Adaptive avatar customization for immersive experiences

Nicolas Olivier

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

Nicolas OLIVIER

Adaptive Avatar Customization

Thèse présentée et soutenue à Rennes, le 31 Mars 2022
Unité de recherche : INRIA Rennes - Bretagne Atlantique et Interdigital

Rapporteurs avant soutenance :

Prof. Céline LOSCOS	Professor, University of Reims Champagne-Ardenne
Prof. Slim OUNI	Associate Professor, University of Lorraine

Composition du Jury :

Président :	Prof. Edmond BOYER	Senior Researcher, INRIA Grenoble Rhône-Alpes
Examinateurs :	Prof. Rachel McDONNELL	Associate Professor, Trinity College Dublin
	Prof. Edmond BOYER	Senior Researcher, INRIA Grenoble Rhône-Alpes
	Prof. Céline LOSCOS	Professor, University of Reims Champagne-Ardenne
	Prof. Slim OUNI	Associate Professor, University of Lorraine
Dir. de thèse :	Prof. Franck MULTON	Senior Researcher, INRIA Rennes
Enc. de thèse :	Dr. Ferran ARGELAGUET	Researcher, INRIA Rennes
	Dr. Fabien DANIEAU	Researcher, Interdigital
	Dr. Quentin AVRIL	Researcher, Interdigital

Table of Contents

Chapitre 1

General Introduction

The notion of preferential appearance is ancient, emerging early with inter-organism interactions, whether for camouflage, deception, or sexual competition [1], [2]. As other evolutionary features of our species, this notion has gained new meanings with the growth of group specific culture [3], for identifying genders, castes, official functions, and for various ceremonial aspects [4]. Appearance also has an unconscious influence on behavior, as shown by the Proteus effect [5]. People could hence change their appearance in order to induce certain desired behavior in themselves, or simply for custom. There was indeed an emergence of the notion of embodiment of a sub part of ourselves through changing our appearance (for work, parties, marriage, etc.), but also of the notion of embodiment of another person or entity, whether for rituals, theater, or other role games [6]. In recent history one of the most common part of these practices – leisure – have gained traction from being restricted to the most privileged part of the population, to the main part of the developed world. This was due to the diminution of the cost of production of clothes caused by the industrial revolution, and the rise of non working time happening from the start of the 19th century [7]-[9]. This was also a period during which the experience of embodiment of other persons have risen massively, through the growth and popularity of the cinematographic and television industries. During the last decades, a practice that had been mainly restricted to the use of clothes, masks, or body paintings, has also known the deep changes brought by computers. It was brought first by computer graphics being used in visual content, and then by the rise of embodiment-based games on personal computers, skyrocketing the use and cultural normality of embodiment and third-person control of characters to heights never seen. In this thesis, we aim to perform character stylization in a fully automatic manner.

FIGURE 1.1 – Sigourney Weaver (right) and her stylized character (left) in Cameron's *Avatar (2009)*

1.1 Context

Personalizable characters have been present in games for several decades, allowing users to stylize (alter) the appearance of their faces, bodies, or clothes of the character they incarnate or interact with. This character stylization is traditionally operating as a combination of attributes (e.g. hair color, facial shape, nose width) in its simplest form (e.g. The Sims 1, World of Warcraft, Snapchat), or a mix of it and an editable linear morphable model, with dimensions related to various facial features (e.g. The Elder Scrolls IV – 2006, Fallout 3 – 2008). Various games bypass the notion of personnalisation by removing most of the concept of appearance from the equation, using mute, first person characters with little background (e.g. Half Life – 1996, Portal – 2007). This has the drawback of limiting multi-user interaction, as appearance is important for it, even simply for distinguishing characters from one another. In the case of movies, with only tens of actors compared to thousands or millions of gamers, personalizing characters does not require automation and can be seen instead as a cost saving factor, since realistic hand crafted personalized embodiments are limited to a few per high budget movies (e.g. The Lord of the Rings – 2001, Avatar – 2009 1.1). Stylized or not, digital doubles presence in published content has grown significantly these last decades, from movies (e.g. Terminator 2 – 1991), to TV series (e.g. The Boys – 2019), and games (e.g. Cyberpunk2077 – 2020). Their applications range from stunt double, to securing or revalorizing intellectual property by modeling deceased people (e.g. Rogue One – 2016, Michael Jackson hologram concert – 2014).

The advances of the last decades have produced radical changes in the way we use and interact with virtual characters. Virtual embodiments moved from being only acted by others, to being near photorealistic and controllable through haptic interfaces in first person view, and from being a persona you pick to play once in a while, to being a constant representation of you on social media. The concept of having a whole digital double is a future strongly pushed forward by commercial products such as the Metaverse. As modern day virtual embodiments grow closer and closer to ourselves, and as we reach times where entirely virtual characters can look realistic on screen, there is a growing interest in being able to persoannalize your own embodiment. This means that there is a need of a stylized character resembling a reference person, whether for not losing the brand value of an actor, or increasing the sentiment of embodiment, intimacy as one's virtual double, and interactivity with others in shared environments. Indeed, in a cultural climate with both strong individuality and an inter-connectivity allowing for a large set of group identities, the personalization of our virtual representations according to one's own preferential aesthetic and beliefs seems like a key aspect of our embodiments. The concept of changing the appearance of your character is already very present and popular in video games [1].

On a wider picture, outside personalisation of oneself, due to our nature as a social species and the inherently rich interaction they promise (since we are designed to interact with humanoids) stylized virtual characters are not only here to stay, but promised to flourish, whether as the faces of "smart" natural language systems of various kinds, or even simply as dynamic decorations of a virtual environment. The richness and diversity their appearance can hold is directly linked to our capabilities in producing large numbers of high quality customized characters. Hence, it narrowly related to the notion of personalization of embodiments, expanding the application domains of such stylization capabilities. The ongoing and consistent growth of the industries of movies, games, and VR experiences promises only a near future with more embodiments and stylized virtual characters.

1.2 Motivation and Goals

Virtual character stylization is about altering the appearance of a virtual character, while conserving some core identity features, to let them be recognizable. An exhaustive set of

1. https ://newzoo.com/insights/articles/u-s-core-gamers-81-of-those-aware-of-in-game-cosmetics-want-to-trade-skins-for-real-world-money/

stylizable features would include face and body visual appearance, animations, as well as the voice, and the language.

Humans as a species endowed with senses, depend heavily on their vision. The richness of that perceptual channel has caused it to be the subject of most of our functionnal, cultural, and aesthetic focus, as could be guessed observing our arts, five of the so called seven higher being heavily dependant on vision (architecture, sculpture, painting, performing, and film). A person's fine variations of appearance (its identity) are mostly contained in their face, these having evolved to be varied and recognizable [10]. As such, aiming to prioritize stylization of the attributes most important in the human perception, we focus in this thesis exclusively on facial appearance and identity, as well as expressions.

Virtual characters can have various degrees of realism, ranging from the simple cartoon characters of messaging apps to photorealist humanoids shown in movies (e.g. The Lord of the Ring (2001-2003), Avatar (2009)). In this thesis, we target a level of realism close to realist video games, aiming to provide a strong baseline which could be used for a game or VR application, or fine tuned and worked on by artists to be used in a context requiring higher photorealism. This thesis was part of the Digital Human project at InterDigital, which aimed to provide near-photorealistic scans of individuals, and had similar contextual targets.

Outside of movies, character stylization have gained momentum with the commercialisation of consumer applications such has Prisma, or FaceApp. The popularity of much simpler filters such as the basic ones of Snapchat being already popular, coupled with the investments towards the production of more smartphones and other consumer products with neural network dedicated hardware showcase the short term board consumer commercial and cultural capabilities of facial stylization methods. The use of neural network based technology has known considerable growth during these last few years. In the artistic world it has been used both as a tool to support creativity [2], and as a way towards art creation by itself, through the use of generative networks [3]. In the middle, lays the field of controllable and tunable image generation, of which the concept of stylization is a key element. It allows the generation of realistic or near realistic images from semantic maps or sketches [11], and the control and mixing of semantic attributes, most notably in the

2. https ://helpx.adobe.com/photoshop/using/neural-filters.html
3. https ://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

concept of faces [4].

Stylization – or style transfer – can be characterized as the mapping of one distribution (as in, a function giving all possible values of some data, with their occurrence frequency) towards another, where both distributions share part of their information. In the literature, style transfer has been mostly modeled as the separation of data into content (what is shared between distributions, identity in the case of faces), and style (everything that is domain specific). This model has allowed the development of tools for converting photos to paintings or drawings (e.g. Prisma), doing a cartoonisation of selfies (e.g. Snapchat), transfering pose and expression between facial photos, converting maps to satellite images [12], and cats into dogs [13]. However, no well formed theory although exist on the distinction of style and content, separation being done empirically, through heuristics [14] or functions learned in a supervised or unsupervised manner. Both style and content, in the context of character appearance, are difficult to define, and remain largely subjective. Heuristics used to model style and content include high/low frequencies, dimensions of a PCA [15], features of layers of convolutional neural networks [14], or feature vectors of facial recognition networks [16].

Whatever the approach is used, producing an output with a given style and producing one with a given content tend to be adversarial goals, as optimizing for one typically negatively impact the other. Changing the style of an input generally affects its global information, hence also its content, and vice versa. In the supervised (paired) learning setting [17], content can be estimated as the shared information between image pairs, and style as their difference, but pairs of stylized and non-stylized faces are too uncommon to allow for such an approach. Instead, unsupervised (unpaired) datasets and methods are used, editing style and content guided by a mixture of features handcrafted or derived from pretrained networks, specially trained generators, or pixel-wise reconstruction losses. The level of unsupervision can vary, from having only style labels, to only having them for photorealist data, while identity labels remain restricted to very few datasets [18]. Some approaches limit themselves to textural and local shape stylization, limiting causes for content data loss, hence also for content preservation [19]. Some other focus on a specific style and design their pipeline around it [20], while others restrict themselves to considering content as global spatial features such as pose [13].

4. https ://artbreeder.com

There exist no such approach showcasing capabilities for transferring arbitrary (e.g. alien, orc, cartoon) styles to a human face from a few examples, with an output reflecting the desired identity and style. Such a method would be desirable, as there usually exist very few examples (often only one) for any given style, especially in the 3D domain. In the 3D domain, there is no approach for stylizing a given face from a few examples, only methods with rules designed for a specific one (e.g. caricatures [21]). Even approaches for 3D stylization from a large set of examples are rare and have limitations such as being mostly textural [20], or being domain specific and untested against their rule based alternatives [22], [23]. There is therefore a need for designing a learning based 3D facial stylization method that can be tested against the existing rule-based state of the art. It would ideally use an existing dataset. In the context of expression transfer, existing approaches either do not take the morphology and identity of the face into account, or do not take inputs, operating only in latent spaces. As facial shape can vary significantly between stylized faces, we require a method addressing both things. Finally, there exists no research on the study of the relation between stylization and identity preservation, a factor we believe to be key to the parametrization and design of stylization methods.

In summary, in this thesis we aim to provides generic methods for facial stylization for both the case where little and a lot of data is available, and measure its impact on facial stylization. Then we aim to provide a morphology aware expression transfer model.

1.3 Thesis Structure and Contributions

This current introduction chapter provided a global understanding of the context and problems addressed during this thesis. Chapter 2 provides a broad background of the literature related to the presented work, both from a technical and perceptual perspective, and gives background on key concepts such as geometric data representation and deep learning. The contributions of each chapter (from 3 to 5) are presented in Figure 1.2.

In Chapter 3, we first focus on a novel rule-based method to stylize a given human facial scan from very few style examples. Indeed, depending on the target application, in accordance with the style of the narrative, one may want to look like a dwarf or an elf in a heroic fantasy world, or like an alien on another planet. 3D style examples of such styles are scarce, while the number of possible styles is large. We explore how convincingly a person's face can be stylized by our method, while remaining recognizable. In a second

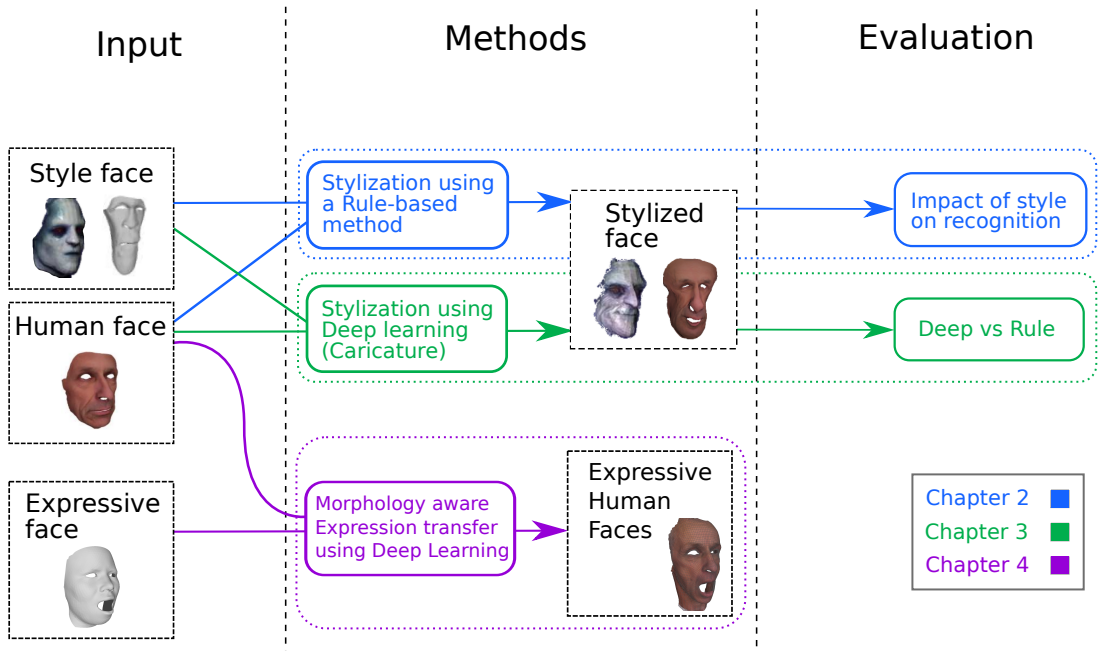


FIGURE 1.2 – Mapping of the chapters of the thesis to their research axes : automatic facial stylization from minimal data and its impact of facial recognition (first row), learning based stylization and its quality versus rule-based methods (second row), and morphology aware expression transfer (third row).

part of this chapter, we present a perceptual study investigating the effect of the degree of stylization on the ability to recognize a stylized person, and the subjective acceptability of stylizations. Results show that recognition rates decrease when the degree of stylization increases, while acceptability of the stylization increases. These results provide insights for achieving good compromises between stylization and recognition, and pave the way to new stylization methods providing a tradeoff between stylization and recognition of the actor.

However, rule-based methods are limited, due to the difficulties in defining style and content. In Chapter 4 , we explore a learning-based approach in the context of a particular application : caricatures. Facial caricature is the art of drawing faces in an exaggerated way to convey emotions such as humor or sarcasm. Automatic caricaturization has been explored both in the 2D and 3D domain. In this chapter, we propose two novel approaches to automatically caricaturize input facial scans, filling gaps in the literature in terms of user-control, caricature style transfer and exploring the use of deep learning for 3D mesh caricaturization. To evaluate and compare these two novel approaches with the state

of the art, we conduct a user study of facial mesh caricaturization techniques, with 49 participants. The obtained results highlights the subjectivity of the caricature perception and the complimentarity of the methods. Finally, we provide insights for automatically generating caricaturized 3D facial meshes.

Having explored rule and learning based approaches for facial stylization, in Chapter 5 we focus on another key perceptual aspect : expressions. Traditional methods transfer expressions without taking into account the facial morphology, which can vary significantly between stylized faces. In this last contribution, we present a new approach for learning-based prediction of facial expressions on 3D facial scans, taking in account their morphology. Improving upon the approach introduced in Chapter 4, our style-based model decouples factors of variations of the 3D face. Current methods for animating faces often involve using blendshapes that are hand-crafted by artists. These approaches are costly and time-consuming to create and compute. Furthermore, when transferred from one face to another, blendshapes can fail to satisfyingly adapt to the target morphology. Leveraging large face scan databases that have recently become available, we train a neural network to transfer facial expressions from one input face to another. We propose a new solution based on deep learning techniques that have successfully been used in the 2D domain.

Finally in Chapter 6 we conclude this manuscript by drawing a general conclusion on the work accomplished, and highlight limitations and a set of promising future research perspectives. The list of publications produced during the thesis can be in found in Appendix A.

Chapitre 2

Background

Facial stylization is closely related to facial appearance, and perception. In this chapter we review the literature on face perception, in the context of virtual characters, in order to look at the factors impacting recognition and the perception of virtual character, and to gain insights into what would constitute style and identity. It is then followed by the literature on facial style transfer. One way of addressing the problem of facial style transfer is to define rules precising what constitutes style and content. If chosen well enough, these rules enable the computation of a convincing stylization result. Yet such rules prove themselves often difficult to design, and provide limited result quality. In order to bypass these limits, a different approach relies on using statistical models to learn the concepts of content (which in our case is the identity) and style. Real world data being most of the time non-linear in nature, non-linear models often perform better than linear ones. Following this line of thought, we first introduce the facial style transfer literature, and look at the limitations of rule-based methods. Then, we review the state of the art on facial representation learning, in two separate parts. The first focuses on linear models. In the second we look at deep learning based style and content separation and transfer in the image and geometrical domains. But first, we will look into the facial perception of virtual characters.

2.1 Face perception

A human face reveals a great deal of information to a perceiver [24]. A look at it enables to identify an individual, and it can reveal mood and intention. Of course, it is not the only way to identify a person. Voice [25], gait [26], [27], silhouette [28] or even clothing can all be used to establish identity, but the face is the easiest and most common way to do it [29]. In the context of virtual character facial stylization, there is a strong need of

perceiving the identity of a face to be similar pre and post stylization, while perceiving it to have the desired style in each case. We first look into facial recognition strategies, then factors impacting recognition, and see how it applies to virtual characters.

2.1.1 Recognition strategies

Global (holistic) and local facial features are both crucial for recognition [30], [31]. Galton *et al.* [24] described holistic perception as "the sum of a multitude of small details, which are viewed in such rapid succession that we seem to perceive them all at a single glance". The concept of holisticity is that faces are recognized better as a "whole" (global appearance), rather than by the recognition of individual parts (nose, eyes, etc.) [32]. This effect is although not confirmed for faces that are inverted [33], or scrambled [32], nor for non-face objects [33] which suggests that holistic encoding is specific to regular faces. However, human attention seems to be attracted quickly by strong local features (large nose, staring eye, etc.). Thus in this case holistic features are perhaps not used, although their perception seems more subtle. Hence we see that two main categories of facial features : holistic and local. They can both be leveraged for the recognition of stylized faces. We now look at which factors impact the use of these features.

2.1.2 Factors Impacting Recognition

When considering real human faces, the hair, face outline, eyes, mouth, and nose have all been identified to be important for perceiving and remembering faces [30], [34]. The nose was found to be insignificant in frontal images [35], but seems at least as important as the eyes or mouth when looking at profiles [30]. The perception and recognition of faces has also been studied in term of aesthetic attributes such as beauty, attractiveness, or pleasantness, where most attractive faces have been found to have the best recognition rate, followed by least attractive then mid range faces [35]. Studies also showed that faces with distinctive features are better retained in memory and are recognized better and faster than typical faces [36]. Average faces are perceived as less distinctive and more attractive [37], [38], although people are more sensitive to small differences in average faces [39]. Overall, the perception of one's face is however not completely understood, but results generally suggest that all parts can contribute to facial recognition, and that strong local features are more recognizable.

While there have been little work on the recognition of stylized characters faces, there is

a rich literature on the recognition of faces outside of our own ethnical group. A large body of research also focuses on understanding cultural differences in face perception. Previous works in this area showed that people recognize faces from their own-race better than faces from other-races, an effect called the Other Race Effect (ORE) [40]. Having experience with faces from other races although does not always cause improvements in recognition for these races [41], [42]. In addition to perceptual experience, motivation to individuate people from the other-race increases the use of holistic strategies when recognizing faces of this race, and seems to be an important factor for ameliorating the ORE [43]. Despite all the studies in this area, the source of the ORE is however still an open question. As a parallel, human and stylized face matching has been shown to be correlated to how human-like the stylized face was, heavily non-human features hindering matching [44].

2.1.3 Perception of Virtual Characters Faces

There is also a large body of work in the Computer Graphics community investigating the perception of virtual characters' faces. It is common to divide the style of a virtual face between realistic and abstract, and to subdivide it into shape, texture, and shading. Shape and texture are two crucial aspects of the appearance of a virtual character [45]-[47]. Shape is more important for perceived realism , texture is the most important for appeal, eeriness, and attractiveness. Lighting can also be an important factor [48]. Furthermore, mismatches between shape and texture realism lead to an uncanny valley effect [45]. Although movement can affect the perception of virtual characters (worsening it if the character is already perceived negatively) much of the information used to evaluate virtual characters is available in a still image [46].

Similarly, stylization can have an important impact on the perception of a character, as Fleming *et al.* [49] showed that partially stylizing a character with the identity of an actor can increase its appeal. Wallraven *et al.* [47] studied the perception of real and computer generated expressive faces with various levels of stylization, and concluded that realistic faces helped having more certainty as to the conveyed expression.

2.1.4 Conclusion

The face has a central importance in recognition, both globally, and as a set of features. Several factors impact recognition, such as the global and feature-wise averageness of

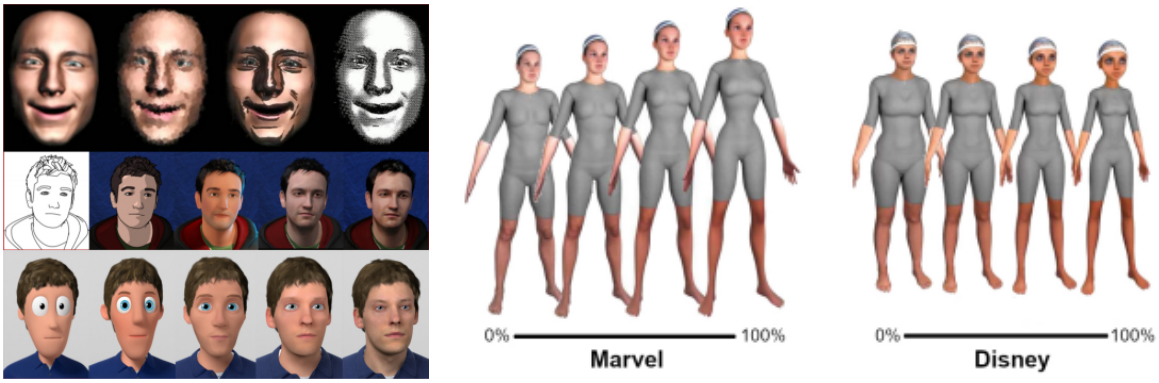


FIGURE 2.1 – Various styles applied to a character through image filters [47], different shadings [46], and different textures and facial shapes [45] (left). Cartoonisation of body scans using two different styles (right) [49]

the face, its race, and the observation viewpoint. In the context of virtual characters, appearance of shape and texture matter for factors such as realism and appeal. To the best of our knowledge, no research has focused on the recognition of stylized faces of virtual characters. We delve into this very topic during this thesis, in Chapter 3. The degree to which a face can be stylized and still remain recognizable is therefore unknown, as well as the degree to which it needs to be stylized to be considered "properly stylized" (looking like an orc, alien, etc. and not a human). Automatic stylization methods can be separated into two main categories : rule-based methods, and learning based method. In the first, a rule is defined to precise what constitutes its style, and the rest, that is typically called content (the identity of the face, in our case). It is then used to change the style while preserving the content. In the second, the stylization rule is not manually defined, but learned from a dataset, using content or style labels to learn what in the data constitute each. We now review the state of the art on the first main family of facial stylization approaches : rule-based methods.

2.2 Rule-based Style Transfer

Rule-based stylization is the approach of stylizing a given data by defining what is considered content, and what is considered style. In the image domain, such methods have moved from using manually defined features, to using pre-learned neural network features. In the 3D domain, methods mostly focus on low level features such as raw vertices and gradients. In the case of facial style transfer, caricaturisation is a context where rule-based

approaches are widely applied, both in the 2D (image) and 3D (geometry) domains. We now review each of these 3 domains : image, 3D, and caricature.

2.2.1 Rule-based Style Transfer for Images

Historically, image style transfer has been accomplished through heuristics such as correspondence maps [50], image analogies [51], frequency separation [52], or edge orientation [53]. Rule based image facial style transfer — whether for portraits or textures — is nowadays usually solved by leveraging pretrained deep neural networks, such as a VGG19 trained on *ImageNet* [54], [55]. Using a content image and a style image, a third image can be computed through optimization, by following a given rule. Gatys *et al.* proposed defining the content as the value of the features of certain layers, while modeling style with features summary statistics (Gram matrices) [56]. Most of rule-based facial style transfer methods are based on the MRF (Markov random fields) style transfer of Li and Wand [57]. It assumes that each pixel in a texture image is entirely characterised by its spatial neighbourhood, which enables to control the image layout at a local level. This makes the result more realistic than with Gatys *et al.* original neural style transfer method. Selim *et al.* [58] proposed to use gain maps to constrain spatial configurations, which can preserve the facial structures while transferring the texture of the style image. Kaur *et al.* [16] improved facial texture transfer by warping the face of the style image to the shape of the face of the content image in order to improve results facial parts wise, while preserving identity using a facial recognition network loss. Champandard's [59] technique can be used to semantically constrain the style transfer using the facial parts. These approaches are although specific to 2D style transfer, and do not transfer when applied on 3D data, as they are dependant on generalist pre-trained neural networks trained on datasets such as *ImageNet*, which has no equivalent in 3D.

2.2.2 Rule-Based Style Transfer for Caricatures

Rule-based caricature methods use *a priori* known procedures to deform a face. Face rule-based methods follow caricature drawing guidelines (e.g. EDFM : Exaggerating the Difference from the Mean) to generate deformed faces with emphasized features. Brennan *et al.* first proposed an implementation of EDFM in two dimensions [60]. They built an interactive system where a user can select facial feature points which are matched against the average feature points, then the distance between them is exaggerated. This algorithm

FIGURE 2.2 – Rule based neural style transfer using the approach of Gatys *et al.*. Content image on the left, style image in the middle, result on the right [56].

was later extended by Akleman *et al.* in 2D and 3D domains [21], [61]. Their software rely on a low-level procedure which requires the user to decide whether the exaggeration of a feature increases likeness or not. In the same spirit, [62] developed a software named *PICASSO* for automatic 3D caricature generation. They used a set of feature points to generate simplified 3D faces before performing EDFM.

EDFM was also used by Blanz *et al.* in an application example of their morphable model [15]. They learn a principal component analysis (PCA) space from 200 3D textured faces. Their system generates caricatures by increasing the distance to the statistical mean in terms of geometry and texture. Statistical dispersion has been taken into account by Mo *et al.* who showed that features should be emphasized proportionally to their standard deviation to preserve likeness [63]. Chen *et al.* created 3D caricatures by fusing 2D caricatures generated using EDFM from different views [64]. Redman's guide [65] not only introduces EDFM but also high levels concepts such as the five head types (oval, triangular, squared, round and long) and the dissociation between local and global exaggeration. These concepts were exploited by Liu *et al.* to perform photo to 3D caricature translation [66]. They applied EDFM with respect to the shape of the head (global scale) and to the distance ratios of a set of feature points (local scale). Face rule-based methods can generate a caricature from an input photograph or a 3D model, but fail at reproducing artistic styles. Different caricaturists would make different caricatures from the same person. The interpretability character of rule-based methods helps avoiding this issue, as it provides user control at a relatively low-level of comprehension, allowing to tune results using artistic knowledge.

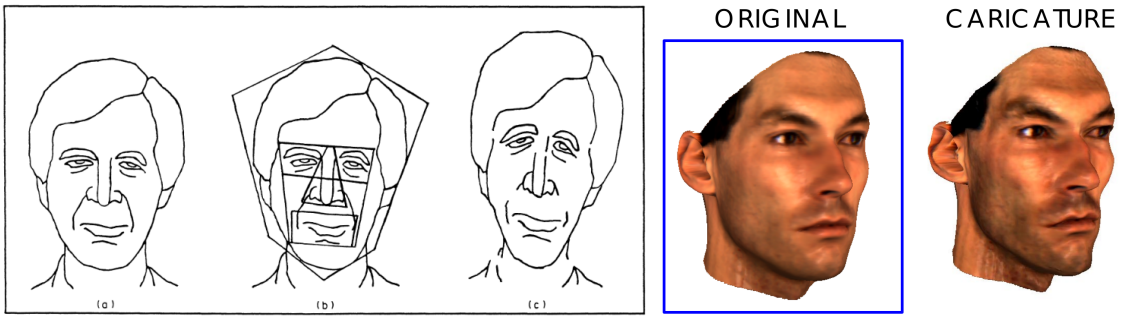


FIGURE 2.3 – Automatic caricature of 2D and 3D faces using the approach of Brennan *et al.* (left) and the approach of Blanz *et al.* (right) [15], [60].

2.2.3 Rule-based Style Transfer on Generic 3D Shapes

Non face specific rule-based methods rely on intrinsic or extracted features of geometrical shapes. Some approaches generalize the concept of caricature beyond the domain of human faces. Eigensatz *et al.* developed a 3D shape editing technique based on principal curvatures manipulation [67]. With no reference model, their method can enhance or reduce the sharpness of a 3D shape. The link between saliency and caricature has been explored by Cimen *et al.* [68]. They introduced a perceptual method for caricaturing 3D shapes based on their saliency using free form deformation technique. A computational approach for surface caricaturization has been presented by Sela *et al.* [69]. They locally scale the gradient field of a mesh by its absolute Gaussian curvature. A reference mesh can be provided to follow the EDFM rule. They show that their method is invariant to isometries, i.e. invariant to poses. General shape rule-based methods can also caricature a 2D or 3D shape without any reference model [67]-[69]. As they do not take into account any statistical information nor the concept of artistic style, they try to link low-level geometry information to high-level caricature concepts, e.g. the fact that the most salient area should be more exaggerated [68].

Other approaches are based on approximating a transformation by segmenting the meshes and matching their parts, then using operations such as deformation or substitution between them [70], [71]. However, the low level and patch-work aspect of these approaches make them limited to simple objects (i.e. furniture).

In the case of body geometry transfer, Fleming *et al.* [49] proposed a stylization technique based on interpolations between the person's 3D scan, an artist-created style template, and an average human mesh. More precisely, they first generate a person-specific style

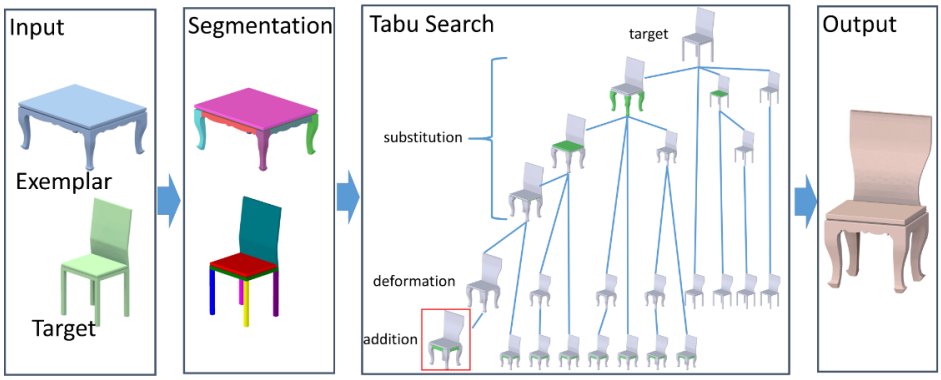


FIGURE 2.4 – Overview of the framework of the stylization method of Lun *et al.* [71]. Shapes are segmented, then altered by a set of elementary substitution and deformation operations.

template by adding half of the Euclidean difference between the average body shape and the person's 3D scan to the style template. Different levels of stylization were then created by interpolating between the person's 3D scan and the person-specific style template.

2.2.4 Conclusion

Rule-based stylization manages to produce convincing results on a variety of domains, such as painting style transfer, facial texture transfer, caricature, or furniture style transfer. It although has limitations. Defining with clear rules what should constitute or not content or style has shown itself to be a hard problem. In some contexts (such as caricatures) the notion of what the style represents can be reasonably modeled, while in most others it has only be achieved through empirical heuristics. In both cases, result quality remains limited, results being often subjectively unconvincing in ways that can not seems to be easily definable by a rule. The limitations of rule-based model could be alleviated by mixing or replacing it with a data driven approach. We now look into a larger part of the literature, focused on learning the concepts of style and content from a given data distribution, instead of trying to manually define it.

2.3 Facial Models

Learning a model of the human face shape provides great advantages for tasks such as face generation, expression transfer, and style transfer (which can all be used for deep

fakes, for instance) [15], [72]. For the latter, most applications in the 3D domain concern facial caricatures. We first review linear models, then the advantages to be brought by using non-linear models. For more detailled information on facial models, the reader may refer to the complete survey of Eggers *et al.* [73]

2.3.1 Linear models

Blanz and Veter originally presented an approach for creating a 3D Morphable Model of the human face [15]. They used principal component analysis (PCA) to infer the distribution of facial geometry for a finite (two hundred) set of face scans. This method allows for the manipulation of face features by manually mapping attributes to vectors in the parameter space. For example, rudimentary expression transfer can be achieved by applying the deformation of an expressive facial mesh onto another. One major limitation of this approach is that the expression is not adapted to the target face. For instance, when applying a mouth-opening expression from one subject to another with a larger jaw, the vertex displacements might differ. To alleviate some of the limitations of this approach, Vlasic *et al.* extend the PCA into a multilinear model which can correlate shape variations caused by identity and expression, and have different linear components for each [74]. Despite several improvements on quality and decoupling over time (e.g. [75], [76]), these approaches fall short when it comes to modelling subtle high resolution facial details. In 2017, Li *et al.* present the *FLAME* model, which combines a linear shape space with articulated parts (jaw, neck, eyeballs) and blendshapes for expressions and pose [72] (Figure 2.5).

Automatic caricature is another case of application of linear models, and can also be seen through the lens of style transfer. Existing linear learning-based methods for caricature generation require paired data as training material, and automatically find rules by relying on pairs of exemplars to learn a mapping between the domain of normal faces and the domain of caricatures. Xie *et al.* [77] proposed a framework that learns a PCA model over 3D caricatures and a Locally Linear Embedding (LLE) model over 2D caricatures, both made by artists. The user can manually create a deformation that is projected into the PCA subspace and refined using the LLE model. Pengfei *et al.* and Liu *et al.* both focused on learning a mapping between the LLE representation of photographs and their corresponding LLE representation of 3D caricatures modeled by artists [78], [79]. In the same vein, but only in the 3D domain, Zhou *et al.* regressed a set of locally

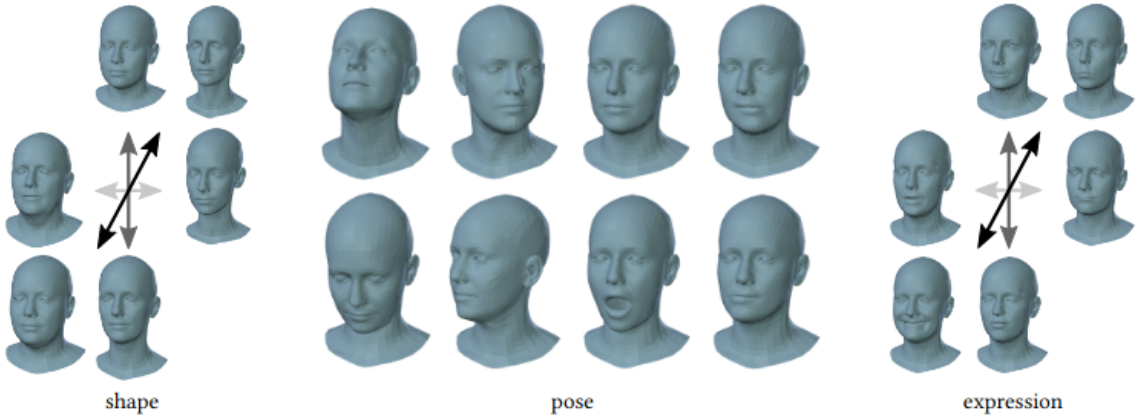


FIGURE 2.5 – Samples from the *FLAME* [72] linear facial shape model, decoupling facial shape, pose, and expression.

linear mappings from sparse exemplars of 3D faces and their corresponding 3D caricature [80]. Clarke *et al.* proposed a physics oriented caricature method [81]. They capture the artistic style of 2D caricatures by learning a pseudo stress-strain model which describes physical properties of virtual materials. All these data-driven approaches are based on paired datasets which requires the manual work of artists. Such datasets are costly to produce, therefore techniques of this kind are hardly applicable.

Because variations of the face shape in the real world are non-linear in nature, we now look at architectures that introduce non-linearity for more accurate representations of the face shape.

2.3.2 Deep Learning Based Models

Deep learning based generative models have two main families : generative adversarial networks (GAN) [82], and autoencoders. Both types of networks are designed to capture the distribution of a particular set of data. GANs are based on two networks, the generator (typically similar to the decoder of an autoencoder) and the discriminator (comparable to a classifier network) which are trained in an adversarial manner, the discriminator being made to learn to discriminate between real data from the dataset, and data generated by the generator, while the generator is taught to generate data able to fool the discriminator. For instance Zhu *et al.* [12] use two discriminators to learn the distribution of two different image domains, and learn two mapping function to map images from one to the other. Using their gan losses as well as a cycle loss, they are able to map one image domain to

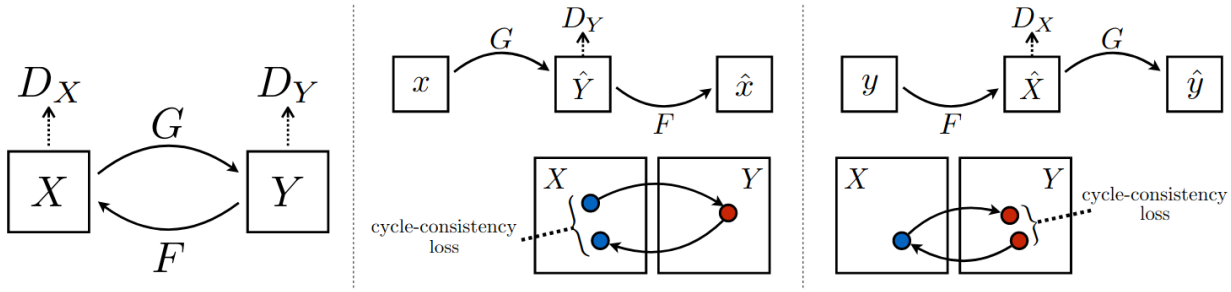


FIGURE 2.6 – Architecture of the style transfer network of Zhu *et al.* [12]. Two cycle-consistent mapping functions are learned between two domains, guided by two discriminators.

the other.

