Morphologically Plausible Deformation Transfer

Jean Basset

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Jean BASSET

Thèse dirigée par **Edmond BOYER**
et codirigée par **Franck MULTON**,Professeur des universités, Université Rennes 2
et co-encadrée par **Stefanie WUHRER**,CR, INRIA

préparée au sein du **Laboratoire Laboratoire Jean Kuntzmann**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Transfert de déformation morphologiquement plausible

Morphologically Plausible Deformation Transfer

Thèse soutenue publiquement le **2 juin 2022**,
devant le jury composé de :

Monsieur TAKU KOMURA
Professeur, The University of Hong Kong, Rapporteur
Monsieur PIERRE HELLIER
Ingénieur HDR, INTERDIGITAL, Rapporteur
Madame CELINE LOSCOS
Professeur des Universités, UNIVERSITE DE REIMS - CHAMPAGNE ARDENNE, Examinatrice
Monsieur ADNANE BOUKHAYMA
Chargé de recherche, INRIA CENTRE RENNES-BRETAGNE ATLANTIQUE, Examinateur
Monsieur JOCELYN CHANUSSOT
Professeur des Universités, GRENOBLE INP, Président

Abstract

With the advances of 3D content generation and capture, and the recent popularity of immersive virtual environments, producing realistic animations of 3D virtual characters has seen an increasing demand. More specifically, automatically applying existing animations on new characters with different body shapes would represent a significant gain of both time and resources for animators. Traditional methods transfer the pose in each frame of the animation to the new character. However, this implies being able to define what makes two poses equivalent. This is not straightforward as poses tend to change depending on the morphology of the character performing them, and as their meaning is highly contextual. In this manuscript, we propose new approaches that transform the identity of a character into a new identity without modifying the character's pose, which does not require defining pose equivalences. When changing the identity of a character, some artifacts may appear, such as interpenetration or loss of self-contacts between body surfaces, *e.g.* the hands touching in a clapping pose. We study how to adapt our methods to correct these artifacts.

We first propose a method that iteratively morphs the identity of a source character in a specific pose to match the identity of a target character. This method allows to naturally mimic the pose of the source character in our results, as the optimization directly starts from the wanted pose. In this method we only apply simple pose corrections in order to preserve self-contacts present in the source and avoid interpenetrations, which allows the method to adapt the results to extreme morphologies.

We then present a deep encoder-decoder architecture that learns from data to predict the identity deformation of a character to match a target character's identity, without changing the pose. We propose self-supervised identity losses, that allow an inference time fine tuning step enabling transfer to identities far from the training set, such as casually clothed humans. Our model generalizes well to complex unseen poses.

Finally, we study the impact of self-contacts between body surfaces on perceived pose equivalences. Indeed, we observed that some self-contacts were only present because of the morphology of the character and were not important to the pose. Preserving all self-contacts in our first method could therefore create artifacts in some cases where unnecessary self-contacts were preserved, significantly modifying the pose. We conduct a study where we present to observers two models of a character mimicking the pose of a source character, one with the same self-contacts as the source, and one with one self-contact removed. We ask observers to select which model best mimics the source pose. We show that poses with different self-contacts

are considered different by observers in most cases, and that this effect is stronger for self-contacts involving the hand than for those involving the arms.

Résumé

Avec les progrès de la génération et de la capture de contenu 3D, et la popularité récente des environnements virtuels immersifs, la production d'animations réalistes de personnages virtuels 3D a connu une demande croissante. En particulier, l'application automatique d'animations existantes sur de nouveaux personnages aux morphologies différentes représenterait un gain de temps et de ressources important pour les animateurs. Les méthodes traditionnelles transfèrent la pose de chaque image de l'animation au nouveau personnage. Toutefois, cela implique de pouvoir définir ce qui rend deux poses équivalentes. Ceci n'est pas simple car les poses ont tendance à changer en fonction de la morphologie du personnage qui les exécute, et leur signification est hautement contextuelle. Dans ce manuscrit, nous proposons de nouvelles approches qui transforment l'identité d'un personnage vers une nouvelle identité sans modifier la pose du personnage, ce qui ne nécessite pas de définir des équivalences de pose. Lors du changement d'identité d'un personnage, certains artefacts peuvent apparaître, comme des collisions ou la perte d'auto-contacts entre les surfaces du corps, comme les mains se touchant dans une pose d'applaudissement. Nous étudions comment adapter nos méthodes pour corriger ces artefacts.

Nous proposons d'abord une méthode qui transforme de manière itérative l'identité d'un personnage source dans une pose spécifique pour la faire correspondre à l'identité d'un personnage cible. Cette méthode permet d'imiter naturellement la pose du personnage source dans nos résultats, puisque l'optimisation part directement de la pose voulue. Dans cette méthode, nous n'appliquons que des corrections de pose simples afin de préserver les auto-contacts présents dans la source et d'éviter les collisions, ce qui permet à la méthode d'adapter les résultats à des morphologies extrêmes.

Nous présentons ensuite une architecture d'encodeur-décodeur profonds qui apprend à partir de données à prédire la déformation de l'identité d'un personnage pour correspondre à l'identité d'un personnage cible, sans changer la pose. Nous proposons des fonctions de pertes sur l'identité auto-supervisées, qui permettent une étape de précision des poids du modèle au moment de l'inférence, permettant le transfert à des identités éloignées de la base de données d'apprentissage, telles que des humains habillés simplement. Notre modèle généralise bien aux poses complexes non vues durant l'entrainement.

Enfin, nous étudions l'impact des auto-contacts entre les surfaces du corps sur la perception des équivalences de pose. En effet, nous avons observé que certains auto-contacts n'étaient présents qu'en raison de la morphologie du personnage et n'étaient pas importants pour la pose. La

préservation de tous les auto-contacts dans notre première méthode peut donc créer des artefacts dans certains cas où des auto-contacts inutiles ont été préservés, modifiant significativement la pose. Nous réalisons une étude dans laquelle nous présentons aux observateurs deux modèles d'un personnage imitant la pose d'un personnage source, l'un avec les mêmes auto-contacts que la source, et l'autre avec un auto-contact en moins. Nous demandons aux observateurs de choisir le modèle qui imite le mieux la pose de la source. Nous montrons que les poses avec différents auto-contacts sont considérées comme différentes par les observateurs dans la plupart des cas, et que cet effet est plus fort pour les auto-contacts impliquant les mains que pour ceux impliquant les bras.

Acknowledgments

This research work was supported by the Inria IPL AVATAR project.

My first thanks go to my supervisors, Edmond, Franck and Stefanie, for giving me the opportunity to do a Ph.D. at Inria, for their trust, their teachings, and their continuous support that made it possible for me to succeed in this thesis. I would also like to thank Adnane and Ludovic, for their important involvement and help with several of this thesis' projects.

I would then like to thank Taku Komura, for accepting to give his external expert opinion on my work at several points of my Ph.D.. Furthermore, I would like to thank my manuscript reviewers, Pierre Hellier and Taku Komura, and the rest of my Ph.D. committee, Adnane Boukhayma, Jocelyn Chanussot, and Celine Loscos for their helpful and insightful comments, and for their kindness that helped relieve the stress of the defense.

I would also like to thank all the people I had the chance to work with during these almost four years. I'm especially grateful to, in no particular order, Joao for the friendly chats and his helpful input on my 3DV paper, Badr for his great internship work, Sergi for the interesting discussions and his help with especially recalcitrant 3D models, Julien for the beer brewing tips and the almost successful attempts to make me run on a regular basis, Matthieu and Vincent for all the pleasant banter, the intense shogi breaks and the youtube wandering, Mathieu for being there for coffee breaks and stress relief even when covid emptied the office, Edmond, Matthieu, Julien and the very selective Saint Naz en Finesse paragliding group, for the morning hike-and-fly excursions and the lunch breaks spent waiting for the right moment to take off, Benjamin for exploring the streets of Newcastle with me, Gabriel for his important help in teaching myself and students good web programming practices. My thanks also go to everyone in the Morpheo team, Boyao, Di, Nitika, Claude, Pierre, Nicolas, Eymeric, Robin, Victoria, Tomas, Sanae, Haroon, Nathalie, Jean-Sébastien, and anyone I already mentioned or might have forgotten, for the relaxing coffee breaks, the interesting weekly talks, the helpful advice, the long evenings at the K'fée des jeux or in the Paul Mistral park, and the seminars and cheese based food on top of Chartreuse and Belledonne's mountains. I would also like to thank my colleagues at my teaching position at the ENSC engineering school; Jérome Saracco, Jean-Marc André, Benoit Le Blanc and the rest of the teaching and administrative staff, for giving me the opportunity to gain significant teaching experience, and the time needed to finish writing my thesis serenely; Gabriel, Hélène, Floriane, Yvan, Alexis, Alix, Pierre, for their good mood, the cakes and their seemingly infinite creativity for wacky shenanigans that made every day at the Ph.Ds' office a surprise.

I would also like to thank the members of the POTIOC Inria team in Bordeaux at the time of my first internship, especially Sol, PA and Martin, for introducing me to the research world and for their friendly welcome.

I am also grateful to everyone in my personal life that was there for me during these intense four years. I would first like to thank my family, especially my parents and my sister, for their support, their love, and for genuinely trying to understand what I was doing in my thesis. I would also like to thank my friends from Bordeaux, Aline, Thibaud, Alex, Charles, Nico, Alexis, Marianne, Aubin, Sébastien, Micka, Capu, Dylan, Kriss, Florian, Sylvain, for your presence and emotional support despite distance, and all the great times in person or by interposed screens. It's an everyday pleasure having you in my life. I am especially grateful to Sébastien, for proofreading important parts of this document, and Aubin, for his technical help during my defense. Finally, thank you Charlotte for following me and being there through all the ups and downs of a Ph.D. student life. Through your presence and your support, you gave me the motivation and the energy to continue this adventure. This would have been so much harder without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Computer graphics aim to numerically and automatically model the 3D world. In the past few decades, creative medias such as animation movies or video games have made 3D content widely available to audiences. Specifically, modeling and animating 3D characters is central to these applications, as these characters are often important elements of 3D contents. More recently, immersive medias have seen fast technical improvements and popularization. These medias concern either Augmented Reality (AR), *i.e.* the introduction of virtual elements in the real world, such as *e.g.* Hololens or Google Glass, and Virtual Reality (VR), *i.e.* the immersion in a fully virtual environment, such as HTC Vive or Oculus Quest Head Mounted Displays. A valuable application of these new tools is the ability to virtually recreate an environment and inter-person interactions. Telepresence in such an environment could for example allow virtual tourism such as visits of monuments, or virtual meetings, such as scientific conferences. This application could help preserve fragile monuments or natural habitats by reducing attendance. It would also greatly reduce the need for plane travel, making tourism and international meetings more accessible and significantly reducing their carbon footprint. In telepresence, users are immersed in the virtual environment through an Avatar that represents them in this world, *e.g.* Figure 1.1. Similarly to traditional medias such as movies or video games, the Avatar and its motion must look realistic, but must also imitate as well as possible the motion of the user in order to preserve the user's immersion.

While realistic animations are thus a cornerstone of virtual 3D contents, they require skilled artists to create, and are expensive both in terms of time

(a) A student learning Tai Chi with a virtual teacher. Figure extracted from [Chua et al., 2003]

(b) Two persons interacting in VR through Avatars. Figure extracted from [Fribourg et al., 2018]

Figure 1.1: Example of users and their Avatar in immersive virtual environments

and money. A lot of interest has been given to automating parts or all of the animations' tedious steps. This helps reduce the overall price of the animation process, and gives more artistic liberty to animators by simplifying repetitive tasks. In particular, being able to apply an already existing animation to a new character with a different body shape (*e.g.* Figure 1.2) is of great interest. This problem is known in Computer Graphics as motion retargeting. This would save a lot of time to artists in designing highly populated environments, where a lot of different characters perform similar motions, or applying existing animations to new user designed characters *e.g.* in the context of a video game. Automatically animating a character would also allow interactive Avatar control, where a virtual character is animated to mimic the motion of an user in real time, which is obviously impossible for an animator.

A popular and simple way to represent a 3D character's motion is to use an intermediary skeleton that approximates the character's body. This skeleton can be deformed by *e.g.* modifying the angles between the different body segments. Applying an existing animation to a new character can thus be done by applying the deformations of the source character's skeleton to the new character's. However, the skeleton does not represent the surface of the character, and thus can not model important information on the motion of characters. For example, contacts between body surfaces or between the body and the environment can be lost, and collisions can appear. In this thesis, we are thus interested in automatically animating a new character to mimic an existing motion, while adapting the motion to the new body shape, and taking into account surface information.

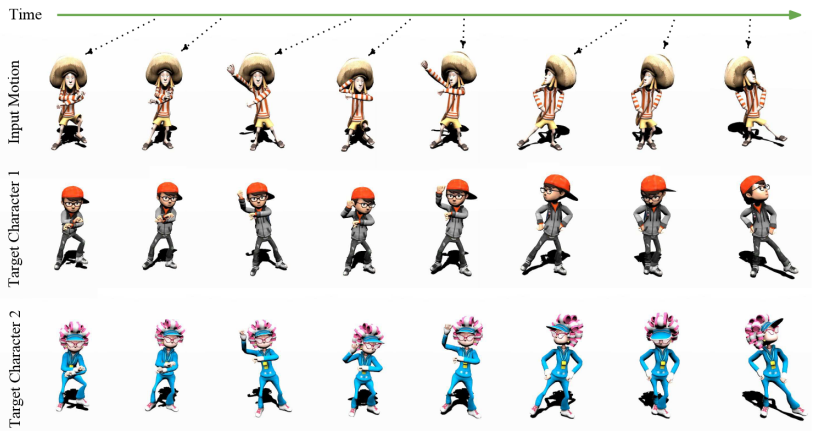


Figure 1.2: Example of an animation (top row) being applied to new characters (middle and bottom rows). Figure extracted from [Villegas et al., 2018]

1.2 Problem Statement

In this manuscript, we explore motion retargeting, *i.e.* how to make a new character mimic an existing animation, such as an Avatar mimicking the motion of a user. To do so, we represent the motion as a succession of static poses, and separately make the new character mimic each pose in order to recreate the full animation. For each pose, we take as input a source and a target 3D characters, and the deformed source character performing a specific pose. Our goal is to generate the target character performing a pose that is similar to the pose of the deformed source character. This approach is known as Deformation Transfer, and is illustrated in Figure 1.3.

