the initial face representation $\mathbf{z}_f$, and optimize for l_E. The updated latent vector $\mathbf{z}_f{}'$ is only different in the part latent sub-vector $\mathbf{z}_2$. As shown above, the edited face is the decoded result of the optimizer after a number of iterations. It has a wider nose since the user has moved the desired nose vertices apart

Our main goal is to ensure the locality of the edits. As such, at any given time, the user either manipulates one vertex (e.g., the tip of the nose) or a pair of symmetric vertices (e.g., corners of the mouth). At every stage, we aim to fulfill the 3D deformation prescribed by the user, maintain the locality of the edit, and preserve past edits. We formalize this as an optimization problem and use the Adam optimizer (Kingma & Ba, 2015) to find the best latent values. Given our editing workflow, the user controls one part at a time, which keeps the deformations local and helps the optimizer work on a smaller subspace and converge in a limited number of runs. After each modification, we start from the current face latent representation and gradually optimize that latent vector to achieve the next edit. To do this, we define the editing loss:

$$l_E = \sum_{i=0}^{N_v} \left\| v_i' - c_i \right\|_2 + w_{reg} \left\| \left(\mathcal{F} - D(\mathbf{z}_f{}')\right) \odot \delta_{\notin V_s} \right\|_1 . \tag{2.7}$$

The first term is the average Euclidean distance between the edited vertices v_i' and the constraints c_i set by the user. N_v corresponds to the number of vertices V_s selected for editing (either 1 or 2 in our workflow). The second term is regularization. We use this term to avoid deviating too much from the current face. We measure the average per-vertex L_1 distance between the two faces, excluding the selected vertices by defining $\delta_{\notin V_s}$ such that entries corresponding V_s are zero and others are one. In addition, w_{reg} is used to adjust the regularization term effect (we use $w_{reg} = 3$). To further fulfill local face editing, we only update the required part of the latent vector by identifying which segment of the face is being modified and will not include the rest of the vector as parameters of the optimizer. For example, if the selected vertices belong to part $\mathcal{P}_i$, only the related part of the latent $\mathbf{z}_i$ is modified. Therefore, the rest of the latent vector would remain the same.

Considering that we aim to integrate our solution into an interactive application, we want to run the optimizer only for a limited number of iterations. We use the Adam optimizer (Kingma & Ba, 2015) to find a solution quickly. Figure 2.4 shows the decrease of the loss through 500 iterations.

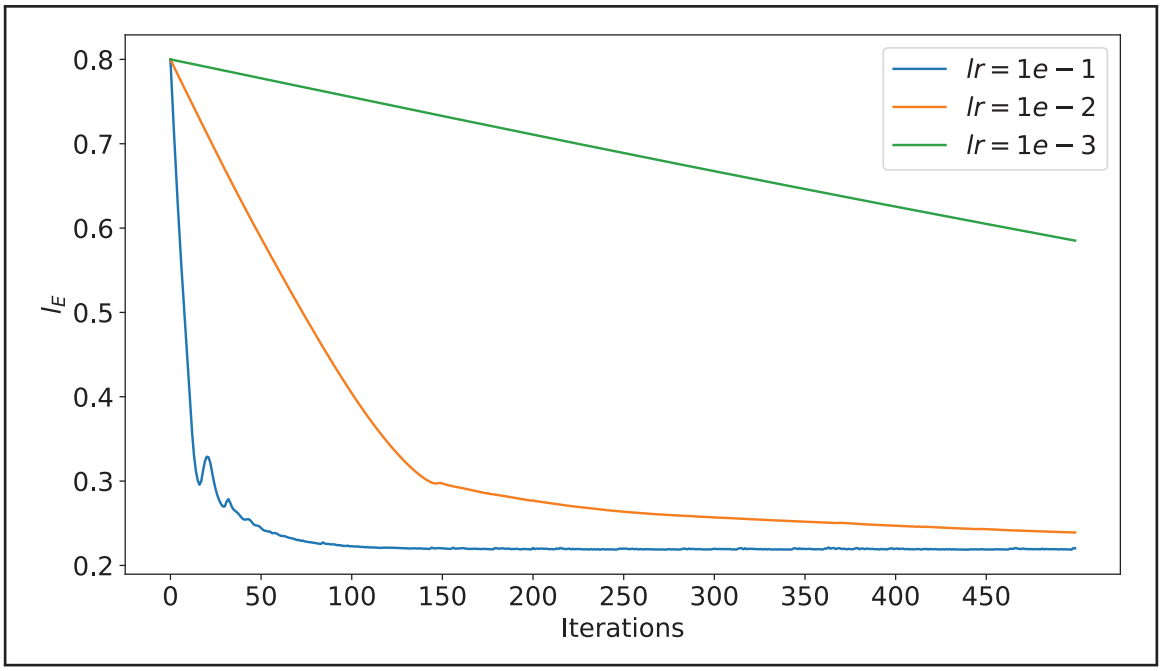

Figure 2.4 Graph of the l_E loss history (Equation 2.7), with different learning rates of the optimizer for vertex-based editing

In this test, we run the optimizer long enough to study its behavior. However, running the model for this many iterations is not suitable for interactive applications, and as we explain in the following, we use a significantly lower number of iterations. We see that with learning rates of 1e-2 and 1e-3 (taking small steps), the learning is predictable yet very slow. The learning rate of 1e-1 (taking larger steps) still makes reasonably stable progression and has the advantage of converging much faster, which is important in our interactive editing context. We can see that the optimizer makes significant progress toward the solution in the first tens of iterations and reaches a plateau around the 100th step, where the per iteration progress becomes negligible. Given our current target hardware (RTX3070), we choose to run the optimizer for 50 iterations with a learning rate of 1e-1. As a result, the process takes about 500 ms and is within the interactive range.

2.8 Texture Generator

In this section, we introduce our texture generator and discuss its properties, the training dataset, and the rationale behind our choice of architecture. Our image generator is specifically designed to generate albedo maps using our current dataset. This generator serves to enhance our 3D face generator (Sec. 2.1) by providing realistic textures. Additionally, it is intended to facilitate the work of content creators by enabling them to produce a large number of diverse faces with textures in a matter of seconds.

2.8.1 Dataset

The dataset utilized in this section contains the albedo maps for each of the 892 3D faces included in our face dataset. Each image has a resolution of $4K$ and is represented in the UV space. Given that all faces share the same mesh topology, they also share the same UV layout. As a result, any new albedo map synthesized by our texture generator can be applied to any face within the dataset or to any other face generated by our 3D face generator.

2.8.2 Network Architecture

We select StyleGAN2-ADA (Karras *et al.*, 2020) as our texture model for several reasons. Firstly, this model is capable of being trained on limited data, making it well-suited for our dataset, which is comparatively small in size. Secondly, StyleGAN2-ADA requires less powerful hardware and VRAM to train, allowing it to be used with commercially available GPUs such as the RTX3070. We follow the network parameters and details outlined in the paper (Karras *et al.*, 2020), including the latent sizes. Despite the fact that StyleGAN2-ADA is more memory-efficient than the original work (Karras *et al.*, 2018), training at high resolutions would still require a significant amount of memory. Furthermore, it would take over 40 days to train the network with $1K$ resolution images on an NVIDIA Tesla V100 GPU. As a result, we opted to train our model on images with a resolution of 256×256 pixels. Therefore, we downscale our $4K$ images before feeding them into the network. However, using a low-resolution albedo map

would compromise the overall quality of our generated faces. To address this issue, we trained a super-resolution network (Wang, Xie, Dong & Shan, 2021) to upscale our images from 256×256 to $1K$ resolution, thereby preserving the quality of our albedo maps. Figure 2.5 illustrates the workflow of our texture generator module.

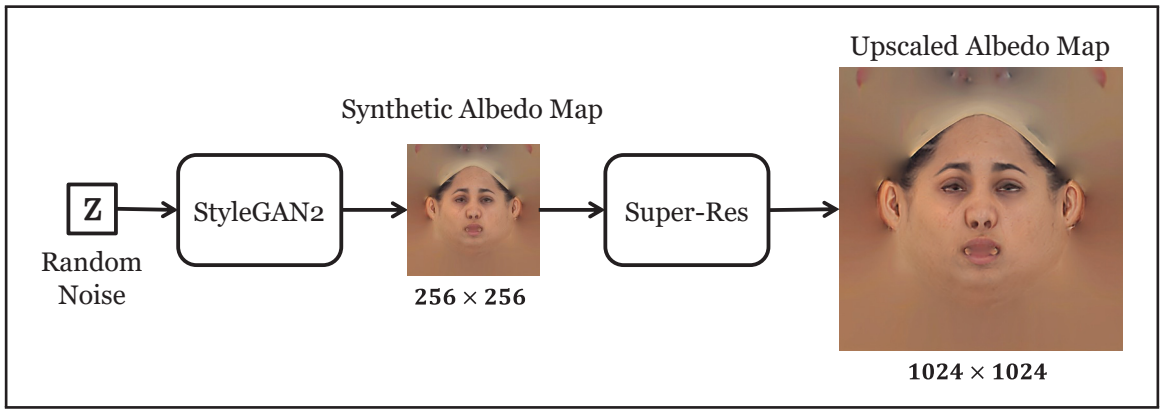

Figure 2.5 Texture generation workflow

2.8.3 Training Details

In the training of the texture generator network, again, we follow the exact steps described in the original work (Karras *et al.*, 2020). Specifically, we employ the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 25e-4 for both the discriminator and generator. Additionally, we use a batch size of 16.