Autoencoders require an encoder, a decoder, and a simple reconstruction loss to compare the output with the input. Most of the information of the dataset (the information shared between each element) is aimed to be encoded into the weights of the network itself, while the remaining (the element specific information) is encoded in a small latent space. Approaches belonging to the style transfer literature typically separate this latent space in two, one part for the style, and one for the content. Style transfer methods often mix both, allowing to stylize specific data instead of random samples thanks to the encoder, and being able to optimize for a given style with the discriminator.

Several contributions have been made in the image domain during the last few years, with applications such as changing human head poses and expressions, turning animal faces into others [13], [83], [84], or human faces into manga faces [85], historically using unpaired one-to-one translation [12], [83], [86], [87], or more recently, unpaired many-to-many translation [13], [84], [88]. In the first case, the goal is to learn a mapping between images of two specific domains, while in the second the aim is to learn one between an arbitrarily large set of image domains. The earliest approaches directly learn mapping networks between each domain to each other, strongly limiting the its scaling, as the relation between the number of mapping networks and the number of domains n is $\mathcal{O}(n^2)$.

Most recent methods work with a partially shared latent space hypothesis, decomposing the latent space into a shared content part, and a class specific style part [13], [84], [88], allowing the use of a shared content encoder and discriminator.

An example of it is FUNIT [13], which uses two encoders, a decoder, and a multi-head

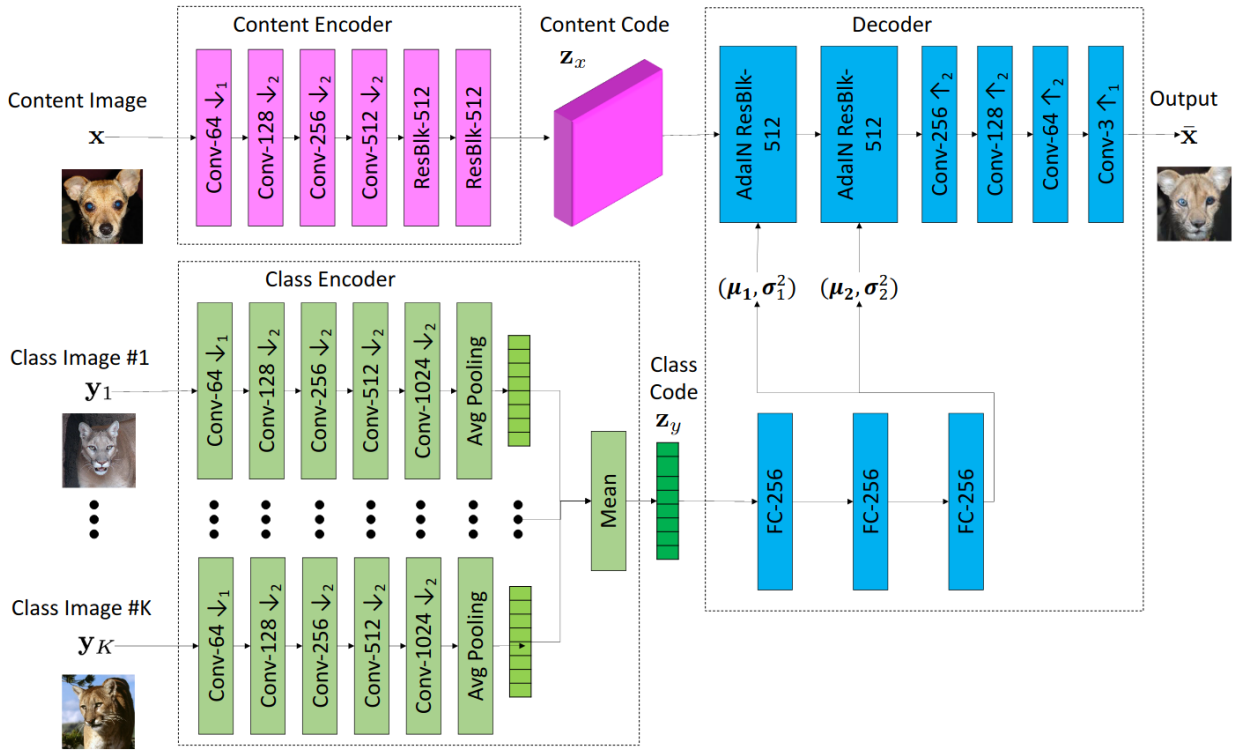


FIGURE 2.7 – Network architecture of the stylization method of Liu *et al.* [13].

discriminator. One encoder encode the content of an image, while the other encode the style. Using a GAN loss, a reconstruction loss, and a discriminator feature loss, the network can be train to learn to map multiple styles from one to another, using the same network weights. The redirection of content information through the content encoder is also guided by its structure, which maintain local information by not going through any fully connected layers, while on the opposite, the information from the style encoder is made to influence the output image globally, by using AdaIN [89] injection.

Other approaches directly learn the transfer functions, content and style being mixed in the same latent space [12], [83], [86]. Some methods apply stylization by using a form of self attention [90] to modulate their feature space [85].

3D Deep Learning

Adapting successful image-to-image translation techniques to 3D geometry could enable better accuracy in transferring attributes from one shape to another. To tackle this ill-posed problem, one of the challenges is to extract style features from the shape structure.

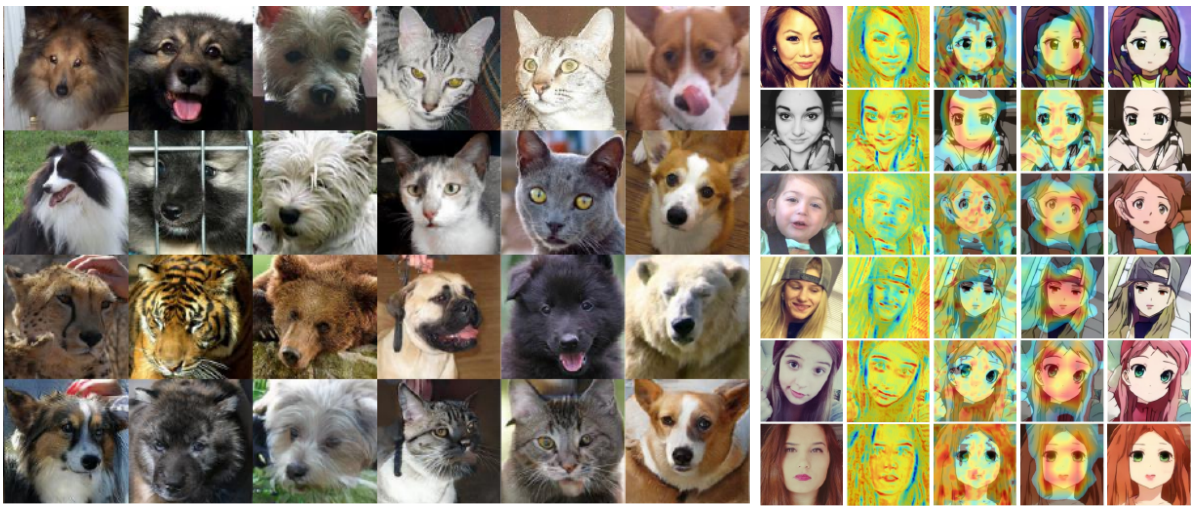


FIGURE 2.8 – Results of the stylization method of Liu *et al.* (left) [13]. Results and attentions maps of the method of Kim *et al.* (right) [85]

Several approaches resort to mapping 3D faces to a 2D domain, and using 2D convolution operators [91]-[93]. Projecting a 3D surface to a 2D plane for 2D convolutions although requires locally deforming distances, which translates to higher computing and memory costs compared to recent 3D convolution approaches, and some high-frequency information loss [94]. Adopting a 3D mesh representation requires application of mesh convolutions defined on non-Euclidean domains (*i.e.* geometric deep learning methodologies). Over the past few years, the field of geometric deep learning has received significant attention [95], [96]. Several approaches have been developed for performing convolutions on 3D meshes over the past few years. Many papers apply spectral graph convolutions by exploiting the connection of the graph Laplacian and the Fourier basis [97], sometimes approximating the spectral filters using truncated Chebyshev polynomials to reduce computational complexity [98], [99]. Several methods apply this approach to 3D faces [100]-[102]. In 2018, Lim et al. introduced SpiralNet, a new convolution operator specialized for 3D meshes [103]. Gong et al. later released SpiralNet++ to refine the approach [104]. This new operator captures local geometric features by using pre-computed spiral sequences on the mesh surface. The N-neighborhood of a vertex is then defined as the N first vertices along the local spiral, and can be used to perform a convolution in the usual manner (Figure 2.9). Akin to a strided 2D convolution, its convolutional receptive field can be increased with no additional memory or compute cost by skipping every K vertex along the spiral. Their results are competitive with previous state-of-the-art methods for reconstruction on 3D face datasets.

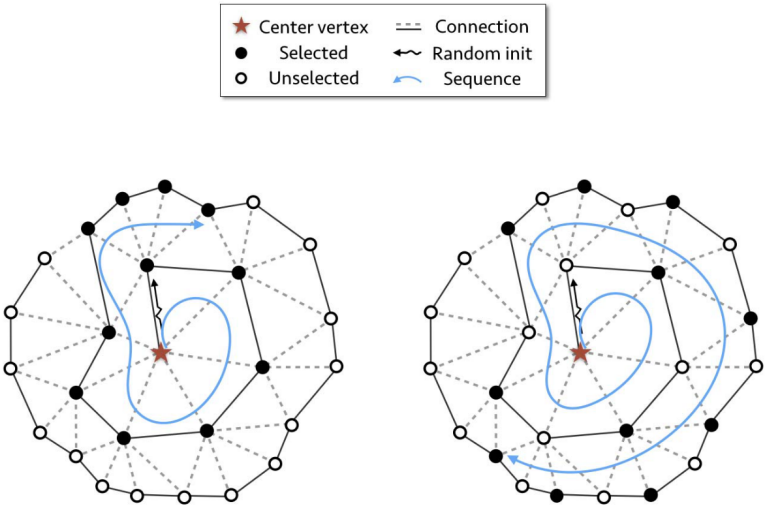


FIGURE 2.9 – The receptive field of a Spiralnet++ convolution, without (left) and with vertex skipping (right) [104].

With the same number of parameters, they outperform previous methods while running several times faster.

Numerous approaches have also been proposed for mesh synthesis and mesh-to-mesh translation in the 3D domaibn using GANs [82]. In order to do shape style transfer, Yin *et al.* [105] proposed a method able to learn general-purpose shape transforms via point displacements. While significant shape changes, e.g., skeleton-to-shape or incomplete-to-complete scans, are possible, this method is supervised and requires paired shapes from two domains. The VAE-CycleGAN of [106] encodes each input set into two separate latent spaces and trains a network to translate codes between the those latent spaces. This allows animation transfer between various shapes of humans, animals, and faces. Yin *et al.* [107] perform cross-domain shape translation by autoencoding them to a shared autocomplete latent space. A translator network is then trained to translate the encoded shapes from one part of the latent space to another, the style being preserved thanks to a feature preservation loss. This method allows furniture style transfer, animation transfer, and low-quality scan to high-quality scan transfer. Segu *et al.* demonstrated a 3D style transfer architecture that uses a PointNet [108] encoder and a decoder with Adaptive Instance normalization (AdaNorm) [109] to perform shape translation on static objects [110].

3D Deep Generative Expression Models

In the context of faces, Ranjan *et al.* were among the first working on 3D deep generative models, using graph convolutions on face meshes and introducing new efficient down-sampling and up-sampling operators [102]. Adopting a multi-scale autoencoder approach, they can accurately represent 3D faces with 75% fewer model parameters than previous attempts (121k to 34k), although showcase no particular style & content separation capabilities. They outperform previous approaches for reconstruction and interpolation of expressions, in addition to demonstrating the ability to synthesize new faces (Figure 2.10). Abrevaya *et al.* use an Auxiliary Classifier GAN [111] to model non-linear variations of 3D face geometry while decoupling identity and expression factors [92]. Given an identity vector, an expression vector and noise, the generator outputs 3D coordinates of the mesh. These coordinates are mapped to a two-dimensional image, which is fed to a discriminator that classifies it into an identity class, an expression class and real & fake classes using standard 2D Convolutional Neural Networks (CNN) techniques. Moschoglou *et al.* introduced 3DFaceGAN for representation, generation and translation of 3D facial surfaces [112]. While they obtain state-of-the-art results at the time for the task of reconstruction, the expression transfer capabilities are more limited. Although these last models achieve state-of-the-art reconstruction results and a decoupled face representation, they require optimizing an input noise vector to fit a given face. Jiang *et al.* address this by adopting an autoencoder architecture, similar to CoMA's [102], with the added ability to decompose the input face into identity and expression latent representations [100]. A fusion module then allows them to reconstruct these representations into a face. Zhang *et al.* extend this decomposition approach with an architecture that enforces distributional independence between identity and expression attributes by design [101]. They obtain great performance on the task of neutralizing the expression or identity of a given mesh. Their method consists in extracting both an identity mesh and an expression mesh from the given face, and adding them together for reconstruction. However, since the decoded identity and expression mesh are respectively expression-agnostic and identity-agnostic by design, adding the two together cannot capture the specific way in which the expression applies to that identity. This severely limits the performance of expression transfer for expressions that differ substantially from the neutral.

FIGURE 2.10 – Comparison of reconstruction of faces between a PCA model and the deep facial model of Ranjan *et al.* [102]

Deep Generative Caricature Models

In the context of caricatures, Chen *et al.* and Liang *et al.* generated 2D caricatures by learning a non linear mapping between photos and corresponding caricatures made by artists [113], [114]. Cao *et al.* proposed a photo to 2D caricature translation framework CariGANs based on a large dataset of over 6000 labeled 2D caricatures [18], and two GANs, namely CariGeoGAN for geometry exaggeration using landmark warping, and CariStyGAN for stylization [20]. CariStyGAN allows to use a reference graphic style, or else, it generate a random style. This framework was first extended by Shi *et al.* [115] with a feature point based warping for geometric exaggeration, then by Gu *et al.* which provides a random set of deformation styles in addition to the random set of graphics styles, offering consequent user control [116]. In the case of unsupervised style transfer, Wu *et al.* [117] then Zhang *et al.* [118] proposed robust methods for 3D caricature reconstruction from meshes, enlarging the set of available in-the-wild 3D caricatures, when used in combination with WebCaricature [18]. Guo *et al.* showed an approach for producing expressive 3D caricatures from photos using a VAE-CycleGAN [23]. Ye *et al.* proposed an end-to-end 3D caricature generation from photos method, using a GAN-based architecture with two symmetrical generators and discriminators [22]. A step of texture stylization is performed with CariStyGAN. The recent works for caricature generation in 3D domain allow to reproduce the style of artists

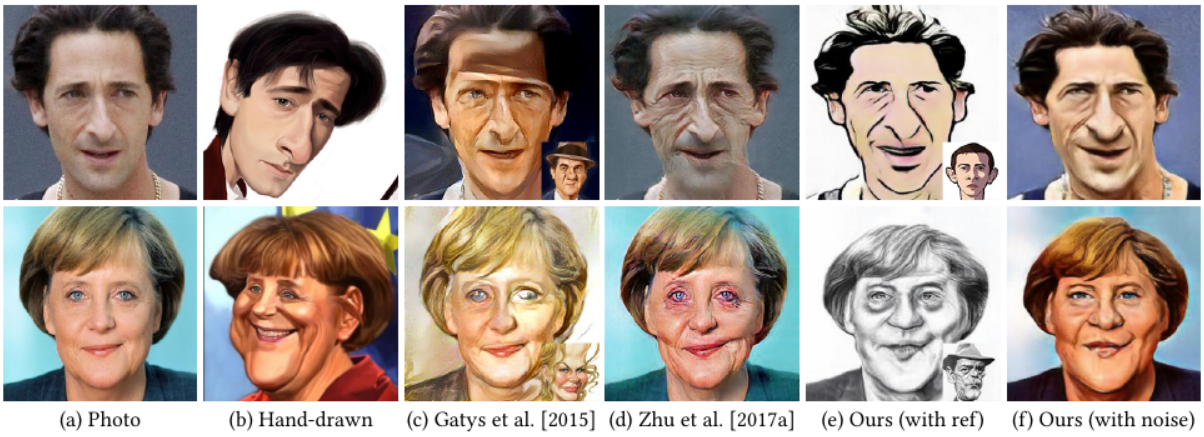


FIGURE 2.11 – Comparison between hand-drawn caricature, and different style transfer methods applied to caricature. This highlight the accuracy of the method of Chen *et al.* [113].

but they do not feature much user control. Ye *et al.* introduced Facial Shape Vectors so the user can choose the facial proportions on the caricature, but this is a quite low-level interaction and thus should be done by an artist [22]. These works also show a weakness from the use of CariStyGAN for texture stylization. CariStyGAN tends to emphasize the shadows and light spots of the photos in order to make the reliefs sharper. In the case of textured 3D models, the shadows and light spots should be induced by the geometry and the lighting conditions, not by the texture albedo. If lighting information is entangled within texture information, changing the lighting condition can make the 3D model appear to be enlighten by non-existent lights.

2.3.3 Conclusion

As we discussed above in Section 2.2, defining a rule-based method capturing all the relevant information for stylizing faces would require understanding how to model every single aspect of it at a low level, which is unfeasible in most cases. Learning based methods instead leverage datasets in order to learn a model of the data, automatically extracting the relevant information. Given enough data to learn, using a linear model of the human faces one can create convincing style transfers of various kind. The non-linear nature of faces although causes such approaches to prove themselves to be limited in several domains, such as expression transfer, failing to capture non-linear relations. Deep learning based methods show themselves more capable in these contexts, although no study has been

TABLE 2.1 – Comparative table of state of the art style transfer methods, with their application domain, key concept, and main limitation.

Name	Conv.	Data	Key concept	Limitation
CycleGAN	2D	Images	Cycle loss	Low-level features
Yi *et al.*	2D	Images	Dual GAN	Retrain for each pair
Liu *et al.*	2D	Images	Shared latent space, Multimodal	Retrain for each pair
Huang *et al.*	2D	Images	Partially shared latent space	Retrain for each pair
Liu *et al.*	2D	Images	Shared GAN and AE, Few-shot	Localized content
Kim *et al.*	2D	Images	Spatial attention	
Choi *et al.*	2D	Images	Partially shared latent space	
Abrevaya *et al.*	2D	3D Faces	Auxiliary discriminators	No encoder
Moschoglou *et al.*	2D	3D Faces	AE + GAN for 3D faces	
Jiang *et al.*	Spectral	3D Faces	Neutralization based	
Zhang *et al.*	Spectral	3D Faces	Neutralization based	Not morphology aware

conducted to compare both approaches on perceptual domains such as caricatures. They are also able to capture the information at a much higher level, allowing applications where linear models fail completely, such as image to image translation. This shows potential for face stylization of various kinds, both in the image and geometrical domains. We include a comparative table showcasing the salient points of style transfer methods mentioned in this section 2.1.

2.4 General Conclusion

The appearance of a virtual character depends on many parameters, such as shape and texture, and is key to perceptual factors such as realism, attractiveness, trust, and agreeableness, most existing studies focusing on the face. The face has a central importance in recognition, both globally, and as a set of features. We focus on the face during this thesis. Its recognition has although not be studied in the context of stylized characters, but there is a large body of literature reporting the importance of race, lighting, observation viewpoint, and both global and local features. The degree to which a face can be stylized and still remain recognizable is therefore unknown. We address this research question during this thesis, in Chapter 3. In terms of ways to stylize, there exist two main families of approaches : rule-based stylization, and data driven stylization. The first aims to manually define what in the data should be considered content, and what should be considered style. It manages to produce convincing results on a variety of domains, such as painting style transfer, facial texture transfer, caricature, or furniture style transfer, although with

limitations. Defining with clear rules what should constitute or not content or style has shown itself to be a tricky problem. Instead of trying to manually define rules, learning based methods estimate them from data. Linear or multi-linear data driven methods allow to capture accurate statistical models of the human face, given human data. While linear data-driven methods can prove themselves to be powerful, but non-linear (deep learning) models allow to model human facial data more accurately, using less parameters and higher level features. This make it possible to manipulate the data in a more semantic way, allowing style transfer between facial images, and geometries.

Chapitre 3

Facial stylization and its impact on face recognition

3.1 Introduction

While the building of accurate and interpretable 3D facial model has received considerable attention, automatic stylization of 3D faces has mostly been studied in the particular case of caricature. There is although a need for such a stylization approach, that would work for any given style, in order to fit a real human character to a specific narrative.

Indeed, depending on the target application, one may want to look like a dwarf or an elf in a heroic fantasy world, or like an alien on another planet (e.g. Sigourney Weaver in Cameron's Avatar). There exist approaches in the 2D domain to learn a particular style from a dataset and use it to stylize faces, but a given style usually exist only as a few examples, especially in the 3D domain. There is therefore considerable interest in having a method that allow stylization from as little as one example. In this chapter, we consider stylization to be the process that, from contents (identities) A and B, produces a content C similar to A, but with the style (human, alien, etc.) of B. In the case of a virtual character, "C similar to A" means that people could recognize A when watching C. For instance in the case of an actor being stylized as an orc, we want the produced orc face to be recognizable as the actor. We design a novel approach to stylize both geometry and texture, in a rule-based manner, allowing to produce faces of various stylization levels, from minimal data. However, despite the rising interest in stylizing virtual characters, no formal studies have been conducted to assess the ability to recognize stylized characters. Can we recognize the human face that has been stylized into a virtual character ? Is there a limit in the stylization that can be applied ?

In the second part of the chapter, we focus on the ability of recognizing a person's face (A) that has been stylized into the face of another non-human virtual character (B). In particular, we create stylized representations (C) of a number of actors, in order to evaluate the ability of viewers to recognize the original actor (A). In this context, we addressed the following questions : What is the effect of the degree of stylization on the ability of viewers to recognize an actor ? What is the effect of the degree of stylization on the acceptability of the stylization by viewers (the judgement of the quality of the result in respect to the original style character) ? E.g., do they consider the result of an orc stylization to be an orc character ? Is the ability of viewers to recognize a stylized actor affected by how much non-human is the virtual character used for the stylization ?

To answer these questions, we conducted a study (N=24) investigating the effect of the degree of stylization on the ability to recognize an actor, and the subjective acceptability of the stylization results.

The work in this chapter resulted in publications in IEEE VR 2019, and ACM SAP 2020 (Appendix A).

3.2 Stylization Method

In this section we introduce the face stylization method used in our experiment. The method is adapted to enable controlling of the degree of stylization. The approach separates the stylization into two parts (geometry and texture) and is illustrated with high quality human facial meshes, captured with the camera system of Danieau *et al.* [44], that makes use of 48 calibrated cameras on a spherical rig, taking a picture at the same time to facilitate the use of photogrametry for reconstruction. Non-human facial meshes are obtained from the *Paragon* [119] free assets. All facial meshes are cropped and retopologized in the same manner.

3.2.1 Geometry Stylization

As mentioned in section 2, the shape of several facial features is an essential part of a face identity (e.g., mouth, nose). Therefore, to transfer the identity of one's face to another, the geometrical particularities should be transferred, whether it is the size of the jaw, the angle of the nose, or the eye-to-eye distance. All meshes are first normalized to

FIGURE 3.1 – Left : actor scanned face ; Right : non-human character face ; Middle : 5 different levels of stylizations, with style levels 0.40, 0.55, 0.70, 0.85, 1.0.

the same topology, and then passed through a geometry relative style transfer method, which computes the variations of a given face from an average human model to capture the particularities of each specific face. As all the human faces in our studies were male Caucasians, we therefore used as average model the default Caucasian male facial model from *MakeHuman* [120].

In this chapter, we are interested in designing a rule-based facial stylization method, that can be used for any given style. Our second focus is on exploring the effect of the degree of stylization, i.e., how much the features of the non-human character are transferred onto the person's face on a continuum from no stylization (the person's face) to a strong stylization (the person's face with strong features of the non-human character), as depicted in Figure 3.1. Our method enables controlling the degree of exaggeration of the person's features onto the non-human character, i.e., producing stylizations on a continuum from the non-human face to the non-human face with features of the corresponding person. We use the following equation :

$$M = M_h + w(M_n - M_a) \tag{3.1}$$

where M is the set of vertices of the final mesh, M_n is the set of vertices of the non-human mesh, M_h the set of vertices of the human mesh and M_a the set of vertices of the average human. The scalar weight w controls the importance of the non-human in the deformation (i.e., the degree of stylization increases with w). Examples of geometry stylization at five different levels are presented in Figure 3.1.

3.2.2 Texture Stylization

Additionally to geometric stylization, texture stylization enables the transfer of one individual's face textural characteristics to a non-human texture. The method computes

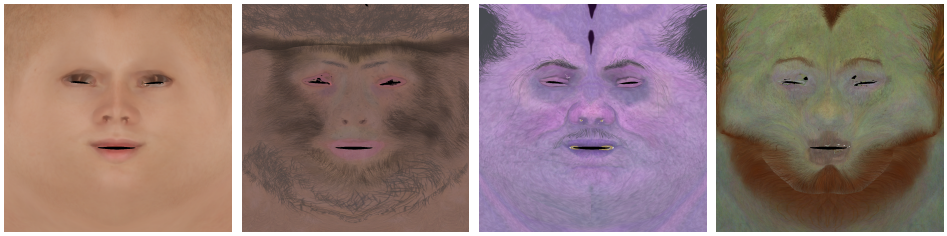


FIGURE 3.2 – The average human texture, and the three style textures used for the study.

FIGURE 3.3 – Left : Original identity ; Right : style texture ; Middle : Stylized textures at style levels 0.40, 0.75, 1.0.

texture differences from an average (reference) human facial texture. As all the actors in our studies were male Caucasians, we used the default Caucasian male texture from MakeHuman [120] as the average facial texture 3.2. Our average human texture display an artificially flawless skin, to enable transferring facial features such as hair, scars or wrinkles. Finally, the texture stylization method is based on a neural network optimization [56], where the following loss function is minimized :

$$loss = ((T - T_h.w_{tex}) - (T_n - T_a.w_{tex}).wc_{tex})^2 \tag{3.2}$$

where T is the result of the style transfer process, T_h is the human texture, T_n is the non-human texture, T_a is the default human texture, w_{tex} is the weight controlling how pronounced the identity features are in the output and wc_{tex} is the weight controlling the balance between content (non-human) and style (identity). We minimize the relative neural style feature difference between the human texture and the average human texture, and the target stylized texture and the average style texture.

We also introduced an additional texture normalization step in order to avoid any influence in our experiments due to differences in colorimetry resulting from the scanning process. In particular, we found that textures generated with the same level of stylization (especially

FIGURE 3.4 – Three stylized textures, un-normalized (top), then normalized (bottom). Normalisation helps with colometry and textural features (e.g. forehead spots, hair)

when not using a full stylization) displayed visually varying degrees of perceived stylization when the normalization step was not used. For instance, the three top examples of Figure 3.4 present differences in intensity, as well as differences in the appearance of some of the key visual features of the non-human model (diamond shape on the forehead), which were more consistent after normalization (right-most three examples). We therefore normalized the original textures (before stylization) by aligning the mean color distribution of the textures used with :

$$T_h = \hat{T}_h - (mean(\hat{T}_h) - targetcolor) \tag{3.3}$$

where $\hat{T}_h$ is the original scanned texture, $mean()$ is the RGB-wise average of an image, and $targetcolor$ is the RGB value that T_h's mean will be aligned on (we used the mean color of a reference texture with good colometry). Examples of texture stylization (including normalization) at 3 different levels are presented in Figure 3.3. Various stylization levels are presumed to mean that there exist a trade-off between style and identity.

3.3 Study : style and identity trade-off

Style transfer raises the question of the trade-off to reach between the original content and the style to ensure that both are identifiable in the stylized content. In the particular

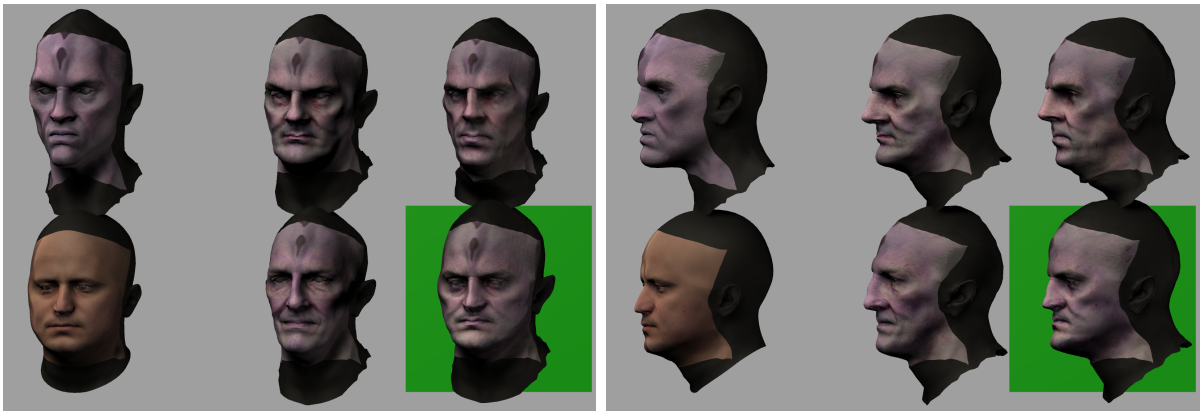


FIGURE 3.5 – Front and profile views for a trial example of the experiment. Participants could switch between the views by pressing spacebar. The face selected by the participant as being the stylized face of the actor is highlighted in green.

context of face stylization, we are exploring the following questions : To what degree a human face can be stylized and still remain recognisable ? To what degree a stylized face can be considered as stylized enough ? To answer these questions, we conducted a study exploring the relationship between the degree of stylization and the recognition performance for a set of non-human styles. In addition to recognition accuracy, we also explored what constitutes an acceptable level of stylization. In summary, the main hypotheses of the experiment were :

H1 - Lower degrees of stylization will result in higher recognition rates.

H2 - Higher degrees of stylization will increase subjective recognition of the original model species. E.g., someone stylized as an orc might not be considered to look like a proper orc at low stylization levels, but will at high stylization levels.

3.3.1 Population

Twenty-four participants took part in the experiment (6 females). They were between 23 and 61 years old (mean and SD age : 41.8±10.9), Caucasian, and were recruited from our laboratory among students and staff. They were all naive to the purpose of the experiment, had normal or correct-to-normal vision, and gave written and informed consent. The study conformed to the declaration of Helsinki. They were not compensated for their participation. None of the participants knew the human faces used in the study.

3.3.2 Stimuli

Ten human face scans were used in the study. In order to reduce recognition due to outliers (bearded face, under-represented gender, etc.), the 10 faces were all unbearded Caucasian males, with neutral facial expressions (age $mean = 44.86$; $std = 7.20$). As eye color could have also been an easily distinguishable feature, it was not transferred during the stylization process.

To explore the effects of the degree of stylization, we used 5 levels of stylization (from low to high, see Figure 3.1). As texture and geometry stylization weights in Equations 3.1 and 3.2 are not necessarily equivalent, we experimentally selected the following weights for the geometry (0.40, 0.55, 0.70, 0.85, 1.00) and texture (0.60, 0.70, 0.80, 0.90, 1.00). We also used 3 non-human faces, obtained from the Paragon [119] free assets : an *Alien*, an *Orc* and a *Monkey* 3.2. They were picked amongst 5 non-human faces used in a pilot study, respectively as the most recognized, the least recognized, and the one which recognition was the least correlated with the others. They were selected for their relative proximity to a normal human face, where the Alien face was the closest to the average human face (vertex to vertex difference), followed by the Monkey, then the Orc.

Regarding the presentation of the faces, we controlled the lighting parameters and the viewing angle. For example, Jonhston et al. [121] showed that bottom lighting makes harder to identity familiar faces. Similarly, Hill and Bruce [122] showed the importance of top lighting for face recognition by matching surface images of faces to determine whether they were identical. Therefore, three directional lights coming from above were used, with intensities 0.9 (front), 0.3 (right), and 0.25 (left). During the experiment, participants could switch between two point-of-views, where meshes were viewed either 30° or 90° from the right (see Figure 3.5).

3.3.3 Protocol

The task required participants to recognize one face (human scan, displayed on the bottom left corner of the screen) among four stylized faces (displayed on the right side of the screen). The non-human mesh was also displayed for reference (top left corner of the screen). Figure 3.5 displays a trial example. Participants were instructed to select the stylized face which they thought matched the human face using keys identified on the keyboard. Character faces were presented at a 30° angle at the beginning of the trial, and

could be switched back and forth to the 90° angle by pressing the space key. Each trial lasted a maximum of 20 seconds, and was displayed on a 24-inch screen located 50cm away from participants.

We used a between-subject design, as each participant was shown stylization results for only one non-human out of three (Alien, Monkey or Orc) during the study. In total participants performed 100 trials, displayed in random order : 10 Actor Faces × 5 Stylization Levels (low to high stylization) × 2 Repetitions. In each trial, the four stylized faces presented to participants all had the same degree of stylization : one of them was always the correct stylized actor, while the three others were selected randomly from the nine remaining actor faces. Before the experiment, one training trial was included to familiarize participants with the user interface, using a human face never used again in the experiment.

Finally, participants were asked to fill in a post-study questionnaire to evaluate their subjective perception of the stylized faces, both in terms of recognizing the actor (Identity subjective rating) and the species of the reference model (Species subjective rating). For each Actor Face, participants were showed simultaneously all 5 stylization levels, as well as both the reference actor and non-human faces, and asked two questions : "From the least stylized face to the most, up to which stylization level can you recognize the individual ?" and "From the most stylized to the least, down to which stylization level would you consider the face to be of the same species than the reference non-human ?" Participants required approximately 35 minutes to complete the entire experiment.

3.3.4 Results

Recognition rates

To analyze recognition rates, we first performed a 3-way mixed-design Repeated Measure Analysis of Variance (ANOVA) with within-subject factors Stylization Level and Actor Face, and between-subject factor Non-Human Model. Effects are then explored further using Neuman-Keuls post-hoc test for pair-wise comparisons of means. We first found a main effect of Stylization Level ($F_{4,88} = 34.99$, $p < 0.0001$), where post-hoc analysis showed that recognition rates were significantly decreasing with increasing stylization levels ($p < 0.05$), except between Levels 3 and 4 (Figure 3.6). More specifically, we found that recognition rates ranged on average from 92% (lowest level of stylization) to 66% (highest level of stylization), where chance level was 25%. We also found a main effect of the Non-Human

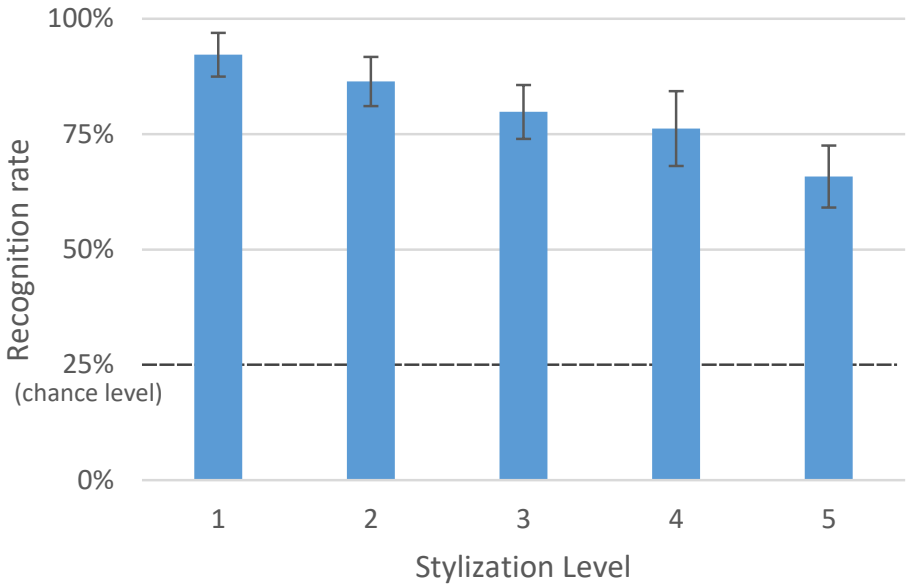


FIGURE 3.6 – Main effect of Stylization Level on recognition rates. Error bars represent standard deviation.

Model presented to participants ($F_{2,22}=3.52$, $p=0.047$), which was not confirmed by the post-hoc analysis (Alien-Monkey : $p=0.656$; Alien-Orc : $p=0.057$; Monkey-Orc : $p=0.059$), even though there seems to be a tendency for actors to be on average less recognized when stylized onto the Orc model than onto the other two. We finally found a main effect of Actor Face ($F_{9,198}=3.955$, $p<0.001$), where post-hoc analysis showed that some actors were on average significantly more recognized than others. Moreover, recognition rates were high on average, even for the least recognized actors, ranging from 73% to 89%. We however did not find any interaction effect between factors.