The first approach to deformation transfer is to deform the pose of the target character to match the source character's pose, while preserving its identity. This is the Pose Transfer strategy (Figure 1.3, right column). This approach has been widely explored in the literature, and naturally corresponds to how humans imitate poses: when a person tries to mimic the pose of an another, for example a student following the motion of a yoga instructor, they move their body as similarly as possible in order to get in the correct pose.

This strategy relies on the assumption that poses are defined consistently across different characters. However, the pose of a character is subjective, and depends on contextual information. For example, the characters in Figure 1.4, extracted from a dataset of captures of human actors [Bogo et al., 2014], are considered to be performing similar poses. However, small variations exist between them, such as the exact angle of the arms or the

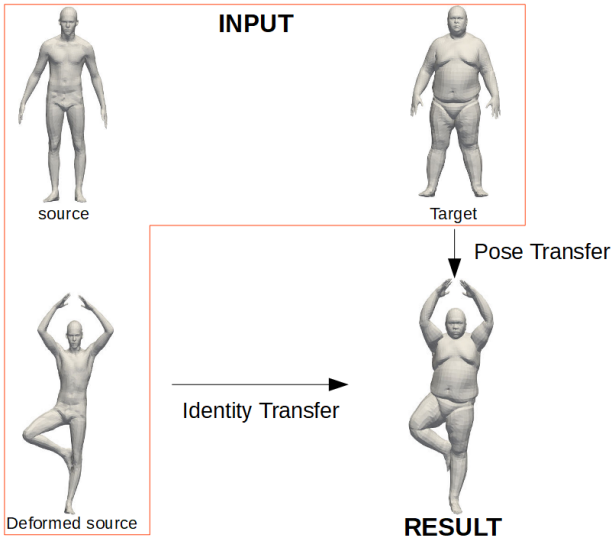


Figure 1.3: Illustration of the deformation transfer problem. Given as input a source and target character, and a deformed source character, we aim to generate the target character performing the deformed source's pose. Two orthogonal approaches can be applied: transferring the pose of the source to the target (Pose Transfer), or transferring the identity of the target to the deformed source (Identity Transfer). Deformation transfer result obtained with the method described in Chapter 4

contact point between the leg and the foot. These poses could thus be considered different in another context, depending on the application. Different approaches have been explored to define more precisely equivalences between poses. A popular approach is to identify constraints that the pose must satisfy in order to be similar to another. These constraints can be *e.g.* angles between body segments, spatial relationships between body parts, contacts with the ground, the environment or the character's own body. When transferring the pose of a character to a new one with a different morphology, respecting the combination of these constraints can result in a complex optimization problem. Moreover, it is not always obvious which constraints give meaning to a specific pose in a given context. For example, while the distance between the hands of a clapping character should be important, it is not obvious whether the distance between the hands and the torso of the character has any importance.

Another orthogonal strategy is to start from the source deformed pose, and to deform the surface of the character in order to match its identity to the target, while preserving its pose. We call this strategy Identity Transfer (see Figure 1.3 bottom row). Contrarily to the pose, identity of a character

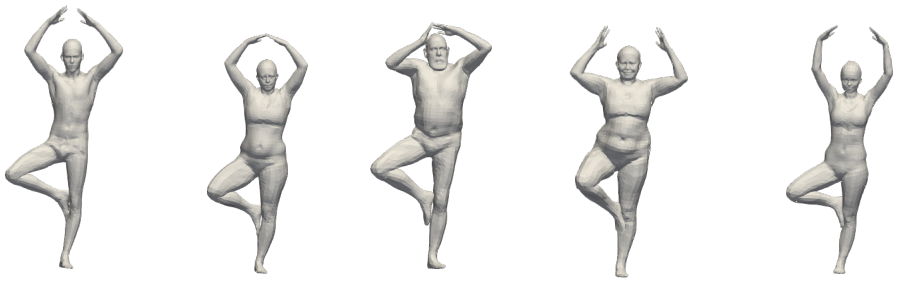


Figure 1.4: Characters with the same pose label extracted from the FAUST [Bogo et al., 2014] dataset

is an objective parameter: it is obvious whether two characters have the same identity. It should thus be possible to parameterize the identity of a character in order to apply it to another character in a different pose.

With this approach, as the optimisation starts from the source character in the correct pose, the pose constraints (*e.g.* body part orientations or contacts) are naturally satisfied at the initial step. This strategy should thus be easier to apply on complex poses as the task of selecting constraints that define the pose is greatly simplified. When deforming the surface to a new identity with a different morphology, artifacts can appear, such as collisions or loss of contact constraints when the body parts inflate or deflate. Small pose corrections should thus be applied to avoid these artifacts. We assume that other constraints, such as general orientation of the body segments or spatial relationships between body parts should not be significantly affected by these corrections. Identity transfer thus requires satisfying significantly fewer constraints than pose transfer.

In this work, we thus focus on identity transfer while preserving important self-contacts between body surfaces, and avoiding collisions. However, while collisions are an obvious problem that must be solved in order to obtain a realistic pose, some contacts may or may not be an important constraint to the meaning of the pose. Indeed, we argue that while some contacts give meaning to the pose, such as the hands touching for a character clapping, others are only present as an effect of the morphology of the character, such as the thighs touching for a character with a wider morphology. We thus also explore in this manuscript how to select which self-contacts are important to the meaning of the pose and which are not.

1.3 Outline and Contributions

In Chapter 2, we present an overview of works that explored how to transfer existing animations to new characters. We briefly discuss methods that solved this problem between 2D characters and between 3D skeletons, and focus the rest of the chapter on 3D surface deformation transfer approaches. We also briefly review existing datasets of 3D human characters that are commonly used in research on this subject. In Chapter 3, we give background information and preliminary discussions on the approaches explored in this thesis. More specifically, we further discuss the identity transfer strategy, and the importance of self-contacts. In the following chapters, we present the main contributions of this thesis.

In Chapter 4, we propose Contact Preserving Identity Transfer, an optimization based identity transfer method that iteratively deforms the surface of a character to match its identity to a target. The method applies small pose corrections in order to avoid interpenetrations of body parts, and to preserve self-contacts present in the source pose. This method shows that the identity transfer strategy gives good results across different categories of characters. This method is able to adapt the pose to the identity of the new character, by allowing distances between body parts to change and new contacts to appear to accommodate to the different morphology. However, several limitations remain. First, the method is slower than state of the art approaches, which limits possible applications. Some artifacts can still appear in the results, such as unnatural surface deformations around the joint for extreme poses or morphologies. Finally, this method naively preserves all self-contacts present in the source pose, which can lead to artifacts as discussed in the previous section. As designing a fast and robust identity transfer approach, and exploring the importance of self-contacts to the meaning of the pose are two difficult problems, we treated them separately in the subsequent chapters.

In Chapter 5, we presented an adaptation of the Identity Transfer strategy to the deep learning framework. Deep learning approaches display several interesting advantages; while the training step can be long, the inference time transfers are close to real time. Moreover, learning from data how human characters deform should give more realistic results and avoid unnatural deformations observed in the previous contribution. Our method, Neural Identity Transfer, is able to generalize to poses unseen at training time better than the state of the art, and is faster than our previous optimization based method.

Finally, in Chapter 6, we explore the importance of self-contacts for pose equivalences. We assume that human observers should be able to de-

termine if two characters are performing the same pose. We thus designed a perceptive study in which we presented observers with different characters performing a similar pose, with differences in the self-contacts preserved. The subjects are asked to select which pose better imitates a target pose. This study shown a tendency of users to give more importance to certain contacts, specifically those implying hands. This study is a first step towards a better understanding of the importance of self-contacts to the pose, and will be extended to a wider panel in a future work.

We give a general conclusion of this manuscript in Chapter 7.

1.4 List of publications

- Basset, J., Wuhrer, S., Boyer, E., and Multon, F. (2019). Contact preserving shape transfer for rigging-free motion retargeting. In Motion, Interaction and Games (pp. 1-10).

  HAL page: https://hal.inria.fr/hal-02293308v2

- Basset, J., Wuhrer, S., Boyer, E., and Multon, F. (2020). Contact preserving shape transfer: Retargeting motion from one shape to another. Computers & Graphics, 89, 11-23.

  HAL page: https://hal.inria.fr/hal-02613783/

- Basset, J., Boukhayma, A., Wuhrer, S., Multon, F., and Boyer, E. (2021). Neural Human Deformation Transfer. In 2021 International Conference on 3D Vision (3DV) (pp. 545-554). IEEE.

  HAL page: https://hal.inria.fr/hal-03440562/

Chapter 2

Related Works

In this chapter, we review state-of-the-art works that are relevant to the problem of motion retargeting, and more specifically to the deformation transfer approach defined in Section 1.2. We first briefly present works that explore how to edit videos or images of 2D characters, in order to create new 2D content from a source 2D character. While we are interested in animating 3D characters, these 2D methods helped introduce interesting concepts and research directions to the motion retargeting literature. We then discuss skeleton based methods, that explore how to transfer the motion of a 3D skeleton to a new one with different bone lengths and/or topology. We then give a more in-depth review of 3D surface deformation transfer methods, that are closer to our contributions. Finally, we briefly present the main datasets of 3D human shapes and poses that are used for deformation transfer or deep learning works.

2.1 2D Motion Retargeting

Video and image editing has received a lot of interest in the last two decades. In this section, we briefly review approaches that have explored how to create new 2D content by transferring parameters of 2D human characters. [Efros et al., 2003] propose to match poses of characters in videos using optical flow. This allows them to match a character to a 2D skeleton that can be used to transfer his motion to a new character. Similarly, [Kemelmacher-Shlizerman et al., 2010] propose to puppeteer 2D images of human faces by retrieving the closest neighbour to the source pose+expression face in a database of faces of the target character. Some works use multiple views of a character performing a motion in order to fit it to a 3D model, which is then manipulated to generate new motions from

a given point of view [Cheung et al., 2004, Xu et al., 2011]. The idea of using 3D data to constrain the 2D motion was also used by [Hornung et al., 2007], who transfer the animation of a motion captured 3D skeleton to the 2D character. Similarly, some works use statistical analysis of 3D datasets to model human body parameters, and use it to modify the appearance and morphology of 2D characters [Zhou et al., 2010b, Jain et al., 2010].

The recent advances of deep learning allowed to drastically improve the performance of 2D motion retargeting methods. With the massive amount of 2D images and videos available, an important branch of works have explored how to learn from data how to deform 2D characters. These approaches are based on various deep learning architectures such as Recurrent Neural Networks (RNNs) [Kappel et al., 2021] Generative Adversarial Networks (GANs) [Chan et al., 2019, Liu et al., 2019] and Variational Auto Encoders (VAEs) [Esser et al., 2018]. Related to the deformation transfer problem, some works have explored how to perform style transfer between images [Gatys et al., 2016, Huang and Belongie, 2017, Huang et al., 2018]. These methods transfer style properties of images to generate new ones, and can be applied to 2D pose transfer [Isola et al., 2017]. [Aberman et al., 2020b] successfully apply the Adaptive Instance Normalization (AdaIN) layer traditionally used in image generation to motion style transfer. This allows them to transfer the style of a motion in a captured 2D video to a 3D skeleton.

While these methods are applied to 2D videos or images, they are still interesting to our 3D deformation transfer problem, as they helped develop important building blocks of deep neural networks and inspired works that solved the same problem in 3D.

2.2 Skeletal Motion Retargeting

With the popularization of marker-based motion capture systems and of 3D content creation in the early 90's, the ability to reuse existing 3D data, such as motion captured on a user or designed by an artist, rapidly gained success. 3D Motion retargeting first appeared as the task of mimicking an animation performed by a skeleton to a new one, with different bone lengths or topology.

The first approaches performed this by solving kinematic constraints on joint positions and ensuring continuity using space time constraints [Gleicher, 1998, Lee and Shin, 1999, Popović and Witkin, 1999, Choi and Ko, 2000]. Another approach consists in defining a morphology-independent representation of the movement, that can be then applied to different new

skeletons. This can be done for example by encoding the movement and the associated constraints on a normalized skeleton [Kulpa et al., 2005, Kulpa and Multon, 2005]. Morphology-independent representations can also allow to retarget to characters with different topology, *e.g.* by using an intermediate skeleton [Monzani et al., 2000] or semantic labels [Hecker et al., 2008].

All these methods generally use predefined kinematic constraints that must be manually tuned. This implies the intervention of an animator or of the user, and can be a heavy step in the animation process. Automatic kinematic constraint detection in the source motion has been proposed to automate the constraint editing problem, with *e.g.* the work of [Le Callennec and Boulic, 2006]. Most of these constraints consist in spatial relationships between body segments, which can be modeled as distance constraints [Al-Asqhar et al., 2013, Bernardin et al., 2017] or as more generalized spatial relationship between joints [Baciu and Iu, 2006]. These methods aim at transferring the topology between body segments of the source motion to the target character, while using generalized inverse kinematics to solve all the corresponding constraints. This idea of modeling the topology between body segments has been extended by introducing an interaction mesh [Ho et al., 2010, Ho et al., 2014]. The interaction mesh connects joints of the skeleton by edges. During retargeting, the authors aim to minimize the local deformations of this mesh, therefore preserving spatial relationships between skeleton joints.

Following the traditional skeletal animation techniques, some works have explored how to transfer the animation of a skeleton to a new different one using tools from deep learning.

A common strategy in the state-of-art is to include a differentiable forward kinematic layer in a deep neural network [Zhou et al., 2016, Villegas et al., 2018, Shi et al., 2020], to predict the final joint positions of the skeletons. This kinematic layer allows to capture the parameters of an input skeleton's motion, and to transfer them to a new skeleton with the same topology and different bone lengths.