CHAPTER 3

RESULTS AND EXPERIMENTS

In this chapter, we first cover the details of our datasets. Then, we present our editing and random face generation results. Afterward, we compare our approach with three state-of-the-art methods showcasing how our approach is better suited for local face modeling. Finally, we experiment with our face and texture generator to see how the faces would look when texture mapped.

3.1 Datasets

We evaluate our approach on two different datasets. The first dataset is composed of 892 scanned faces that share the same mesh topology. This is our primary dataset used to show the results in this thesis. We segment all the faces into the parts shown in Figure 2.1. Similar to the work of Ghafourzadeh *et al.* (2020), the segmentation is an offline process done manually and based on artists' feedback. More specfically, first, we use the help of artists to define the segments based on how they want to edit the face, then we build a list of triangles representing each part of the face. We use FaceWarehouse (Cao *et al.*, 2014) as our second dataset composed of 150 heads that also have the same mesh topology. The purpose of using this dataset is to evaluate the generalization capacity of our approach when trained on a relatively small dataset. Moreover, training independently on each of these two datasets demonstrated that our approach is not dependent on a specific one. In addition, we train the CoMA (Ranjan *et al.*, 2018) model with the FaceWarehouse dataset to compare it with ours. We use 80% of a dataset for the training set and the rest for validation.

3.2 Results

In Figure 3.1, we show some examples of vertex editing (Sec. 2.7). We use the version of our model that was trained with our primary dataset. We modify two different initial faces (initial 1 and 2) with different operations. It can be the movement of one vertex or two vertices. In the case of two control vertices, the user selects one vertex, and our system automatically selects the

symmetric one. Similarly, the user moves one vertex, and our system moves the second one in a symmetric manner. In each example, we modify one part of the face and, thus, one part of the latent vector. We see that the resulting deformations on the face are local. Moreover, we observe that even when moving only one or two vertices, we can effectively edit the face and benefit from our editing-friendly latent space and workflow. In addition, in Figure 3.2, we demonstrate some of faces that are the randomly generated from normal distribution $\mathcal{N}(0, 1)$ using our 3D face generator (Sec. 2.1). We do not apply any clamping to the scalars of the randomly generated latent vectors. As is typical from such normal distributions, most scalars are within the $[-1, +1]$ range, but others could be further away from the average. We can observe the broad range of facial characteristics that our generator can achieve and mix together. We use the model trained with our primary dataset for this case as well.

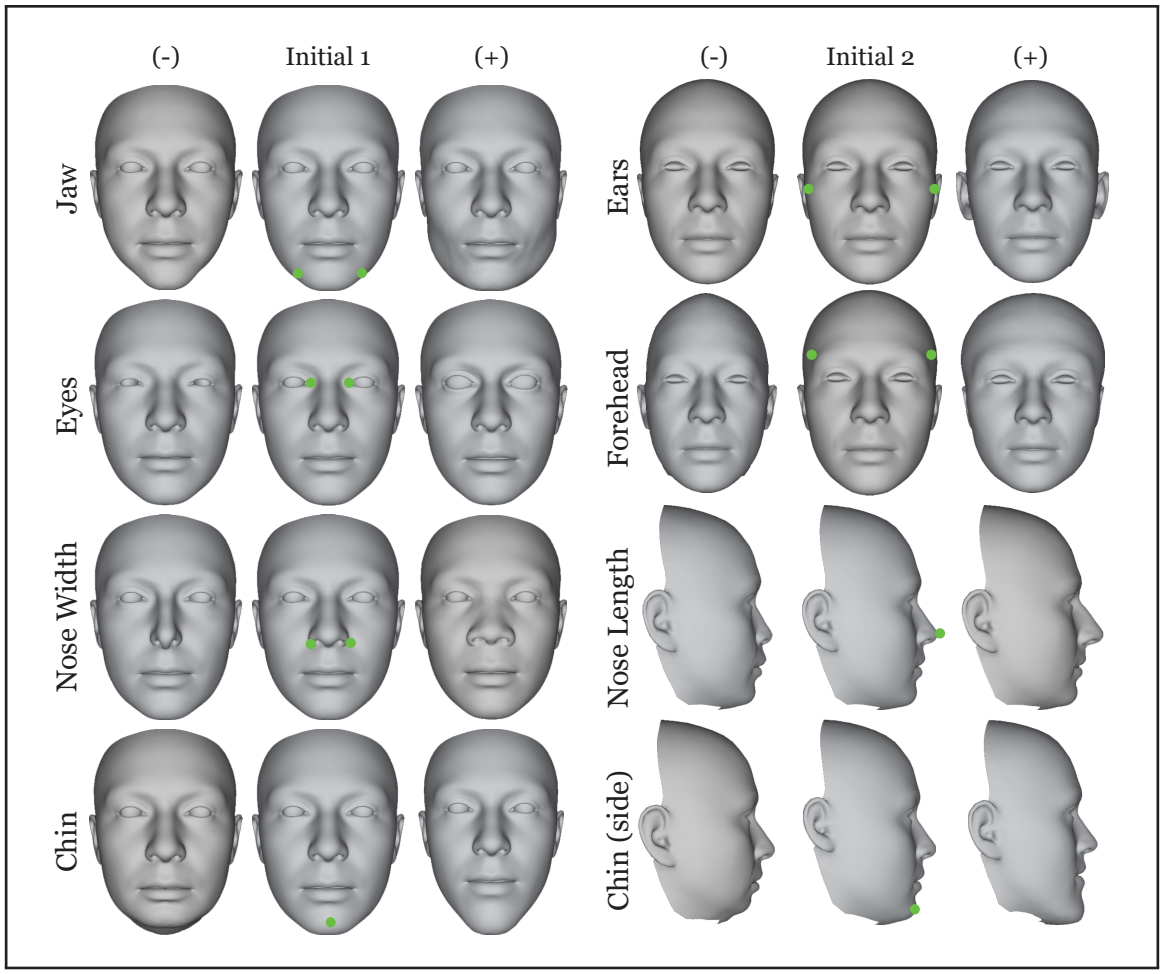

Figure 3.1 Vertex editing: Results of various face operations

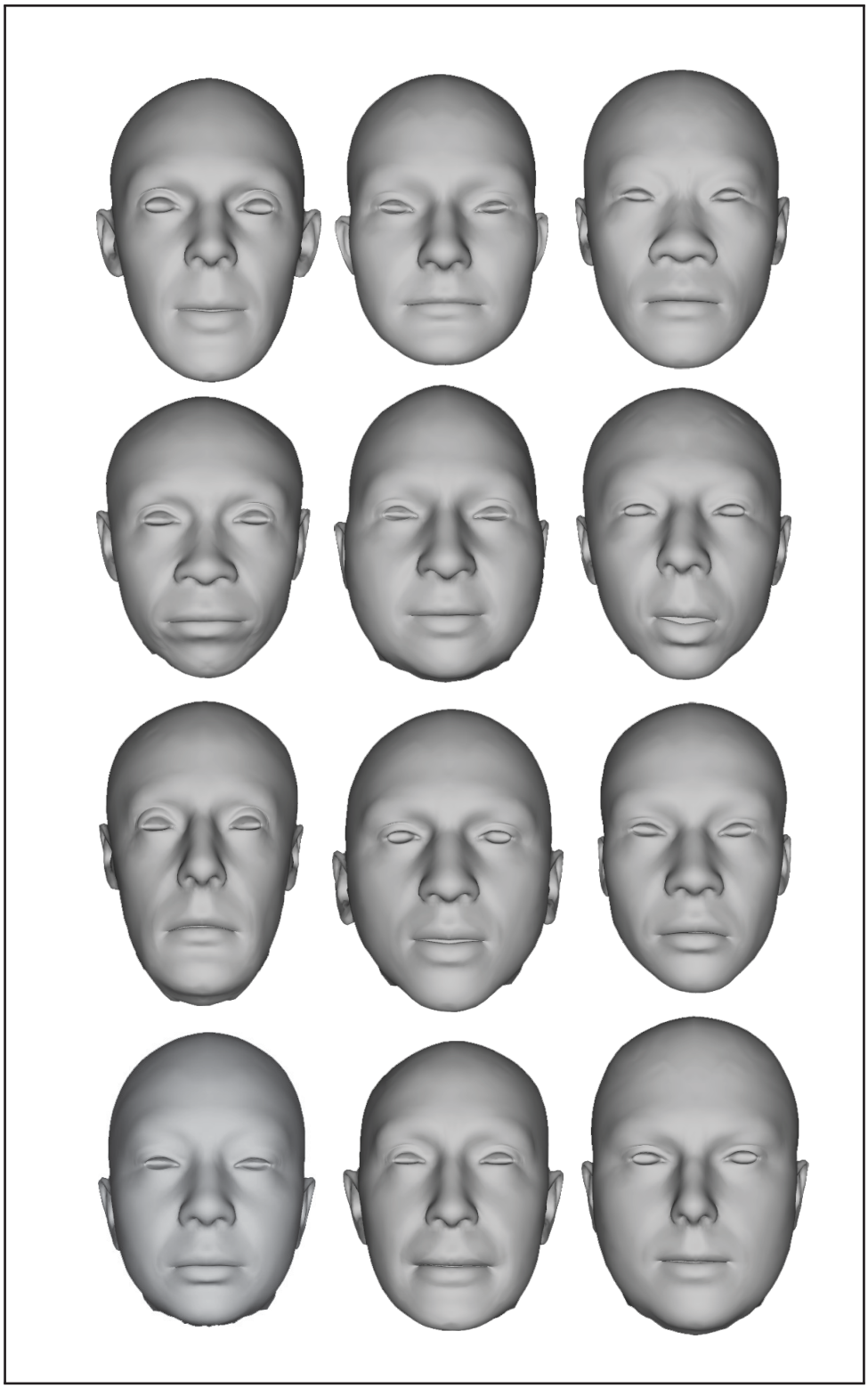

Figure 3.2 Examples of using our method to generate faces randomly by sampling from $\mathcal{N}(0, 1)$ in our latent space.