In order to further understand the relation between recognition rates and the degree of stylization, we then computed person correlations between these two variables (averaged over the non-human models for each participant, as we did not find a significant effect of this factor in the first analysis). Results showed that recognition rates are negatively correlated with the level of stylization ($r=-0.59$, $p<0.001$). All these results therefore support **H1**, showing that *lower degrees of stylization result in higher recognition rates.*

Subjective scores

At the end of the experiment, participants were asked to provide subjective ratings for each Actor Face regarding the stylization level from which they considered that they could not recognize each individual (Identity subjective rating), as well as the level from which they considered that the stylized face was not anymore from the same species as the reference non-human (Species subjective rating). The distributions of both Identity and Species subjective ratings are displayed in Figure 3.7.

We first analyzed the Stylization Level selected by participants for each Actor Face for both questions by performing two mixed RM ANOVA with within-subject factor Actor Face and between-subject factor Non-Human model. We only found a main effect of Actor Face for the Identity subjective ratings ($F_{9,189} = 2.777$, $p < 0.01$), where post-hoc analysis showed that overall there were mostly no significant differences between actors, except from slight differences between a few extremes. These results suggest that the displayed model or actor did not influence the level at which participants considered that they could not recognize an individual or species.

To further explore this data, we transformed the subjective ratings to represent the probability of recognizing the identity or species at a given level stylization. For the identity, answers were converted to a binary format where a stylization level for an individual was assigned a value of 1 if it was below or equal to the level selected by the participant (i.e., participants answered that they were able to recognize the identity at this given level of stylization), and 0 otherwise. The opposite was performed for the species subjective ratings, meaning that the species of the model at a given level of stylization would be considered to be recognized by participants if it was above or equal to the subjective rating. We then performed two mixed RM ANOVA with within-subject factor Stylization Level and between-subject factor Non-Human model. We found a main effect of Stylization Level for both identity ($F_{4,84} = 150.04$, $p < 0.0001$) and species ($F_{4,84} = 100.43$, $p < 0.0001$), where post-hoc analyses showed that all the stylization levels were significantly different and decreasing for identity (all $p < 0.05$), and significantly different and increasing for style (all $p < 0.001$) except for the two highest levels, which supports **H2**. These results are displayed in Figure 3.8, concurrently with the objective recognition rates analyzed in Section 3.3.4. It is also interesting to notice that participants' subjective perception of their ability to recognize an identity was drastically lower than their objective performance, especially for higher stylization levels, which is further discussed in Section 4.5. Finally,

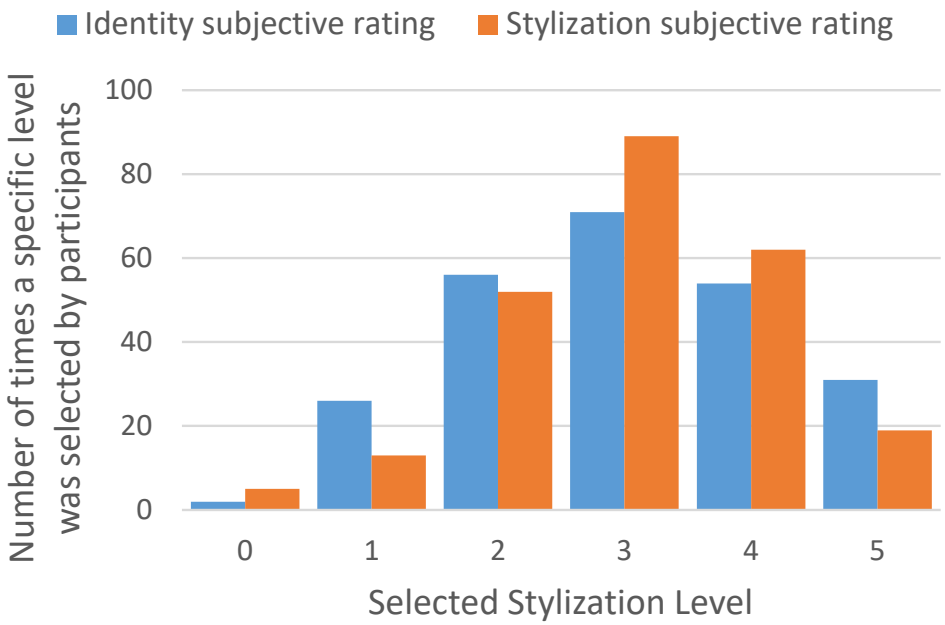


FIGURE 3.7 – Distribution of the Identity and Stylization subjective ratings.

identity subjective ratings were found to be negatively correlated with the species subjective rating ($r = -0.74$, $p < 0$), suggesting that identity preservation and stylization might be based on a compromise.

Non-uniformity of the choices repartition

To explore whether the distinctiveness of the actors' faces (i.e., in term of how different they are from the other faces) influenced user performance, we computed for each actor scan the geometrical difference to the average of the ten faces used in the study (sum of vertex-to-vertex distances). We then explored potential correlations with the number of times each actor face was selected in the experiment, with the average recognition rate per actor, as well as with the average precision per actor (number of correct recognition divided by number of times the given face was selected). We only found a slightly significant positive correlation of the geometrical difference with precision ($r = 0.63$, $p = 0.0497$), suggesting that more distinctive faces (in terms of geometric distances) might tend to be slightly more accurately recognized than less distinctive ones.

However, as geometrical differences might not capture the subtle idiosyncrasies of a person's face, we also decided to explore facial differences in terms of features extracted from a view of each human face using a facial recognition neural network [123]. With this metric, we

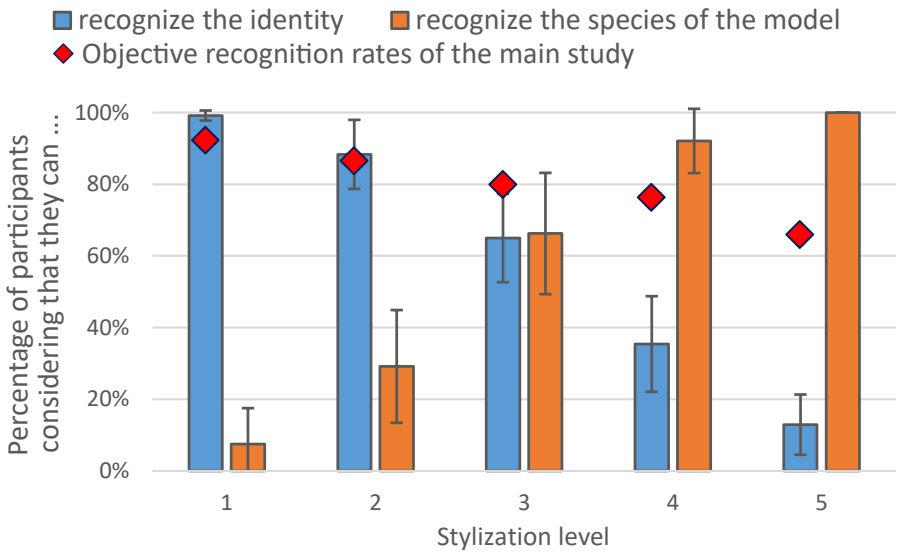


FIGURE 3.8 – Percentage of participants considering that they can recognize the identity or the style at a given stylization level. It is interesting to notice that subjective ratings are usually lower than objective performance, especially for higher stylization levels.

also found a similar positive correlation between how different the face was and precision ($r = 0.61$, $p = 0.0603$). Interestingly, we also found a significant negative correlation with the number of times a face was chosen ($r = -0.79$, $p < 0.01$), suggesting that less distinctive faces were more often chosen by participants.

These results suggest that novel metrics (e.g., neural network based methods) can provide face similarity metrics similar to common geometric methods, despite working in other domains (e.g., 2D domains for images used with neural networks). This therefore opens a possibility for novel manners of analyzing such effects, which is further discussed in Section 4.5.

Time spent

Participants spent an average of 10.8 seconds on each trial (min=5.0, max=16.5), with an average total time of the study of 18 minutes. Correlations between the total mean time spent and performance were measured for each style weight level and are all negative, but only significant for the lower level of stylization ($r = -0.535$, $p < 0.01$). This result suggests that participants needing time to recognize an identity perform worse on average, especially when faces are only partially stylized.

3.4 Discussion

In this chapter, we have studied the impact of stylization on recognition, for different identities and stylization levels, when both facial geometrical and textural styles are transferred. More precisely, we conducted a study involving three different styles and five stylization levels. Recognition was found to decrease linearly with the level of stylization, ranging from 93% recognition accuracy at the lowest level of stylization to 66% for the highest level. A similar tendency was observed from subjective ratings collected at the end of the experiment, where participants were directly asked to select the level of stylization up to which they considered that they could recognize the individual. Interestingly, while the average recognition rate reached 66% for the highest level of stylization in our experiment (where chance level was 25%), participants' subjective perception of their ability to recognize an identity was drastically lower, and below 50% for the two highest levels of stylization (respectively 35% and 13%). This result suggests that while participants might consider that they are not able to recognize a stylized individual when it is presented alone, the ability to recognize the person might actually be higher when presented amongst other stylized individuals. However, the limit to which this is possible, and facial features potentially responsible for such differences, cannot be determined from the experiment presented here, and would therefore require to be explored in further studies.

As for the perception of human faces, we also found slight differences in the recognition of stylized individuals. These differences were however small, and interestingly recognition rates were always relatively high (min : 73% ; max : 89%) and all above chance level. It is also important to point that the experiment was conducted using only 10 human faces, which were selected to avoid introducing any recognition bias (e.g., ethnicity, age, gender). However, future experiments involving a broader number and variety of faces would allow stronger generalisation of the findings to other faces.

Similarly, in order to explore potential effects of the target character face, we also used three different non-human faces in our experiment, which were selected for their relative proximity to a normal human face (the Alien face was the closest to the average human face, followed by the Monkey, then the Orc). Unexpectedly, we did not find any effect of the character model on participants' objective or subjective ability to recognize individuals. However, we observed a slight tendency for individuals to be on average less recognized when stylized onto the Orc model, most different from the average human face. Despite

being non-significant, this tendency raises the question of how different the character face can be while stylizing an individual, and the limits of the stylization.

Concurrently to the ability of observers to recognize the stylized actor, we also collected subjective ratings regarding whether observers considered the stylized faces to be of the same species than the reference non-human character. This subjective rating was positively influenced by the degree of stylization, where only 8% of the participants answered that they considered the lowest level of stylization to be of the same species, while 92.5% considered it to be the case for the highest level of stylization. The degree to which participants were convinced by the style of the result was also found to be negatively correlated to the degree to which the actor face was recognized, therefore suggesting that a balance might be required to enable observers to recognize the actor while providing sufficient cues about the type of character he/she is stylized into. It is also possible that achieving both high recognition and stylization might be currently limited by the stylization method available. Other stylization approaches could therefore be explored in the future to reduce this limitation.

As studies in Psychology demonstrated that distinctive faces affect recognition [36], we also explored a possible influence of the distinctiveness of the actors' faces on participants' ability to recognize stylized faces. We looked into this by assessing the level of distinctiveness using two methods : the geometrical distance of each face to the average of the faces, and the distance of their projection in a facial recognition neural network to the average of the faces. We found some small correlations between distinctiveness and recognition precision for both methods, suggesting that more distinctive faces might tend to be slightly more accurately recognized when stylized than less distinctive ones. While these tendencies are similar to results observed for facial recognition, it also suggests that novel metrics, such as those based on neural network methods, might provide novel manners of analyzing such effects. Also, beyond considering the limits of a metric for measuring objective face distance, face perception remains a subjective matter not only based on the position of faces in a shape or texture space, but also on aspects such as attractiveness and familiarity [124], which could also be evaluated using such metrics in the future.

Despite providing the first insights about the influence of stylization level on the perception of faces, the controlled settings of our study might have introduced some limitations which should be explored in further studies. For instance, we presented participants with only 2 possible points of view during the experiment, as well as with a static lighting and no

facial animation. As recognition of an individual involves more features than a static face, future studies could explore the influence of other factors on the ability of viewers to recognize an individual, including facial expressions, body shapes to motion, or the tone of the voice. Also, for this first study on the topic, the stylization process applied a global stylization to the face, as to our knowledge there is no information about how stylizing different parts would affect recognition, while it was simultaneously restricted for some areas of the face (i.e., no ear or eye stylization), which might have reduced recognition abilities. However it would be interesting in the future to identify which are the features most important for stylization, which would open the door to more localized stylizations. Finally, our experiment involved the use of a single facial style transfer method, as there are currently no other method available in the literature. With the growing interest for this field of study, new methods might appear in the future and would benefit from comparison using our experimental protocol as a baseline.

3.5 Conclusion

In this chapter, we have presented a novel rule based method allowing geometrical and textural stylization of faces from one example. It works in a relative manner, based on an average human face, and an average style face, making use of raw vertices as well as deep pre-learned features. Then, we have investigated the recognizability of stylized faces for different identities and stylization levels. We have presented a study, based on 3 character faces and 5 stylization levels, which main goal was to measure the balance between style and recognition. Recognition was found to linearly decrease with stylization, while participants' ability to match the stylized content with the original character species increased linearly with the level of stylization. Recognition is thus inversely proportional to being convinced by the style. These results provide new insights about necessary compromises between stylization and recognition, and pave the way for the new field of study of 3D facial style transfer, combining both knowledge in facial perception and computer graphics. We explored the interest of rule based methods for cases where few examples are available, and some of the limits of stylizing faces through the first user study on facial recognition of strongly stylized faces. How much these limits are inherent to stylization, or to rule-based stylization, remains to be seen. Further work will include designing a learning based approach, and comparing it to rule-based methods.

Chapitre 4

Comparing rule-based and deep learning based stylization

4.1 Introduction

In the previous chapter, we explored the impact of facial stylization on recognition, using a rule-based stylization approach. In the field of automatic image style transfer, deep learning based approaches have overtaken existing rule-based methods in several domains, by a large margin. In this chapter, we explore whether this approach could bring similar improvements in the 3D facial domain. We focus in particular on a kind of 3D facial style transfer that received significant attention in the literature, and has dedicaced rule-based approaches : caricatures.

Caricatures have been used for centuries to convey humor or sarcasm. References can be found during the Antiquity with Aristotle referring to these artists as “grotesque”, or in the works of Leonardo Da Vinci who was eagerly looking for people with deformities to use as models. Caricature can be defined as the art of drawing persons (usually faces) in a simplified or exaggerated way through sketching, pencil strokes, or other artistic drawings. Caricatures have been commonly used to entertain people, to laugh at politics or as a gift or souvenir sketched by street artists. Automatically generating caricatured avatars is a key issue, as having artists manually creating caricatured avatars would not be feasible for applications involving large numbers of users, such as games or photo filters apps. Let us consider a 3D mesh representing the user’s face (either using 3D scanning or computer vision methods to build 3D shape from a minimum set of images). An automatic caricature system should maintain the relative geometric location of facial components, while emphasizing the subject’s facial features distinct from others. While

different caricature experts would generate different styles of faces (more or less cartoonish style for example), they would all be exaggerating facial traits of the individual [60], [63], [114]. The ability of creating a variety of plausible caricatures for each single face is therefore a key challenge when automatically generating caricatures, as different artists would create visually different caricatures, which should also be taken into account when evaluating the subjective quality of the results.

Previous works for the generation of 3D caricatures can be separated into two main families : interactive and automatic methods. Interactive methods offer tools to caricature experts to design the resulting caricature [61], [125]-[127], while fully automatic methods use hand-crafted rules [60], [63], [114], often derived from the drawing procedure of artists. However, these approaches are typically restricted to a particular artistic style, e.g., sketch or a certain cartoon, and predefined templates of exaggeration. From the works in the literature in other domains, two different solutions could be envisioned to automatically generate caricatures. First, in the context of exaggerating distinct features, [69] proposed a generic method to exaggerate the differences between the 3D scan of an object and an average template model of such type of object. However, this method has never been formally evaluated for human faces. Second, deep learning methods could be considered. As mentioned above, automatic methods mainly use hand-crafted rules that may fail to capture some complex choices made by caricature experts. In contrast, generative adversarial networks (GANs) are a promising mean to attempt to learn these choices based on a set of examples made by experts, without being limited to hand-crafted rules, but it has been never applied for the generation of 3D caricatures. The main goal of this chapter is to propose and evaluate novel methods for the automatic generation of 3D caricatures from real 3D facial scans, first with a rule-based method, in order to keep tunable and interpretable parameters, and a deep learning method, to leverage real caricature data and hence generate caricatures closer to real ones. The main hypotheses we wish to address in this chapter are :

— **H1** : the specialization of generic exaggeration methods for human faces should enable to produce convincing caricatures. To this end, we adapted the generic method proposed by [69] in order to generate caricatures by exaggerating facial features from a 3D face scan. This method has two main stages, one based on a curvature EDFM (Exaggerating the Difference From the Mean), and another based on a nearest-neighbors search in a 3D caricature dataset, to apply the proportion exaggeration.

— **H2** : deep learning should enable to overcome some of the limitations of rule-based methods by their ability to generalize based on a set of examples. Thus, we designed a method leveraging advances in the field of GAN-based style transfer, which has shown great success in the 2D domain, for instance on drawn caricatures [20].

— **H3** : both methods should reach and overcome the state of the art results when trying to automatically generate caricatures from a human face 3D scan. To assess the advantages and disadvantages of the proposed methods, we conducted a perceptual study considering the base method proposed by [69] and an additional EDFM method [21].

The remainder of the chapter is structured as follows. Sections 4.2 and 4.3 present the proposed rule-based and deep learning-based caricaturization methods respectively. Then, Section 4.4 presents the perceptual evaluation of the proposed methods with state-of-the-art methods. Finally, we discuss the results and provide insights on the automatic caricature generation in Section 4.5.

The work in this chapter resulted in a publication in Frontiers in Virtual Reality A.

4.2 Rule-based User-Controlled Caricaturization

We present a novel method featuring short computation time and providing meaningful user control over the generated caricatures. It is based on two main modules depicted in Figure 4.1 (in green and in yellow). First, a curvature exaggeration module (in green) enhances the facial lines by applying EDFM technique to the main PCA scores of the mesh gradients of the input face. This emphasizes only the 3D surface details such as ridges, peaks and folds, and does not affect the global shape of the face (such as eyes, nose and mouth relative positions). Second, a proportion exaggeration module (in yellow) leverages compositions of real artists (see Section 4.2.1) to caricature the general shape of the face. It projects the input face into a 3D caricature shape space thanks to a kNN regressor. This process applies a smooth and large scale deformation to the input face while preserving its local features. The curvature exaggeration and proportion exaggeration modules are thus complementary. They are combined to provide the user with a bilateral control (small

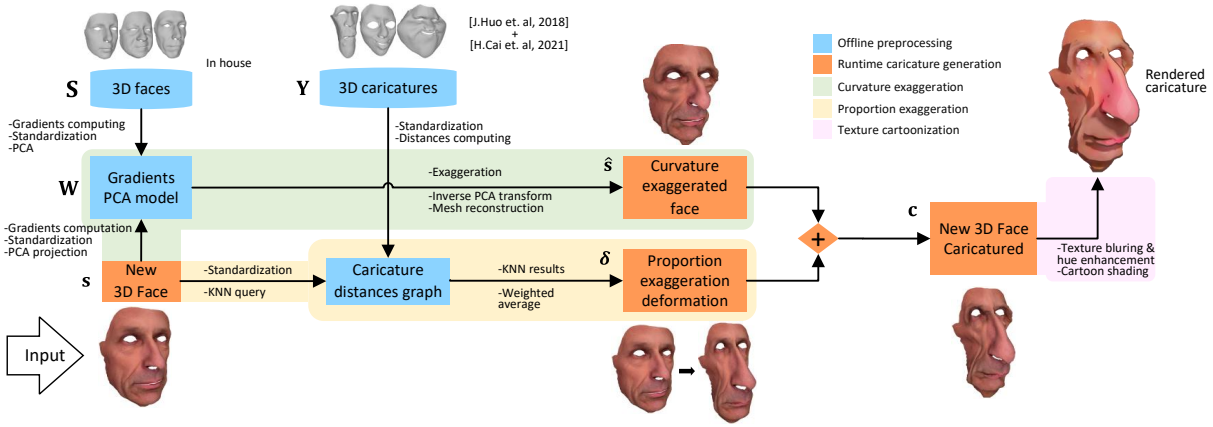


FIGURE 4.1 – Overview of our user-controlled method presented in Section 4.2. Arrows and diamond shapes represent algorithms while boxes represent data. Offline and online processing are represented by the blue and orange colors, respectively. Green, yellow and pink highlights show the different modules which compose the core of the user-controlled caricature system. For simplification purposes, the face segmentation is not shown.

scale versus large scale) over the resulting caricature. Lastly, an optional texture blurring and contrast enhancement module (in pink) makes the resulting caricature less realistic and more graphic. The reason behind this step is to make the result more acceptable for human observers. As showed by Zell *et al.* [128], we use texture blurring because it increases the appeal and lowers the eeriness of a virtual character. The increase in contrast is meant to make the caricatures less realistic, but one could have used another technique to this end. In addition to these modules, our user-controlled method features semantic mesh segmentation in four regions (see Section 4.2.2).

4.2.1 Datasets

Realistic 3D faces were sampled from the LSFM dataset [129] which contains nearly 10k distinct 3D faces. In order to have textured meshes, we completed this set with 300 in-house 3D face scans. Their topologies are unified through automatic facial landmarking and geometry fitting [44]. To build our 3D caricatured mesh dataset, we run the 2D to 3D caricature inference method of Zhang *et al.* [118] on the WebCaricature dataset [18], which enables to extract the 3D caricatured face mesh from each 2D image. The WebCaricature dataset contains over 6k 2D caricatures. All faces were then registered, in order to have a fixed topology [130].

4.2.2 Facial Segmentation

In face modeling, cartoonization and caricaturing, semantic segmentation is a popular technique for increasing expressivity and user interaction [15], [79], [80]. In the proposed system, the 3D faces are segmented using the scheme proposed by Blanz *et al.* [15] *i.e.* in four regions : the eyes, the nose, the mouth and the rest of the face. This semantic segmentation allows the user to choose whether to emphasize or not a facial part. In total, the method expose ten parameters to the user : one scalar is used for the strength of the gradient EDFM and another one for the amount of deformation from the kNN regressor to be added. Those two weights are tunable for each of the five regions (four masks and full face). Segmenting the domain also allows to break the inherent linearity of PCA by learning different subspaces.

4.2.3 Curvature exaggeration

To emphasize the small scale features of the input 3D face, the curvature exaggeration module performs EDFM on the mesh gradient. In the process, we use PCA as a mean to reduce high frequencies (Figure 4.2).

— **Offline preprocessing.** The edge-based gradient operator $\mathbf{E}$ (see Appendix 1) is used to compute the gradients $\mathbf{g}$ of each face mesh $\mathbf{s}$ of our custom 3D face dataset (Section 4.2.1). Following the results of [63] showing that low-variance features should be more taken into account, the gradients $\mathbf{G}$ are standardized : $\mathbf{G}^{\text{std}} = \frac{\mathbf{G}-\overline{\mathbf{g}}}{\sigma_{\mathbf{G}}}$. Then, a PCA is performed on the standardized gradients leading to the principal components $\mathbf{W}$ and each PCA scores $\mathbf{t}$ such that $\mathbf{t} = \mathbf{g}^{\text{std}} \cdot \mathbf{W}$.

— **Runtime curvature exaggeration.** The input face mesh $\mathbf{s}$ is standardized then projected in to the PCA space learnt offline. EDFM technique is applied with a factor f_{grad} given by the user. To prevent from noise, we weight the result by the normalized standard deviation associated to each principal components $\boldsymbol{\sigma} = \sqrt{\frac{\boldsymbol{\lambda}_{\mathbf{C}}}{\max(\boldsymbol{\lambda}_{\mathbf{C}})}}$. The exaggerated PCA scores are obtained as $\hat{\mathbf{t}} = \mathbf{t} \cdot \max\left(f_{\text{grad}} \cdot \boldsymbol{\sigma}, 1\right)$. The exaggerated gradient is then recovered as $\hat{\mathbf{g}} = \overline{\mathbf{g}} + \boldsymbol{\sigma}_{\mathbf{G}} \cdot (\hat{\mathbf{t}} \cdot \mathbf{W}^{\mathsf{T}})$. The gradients exaggerated mesh $\hat{\mathbf{s}}$ is eventually reconstructed at the least squares sense by setting the border vertices fixed (the border of the eyes, the nostrils, the inner lips and the contour of the head), as described in Appendix 1.

4.2.4 Proportion exaggeration

The proportion exaggeration module leverages the 3D caricatures (see Section 4.2.1) to sample a deformation that matches the input face difference from the mean using a kNN. Thus, it can be seen as an example-based version of EDFM. We argue that the sampled deformation contains mainly low frequencies and adding it to an input face will modify very little its surface curvatures. We observed that the 3D caricatures have more diverse global shapes than our 3D faces while being much smoother. In addition, the kNN regression also contributes to smooth out the deformation by averaging the k nearest neighbors. The process works as follows :

— **Offline preprocessing.** The 3D caricatures are first standardized using the standard deviation of our 3D faces to make the low-variance areas more important [63]. Then, we fit a kNN regressor using a cosine distance metric, as we mainly seek to find directions of deformation rather than amplitudes of deformation. The amplitude tuning is reserved for the user.

— **Runtime proportion exaggeration.** The input face is standardized then projected into the 3D caricature space with the kNN regressor using barycentric weights. The obtained deformation $\boldsymbol{\delta}^{\text{std}}$ is weighted by the 3D face standard deviation $\boldsymbol{\sigma}_{\mathbf{S}}$ and by a user-defined scalar f_{prop} for amplitude tuning. Eventually, we add this deformation to the curvature exaggerated face to get the vertex positions of the resulting generated caricature $\mathbf{c}$:

$$\mathbf{c} = \hat{\mathbf{s}} + \boldsymbol{\delta} \qquad \text{with} \qquad \boldsymbol{\delta} = f_{\text{prop}} \cdot \boldsymbol{\delta}^{\text{std}} \cdot \boldsymbol{\sigma}_{\mathbf{S}} \tag{4.1}$$

4.2.5 Results

In this section, the results of both the curvature exaggeration module and the proportion exaggeration module are presented and compared to those of their most similar existing approaches. We compare the curvature exaggeration module to Sela's method [69] because they fix the positions of border vertices and therefore tend to preserve the proportions of the caricatured faces. Our proportion exaggeration module is compared to the baseline 3D position EDFM introduced in the seminal work of Blanz *et al.* [15].

— **Curvature exaggeration module.** The benefit of the PCA-based denoising mechanism is visible in Figure 4.2 between column b), and column c) and d). Without PCA, the EDFM technique magnifies the existing high-frequencies of the face's

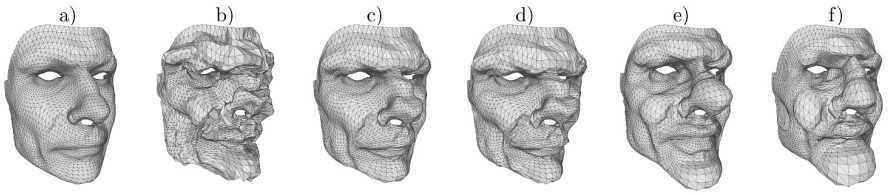


FIGURE 4.2 – Different curvature exaggeration techniques : a) Orignal 3D Mesh. b) Naive gradient EDFM without segmentation ($f_{grad} = 5$). c) Gradient EDFM with PCA denoising, without segmentation and d) with segmentation ($f_{grad} = 5$). e) [69]'s method, without reference model ($\gamma = 0.3$) and f) with the mean face as reference model ($\beta = 4$).

difference from the mean. With PCA, the noise is removed but the exaggeration of facial lines remains. The use of a segmented model not only enables to provide more user-control, but also to emphasize the curvatures more locally. This effect can be noticed when comparing the results c) and d) in Figure 4.2. Sela's [69] method successfully preserves the position of the eyes, the nostrils, the inner lips and the contour of the face. However other parts such as the nose, the lips and the chin seem greatly inflated and displaced which should not belong to facial lines enhancement. Conversely, our curvature exaggeration module modifies the vertex positions such that it only enhances the fine curvature details.

— **Proportion exaggeration module.** Figure 4.3 shows the effect of modifying k on the results of our proportion exaggeration module. Visually, the parameter k of the kNN regressor has less impact than we expected. However, it appears that a small value of k (≤ 5) tends to introduce high-frequencies and vertex entanglement while larges values of k (≥ 1000) seem to produce less vivid results. We fixed $k = 40$ in our experiments.

The semantic segmentation has also an impact on our proportion exaggeration module. In Figure 4.4, the results with segmentation (column c) seem more caricatural but also more expressive than without segmentation (column b). Expressiveness is not intended by the proposed method since the focus is on neutral expression caricature generation. Nevertheless, we decided to conserve the segmentation scheme for the proportion exaggeration module. We also compare the proportion exaggeration algorithm to the baseline PCA-based EDFM on 3D coordinates proposed by [15] (column d). Our method clearly generates more diverse and inhomogeneous shapes than Blanz's [15] approach. It is also noticeable that less high-frequency details are added than with the baseline method, which is what we aim at.

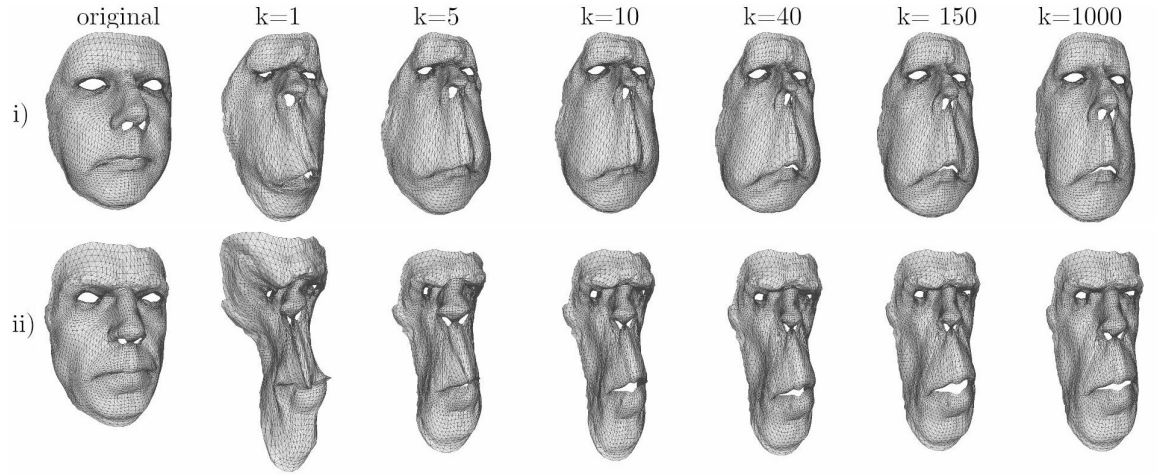


FIGURE 4.3 – A comparison of results with different values of k for the kNN algorithm of the proportion exaggeration module. The first column shows the original facial mesh. Here, the caricatures are generated with $f_{proportions} = 2$.

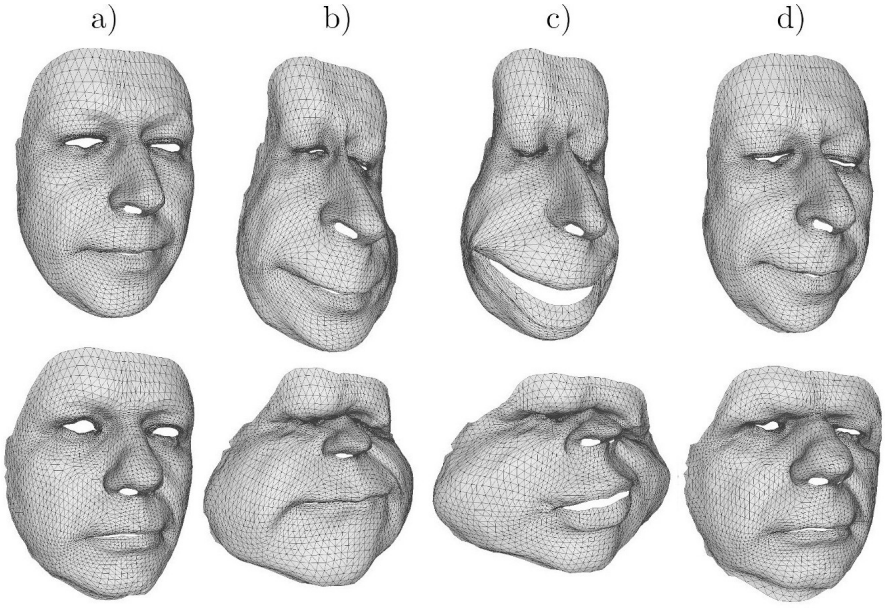


FIGURE 4.4 – A comparison between proportion exaggeration techniques on two facial meshes. a) Original facial mesh. b) Our proportion exaggeration algorithm without segmentation and c) with segmentation. d) Baseline PCA-based 3D positions EDFM [15].

Rule-based methods allow the use of controllable and interpretable parameters, but are limited to capture information about caricature styles. We now introduce a second approach, aiming to solve this issue through the use of deep learning.

4.3 Deep learning based Automatic Caricaturization

Supervised learning based methods require a large paired mesh-to-caricature dataset, that are highly consuming in term of both time and means to build. Instead, we consider the case of an unpaired learning-based approach, taking advantage of our 3D datasets of both neutral and caricatured faces [118] (cf. Section 4.2.1). Our network architecture is based on the shared content space assumption of Liu *et al.* [13], that we adapt to the context of 3D data through the use of 3D convolutions of Bouristas *et al.* [94], which define 3D convolution neighborhoods.

4.3.1 Framework overview

Let us consider meshes of different styles (e.g. scans and caricatures), all sharing the same mesh topology. We represent our faces with raw 3D coordinates, and encode them using a recent 3D convolutional operator [94]. Given a mesh $x \in X$ and an arbitrary style $y \in Y$, our goal is to train a single generator G that can generate diverse meshes of each style y that corresponds to the mesh x. We generate style-specific vectors in the learned space of each style and train G to reflect these vectors. Figure 4.5 illustrates an overview of our framework, which consists of three modules described below.

Generator. Our generator G translates an input mesh x into an output mesh $G(x, s)$ reflecting a style-specific style code s, which is provided by the style encoder E. We use adaptive instance normalization (AdaIN) [89] to inject s into G. We observe that s can represent any style, which removes the necessity of providing y to G and allows G to synthesize meshes of all domains.

Style encoder. Given a mesh x, our encoder E extracts the style codes $s = E(x)$. Similar to [13], our style encoder benefits from the multi-task learning setup. E can produce diverse style codes using different reference meshs. This allows G to synthesize an output mesh reflecting the style code s of a reference mesh x.

Discriminator. Our discriminator D is a multitask discriminator [13], [88], [131], which

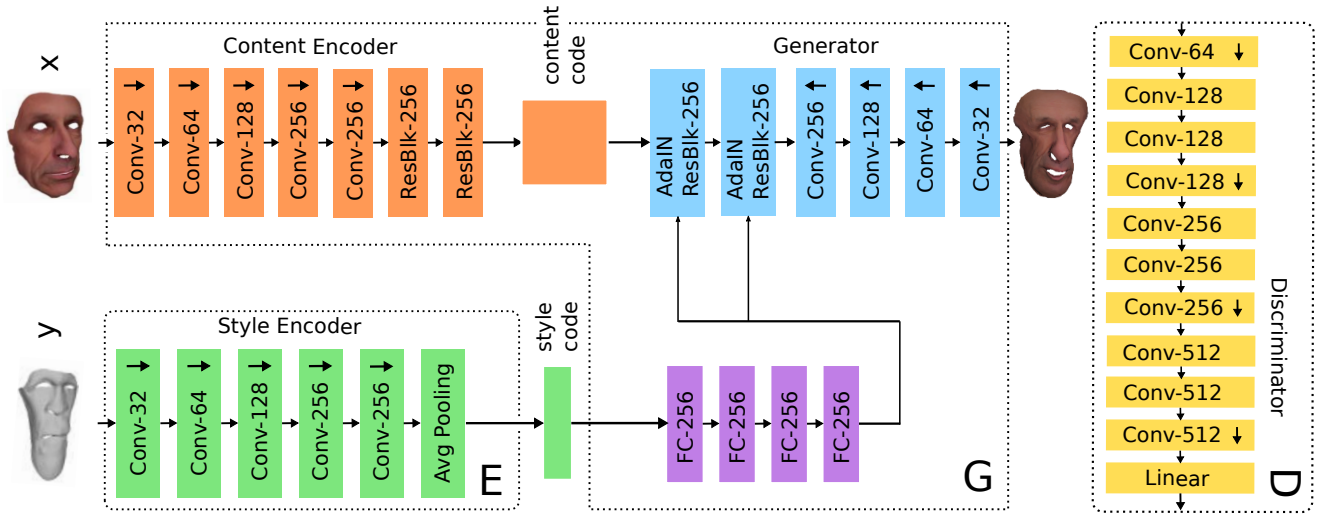


FIGURE 4.5 – Overview of the network. A facial scan's identity is encoded along with the style of a caricature mesh, in order to produce the caricatured face. Textures are not processed, and presented for illustration purpose only. E represent the Style Encoder, G the Generator, and D the Discriminator

consists of multiple output branches. Each branch D_y learns a binary classification determining whether a mesh x is a mesh from the dataset of style y or a fake mesh $G(x, s)$ produced by G.

4.3.2 Training objectives

Given a mesh $x \in X$ and its original style $y \in Y$, we train our framework using the following objectives :

— **Adversarial objective.** During training, we sample a mesh a and generate its style code $s = E(a)$. The generator G takes a mesh x and s as inputs and learns to generate an output mesh $G(x, s)$ that is indistinguishable from real meshes of the style y, via a classical adversarial loss [132] :

$$L_{adv} = E_{x,y}\left[logD_y(x)\right] + E_{x,\tilde{y}}\left[log(1 - D_{\tilde{y}}(G(x, s))\right]$$

where $D_y(\cdot)$ denotes the output of D corresponding to the style y.