[Aberman et al., 2020a] explore how to perform retargeting between skeletons with different topologies, by reducing the skeletons of their inputs to a primal skeleton similar for all characters. They also introduce a skeleton convolution operator. Their method allows to retarget motions between skeletons with similar global structure (*e.g.* same number and position of limbs), but different topologies.

As these methods retarget the motion at the skeleton level, they do not account for interactions between body surfaces. This can result in interpenetrations or loss of important self-contacts when animating a skinned character using the skeleton animation. [Villegas et al., 2021] propose a RNN

architecture that retarget the motion between skeletons, while solving constraints on the resulting skinned surface motion to preserve self-contacts and avoid interpenetrations.

2.3 Surface Mesh Deformation Transfer

Skeleton-based approaches have difficulties dealing with pose features that concern the surface of a character. This can result in artifacts, such as interpenetrations between body surfaces. Moreover, applying the motion of a skeleton to a surface character can be a tedious task, implying several intermediary steps such as rigging and skinning of the target character. Closer to our method, a lot of interest has been given to transferring directly the deformation of the 3D surface of the characters, generally in the form of 3D meshes.

2.3.1 Optimization Based Deformation Transfer

An important category of 3D mesh deformation transfer methods encode the pose of the source character as a deformation of the source surface mesh, and transfer this deformation to the target surface mesh, using an optimization framework.

[Sumner and Popović, 2004] propose to compute the transformation of each triangle in the deformed source mesh, and to apply these transformations to the triangles of the target mesh through a correspondence map. This method was extended to multi-component objects by [Zhou et al., 2010a]. Other methods similarly proposed to encode deformations of the source character's surface to transfer it to the target character using correspondence maps [Zayer et al., 2005, Zhao et al., 2011]. A different approach is to encode the deformation of the space in which the source character is embedded, and apply a similar space deformation to the target character. This can be done using *e.g.* coarse tetrahedral control meshes [Zhao et al., 2009], cage structures [Chen et al., 2010] or harmonic maps [Ben-Chen et al., 2009]. These methods only needs global landmark correspondences between the source and the target, and as such can be applied to transfer between shapes with different topologies and geometries.

Similarly to skeleton based approaches, some surface deformation transfer methods have focused on automatically detecting and preserving constraints on the poses of the characters, such as contacts or distances between body segments. For instance [Liu et al., 2018] introduced the context graph, an extension of the interaction mesh proposed for skeleton joint centers [Ho

et al., 2010] to body surfaces. The edges of this context graph link nodes on the body surface of the characters. By minimizing the deformation of the context graph, the method preserves the distances between body surfaces during deformation transfer. Using distance constraints to preserve context was also explored by [Jin et al., 2018], who proposed the Aura mesh, a volumetric mesh enclosing the body surface with a fixed offset. Spatial relationships are then detected as the interpenetration of this Aura mesh, and preserved in the deformation transfer result.

Some works focus on generating more realistic animations by taking into account physics-based constraints, such as balance [Lyard and Magnenat-Thalmann, 2008]. In their work, [Al Borno et al., 2018] propose to apply physics-based forces on the characters. As a result, the pose adapts to the morphology of the target character, *e.g.* the extent of a kick motion depends on the corpulence.

2.3.1.1 Hybrid Approaches Combining Skeleton and Surface

In order to benefit from both worlds, a body of work combines skeleton and surface constraints in the pose transfer. [Molla et al., 2018] use a skeleton to model the pose with joint angles, and combines it with a simple surface representation with bounding volumes. They introduce egocentric planes to ensure that the topology between body parts is preserved in the result character. Other methods use a complete surface mesh together with a skeleton [Huang et al., 2013, Le Naour et al., 2019] to control the surface mesh deformation while preserving the coherence with the skeleton topology. By satisfying both skeletal and surface constraints, natural animations and poses can be generated. [Shi et al., 2007] combine constraints on the skeleton (*e.g.* limb lengths and joint limits) and on the surface of the body (*e.g.* self-collisions) to generate plausible new poses.

2.3.1.2 Spectral Style Transfer

Spectral Geometry has been a widely explored field in computer graphics, and proposed numerous applications to 3D meshes [Zhang et al., 2010]. Some works from this field explored how to use spectral information, such as the eigendecomposition of the Laplacian operator, in order to deform 3D shapes [Rong et al., 2008]. These works build from the well know observation that eigenvectors of the spectrum associated with small (respectively large) eigenvalues encode low (respectively high) frequency details [Dey et al., 2012]. In the case of 3D shapes, low frequencies correspond to the general pose of the shape, and high frequencies correspond to the surface

details. It is thus possible to deform these different levels of details by acting on the corresponding frequencies of the spectrum.

From this observation, spectral mesh deformation has naturally been applied to the deformation transfer problem. [Lévy, 2006] compute the eigendecomposition of the Laplacian operator for the source pose and the target identity characters, and project their geometry on the resulting eigenfunction bases. They then reconstruct a 3D shape using the low frequency projections (*i.e.* the pose information) of the source pose character and the high frequency projections (*i.e.* the surface details) of the target identity character. This method is able to perform deformation transfer between near isometric shapes, that have relatively similar Laplacian eigendecompositions, but tends to fail for non isometric shapes where the eigenfunctions are inconsistent. Several works have explored how to improve this application of spectral mesh processing. [Kovnatsky et al., 2013] propose to use an approximate common eigenbasis of the Laplacian of the two models, to avoid inconsistent eigenfunctions. [Yin et al., 2015] use the low frequencies of the source pose character spectrum as handles for a classical deformation problem, and preserves the surface details of the target identity using Laplacian coordinates. More recently, [Cosmo et al., 2019] introduced Isospectralization, where they deform a mesh in order to align its Laplacian spectrum to a target one, using modern differentiable programming tools. The authors show several applications of their method, such as shape matching and style transfer.

Most of these methods can be adapted to perform deformation transfer between meshes with different connectivity, widening the possible applications. However, the Laplacian spectrum is known to be sensitive to noise. Moreover, it struggles to encode the fine details of the shapes, resulting in either over-smoothed transfer results [Lévy, 2006, Kovnatsky et al., 2013, Cosmo et al., 2019] or requiring additional elements to transfer these fine details [Yin et al., 2015].

2.3.2 Data Driven Approaches without Deep Learning

With the availability of 3D characters datasets, approaches have leveraged existing 3D data to understand how to animate characters and perform applications such as deformation transfer.

2.3.2.1 Semantic Deformation Transfer

Semantic deformation transfer was first explored by [Baran et al., 2009]. This approach allows to perform deformation transfer between very different classes of shapes, *e.g.* human to animal characters, by focusing on semantic properties of the shapes and not their literal deformation. The key idea is to select a number of previously existing poses of the characters, and label them with similar semantic information, *e.g.* "raising a leg". A shape space of each character is then created, and a linear map between the shape spaces is computed using the labelled examples. Deformation transfer is then performed by encoding a new pose of the source in its shape space, and mapping it to the shape space of the target before reconstructing it.

[Boukhayma et al., 2017] use this approach to map motions between subjects, using sparse correspondences of key poses for each motions. The authors use Gaussian Process regression for the motion mapping step of their method, which improves the accuracy and generalization capability of their model. [Rhodin et al., 2014] use predefined correspondences between surfaces and sparse point clouds to create a mapping between the two. This allows the user to interactively control a 3D character's mesh using input from consumer devices, *e.g.* skeleton approximation from a Microsoft Kinect. The authors extend this approach by mapping user motions to motions of the 3D characters to be animated, and leveraging wave properties of the motions (amplitude frequency and phase) for animation control [Rhodin et al., 2015]. Other works use a similar approach and propose blendshapes, i.e. base expressions of a target face, and estimate from captured user video weights associated with each blendshape on the source to create the resulting target expression [Bouaziz et al., 2013].

All these methods neither require skeletons nor point-to-point correspondences between sources and targets. They can thus be applied between very different classes of shapes, and used to control animations from consumer capture systems. However, they require heavy preprocessing to define correspondences between source and target characters, and the quality of their results is highly dependent on the quality of this initial step.

2.3.2.2 Statistical Shape Models

To make use of the large amount of 3D data available, an important line of works have focused on using statistical analysis of large datasets in order to derive parameters that control the shape of 3D characters. Early works in this field model body shape deformations of human characters and are able to morph the shapes to generate new characters [Allen et al., 2003,Seo et al.,

2003]. By animating a model with standard animation techniques, and modifying its body shape, these methods are able to perform deformation transfer.

Following this strategy, several works have focused on modeling separately shape and pose parameters of 3D characters. These methods naturally allow deformation transfer by combining the shape parameter of a character with the pose parameter of another. SCAPE [Anguelov et al., 2005] is the first model to do this. In this model, pose and shape deformations are modeled using deformations of triangles. Several works have based their models on SCAPE and aimed to improve it, *e.g.* by taking into account the interaction between shape and pose parameters in generated deformations [Hasler et al., 2009,Chen et al., 2013], or by modeling dynamic deformations of the bodies [Pons-Moll et al., 2015]. Other approaches encode the shape of the human models using Principal Component Analysis (PCA), and use Linear Blend Skinning (LBS) to deform their pose [Neophytou and Hilton, 2013, Pishchulin et al., 2017]. The popular work of [Loper et al., 2015], SMPL, use a similar approach combining PCA and LBS equations, in a skeleton driven way. Their model is more accurate than SCAPE, easy to use, and is compatible with existing rendering engines. Several works have extended the SMPL model, *e.g.* to include fully articulated hands and facial expressions [Pavlakos et al., 2019], or to model the impact of body shape on pose deformations [Osman et al., 2020]. [Zuffi et al., 2017] use a similar approach than SMPL to create a shape and pose model for 3D animal models, SMAL.

These methods generate realistic human shapes and poses, and can be naturally applied to deformation transfer between characters encoded in their models by exchanging shape and pose parameters. However, transferring between arbitrary characters is more complex and requires the preprocessing step of fitting the characters to the model. Moreover, these methods usually do not encode surface constraints, which can result in collision artifacts for complex poses and/or extreme shapes.

2.3.3 Deep Learning Approaches

2.3.3.1 Learning From 3D Data

Deep learning methods have naturally been applied to learn from data how 3D shapes deform. A lot of works have explored how to use 3D data to train generative deep learning architectures, and reviewing all of them is beyond the scope of this work. In this section we review the main categories of

these approaches that have been applied to learn deep deformation transfer models.

[Jiang et al., 2020] represent 3D data by encoding the vertices of their meshes in a lower dimensional feature, based on an anatomical hierarchical segmentation, and train their model on those features. Similarly, [Tan et al., 2018] represent meshes with rotation invariant features and train a variational encoder with fully connected layers on these features.

Convolutional Neural Networks (CNNs) have showed impressive generative capacity for 2D images, thanks to their hierarchical structure that allows them to exploit different scales of details of the images. However, they are difficult to adapt to 3D data, as they can not be directly applied to irregular graph-like structures with no ordering, such as 3D meshes. A lot of works have explored how to adapt CNNs to the 3D domain. A first approach is to map the 3D data to a 2D representation, *e.g.* a rendered image [Su et al., 2015] or a height map [Abrevaya et al., 2018a], for which classical CNNs can be applied. This intermediate step however adds complexity to the architectures, and can be a source of error if the 3D to 2D mapping is not perfect. [Defferrard et al., 2016] define convolution layers in the spectral domain of the input meshes. Their approach has been successfully applied to learning face [Ranjan et al., 2018] and human body [Tretschk et al., 2020] models. Ranjan *et al.* also introduce down and up sampling layers adapted to 3D mesh data based on quadratic edge collapse. These methods also require an intermediary representation, and other approaches have explored how to adapt the convolution layers directly to the graph structure of the meshes. [Bouritsas et al., 2019, Gong et al., 2019] use a spiral ordering of the neighbourhood of each vertex to apply convolutions to 3D meshes. [Verma et al., 2018] propose a new graph convolution operator, that dynamically determines from the learned features correspondences between filter weights and local graph neighbourhood.

Other approaches have explored architectures that can be applied directly to point clouds independently from vertex ordering [Klokov and Lempitsky, 2017]. An important contribution in this direction is the PointNet architecture proposed by [Qi et al., 2017]. This architecture uses shared weights Multi Linear Perceptrons (MLP) to learn per-point spatial encoding and max pooling layers to obtain a global feature. PointNet has rapidly been used in a wide variety of state of the art approaches, such as classification, segmentation, or generative models predicting deformations of 3D shapes [Cosmo et al., 2020, Wang et al., 2020].

More recently, a lot of interest has been given to methods that learn implicit functions of the 3D surfaces. These methods can be applied independently from topology of the input 3D models. They are based on

different implicit representations of the 3D surface, such as Signed Distance Functions (SDF) [Park et al., 2019], occupancy [Mescheder et al., 2019, Mihajlovic et al., 2021, Deng et al., 2020] or implicit fields [Chen and Zhang, 2019].

2.3.3.2 Deep Deformation Transfer

Based on the advances of deep learning, and its adaptations to the 3D domain, a lot of works have explored how to perform deformation transfer between 3D characters using Deep Neural Networks.

Deep learning approaches have taken inspiration from the deformation transfer literature. [Gao et al., 2018] adapted the semantic deformation transfer approach described in Section 2.3.2.1. They trained two autoencoders for the source and the target shapes, and used a GAN to map the latent code of a deformed source to the latent code of the deformed target. Similarly to semantic deformation transfer, this leads to satisfying results, but the model needs to be retrained for each new shape pair. [Marin et al., 2020] extend the spectral style transfer approach described in Section 2.3.1.2 to a deep learning architecture by the mapping latent space of an autoencoder to eigenvectors of the spectrum of a shape. The Neural Pose Transfer (NPT) approach [Wang et al., 2020] takes inspirations from 2D deep style transfer techniques and uses spatially adaptive instance normalisation [Huang and Belongie, 2017] to perform pose transfer between human characters. Their architecture is based on PointNet, making their method applicable to unordered meshes. Recently, [Lombardi et al., 2021] proposed to learn a disentangled representation of human shape and pose using implicit SDF. Their model is able to perform deformation transfer by exchanging shape and pose parameters of two different characters.