3.3 Ablation Studies

We perform two ablation studies. First, we cumulatively remove key parts one after the other, and second, we remove them one by one. These are two complementary analysis, and each helps us gain better insight into different elements of our approach. We evaluate the locality of the edits in each step with the following two quantitative measures: (i) The average vertex displacement inside the edited part $\overline{\delta}_{in}$. (ii) The average vertex displacement outside of the mentioned part $\overline{\delta}_{out}$ (vertices located in other parts of the face), $\overline{\delta}_{in}$ shows how much the part has changed locally, and $\overline{\delta}_{out}$ is essential for determining if the changes did not cause unwanted changes in other areas.

3.3.1 Cumulative

Figure 3.3 shows the effect of different elements of our approach: the regularization term (Equation 2.7), the local latent optimization, and the contrastive loss (Equation 2.6). In (a), we see our current approach, where all of the elements are present. In (b), we remove the regularization term from the optimizer. This change increases the unwanted global deformations (see the mouth and $\overline{\delta}_{out}$ in Table 3.1 which increases from 0.18 mm to 0.34 mm). In (c), on top of removing the regularization term, we also update the whole latent vector (global latent optimization) instead of only updating the part's sub-vector (local latent optimization). By comparing (b) and (c), we see how effectively our local latent optimization works. It performs much better than the global latent optimization because it mitigates most of the global deformations visible in (c). Finally, it is shown in (d) that when we exclude all the previous factors and also train our model without the contrastive loss, the deformations become even less local. For both (c) and (d), we see in Table 3.1 that the global deformation $\overline{\delta}_{out}$ drastically increases (0.18 mm to 4.61 mm and 6.92 mm).

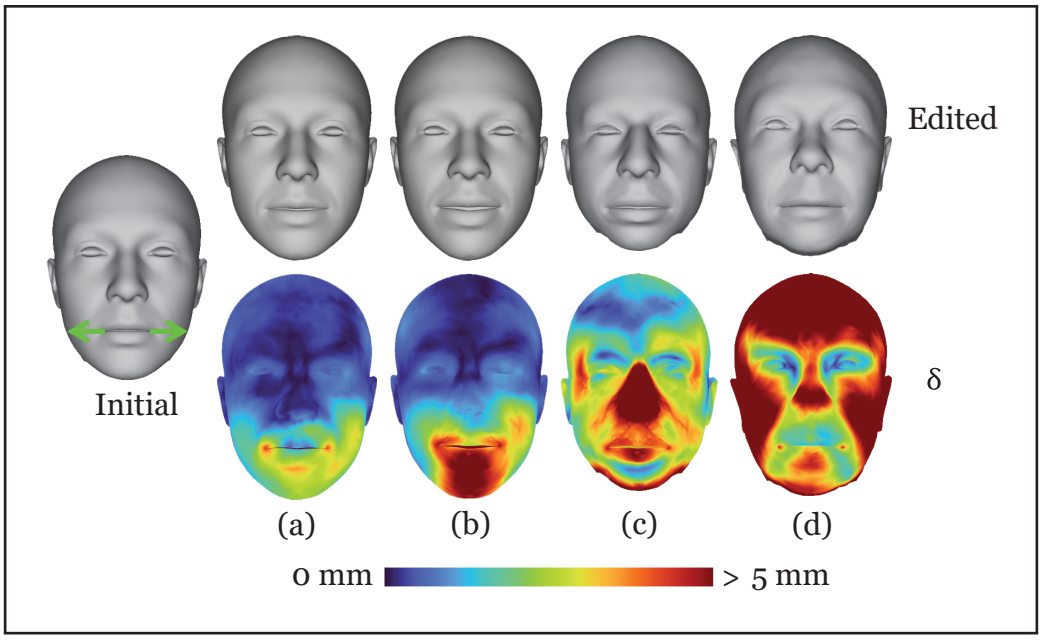

Figure 3.3 Ablation study: showing the effect of each element of our approach. (a) Our current approach. (b) No regularization term (Equation 2.7). (c) Like (b) but global latent optimization instead of local. (d) Like (c) but the model is trained without the contrastive loss (Equation 2.6). δ is the vertex displacement

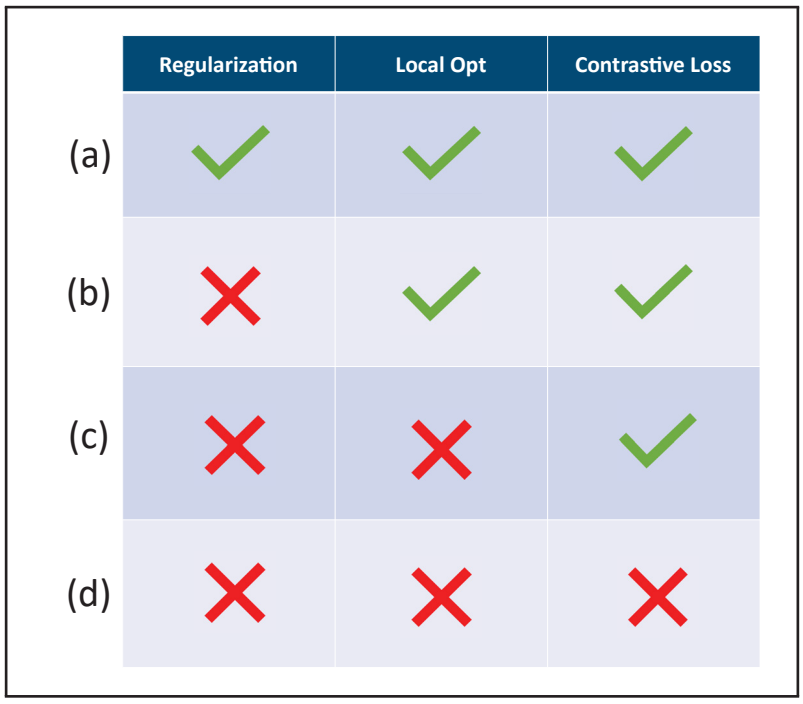

Figure 3.4 Cumulative ablation study summary

3.3.2 One by One

Figure 3.5 reflects the individual absence of each pivotal part of the method when editing the ears of the subject. In (a), we see our current approach. In (b), we again remove the regularization term. Depending on the edit, the regularization term sometimes makes a big difference (as we saw in Figure 3.3 (b)), while here we see that our model is already benefiting from a local latent space, and removing the regularizer makes little difference (only a slight increase of global deformation $\overline{\delta}_{out}$ from 0.19 mm to 0.22 mm, see Table 3.1). In (c), we only replace the local optimization with a global optimization (optimizing the whole latent vector instead of the part sub-vector). Other elements remain the same as in (a). We observe that there are more unwanted global deformations. Furthermore, in terms of speed, the same optimization task takes 37% more time to compute. This is because we are optimizing the whole latent vector. Hence we have more variables to tune. Finally, in (d), we replace the model in (a) with a model that was trained without the contrastive loss. The unwanted deformations become even more problematic. Similarly to what we observed for the cumulative ablation study, for both (c) and (d), the global deformation increases further more, even if in this case we make the changes one by one ($\overline{\delta}_{out}$ increases from 0.19 mm to 0.26 mm and 0.48 mm).

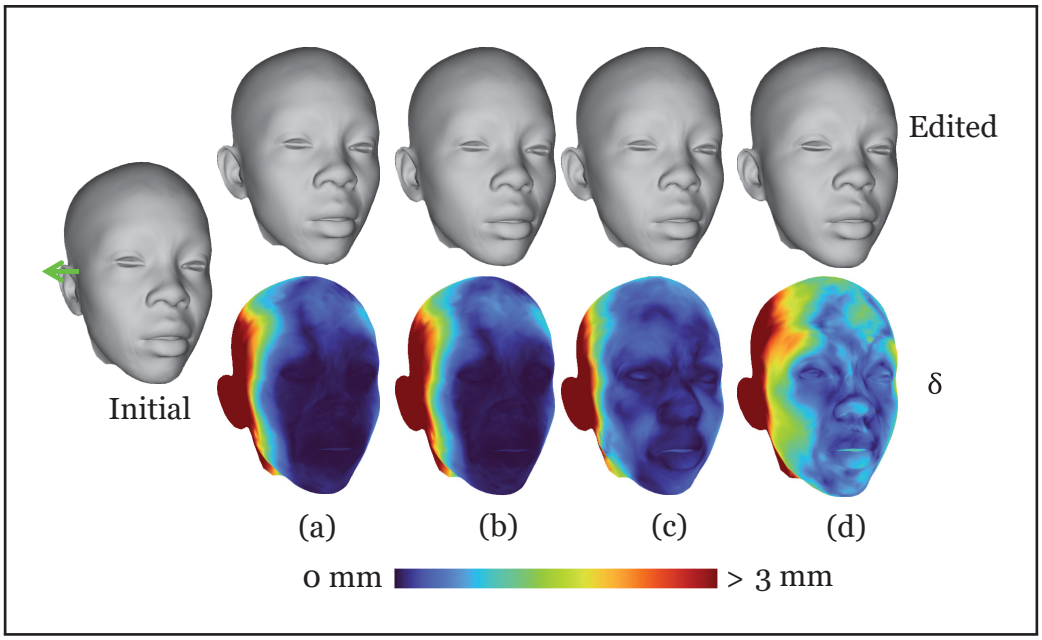

Figure 3.5 Ablation study: showing the effect of each element of our approach. (a) Our current approach. (b) No regularization term (Equation 2.7). (c) Global optimization instead of local. (d) As (a) but the model is trained without the contrastive loss (Equation 2.6). δ is vertex displacement