— **Reconstruction and cycle losses.** To guarantee that the generated mesh $G(x, s)$ properly preserves the style-invariant characteristics (e.g. identity) of its input mesh

x, we employ the cycle consistency loss [12], [133], [134]

$$L_{cyc} = E_{x,y,\tilde{y}} \left[\|x - G(G(x, \tilde{s}), \hat{s})\|^1 \right]$$

where $\hat{s} = E_y(x)$ is the estimated style code of the input mesh x, $\tilde{y}$ and $\tilde{s}$ are the style and estimated style codes of another mesh than x. By encouraging the generator G to reconstruct the input mesh x with its estimated style code $\hat{s}$, G learns to preserve the original characteristics of x while changing its style faithfully. In a similar goal of preserving style invariant characteristics, we use a reconstruction loss

$$L_r = E_{x,y} \left[\|x - G(x, \hat{s})\|^1 \right]$$

where $\hat{s} = E_y(x)$ is the estimated style code of the input mesh x.

— **Full objective.** Our objective function can be summarized as follows :

$$\min_{G,F,E} \max_{D}$$

$$L_{adv} + \lambda_{cyc} \cdot L_{cyc} + \lambda_r \cdot L_r$$

where λ_r and λ_{cyc} are hyper parameters for each term. We use the Adam Optimizer [135].

4.3.3 Results

We trained the network on 10k human facial scans [136], and 6k 3D caricatures [118] for 50k iterations on a Titan X Pascal (4h, 8Go). Results of the approach are visible in Figure 4.6. The original faces (top row) are encoded using the network illustrated in Figure 4.5 along with a random caricature of the dataset, producing the caricatured face (bottom row). Facial proportions are hence exaggerated according to the distribution of the neutral and caricatured faces learned during the training stage.

4.4 User study

In order to assess the subjective quality of the caricatures generated by the previously described methods, we have conducted a perceptual study. The goal of the perceptual study was to subjectively rank the generated caricatures based on the perceived quality of

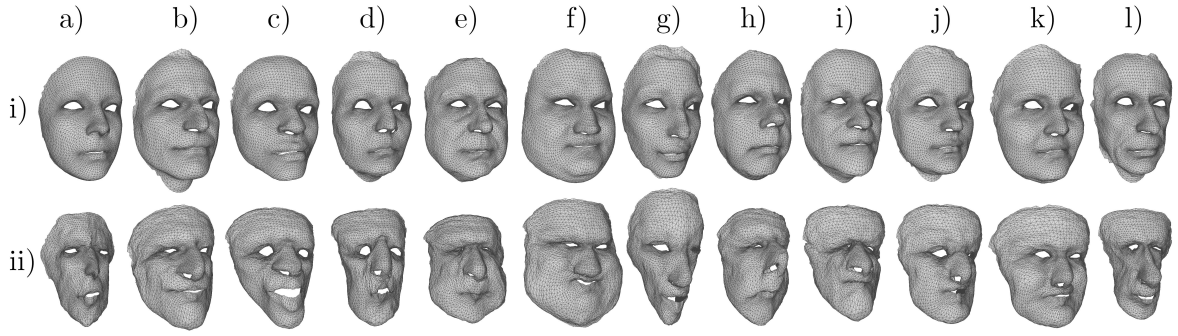


FIGURE 4.6 – Deep learning based caricatures for a number of facial scan examples.

the caricatures. In addition to the two methods described in Sections 4.2 and 4.3, we also considered two baseline methods, the method from [69] and a EDFM method [15].

4.4.1 Participants

Fourty-nine participants took part in the experiment (9 females). They were between 18 and 63 years old (mean and STD age : 31.0±11.3), and were recruited from our laboratory among students and staff. They were all naive to the purpose of the experiment, had normal or correct-to-normal vision, and gave written and informed consent. The study conformed to the declaration of Helsinki. Participants were not compensated for their participation and none of the participants knew the human faces used in the study.

4.4.2 Stimuli

Figure **??** and the top part of Figure 4.6 presents the 12 human face scans (Identity factor) used in the study (4 females, 8 males). They were caricatured using 5 different approaches (Method factor) : the learning-based approach (Deep) presented in Section 4.3, two variations of the rule-based approach presented in Section 4.2 (see Table 4.1), and two state-of-the-art caricaturization methods – EDFM [15] and Sela [69] **??**. For each face (original and caricatured), we used the cartoonization module presented in Section 4.2. The texture blurring is expected to reduce the mismatch of realism between the shape and the texture and therefore make the caricature more acceptable to human observers [128]. Stimuli was rendered with a rotation of 30° around the vertical axis. We considered only the facial mask, hence other facial attributes such as eyes and hair were not displayed.

TABLE 4.1 – Parameters sets of the two variations of our rule-based method used in the user study (Section 4.4). The first variation targets more the proportions while the second strongly exaggerates the curvatures. These parameter sets aim at exploring the range of user control provided to the user. A number of other variations could have been proposed, but we meet complexity restrictions for the user study.

	Exaggeration type	**Eyes**	**Nose**	**Mouth**	**Rest**	**Full face**
Rule-based 1	curvatures	0.5	0	0	0	4
	proportions	1	0.5	0.75	0	0.75
Rule-based 2	curvatures	0	0	3	2	8
	proportions	0	1.75	0	0	0

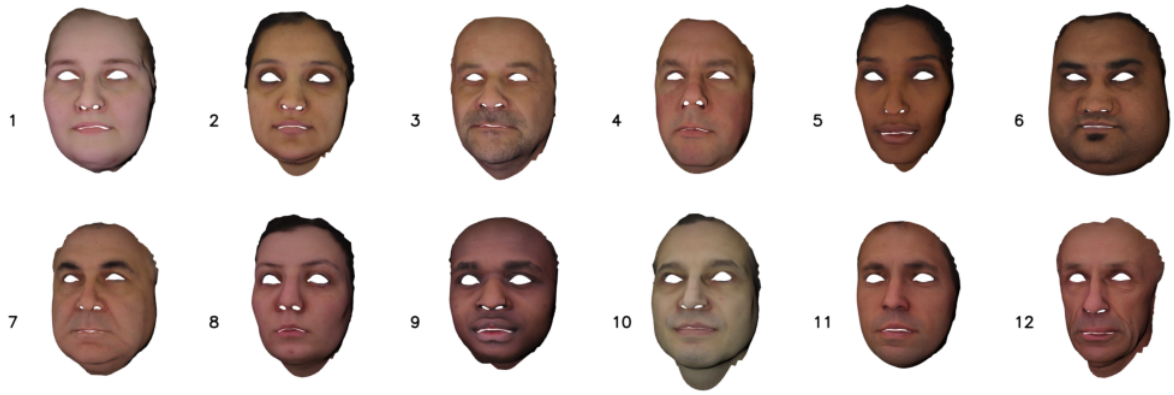


FIGURE 4.7 – The 12 scans used for the study

4.4.3 Protocol

The perceptual study consisted in two parts. The first part of the study assessed the results produced by each method for each face, according to participant's preferences. For each human facial scan, participants were presented with the original face and the caricatures generated with the five methods. They were asked to rank all five caricatures from the best to the worst caricature. The order of the scans and the presentation of the caricatures was randomized for each participant and each facial scan was only presented once, for a total of 12 trials. The second part of the study aimed at evaluating globally each of the 5 methods. For each method, the caricaturization results (12 facial scans) were displayed at once. Participants were asked to indicate how much they agreed to three statements using

FIGURE 4.8 – The caricatures of 6 of the study scans, using the 5 methods

5-point Likert scales. The statements were “They preserve the identity of the person”, “They correspond to what would be expected of a caricature”, “I like the results”. There was no time limit for any of the two parts, and the evaluation was conducted online using the PsyToolkit software [137], [138]. We include a sample view of the ranking task in the appendix of this thesis, in Figure B.6. A render of all 12 caricatures for each method can also be seen there, on Figures B.1, B.2, B.3, and B.5.

FIGURE 4.9 – The five best caricatures (with the best mean ranks ; identities 7, 6, 9, 12, 2)

FIGURE 4.10 – The five worst caricatures (with the worst mean ranks ; identities 6, 7, 5, 11, 12)

4.4.4 Results

Average rankings

To analyse ranking distributions (Figure 4.13), we first performed a Friedman test with the within-subject factor Method (using the average rank between all 12 scans). We found an effect of the Method on average ranking ($\chi^2 = 12.21; p < 0.05$). The effect is then explored further using a Wilcoxon post-hoc test for pair-wise comparisons. We found significant differences only between EDFM and Deep,Geo.1,Sela (all $p < 0.05$). We found that per method, average rankings vary between 2.81 (EDFM) and 3.12 (Deep). In order to determine whether ranking distributions per method differed with identities, we used a Friedman test with within-subject factors Method and Identity. Out of 12 distinct identities, 6 (identities 2, 5, 6, 7, 11, 12) showed significantly different rankings between methods. This is in most cases (5 out of 6) due to worse than average performance from a set of methods, usually Deep or Sela.

Top rankings

We measured Top-1, Top-2, and Top-3 rank differences per method, using Friedman tests, Top-X rankings being the amount of times the techniques was ranked X or lower (lower is better, Figure 4.12). We found no significant differences for Top-1 ($\chi^2 = 4.14; p = 0.38$)

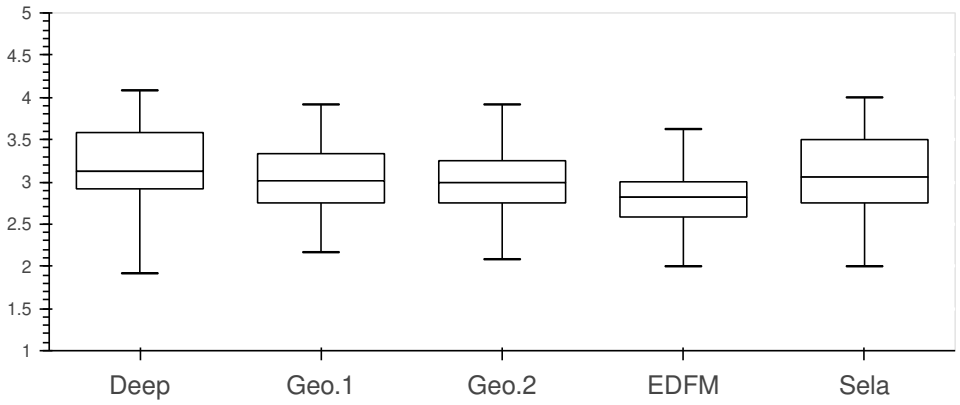


FIGURE 4.11 – Boxplot of the average rankings over participants, per method. Rankings range from 1 to 5.

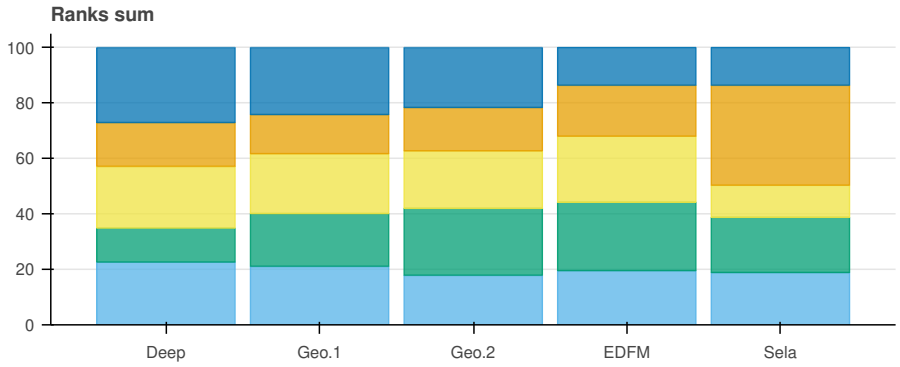


FIGURE 4.12 – Caricature ranking distribution across all participants, per method. Top-1 to Top-5 rankings respectively shown in light blue, green, yellow, orange, and blue

rankings, but an effect was found for both Top-2 ($\chi^2 = 9.74; p < 0.05$) and Top-3 rankings ($\chi^2 = 34.60; p < 0.001$). The effect for Top-2 and Top-3 rankings is then explored using a Wilcoxon post-hoc test. For Top-2 rankings, we found that EDFM was chosen significantly more often as 1st or 2nd choice than Deep ($p < 0.05$) and Sela ($p < 0.01$). For Top-3 rankings, we found a similar preference for EDFM over Deep, Geo.1, and Sela ($p < 0.05$), as well as a significant lower preference for Sela over all others ($p < 0.05$).

Variations between participants

We looked into participant-wise preferences for caricature methods using a Friedman test on ranking choices of each participant, individually. Out of 49 participants, separate Friedman tests on their Top-1 rankings showed that only 12 had a significant preference towards a set of methods, and out of these only 4 towards a specific one. These numbers

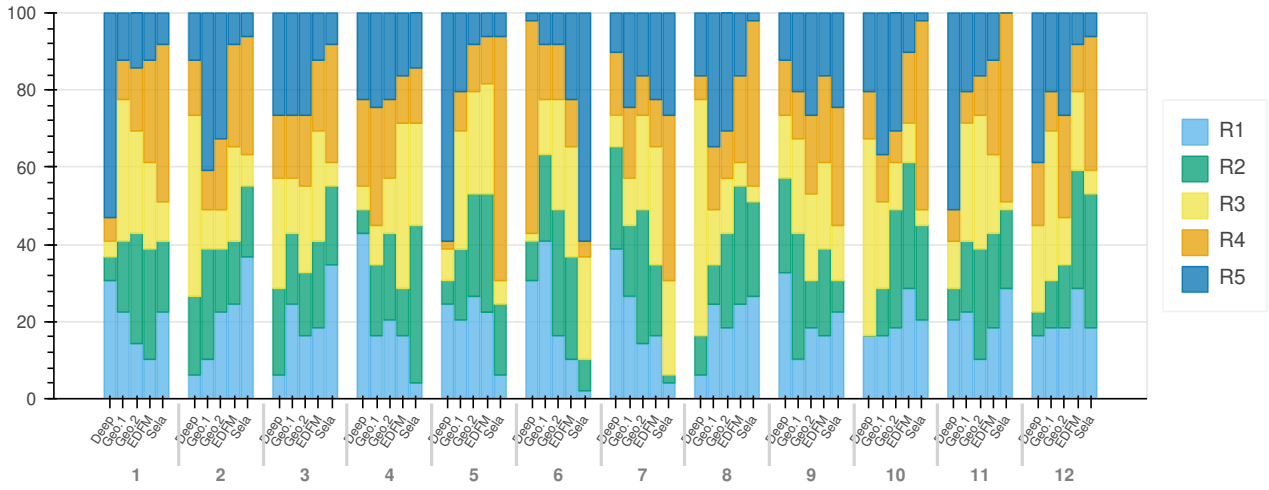


FIGURE 4.13 – Average rankings, per Method and Identity. R1 to R5 are the ranks 1 to rank 5. Note the high variance per face and method

are too low to show anything conclusive in that regard.

Subjective scores

Subjective ratings results were analyzed separately using a one-way ANOVA with within-subject factor Method on the data of each question. All subjective results differences between methods were found to be significant (p values of $5.7e-6$, $7.35e-6$, and $2.28e-5$). We conducted sepearate post-hoc analyses using Wilcoxon. For the statement "They preserve the identity of the person" (Figure 4.14), significantly different groups of method were Deep, Sela (mean=3), and Geo.1, EDFM (mean=2.3). The method Geo.2 (mean=2.6) was not significantly different from others. For the statements "They correspond to what would be expected of a caricature" (Figure 4.15) and "I like the results" (Figure 4.16), the only significant differences were between the group of Geo.1, Geo.2, EDFM, and Sela, Deep being in between.

4.5 Discussion

In this chapter, we have proposed two novel caricaturization methods. One leveraging the capabilities of deep style transfer networks for caricaturization (Deep), and the two remaining are variations of a gradient-based EDFM, with and without the use of a data

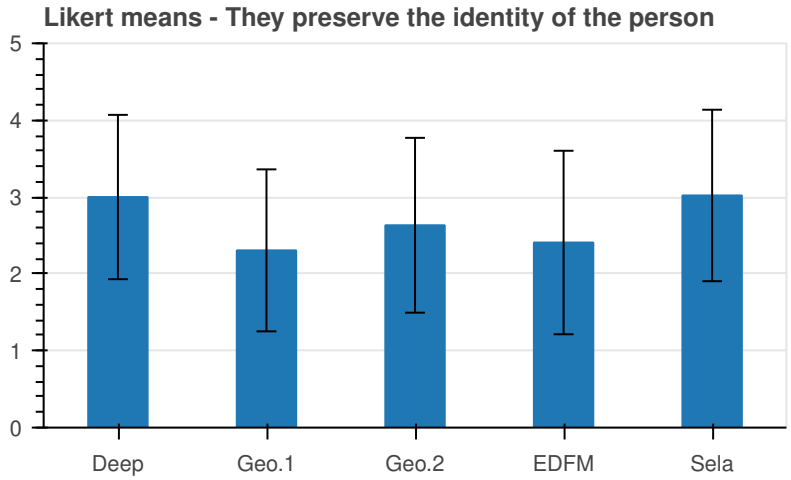


FIGURE 4.14 – Average Likert ratings for the statement "They preserve the identity of the person". Deep, Sela and Geo.1, EDFM are significantly different.

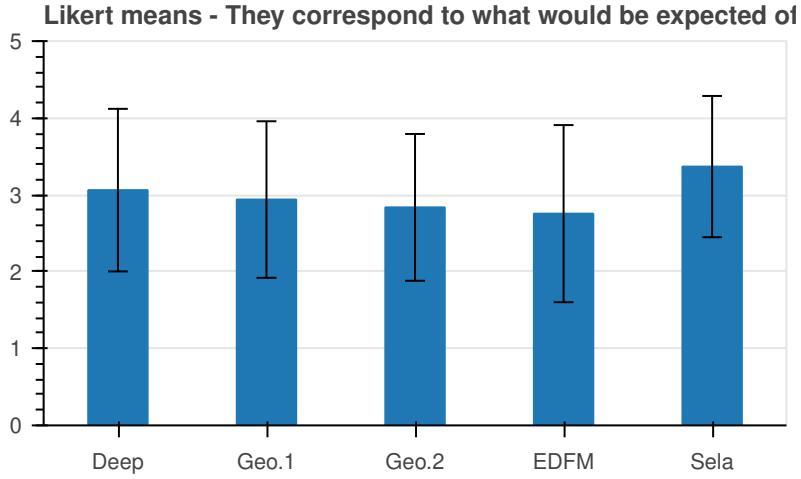


FIGURE 4.15 – Average Likert ratings for the statement "I like the results". Geo.1, Geo.2, EDFM and Sela are significantly different.

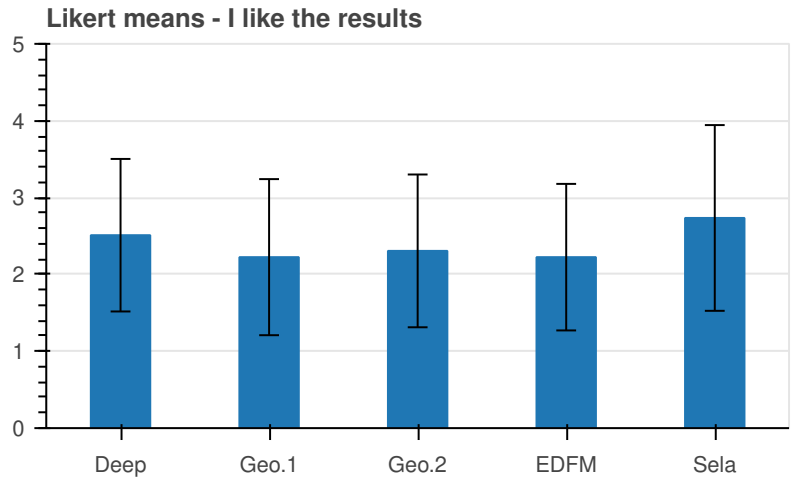


FIGURE 4.16 – Average Likert ratings for the statement "They correspond to what would be expected of a caricature". Geo.1, Geo.2, EDFM and Sela are significantly different.

driven face shape stylization (Geo.1 and Geo.2).

The proposed methods, and two additional methods from the literature were evaluated through a user study considering 12 different facial scans and the corresponding caricature generated from these different methods. Overall, the results showed that all methods achieved similar performances, average ratings going from 2.82 to 3.12 (lower is better). An observation from the results is that in general, there was not a method which was significantly superior to the others. The results considering only the method (see Figure 4.12) show a fairly distributed results, although Deep and Sela approaches seem to generate a higher number of "badly ranked caricatures" (4^{th} and 5^{th} ranks). This observation matches with the global appreciation from participants, as EDFM, Geo.1 and Geo.2 got slightly higher scores. While this result could suggest that some of the methods worked best from some facial scans than others, the results split by Identity do not totally support this hypothesis (see Figure 4.13). Looking at the top 5 worst ranked caricatures (Figure 4.10), we can identify several cases in which the method considered could have generated undesired results. The facial features of face 6 interpenetrate each other when using Sela, and the borders of face 7 are spread too widely using the same method. On face 5, eye size difference is too greatly exaggerated with the method Deep. These generated faces rated significantly worse than others on average can be easily identified, opening possibilities of a manual or automatic filtering protocol. Nevertheless, these results seem to evidence that some methods had a particular bad performance on some of the facial scans.

Yet, this did not happen consistently. Each caricaturization method had a pre-defined set of meta-parameters. The choosen configuration could have suited better some faces than others, generating caricatures of different qualities.

Another potential explanation for the results is that the task was too hard and subjective, choices ending up being random. Using faces with no hair or eyes might have even increased the complexity of the task. Indeed, some participants explicitly stated that the task was difficult, especially as they were judging textured facial masks instead of full faces. Nevertheless, this potential user preference does not seem to be linked with any particular caricaturization method. Looking at participant preferences, only 12 participants out of 49 showed a significant rating variation between methods ranked 1st. Looking at results on subjective questions, the two worse rated (Deep and Sela) methods rank-wise (being also those with the worst rated specific caricatures) were rated significantly higher both at "They correspond to what would be expected of a caricature" and "I like the result", where caricatures of each method were presented globally, suggesting that without their bad results on specific faces – which might be less visible when presented amongst all the others – they could actually have ranked higher than other methods.

Considering these findings, we issue the following guidelines for choosing a method to generate caricatures automatically.

— If the main goal is to generate caricatures with a given set of parameters, no specific style, and as little variance as possible in quality, an EDFM-based method is the most suitable.
— If there is still no specific style required, but more tolerance to variance in quality (for instance if it is possible to tune the generated faces when they are unsatisfying), we recommend the approach of Sela, rated very similarly to EDFM on average in the rankings task, and significantly more on the subjective questionnaire.
— If a specific caricature style is required, the Deep approach will offer results comparable with Sela both in the ranking task and the questionnaire.
— Finally, if there is a need to target a specific user, the best solution is to use the panel of available methods, and leave the choice to them.

Caricatures provide a style whose notion can be understood as an "accentuation of facial features", allowing manually defined rules to achieve comparable performance to learning-based approaches. Other stylistic facial domains, such as aliens or anthropomorphic

animals could have more to gain from learning. Such non-realistic 3D facial data is although currently very scarce.

4.6 Conclusion

In this chapter we have introduced two novel approaches to automatically generate caricatures from 3D facial scans. The first method mixes EDFM-based curvature deformation and data driven proportion deformation, while the second method is based on a domain-to-domain translation deep neural network. We made three main hypothesis : that human facial caricatures produced from generic shape caricatures methods would be competitive (H1), that a deep learning based approach would help overcome some limitations of rule-based methods (H2), and both methods we proposed would reach and overcome the state of the art in automatic caricature generation (H3). Then, we present and discuss a perceptual study aiming to assess the quality of the generated caricatures. The results of the study support hypotheses H1 and H2, as the perceptual study demonstrated no significant preference of the subjects for any of the tested methods, for the proposed human faces. Although this result shows that the two proposed methods reached state of the art performance (H3), the perceptual study did not show a clear winner, highlighting the difficulty to simulate and evaluate such artistic caricatures for which a large variety of styles and solutions exists. Overall, the results showed that the different evaluated methods performed in a similar way, although there performance could vary with respect to the facial scan used. This result illustrates both the subjectivity of evaluating caricaturization performance, along with the complementarity of using different approaches, producing different styles of caricatures. Future work could involve looking into automatic detection of the worse cases of automatic caricaturization, to apply a correction or a filter, or exploring learned-based automatic caricaturization by learning on different caricature styles, and setting up a network able to generate faces of a given style. We believe this study of the extended state of the art have helped grow and precise the landscape of automatic caricaturization approaches, and 3D facial stylization in general, and that our work provides interesting insights and guidelines for the automatic generation of caricatures that will help practitioners and inspire future research.

In future work, we will look at whether the geometrical style transfer network we introduced can inspire the design of a network answering another problematic : morphology aware expression transfer.

Chapitre 5

Morphology aware expression transfer

5.1 Introduction

In this thesis, we have introduced approaches for the automatic stylization of 3D faces. However, such methods are restricted to neutral faces, while virtual characters are typically expressive, being used to communicate with users. Hence there is a need for an approach that would allow to transfer expressions to stylized faces. This would need to be done in a morphology aware manner, as an alien most probably does not smile the same way a human does, the same way that a very thin and young person does not smile the same way that an older and more corpulent person does. In this chapter, we introduce such a method, able to transfer expressions between different human faces. Data of expressive stylized faces being considerably lacking compared to expressive human faces, in this chapter we focus on the later.

Achieving a convincing representation of an expressive 3D face is a difficult task. Countless attempts have resulted in appearances that, while technically impressive, fall just short of fooling the audience and end up in the so-called *uncanny valley*. This is due to our considerable ability to effortlessly infer social cues and gather information when looking at human faces [139]. Variations of the 3D face shape are based on several factors, including identity, pose, expression and age. An approach that is successfully able to decompose faces into these factors of variations would bring a better understanding of the face semantic, and eventually improve many applications such as expression transfer, expression extrapolation, performance retargeting, avatar creation [44] and personalization [140], facial recognition, or aging and de-aging. This work focuses on two factors : the identity and the expression.

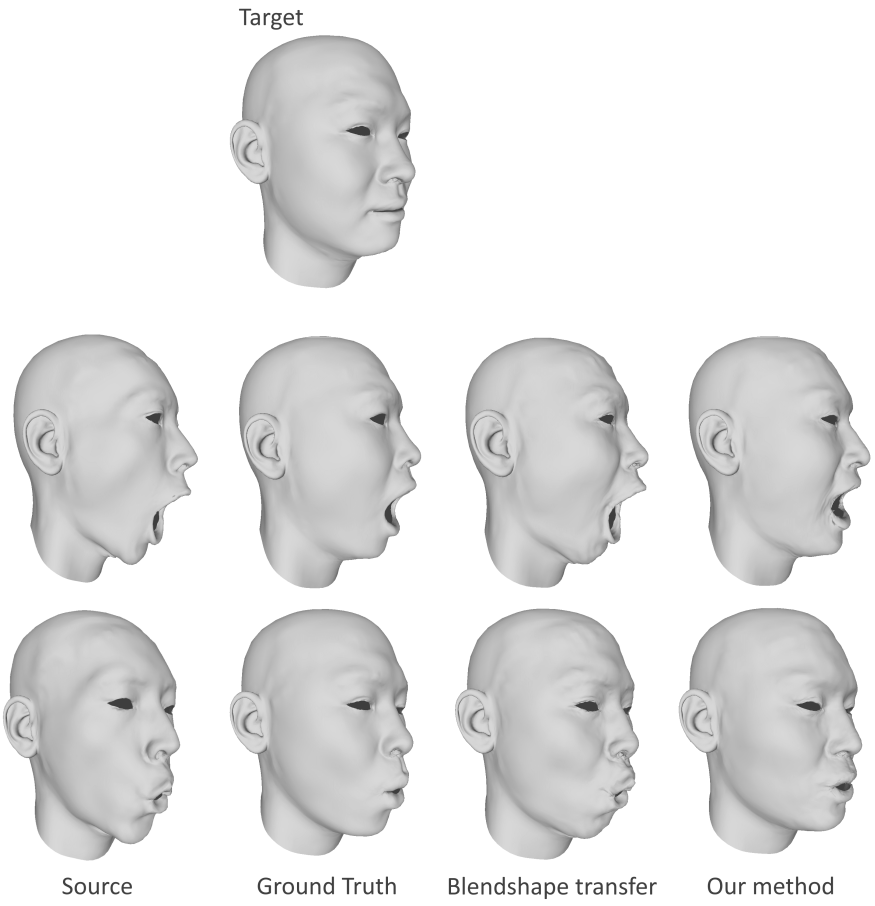


FIGURE 5.1 – Blendshape transfer compared to our method. Results for the expressions *mouth_stretch* (row 2) and *lip_funneler* (row 3) expressions for two subjects of the FaceScape dataset are shown.

Our goal is to add expressions to a given 3D face model while preserving its identity and respecting its morphological constraints.

Current methods for bringing facial expressions to 3D faces often involve a lot of manual artistic work. While some high-budget productions can afford to have expressions manually sculpted by artists, this is a limiting factor for many applications and it is hardly scalable. Tools have been developed to adapt expressions automatically from one face to another, such as Blendshape transfer [130]. However, most of these automatic methods fail to appropriately take into account the specific morphology of the person, resulting in inaccurate depictions.

The limits of expression transfer methods adapting poorly to different facial morphologies than the source causes issues for realistic human faces, but is even more limiting for non-realistic faces, as their proportions typically have more variations. There currently is no public dataset of expressive stylized 3D faces, as there can be for realistic expressive 3D faces, but we believe morphology aware expression transfer methods to be key to leveraging such datasets, when available.

In this work, we apply style transfer techniques to the 3D domain in order to decompose face geometries into structural and stylistic features. Adopting the language of image-to-image translation methods, we designate *content* as the identity dependent features of the face, and *style* as the elements that vary with facial expressions. We apply this approach to decompose identity and expression using only expression labels, allowing morphology aware expression transfer between a face with a given identity, and another with a given expression. We present an architecture that is able to decompose a given 3D face into one latent code for each, and map these latent representations back into a face. We adapt successful image-to-image translations techniques, such as the multitask adversarial discriminator introduced by Liu et al. [13] to the 3D domain. Limiting ourselves to geometry (i.e. no texture), we achieve state-of-the-art results for decomposition and reconstruction of expressive 3D faces with a flexible approach that could be applied to other domains. This allows us to accurately transfer expressions in a morphology aware manner.

The work in this chapter resulted in a submission to Image and Vision Computing A.

5.2 Morphology Aware Expression Transfer

We propose an architecture that adapts successful image-to-image translation techniques to 3D geometry in order to map local features of a 3D face shape (content) from one facial expression to another (style). Specifically, we adapt the base architecture of FUNIT [13], using SpiralNet++ [104] for our convolution and sampling operators. First, we define several terms and notations used in our approach. We then present our architecture and the objectives that are optimized during training.

5.2.1 Notations and Definitions

In our approach, we adopt language analogous to style transfer and image-to-image translations papers. We designate the *content* of the faces as their identity : these are the structural features of the face that do not depend on the facial expressions and should be preserved when performing style transfer. The term *style* corresponds to the features that vary with facial expressions.

Let χ denote a domain of 3D shape, defined as triangular meshes. All shapes within these domains have previously been fitted to the same topology. Thus, any sample $x \in \chi$ is represented by a matrix of 3D coordinates of shape $(V, 3)$, $\sharp V$ being the number of vertices in our topology.

We propose to learn encoding functions E_c and E_s that map a sample $x \in \chi$ to its respective content and style latent codes $E_c(x)$ and $E_s(x)$. We also learn a corresponding decoding function Dec that reconstructs a 3D geometry from the latent codes. These functions should ideally satisfy the reconstruction constraint $Dec(E_c(x), E_s(x)) = x$, meaning that we are able to encode and decode our geometry in a lossless manner. To achieve expression transfer, we use the common method of encoding a source mesh, swapping its style code with one that corresponds to a different expression, then decode to reconstruct a mesh.

Average vertex distance : we use average vertex distance to measure distances between two geometries. Given two triangle meshes x and y with V vertices under the same topology, the average vertex distance is defined as :

$$\text{AVD}(x, y) = \frac{1}{V} \sum_{i=1}^{V} ||x_i - y_i||_2 \tag{5.1}$$

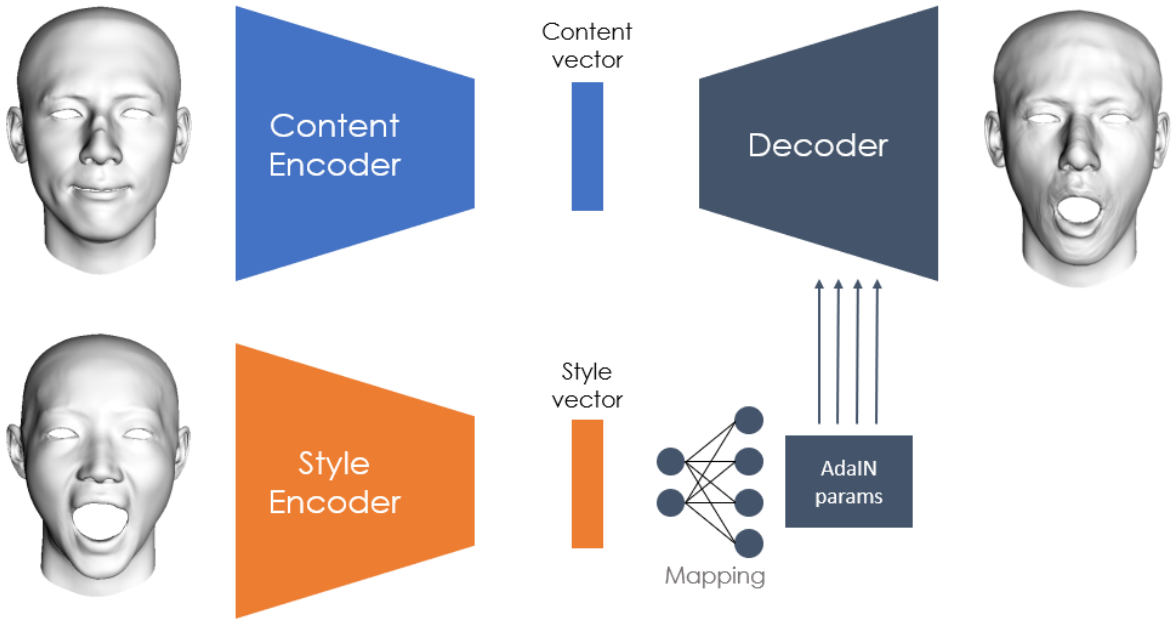
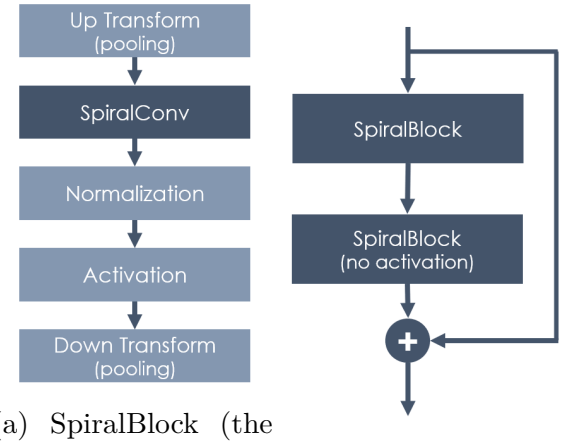


FIGURE 5.2 – Our network architecture. The encoders extract and compress the features of their input into low-dimensional content and style vectors. A decoder reconstructs a mesh from these compact representations. The style information is passed to the decoder through AdaIN normalization layers. More details of the architecture are given in the Appendix (Section C).

where x_i and y_i respectively denote the ith vertex of x and y and $||\cdot||_2$ is the L2 Euclidean distance.

5.2.2 Architecture

We introduce an autoencoder that is able to separate the content (identity) and style (expression) features of the input facial shapes. We take the identity and expressive faces as input in of the encoders, and generate the resulting face with the decoder. The complete architecture is presented in Figure 5.2. The specific composition of the modules is given in the appendix (Section C, Figures C.1 and C.2), with tweaks depending on the dataset used for training. We adopt SpiralNet++ [104] for all of our convolutions and pooling [141] layers and use SpiralBlock as the main building block of our mapping functions. A SpiralBlock (Figure 5.3a) is composed of an optional up-sampling pooling layer, followed by a convolutional layer, normalization, activation and an optional down-sampling layer. Additionally, we use the SpiralResBlock variant described in Figure 5.3b.



(a) SpiralBlock (the layers depicted in light blue are all optional)

(b) SpiralResBlock

FIGURE 5.3 – The building blocks of the network.

SpiralNet-based Autoencoder for 3D Faces

The autoencoder is composed of 3 parts : The content encoder, that takes a face as input, and extract its content (identity), the style encoder, that takes another face and extract its style (expression), and finally the decoder, which generates a face from the outputs of the content and style encoders (see Figure 5.2).

Content Encoder : the content encoder is composed of several downscaling SpiralBlocks (the exact number depends on the dataset, see appendix C), two SpiralResBlocks and a Multi-Layer Perceptron (MLP). The SpiralBlocks capture local information in the geometry. We then flatten the spatial and feature dimensions together and use the MLP to compress the information into the content representation.

Style Encoder : similarly, the style encoder is made up of a series of downscaling SpiralBlocks. However, as style information must be global in nature, instead of flattening the spatial and feature dimensions together, we compute the mean along the spatial dimension. Similarly to the content encoder, the resulting vector is then fed to a MLP in order to control the size of the latent space.