A lot of interest has recently been given to autoencoders to disentangle shape and pose parameters [Jiang et al., 2020, Tretschk et al., 2020], or identity and expression parameters in the case of 3D faces [Ranjan et al., 2018, Abrevaya et al., 2018a]. Similarly to statistical models presented in Section 2.3.2.2, these methods naturally allow deformation transfer by reconstructing a character from the shape and pose parameters of two different characters. A common problem these methods encounter is the lack of large real-world datasets with pose labels. To remedy this, [Cosmo et al., 2020] present LIMP, a supervised model that allows to train from small-scale datasets. LIMP is built on the hypothesis that human pose deformations are near-isometric, and thus propose to preserve geodesic distances on the characters' surfaces. The computational cost of their metrics makes LIMP unscalable to large datasets. Other methods explored

unsupervised approaches to avoid this limitation. [Aumentado-Armstrong et al., 2019, Aumentado-Armstrong et al., 2021] use the LBO spectrum to define intrinsic shapes in a way invariant to isometric pose deformation. [Zhou et al., 2020] create pseudo ground truth for the pose transfer between two characters, by applying on the fly As Rigid As Possible deformations [Sorkine and Alexa, 2007] during training.

2.4 Datasets

With the development of 3D acquisition techniques and synthetic data creation, a wide variety of datasets of 3D human characters has been made available for research purposes over the past two decades, and it is beyond the scope of this work to list all of them. In this section we simply review some of the main datasets of 3D human body models that are widely used in the state-of-the-art to understand how human characters deform, and train deep deformation transfer models.

One of the first consequent datasets of 3D scans of full human bodies is the CAESAR dataset [Robinette et al., 1999]. This project aimed at gathering anthropometric data of populations of NATO countries in north America, Netherlands and Italy, and resulted in a large number of full body scans of participants in a few key poses. The resulting dataset was widely used in human statistical analysis, and was used to train models such as SMPL and SCAPE. Other more recent datasets aim at providing scans of real human characters. These works use multi-view videos of actors to reconstruct 3D human characters with good precision [Gkalelis et al., 2009, Sigal et al., 2010, Yang et al., 2016]. [Ionescu et al., 2013] propose the Human3.6M dataset, containing 3.6 millions of 3D human poses captured on actors. This significant size increase compared to previous datasets allows an important gain of performance for their accompanying statistical model. FAUST [Bogo et al., 2014] contains scans from 10 actors performing a variety of poses, and is used as benchmark for shape registration. This dataset has been extended to contain dynamic data by [Bogo et al., 2017]. [von Marcard et al., 2018] provide the 3D Poses in the Wild (3DPW) datasets, which contains 2D frames of in the wild clothed characters, associated with 3D meshes reconstructed with their state-of-the-art video to 3D method.

Using recent animation techniques or models of human shape and pose, other works create synthetic datasets of 3D human characters. In order to help gather and unify a large dataset in a single representation, [Mahmood et al., 2019] present the AMASS dataset. They gather 15 popular motion

capture dataset, and convert the data to 3D human meshes that they parameterize to fit the SMPL model. [Müller et al., 2021] propose to associate 3D models with 2D images, focusing on poses with self-contacts between body surfaces. They do this by selecting 3D models from AMASS or from refined 3D scans with self-contacts. They then associate a 2D image to each model by asking subjects to take pictures of themselves imitating the pose of a 3D model. As these datasets use pose information from motion capture and shape information from human body models such as SMPL, they can be considered semi-synthetic. [Pumarola et al., 2019] present a large dataset of 3D synthetic clothed humans, and use it to generate annotated 2D videos. Adobe have publicly shared the Mixamo [mix,] dataset, which contains a wide variety of artist designed skeletal motions, and 3D characters. The motions can be applied to any character in the dataset and to user created characters.

Chapter 3

Preliminary

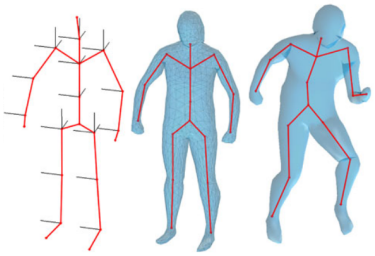
In this chapter, we present background information relevant to the remainder of this manuscript. In Section 3.1, we present our choice for 3D virtual human representation. In Section 3.2 we give more details on the identity transfer strategy presented in the introduction (Section 1.2) and propose geometrical parameters encoding the identity of a character. Finally, in Section 3.3, we present how self-contacts can help to define equivalences between poses.

3.1 3D Shape Representation

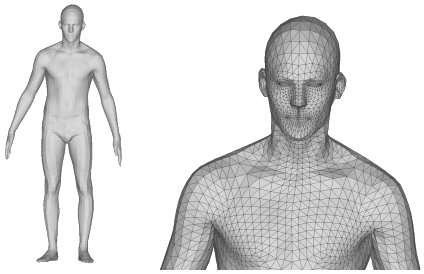
Automatically animating 3D characters requires representations of the 3D shapes that allow numerical computations. To this end, several representations have been used by state-of-the-art works.

A simple representation is to use an approximation of a character's skeleton. This skeleton is composed of a hierarchy nodes, called "joints", linked by edges, called "bones" (see Figure 3.1a, left). A skeleton can be animated by acting on the angle between joints, and moving accordingly all joints lower down the hierarchy. Skeleton animations can preserve constraints, such as contacts between joints and the environment, with the inverse kinematics method. In this method, the animator first places the constraints (such as a contact), and then uses optimisation to determine joint parameters that poses the skeleton while preserving the constraints. This representation has been widely used as a basis to retarget motions between simple 3D characters (see Section 2.2).

However representing 3D characters with a skeleton poses several limitations. The first and obvious one is that a skeleton is a very simple approximation of the character's body. In most applications of 3D anima-



(a) Left to right: 3D skeleton, skinned character, example skeleton pose. Figure extracted from [Pons-Moll and Rosenhahn, 2011]



(b) Surface mesh of a 3D character, and close-up with visible mesh edges

Figure 3.1: Example of classical 3D character representations

tions (*e.g.* movies, games or VR), we are interested in animating a more detailed 3D character, with a textured surface outside its skeleton. Skeletons can be used to control the animation of a character's body surface. This is traditionally done by associating a skeleton to a 3D mesh (*i.e.* a collection of vertices linked by edges, see Figure 3.1b) representing the surface of the character, with the process known as rigging [Magnenat-Thalmann et al., 1988, Baran and Popović, 2007] (see Figure 3.1a). Rigging determines the parameters of the deformation of the mesh depending on the skeleton pose. In particular, the skinning step associates each vertex of the mesh to one or several bones of the skeletons, and sets corresponding weights. This process is known to be long and tedious for animators, and animations obtained with it are still prone to artifacts. In particular, skinning weights are difficult to optimize, which often leads to unnatural deformations of the body surface. Moreover, as the skeleton itself does not encode surface information, collisions can appear when using it to control a surface animation.

To avoid these limitations, a lot of works have explored how to deform directly the surface of the 3D characters in order to create new poses or animations (see Section 2.3). In this work, we aim to generate realistic body surface animations, that take into account self-contacts and prevent collisions. Therefore, for the contributions presented in this document, we represent 3D shapes using triangular surface meshes. Input meshes are defined as $\mathcal{V} = (\mathbf{V}, E, F)$, where E is the set of edges of the mesh, F is the set of triangular faces, and $\mathbf{V} = (v_1, ..., v_n)$ is the set of mesh vertices with v_i the 3D coordinates of vertex i (see Figure 3.1b). Note that other representations can be used for 3D shapes, however meshes are a relatively

simple and intuitive representation, and are compatible with most recent animations tools.

3.2 Identity Transfer

As defined earlier in this document, deformation transfer takes as input a source character in a deformed pose and a target character, and aims to generate the target character performing the source character's pose. To do so, one strategy is to deform the pose of the target character to make it mimic the source pose. Another is to deform the surface of the source character in the deformed source to make its identity match that of the target character, while preserving its pose. We refer to the former as pose transfer and to the latter as identity transfer. They are illustrated in figure 1.3.

While both solutions should give the same result, in practice literature on the subject mostly explore the pose transfer strategy. In this section, we give a formal definition of the pose and identity transfer approaches. We discuss advantages of identity transfer, and propose an approach for applying this strategy in practice.

3.2.1 Pose Transfer vs Identity Transfer

Pose transfer assumes that pose is identifiable consistently across different characters. Under this assumption, pose can be parameterized by geometric properties of the 3D shape, such as the deformation of the source's mesh triangles [Sumner and Popović, 2004]. Deformation transfer then boils down to transferring the deformed pose using these parameters. However the assumption that pose can be defined consistently across different characters is arguable. Exact correspondences between body poses are subjective and contextual; small variations of the pose can or not have an effect on its semantic meaning, as discussed in Section 1.2.

Symmetrically, we can make the assumption that the shape of a human character can be defined independently from the pose, *i.e.* by its identity. Identity of a character is an objective information: it is clear that a same person at different time stamps of an animation should have the same identity. In this case, deformation transfer can be done by transferring the identity of a new character to a character already in a deformed pose.

It is important to note that both approach are approximations, and that pose and identity of a character are intrinsically entangled notions. However, when transferring identity, we avoid the difficult task of defining

equivalences between poses. By directly considering the correct pose and just modifying identity properties, identity transfer should allow to better adapt to any pose.

3.2.2 Identity Parameters

Identity transfer requires being able to parameterize the identity independently from the pose, using geometrical information. In this work, we propose to leverage common hypotheses on human deformations to identify such parameters.

3.2.2.1 Near-Isometry

A common assumption is that pose deformations of a human character are near-isometric [Elad and Kimmel, 2001,Aubry et al., 2011]. This means that distances on the surface of the body (geodesics distances) are preserved. For example, the distance on the surface of the body between a point on the hip of a character and a point on his arm stays roughly the same if the character raises its arm. However, the distance between the same points on a bigger character will be different than the distance on the first character, due to the wider body surface. This hypothesis has been successfully used to encode identity in works such as LIMP [Cosmo et al., 2020].

Building on this observation, we assume that geometric features of a 3D human mesh $\mathcal{V}$ that are isometry invariant encode the identity of the character independently of the pose. We refer this hypothesis as "near-isometry" in the remainder of this document. It is noteworthy to mention that this hypothesis is an approximation; folds around the joints and soft tissue deformations due to *e.g.* muscles contracting or breathing can create local variations of the isometry for the same character in two different poses.

We propose to use local Laplacian coordinates as an isometry invariant feature. The goal of this representation is to encode, for each vertex, its offset w.r.t. its neighbours in a local coordinate frame. This representation was first presented by [Wuhrer et al., 2012], who used it to represent the identity of characters in a pose independent manner. It was also successfully used to reconstruct the geometry of human bodies [Yin et al., 2015,Lifkooee et al., 2019].

First, the uniform Laplacian matrix L of the mesh $\mathcal{V}$ is computed as

$$L_{ij} = \begin{cases} -1 & \text{if } i = j \\ \frac{1}{deg(v_i)} & \text{if } v_j \in N_1(v_i) \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

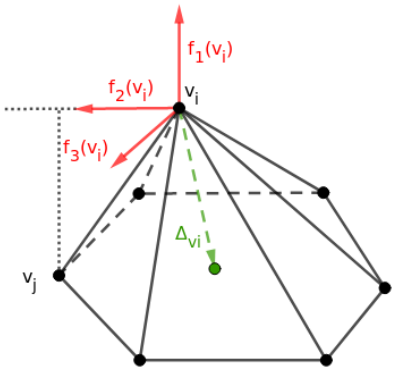


Figure 3.2: Local coordinate system (red) of Δ_i (green) at v_i

with $v_i \in V$ the vertices of mesh V, $N_1(v_i)$ the one ring neighbourhood of vertex v_i (*i.e.* the vertices directly neighbours to v_i), and $deg(v_i)$ the number of vertices in $N_1(v_i)$.

This matrix is used to compute the Laplacian offsets Δ_i of each vertex v_i, as:

$$\begin{pmatrix} \Delta_1 \\ \vdots \\ \Delta_n \end{pmatrix} = L \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \sum_{v_j \in N_1(v_1)} \frac{1}{deg(v_1)} v_j - v_1 \\ \vdots \\ \sum_{v_j \in N_1(v_n)} \frac{1}{deg(v_n)} v_j - v_n \end{pmatrix}, \tag{3.2}$$

These offsets are then made isometry invariant (and thus pose invariant) by expressing them in a local coordinate system for each vertex. This coordinate system is composed of the normal vector of the surface at vertex v_i (called $f_1(v_i)$), a projection of a fixed vertex neighbour of v_i in the orthogonal plane of the normal (called $f_2(v_i)$), and their cross product (called $f_3(v_i)$). The three vectors are normalized to create the local coordinate system at vertex v_i (see Figure 3.2).

This coordinate system is invariant to translation and rotation of the neighbourhood of v_i. The resulting local offsets are therefore invariant to isometric deformations. The local Laplacian offsets are thus geometrical properties of the mesh that are isometry invariant. Following the near isometry hypothesis, they can thus be used as a parameter of the identity.

3.2.2.2 Body Part Rigidity

Using the representation described in the previous section, we are able to encode and transfer local identity properties of the characters. However,

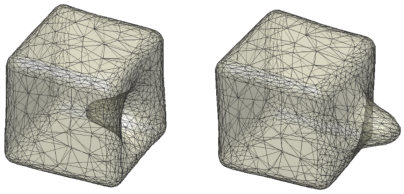


Figure 3.3: Two isometric 3D shapes with drastically different volumes. Figure extracted from [Cosmo et al., 2019]

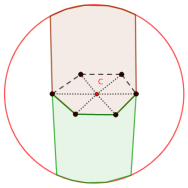
local isometry does not encode the global volume of the shape, and two isometric shapes can have drastically different volumes (see Figure 3.3). Local isometry is therefore not sufficient to encode the identity of a character.