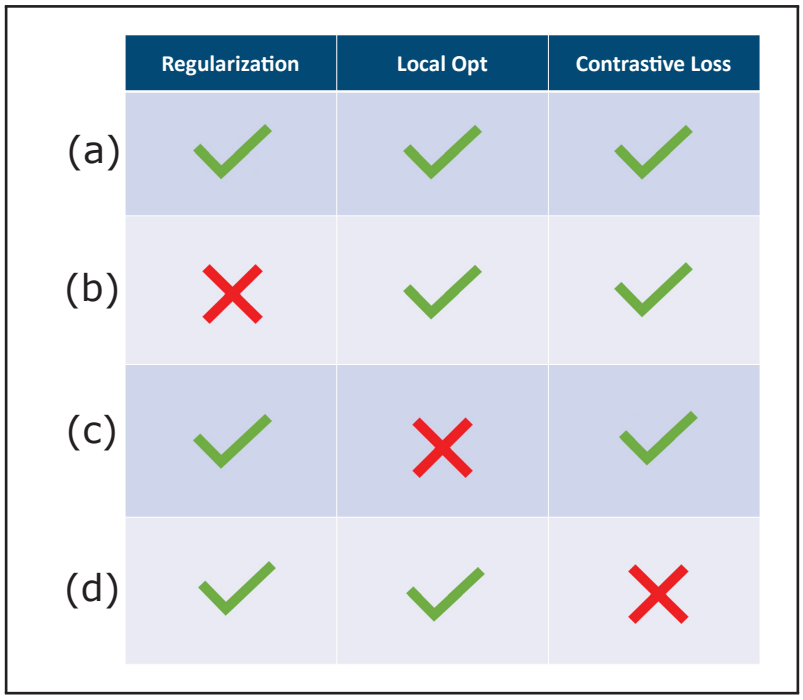

Figure 3.6 One by one ablation study summary

In Table 3.1, we measure the average *local* vertex displacement (inside the part, $\overline{\delta}_{in}$) compared to the average *global* vertex displacement (outside the edited part, $\overline{\delta}_{out}$) which we want to minimize to maintain the locality of the edit.

Table 3.1 Quantitative results of ablation studies. Average vertex displacement inside ($\overline{\delta}_{in}$) and outside ($\overline{\delta}_{out}$) the edited part.

Figure, Part		$\overline{\delta}_{in}$	$\overline{\delta}_{out}$
Fig. 3.3, Mouth	(a)	4.31 mm	0.18 mm
	(b)	5.00 mm	0.34 mm
	(c)	3.54 mm	4.61 mm
	(d)	7.60 mm	6.92 mm
Fig. 3.5, Ears	(a)	6.41 mm	0.19 mm
	(b)	6.75 mm	0.22 mm
	(c)	6.56 mm	0.26 mm
	(d)	7.17 mm	0.48 mm

3.4 Comparisons

In this section, we present a number of comparisons with various models ranging from very recent ones (Sec. 3.4.2) to more classical methods (Sec. 3.4.4). We compare all the models qualitatively. We also introduce a measure to compare some of the models quantitatively. More specifically, we use this measure to compare our model with the DNN models (Ranjan *et al.*, 2018; Jung *et al.*, 2022) since we can ensure that we can make the experiment environment as similar as possible.

3.4.1 Comparison with CoMA (Ranjan *et al.*, 2018)

We compare our method with CoMA (Ranjan *et al.*, 2018) as it also uses graph convolution and mesh sampling. We compare it to the VAE version of CoMA since, as our approach, it has a better interpolation space. This is because the VAE version is not only focused on reconstruction

tasks in contrast to the plain autoencoder version. In order to conduct different comparisons, we first need to train both models with the same dataset. We selected the FaceWarehouse dataset (Cao *et al.*, 2014) because the CoMA dataset (Ranjan *et al.*, 2018) has an insufficient number of subjects in the neutral pose, and a reasonable number of neutral poses is necessary to train a neural network for editing. We think this is a fair comparison since neither model has been designed around the FaceWarehouse dataset.

We first compare the reconstruction capabilities of the models. Our model has an average error of 1.40 mm on the training dataset (80% of the dataset) and 1.77 mm on the unseen samples (20% of the dataset). In the case of CoMA, it is 2.10 mm and 2.28 mm respectively. We can observe that even though our model's main task is not reconstruction, it has a better performance than CoMA.

Next, we want to check each model's output in various face editing scenarios. We use our vertex editing approach (Sec. 2.7) with both our model and CoMA. As shown in Figure 3.7, we edited different areas of the face with both models. Our model has a much more editing-friendly latent space and results in localized changes on the face. For example, we can see that CoMA (Ranjan *et al.*, 2018) introduces a lot of non-local movement: on the forehead when editing the mouth, toward the chin/jaw when editing the nose, and on the ears when editing the forehead. Furthermore, this test shows that using a regularization term with vertex-editing (Sec. 2.7) is not enough to prevent global changes, as apparent in the results of CoMA (Ranjan *et al.*, 2018). Table 3.2 verifies this with quantitative measures: the vertex displacement outside of the edited part ($\overline{\delta}_{out}$) is roughly an order of magnitude larger for CoMA (Ranjan *et al.*, 2018). Therefore, the complexity introduced in our model is justified. Finally, we can conclude that newer extensions of CoMA (Li *et al.*, 2019; Tan *et al.*, 2018; Yuan *et al.*, 2019) that only focus on the face and expression reconstruction also suffer from a non-editing-friendly latent space.

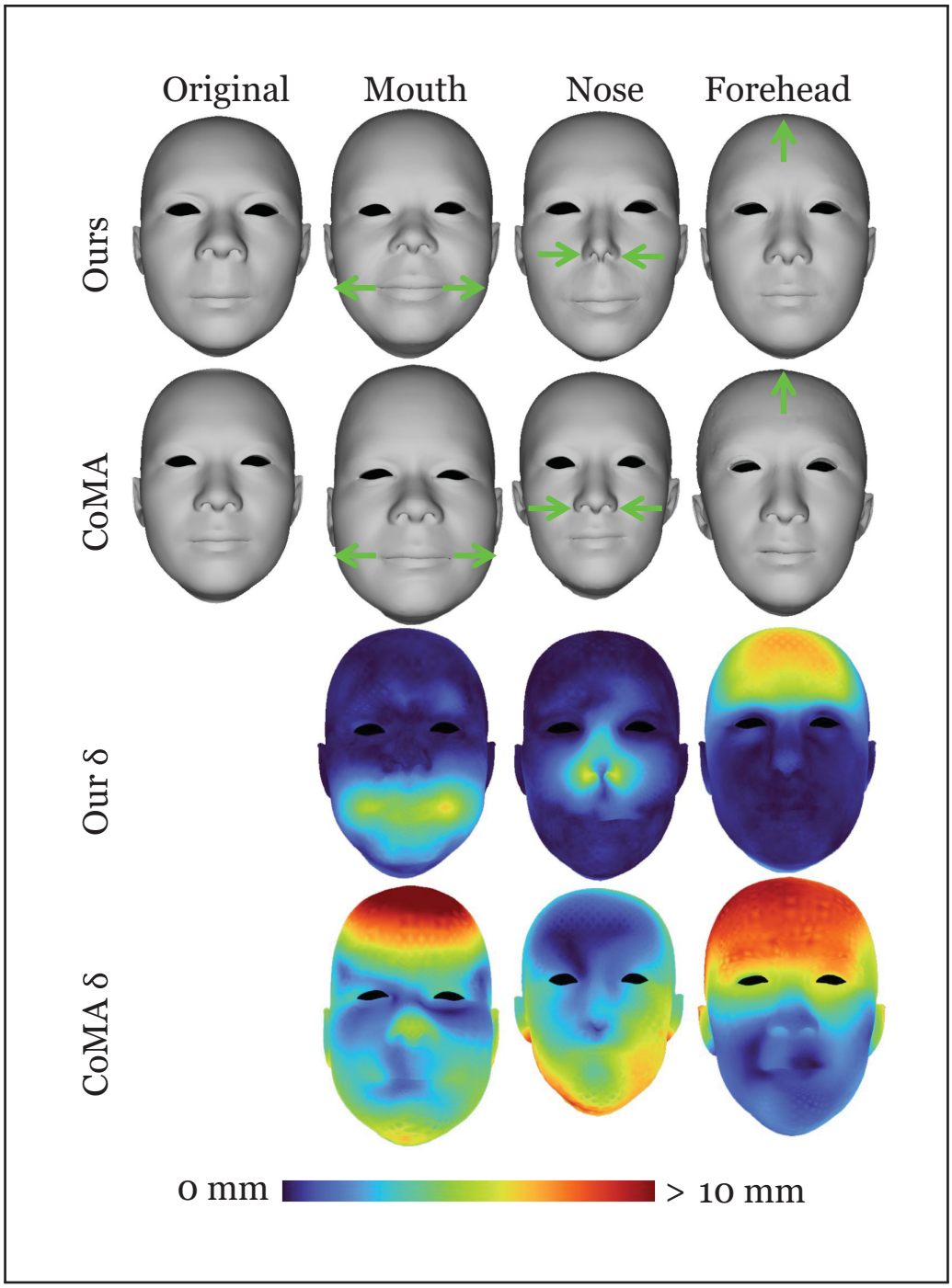

Figure 3.7 Comparison with Ranjan *et al.* (2018). δ is vertex displacement

Table 3.2 Comparing the locality of the editing for our approach against CoMA (Ranjan *et al.*, 2018) quantitatively (average vertex displacement inside, $\overline{\delta}_{in}$, and outside, $\overline{\delta}_{out}$, the edited part).