Decoder : the decoder upscales the content code into a 3D geometry using SpiralBlocks that are conditioned by the style code using Adaptive Instance normalization (AdaIN) [142]. Given a sample x that is passing through the network, AdaIN first normalizes the activations in each channel of x to a zero mean and unit variance. The activations are then

scaled on a per-channel basis. We use a mapping function M that maps a style code y into (μ, σ) parameters for every channel of each AdaIN layer. Hence the following equation :

$$\text{AdaIN}(x, y) = M_\sigma(y)\frac{x - \mu(x)}{\sigma(x)} + M_\mu(y) \tag{5.2}$$

M is a learned affine function composed of multiple fully connected layers, taking the style latent code as input. Since the AdaIN transformation operates on whole channels, the style code alters global appearance information while the local features (e.g. the shape of the chin) are determined by the content code. The mapping function and decoder are not conditioned by a discrete class label. As a result, style codes are not domain-specific, in contrast with several existing methods for image translation [143], image generation [88] and shape transfer [110]. Having a single style space allows us to interpolate between style codes.

Together, the content encoder, style encoder, style mapping and decoder make up the generator G of our adversarial network.

Discriminator

Similar to FUNIT [13], we implement a multi-task adversarial discriminator D. Its role is to both enforce that the output mesh belongs to the distribution of the target style class, and that its geometry cannot be distinguished from a real scan. D solves as many binary classification tasks as there are style classes in the dataset. For each of these style classes, D outputs a classification of whether the geometry is a real sample of that class, or a translation output from the generator.

Let s denote a style class.

— When updating the discriminator with a translation output of class s from the generator, we penalize D if and only if its s-th output is positive and ignore the predictions for other classes. Given a real geometry of style class s, we penalize it if its s-th output is negative. This way, the discriminator learns to distinguish real from generated meshes.
— When updating the generator, we perform a style translation using a sample of style class s. We then penalize G if the s-th prediction output from D is negative. This encourages the generator to output realistic meshes.

Similarly to our encoders, the discriminator is composed of SpiralBlocks that gradually downscale its input mesh in the vertex dimension while adding features, followed by a linear layer for classification. More details can be found in Appendix C.

5.2.3 Loss Functions

In this section, we define the loss functions that are used for training our network. Let x and s denote two samples respectively taken from our content and style sets. Let x_r denote the reconstructed mesh obtained by encoding and decoding x :

$$x_r = Dec(E_c(x), E_s(x)) \tag{5.3}$$

We also define x_t, the mesh obtained after translating x into the style class of s :

$$x_t = Dec(E_c(x), E_s(s)) \tag{5.4}$$

Reconstruction loss : we define the reconstruction loss for our generator as the following :

$$L_{\text{rec}}(G) = ||x_r - x||_2 \tag{5.5}$$

By minimizing this loss, we force the encoders and decoder to extract relevant information in order to compress to and from our latent spaces with the least possible loss.

Cycle consistency loss : we adopt the cycle consistency loss [144] [145], defined as :

$$L_{\text{cycle}}(G) = ||Dec(E_c(x_t), E_s(x)) - x||_2 \tag{5.6}$$

Its purpose is to ensure that the generator is able to translate x to the style class of s and back to its original style class with minimal content information loss.

Style reconstruction loss : we introduce a novel style reconstruction loss to encourage G to preserve the specific style features of the input style mesh in the style latent space :

$$L_{\text{srec}}(G) = ||E_s(x_t) - E_s(x)||_1 \tag{5.7}$$

Adversarial loss : our adversarial loss is a conditional loss given by :

$$L_{\text{adv}}(G, D) = \text{E}\left[-\log(D_s(s)) \right] + \text{E}\left[\log(1 - D_s(x_t)) \right] \tag{5.8}$$

where $D_s(\cdot)$ denotes the discriminator output for the style class of s and E[$\cdot$] the mean over the current batch.

Feature matching loss : we use a feature matching loss that leverages our multi-task adversarial discriminator to encourage our generator to output meshes that belong to the correct style class :

$$\begin{aligned} L_{\text{feat}}(G) =& \text{E}\left[\, ||D^f(x_r) - D^f(x)||_1 \, \right] \\ &+ \text{E}\left[\, ||D^f(x_t) - D^f(s)||_1 \, \right] \end{aligned} \tag{5.9}$$

where $D^f(\cdot)$ denotes the last feature layer of the discriminator, prior to classification. By minimizing this loss, we enforce that the generator preserves style features when reconstructing the input and includes style features of the target when translating.

Discriminator regularization loss : we regularize the training by adopting the gradient penalty regularization loss L_{reg} introduced by Mescheder et al. [131] and used in FUNIT [13].

Laplacian Smoothing loss : finally, we use a Laplacian Smoothing loss on the translated mesh [146] [147]. Specifically, we use Pytorch3D's implementation of uniform mesh Laplacian Smoothing, which consists in minimizing the distance between every vertex and the centroid of its neighbors **Johnson2020**. Empirically, we have determined that this constraint makes the generator perform better at reconstruction at the cost of some high frequency resolution.

Full objective : our network is trained by solving the following minimax optimization problem :

$$\begin{aligned} \min_D \max_G L_{\text{reg}}(D) + \lambda_{\text{adv}} L_{\text{adv}}(G, D) + \lambda_{\text{feat}} L_{\text{feat}}(G) \\ + \lambda_{\text{srec}} L_{\text{srec}}(G) + L_{\text{rec}}(G) + L_{\text{cycle}}(G) \end{aligned} \tag{5.10}$$

where λ_{adv}, λ_{feat} and λ_{srec}, are hyper-parameters. The discriminator is trained for one iteration every time we train the generator.

5.3 Results

This section reports the evaluation of our proposed method. We first describe the datasets that were used. Then, we give details for the implementation with which the evaluation was conducted. Finally, our results are evaluated in a series of experiments and compared to existing methods.

5.3.1 Datasets

We evaluate the capabilities of our approach on two publicly available datasets :

FaceScape [148] : this dataset includes 16,940 topologically uniformed 3D face models, captured from 847 subjects performing 20 facial expressions. Displacement and texture maps are also available, though they are not used in this work. Out of the 847 subjects, we discard 40 due to issues with some of the scans. We select 10% of the remaining scans as our test set.

CoMA [102] : this dataset contains dynamic sequences of 12 subjects, each performing 12 facial expressions. In total, it comprises 144 sequences which add up to more than 20k face scans. While the small number of subjects does not allow for great generalization of identity features, the facial expressions are more extreme and asymmetrical than those of FaceScape. In our case, we need discrete expression labels for training. First, we select the first frame of each sequence as samples of neutral expression for the subject. Then, on each sequence, we select the frame with the largest average vertex distance to the first (neutral) frame. Since it is not always the best match to represent the expression, we manually verify all sequences and adjust this selection. We also split up the *mouth_up*, *mouth_middle* and *mouth_down* expressions into their left and right variants, adding up to a total of 17 expressions, though some do not exist for all subjects (see Figure C.3). Finally, we sample 10 frames before and after the selected one in each sequence to add diversity and noise. We obtain a total of 3, 720 scans, which we randomly split into train and test sets by a 9 :1 ratio.

For fair comparison, we limit the dimensionality of our latent space to what was showcased in the result section of each work we compare with. On the CoMA dataset, both our content and style spaces have 4 dimensions, adding up to a total dimensionality of 8. On the FaceScape dataset, Kacem et al. [149], use a single latent space of 25 dimensions.

However, this latent space is only used to represent neutral faces. Nonetheless, we limit ourselves to 20 content dimensions and 5 style dimensions for these comparisons.

5.3.2 Implementation Details

We set the weights in Equation (5.10) to $\lambda_{\text{adv}} = 1.0$, $\lambda_{\text{feat}} = 1.0$, $\lambda_{\text{srec}} = 0.4$. All of our spiral convolutions use a sequence length of 9 with no dilation (refer to 2.3.2 or [104] for more information). We train our generator and discriminator using ADAM optimizers [150], with a learning rate of 1e−4 and a weight decay of 5e−5. All weights are initialized using the Kaiming method [151]. We train with a batch size of 8, until no significant improvements are seen on our reconstruction, neutralization and style transfer metrics. The training duration depends on the dataset. On FaceScape, we train for 70 epochs over a duration of approximately 36 hours. On CoMA, the network is trained for 480 epochs over 8 hours. Fine-tuning was done on a NVIDIA GeForce RTX 2080 Ti, a Tesla P100 GPU and a Tesla V100. The final training runs were done on the 2080 Ti for CoMA, Tesla V100 for FaceScape.

5.3.3 Experiments

First, we assess our autoencoder's ability to reconstruct an input mesh and compare it with state-of-the-art methods. Second, we conduct the expression neutralization task and compare our results with several baselines. Third, we evaluate our performance on the more general expression transfer task, and we compare the results to a blendshape transfer method.

Reconstruction

In order to ensure that our autoencoder is able to compress the information with as little information loss as possible, we calculate the reconstruction error as follows :

$$E_{\text{rec}} = \frac{1}{|S|} \sum_{x \in S} \text{AVD}(x, Dec(E_c(x), E_s(x))) \tag{5.11}$$

where S is a testing set of face geometries.

Figure 5.5 shows error maps of reconstruction samples on the CoMA dataset. In Table 5.1, we list quantitative results compared with other methods for both reconstruction

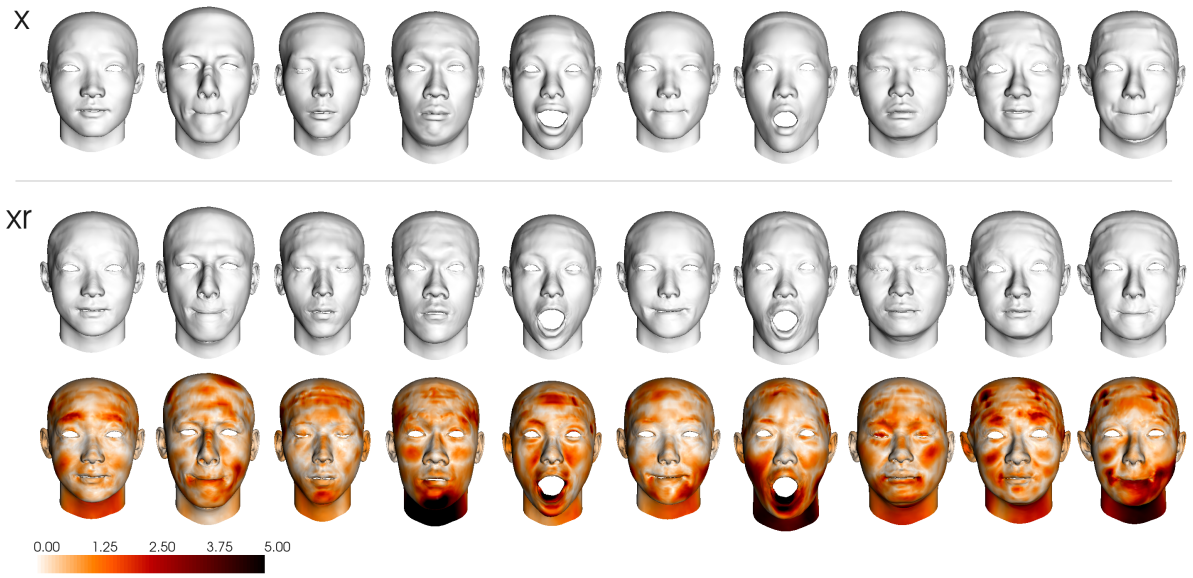


FIGURE 5.4 – Reconstruction examples on FaceScape, with latent dimensions $(20, 5)$. The first row shows the input mesh. The reconstruction is displayed in the second row, while the third row shows an error map relative to the input. Values are in millimeters.

and disentanglement of identity and expression. All methods use a total latent space size of 8. On CoMA, we attain reconstruction results in the range of other disentanglement methods : we perform better than FLAME [72] and Jiang et al. [100] but worse than Zhang et al. [101]. Note that the latter benefits from being able to use the entire CoMA database, while we only select a few frames per sequence to fit discrete expression labels. We also do not require each mesh to be paired with its neutral ground truth for training. For comparison, we include the results of the original SpiralNet++ [104] architecture, which obtains better performance on the reconstruction but does not disentangle identity and expression.

On the FaceScape dataset, we obtain a mean reconstruction error of 0.81 mm, with a standard deviation of 0.25 mm and a median of 0.76 mm. Metrics from other work are not available for comparison. Examples of reconstructions are given in Figure 5.4. We can observe that the error is spread over the face, with notable patches on the neck, eyebrows, lips and cheeks. On the *mouth_stretch* expression (columns 5 and 7), the network visibly struggles to capture the geometry and position of the lower lip.

TABLE 5.1 – Reconstruction error on the CoMA dataset (mm). The SpiralNet++ method does not disentangle identity and expression.

Model	Mean	Median
SpiralNet++ [104]	0.54 ± 0.66	0.32
FLAME [72]	1.45 ± 1.65	0.87
Jiang et al. [100]	1.41 ± 1.64	1.02
Zhang et al. [101]	$\mathbf{0.67 \pm 0.75}$	**0.43**
Our method	0.83 ± 0.21	0.77

FIGURE 5.5 – Reconstruction examples on CoMA. The first row shows the input mesh. The reconstruction is displayed in the second row, while the third row shows an error map relative to the input. Values are in millimeters.

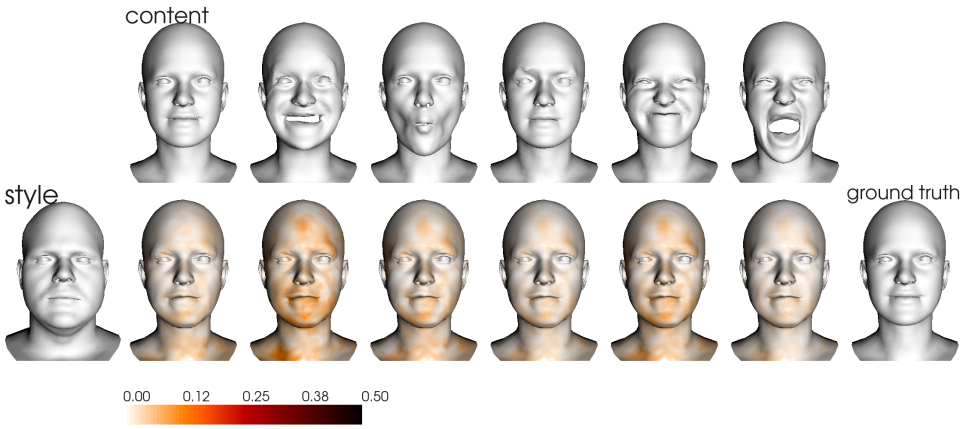


FIGURE 5.6 – Variations of neutralized meshes for one subject of CoMA. The error map is relative to the mean of the outputs. Values are in millimeters. The identity decomposition error is the mean of this standard deviation computed for each subject.

Expression Neutralization

We evaluate our method on the expression neutralization task using two metrics. We first adapt the identity decomposition error introduced by Jiang et al. [100] (and adopted by Zhang et al. [101]) to our approach. Let $x_{c,s}$ denote a mesh of content class c and style class s. We compute the identity decomposition as follows : for each content class c, we randomly select another content class c' and apply its neutral style onto each $(x_{c,s})_{s\in S}$. Then, we calculate the standard deviation of the resulting meshes using AVD : $\sigma_c = \underset{s\in S}{\text{std}}\ Dec(E_c(x_{c,s}), E_s(x_{c',\text{neutral}}))$. Since we are effectively applying the same neutral style on meshes $(x_{c,s})_{s\in S}$, which only differ in style, a successfully trained network is supposed to yield identical outputs. A visualization for a subject of the CoMA dataset is given in Figure 5.6. We report the mean of these deviations as our identity decomposition metric (Equation (5.12)). For quantitative results, see Table 5.2.

$$E_{\text{id\_decomp}} = \frac{1}{|C|}\sum_{c\in C}\sigma_c \tag{5.12}$$

In addition, we compare our results to existing methods on the FaceScape dataset by reporting the mean error between neutralized faces and the corresponding ground truth neutral. More precisely, we apply the following process : we randomly draw n triplets $(c, c', s)_{c\in C,\ c'\in C\setminus\{c\},\ s\in S\setminus\{\text{neutral}\}}$, i.e. two different content classes and a non-neutral style class. For each triplet (c, c', s), we apply the style of $x_{c',\text{neutral}}$ onto $x_{c,s}$ and compare the

TABLE 5.2 – Identity decomposition error on the CoMA dataset (mm).

Model	Mean	Median
FLAME [72]	0.599	0.591
Jiang et al. [100]	0.102	0.096
Zhang et al. [101]	0.019	0.020
Our method	**0.018**	**0.018**

TABLE 5.3 – Neutralization error (mm). Standard deviations and medians for other methods are not provided.

Model	Dataset	mean $\pm$ std (median)
Ranjan et al. [102]	FaceScape	2.88
Kacem et al. [149]	FaceScape	2.02
Our method	**FaceScape**	$\mathbf{1.47 \pm 1.43\ (1.31)}$
Ranjan et al. [102]	CoMA	3.28
Kacem et al. [149]	CoMA	2.73
Our method	**CoMA**	$\mathbf{0.98 \pm 0.46\ (0.81)}$

output with ground truth $x_{c,\text{neutral}}$ (Equation (5.13)). Results are presented in Table 5.3.

$$E_{\text{neu}} = \frac{1}{n} \sum_{(c,c',s)} \text{AVD}(x_{c,\text{neutral}}, Dec(E_c(x_{c,s}), E_s(x_{c',\text{neutral}}))) \tag{5.13}$$

Note that our model is trained in the unpaired setting. We do not explicitly pair expressive faces with their neutral counterpart, contrary to the compared methods. These metrics allow us to evaluate our method against the state of the art for expression neutralization. However, this task is only one particular case of our network's capabilities.

Expression Transfer

We now evaluate our model on the more general expression transfer task. We introduce a metric similar to the neutralization error above and provide a baseline. We randomly draw n triplets exactly as described above : each triplet (c, c', s) contains two different content classes and a non-neutral style class. This time, the style of $x_{c',s}$ is applied onto $x_{c,\text{neutral}} : x_t = Dec(E_c(x_{c,\text{neutral}}), E_s(x_{c',s}))$. The output x_t is compared to the ground truth $x_{c,s}$.

$$E_{\text{transfer}} = \frac{1}{n} \sum_{(c,c',s)} \text{AVD}(x_{c,s}, x_t) \tag{5.14}$$

In Table 5.4, we report our expression transfer error on the FaceScape dataset for each style class s. It can be observed that some expressions are associated with much higher transfer errors than others. The most difficult expressions seem to be the ones farthest apart from neutral. Figure 5.7 shows visual examples for *smile* and *mouth_stretch*, the expressions with the lowest and highest error. It illustrates one of the challenges of the task particularly well. The first subject in the *mouth_stretch* expression opens his jaw less than most subjects in the dataset. In the transfer output, the lower part of the face is lower than it should be, because the corresponding $x_{c',s}$ (not shown) whose style is transferred has a more open jaw. During training, this would penalize the network. However, this difference could perhaps be attributed to a different interpretation of the expression by the subject (i.e. for the same expression, two scans can be different due to factors other than identity).

Finally we compare our method to traditional blendshape transfer [130]. Results are depicted in Figure 5.8. A visual comparison to our method is shown on Figure 5.1. For the two selected expressions, the blendshape transfer method applies the expression geometry

TABLE 5.4 – Expression transfer errors on FaceScape (mm) with our method and the blendshape transfer.

Expression	Our Method	Bs transfer
smile	**1.23**	2.11
anger	**1.32**	3.25
sadness	**1.45**	4.10
grin	**1.57**	2.60
mouth_stretch	**2.02**	3.26
mouth_left	**1.61**	2.78
mouth_right	**1.54**	2.45
dimpler	**1.42**	2.20
jaw_left	**1.53**	2.24
jaw_right	**1.63**	2.59
jaw_forward	**1.62**	2.43
chin_raiser	**1.65**	2.82
lip_puckerer	**1.56**	2.36
lip_funneler	**1.80**	2.64
lip_roll	**1.52**	2.50
cheek_blowing	**1.77**	3.55
eye_closed	**1.23**	1.86
brow_raiser	**1.30**	3.06
brow_lower	**1.41**	3.27
all	**1.54**	2.74

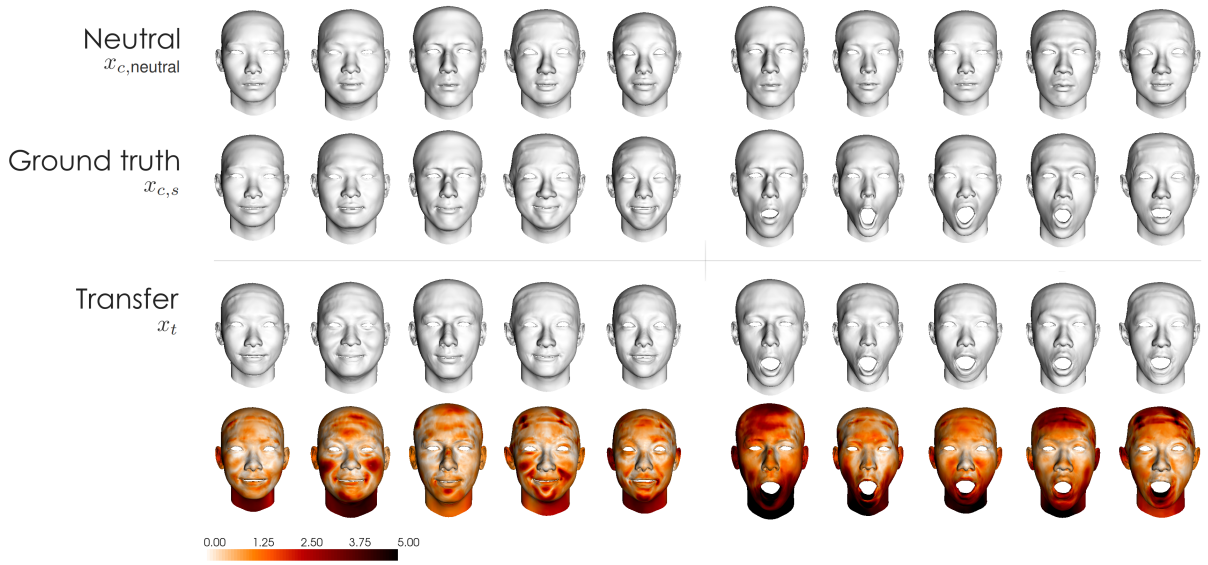


FIGURE 5.7 – Expression transfer results for two expressions (smile on the left, mouth_stretch on the right). The transfer outputs are computed as described in Section 5.3.3.

to the target as it is on the source face, without taking into account the target morphology. Our method leads to more realistic outputs that look closer to the ground truth. Objective measurements shows that our method has smaller errors, as reported in Table 5.4.

5.4 General Discussion and Perspectives

We have developed a network architecture that adapts style transfer techniques to 3D faces and demonstrated results that are competitive with the state of the art reconstruction, neutralization and expression transfer. However, there are several ways in which this work could be extended.

Interpolation : we visually investigated our model's ability to interpolate between two faces. Leveraging our low-dimensional representations, we perform linear interpolations in the content and style spaces. Let A and B denote two style codes obtained by encoding two scans of the same subject in a different facial expression. To interpolate, we move along the style vector $\overrightarrow{AB}$ in fixed increments of $s \times ||\overrightarrow{AB}||$ in the style space. Similarly, we can interpolate in content space. In Appendix C, Figure C.4 shows results of interpolating between expressions of the same subject on the FaceScape dataset. In Figure C.5, we interpolate between different subjects in their neutral expression.

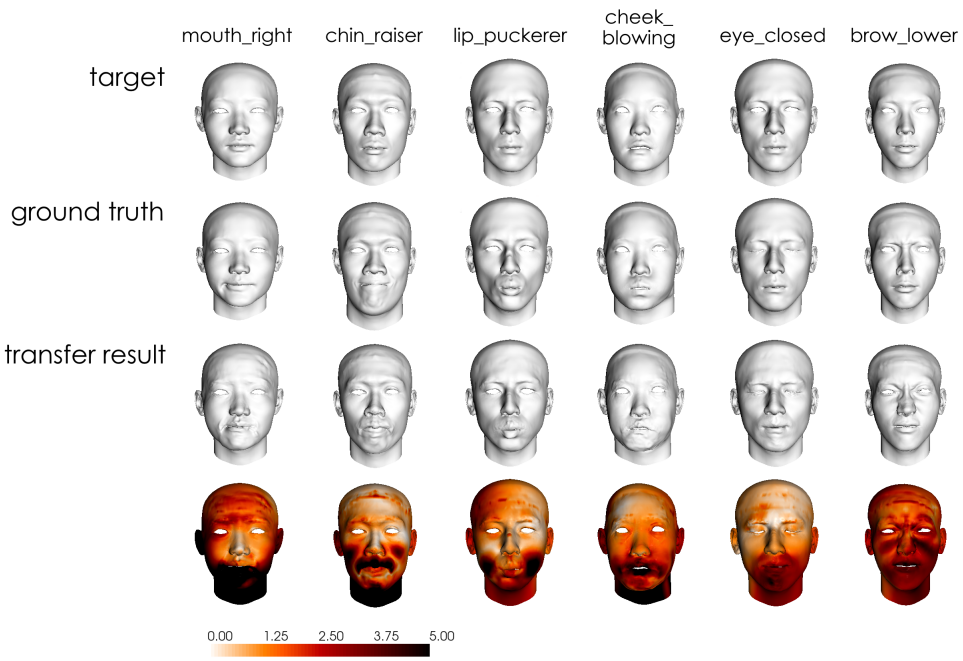


FIGURE 5.8 – Blendshape transfer results. Expressions are tranfered from a source (not shown) to the target

Geometric approaches (e.g. blendshapes) have explicit control over the intensity of the expression being applied. On the contrary, our method in its current state does not have any additional constraint on the latent spaces. Hence the results of these interpolations are completely dependent on the structure of the data on which the network was trained. We find that the network learns a much more regular representation for the content space than the style space. We attribute this to the large amount of subjects in the FaceScape dataset, in comparison to the more limited variation of the 20 expressions.

Extrapolation : the previous interpolation method was extended to investigate the network's ability to extrapolate from the training distribution. In Figure 5.9, we extrapolate in style space along a vector going from the neutral code of a subject to its style code for another expression. We find that while the network can sometimes succeed in amplifying the expression (e.g. *mouth_stretch* and *cheek_blowing* on Figure 5.9), it is also susceptible to changing some features in unexpected ways (e.g. the *dimpler* expression on Figure 5.9).

Figure 5.10 shows extrapolation in the content space. Our method is able to amplify the differences between two faces. Similarly to our findings for interpolation, we can observe

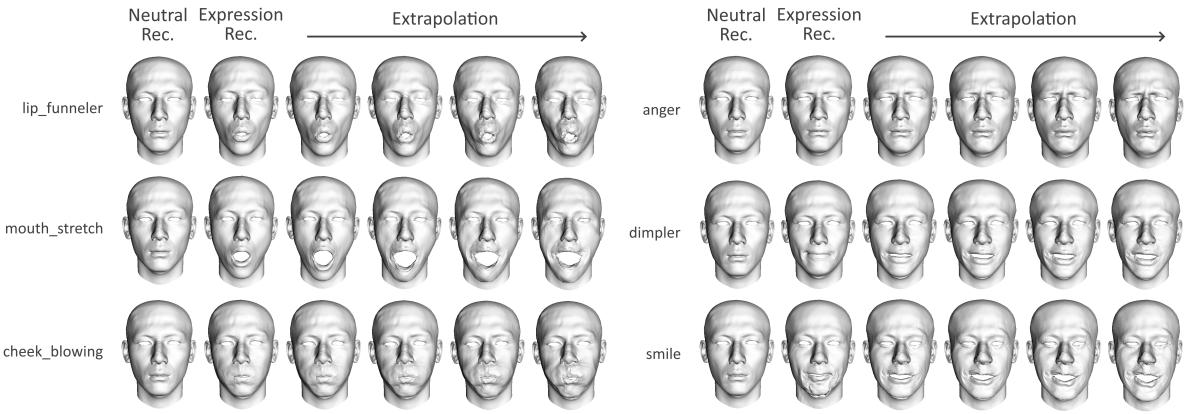


FIGURE 5.9 – Extrapolation in style space. The two leftmost columns are reconstructions of the neutral and expression scans. We gradually extrapolate along the style vector ($s = 0.5$).

that the extrapolations in content space are less prone to adding noise and warping unwanted features than those in the style space. These interpolation and extrapolation experiments are only preliminary tests. Due to the subjective nature of this task, a user study would need to be conducted to validate these results.

Increased realism : while we currently only predict geometry, one could use the data available in FaceScape to learn prediction of displacement maps and textures from the latent codes and generate much more realistic faces. Our network's ability to decouple identity and expression factors could be well-suited to modelling the subtle expression-specific texture variations, a capability demonstrated by Chandran et al. [152].

Investigate bias : it is common for deep learning methods to display biases in the output distribution. While the gender of the subjects in the FaceScape dataset is distributed quite evenly, there is a clear bias toward 18 to 24 year-old Asians. It would be beneficial to study how this bias is reflected in the outputs and whether the network is able to generalize to more diverse faces.

Generative model : currently, sampling random content or style codes is very unlikely to provide satisfying results. In order to add this generative ability to the network, our autoencoder could be turned into a Variational Autoencoder [153] by adding a regularization term (e.g. Kullback-Leibler divergence loss) and encoding distributions instead of single points. This could also improve our ability to interpolate in latent space.

Unseen expressions : we could take a step further toward adapting the performance

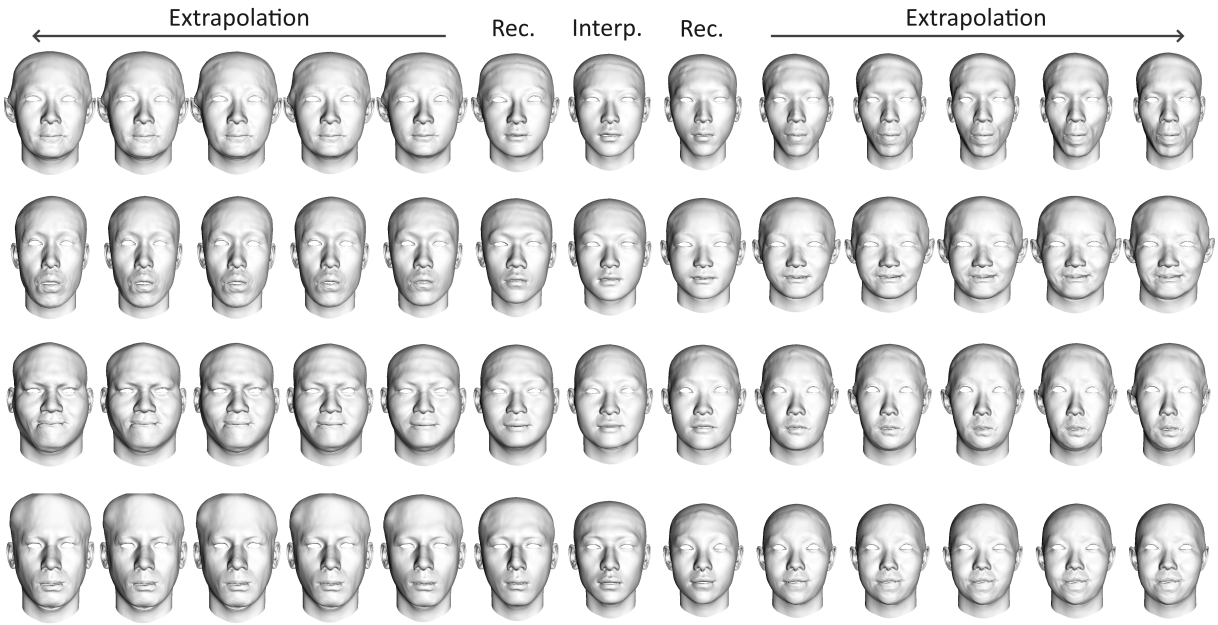


FIGURE 5.10 – Extrapolation in content space. The reconstructions of two neutral scans are shown in the columns marked "Rec.". We move along their relative content vector from the middle to the outer columns ($s = 0.5$).

of image translation networks to the 3D domain by adding few-shot inference for unseen style classes. Given a few samples of a new style class, the model would be able to translate meshes to that new class. This capability is demonstrated in 2D by the FUNIT model [13].

Application to other domains : finally, while our approach was only tested on face datasets, applying it to data from other domains could yield interesting results. The image-to-image translation models from which it is inspired have been able to operate on animals, shapes of shoes and handbags, seasonal variations of landscapes, painting styles, etc. The challenge here will be the availability of large, labeled 3D datasets.

5.5 Conclusion

In this chapter we have presented a novel 3D face model representation decomposing and encoding separately facial identity and facial expression. The proposed method is an adaptation to the 3D domain of existing work for transferring features of images from one style class to another, awllowing morphology aware expression transfer. Adapting

several existing techniques to the 3D domain, this new style-based adversarial autoencoder architecture can capture identity and expression features of the input 3D face in separated low-dimensional spaces. Two encoders respectively extract the content (identity) and style (expression) information, which the decoder takes as inputs to reconstruct a 3D face. Additionally, a discriminator is used in an adversarial training scheme to regularize the training, enforcing the output to be realistic and of the correct style class. This method is shown to be better with previous state-of-the-art approaches on several tasks, outperforming them on several metrics, such as identity decomposition and neutralisation.

This method could bolster the creation of digital characters, allowing to accurately transfer expressions between various facial morphologies, realistic or stylized. The architecture of the method containing nothing specific to expressions, it also has the potential to be adapted to applied to other style transfer applications for 3D models.

Chapitre 6

Conclusion

In this thesis, we explored the topic of virtual character stylization. Our goals were to address the problem of automatic character stylization in the context of both high and low data availability, as well as expression transfer between faces with varied morphologies. We presented a novel method for both character stylization context along with users studies to examine their performance and limitations, and proposed a novel method for facial morphology aware expression transfer, that beat existing methods on several metrics. These three contributions are summarized in the next section. Section 6.2 presents different future research perspectives.

6.1 Contributions

6.1.1 Rule based stylization

In Chapter 2, we explored the link between facial stylization and recognition. This question is important, as in many cases a user may wish to stylize their face, while remaining recognizable. We thus proposed a novel stylization method, in which we decomposed the facial representation into the classical mesh plus texture, and focused on each of these individually. The geometry is deformed with a style and identity decomposition, a reference stylized face being used as the style, and the identity of the person's geometry being applied to it, by computing its euclidian difference to an average human face. For texture stylization, we used an approach of a similar relative nature, but this time modify the style texture through optimization of a generalist network's features, inspired by the work of Gatys [14]. We then leveraged this method to conduct a user study, in order to measure the impact of the level of stylization of a human face on the recognisability of the person, and the acceptability of the output face as "stylized". We laid two main hypothesis, the

first being that the most a face is stylized, the less the person is recognizable, the second being that faces are accepted as "stylized" in inverse relation to it. Both hypothesis were verified through our experimental results, both rates varying in an inverse near linear fashion. Recognition rate (out of four choices) varied from 95% at the highest to 65% at the lowest, while style acceptance rates varied from 10% to 100%. No style level although maximized both metrics, the two style levels maximizing both corresponding to 75%, 92%, and 80%, 65% respectively for identity and style.

6.1.2 Learning based stylization

Rule based methods have empirically proven to perform more weakly in several fields compared to learning based approaches, notably in the domain of the stylization of 2D content. Looking to build a stronger stylization method, we took inspiration from approaches in the 2D domain and developed a GAN based domain translation neural network, aiming to learn to separate style and identity from a large set of unpaired 3D faces. As learning based methods require a significant quantity of data, the work of Chapter 4 focused on caricatures, the only style domain where such quantity of data were available, leveraging a recently published dataset. We then conducted a perceptual experiment to measure the performance of our approach, we compared it against three rule based state of the art methods during a two part user study aiming to rank each approach. From the first part of the study, we found nearly no difference in ranking between the four methods, average rankings being highly similar, and only one method being measured significantly better than others. Looking at rankings per face used in the study, we observed a very large variation of each method's ranking per face, some being ranked considerably higher — or lower — for certain faces. We also observed cases where the faces produced by some method where both ranked most first, and last, considering all participants answers. These observations highlight the high subjectivity of a stylization task, perhaps even more in the case of caricatures, as there can exist numerous valid caricature styles. From these empirical results, we issued several guidelines for choosing a caricaturisation methods, depending on the available data, the tolerance on the variance in subjective quality, and the specificity of the desired style.

6.1.3 Morphology aware expression transfer

In the two previous contributions, we introduced two methods to apply a stylization to fixed neutral faces, but virtual characters are typically interactive, and need to be able to express a range of facial emotions and expressions. Traditional expression transfer methods like blenshape transfer fail in cases where morphologies differ significantly. More advanced approaches for identity and expression separation have various limitations, such as working only in a latent space, and not adapting the expression to the morphology. We introduced a novel approach without these limitations, to pave the way towards expression transfer for stylized faces, building on the approach we used for caricatures. Likewise, our method is based on semi-supervised training (requiring only expression labels) of an autoencoder guided by a GAN. We improve both identity and expression preservation, and compare it with the state of the art in expression transfer, obtaining better reconstruction error, identity decomposition, and neutralization performance.

6.2 Future work and perspectives

In this thesis, we have explored multiple approaches addressing the accurate stylization of virtual characters, as well as a method for morphology-aware expression transfer. However, there are numerous ways in which this work on virtual characters stylization could be extended. Here we present a set of possible short term, medium, and long term research perspectives on the topic of virtual character stylization. In the first, we propose improvements and hypotheses that could readily be tested, directly following this thesis' work, requiring more experiments and user studies, but no more technical research. In the second, we present approaches that could be derived from existing literature, and combined and adapted to the domain of character stylization. In the last, we suggest a method that would require solving some central challenges in the field of generative networks.