The human body volume is not equally distributed among its body parts. Moreover, a same body part can have a proportionally very different volume between two different characters. For example, the thighs of a cyclist will be significantly bigger than the average person's, in proportion to the rest of the body. Global preservation of the body volume is thus an under-constrained problem, and can lead to unnatural distribution of the volume.

Another approach is to make the hypothesis that body parts of the human body deform near rigidly for a same character in different poses. By preserving the rigidity of each body part separately, one can enforce their volume to be consistent, while allowing for non-rigid deformation of the global body. While the volume of a body part does not stay perfectly constant during pose deformations, due to *e.g.* breathing or muscle deformations, these changes can be neglected and this approximation led to good result in the literature [Sorkine and Alexa, 2007, Jiang et al., 2020] and in our experiments (see Chapters 4 and 5).

In the remainder of this document, we refer to "body parts" as the segmentation of rigidly deforming body segments. We segmented our template in 17 body parts corresponding to the rigidly deforming parts of the human body(see Figure 3.4a). Some applications of this segmentation, such as volume computation, require a closed mesh. We thus close our body segments by computing the centroid of the seam between two segments and generating new triangles as shown in Figure 3.4b. This segmentation lead to convincing results in our experiments (see Chapters 4 and 5). Note however that this segmentation could be further refined depending on the detail level needed for specific applications. For instance, for an application involving precise hand movements, the hands could be segmented into smaller rigid parts, for each phalanx of the fingers.

(a) Body part segmentation



(b) The seam is closed by computing its centroid C and generating new triangles between C and the vertices located on the seam.

Figure 3.4: Body part segmentation on the template and close up of a seam.

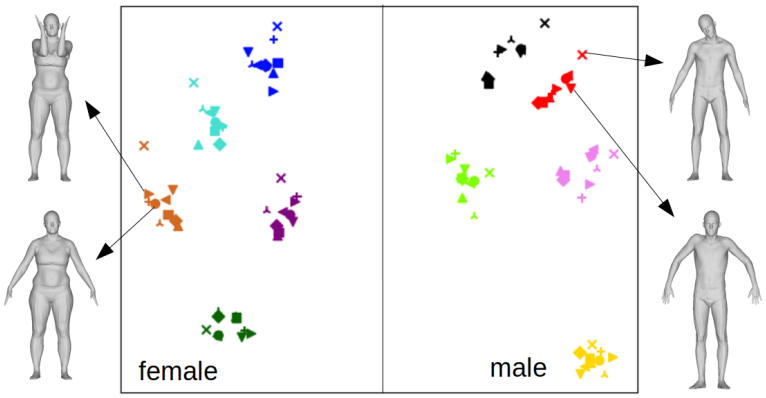
3.2.2.3 Hypotheses Validation

In this section we use the FAUST dataset [Bogo et al., 2014] to test the hypotheses presented above. This dataset contains 100 models of 10 identities performing the same 10 poses. This allows testing which geometric properties stays consistent across changes in identity or pose.

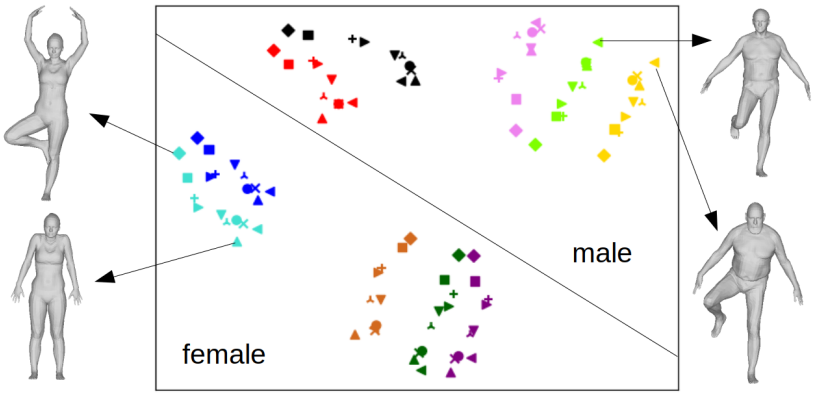
In order to test the hypotheses, we compute specific geometric properties for each 3D models of the FAUST dataset. We then apply the T-distributed Stochastic Neighbor Embedding (T-SNE) dimension reduction algorithm [Van der Maaten and Hinton, 2008] to these properties in order to express them in a two dimensional space. The T-SNE algorithm aims to preserve proximity between points: points that were close in the high input dimension of the algorithm should stay close in the output lower dimension.

We present the result of the T-SNE algorithm in figure 3.5. In this figure, we embed geometrical properties of the 100 models from FAUST in a two dimensional space. Points with the same color correspond to models with the same identity, and points with the same symbol correspond to models performing the same pose.

In Figure 3.5a we test the near-isometry hypothesis presented in Section 3.2.2.1. For each model, we compute the local Laplacian offsets of each vertex and store them in a $(n \times 3)$ length list, with $n = 6890$ the number of vertices of the FAUST models. We apply T-SNE to this list. We observe that resulting points are clearly grouped by color, thus by identity of the corresponding 3D model. This implies that 3D models of a human character with a specific identity have similar local Laplacian offsets. Therefore,



(a) Local laplacian coordinates



(b) Distances between vertices belonging to the same body part

Figure 3.5: T-SNE dimensionality reduction applied to local Laplacian (3.5a) and intra body part distances (3.5b). All parameters are computed on the FAUST dataset, containing 10 identities performing 10 poses. In each figure, marker colors indicate identities, and marker shapes indicate poses.

this tends to validate the near-isometry hypothesis.

In Figure 3.5b we explore the body part rigidity hypothesis. For each model, we use the segmentation of body parts presented in Figure 3.4. For each body part, we encode the Euclidean distance between each pair of vertices belonging to the body part. These distances should stay similar if the body part deforms rigidly. We apply T-SNE on the list of all these distances for each model. We observe that points corresponding to models with the same identity are roughly grouped by the T-SNE algorithm. This tends to validate the rigidity hypothesis.

It is important mentioning that while these hypotheses are reasonable, as demonstrated by the T-SNE results, they are approximations. As already discussed, joints of the human body can deform non-isometrically with changes in pose, and body parts can deform non-rigidly due to *e.g.* muscles contracting or breathing. However, these hypotheses lead to good results in our experiments, as can be seen in Chapters 4 and 5. They are therefore a promising first step in defining a more robust identity parameter, that is fully pose independent.

3.3 Self-contacts

As stated in Section 3.2.1, defining equivalences between poses is a difficult task. State-of-the-art methods rely on heuristic to define equivalences. A simple skeleton retargeting approach can for example consider that two skeletons with similar joint angles have the same pose, as the Euclidean distance between the joints of the two would be as low as possible. However, studies on human perception of pose similarity showed that a low Euclidean distance does not necessarily imply pose equivalence [Tang et al., 2008, Durupinar, 2021]. Moreover, as skeletons do not directly encode surface information, simply preserving joint angles can cause interpenetrations and loss of constraints such as foot/ground contacts.

Another popular constraint explored by methods that aim to adapt poses to the morphology of new characters is to encode and preserve relative positions of body parts. This approach is well illustrated by the interaction mesh proposed by [Ho et al., 2010]. In this representation, all joints of the skeleton of characters are linked by edges, and their relative positions are preserved by minimizing changes in the Laplacian coordinates of this interaction mesh. Similar approaches were applied to 3D meshes, by considering vertices on the surface of the characters instead of skeleton joints [Liu et al., 2018,Jin et al., 2018]. However, this approach tends to preserve exactly distances between body surfaces. This is efficient to avoid collisions and to preserve some important constraints such as contact, but can induce artifacts when transferring between very different morphologies. For example, when transferring a pose with close body surface interactions to a wider characters, some distances between surfaces must be reduced in order to adapt the pose to the new morphology.

With the identity transfer strategy, this problem is simplified; as the initial step of the algorithm is the source character already in the deformed pose, the general relative distances between body part is naturally respected. Deforming the surface of the body to transfer a new identity

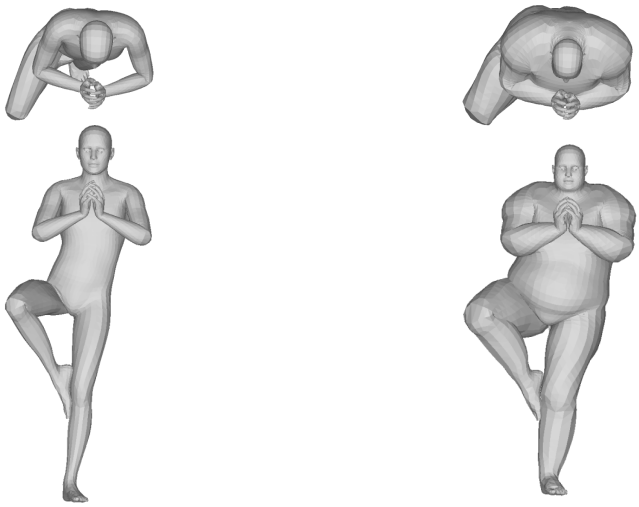


Figure 3.6: Characters with very different morphology performing a similar pose with different self-contacts, viewed from front and from top. Meshes generated by an artist, courtesy of [Liu et al., 2018]

should not significantly change these global spatial relationships. However, collisions might appear, and contacts between surfaces be lost, because of the body parts inflating or deflating with the new morphology. Even with identity transfer, it is thus necessary to apply slight pose corrections in order to avoid artifacts. We argue that simply correcting interpenetrations and preserving self-contacts that were present in the source is enough to preserve the pose of the source character in the result.

However, while this approach gives satisfying results, limitations remain. Indeed, preserving all contacts present in the source pose is not always pertinent. This is illustrated in Figure 3.6: In this figure, we present two characters with very different morphology performing the same pose. This pose contains several self-contacts. Some of them seem obviously important to the meaning of the pose, such as the contacts between the hands or between the foot and the leg. However, in the top view we can observe that the contact between the arms and the torso is only present in the right, wider character. We argue that this contact simply appears to adapt to the morphology of this character and does not bring meaning to the pose; we consider that the left character is in the same pose even though this contact disappeared. This example illustrates that while some self-contacts must be preserved in order to correctly mimic the pose, some do not have this importance. Preserving all self-contacts present in the source pose can thus create artifacts and significantly alter the pose.

Chapter 4

Contact Preserving Identity Transfer

4.1 Introduction

In Chapter 3 we compared two orthogonal strategies for the deformation transfer problem; the pose and identity transfer directions (see Section 3.2). We argued that identity transfer should result in simpler deformations than its counterpart, and should thus be able to generalize to complex poses. Identity transfer also allows to avoid the task of defining equivalences between poses, by considering the pose of the source character and simply adapting it to the new morphology. We argue that this adaptation only needs to preserve simple constraints, such as self-contacts (see Section 3.3) in order to preserve the contextual meaning of the pose. In this chapter, we inquire these assumptions by proposing an optimization-based identity transfer method with a simple pose adaptation step, that aims to preserve self-contacts and avoid interpenetrations.

Existing works on optimization-based deformation transfer can be sorted broadly in two main categories. First, skeletal deformation transfer aims to adapt the joint angles of the character in order to satisfy kinematic constraints either edited manually [Gleicher, 1998, Kulpa et al., 2005] or automatically built based on geometric constraints between body parts [Ho et al., 2010]. With these approaches, since only the skeleton of the character is animated, and not its body surface, it is difficult to prevent collisions or more generally respect distance constraints of the body surface. Second, surface-based deformation transfer considers surface deformations, typically mesh deformations, when transferring the pose of a source character to a target one [Sumner and Popović, 2004].

As we focus in this work on the handling of surface-to-surface contacts, we chose to apply our method directly at the surface level. Some works following this approach also explore how to detect and preserve distance constraints between body surfaces [Jin et al., 2018, Liu et al., 2018]. These methods make the assumption that distances between body segments should be preserved exactly during deformation transfer. While this successfully avoids interpenetrations and helps preserving the pose's meaning, this can result in artifacts. Indeed, these distances do not necessarily relate to the contextual meaning of the pose, but can result from intrinsic shape constraints, such as surface contacts due to corpulence. By simplifying this assumption and only preserving existing contacts, we aim to preserve the contextual meaning while allowing adaptations due to the morphology.

Given as input a source character in both a standard and a deformed pose and a target character in a standard pose, we propose to morph the surface of the source character to match its identity to the target character. During the deformation we avoid changes of the pose, simply correcting interpenetration artifacts and preserving contacts already present in the source. By starting from the correct pose and modifying it as little as possible, we consider that important constraints of the pose are naturally preserved in the result character. As a consequence, surface contacts due to shape differences in body sizes can be handled by design with our approach, as we demonstrate in our results (see Section 4.4). To allow for motion retargeting, continuity between subsequent poses of a motion sequence is encouraged in a post-process. We show experimentally that our approach can be used to transfer identities between a wide range of characters, in any class of shape, *e.g.* human or animal characters.

4.2 Related Work

In this section, we briefly review works that aim to detect and preserve contacts involving the body surface of characters. A complete review of these methods is beyond the scope of this chapter, and we simply aim to build intuition on contact aware 3D character modeling. For more details on methods specifically applied to the deformation transfer problem, such as [Liu et al., 2018, Jin et al., 2018, Villegas et al., 2021], we refer the reader to Chapter 2.

Humans mainly interact with their environment using their hands. Reconstructing and animating accurate 3D models of hands is thus an important step of 3D animation. A lot of works have specifically explored how to take into account (self-)contacts and collisions in 3D hand models.

For example, [Ballan et al., 2012] reconstruct interacting 3D hand models from 2D images, and focus on avoiding surface interpenetrations. They detect contacts with bounding volumes hierarchy [Teschner et al., 2005], and propose contact terms that penalizes interpenetrations. Other works aim to reconstruct a hand grabbing an object [Tzionas et al., 2016,Sridhar et al., 2016,Hasson et al., 2019]. These works propose to combine attraction and repulsion terms in their optimization in order to avoid collisions while preserving the contact between the object and the hand.