Model	Mouth	Nose	Forehead
	$\overline{\delta}_{in}$, $\overline{\delta}_{out}$	$\overline{\delta}_{in}$, $\overline{\delta}_{out}$	$\overline{\delta}_{in}$, $\overline{\delta}_{out}$
Fig. 3.7, Ours	4.71, 0.39 mm	3.44, 0.28 mm	3.73, 0.36 mm
Fig. 3.7, CoMA	2.47, 3.14 mm	1.96, 3.12 mm	7.54, 2.54 mm

3.4.2 Comparison with Jung *et al.* (2022)

In order to compare our method with that of Jung *et al.* (2022), we train the model with our primary dataset but keep the hyperparameters the same as the original work.

The original dataset of Jung *et al.* (2022) contains 1268 meshes for the training set compared to ours with 713 faces (80% of the dataset). In addition, our faces have 64% fewer vertices, but the model converges to a similar loss. Consequently, even when we train the model of Jung *et al.* (2022) on the whole dataset of 892 faces (and not only on a training set), the reconstruction error is 5.56 mm, which is very high compared to our model, where the reconstruction error is 1.22 mm on the training set and 1.51 mm on the validation set. We also should mention that their method has a latent dimension of 128, which is about double larger than ours.

To compare the face editing capabilities of the models, we use both the "point-handle-based editing" of their work and our method (Sec. 2.7) to make similar modifications to an initial face. The initial face is not identical since the reconstruction power of the models differs. We run each method for 50 iterations. Both take 500 ms to converge on average on an RTX3070. The results are shown in Figure 3.8.

We observe that while the method of Jung *et al.* (2022) can converge to a solution, it cannot prevent the unwanted changes that appear globally on the face. In Table 3.3, we observe that this method has a noticeably higher vertex displacement outside of the edited part ($\overline{\delta}_{out}$ is roughly a order of magnitude larger) in all scenarios.

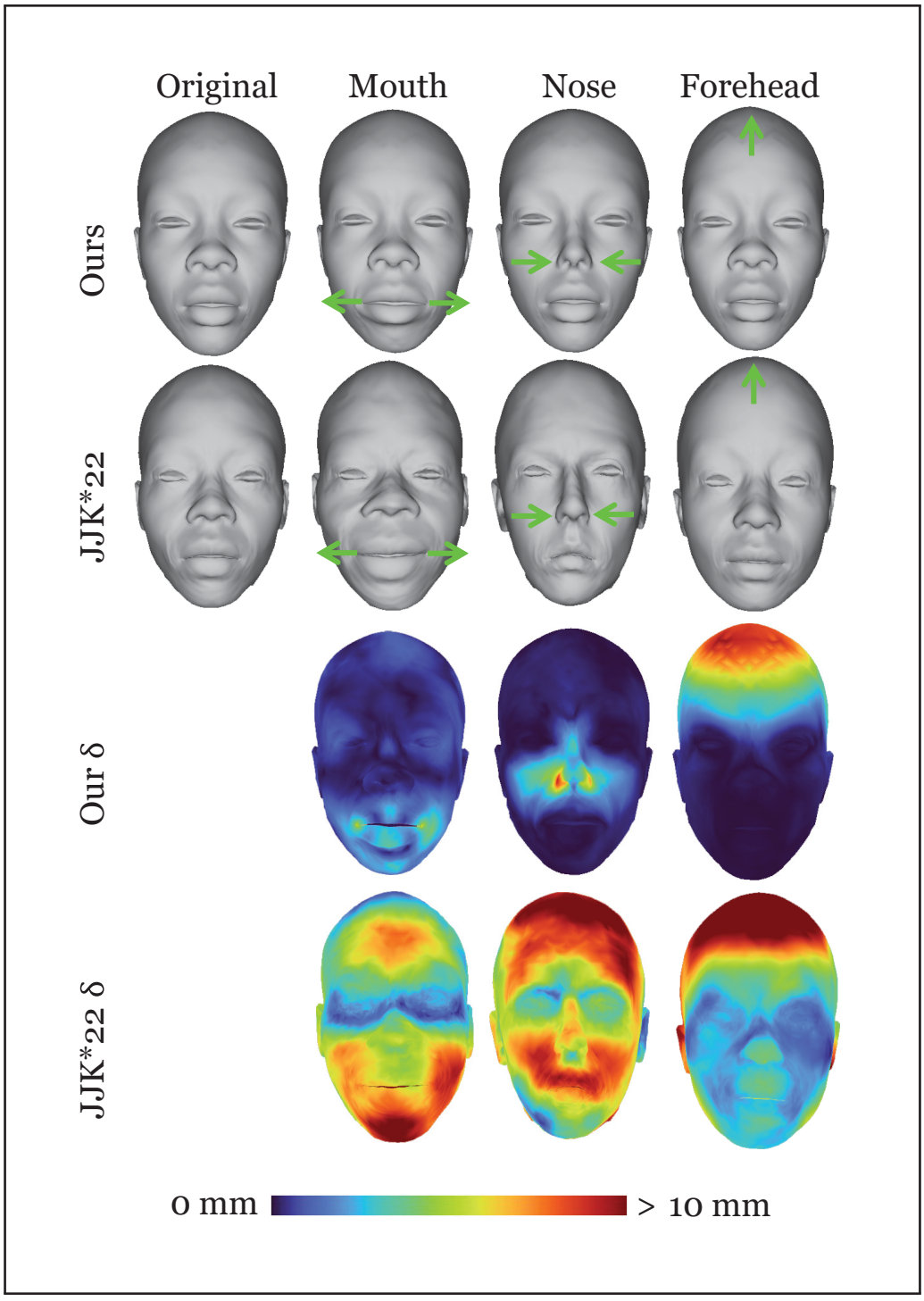

Figure 3.8 Comparison with Jung *et al.* (2022). δ is vertex displacement

Table 3.3 Comparing the locality of the editing for our approach against Jung *et al.* (2022) quantitatively (average vertex displacement inside, $\overline{\delta}_{in}$, and outside, $\overline{\delta}_{out}$, the edited part).

Model	Mouth	Nose	Forehead
	$\overline{\delta}_{in}$, $\overline{\delta}_{out}$	$\overline{\delta}_{in}$, $\overline{\delta}_{out}$	$\overline{\delta}_{in}$, $\overline{\delta}_{out}$
Fig. 3.8, Ours	3.40, 0.92 mm	2.55, 0.18 mm	4.75, 0.23 mm
Fig. 3.8, Jung *et al.* (2022)	6.90, 3.04 mm	4.32, 3.55 mm	7.29, 2.68 mm

3.4.3 Comparison with Ghafourzadeh *et al.* (2021)

We compare our approach with the localized 3DMM editing method of Ghafourzadeh *et al.* (2021) since both offer localized vertex-based editing. We fit that 3DMM model with our primary dataset to evaluate its face modeling application directly. The Ghafourzadeh *et al.* (2021) method is a clustered PCA-based approach that aims to pick the best eigenvectors that bring the reconstruction loss to under 1 mm. The budget is 50 eigenvectors on average as it differs for female and male heads. Also, it segments the face, similar to our method. The method of Ghafourzadeh *et al.* (2021) deforms each part in isolation and uses an additional smooth blending step to merge the generated part into the face. As a result, it ensures that the deformation only appears locally. Therefore, the method of Ghafourzadeh *et al.* (2021) performs better regarding reconstruction quality and keeping the changes local. Nonetheless, in Figure 3.9, we observe that our model deforms the face more naturally and meaningfully while keeping the changes comparably local.

For this comparison, we selected an initial face that is edited by both models. The face has an African ethnicity because we wanted to examine how well each model generalizes and works with an underrepresented face in a dataset (our dataset only contains 19 faces of this ethnicity out of 892 total). For the nose editing, we observe that while both models can decrease the width of the nose, our model preserves the shape of the nose, but theirs (Ghafourzadeh *et al.*, 2021) cannot achieve the same. For mouth editing, we try to close the gap between the lips of the subject. We can see that the PCA model does not generalize well and fails to close the gap. On the other hand, our model manages to achieve its goal naturally while not getting too far from

the original shape of the mouth. When editing the cheeks, we move one vertex on each cheek to create a chubbier or skinnier face. Our model achieves noticeably more plausible outputs. In contrast, the PCA model fails to output any visible changes on the face when it tries to make it skinnier. We find similar results when changing other features of the face. In conclusion, the method of Ghafourzadeh *et al.* (2021) either cannot make visible changes or, when pushed too far, results in uncanny and linear deformations.