6.2.1 Short-term perspectives

Few-shot 3D style transfer networks

In this thesis, we proposed a deep learning based approach for 3D mesh facial style transfer, that we first applied to caricatures. This method although only works in a class wise setting, mapping each human face to one caricature, and is not able to learn different caricature styles in an unsupervised manner. This limitation also implies that the network would not

be able to adapt to unseen style, as its style space is collapsed to only two modes, human and caricature. One of the first short-term perspectives to increase the value of such an approach would be to include properties allowing a one-to-many mapping instead, in order to allow the stylized face to take the specific kind of style of a given caricature. This kind of capabilities have been shown in the image domain [13]. The simplest hypothesis we could make as to why it does not work in our case is that our number of classes (human and caricature) is too low, Liu *et al.* using from 85 to 444 of them in their tests [13]. This must force the network to learn a complex style latent space in a supervised manner, allowing it to adapt more easily to intra-class variations in an unsupervised manner. With only two classes, this is considerably more difficult A solution to it would be to separate the style data into different classes, whether manually or through means of a clustering algorithm, to have more classes for training. Such manual work would although have a considerable time cost and remain highly subjective, and there are no obvious metrics that could be used for clustering. We experimented with the idea of artificially adding style classes by training the network with both neutral human faces, caricatures, and expressive human faces, but it did not bring any improvement. The fact that a given caricature can be encoded and decoded with little loss, but its specific style not applied to a given face, implies that part of its style information flow through the content encoder instead of the style encoder. Considering this, a simple scheme would be to bottleneck the content latent space, as we did in our work on expressions. This would makes it harder for style information to flow through the content encoder in order to satisfy the reconstruction loss, guiding it instead towards the style encoder.

Perceptual aspects

On the perceptual side, there remains many questions related to the perception and recognisability of stylized virtual characters. For instance we have shown the requirements for a moderate stylization if the character needs to remain easily recognizable, measuring the relation between style strength and recognisability in a global manner. Instead of considering its global impact, it could be interesting to instead focus on individual facial features, some parts of the human appearance perhaps being more tolerant to stylization than others. This would help us to understand how to perceptually perform partial stylization (which has shown itself to be necessary in most cases) to focus only on some features (e.g. the eyes, nose, texture, etc.), allowing better recognition for the same perceived style level. Recognition is performed not only using facial appearance features,

but all parts of what can be perceived from a person.

Another focus could be on the possible improvements brought by facial or full body movement, as it has been shown that it could sometimes be enough to recognize an individual [26]. In the case of emotions, combining facial and body movement also helps with recognition [154]. This brings the question of how much body or facial movement could improve stylized character stylization. A study similar to the one of Chapter 2 but designed with moving characters could reveal higher recognition rates at similar stylization levels. Virtual characters being often in movement, this could allow stronger appearance stylization.

On a similar note, a character's voice often have a central importance, and could certainly be used for better recognition. As stylized characters in content such as games or movies typically talk, there could be an interest in studying the recognition of stylized characters in this specific context. This also bring into view the topic of voice stylization, which would although be a longer term perspective, Finally, race bias theory show that difficulties distinguishing faces from other ethnicities diminish with enough exposure to them. A study could be made to evaluate whether as people get exposed to more stylized faces, they get better at recognizing them. If so, it would be interesting to identify if factors other than time impact it, such as how much they are shown talking, moving, or if the exposition is gradual or not, and if these can be leveraged in order to accelerate the adaptation by using specific techniques.

6.2.2 Medium-term perspectives

Leveraging 2D data for 3D style transfer

As handcrafted 3D data remain extremely scarce, heavily limiting possibilities of learning a particular notion of style, future work could include focusing on 2D data instead, and then mapping it to the 3D space. This brings the question of how to leverage 2D data in order to generate stylized 3D content. An answer could be a two step method, where a facial image is first stylized in the 2D domain, and then a 3D version of the result is inferred. Inferring 3D content from 2D data in this case could be done in three main ways : The first would be inspired from the literature on human 3D facial shape inference from 2D, and use a differentiable renderer [155], [156], which is a renderer that work using only differentiable operations, allowing to compute gradient, and thus to use it for deep learning.

In this family of method, the geometry, texture, pose, and lighting are inferred from the latent code of a pre-trained generative network, using a network optimized by comparing the rendered result to the corresponding generated image. After training, the network can therefore estimate both geometry and texture from the latent code of a stylized 2D face, allowing to produce 3D stylized faces after performing the stylisation in the 2D domain. The second approach would be to learn the 2D stylization from both real and realistic rendered human faces, and then learn a mapping from human latent to 3D human scans, using the meshes used for the renders as ground truth. Using a large enough number of 3D stylized faces in the same manner, and taking advantage of the shared latent features, would allow to get an approximation of the non-human spaces present in the latent space. While this would allow for more precised loss than using a renderer, as there would be no information lost in rendering parameter inference and projection, this would probably require a non-negligible amount of 3D stylized data. The two main limits of this approach would be getting renders close enough to real photos, and particularly getting enough 3D stylized faces to teach the network to produce accurate 3D faces from all kind of latent codes. Indeed, the required amount would although probably be considerably lower than in a pure 3D stylization case, as this method would leverage large quantities of 2D data. Hence we would recommend the first approach in a case where there truly no stylized 3D data available, and the second if there is some, but not enough for a fully 3D approach. The third main way is perhaps the most promising. The generative network could directly learn the notion of camera pose by generating its data using camera poses [157]. In this approach, the generative network output images pixel by pixel, taking directional rays as input, instead of using a decoder like architecture and generating all the image at once.

Learning the expressions of non-realistic faces

While most research on facial stylization focus on neutral faces, expressions are a crucial aspect of a face and its perception. Despite work on morphology aware facial expression transfer, it has not been applied on stylized faces such as orcs, or aliens, hence there is no certainty on how well expressions such as smiling or frowning could be transferred, and perceived. Expression wise, the lack of stylized expressive 3D faces is even stronger than the lack of neutral 3D faces. Following the same principle as above, we believe that 2D stylized faces could be leveraged to perform not only style transfer, but expression transfer in the 2D domain, using the latent space of a generative network. The expressive 3D face

could then be reconstructed the same way, either through semi-supervised training with partial labels, or using a differentiable renderer. Transferring the specific way that one convey emotions with their face to a stylized face is an unanswered topic as well. Some individuals have very recognizable facial expressions, which we might wish to transfer. In the case where enough facial expression data is available, we although believe this problem to be directly aligned with the one of transferring specific caricature styles of Part 6.2.1.

6.2.3 Long-term perspectives

Full body stylization

There remain many questions to answer on the topic of stylization, one significant being how to stylize the entirety of the body. While facial stylization is certainly the most important aspect to focus on, it is not the only one. We presented in the last section ways in which 3D stylized faces could produced from only 2D data. Full body images could be leveraged the same way, although it would prove more difficult, as full body data is less numerous than facial data, and existing generative networks prove too limited to capture its important spatial variations. Indeed, they have showcased remarkable performance in some contexts, such as faces, but their capabilities to model data with higher spatial variation, such as buildings, animal pictures, or even images of randomly placed simple shapes, while sometimes quite convincing, have show themselves to be lower. Some variations of the approach has been tried in order to solve the issue, such as introducing an attention mechanism to the generative model, showing improved results on several datasets [158], but still not on levels comparable to generated faces. Recent approaches at modelling image dataset distribution explicitly model it as a projection of three-dimensional data in order to improve camera movement in the latent space [157], [159], which might also prove itself to be a useful prior modelling other kind of spatial variation, although these specific papers do not explore this direction. It would require architectural advances. The problem might also be set aside aside by using highly aligned data (with little pose variation), which would although be very challenging to obtain, as data scrapped from the internet does not fit this constraint. Another way to answer this limitation could lie in using data paired from different domains, such as image and text pairs, allowing to learn a generative model of both images and text, each providing semantic cues to the other, working as soft labels. Recent research in this domain have shown significant capabilities for generating

spatially complex images [160]. Their capabilities at generating full bodies although again fall short of comparison with faces generation. This direction of research would also open the door to text based stylization, where one could create various style by simply writing text, without necessarily the need of an example stylized character image. This perspective of using image based learning approaches for stylization also implies that these would be readily available for anyone with a camera, instead of being restricted to those with an available 3D scan system. This follow the same path as general facial reconstruction systems, which are now able to produce convincing scans from as little as one photo thanks to learning precise mappings between the 3D and image domain [161].

To end this thesis on a more general note, virtual characters have become ubiquitous in entertainment and are steadily gaining ground in more socially oriented applications. The ability to represent and stylize one's own virtual character is promised to be key to a large amount of virtual interactive worlds, and while the combination of existing capture approaches combined with the development of affordable acquisition devices (smartphones) already address most of the first point, character stylization methods remain far and between. While the short term solution to such stylization is in learning-based methods leveraging 3D data, or even rule-based approaches, in the longer run we believe that they ability to interpret 2D data in a 3D aware manner to be key, the latter being vastly more numerous and easier to produce. At present, we hope that the solutions proposed in this thesis will be useful to allow the automatic generation of expressive stylized characters, and their limits, and will give insights towards an extension of it to the whole body, and from 2D content.

Annexe A

Author's publications

This thesis manuscript is based upon the following publications.

Articles

Conference papers

F. Danieau, I. Gubins, N. Olivier, O. Dumas, B. Denis, T. Lopez, N. Mollet, B. Frager, Q. Avril, "Automatic Generation and Stylization of 3D Facial Rigs.", *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Pages 784–92. 2019

N. Olivier, L. Hoyet, F. Argelaguet, F. Danieau, Q. Avril, P. Guillotel, A. Lécuyer, F. Multon, "The impact of stylization on face recognition ", *ACM Symposium on Applied Perception (SAP)*, Article No. : 15, Pages 1–9, 2020.

Journal papers

N. Olivier, G. Kerbiriou, F. Argelaguet, Q. Avril, F. Danieau, P. Guillotel, L. Hoyet, F. Multon, "Study on automatic 3D facial caricaturization : from rules to deep learning ", *Frontiers in Virtual Reality : Creating Lifelike Digital Humans*, 2021

N. Olivier, K. Baert, F. Danieau, F. Multon, Q. Avril, "FaceTuneGAN : Face Autoencoder for Convolutional Expression Transfer Using Neural Generative Adversarial Networks", *Computer & Graphics*, Submitted

Patents Applications

N. Olivier, F. Danieau, Q. Avril, P. Guillotel, F. Argelaguet Sanz, A. Lecuyer, F. Multon, L. Hoyet, "Unsupervised Deep Learning-based caricature generation of facial 3D meshes", *2021ID00168*, 2021.

F. Danieau, G. Kerbiriou, Q. Avril, N. Olivier, "Automatic caricature generation of facial 3D meshes", *2020ID00287*, 2020.

Annexe B

Additional Figures for Chap 3

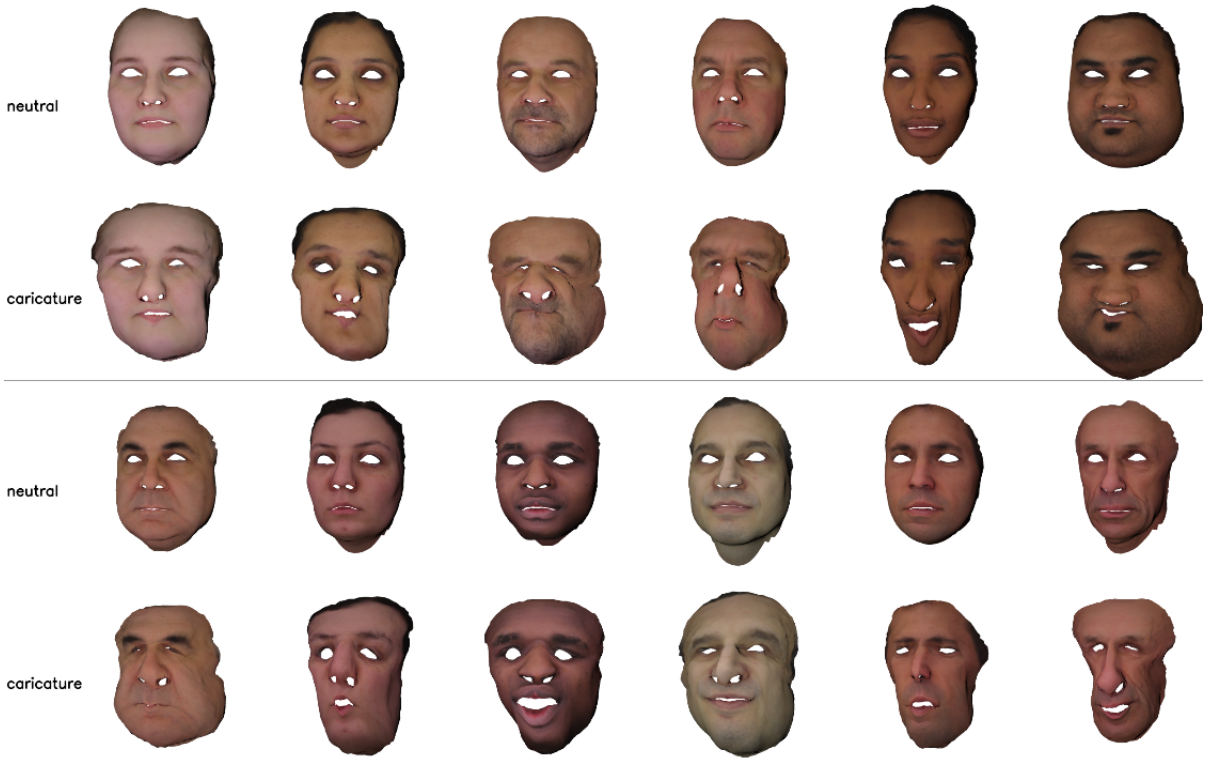


Figure B.1 – Caricaturization using Deep

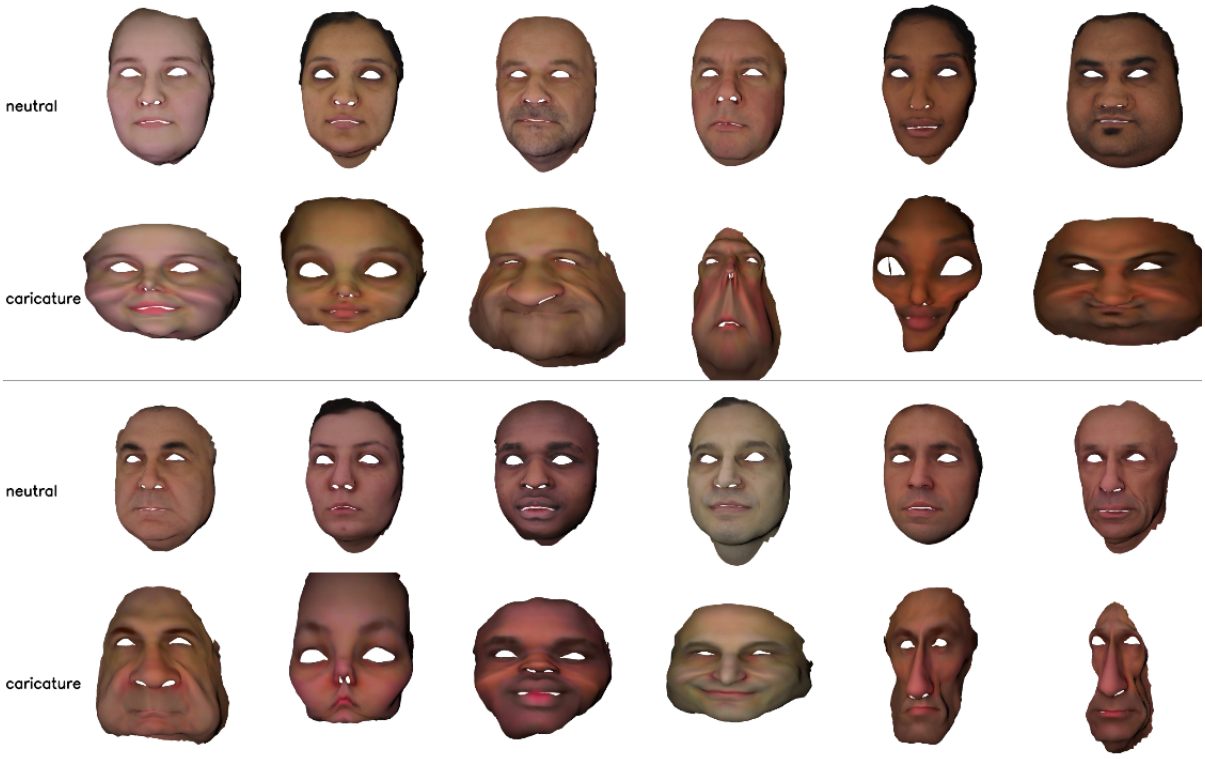


FIGURE B.2 – Caricaturization using Geo.1

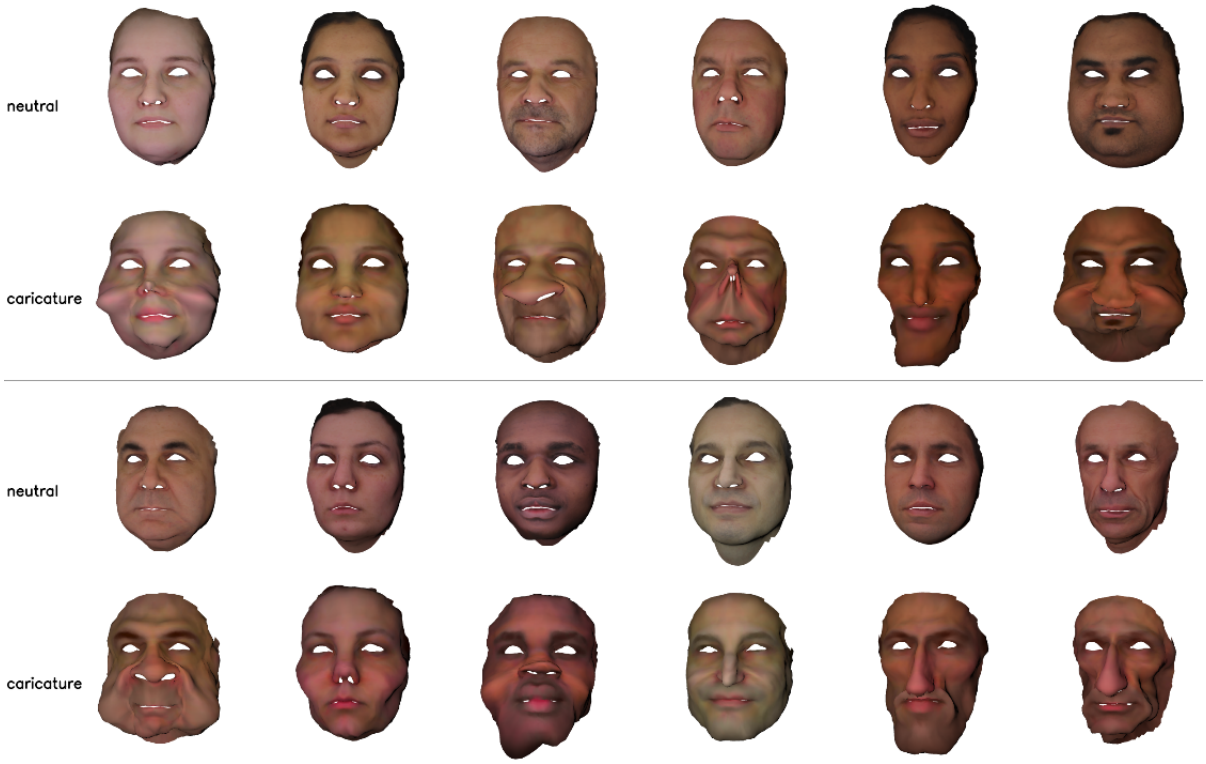


FIGURE B.3 – Caricaturization using Geo.2

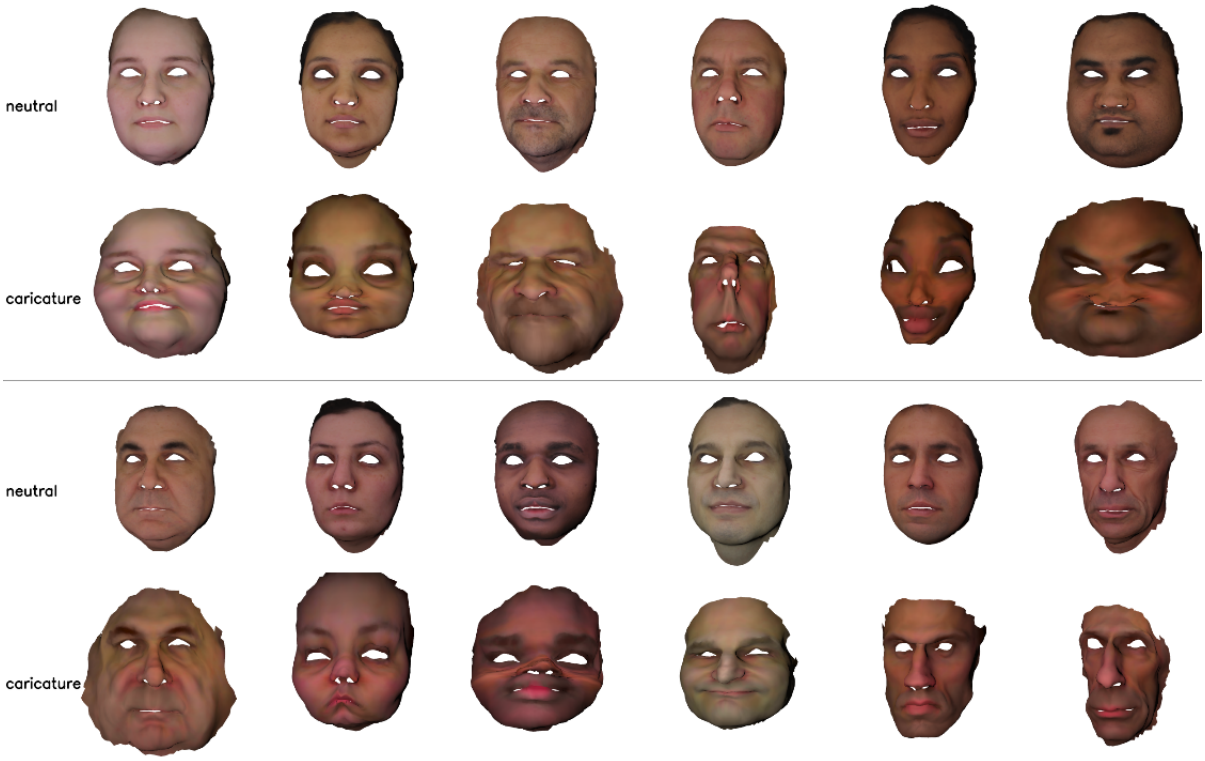


FIGURE B.4 – Caricaturization using EDFM

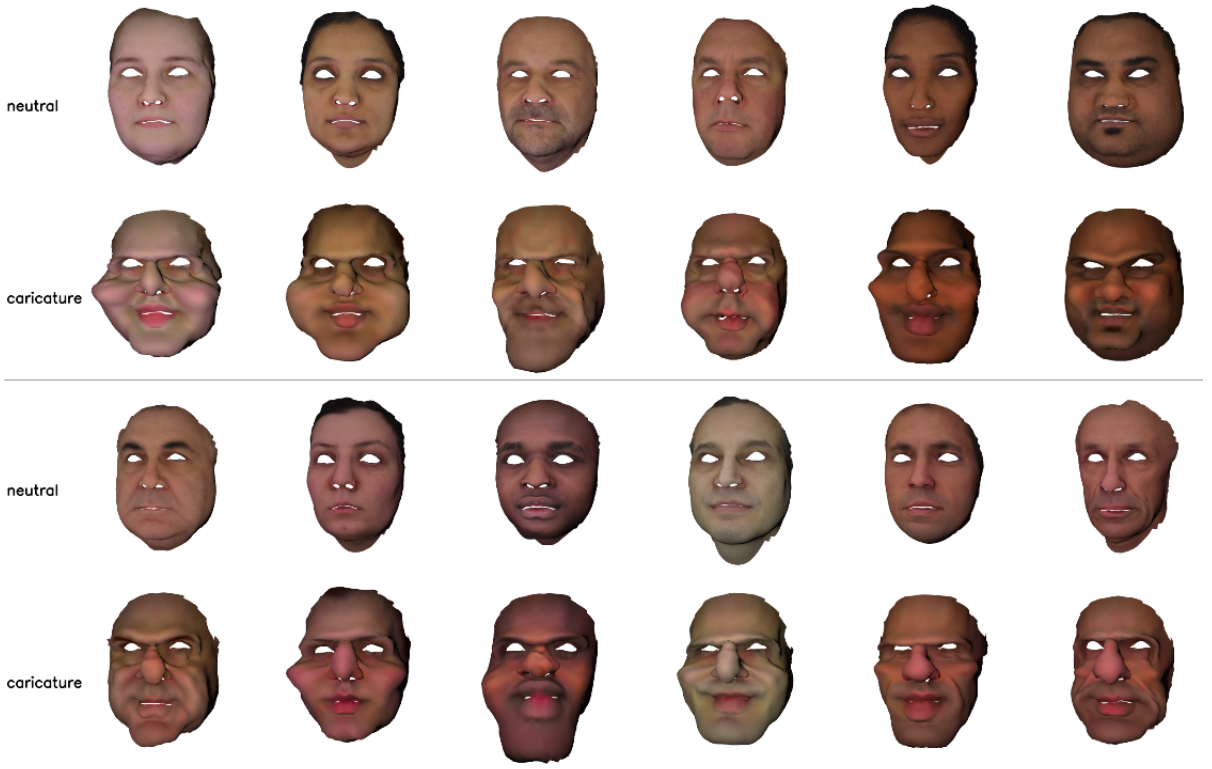


FIGURE B.5 – Caricaturization using Sela

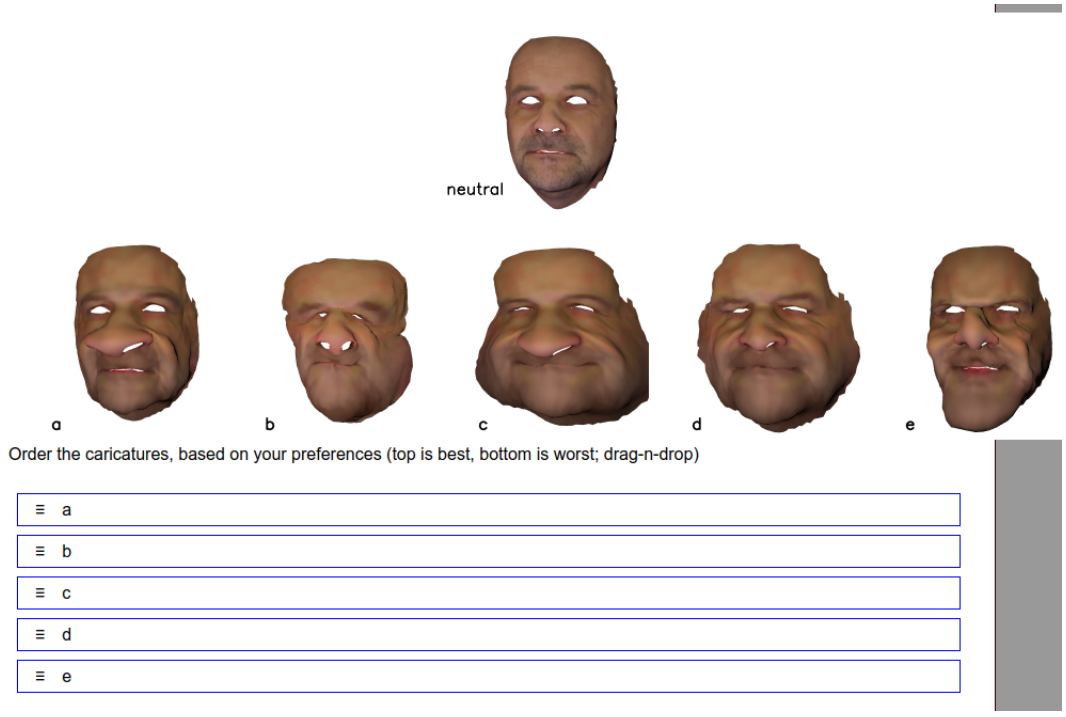


FIGURE B.6 – Sample of the ranking task of the user study

Annexe C

Architecture Details and Additional Figures for Chap 4

This section details the specific layers of our expression transfer network. Due to the different amount of vertices in the datasets (27k in FaceScape, 5k in CoMA), the architecture is tweaked differently for each. Figure C.1 and C.2 respectively depict the composition of our network modules for training on CoMA and FaceScape. All linear layers use ReLU activations, except for the final discriminator layer which uses a Softmax activation. All SpiralBlocks and SpiralResBlocks use Exponential Linear Unit (ELU) activations [162]. The style encoder and discriminator use no normalization. The content encoder uses Instance Normalization layers on all SpiralBlocks and SpiralResBlocks. The decoder uses Adaptive Instance Normalization on its two first SpiralResBlocks and first SpiralBlock to inject the style information.

The mapping function M that converts the style codes into parameters for the AdaIN normalization layers of the decoder is composed of fully-connected layers. No normalization is used, and all layers use ReLU activations except for the last one. On CoMA, we use the following configuration : $4 \rightarrow 128 \rightarrow 128 \rightarrow 16 \rightarrow N_{\text{AdaIN}}$ where N_{AdaIN} is the number of AdaIN parameters in the decoder. On FaceScape, we use more layers and a larger size : $5 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 16 \rightarrow N_{\text{AdaIN}}$.

C.1 CoMA dataset

Figure C.3 shows one selected frame for each (subject, expression) pair of the CoMA [102] dataset.

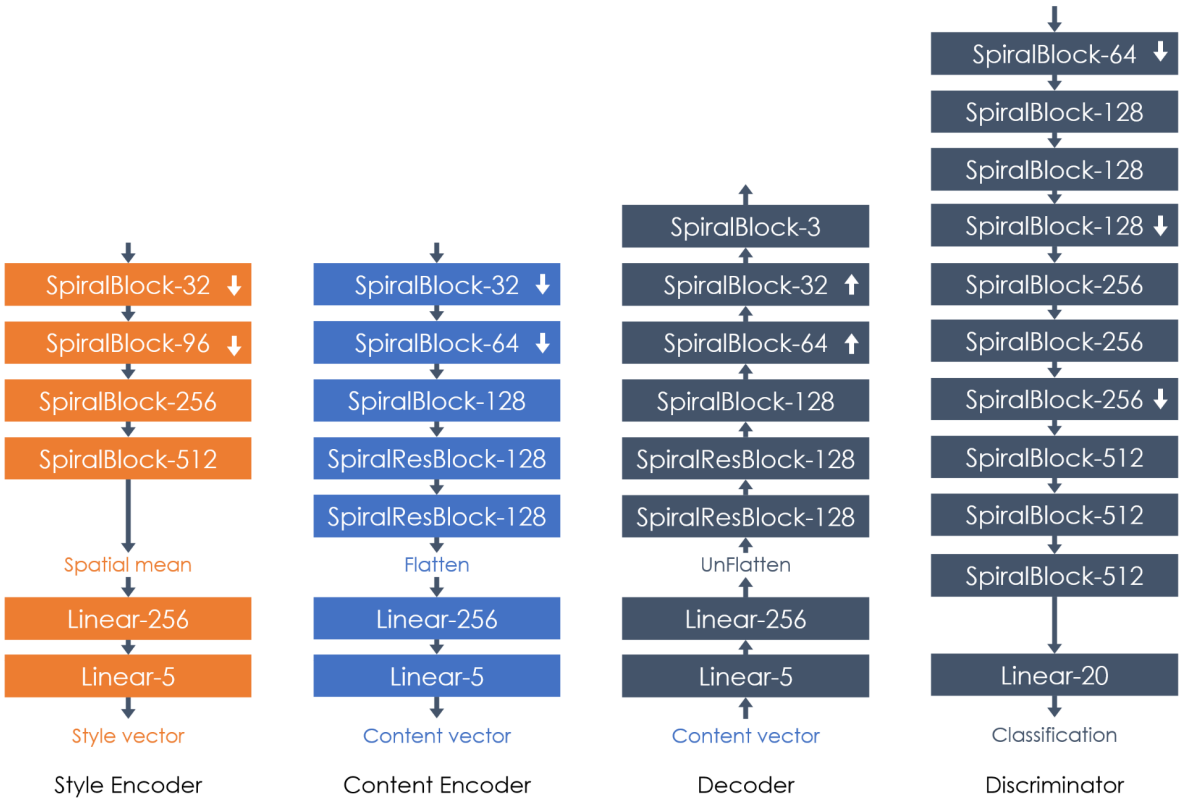


FIGURE C.1 – Composition of each module for training on FaceScape. The down/up arrows respectively indicate down-sampling and up-sampling layers, always by a factor of 4.

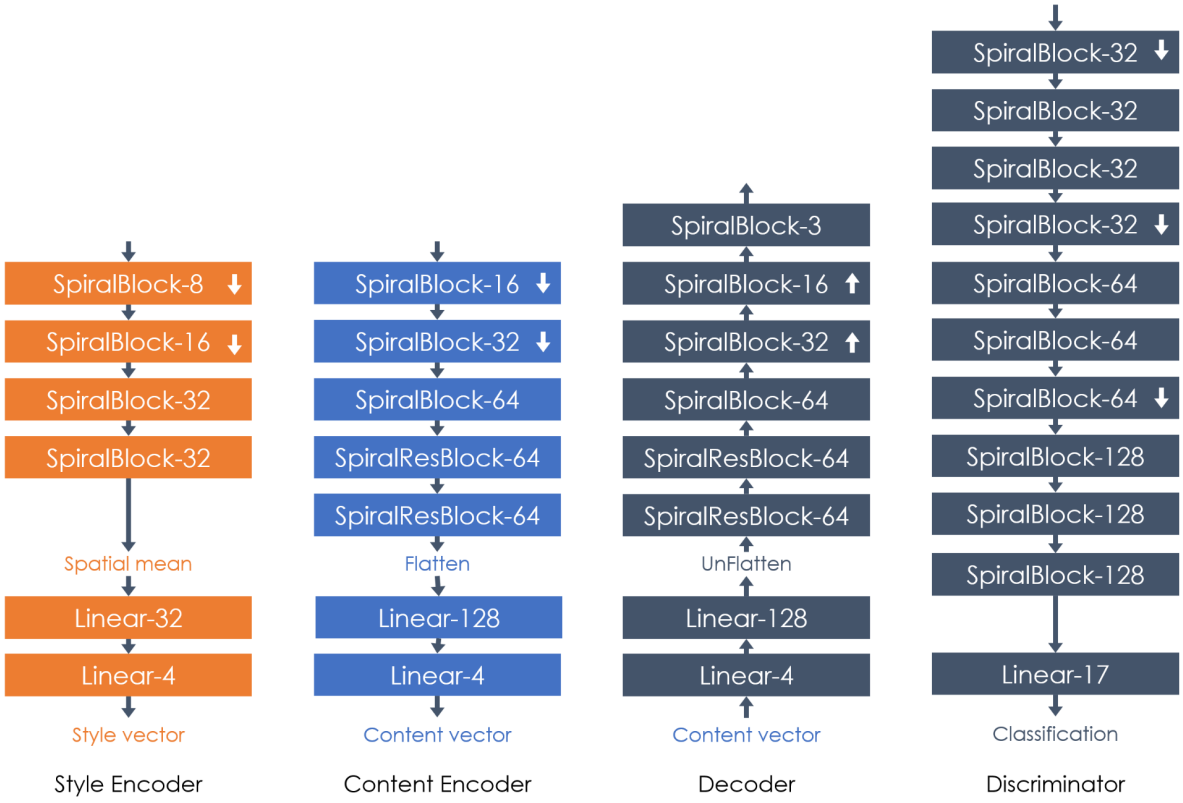


FIGURE C.2 – Network modules for training on CoMA. The down/up arrows respectively indicate down-sampling and up-sampling layers, always by a factor of 4.

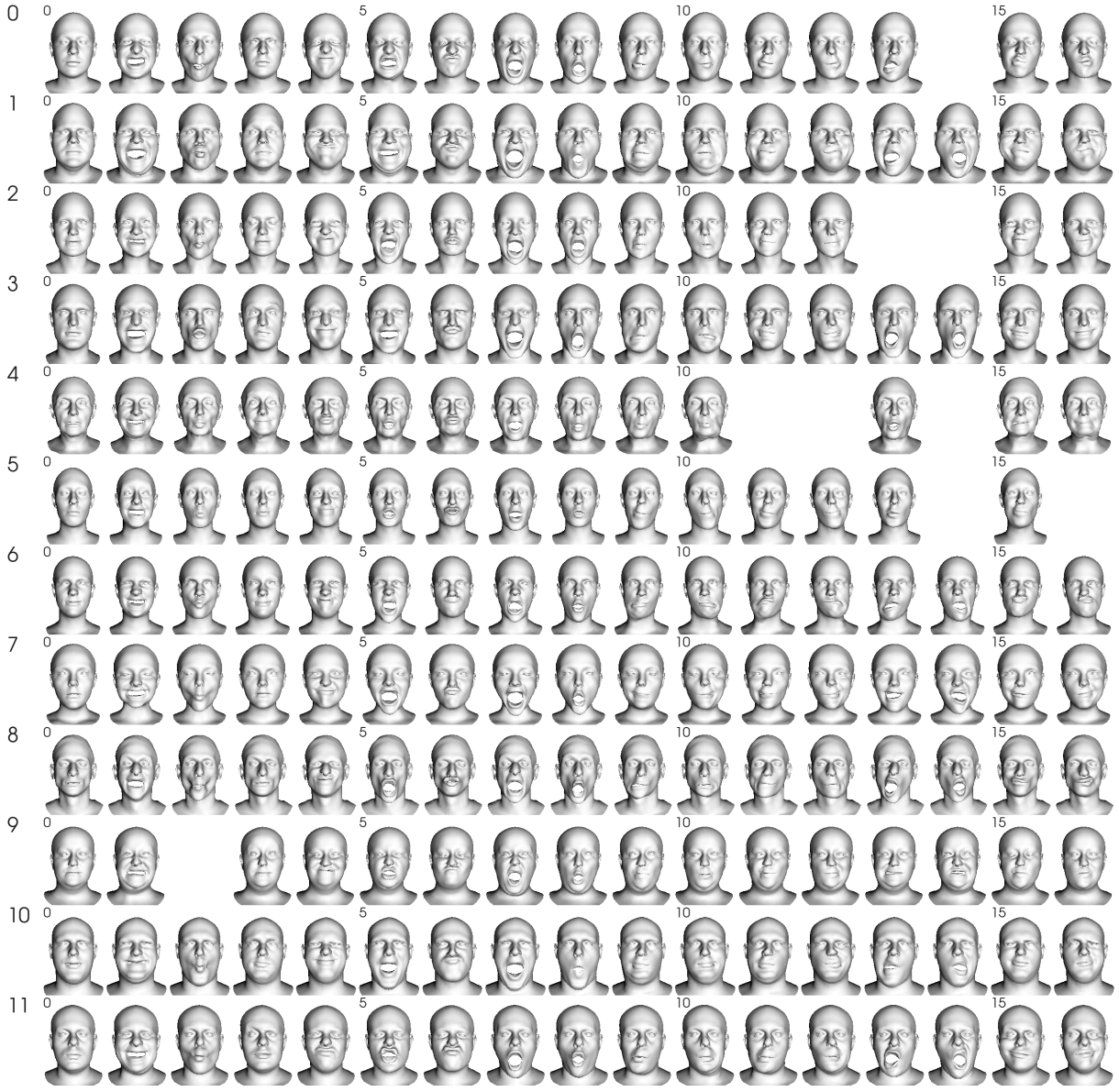


FIGURE C.3 – Selected frames on CoMA. Entries are missing when the expression performed by the subject varies significantly from the rest.