Closer to our work, interest has been given to reconstructing full body human 3D models that take (self-)contacts into account. [Pavlakos et al., 2019] propose SMPL-X. This work aims to include expressive faces and hand models in the statistical body model SMPL [Loper et al., 2015], and to avoid collisions between body surfaces. Similarly to previously cited works on hand models, they use bounding volumes hierarchy and penalize detected interpenetrations. Several works have built on SMPL-X for contact aware applications: POSA [Hassan et al., 2021] extends this model by encoding the contacts between the body surface and the environment, and [Müller et al., 2021] use SMPL-X to learn to reconstruct meshes from 2D images of poses with self-contacts. Other works have explored how to account for contacts and collisions in animations of physics-based skinned characters [McAdams et al., 2011,Kadleček et al., 2016], but are computationally expensive. More recently, [Komaritzan and Botsch, 2019] proposed a similar method using efficient contact detection approach [Teschner et al., 2003] and obtained real time performances.

In the method presented in this chapter, we use tools from these works to detect and handle contacts during identity transfer. We use the efficient collision detection from [Teschner et al., 2003], and a contact energy based on repulsion and attraction terms inspired by [Hasson et al., 2019].

4.3 Method

4.3.1 Overview

Our goal is to make a target character reproduce the motion of a source character. To this aim we consider as inputs: the source character mesh in a *standard pose* (e.g. A-pose), the same source character in a flow of *deformed poses* we wish to duplicate with a target character, and the target character mesh in the standard pose. In this chapter, we consider the motion as a continuous sequence of static poses. Consequently, for each pose of the source character, the deformation transfer process should compute a de-

formed pose adapted to the target character, while preserving continuity in the resulting sequence. All the source and target meshes are assumed to be in correspondence through a single mesh graph. We segment this common mesh topology in a number of body parts, as discussed in Section 3.2.2.2 and Figure 3.4. Note that our method does not require a rigged skeleton.

We proceed by first transferring the target identity to each deformed input pose independently, and by subsequently smoothing the resulting animation to encourage continuity in a post-process. For each static frame, starting from the source character in the deformed pose, our approach morphs its surface until its identity fits the target character identity, while preserving the surface contacts present in the source deformed pose, and correcting interpenetration artifacts. This way, we transfer identities at the desired poses instead of transferring poses to the desired identities, as is traditionally done in the existing works.

Input meshes are defined as $\mathcal{V} = (\mathbf{V}, E, F)$ (see Section 3.1). We define the rigid body part transformations $\mathbf{\Theta} = \{R_P\}_{P \in BP}$, where R_P is the rotation associated with the body part $P \in BP$ and BP is the set of body parts. To perform identity transfer, we cast the problem as an optimization over the vertex positions $\mathbf{V}$ and the rigid transformations $\mathbf{\Theta}$, and with respect to three energy terms that account for local and more global identity properties as well as surface contacts:

$$\underset{\mathbf{V},\mathbf{\Theta}}{\operatorname{argmin}} \; [\gamma_{iso} E_{iso}(\mathbf{V}) + \gamma_{vol} E_{vol}(\mathbf{V}) + \gamma_C E_C(\mathbf{\Theta})]. \tag{4.1}$$

The terms E_{iso} and E_{vol} penalize the discrepancy in isometry and volume with the target character, and are minimized w.r.t. $\mathbf{V}$. This creates non-isometric deformations that change the identity of the character to the target identity. The term E_C penalizes collisions of surfaces and loss of contacts present in the source pose, and is minimized w.r.t. $\mathbf{\Theta}$. This term induces near-isometric and rigid deformations of the body parts, and thus should deform the pose of the character without impacting its identity. The weights γ_{iso}, γ_{vol}, and γ_C modulate the influence of each energy term.

To facilitate the identity transfer, before optimizing Expression 4.1, we compute the height of the source and target characters using their provided standard poses, and pre-scale the deformed source mesh to the height of the target mesh.

4.3.2 Identity and Pose Optimization

4.3.2.1 Local Isometry

Our first energy term E_{iso} decreases if the input and target mesh are near-isometric. We build it on the isometry invariant feature inspired by [Wuhrer et al., 2012] and presented in Section 3.2.2.1, the local Laplacian coordinates Δ_i (see Equation 3.2).

In a preliminary step of the algorithm, we compute the target shape representation $\Omega^T = \{\omega_1^T(v_i^T), \omega_2^T(v_i^T), \omega_3^T(v_i^T)\}_{v_i^T \in V^T}$ at each vertex v_i of the target character in the standard pose, where $\omega_k^T(v_i^T)$ are the target's Laplacian coordinates Δ_i^T expressed in the their local coordinate system (f_1^T, f_2^T, f_3^T) (see Figure 3.2).

During optimization, we express back Ω^T in canonical coordinates using the local coordinate systems of the source mesh $\mathbf{V}$. We obtain the target Laplacian offsets $\Delta^{T\prime}$. Our isometry energy is then expressed as:

$$E_{iso}(\mathbf{V}) = \sum_{v_i \in V} \left(\Delta_i - \Delta_i^{T\prime}\right)^2, \tag{4.2}$$

with $\Delta_i^{T\prime} = \omega_1^T(v_i^T) f_1(v_i) + \omega_2^T(v_i^T) f_2(v_i) + \omega_3^T(v_i^T) f_3(v_i)$.

A direct minimization of the above term results in a non-linear and complex optimization. In practice, given a fixed local configuration, i.e. $(f_1(v_i), f_2(v_i), f_3(v_i), v_j \in N_1(v_i))$, around v_i, Expression 4.2 is minimized by moving v_i towards the optimal position $\hat{v}_i$:

$$\hat{v}_i = \sum_{v_j \in N_1(v_i)} \frac{v_j}{deg(v_i)} - \big(\omega_1^T(v_i^T) f_1(v_i) + \omega_2^T(v_i^T) f_2(v_i) + \omega_3^T(v_i^T) f_3(v_i)\big). \tag{4.3}$$

Hence, we proceed iteratively in two steps: (i) vertices are moved in the optimal direction $\hat{v}_i - v_i$; (ii) local configurations are re-estimated. Details on the iterative solving are given in section 4.3.3.

4.3.2.2 Body Part Volume

With the energy presented in the previous section, we enforce our result and the target identity to be locally near-isometric. However, as stated in section 3.2.2.2, local isometry is not enough to encode the identity since the global volume of the shape is not preserved. In order to remedy this issue, we add an energy term to our optimization, that constrain corresponding

body parts between our result and the target character to have a similar volume.

We make use of the body part segmentation presented in Figure 3.4. For each body part, computing its volume requires it to be a closed mesh. To this purpose, body parts are closed by computing the centroid of the seam between two neighboring body parts, and by generating triangles between this centroid and the vertices on the seam (see Figure 3.4b).

With closed body part meshes, their volume can be computed as the sum of the signed volumes of the tetrahedrons formed by the body part's triangle faces and the origin O [Zhang and Chen, 2001]. Let $\{v_i, v_j, v_k\}$ be an oriented triangle and O the origin, the signed volume of the tetrahedron $\{O, v_i, v_j, v_k\}$ writes:

$$\begin{aligned} V_{Oijk} =& \frac{1}{6}(-x_k y_j z_i + x_j y_k z_i + x_k y_i z_j \\ & - x_i y_k z_j - x_j y_i z_k + x_i y_j z_k), \end{aligned} \tag{4.4}$$

where (x_i, y_i, z_i) are the 3D coordinates of v_i. The volume of a body part P is then $V_P = \sum_{T \in P} V_{O,T}$, where $T \in P$ denotes the triangles composing body part P.

Given the body part segmentation, the volume energy term measures the discrepancy between body part volumes on the deformed source shape and on the target shape:

$$E_{Vol}(\mathbf{V}) = \sum_{P \in BP} (V_P - V_P^T)^2. \tag{4.5}$$

with BP the set of body parts of the meshes, V_P (respectively V_P^T) the volume of body part P for the input mesh (respectively the target mesh).

4.3.2.3 Contacts

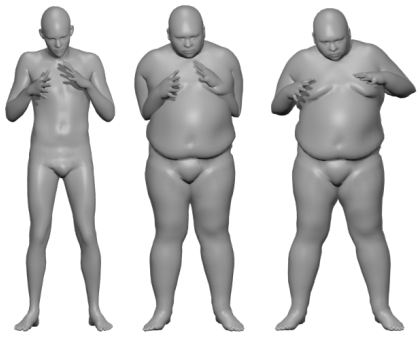
The isometry and volume terms previously presented help to accurately deform the source character to match its identity to the target. However, as discussed in Section 3.3, differences in morphology can make it difficult for the target character to correctly reproduce the source pose. Only transforming the identity of a character while maintaining the pose strictly similar is likely to create artifacts such as interpenetrations or loss of contextually important contacts.

To address this issue, our method includes a contact energy term. This term aims to maintain all contacts present in the source pose, while not introducing interpenetrations. Our contact term is built on the contact loss

presented in [Hasson et al., 2019]. It is composed of a repulsion term that increases when surface inter-penetrations occur, and an attraction term that increases when a contact present in the source pose is lost:

$$E_C(\mathbf{V}) = \gamma_r E_r(\mathbf{V}) + \gamma_a E_a(\mathbf{V}), \tag{4.6}$$

where E_r and E_a are the repulsion and attraction term respectively, with associated weights γ_r and γ_a.



(a) Left to right: source pose, transfer result without the contribution of a repulsion term, and with such a contribution (target from Figure 4.3c). Notice on the right the arms that do not penetrate the torso anymore, and the wider gap between legs to avoid thighs colliding

(b) Left to right: source pose, transfer result without the contribution of an attraction term, and with such a contribution (target from Figure 4.4b). Notice on the right the foot that does not penetrate the leg thanks to the repulsion term, but still remains in contact thanks to the attraction term

Figure 4.1: Contribution of the contact energy terms.

Repulsion Term Our body part segmentation (see Section 3.2.2.2) enables us to follow the rigid members of the human body during the defor-

mation. As such, if the source and target characters have correct poses, no interpenetration should appear inside a given body part. We thus test only interpenetrations between a vertex and all body parts but the one it belongs to. The repulsion term also considers collisions with the ground and is defined as

$$E_r(\mathbf{V}) = \gamma_r \sum_{P \in BP} \sum_{\substack{v_i \in \mathbf{V} \setminus P \\ v_i \in Int(P)}} d(v_i, P)^2 + \gamma_{rg} \sum_{\substack{v_i \in \mathbf{V} \\ v_i \in Int(G)}} d(v_i, G)^2, \tag{4.7}$$

where BP is the body part set, G is the ground, $Int(X)$ is the interior of object X, and $d(v, X)$ is the minimum distance between the vertex v and the object X; $d(v, X) = \inf_{w \in X} \|v - w\|_2$. The effect of the repulsion term is illustrated in Figure 4.1a.

Detecting interpenetrations in a mesh is a computationally heavy task. In this work, we use the point-tetrahedron collision test with spatial hashing method described in [Teschner et al., 2003]. Given a tetrahedral mesh, this method defines a hash-function that maps every object (vertices and tetrahedrons) to a 1D index. The function is designed such that objects mapped to the same index are the ones located in the same region in 3D space, and must be tested for collision. This allows to significantly reduce the number of collision tests to be performed, and has been applied in real-time animation pipelines [Komaritzan and Botsch, 2019].

Employing this collision test requires a tetrahedral mesh. We use the method presented in Section 4.3.2.2 to close the body parts' meshes. This method requires convex objects, which is the case for most of our segmented body parts. However, hands and feet are not convex, due to fingers and toes, and our approximation could consequently lead to important errors. In our experiments, we mostly use poses from SMPL that do not encode the movement of the fingers or toes, so the approximation did not generate artifacts. If more detailed finger or toe poses are required, it is possible to add each phalanx of the fingers and toes to the body part segmentation to make the approximation more robust, as stated in Section 3.2.2.2.

Attraction Term As stated in Section 3.3, interactions between body segments, such as contacts, can be important to define the meaning of the pose. In the work presented in this chapter, we choose to preserve all contacts present in the source pose in our result, in order to avoid losing meaningful information about the pose.

In a preliminary step, we encode those contacts in the source. We define a contact threshold proportional to the height of the character. Vertices

that are under this threshold distance from a surface are considered in contact with the surface. Similarly to interpenetrations, we consider that no important contacts should appear inside a given body part, and thus only encode contacts between different body parts. For each vertex under the contact threshold distance of a surface, we encode the contact as the couple of the vertex and its closest vertex on the surface. The attraction term also forces the vertices at ground level to stay at ground level in the result. However, we don't enforce a precise position on the ground as long as the vertex is at ground level; this allows for example moving the legs of the character to adapt to a new morphology.

The attraction term increases when the distance between vertices in contact in the source exceeds the fixed contact threshold as:

$$\begin{aligned} E_a(\mathbf{V}) =& \gamma_a \sum_{(v_i,v_j)\in C} \max[(d(v_i,v_j)-T),0]^2 \\ &+\gamma_{ag} \sum_{v_i \in C_G} \max[(d(v_i,G)-T),0]^2, \end{aligned} \tag{4.8}$$

where C is the set of pairs of vertices in contact, C_G are the vertices in contact with the ground, and T is the contact threshold. The effect of the attraction term is illustrated in Figure 4.1b.

Rigid Formulation The goal of the contact energy (Equation 4.6) is to adapt the source pose to the morphology of the target character. As such, it must slightly deform the pose, while keeping the identity constant. As discussed in Section 3.2.2.2, rigid deformations of the body parts of the character change the pose but not the identity. Hence, we want to minimize Equation 4.6 w.r.t. rigid body part deformations, in order to preserve the identity.