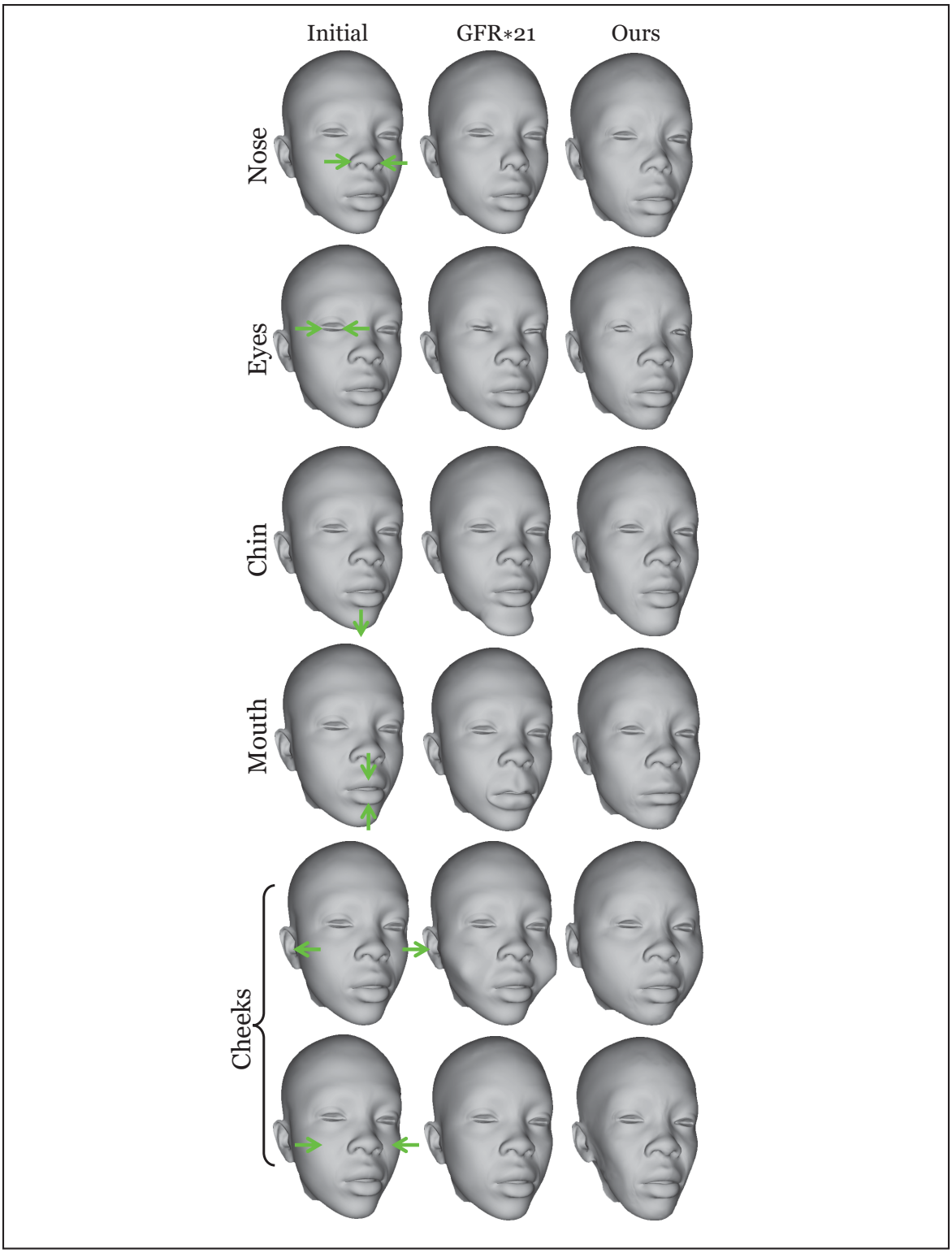

Figure 3.9 Comparison with Ghafourzadeh *et al.* (2021): Editing different features of the face. The green arrows indicate the edited vertices and direction of editing

3.4.4 Comparison with direct deformation

Commercial software allow to deform surfaces using various techniques. Such techniques typically do not rely on prior knowledge of the deformed object, as opposed to our approach which uses a dataset of faces.

We selected the well-known Laplacian surface editing (LSE) technique (Sorkine *et al.*, 2004) to evaluate how our results can compare to such non-data-driven deformation techniques. We pick the same face that was used in the previous comparison. Editing with LSE required more manual work in defining a proper "deformable" region (the vertices which are solved) and a "fixed" region (boundary conditions) for each edit to restrict the deformation to where we expect it. We pulled on the same vertices as in our approach and moved them trying to achieve a deformation similar to the one from our results (see Figure 3.10). The method works well in terms of keeping the deformations local, but this is at the expense of requiring a manual design of the fixed region for each individual edit. Furthermore, the changes may look overly linear and unrealistic. The mouth, chin, and cheeks are such examples. Another reason the results look unnatural is that the method is general and does not take into account that we want to remain within the manifold of realistic faces defined by a dataset. To a certain extent, one can circumvent these issues by carefully moving the handles and defining the fixed regions iteratively and by trial and error. Nonetheless, this process can become time-consuming and more similar to manual 3D face modeling applications.

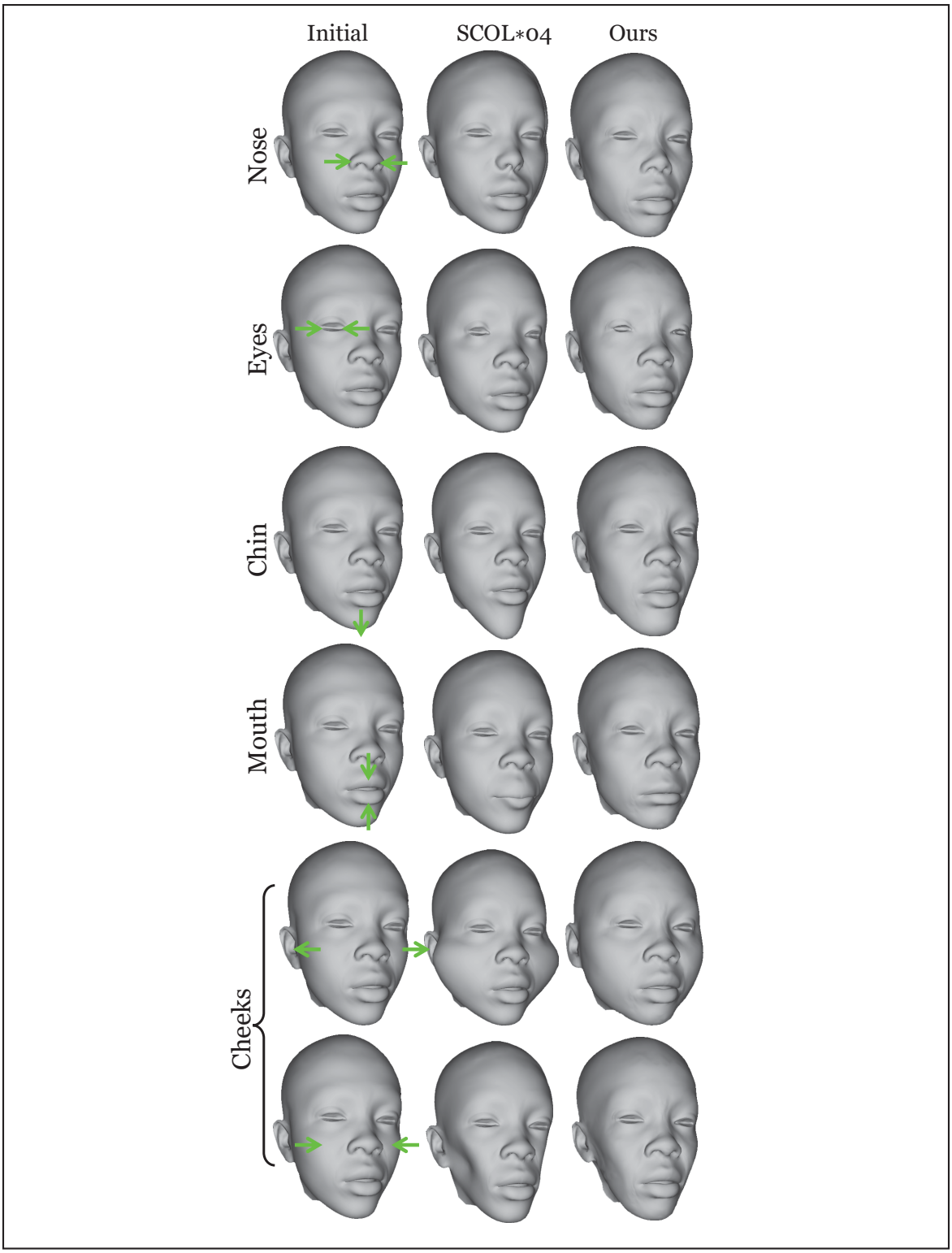

Figure 3.10 Comparison with Sorkine *et al.* (2004): Editing different features of the face. The green arrows indicate the edited vertices and direction of editing

3.5 Texturing the Faces

In this section, an experiment was conducted to evaluate the results of combining the face and texture generator. Initially, a set of 3D faces was generated randomly, as described in Sec. 2.6. Similarly, a set of 2D albedo maps was generated using the texture generator. As explained in Sec. 2.8, the images were upscaled to $1K$ for improved visual fidelity. Figure 3.11 demonstrates the outcome of this experiment. The results of this study demonstrate a high level of diversity among the generated faces. The faces exhibit a wide range of ages, skin tones, ethnicities, and genders. Despite the fact that the training of the texture and face generator, was not directly linked, meaning that sets of faces and textures were generated separately and randomly, the results appear realistic and plausible. These promising findings suggest that future work could explicitly relate geometry and textures to create a more robust 3D face generator.