C.2 Interpolations

Expressions and identities interpolations are provided in Figures C.4 and C.5 respectively.

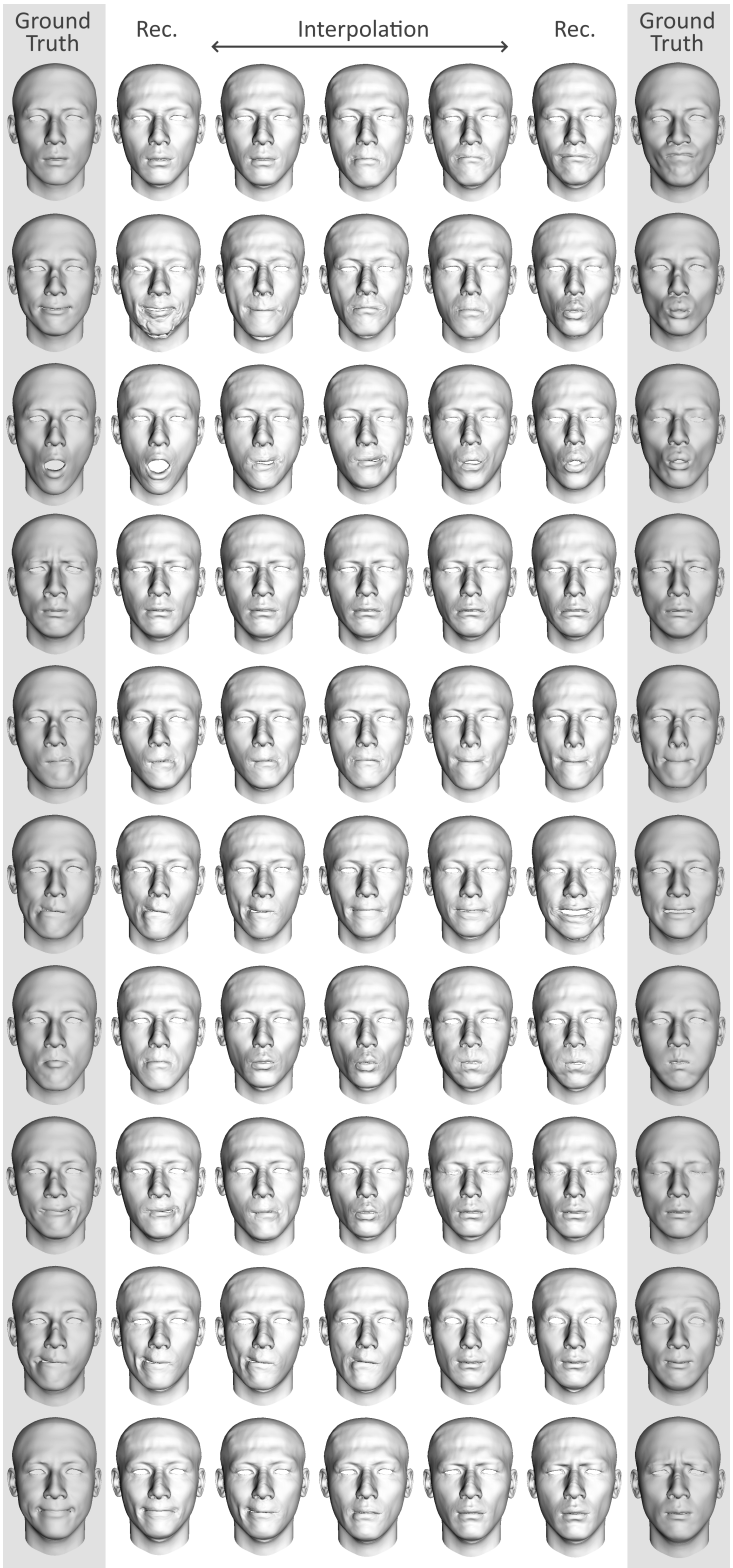


FIGURE C.4 – Expression interpolation results on FaceScape (style space). The ground truth scans are provided on the leftmost and rightmost columns. The second and second-to-last columns show their reconstruction. In the middle three columns, we interpolate along the style vector ($s = 0.25$).

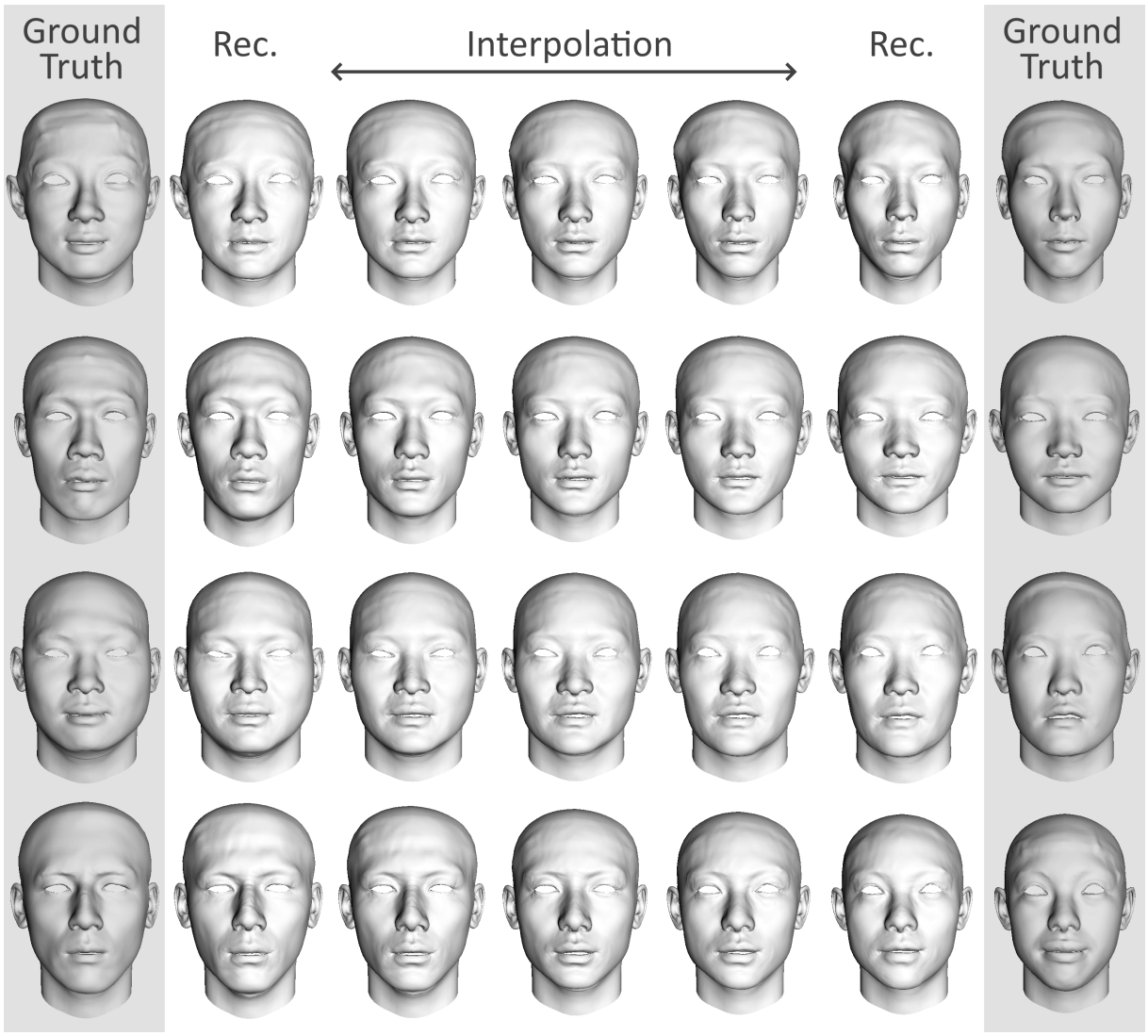


FIGURE C.5 – Identity interpolation results on FaceScape (content space). The ground truth scans are provided on the leftmost and rightmost columns. The second and second-to-last columns show their reconstruction. In the middle three columns, we interpolate along the content vector ($s = 0.25$).

BIBLIOGRAPHIE

[1] B. Wang, F. Xia, M. S. Engel, V. Perrichot, G. Shi, H. Zhang, J. Chen, E. A. Jarzembowski, T. Wappler et J. Rust, « Debris-carrying camouflage among diverse lineages of Cretaceous insects », t. 2, *6*, juin 2016. doi : 10.1126/sciadv.1501918.

[2] R. Garrouste, S. Hugel, L. Jacquelin, P. Rostan, J.-S. Steyer, L. Desutter-Grandcolas et A. Nel, « Insect mimicry of plants dates back to the Permian », t. 7, *1*, déc. 2016. doi : 10.1038/ncomms13735.

[3] N. Creanza, O. Kolodny et M. W. Feldman, « Cultural evolutionary theory : How culture evolves and why it matters », t. 114, *30*, p. 7782-7789, juill. 2017. doi : 10.1073/pnas.1620732114.

[4] M. J. Horn, *The second skin : An interdisciplinary study of clothing*. 1967.

[5] N. Yee, J. N. Bailenson et N. Ducheneaut, « The Proteus Effect », *Communication Research*, t. 36, *2*, p. 285-312, jan. 2009. doi : 10.1177/0093650208330254.

[6] T. Todorović, T. Toporišič, A. P. Čuden et and, « Clothes and Costumes as Form of Nonverbal Communication », t. 57, *4*, p. 321-333, déc. 2014. doi : 10.14502/tekstilec2014.57.321-333.

[7] M. J. Horn, *The second skin : An interdisciplinary study of clothing*. Columbia University Press, 1967.

[8] G. S. Cross, *Worktime and Industrialization : An International History, Labor and Social Change.* Temple University Press, 1988.

[9] J. A. Nobuya Haraguchi Khuong Minh Vu, « Accelerated Globalization and The Dynamics of Deindustrialization », 2018.

[10] M. J. Sheehan et M. W. Nachman, « Morphological and population genomic evidence that human faces have evolved to signal individual identity », in *Nature Communications*, t. 5, déc. 2014, p. 4800. doi : 10.1038/ncomms5800. adresse : http://www.nature.com/articles/ncomms5800 (visité le 21/02/2020).

[11] T. Park, M.-Y. Liu, T.-C. Wang et J.-Y. Zhu, « Semantic Image Synthesis with Spatially-Adaptive Normalization », in *arXiv :1903.07291 [cs]*, arXiv : 1903.07291, mars 2019. adresse : http://arxiv.org/abs/1903.07291 (visité le 29/03/2019).

[12] J.-Y. Zhu, T. Park, P. Isola et A. A. Efros, « Unpaired image-to-image translation using cycle-consistent adversarial networks », in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. doi : 10.1109/ICCV.2017.244.

[13] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen et J. Kautz, « Few-Shot Unsupervised Image-to-Image Translation », in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, p. 10 550-10 559. doi : 10.1109/ICCV.2019.01065.

[14] L. A. Gatys, A. S. Ecker et M. Bethge, « Image Style Transfer Using Convolutional Neural Networks », in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, t. 2016-Decem, IEEE, 2016, p. 2414-2423, isbn : 9781467388504. doi : 10.1109/CVPR.2016.265. adresse : http://ieeexplore.ieee.org/document/7780634/.

[15] V. Blanz et T. Vetter, « A morphable model for the synthesis of 3D faces », in *ACM SIGGRAPH*, 1999, p. 187-194, isbn : 978-0-201-48560-8. doi : 10.1145/311535.311556. adresse : http://portal.acm.org/citation.cfm?doid=311535.311556 (visité le 17/07/2019).

[16] P. Kaur, H. Zhang et K. J. Dana, « Photo-realistic Facial Texture Transfer », in *arXiv :1706.04306 [cs]*, arXiv : 1706.04306, juin 2017. adresse : http://arxiv.org/abs/1706.04306 (visité le 08/02/2019).

[17] P. Isola, J.-Y. Zhu, T. Zhou et A. A. Efros, « Image-to-Image Translation with Conditional Adversarial Networks », in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] J. Huo, W. Li, Y. Shi, Y. Gao et H. Yin, « WebCaricature : a benchmark for caricature recognition », in *British Machine Vision Conference*, ArXiv : 1703.03230, 2018.

[19] J. N. M. Pinkney et D. Adler, *Resolution Dependent GAN Interpolation for Controllable Image Synthesis Between Domains*, 2020. eprint : arXiv:2010.05334.

[20] K. Cao, J. Liao et L. Yuan, « CariGANs : Unpaired Photo-to-Caricature Translation », en, *ACM Transactions on Graphics*, t. 37, *6*, p. 1-14, déc. 2018, issn : 07300301. (visité le 29/11/2019).

[21] E. Akleman et J. Reisch, « Modeling expressive 3D caricatures », in *ACM SIGGRAPH*, 2004. doi : 10.1145/1186223.1186299.

[22] Z. Ye, R. Yi, M. Yu, J. Zhang, Y. Lai et Y. Liu, « 3D-CariGAN : An End-to-End Solution to 3D Caricature Generation from Face Photos », in *Computing Research Repository (CoRR)*, ArXiv : 2003.06841, 2020. arXiv : 2003.06841. adresse : https://arxiv.org/abs/2003.06841.

[23] Y. Guo, L. Jiang, L. Cai et J. Zhang, « 3D magic mirror : Automatic video to 3D caricature translation », in *CoRR*, arXiv : 1906.00544, t. abs/1906.00544, 2019. adresse : http://arxiv.org/abs/1906.00544 (visité le 02/12/2021).

[24] S. F. Galton, « Composite portraits, made by combining those of many different persons into a single, resultant figure », *Journal of the Anthropological Institute*, 1879.

[25] J. M. P. I. Pollack et W. H. Sumby, « On the Identification of Speakers by Voice », *The Journal of the Acoustical Society of America 26, 403*, 1954.

[26] L. Hoyet, K. Ryall, K. Zibrek, H. Park, J. Lee, J. Hodgins et C. O'Sullivan, « Evaluating the distinctiveness and attractiveness of human motions on realistic virtual bodies », in *ACM Transactions on Graphics (TOG)*, t. 32, nov. 2013, p. 1-11. doi : 10.1145/2508363.2508367. adresse : http://dl.acm.org/citation.cfm?doid=2508363.2508367 (visité le 21/02/2020).

[27] G. Johansson, « Visual perception of biological motion and a model for its analysis », *Perception & Psychophysics*, 1973.

[28] G. J. Kaufman et K. J. Breeding, « The Automatic Recognition of Human Faces from Profile Silhouettes », *IEEE Transactions on Systems, Man, and Cybernetics*, t. SMC-6, *2*, p. 113-121, 1976. doi : 10.1109/tsmc.1976.5409181.

[29] V. Bruce et A. Young, « Understanding face recognition », in *British Journal of Psychology*, t. 77, 1986, p. 305-327. doi : 10.1111/j.2044-8295.1986.tb02199.x. adresse : https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1986.tb02199.x (visité le 21/02/2020).

[30] B. V., *Recognizing faces*. Lawrence Erlbaum Associates, 1988.

[31] H. Leder et V. Bruce, « Local and Relational Aspects of Face Distinctiveness », in *The Quarterly Journal of Experimental Psychology Section A*, t. 51, août 1998, p. 449-473. doi : 10.1080/713755777. adresse : http://journals.sagepub.com/doi/10.1080/713755777 (visité le 21/02/2020).

[32] J. W. Tanaka et M. J. Farah, « Parts and Wholes in Face Recognition », in *The Quarterly Journal of Experimental Psychology Section A*, t. 46, mai 1993, p. 225-245. doi : 10.1080/14640749308401045. adresse : http://journals.sagepub.com/doi/10.1080/14640749308401045 (visité le 21/02/2020).

[33] R. K. Yin, « Looking at upside-down faces. », in *Journal of Experimental Psychology*, t. 81, 1969, p. 141-145. doi : 10.1037/h0027474. adresse : http://doi.apa.org/getdoi.cfm?doi=10.1037/h0027474 (visité le 21/02/2020).

[34] H. D. E. G. M. Davies et J. W. Shepherd, « Studies of cue saliency », *Perceiving and remembering faces, vol 96, University of Illinois Press*, 1981.

[35] W. Zhao, R. Chellappa, P. J. Phillips et A. Rosenfeld, « Face recognition : A literature survey », in *ACM Computing Surveys*, t. 35, déc. 2003, p. 399-458. doi : 10.1145/954339.954342. adresse : http://portal.acm.org/citation.cfm?doid=954339.954342 (visité le 21/11/2019).

[36] V. Bruce, M. A. Burton et N. Dench, « What's Distinctive about a Distinctive Face ? », in *The Quarterly Journal of Experimental Psychology Section A*, t. 47, fév. 1994, p. 119-141. doi : 10.1080/14640749408401146. adresse : http://journals.sagepub.com/doi/10.1080/14640749408401146 (visité le 21/02/2020).

[37] K. A. Deffenbacher, T. Vetter, J. Johanson et A. J. O'Toole, « Facial Aging, Attractiveness, and Distinctiveness », in *Perception*, t. 27, oct. 1998, p. 1233-1243. doi : 10.1068/p271233. adresse : http://journals.sagepub.com/doi/10.1068/p271233 (visité le 21/02/2020).

[38] T. Valentine, S. Darling et M. Donnelly, « Why are average faces attractive ? The effect of view and averageness on the attractiveness of female faces », in *Psychonomic Bulletin & Review*, t. 11, juin 2004, p. 482-487. doi : 10.3758/BF03196599. adresse : http://link.springer.com/10.3758/BF03196599 (visité le 21/02/2020).

[39] T. V. H. A. W. Volker Blanz Alice J. O'Toole, « On The Other Side of the Mean : The Perception of Dissimilarity in Human Faces », *Perception*, 2000.

[40] C. A. Meissner et J. C. Brigham, « Thirty years of investigating the own-race bias in memory for faces : A meta-analytic review. », in *Psychology, Public Policy, and Law*, t. 7, 2001, p. 3-35. doi : 10.1037/1076-8971.7.1.3. adresse : http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-8971.7.1.3 (visité le 21/02/2020).

[41] P. Chiroro et T. Valentine, « An Investigation of the Contact Hypothesis of the Own-race Bias in Face Recognition », in *The Quarterly Journal of Experimental Psychology Section A*, t. 48, nov. 1995, p. 879-894. doi : 10.1080/14640749508401421. adresse : http://journals.sagepub.com/doi/10.1080/14640749508401421 (visité le 21/02/2020).

[42] W.-J. Ng et R. C. Lindsay, « Cross-Race Facial Recognition : Failure of the Contact Hypothesis », in *Journal of Cross-Cultural Psychology*, t. 25, juin 1994, p. 217-232. doi : 10.1177/0022022194252004. adresse : http://journals.sagepub.com/doi/10.1177/0022022194252004 (visité le 21/02/2020).

[43] D. T. Levin, « Race as a visual feature : Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. », in *Journal of Experimental Psychology : General*, t. 129, 2000, p. 559-574. doi : 10.1037/0096-3445.129.4.559. adresse : http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.129.4.559 (visité le 21/02/2020).

[44] F. Danieau, I. Gubins, N. Olivier, O. Dumas, B. Denis, T. Lopez, N. Mollet, B. Frager et Q. Avril, « Automatic generation and stylization of 3D facial rigs », in *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, mars 2019, p. 784-792. doi : 10.1109/VR.2019.8798208.

[45] E. Zell, C. Aliaga, A. Jarabo, K. Zibrek, D. Gutierrez, R. McDonnell et M. Botsch, « To stylize or not to stylize ? : the effect of shape and material stylization on the perception of computer-generated faces », en, *ACM Transactions on Graphics*, t. 34, *6*, p. 1-12, oct. 2015, issn : 07300301. (visité le 20/03/2019).

[46] R. McDonnell, M. Breidt et H. H. Bülthoff, « Render me real ? : investigating the effect of render style on the perception of animated virtual humans », in *ACM Transactions on Graphics (TOG)*, t. 31, juill. 2012, p. 1-11. doi : 10.1145/2185520.2185587. adresse : http://dl.acm.org/citation.cfm?doid=2185520.2185587 (visité le 28/03/2019).

[47] C. Wallraven, H. H. Bülthoff, D. W. Cunningham, J. Fischer et D. Bartz, « Evaluation of real-world and computer-generated stylized facial expressions », t. 4, *3*, p. 16, nov. 2007. doi : 10.1145/1278387.1278390.

[48] P. Wisessing, K. Zibrek, D. W. Cunningham, J. Dingliana et R. McDonnell, « Enlighten Me », *ACM Transactions on Graphics*, t. 39, *3*, p. 1-12, juin 2020. doi : 10.1145/3383195.

[49] R. Fleming, B. J. Mohler, J. Romero, M. J. Black et M. Breidt, « Appealing Female Avatars from 3D Body Scans : Perceptual Effects of Stylization : » in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016, p. 333-343, isbn : 978-989-758-175-5. doi : 10.5220/0005683903330343. adresse : http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005683903330343 (visité le 13/03/2019).

[50] A. A. Efros et W. T. Freeman, « Image quilting for texture synthesis and transfer », ACM Press, 2001. doi : 10.1145/383259.383296.

[51] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless et D. H. Salesin, « Image analogies », ACM Press, 2001. doi : 10.1145/383259.383295.

[52] M. Ashikhmin, « Fast texture transfer », t. 23, *4*, p. 38-43, juill. 2003. doi : 10.1109/mcg.2003.1210863.

[53] H. Lee, S. Seo, S. Ryoo et K. Yoon, « Directional texture transfer », ACM Press, 2010. doi : 10.1145/1809939.1809945.

[54] K. Simonyan et A. Zisserman, « Very Deep Convolutional Networks for Large-Scale Image Recognition », in *International Conference on Learning Representations*, 2015.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li et L. Fei-Fei, « Imagenet : A large-scale hierarchical image database », in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, p. 248-255.

[56] L. A. Gatys, A. S. Ecker et M. Bethge, « A Neural Algorithm of Artistic Style », in *arXiv :1508.06576 [cs, q-bio]*, arXiv : 1508.06576, sept. 2015. adresse : http://arxiv.org/abs/1508.06576 (visité le 20/12/2019).

[57] C. Li et M. Wand, « Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis », in *arXiv :1601.04589 [cs]*, arXiv : 1601.04589, jan. 2016. adresse : http://arxiv.org/abs/1601.04589 (visité le 05/04/2019).

[58] A. Selim, M. Elgharib et L. Doyle, « Painting style transfer for head portraits using convolutional neural networks », in *ACM Transactions on Graphics (TOG)*, t. 35, juill. 2016, p. 1-18. doi : 10.1145/2897824.2925968. adresse : http://dl.acm.org/citation.cfm?doid=2897824.2925968 (visité le 25/11/2019).

[59] A. J. Champandard, « Semantic Style Transfer and Turning Two-Bit Doodles into Fine Artworks », in *ArXiv*, 2016.

[60] S. E. Brennan, « Caricature Generator : The Dynamic Exaggeration of Faces by Computer », in *Leonardo*, t. 18, 1985, p. 170-178.

[61] E. Akleman, « Making caricatures with morphing », in *ACM SIGGRAPH*, 1997. doi : 10.1145/259081.259231.

[62] T. Fujiwara, H. Koshimizu, K. Fujimura, G. Fujita, Y. Noguchi et N. Ishikawa, « A method for 3D face modeling and caricatured figure generation », in *IEEE International Conference on Multimedia and Expo*, t. 2, 2002, 137-140 vol.2. doi : 10.1109/ICME.2002.1035531.

[63] Z. Mo, J. P. Lewis et U. Neumann, « Improved automatic caricature by feature normalization and exaggeration », in *ACM SIGGRAPH*, 2004. doi : 10.1145/1186223.1186294.

[64] Y.-L. Chen, W.-H. Liao et P.-Y. Chiang, « Generation of 3D Caricature by Fusing Caricature Images », in *IEEE International Conference on Systems, Man and Cybernetics*, oct. 2006. doi : 10.1109/icsmc.2006.384498.

[65] L. Redman, *How to draw caricatures*, M.-H. Contemporary, éd. 1984.

[66] S. Liu, J. Wang, M. Zhang et Z. Wang, « Three-dimensional cartoon facial animation based on art rules », in *The Visual Computer*, t. 29, nov. 2012, p. 1135-1149. doi : 10.1007/s00371-012-0756-2.

[67] M. Eigensatz, R. W. Sumner et M. Pauly, « Curvature-Domain Shape Processing », in *Computer Graphics Forum*, t. 27, avr. 2008, p. 241-250. doi : 10.1111/j.1467-8659.2008.01121.x.

[68] G. Cimen, A. Bulbul, B. Ozguc et T. Capin, « Perceptual Caricaturization of 3D Models », in *Computer and Information Sciences III*, oct. 2012, p. 201-207. doi : 10.1007/978-1-4471-4594-3_21.

[69] M. Sela, Y. Aflalo et R. Kimmel, « Computational caricaturization of surfaces », in *Computer Vision and Image Understanding*, t. 141, déc. 2015, p. 1-17. doi : 10.1016/j.cviu.2015.05.013.

[70] C. Ma, H. Huang, A. Sheffer, E. Kalogerakis et R. Wang, « Analogy-driven 3D style transfer : Analogy-driven 3D style transfer », in *Computer Graphics Forum*, t. 33, mai 2014, p. 175-184. doi : 10.1111/cgf.12307. adresse : http://doi.wiley.com/10.1111/cgf.12307 (visité le 08/02/2019).

[71] Z. Lun, E. Kalogerakis, R. Wang et A. Sheffer, « Functionality preserving shape style transfer », in *ACM Transactions on Graphics (TOG)*, t. 35, nov. 2016, p. 1-14. doi : 10.1145/2980179.2980237. adresse : http://dl.acm.org/citation.cfm?doid=2980179.2980237 (visité le 08/02/2019).

[72] T. Li, T. Bolkart, M. J. Black, H. Li et J. Romero, « Learning a model of facial shape and expression from 4D scans », in *ACM Transactions on Graphics*, t. 36, 2017, p. 1-17. doi : 10.1145/3130800.3130813. adresse : https://dl.acm.org/doi/10.1145/3130800.3130813.

[73] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz et T. Vetter, « 3D Morphable Face Models – Past, Present and Future », in *arXiv :1909.01815 [cs]*, arXiv : 1909.01815, sept. 2019. adresse : http://arxiv.org/abs/1909.01815 (visité le 27/01/2020).

[74] D. Vlasic, M. Brand, H. Pfister et J. Popović, « Face transfer with multilinear models », in *ACM Transactions on Graphics*, t. 24, 2005, p. 426-433. doi : 10.1145/1073204.1073209.

[75] M. Wang, Y. Panagakis, P. Snape et S. Zafeiriou, « Learning the multilinear structure of visual data », in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, t. 2017-Janua, 2017, p. 6053-6061, isbn : 9781538604571. doi : 10.1109/CVPR.2017.641.

[76] T. Bolkart et S. Wuhrer, « A robust multilinear model learning framework for 3D faces », in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, t. 2016-Decem, 2016, p. 4911-4919, isbn : 9781467388504. doi : 10.1109/CVPR.2016.531.

[77] J. Xie, Y. Chen, J. Liu, C. Miao et X. Gao, « Interactive 3D caricature generation based on double sampling », in *ACM Multimedia*, 2009. doi : 10.1145/1631272.1631403.

[78] P. Li, Y. Chen, J. Liu et G. Fu, « 3D caricature generation by manifold learning », in *IEEE International Conference on Multimedia and Expo*, 2008. doi : 10.1109/icme.2008.4607591.

[79] J. Liu, Y. Chen, C. Miao, J. Xie, C. X. Ling, X. Gao et W. Gao, « Semi-Supervised Learning in Reconstructed Manifold Space for 3D Caricature Generation », in *Computer Graphics Forum*, t. 28, déc. 2009, p. 2104-2116. doi : 10.1111/j.1467-8659.2009.01418.x.

[80] J. Zhou, X. Tong, Z. Liu et B. Guo, « 3D cartoon face generation by local deformation mapping », in *The Visual Computer*, t. 32, juin 2016, p. 717-727. doi : 10.1007/s00371-016-1265-5.

[81] L Clarke, M. Chen et B Mora, « Automatic generation of 3D caricatures based on artistic deformation styles », in *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, t. 17, juin 2011, p. 808-821. doi : 10.1109/tvcg.2010.76.

[82] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville et Y. Bengio, « Generative Adversarial Nets », in *Conference on Neural Information Processing Systems (NIPS)*, ArXiv : 1406.2661,

2014, p. 2672-2680. adresse : https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[83] M.-Y. Liu, T. Breuel et J. Kautz, « Unsupervised image-to-image translation networks », in *Conference on Neural Information Processing Systems (NIPS)*, déc. 2017, p. 700-708, isbn : 978-1-5108-6096-4. doi : 10.5555/3294771.3294838. (visité le 21/02/2020).

[84] X. Huang, M.-Y. Liu, S. Belongie et J. Kautz, « Multimodal unsupervised image-to-image translation », in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. 179-196, isbn : 9783030012199. doi : 10.1007/978-3-030-01219-9_11.

[85] J. Kim, M. Kim, H. Kang et K. Lee, « U-GAT-IT : Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation », in *arXiv :1907.10830 [cs, eess]*, arXiv : 1907.10830, juill. 2019. adresse : http://arxiv.org/abs/1907.10830 (visité le 01/08/2019).

[86] Z. Yi, H. Zhang, P. Tan et M. Gong, « DualGAN : Unsupervised Dual Learning for Image-to-Image Translation », in *IEEE/CVF International Conference on Computer Vision (ICCV)*, oct. 2017. doi : 10.1109/iccv.2017.310.

[87] Y. Taigman, A. Polyak et L. Wolf, « Unsupervised Cross-Domain Image Generation », in *International Conference on Learning Representations (ICLR)*, ArXiv : 1611.02200, 2017.

[88] Y. Choi, Y. Uh, J. Yoo et J.-W. Ha, « StarGAN v2 : diverse image synthesis for multiple domains », in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2020, p. 8185-8194, isbn : 9781728171685. doi : 10.1109/CVPR42600.2020.00821. adresse : https://ieeexplore.ieee.org/document/9157662/ (visité le 02/12/2021).

[89] X. Huang et S. Belongie, « Arbitrary style transfer in real-time with adaptive instance normalization », in *IEEE/CVF International Conference on Computer Vision (ICCV)*, oct. 2017, p. 1510-1519, isbn : 9781538610329. doi : 10.1109/ICCV.2017.167. adresse : http://ieeexplore.ieee.org/document/8237429/ (visité le 02/12/2021).

[90] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser et I. Polosukhin, « Attention is All you Need », in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, éd., t. 30, Curran Associates, Inc., 2017.

[91] A. Sinha, J. Bai et K. Ramani, « Deep learning 3D shape surfaces using geometry images », in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, t. 9910 LNCS, 2016, p. 223-240, isbn : 9783319464657. doi : 10.1007/978-3-319-46466-4_14.

[92] V. F. Abrevaya, A. Boukhayma, S. Wuhrer et E. Boyer, « A decoupled 3D facial shape model by adversarial training », in *Proceedings of the IEEE International Conference on Computer Vision*, t. 2019-Octob, IEEE, 2019, p. 9418-9427, isbn : 9781728148038. doi : 10.1109/ICCV.2019.00951. arXiv : 1902.03619. adresse : https://hal.archives-ouvertes.fr/hal-02064711.

[93] S. Moschoglou, S. Ploumpis, M. Nicolaou, A. Papaioannou et S. Zafeiriou, « 3DFaceGAN : Adversarial Nets for 3D Face Representation, Generation, and Translation », in *International Journal of Computer Vision*, t. 128, nov. 2020. doi : 10.1007/s11263-020-01329-8.

[94] S. Gong, L. Chen, M. Bronstein et S. Zafeiriou, « SpiralNet++ : A Fast and Highly Efficient Mesh Convolution Operator », in *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, p. 4141-4148. doi : 10.1109/ICCVW.2019.00509.

[95] H. Maron, M. Galun, N. Aigerman, M. Trope, N. Dym, E. Yumer, V. G. Kim et Y. Lipman, « Convolutional neural networks on surfaces via seamless toric covers », in *ACM Transactions on Graphics (TOG)*, t. 36, juill. 2017, p. 1-10. doi : 10.1145/3072959.3073616.

[96] O. Litany, T. Remez, E. Rodola, A. Bronstein et M. Bronstein, « Deep Functional Maps : Structured Prediction for Dense Shape Correspondence », in *IEEE/CVF International Conference on Computer Vision (ICCV)*, oct. 2017. doi : 10.1109/iccv.2017.603.

[97] J. Bruna, W. Zaremba, A. Szlam et Y. LeCun, « Spectral networks and deep locally connected networks on graphs », *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014. arXiv : 1312.6203. adresse : http://arxiv.org/abs/1312.6203.

[98] M. Defferrard, X. Bresson et P. Vandergheynst, « Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering », in *arXiv :1606.09375 [cs, stat]*, arXiv : 1606.09375, juin 2016. adresse : http://arxiv.org/abs/1606.09375 (visité le 13/03/2019).

[99] T. N. Kipf et M. Welling, « Semi-Supervised Classification with Graph Convolutional Networks », in *International Conference on Learning Representations, (ICLR)*, 2017. adresse : https://openreview.net/forum?id=SJU4ayYgl (visité le 14/12/2021).

[100] Z. H. Jiang, Q. Wu, K. Chen et J. Zhang, « Disentangled representation learning for 3D face shape », in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, t. 2019-June, 2019, p. 11 949-11 958, isbn : 9781728132938. doi : 10.1109/CVPR.2019.01223. arXiv : 1902.09887. adresse : http://arxiv.org/abs/1902.09887.

[101] Z. Zhang, C. Yu, H. Li, J. Sun et F. Liu, « Learning Distribution Independent Latent Representation for 3D Face Disentanglement », in *Proceedings - 2020 International Conference on 3D Vision, 3DV 2020*, 2020, p. 848-857, isbn : 9781728181288. doi : 10.1109/3DV50981.2020.00095.

[102] A. Ranjan, T. Bolkart, S. Sanyal et M. J. Black, « Generating 3D faces using Convolutional Mesh Autoencoders », in *arXiv :1807.10267 [cs]*, arXiv : 1807.10267, juill. 2018. adresse : http://arxiv.org/abs/1807.10267 (visité le 06/02/2019).

[103] I. Lim, A. Dielen, M. Campen et L. Kobbelt, « A Simple Approach to Intrinsic Correspondence Learning on Unstructured 3D Meshes », *arXiv :1809.06664 [cs]*, 26 sept. 2018. arXiv : 1809.06664. adresse : http://arxiv.org/abs/1809.06664 (visité le 30/08/2021).

[104] S. Gong, L. Chen, M. Bronstein et S. Zafeiriou, « SpiralNet++ : A Fast and Highly Efficient Mesh Convolution Operator », en, *arXiv :1911.05856 [cs]*, nov. 2019. (visité le 10/02/2020).

[105] K. Yin, H. Huang, D. Cohen-Or et H. Zhang, « P2P-NET : bidirectional point displacement net for shape transform », in *ACM Transactions on Graphics (TOG)*, t. 37, juill. 2018, 152 :1-152 :13. DOI : 10.1145/3197517.3201288. adresse : https://doi.org/10.1145/3197517.3201288 (visité le 21/02/2020).

[106] L. Gao, J. Yang, Y.-L. Qiao, Y.-K. Lai, P. L. Rosin, W. Xu et S. Xia, « Automatic unpaired shape deformation transfer », in *ACM Transactions on Graphics (TOG)*, t. 37, déc. 2018, p. 1-15. DOI : 10.1145/3272127.3275028. adresse : http://dl.acm.org/citation.cfm?doid=3272127.3275028 (visité le 10/04/2019).

[107] K. Yin, Z. Chen, H. Huang, D. Cohen-Or et H. Zhang, « LOGAN : Unpaired Shape Transform in Latent Overcomplete Space », in *arXiv :1903.10170 [cs]*, arXiv : 1903.10170, mars 2019. adresse : http://arxiv.org/abs/1903.10170 (visité le 10/04/2019).

[108] C. R. Qi, H. Su, K. Mo et L. J. Guibas, « PointNet : Deep learning on point sets for 3D classification and segmentation », *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, t. 2017-Janua, p. 77-85, 2017. DOI : 10.1109/CVPR.2017.16. arXiv : 1612.00593. adresse : http://arxiv.org/abs/1612.00593.

[109] J. Xu, X. Sun, Z. Zhang, G. Zhao et J. Lin, « Understanding and improving layer normalization », *Advances in Neural Information Processing Systems*, t. 32, 2019, ISSN : 10495258. arXiv : 1911.07013. adresse : http://arxiv.org/abs/1911.07013.