To do so, we use the same body part segmentation as in previous sections. Body parts are ordered in a tree hierarchy, with the crotch as the root. We then define a rotation for each body part $\Theta = \{R_P\}_{P\in BP}$. These rotations are applied to a body part and its children, around a "joint" defined as the centroid of the seam between the body part and its parent. The root body part rotates around its centroid.

By minimizing the contact energy w.r.t. these rotations, each body part deforms rigidly. The contact energy becomes:

$$E_C(\mathbf{V}(\mathbf{\Theta})) = \gamma_r E_r(\mathbf{V}(\mathbf{\Theta})) + \gamma_a E_a(\mathbf{V}(\mathbf{\Theta})), \tag{4.9}$$

using the mesh vertex positions $\mathbf{V}$ as functions of the rotations $\mathbf{\Theta}$.

4.3.3 Iterative Solving

Optimizing the full sum of energies in Expression 4.1 appears difficult in practice since the isometry term $E_{iso}(\mathbf{V})$ (see Expression 4.2) is non-linear. This results from the fact that the differential coordinates ω_i that encode the identity are expressed in a local coordinate system, which depends on the position of the vertices of the mesh. Hence moving a vertex also transforms its local frame. We therefore minimize Expression 4.1 iteratively. In a first step vertices $\mathbf{V}$ are moved with respect to the target identity information, then the pose is optimized using the rotations $\mathbf{\Theta}$ in order to satisfy the contact constraints and finally local frames are re-estimated. This is iterated until the absolute difference in the sum of energies between two successive iterations is below a threshold. The first two steps are detailed below.

The first step aims to optimize the isometry term $E_{iso}(\mathbf{V})$ and the volume preservation term $E_{vol}(\mathbf{V})$, Expressions 4.2 and 4.5 respectively. To this purpose vertices are moved in a direction that accounts for both terms:

$$v_i' = v_i + \epsilon(\gamma_i d_i(v_i) + \gamma_v d_v(v_i)), \tag{4.10}$$

with v_i' the new position of v_i, γ_i and γ_v the weights associated to the directions d_i and d_v, respectively, and ϵ a displacement offset function. The isometry direction d_i is the direction towards the optimal position as defined in Equation 4.3. The volume direction d_v is computed based on the Stokes' Theorem and its resulting divergence theorem. That is, d_v is the direction of the normal n_i of the surface at vertex v_i, and the offset by which we move v_i is the difference in volume of the body part containing v_i between the target and the current shape. This leads to $d_v(v_i) = (V_P^T - V_P)\ n_i$.

The second step of the iterative framework aims to minimize the contact energy defined in Expression 4.9. Auto-differentiation is used to obtain the gradient of the contact energy E_C w.r.t. rotations of the body parts $\mathbf{\Theta}$. We then apply a gradient descent iteration to the rotations. Since the deformation at each iteration is relatively small, this slight correction is enough.

4.3.4 Adaptation to Motion Sequences

The approach presented in the previous sections enables to transfer the identity of a target character to a source character in a given static pose. When considering motion sequences, we apply a per-frame strategy and transfer the target identity to each frame of the sequence independently.

This strategy has proven to result in realistic animations when no pose corrections are necessary during the identity transfer for any frame of the se-

quence. However, when pose corrections are necessary, artifacts can appear in the resulting animation, such as foot-skating and jitter. In this section, we describe these artifacts, and present solutions to adapt our method to tackle this issue.

4.3.4.1 Temporally Consistent Ground Contact

The pose correction applied to shapes can give rise to inconsistent ground contacts. For example, when transferring motion from a skinny character to a corpulent one, the gap between the legs can be widened to avoid potential collisions between the thighs. When such a correction occurs, the feet positions can deviate in consecutive frames, leading to the so-called "foot-skating" artifacts in the resulting animation.

To fix this problem, we modify the attraction term in Equation 4.8. For static data, ground contacts are enforced by constraining the concerned vertex heights to be at the ground level. Although this avoids collision or loss of contact between the foot and the ground, it does not guarantee the foot to remain at a fixed position, hence yielding foot-skating artifacts. For a continuous sequence of poses, we store the ground contact position of the source pose at each frame and compare it with the previous frame. When detecting a ground contact that was already a ground contact in the previous frame, we consider that the associated vertex should remain at the same position. Therefore, the ground contact term for such a vertex v_i becomes: $\max[(d(v_i^t, v_i^{t-1}) - T), 0]^2$ where t denotes the current frame time within the motion sequence, and v_i^t (respectively v_i^{t-1}) corresponds to the vertex v_i at the frame time t (respectively $t-1$).

4.3.4.2 Animation Smoothing

Independent pose corrections on consecutive frames can result in large and irregular movement of body parts, resulting in jittery and unrealistic animations. In the remainder of this section, we describe several post-processing strategies that we applied to smooth the resulting animation. We compare the effectiveness of each approach in Section 4.4.5.

Low Pass Filtering The first solution we explored is to remove the high-frequency displacements of each vertex of the 3D models, by applying a simple low pass filter to the animation. As the irregular and rapid movements that cause the jitter are contained in these high frequencies, this filtering should help smooth the animations.

In practice, we apply the filter using a five-frame rolling average to each vertex of the model.

Discrete Cosine Transform [Akhter et al., 2010] show that the PCA basis learned from human motion sequences converges towards the basis of the Discrete Cosine Transform (DCT). This property was used to combine a DCT basis for temporal data with a spatial shape basis computed using PCA to create a model encoding spatiotemporal data [Akhter et al., 2012]. This model allows to globally smooth motion sequences by removing first the high-frequencies of the temporal DCT basis, and second the basis vectors of the PCA shape space corresponding to small eigenvalues. This approach has been successfully applied to smoothing of point trajectories, *e.g.* facial mesh animations [Abrevaya et al., 2018b].

In our experiments, we apply this approach to retargeted animation sequences. Since the sequences we consider are relatively short, we only project over the DCT basis and remove high frequencies.

De Boor Spline Approximation The last approach we explored smooths locally the retargeted animation by post-processing the trajectory of each mesh vertex independently. To this end, each trajectory is approximated by a spline [Craven and Wahba, 1978], that have shown to be effective in trajectory smoothing [Egerstedt and Martin, 2001]. We use the original de Boor algorithm [De Boor, 1978] to approximate vertex trajectories with a regularization term based on the curve's second derivatives.

4.3.5 Implementation Details

We implemented our algorithm in Julia, and use a Python implementation [Prilepin, 2020] of de Boor's smoothing algorithm to post-process the trajectory of each mesh vertex.

4.3.5.1 Parameter Settings

Our method has a number of parameters that need to be adjusted. The weights of the different energy terms in Equation 4.1 are set to $\gamma_{Shape} = \gamma_{Vol} = \gamma_C = 1$, the parameters weighing the influence of the contact and repulsion terms in the contact energy of Equation 4.6 are set to $\gamma_r = \gamma_a = 1$, and the weights handling ground contact in Equations 4.7 and 4.8, respectively, are set to $\gamma_{rg} = \gamma_{ag} = 0.1$. The offset weight ϵ in Equation 4.10 is set to 0.3, and the parameter p employed during spline smoothing to 0.1.

4.3.5.2 Computation Times

The computation times of our method are highly dependent on the surface interactions present in the pose transfer, *i.e.* contacts and possibly colliding surfaces. This is due to the high computational cost of the contact energy.

The computation times reported in this section correspond to the method ran for static poses, on a PC with an Intel Xeon E5-2623 v3s and 32GB of RAM. When the deformed pose is free of any body-to-body interactions, the method requires around 5 minutes. In the example shown in Figure 4.3, some corrections are needed due to body-to-body surface collisions, and our method takes around 15 minutes to run. In the example shown in Figure 4.10b, contacts in the deformed pose of the source need to be maintained, and our method takes around 20 minutes to run. Finally, when the method needs to both maintain contacts in the deformed pose and avoid surface collision, such as in Figure 4.10c, it runs in about 24 minutes. All experiments applied on static animals models based on SMAL take less than 5 minutes for the transfer, due to the lower resolution of the meshes.

These computation times are approximate, and can change depending on the hardware running the method, and on the resolution of the input meshes.

4.4 Evaluation

In this section, we present results of our method. We discuss its strengths and limitations, and give an overview of interesting future improvements. We first introduce the data used in our evaluations (Section 4.4.1) and give the implementation details (Section 4.3.5). We illustrate results of our method and its generality by applying it to different categories of shapes, such as minimally and casually dressed humans, and wild animals (Sections 4.4.2, 4.4.3 and 4.4.4). We present result of our adaptation to motion sequences in Section 4.4.5. We then qualitatively compare our results to state-of-the-art works in Section 4.4.6.

4.4.1 Data

To demonstrate the generality of our approach, we evaluate our method on two different shape classes. First, human characters, both in a minimally dressed scenario and in a casually dressed one. This class is the most commonly considered in retargeting applications, and the minimally dressed scenario allows in particular comparisons to the state of the art. The second

class of shapes we consider are wild animals. Wild animal shapes are interesting as they can exhibit very different morphologies while still respecting our assumption of near-isometric deformations during motion.

As input, we require source and target character meshes in correspondence. For humans, the correspondence is established using the SMPL template (6890 vertices and 13776 faces) [Loper et al., 2015]. This template is segmented into the 17 body parts shown in Figure 3.4. For animal models, the correspondence is established using the SMAL template (3889 vertices and 7774 faces) [Zuffi et al., 2017]. We segment this template into 24 body parts as shown in Figure 4.2.

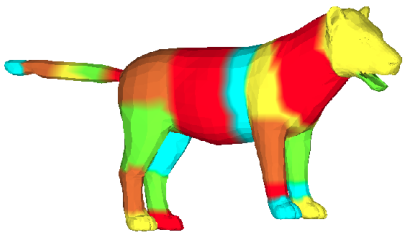


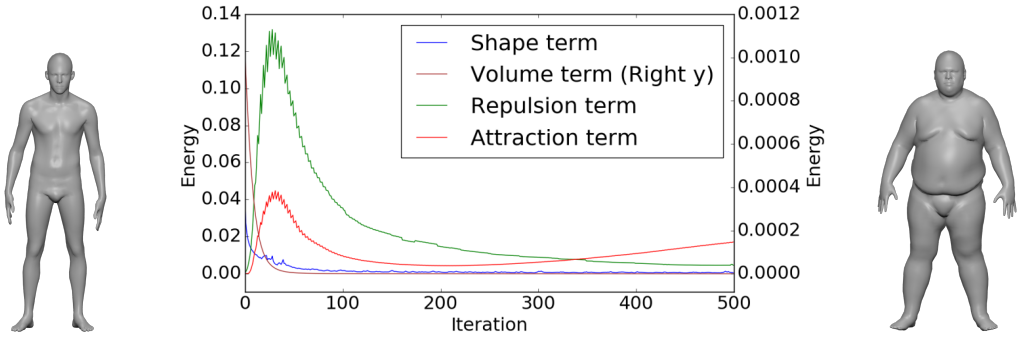
Figure 4.2: Body part segmentation of the SMAL animal body template [Zuffi et al., 2017]

For minimally dressed human characters, we use the example animations provided with SMPL, Faust [Bogo et al., 2014], Dyna [Pons-Moll et al., 2015], and models from Liu et al. [Liu et al., 2018] fitted to the SMPL template. For dressed humans we use meshes from 3D Poses in the Wild [von Marcard et al., 2018] that are already fitted to the SMPL template. For animal models, we create different poses and shapes using the statistical model SMAL.

4.4.2 Minimally Dressed Humans

This section discusses the convergence behaviour of our method, and shows qualitative results for minimally dressed human models.

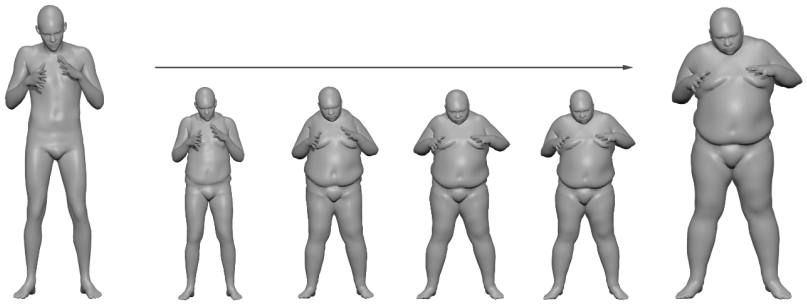
Figure 4.3 illustrates the iterative process of our method and shows intermediate results (4.3d). The identity and volume evolve quickly to match those of the target, while the contact term avoids interpenetration here by widening the gap between the legs and raising the arms. Figure 4.3b shows the evolution of each energy term during this transfer. Note that the identity fidelity (Eq. 4.2) and volume preservation (Eq. 4.5) terms decrease rapidly in the first iterations. The initial spike of the repulsion term (Eq. 4.7) is due to interpenetrations that appear as the morphology



(a) standard source

(b) Evolution of the energy terms during the iterative minimization

(c) standard target



(d) Left to right: source pose, result after 5, 40, 150 and 400 iterations, final result (500 iterations)

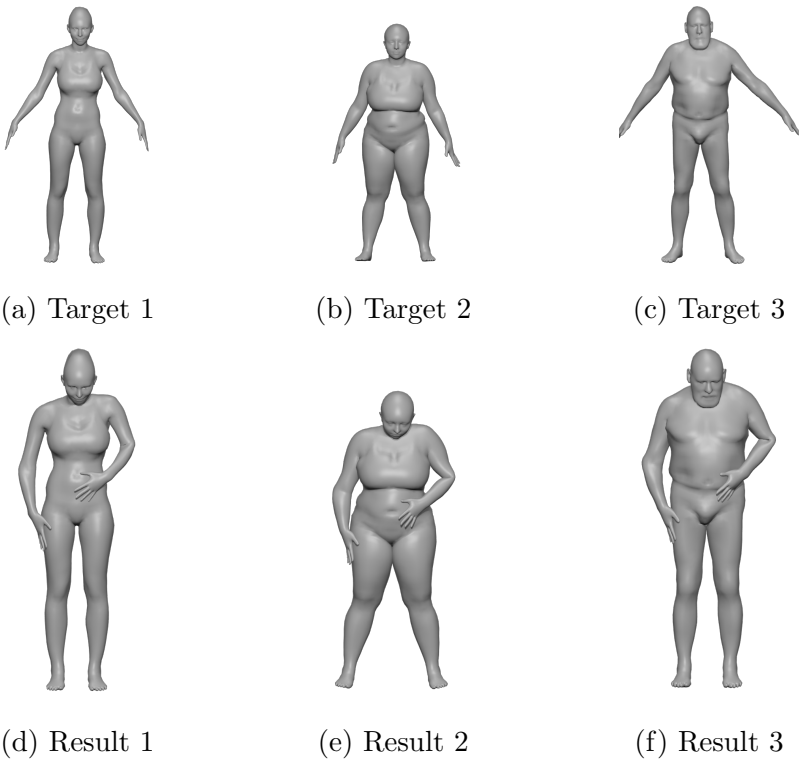
Figure 4.3: Evolution of the identity transfer from a thin to a larger character through the iterations.

changes. The correction of interpenetrations causes loss of contacts around the armpits, explaining the slight increase in the attraction term (Eq. 4.8). Our iterative process efficiently minimizes the identity and volume energies, while maintaining the contact energy at a reasonable level.