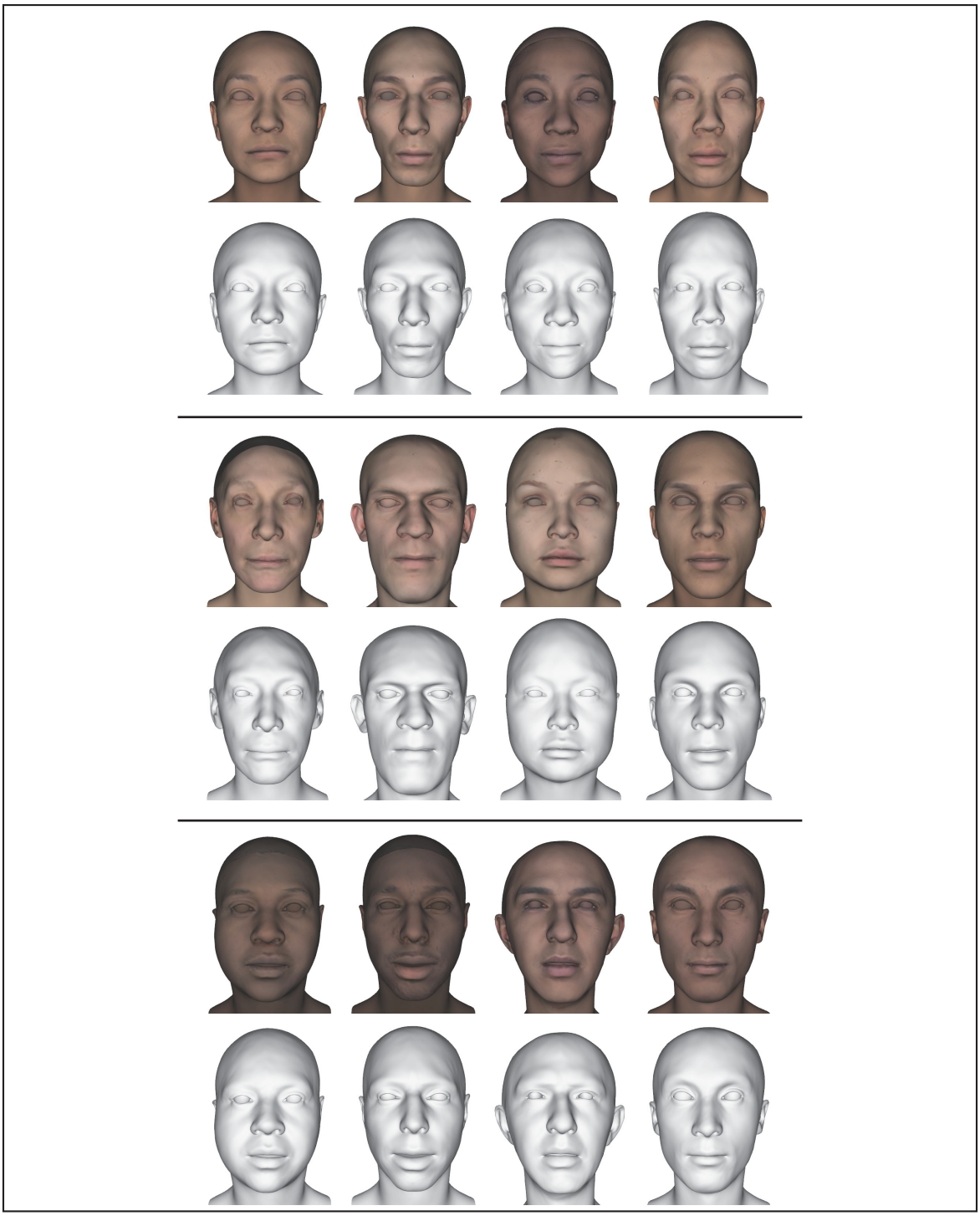

Figure 3.11 Examples of our texture generator (Sec. 2.8) coloring the faces that were randomly generated by our face generator (Sec. 2.1)

CHAPTER 4

LIMITATIONS

Even if our approach compares favorably against other methods in terms of the locality of the edits, some minor deformation (maximum of < 1 mm on average) of other parts of the face can still be observed. We observe this in Figure 3.7 and Figure 3.8, where we show deformation heat maps in our model's outputs. When the mouth area is changed, we can see about 1 mm of unwanted changes in the forehead and eye areas.

Regarding reconstruction accuracy, DNN models such as ours need more data to improve the model accuracy. This can be seen in our tests where the average reconstruction error for the dataset of 150 faces (1.44 mm training dataset and 1.77 mm validation dataset) is larger than the average error from the dataset containing 892 faces (1.22 mm training dataset and 1.51 mm validation dataset). Figure 4.1 also depicts the reconstruction error of a number of subjects in our dataset.

Still, along the lines of the reconstruction error, should a user want to edit a face that was manually modified in a modeling application, we will need to first encode that face to latent space, and thus inducing a reconstruction error. As such, for faces created outside of our system, the benefit of the advanced editing power comes at the expense of a slight global deformation of the face (the reconstruction error for the validation set of our primary dataset is 1.51 mm).

Finally, our optimization based methods will inherently be slower compared to direct inference methods. Nevertheless, optimization based methods provide more control to the user and a performance of 0.5 seconds for meshes of around $8,000$ vertices is still very practical and even for meshes with $3\times$ higher vertex count, our method scales linearly achieving 1.5 seconds to compute the target face.

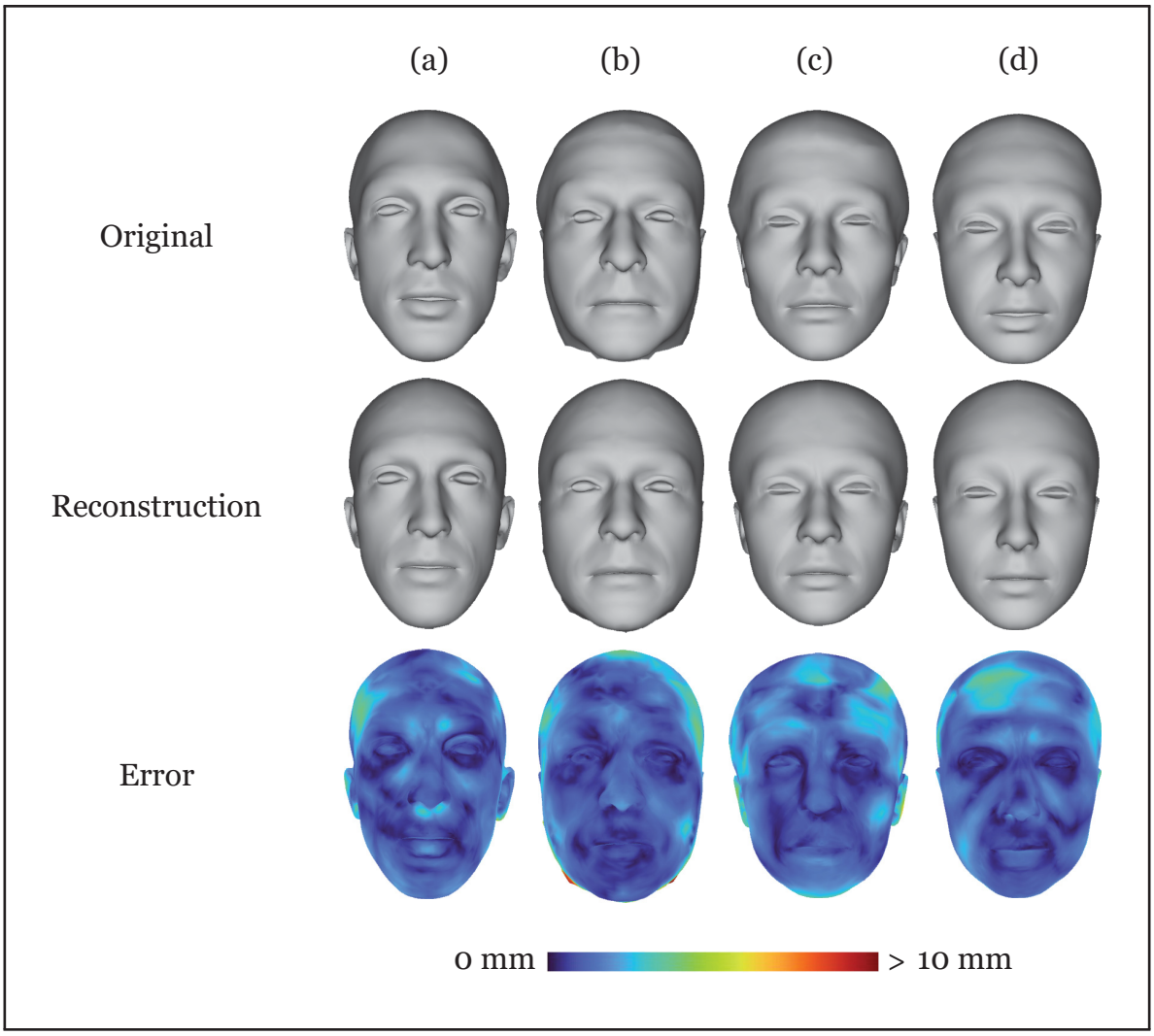

Figure 4.1 Examples of our model's reconstruction error. The heat map is a visualization of the L_2 distance

CONCLUSION AND RECOMMENDATIONS

We presented a novel variational autoencoder architecture to disentangle facial features in the latent space. We feed the separate parts of the face mesh to individual encoders while we use a single decoder to reconstruct the facial mesh. We take advantage of the graph neural networks to improve the learning from meshes. As such, we can successfully train even with datasets containing only 150 meshes.