[110] M. Segu, M. Grinvald, R. Siegwart et F. Tombari, « 3DSNet : Unsupervised Shape-to-Shape 3D Style Transfer », 2020. arXiv : 2011.13388. adresse : http://arxiv.org/abs/2011.13388.

[111] A. Odena, C. Olah et J. Shlens, « Conditional image synthesis with auxiliary classifier gans », in *34th International Conference on Machine Learning, ICML 2017*, t. 6, 2017, p. 4043-4055, ISBN : 9781510855144. arXiv : 1610.09585. adresse : http://arxiv.org/abs/1610.09585.

[112] S. Moschoglou, S. Ploumpis, M. A. Nicolaou, A. Papaioannou et S. Zafeiriou, « 3DFaceGAN : Adversarial Nets for 3D Face Representation, Generation, and Translation », *International Journal of Computer Vision*, t. 128, *10-11*, p. 2534-2551, 2020, ISSN : 15731405. DOI : 10.1007/s11263-020-01329-8. arXiv : 1905.00307. adresse : http://arxiv.org/abs/1905.00307.

[113] H. Chen, Y.-Q. Xu, H.-Y. Shum, S.-C. Zhu et N.-N. Zheng, « Example-based facial sketch generation with non-parametric sampling », in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2001. doi : 10.1109/iccv.2001.937657.

[114] L. Liang, H. Chen, Y.-Q. Xu et H.-Y. Shum, « Example-based caricature generation with exaggeration », in *Pacific Conference on Computer Graphics and Applications*, 2002. doi : 10.1109/pccga.2002.1167882.

[115] Y. Shi, D. Deb et A. K. Jain, « WarpGAN : Automatic Caricature Generation », in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2019. doi : 10.1109/cvpr.2019.01102.

[116] Z. Gu, C. Dong, J. Huo, W. Li et Y. Gao, « CariMe : Unpaired Caricature Generation with Multiple Exaggerations », in *IEEE Transactions on Multimedia*, 2021, p. 1-1. doi : 10.1109/TMM.2021.3086722.

[117] Q. Wu, J. Zhang, Y.-K. Lai, J. Zheng et J. Cai, « Alive Caricature from 2D to 3D », in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2018. doi : 10.1109/cvpr.2018.00766.

[118] H. Cai, Y. Guo, Z. Peng et J. Zhang, « Landmark Detection and 3D Face Reconstruction for Caricature using a Nonlinear Parametric Model », in *Graphical Models*, t. 115, avr. 2021, p. 101 103. doi : 10.1016/j.gmod.2021.101103.

[119] E. Games, *Paragon Assets*, 2018.

[120] T. M. team, *MakeHuman*, version 1.1.1, 2000–2019.

[121] A. Johnston, H. Hill et N. Carman, « Recognising Faces : Effects of Lighting Direction, Inversion, and Brightness Reversal », in *Perception*, t. 21, juin 1992, p. 365-375. doi : 10.1068/p210365. adresse : http://journals.sagepub.com/doi/10.1068/p210365 (visité le 21/02/2020).

[122] H. Hill et V. Bruce, « The effects of lighting on the perception of facial surfaces. », in *Journal of Experimental Psychology : Human Perception and Performance*, t. 22, 1996, p. 986-1004. doi : 10.1037/0096-1523.22.4.986. adresse : http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-1523.22.4.986 (visité le 21/02/2020).

[123] D. E. King, « Dlib-ml : A Machine Learning Toolkit », *Journal of Machine Learning Research*, t. 10, p. 1755-1758, 2009.

[124] F. N. N. Melissa Peskin, « Familiarity breeds attraction : Effects of exposure on the attractiveness of typical and distinctive faces », *Perception*, 2002.

[125] E. Akleman, J. Palmer et R. Logan, « Making extreme caricatures with a new interactive 2D deformation technique with simplicial complexes », in *In Proceedings of Visual*, 2000, p. 100-105. doi : 10.1.1.85.2369.

[126] H. Chen, N.-N. Zheng, L. Liang, Y. Li, Y.-Q. Xu et H.-Y. Shum, « PicToon », in *ACM Multimedia*, 2002. doi : 10.1145/641007.641040.

[127] B. Gooch, E. Reinhard et A. Gooch, « Human facial illustrations », in *ACM Transactions on Graphics (TOG)*, t. 23, jan. 2004, p. 27-44. doi : 10.1145/966131.966133.

[128] E. Zell, C. Aliaga, A. Jarabo, K. Zibrek, D. Gutierrez, R. McDonnell et M. Botsch, « To stylize or not to stylize ? the effect of shape and material stylization on the perception of computer-generated faces », in *ACM Transactions on Graphics (TOG)*, t. 34, oct. 2015, 184 :1-184 :12. doi : 10.1145/2816795.2818126. adresse : https://doi.org/10.1145/2816795.2818126 (visité le 02/12/2021).

[129] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah et D. Dunaway, « A 3D Morphable Model Learnt from 10, 000 Faces », in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2016. doi : 10.1109/cvpr.2016.598.

[130] R. W. Sumner et J. Popović, « Deformation transfer for triangle meshes », in *ACM SIGGRAPH*, 2004. doi : 10.1145/1186562.1015736.

[131] L. M. Mescheder, A. G. 0001 et S. Nowozin, « Which Training Methods for GANs do actually Converge ? », in *International Conference on Machine Learning (ICML)*, ArXiv : 1801.04406, t. 80, 2018, p. 3478-3487.

[132] M. Arjovsky, S. Chintala et L. Bottou, « Wasserstein Generative Adversarial Networks », in *International Conference on Machine Learning (ICML)*, ArXiv : 1701.07875, t. 70, 2017, p. 214-223.

[133] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim et J. Choo, « StarGAN : unified generative adversarial networks for multi-domain image-to-image translation », in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2018, p. 8789-8797, isbn : 9781538664209. doi : 10.1109/CVPR.2018.

00916. adresse : https://ieeexplore.ieee.org/document/8579014/ (visité le 02/12/2021).

[134] T. Kim, M. Cha, H. Kim, J. K. Lee et J. Kim, « Learning to Discover Cross-Domain Relations with Generative Adversarial Networks », in *International Conference on Machine Learning (ICML)*, 2017, p. 1857-1865. doi : 10.5555/3305381.3305573.

[135] D. P. Kingma et J. Ba, « Adam : A Method for Stochastic Optimization », in *International Conference on Learning Representations, (ICLR)*, ArXiv : 1412.6980, 2015. adresse : http://arxiv.org/abs/1412.6980.

[136] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah et D. Dunaway, « A 3D Morphable Model Learnt from 10,000 Faces », in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2016, p. 5543-5552, isbn : 978-1-4673-8851-1. doi : 10.1109/CVPR.2016.598. adresse : https://ieeexplore.ieee.org/document/7780967/ (visité le 15/07/2020).

[137] G. Stoet, « PsyToolkit : A software package for programming psychological experiments using Linux », in *Behavior Research Methods*, t. 42, nov. 2010, p. 1096-1104. doi : 10.3758/brm.42.4.1096.

[138] ——, « PsyToolkit : A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments », in *Teaching of Psychology*, t. 44, 2017, p. 24-31. doi : 10.1177/0098628316677643. eprint : https://doi.org/10.1177/0098628316677643. adresse : https://doi.org/10.1177/0098628316677643.

[139] A. C. Little, B. C. Jones et L. M. Debruine, « The many faces of research on face perception », *Philosophical Transactions of the Royal Society B : Biological Sciences*, t. 366, *1571*, p. 1634-1637, 2011, issn : 09628436. doi : 10.1098/rstb.2010.0386. adresse : https://royalsocietypublishing.org/doi/10.1098/rstb.2010.0386.

[140] N. Olivier, L. Hoyet, F. Danieau, F. Argelaguet, Q. Avril, A. Lecuyer, P. Guillotel et F. Multon, « The impact of stylization on face recognition », in *Proceedings - SAP 2020 : ACM Symposium on Applied Perception*, ACM, 2020, p. 1-9, isbn : 9781450376181. doi : 10.1145/3385955.3407930. adresse : https://dl.acm.org/doi/10.1145/3385955.3407930.

[141] M. Garland et P. S. Heckbert, « Surface simplification using quadric error metrics », in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997*, ACM Press, 1997, p. 209-216, isbn : 0897918967. doi : 10.1145/258734.258849.

[142] X. Huang et S. Belongie, « Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization », in *Proceedings of the IEEE International Conference on Computer Vision*, t. 2017-Octob, IEEE, 2017, p. 1510-1519, isbn : 9781538610329. doi : 10.1109/ICCV.2017.167. arXiv : 1703.06868. adresse : http://ieeexplore.ieee.org/document/8237429/.

[143] L. Hui, X. Li, J. Chen, H. He et J. Yang, « Unsupervised Multi-Domain Image Translation with Domain-Specific Encoders/Decoders », in *Proceedings - International Conference on Pattern Recognition*, t. 2018-Augus, 2018, p. 2044-2049, isbn : 9781538637883. doi : 10.1109/ICPR.2018.8545169. arXiv : 1712.02050.

[144] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang et A. A. Efros, « Learning dense correspondence via 3D-guided cycle consistency », *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, t. 2016-Decem, p. 117-126, 2016, issn : 10636919. doi : 10.1109/CVPR.2016.20. arXiv : 1604.05383. adresse : http://arxiv.org/abs/1604.05383.

[145] J. Y. Zhu, T. Park, P. Isola et A. A. Efros, « Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks », *Proceedings of the IEEE International Conference on Computer Vision*, t. 2017-Octob, p. 2242-2251, 2017, issn : 15505499. doi : 10.1109/ICCV.2017.244. arXiv : 1703.10593. adresse : http://arxiv.org/abs/1703.10593.

[146] M. Desbrun, M. Meyer, P. Schröder et A. H. Barr, « Implicit fairing of irregular meshes using diffusion and curvature flow », in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999*, ACM Press, 1999, p. 317-324, isbn : 0201485605. doi : 10.1145/311535.311576. adresse : http://portal.acm.org/citation.cfm?doid=311535.311576.

[147] A. Nealen, T. Igarashi, O. Sorkine et M. Alexa, « Laplacian mesh optimization », *Proceedings - GRAPHITE 2006 : 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, p. 381-389, 2006. doi : 10.1145/1174429.1174494. adresse : http://portal.acm.org/

citation.cfm?doid=1174429.1174494http://dl.acm.org/citation.cfm?id=1174494.

[148] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang et X. Cao, « FaceScape : A large-scale high quality 3D face dataset and detailed riggable 3D face prediction », *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 598-607, 2020, ISSN : 10636919. DOI : 10.1109/CVPR42600.2020.00068. arXiv : 2003.13989.

[149] A. Kacem, K. Cherenkova et D. Aouada, « Disentangled Face Identity Representations for joint 3D Face Recognition and Expression Neutralisation », 2021. arXiv : 2104.10273. adresse : http://arxiv.org/abs/2104.10273.

[150] D. P. Kingma et J. L. Ba, « Adam : A method for stochastic optimization », *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. arXiv : 1412.6980.

[151] K. He, X. Zhang, S. Ren et J. Sun, « Delving deep into rectifiers : Surpassing human-level performance on imagenet classification », in *Proceedings of the IEEE International Conference on Computer Vision*, t. 2015 Inter, IEEE, 2015, p. 1026-1034, ISBN : 9781467383912. DOI : 10.1109/ICCV.2015.123. arXiv : 1502.01852. adresse : http://ieeexplore.ieee.org/document/7410480/.

[152] P. Chandran, D. Bradley, M. Gross et T. Beeler, « Semantic Deep Face Models », *Proceedings - 2020 International Conference on 3D Vision, 3DV 2020*, p. 345-354, 2020. DOI : 10.1109/3DV50981.2020.00044.

[153] D. P. Kingma et M. Welling, « Auto-encoding variational bayes », in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014. arXiv : 1312.6114.

[154] C. Ennis, L. Hoyet, A. Egges et R. McDonnell, « Emotion Capture : Emotionally Expressive Characters for Games », in *Proceedings of Motion on Games*, ACM, nov. 2013. DOI : 10.1145/2522628.2522633.

[155] S. Wu, C. Rupprecht et A. Vedaldi, « Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild », in *arXiv :1911.11130 [cs]*, arXiv : 1911.11130, mars 2020. adresse : http://arxiv.org/abs/1911.11130 (visité le 09/11/2020).

[156] Y. Shi, D. Aggarwal et A. K. Jain, « Lifting 2D StyleGAN for 3D-Aware Face Generation », in *arXiv :2011.13126 [cs]*, arXiv : 2011.13126, nov. 2020. adresse : http://arxiv.org/abs/2011.13126 (visité le 07/12/2020).

[157] J. Gu, L. Liu, P. Wang et C. Theobalt, *StyleNeRF : A Style-based 3D-Aware Generator for High-resolution Image Synthesis*, 2021. arXiv : 2110.08985 [cs.CV].

[158] D. A. Hudson et C. L. Zitnick, « Generative Adversarial Transformers », *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 2021.

[159] P. Zhou, L. Xie, B. Ni et Q. Tian, « CIPS-3D : A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis », 2021. arXiv : 2110.09788.

[160] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang et N. Duan, *NÜWA : Visual Synthesis Pre-training for Neural visUal World creAtion*, 2021. arXiv : 2111.12417 [cs.CV].

[161] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, M. Johnson, V. Estellers, T. J. Cashman et J. Shotton, *Fake It Till You Make It : Face analysis in the wild using synthetic data alone*, 2021. arXiv : 2109.15102 [cs.CV].

[162] D. A. Clevert, T. Unterthiner et S. Hochreiter, « Fast and accurate deep network learning by exponential linear units (ELUs) », in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016. arXiv : 1511.07289.

TABLE DES FIGURES

Liste des tableaux

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Personnalisation adaptative d'avatar

Mot clés : Transfert de style, personnage virtuel, apprentissage profond, perception, infographie

Résumé : La notion d'apparence préférentielle est ancienne, émergeant très tôt avec les interactions entre organismes, que ce soit pour le camouflage, la tromperie ou la compétition sexuelle [1], [2]. Comme d'autres caractéristiques évolutives de notre espèce, cette notion a acquis de nouvelles significations avec le développement de la culture [3], pour identifier les genres, les castes, les fonctions officielles, et pour divers aspects cérémoniels [4]. Au cours des dernières décennies, une pratique qui avait été principalement limitée à l'utilisation de vêtements, de masques ou de peintures corporelles, a également connu les profonds changements apportés par les ordinateurs. D'abord avec l'utilisation de l'infographie dans le contenu visuel, puis avec l'essor des jeux basés sur l'incarnation sur les ordinateurs personnels, qui ont propulsé l'incarnation et l'utilisation de personnages virtuels à des niveaux jamais atteints.

Dans cette thèse, nous cherchons à réaliser la stylisation des personnages de manière entièrement automatique.

Qu'ils soient stylisés ou non, la présence de doubles numériques dans le contenu publié a connu une croissance significative ces dernières décennies, qu'il s'agisse de films (par exemple Avatar - 2009, Terminator 2 - 1991), séries télévisées (par exemple The Boys - 2019), ou jeux (par exemple Cyberpunk2077 - 2020).

Aavoir un double numérique complet est un concept fortement poussé en avant par des produits commerciaux comme le Metaverse. Il y a un intérêt à disposer d'un personnage stylisé ressemblant à une personne de référence, tout en conservant certaines caractéristiques identitaires essentielles, afin qu'il soit reconnaissable. Cela permet de ne pas perdre la valeur de marque d'un acteur, ou d'augmenter le sentiment d'incarnation, l'intimité avec son double virtuel, et l'interactivité avec les autres dans des environnements partagés.

Dans cette thèse, nous visons un niveau de réalisme proche de celui des jeux vidéo réalistes, afin de fournir une base de référence solide qui pourrait être utilisée pour un jeu ou une application VR. Elle pourrais également être affinée et travaillée par des artistes pour être utilisée dans un contexte nécessitant un photoréalisme plus élevé.

L'utilisation de la technologie basée sur les réseaux neuronaux a connu une croissance considérable au cours de ces dernières années. Dans le monde artistique, elle a été utilisée à la fois comme un outil de soutien à la créativité et comme un moyen de créer de l'art en soi, grâce à l'utilisation de réseaux génératifs. Les réseaux neuronaux ont été largement exploités et conçus pour la stylisation du contenu [14], notamment dans des apps (Prisma, FaceApp).

La stylisation – ou le transfert de style – peut être caractérisée comme le mapping d'une distribution (par exemple, une fonction donnant toutes les valeurs possibles d'une donnée, avec leur fréquence d'occurrence) vers une autre, où les deux distributions partagent une partie de leur information. Dans la littérature, le transfert de style a été principalement modélisé comme la séparation des données

en contenu (ce qui est partagé entre les distributions, l'identité dans le cas des visages), et en style (tout ce qui est spécifique au domaine).

Cependant, il n'existe pas de théorie bien formée sur la distinction du style et du contenu, la séparation se faisant de manière empirique, par le biais d'heuristiques [14] ou de fonctions apprises de manière supervisée ou non supervisée. Le style et le contenu, dans le contexte de l'apparence des personnages, sont difficiles à définir et restent largement subjectifs.

Il n'existe aucune approche de ce type permettant de transférer des styles arbitraires (par exemple : extraterrestres, orcs, dessins animés) à un visage humain à partir de quelques exemples, avec un résultat reflétant l'identité et le style souhaités. Une telle méthode serait souhaitable, car il existe généralement très peu d'exemples (souvent un seul) pour un style donné, surtout dans le domaine de la 3D. Dans le contexte du transfert d'expression, les approches existantes ne prennent pas en compte la morphologie et l'identité du visage, ou ne prennent pas de paramètres d'entrées, opérant uniquement dans des espaces latents. Comme la forme du visage peut varier considérablement entre les visages stylisés, nous avons besoin d'une méthode qui tienne compte de ces deux éléments.

Enfin, il n'existe aucune recherche sur l'étude de la relation entre la stylisation et la préservation de l'identité, un facteur que nous pensons être clé pour la paramétrisation et la conception des méthodes de stylisation.

Dans cette thèse, nous avons exploré le sujet de la stylisation des personnages virtuels. Nos objectifs étaient d'aborder le problème de la stylisation automatique des personnages dans le contexte d'une disponibilité élevée ou faible des données, ainsi que le transfert d'expression entre des visages aux morphologies variées. Nous avons présenté une nouvelle méthode pour deux contextes de stylisation de personnages ainsi que des études d'utilisateurs afin d'examiner leurs performances et leurs limites. Nous avons également proposé une nouvelle méthode pour le transfert d'expression entre visages en tenant compte de la morphologie du visage, qui surpasse les méthodes existantes sur plusieurs métriques.

Dans notre première contribution, nous avons exploré le lien entre la stylisation et la reconnaissance du visage. Cette question est importante, car dans de nombreux cas, un utilisateur peut souhaiter styliser son visage, tout en restant reconnaissable. Nous avons donc proposé une nouvelle méthode de stylisation, dans laquelle nous avons décomposé la représentation du visage en un maillage classique et une texture, et nous nous sommes concentrés sur chacun de ces éléments individuellement. La géométrie est modifiée avec une décomposition en style et identité, un visage stylisé de référence étant utilisé comme style, et l'identité de la géométrie de la personne lui étant appliquée, en calculant sa différence euclidienne par rapport à un visage humain moyen. Pour la stylisation de la texture, nous avons utilisé une approche de même nature relative, mais cette fois-ci en modifiant la texture du style par l'optimisation des caractéristiques d'un réseau généraliste, inspiré des travaux de Gatys [14]. Nous avons ensuite exploité cette méthode pour mener une étude utilisateur, afin de mesurer l'impact du niveau de stylisation d'un visage humain sur la reconnaissabilité de la personne, et l'acceptabilité du visage de sortie comme "stylisé". Nous avons posé deux hypothèses principales, la première étant que plus un visage est stylisé, moins la personne est reconnaissable, la seconde étant que les visages sont plus facilement acceptés comme "stylisés" si la stylization est forte. Ces deux hypothèses ont été vérifiées par nos résultats expérimentaux, les deux taux variant de manière inverse et quasi linéaire. Le taux de reconnaissance (sur quatre choix) variait de 95% au maximum à 65% au minimum, tandis que les taux d'acceptation du style variaient de 10% à 100%. Aucun niveau de stylisation ne maximise les deux mesures, les deux niveaux de stylisation avec les meilleurs compromis correspon-

dant à 75%, 92%, et 80%, 65% respectivement pour l'identité et le style.

Les méthodes basées sur des règles se sont avérées empiriquement moins performantes dans plusieurs domaines que les approches basées sur l'apprentissage, notamment dans le domaine de la stylisation de contenu 2D. Dans le but de construire une méthode de stylisation plus puissante, nous nous sommes inspirés d'approches dans le domaine 2D et avons développé un réseau neuronal de traduction de domaine basé sur le GAN, visant à apprendre à séparer le style et l'identité d'un large ensemble de visages 3D non appariés. Comme les méthodes basées sur l'apprentissage nécessitent une quantité importante de données, le travail du chapitre 4 s'est concentré sur les caricatures, le seul domaine de style pour lequel une telle quantité de données était disponible, en exploitant un ensemble de données récemment publié. Nous avons ensuite mené une expérience perceptive pour mesurer la performance de notre approche, nous l'avons comparée à trois méthodes de l'état de l'art basées sur des règles au cours d'une étude utilisateur en deux parties visant à classer chaque approche. Dans la première partie de l'étude, nous avons constaté qu'il n'y avait pratiquement aucune différence de classement entre les quatre méthodes, les classements moyens étant très similaires, et seule une méthode a été jugée significativement meilleure que les autres. En examinant les classements par visage utilisés dans l'étude, nous avons observé une très grande variation du classement de chaque méthode par visage, certaines étant classées considérablement plus haut - ou plus bas - pour certains visages. Nous avons également observé des cas où les visages produits par une méthode étaient à la fois classés en premier et en dernier, en considérant toutes les réponses des participants. Ces observations soulignent la grande subjectivité d'une tâche de stylisation, peut-être encore plus dans le cas des caricatures, car il peut exister de nombreux styles de caricatures valides. À partir de ces résultats empiriques, nous avons émis plusieurs lignes directrices pour le choix d'une méthode de caricaturisation, en fonction des données disponibles, de la tolérance sur la variance de la qualité subjective, et de la spécificité du style souhaité.

Dans les deux paragraphes précédents, nous avons présenté deux méthodes permettant d'appliquer une stylisation à des visages neutres fixes, mais les personnages virtuels sont généralement interactifs et doivent être capables d'exprimer toute une gamme d'émotions et d'expressions faciales. Les méthodes traditionnelles de transfert d'expression, comme le transfert de blenshape, échouent dans les cas où les morphologies diffèrent de manière significative. Les approches plus avancées de séparation de l'identité et de l'expression présentent diverses limites, comme le fait de ne travailler que dans un espace latent et de ne pas adapter l'expression à la morphologie. Nous avons introduit une nouvelle approche sans ces limites, pour ouvrir la voie au transfert d'expression pour les visages stylisés, en s'appuyant sur l'approche que nous avons utilisée pour les caricatures. De même, notre méthode est basée sur l'entraînement semi-supervisé (ne nécessitant que des étiquettes d'expression) d'un autoencodeur guidé par un GAN. Nous avons amélioré la préservation de l'identité et de l'expression, et l'avons comparée à l'état de l'art en matière de transfert d'expression. Nous obtenons de meilleures performances en matière d'erreur de reconstruction, de décomposition de l'identité et de neutralisation.

Il existe de nombreuses façons d'étendre le travail présenté sur la stylisation des personnages virtuels.

Par exemple, nous avons proposé une approche basée sur l'apprentissage profond pour le transfert de style facial en maillage 3D, que nous avons d'abord appliquée aux caricatures. Cette méthode, bien qu'elle ne fonctionne que dans un cadre de classe, a la limitation de ne pas être capable de s'adapter à un style non connu, car son espace de style est réduit à seulement deux modes, humain et caricature. L'une des premières perspec-

tives à court terme pour augmenter la valeur d'une telle approche serait d'inclure des propriétés permettant une correspondance de un à plusieurs, afin de permettre au visage stylisé d'adopter le style spécifique d'une caricature donnée. Ce type de capacités a été démontré dans le domaine de l'image : [13].

Du point de vue perceptif, il reste de nombreuses questions liées à la perception et à la reconnaissance des personnages virtuels stylisés. Par exemple, nous avons montré les exigences d'une stylisation modérée si le personnage doit rester facilement reconnaissable, en mesurant la relation entre la force du style et la reconnaissabilité d'une manière globale. Au lieu de considérer son impact global, il pourrait être intéressant de se concentrer sur les caractéristiques faciales individuelles, certaines parties de l'apparence humaine étant peut-être plus tolérantes à la stylisation que d'autres. Cela nous aiderait à comprendre comment effectuer une stylisation partielle sur le plan perceptif (qui s'est avérée nécessaire dans la plupart des cas) pour se concentrer uniquement sur certaines caractéristiques (par exemple : les yeux, le nez, la texture, etc.), ce qui permettrait une meilleure reconnaissance pour le même niveau de style perçu.

Les données 3D restant extrêmement rares, ce qui limite fortement les possibilités d'apprentissage d'une notion particulière de style, les travaux futurs pourraient consister à se concentrer plutôt sur les données 2D, puis à les transposer dans l'espace 3D. La question se pose alors de savoir comment exploiter les données 2D afin de générer un contenu 3D stylisé. Une réponse pourrait être une méthode en deux étapes, où une image faciale est d'abord stylisée dans le domaine 2D, puis une version 3D du résultat est déduite. Dans ce cas, l'inférence de contenu 3D à partir de données 2D pourrait s'inspirer de la littérature sur l'inférence de la forme du visage humain en 3D à partir de la 2D, et utiliserait un moteur de rendu différentiable [155], [156]. Le réseau génératif pourrait sinon directement intégrer la notion de pose de caméra, générant ses données en la prenant en compte [157].

Les travaux sur le transfert des expressions faciales en fonction de la morphologie n'ont pas été appliqués à des visages stylisés tels que les orcs ou les extraterrestres. Il n'y a donc aucune certitude sur la façon dont des expressions telles que le sourire ou le froncement des sourcils peuvent être transférées et perçues.

Il reste de nombreuses questions à résoudre sur le sujet de la stylisation, l'une d'entre elles étant de savoir comment styliser l'ensemble du corps. Nous avons présenté dans la dernière section les moyens de produire des visages stylisés en 3D à partir de données 2D uniquement. Les images du corps entier pourraient être exploitées de la même manière, bien que cela s'avère plus difficile, car les réseaux génératifs existants s'avèrent trop limités pour capturer leurs importantes variations spatiales. En effet, ils ont montré des performances remarquables dans certains contextes, comme les visages, mais leurs capacités à modéliser des données à plus forte variation spatiale se sont révélées plus faibles.

Pour terminer cette thèse sur une note plus générale, les personnages virtuels sont devenus omniprésents dans le domaine du divertissement et ne cessent de gagner du terrain dans les applications à vocation plus sociale. La capacité de représenter et de styliser son propre personnage virtuel promet d'être la clé d'un grand nombre de mondes interactifs virtuels. La solution à court terme pour une telle stylisation réside dans des méthodes basées sur l'apprentissage exploitant des données 3D, ou même des approches basées sur des règles. A plus long terme, nous pensons que la capacité d'interpréter des données 2D d'une manière consciente de la 3D sera la clef, ces dernières étant beaucoup plus nombreuses et plus faciles à produire. À l'heure actuelle, nous espérons que les solutions proposées dans cette thèse seront utiles pour permettre la génération automatique de personnages stylisés expressifs, ainsi que pour montrer leurs limites, et qu'elles donneront des indications en vue d'une extension à l'ensemble du corps, et à partir de contenu 2D.

Title: Adaptive Avatar Customization

Keywords: Style Transfer, Virtual Character, Deep Learning, Perception, Computer Graphics

Abstract: The notion of preferential appearance is ancient, emerging early with inter-organism interactions, whether for camouflage, deception, or sexual competition [1], [2]. As other evolutionary features of our species, this notion has gained new meanings with the growth of group specific culture [3], for identifying genders, castes, official functions, and for various ceremonial aspects [4]. During the last decades, a practice that had been mainly restricted to the use of clothes, masks, or body paintings, has also known the deep changes brought by computers. It was brought first by computer graphics being used in visual content, and then by the rise of embodiment-based games on personal computers, skyrocketing the use and cultural normality of embodiment and third-person control of characters to heights never seen.

In this thesis, we aim to perform character stylization in a fully automatic manner.

Personalizable characters have been present in games for several decades, allowing users to stylize (alter) the appearance of their faces, bodies, or clothes of the character they incarnate or interact with. In the case of movies, with only tens of actors compared to thousands or millions of gamers, personalizing characters does not require automation and can be seen instead as a cost saving factor, since realistic hand crafted personalized embodiments are limited to a few per high budget movies (e.g. The Lord of the Rings – 2001, Avatar – 2009 1.1). Stylized or not, digital doubles presence in published content has grown significantly these last decades, from movies (e.g. Terminator 2 – 1991), to TV series (e.g. The Boys – 2019), and games (e.g. Cyberpunk2077 – 2020).

The concept of having a whole digital double is a future strongly pushed forward by commercial products such as the Metaverse. As modern day virtual embodiments grow closer and closer to ourselves, and as we reach times where entirely virtual characters can look realistic on screen, there is a growing interest in being able to personalize your own embodiment. This means that there is a need of a stylized character resembling a reference person, while conserving some core identity features, to let them be recognizable. This allows to not lose the brand value of an actor, or increase the sentiment of embodiment, intimacy as one's virtual double, and interactivity with others in shared environments.

Virtual characters can have various degrees of realism, ranging from the simple cartoon characters of messaging apps to photorealist humanoids shown in movies (e.g. The Lord of the Ring (2001-2003), Avatar (2009)). In this thesis, we target a level of realism close to realist video games, aiming to provide a strong baseline which could be used for a game or VR application, or fine tuned and worked on by artists to be used in a context requiring higher photorealism.

The use of neural network based technology has known considerable growth during these last few years. In the artistic world it has been used both as a tool to support creativity, and as a way towards art creation by itself, through the use of generative networks. Neural networks have been heavily leveraged and designed for content stylization [14].

Stylization – or style transfer – can be characterized as the mapping of one distribution (as in, a function giving all possible values of some data, with their occurrence frequency) towards another, where both distributions share part of their information. In the literature, style trans-

fer has been mostly modeled as the separation of data into content (what is shared between distributions, identity in the case of faces), and style (everything that is domain specific).

However, no well formed theory although exist on the distinction of style and content, separation being done empirically, through heuristics [14] or functions learned in a supervised or unsupervised manner. Both style and content, in the context of character appearance, are difficult to define, and remain largely subjective.

There exist no such approach showcasing capabilities for transferring arbitrary (e.g. alien, orc, cartoon) styles to a human face from a few examples, with an output reflecting the desired identity and style. Such a method would be desirable, as there usually exist very few examples (often only one) for any given style, especially in the 3D domain. In the context of expression transfer, existing approaches either do not take the morphology and identity of the face into account, or do not take inputs, operating only in latent spaces. As facial shape can vary significantly between stylized faces, we require a method addressing both things.

Finally, there exists no research on the study of the relation between stylization and identity preservation, a factor we believe to be key to the parametrization and design of stylization methods.

In this thesis, we explored the topic of virtual character stylization. Our goals were to address the problem of automatic character stylization in the context of both high and low data availability, as well as expression transfer between faces with varied morphologies. We presented a novel method for both character stylization context along with users studies to examine their performance and limitations, and proposed a novel method for facial morphology aware expression transfer, that beat existing methods on several metrics.

In Chapter 2, we explored the link between facial stylization and recognition. This question is important, as in many cases a user may wish to stylize their face, while remaining recognizable. We thus proposed a novel stylization method, in which we decomposed the facial representation into the classical mesh plus texture, and focused on each of these individually. The geometry is deformed with a style and identity decomposition, a reference stylized face being used as the style, and the identity of the person's geometry being applied to it, by computing its euclidian difference to an average human face. For texture stylization, we used an approach of a similar relative nature, but this time modify the style texture through optimization of a generalist network's features, inspired by the work of Gatys [14]. We then leveraged this method to conduct a user study, in order to measure the impact of the level of stylization of a human face on the recognisability of the person, and the acceptability of the output face as "stylized". We laid two main hypothesis, the first being that the most a face is stylized, the less the person is recognizable, the second being that faces are accepted as "stylized" in inverse relation to it. Both hypothesis were verified through our experimental results, both rates varying in an inverse near linear fashion. Recognition rate (out of four choices) varied from 95% at the highest to 65% at the lowest, while style acceptance rates varied from 10% to 100%. No style level although maximized both metrics, the two style levels maximizing both corresponding to 75%, 92%, and 80%, 65% respectively for identity and style.

Rule based methods have empirically proven to perform more weakly in several fields compared to learning based approaches, notably in the domain of the stylization of 2D content. Looking to build a stronger stylization method, we took inspiration from approaches in the 2D domain and developed a GAN based domain translation neural network, aiming to learn to separate style and identity from a large set of unpaired 3D faces. As learning based methods require a significant quantity of data, the work of Chapter 4 focused on caricatures, the only style domain where such

quantity of data were available, leveraging a recently published dataset. We then conducted a perceptual experiment to measure the performance of our approach, we compared it against three rule based state of the art methods during a two part user study aiming to rank each approach. From the first part of the study, we found nearly no difference in ranking between the four methods, average rankings being highly similar, and only one method being measured significantly better than others. Looking at rankings per face used in the study, we observed a very large variation of each method's ranking per face, some being ranked considerably higher — or lower — for certain faces. We also observed cases where the faces produced by some method where both ranked most first, and last, considering all participants answers. These observations highlight the high subjectivity of a stylization task, perhaps even more in the case of caricatures, as there can exist numerous valid caricature styles. From these empirical results, we issued several guidelines for choosing a caricaturisation methods, depending on the available data, the tolerance on the variance in subjective quality, and the specificity of the desired style.

In the two previous contributions, we introduced two methods to apply a stylization to fixed neutral faces, but virtual characters are typically interactive, and need to be able to express a range of facial emotions and expressions. Traditional expression transfer methods like blenshape transfer fail in cases where morphologies differ significantly. More advanced approaches for identity and expression separation have various limitations, such as working only in a latent space, and not adapting the expression to the morphology. We introduced a novel approach without these limitations, to pave the way towards expression transfer for stylized faces, building on the approach we used for caricatures. Likewise, our method is based on semi-supervised training (requiring only expression labels) of an autoencoder guided by a GAN. We improve both identity and expression preservation, and compare it with the state of the art in expression transfer, obtaining better reconstruction error, identity decomposition, and neutralization performance.

In this thesis, we have explored multiple approaches addressing the accurate stylization of virtual characters, as well as a method for morphology-aware expression transfer. However, there are numerous ways in which this work on virtual characters stylization could be extended.

In this thesis, we proposed a deep learning based approach for 3D mesh facial style transfer, that we first applied to caricatures. This method although only works in a class wise setting, and has the limitation of not being able to adapt to unseen style, as its style space is collapsed to only two modes, human and caricature. One of the first short-term perspectives to increase the value of such an approach would be to include properties allowing a one-to-many mapping instead, in order to allow the stylized face to take the specific kind of style of a given caricature. This kind of capabilities have been shown in the image domain [13].

On the perceptual side, there remains many questions related to the perception and recognisability of stylized virtual characters. For instance we have shown the requirements for a moderate stylization if the character needs to remain easily recognizable, measuring the relation between style strength and recognisability in a global manner. Instead of considering its global impact, it could be interesting to instead focus on individual facial features, some parts of the human appearance perhaps being more tolerant to stylization than others. This would help us to understand how to perceptually perform partial stylization (which has shown itself to be necessary in most cases) to focus only on some features (e.g. the eyes, nose, texture, etc.), allowing better recognition for the same perceived style level.

As handcrafted 3D data remain extremely scarce, heavily limiting possibilities of learning a particular notion of style, future work could include focusing on 2D data instead, and then

mapping it to the 3D space. This brings the question of how to leverage 2D data in order to generate stylized 3D content. An answer could be a two step method, where a facial image is first stylized in the 2D domain, and then a 3D version of the result is inferred. Inferring 3D content from 2D data in this case could for instance be inspired from the literature on human 3D facial shape inference from 2D, and use a differentiable renderer [155], [156], which is a renderer that work using only differentiable operations, allowing to compute gradient, and thus to use it for deep learning. The generative network could also directly learn the notion of camera pose by generating its data using camera poses [157].

While most research on facial stylization focus on neutral faces, expressions are a crucial aspect of a face and its perception. Despite work on morphology aware facial expression transfer, it has not been applied on stylized faces such as orcs, or aliens, hence there is no certainty on how well expressions such as smiling or frowning could be transferred, and perceived.

There remain many questions to answer on the topic of stylization, one significant being how to stylize the entirety of the body. While facial stylization is certainly the most important aspect to focus on, it is not the only one. We presented in the last section ways in which 3D stylized faces could produced from only 2D data. Full body images could be leveraged the same way, although it would prove more difficult, as full body data is less numerous than facial data, and existing generative networks prove too limited to capture its important spatial variations. Indeed, they have showcased remarkable performance in some contexts, such as faces, but their capabilities to model data with higher spatial variation, have show themselves to be lower.

To end this thesis on a more general note, virtual characters have become ubiquitous in entertainment and are steadily gaining ground in more socially oriented applications. The ability to represent and stylize one's own virtual character is promised to be key to a large amount of virtual interactive worlds, and while the combination of existing capture approaches combined with the development of affordable acquisition devices (smartphones) already address most of the first point, character stylization methods remain far and between. While the short term solution to such stylization is in learning-based methods leveraging 3D data, or even rule-based approaches, in the longer run we believe that they ability to interpret 2D data in a 3D aware manner to be key, the latter being vastly more numerous and easier to produce. At present, we hope that the solutions proposed in this thesis will be useful to allow the automatic generation of expressive stylized characters, and their limits, and will give insights towards an extension of it to the whole body, and from 2D content.