Our architecture leverages the disentanglement properties of VAEs. Furthermore, given our new loss function, the network learns a latent space where each variable of the latent space influences a local region of the face mesh. Given our disentangled latent variables, our decoder is effective in sampling random faces as well as conducting face editing. For the face editing application, we developed a new loss function and a process that ensure that the face editing will remain local. The user can thus push and pull on a vertex and see the deformed face in interactive time. Thanks to our tailored latent space, the random face sampling and the facial editing both reconstruct faces that are realistic variations of the faces from the training dataset. We validated that our whole DNN architecture and learning strategy are not dependent on a specific dataset by successfully training it independently on our dataset as well as the FaceWarehouse dataset (Cao *et al.*, 2014).

Finally, we compared our approach with state-of-the-art methods in the application of facial editing. These comparisons demonstrated that our network has a better generalization property compared to 3DMM methods. Furthermore, our approach provides local editing while other DNN methods deform the face globally, for example, deforming the ears when editing the nose or mouth.

For now, we allocate the same number of variables for each face part in the latent vector. For future work, we would want to derive a strategy to automatically decide how many latent variables are necessary for each part. Ghafourzadeh *et al.* (2020) derived such a technique, but in

the context of 3DMM made out of PCA eigenvectors. Deriving such a strategy is quite different for DNNs.

Similarly, we give equal weight to each part of the face in the contrastive loss (Equation 2.6). Our model produces reasonable results with this approach. However, we can devise better methods to give a different weight to each part to make sure they have the right amount of locality in editing. We can achieve this through different methods, such as analytical ones (based on each part's vertex count) or even by including the weight of each part in the training loss so that the model can balance them automatically.

Another avenue for future work lies in the automatic adjustment of the graph convolution aspects of the learning, similar to the work of Li *et al.* (2019). We feel that adjusting the K factor (Equation 2.1) adaptively based on the mesh density of each part would improve the learning, locality, and generalization aspects of our approach.

Finally, we should mention a problem we encountered when working with higher-resolution meshes (50k vertices). We noticed some patterns and jagged edges that were appearing around the mouth and ears. We think the down-sampling was the culprit. Since the down-sampling operation of CoMA (Ranjan *et al.*, 2018) aims to be differentiable, it picks an aggressive strategy that changes the mesh topology drastically. We noticed that the edges are apparent in the down-sampled version of the heads as well. Therefore, we think it would be fruitful to invest in designing a new down-sampling operation that resolves this issue. Also, we can consider using other graph convolution operators and mesh strategies.

BIBLIOGRAPHY

Aliari, M. A., Beauchamp, A., Popa, T. & Paquette, E. (2023). Face Editing Using Part-Based Optimization of the Latent Space. *Computer Graphics Forum*, 42(2), 269-279.

Bagautdinov, T., Wu, C., Saragih, J., Fua, P. & Sheikh, Y. (2018, June). Modeling Facial Geometry Using Compositional VAEs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bao, L., Lin, X., Chen, Y., Zhang, H., Wang, S., Zhe, X., Kang, D., Huang, H., Jiang, X., Wang, J. et al. (2021). High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies. *ACM Transactions on Graphics (TOG)*, 41(1), 1–21.

Blanz, V. & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of SIGGRAPH 99*, (Annual Conference Series), 187–194.

Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A. & Dunaway, D. (2016). A 3D morphable model learnt from 10,000 faces. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5543–5552.

Booth, J., Roussos, A., Ponniah, A., Dunaway, D. & Zafeiriou, S. (2018). Large scale 3D morphable models. *International Journal of Computer Vision*, 126(2), 233–254.

Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M. & Zafeiriou, S. (2019). Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7213–7222.

Cao, C., Weng, Y., Zhou, S., Tong, Y. & Zhou, K. (2014). FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3), 413–425.

Chandran, P., Zoss, G., Gross, M., Gotardo, P. & Bradley, D. (2022). Shape Transformers: Topology-Independent 3D Shape Models Using Transformers. *Computer Graphics Forum*, 41(2), 195-207.

Cheng, S., Bronstein, M., Zhou, Y., Kotsia, I., Pantic, M. & Zafeiriou, S. (2019). Meshgan: Non-linear 3D morphable models of faces. *arXiv preprint arXiv:1903.10384*, 1–10.

Clevert, D., Unterthiner, T. & Hochreiter, S. (2016). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *Proceedings of ICLR*.

Defferrard, M., Bresson, X. & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, (NIPS'16), 3844–3852.

Deng, Y., Yang, J., Chen, D., Wen, F. & Tong, X. (2020, June). Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dhariwal, P. & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. *CoRR*, abs/2105.05233.

Egger, B., Smith, W. A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S. et al. (2020). 3D morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5), 1–38.

Fernandez-Abrevaya, V. (2020). *Large-scale learning of shape and motion models for the 3D face*. (Theses, Université Grenoble Alpes).

Ghafourzadeh, D., Rahgoshay, C., Fallahdoust, S., Beauchamp, A., Aubame, A., Popa, T. & Paquette, E. (2020). Part-Based 3D Face Morphable Model with Anthropometric Local Control. *Proceedings of Graphics Interface 2020*, pp. 7 – 16.

Ghafourzadeh, D., Fallahdoust, S., Rahgoshay, C., Beauchamp, A., Aubame, A., Popa, T. & Paquette, E. (2021). Local Control Editing Paradigms for Part-based 3D Face Morphable Models. *Computer Animation and Virtual Worlds*, e2028.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27.

Guan, Y., Jahan, T. & van Kaick, O. (2020). Generalized Autoencoder for Volumetric Shape Generation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1082–1088.

Ho, J., Jain, A. & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *CoRR*, abs/2006.11239.

Jung, Y., Jang, W., Kim, S., Yang, J., Tong, X. & Lee, S. (2022). Deep Deformable 3D Caricatures with Learned Shape Control. *ACM SIGGRAPH 2022 Conference Proceedings*.

Karras, T., Laine, S. & Aila, T. (2018). A Style-Based Generator Architecture for Generative Adversarial Networks. *CoRR*, abs/1812.04948.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. & Aila, T. (2020). Training Generative Adversarial Networks with Limited Data. *CoRR*, abs/2006.06676.

Kingma, D. P. & Ba, J. (2015). *Adam: A Method for Stochastic Optimization*. Proceedings of ICLR.

Kingma, D. P. & Welling, M. (2014). Auto-Encoding Variational Bayes. *Proceedings of ICLR*.

Li, K., Liu, J., Lai, Y.-K. & Yang, J. (2019). Generating 3D Faces using Multi-column Graph Convolutional Networks. *Computer Graphics Forum*, 38(7), 215–224.

Li, R., Bladin, K., Zhao, Y., Chinara, C., Ingraham, O., Xiang, P., Ren, X., Prasad, P., Kishore, B., Xing, J. & Li, H. (2020). Learning Formation of Physically-Based Face Attributes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3407–3416.

Li, T., Bolkart, T., Black, M. J., Li, H. & Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)*, 36(6), 194–1.

Ploumpis, S., Wang, H., Pears, N., Smith, W. A. & Zafeiriou, S. (2019). Combining 3D morphable models: A large scale face-and-head model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10934–10943.

Ploumpis, S., Ververas, E., Sullivan, E. O., Moschoglou, S., Wang, H., Pears, N., Smith, W. A. P., Gecer, B. & Zafeiriou, S. (2021). Towards a Complete 3D Morphable Model of the Human Head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11), 4142-4160.

Ranjan, A., Bolkart, T., Sanyal, S. & Black, M. J. (2018). Generating 3D Faces using Convolutional Mesh Autoencoders. *Computer Vision – ECCV 2018*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *CoRR*, abs/2112.10752.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (pp. 318–362). Cambridge, MA, USA: MIT Press.

Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C. & Seidel, H.-P. (2004). Laplacian Surface Editing. *Proceedings of the EUROGRAPHICS/ACM SIGGRAPH Symposium on Geometry Processing*, pp. 179–188.

Tan, Q., Gao, L., Lai, Y.-K. & Xia, S. (2018, June). Variational Autoencoders for Deforming 3D Mesh Models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5841–5850.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, abs/1706.03762.

Vesdapunt, N., Rundle, M., Wu, H. & Wang, B. (2020). JNR: Joint-Based Neural Rig Representation for Compact 3D Face Modeling. *Computer Vision – ECCV 2020*, (Lecture Notes in Computer Science), 389–405.

Wang, L., Chen, Z., Yu, T., Ma, C., Li, L. & Liu, Y. (2022). FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20333–20342.

Wang, X., Xie, L., Dong, C. & Shan, Y. (2021). Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905-1914.

Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R. & Cao, X. (2020). Facescape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction. *Proceedings of the ieee/cvf conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 601–610.

Yuan, C., Li, K., Lai, Y.-K., Liu, Y. & Yang, J. (2019). 3D Face Representation and Reconstruction with Multi-scale Graph Convolutional Autoencoders. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1558–1563.

Zhan, F., Yu, Y., Wu, R., Zhang, J. & Lu, S. (2021). Multimodal Image Synthesis and Editing: A Survey. *CoRR*, abs/2112.13592.