Figure 4.4 shows results of identity transfer from the source of Figure 4.8 (left) to characters with varying morphology. Notice the evolution of the space between the arms and the torso depending on the morphology of the target; while this gap is marked for skinny characters, it is much narrower or even non-existent for larger bodies. Our method thus adapts the pose to the different morphologies of the target characters.

4.4.3 Casually Dressed Humans

In this section, we show results of our method applied to human characters with casual clothing. Figure 4.5 shows three frames obtained when



(a) Target 1 (b) Target 2 (c) Target 3

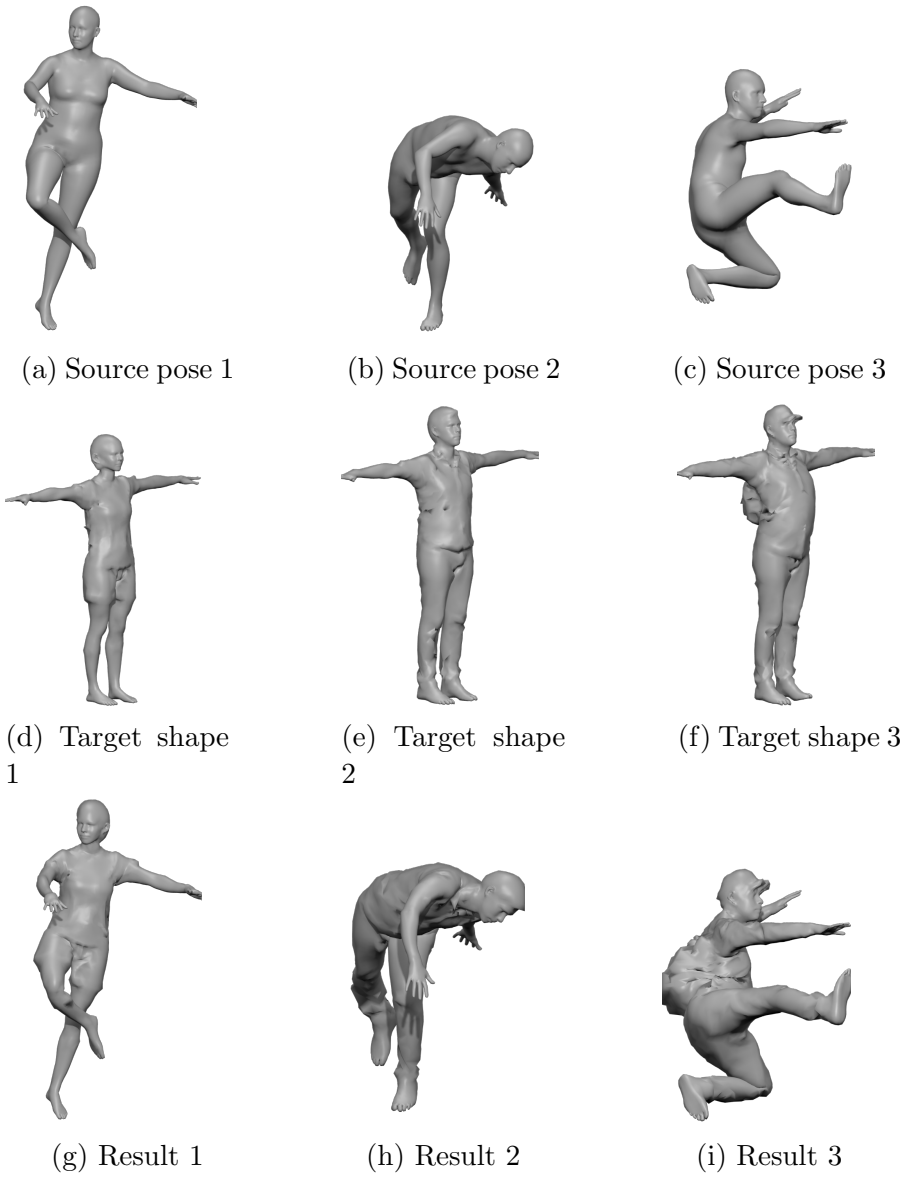
(d) Result 1 (e) Result 2 (f) Result 3

Figure 4.4: Identity transfer results on several characters of the deformed source pose shown in Figure 4.8 (left)

transferring the identity of clothed characters from the 3DPW dataset to new poses. In all three results, the cloth details, including wrinkles present in the standard pose of the target, are transferred to the deformed pose. Furthermore, the method can transfer hair, shown in Figures 4.5g and 4.5h, and even accessories such as the backpack and baseball cap in Figure 4.5i.

4.4.4 Animals

This section further illustrate the generalisation capacity of our method, by applying it to a new complex category of shapes; wild animals (see Figure 4.6). Figure 4.6f shows results of transferring the identity of a lion and a hippopotamus to the pose of a fox. Despite important differences in the morphology and volume distributions among body parts for the different animals, the resulting models are plausible overall. Note that the characteristics of the heads, trunk and legs are maintained in the resulting model for the lion and the hippopotamus. However, some artifacts occur for body parts with smaller volume, such as the tail of the hippopotamus, which is



(a) Source pose 1 (b) Source pose 2 (c) Source pose 3

(d) Target shape 1 (e) Target shape 2 (f) Target shape 3

(g) Result 1 (h) Result 2 (i) Result 3

Figure 4.5: Results of the method from sample poses of SMPL to clothed characters of 3DPW.

elongated after the transfer. The reason for this is that we match the volume of the body parts to the target, but not their lengths: the tail of the fox has its volume shrunk to match that of the hippopotamus, but keeps roughly the same length.

Figure 4.6k shows results of transferring the identities of a lion, a fox and

(a) Source

(b) Target 1

(c) Target 2

(d) Result 1

(e) Result 2

(f) Results of identity transfer to a running fox pose

(g) Source pose

(h) Result for the target shown in Figure 4.6b

(i) Result for the target shown in Figure 4.6a

(j) Result for the target shown in Figure 4.6c

(k) Results of identity transfer paw licking feline pose

Figure 4.6: Identity transfer results on animal models taken from SMAL

a hippopotamus to a pose with contacts from a feline. The results show that the method is able to preserve contact constraints even for animals with significantly different morphologies.

This experiment shows that our method can be applied to transfer identity between very different shapes, such as animals of different species. It only needs the source and target models to have the same general morphology; number and length of the limbs.

4.4.5 Animations

This section compares the different strategies to adapt our method to continuous motion sequences, as introduced in Section 4.3.4.2, and shows some qualitative results.

	head displacement (cm)		arm volume (dm^3)	
	mean	std.	mean	std.
source animation	0.345	0.220	/	/
target character	/	/	2.315	/
no post-process	1.094	0.900	2.221	0.101
simple post-process	0.616	0.398	2.166	0.150
50% DCT	0.835	0.605	2.220	0.104
25% DCT	0.718	0.458	2.215	0.123
10% DCT	0.410	0.260	2.147	0.440
5% DCT	0.237	0.148	1.913	0.551
spline smoothing	**0.553**	**0.343**	**2.175**	**0.176**

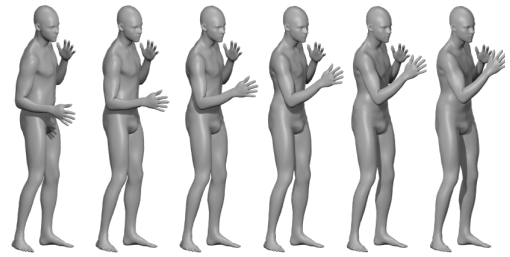
Table 4.1: Comparisons between different smoothing approaches. Mean and standard deviation of the displacement of a vertex on the middle of the forehead between two consecutive frames, and of the volume of the right forearm, evaluated for the motion sequence in Figure 4.7.

Table 4.1 provides quantitative measurements over the motion sequence obtained with the input from Figure 4.7. We measured (1) the displacement of a vertex located in the middle of the forehead between two consecutive frames (mean and standard deviation), and (2) the volume of the right forearm during the animation (mean and standard deviation). For a correct retargeting, we expect the mean and standard deviation of the displacement of the vertex on the forehead to be similar to the ones in the source animation. A higher standard deviation indicates large irregular displacements, and thus jitter. A lower standard deviation indicates that the animation has been over-smoothed and lost movements present in the source. The volume of the forearm should be close to its volume in the target character. It should also stay relatively constant during the animation, which should be shown by a low standard deviation of this measure.

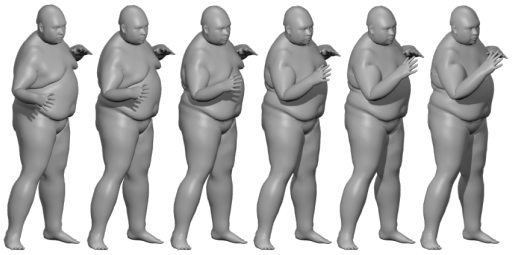
We computed these measures for the resulting animation before post-processing, and with the smoothing approaches described in Section 4.3.4.2: the low-pass filter, the DCT smoothing method, with different percentages of low frequencies retained, and the spline smoothing method.

Without post-processing, the vertex on the forehead performs large motions that lead to jittering artifacts. While the low pass filter reduces the jitter, it is still perceptible on the resulting animation. When smoothing with DCT, keeping a high percentage of components does not remove the jitter on the head. When keeping a low percentage of components, however, artifacts such as volume shrinking on the arm appear, as can be seen by the in-

creasing standard deviation of the arm volume. In contrast, the spline-based smoothing leads to a result without apparent jitter while preserving the arm volume. Figure 4.7 shows the corresponding motion retargeting result. We thus chose to smooth our animation results using the splined-based method. We refer the reader to the supplementary material found at this address for better visualization of these results: https://hal.archives-ouvertes.fr/hal-02613783/file/ContactPreservingShapeTransfer_supplementary.mp4.



(a) Source animation



(b) Result with spline smoothing in post-processing

Figure 4.7: Result of transferring a punching animation to the target in Figure 4.3c, using the spline smoothing in post-processing

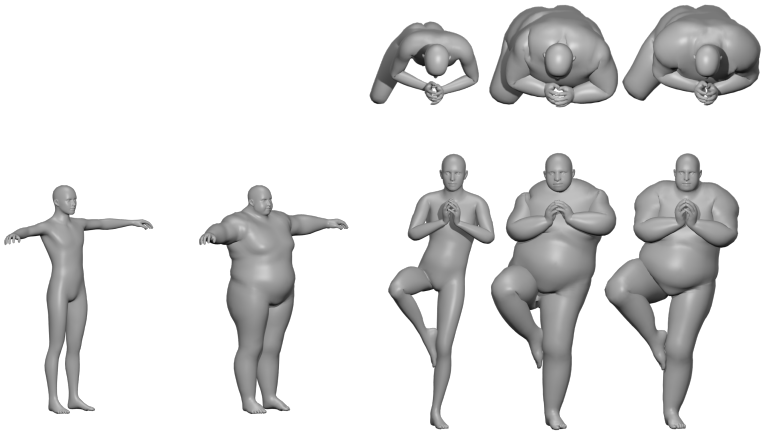
4.4.6 Comparisons

In this section, we compare results of our method with previous works from the literature. First, our results are compared to those of a skeleton-based approach where joint angles are directly applied to a new character. Second, we applied our method to character meshes used in previous surface mesh retargeting methods, namely context graphs [Liu et al., 2018] and AuraMesh [Jin et al., 2018], and compare our results to those obtained in these previous works.

Figure 4.8: Comparison to a skeleton retargeting baseline. Left: The source deformed pose generated by manually tuning SMPL shape and pose parameters. Center: The same SMPL pose parameters applied to new shape parameters. Right: The result with our method

Figure 4.8 compares our method to a skeleton retargeting baseline. The source pose (left) was generated by hand-tuning SMPL pose and shape parameters. Applying the same pose parameters to a character with different morphology leads the result of the baseline shown in the center. This straightforward approach leads to artifacts: the left hand enters the belly, and the contact between the right hand and the hip is incorrect. The result of our method is presented in the right of the figure. The artifacts reported with the baseline do not occur. Moreover, notice that the space between the arms and the body shrinks during the transfer. This demonstrates that the method was able to find a solution without artificially spreading the arms far from the torso to preserve the distances associated with the thin source character.

Figure 4.9 depicts results obtained with our method when applied to 3D models used in [Liu et al., 2018]. Our results are compared to those obtained by an artist (artist performance initially reported in [Liu et al., 2018, Figure 6]). Note that even with a relatively large change in morphology, our result is close to the solution proposed by an artist. In particular, when viewed from above, one can see that the artist created new contacts between the arms and the body. These additional contacts did not change the contextual meaning of the pose, but have been introduced to adapt to the morphology of the target character. These additional contacts have also been mostly recovered by our method, compared to the context graph method, which aims to preserve distances observed with the source charac-

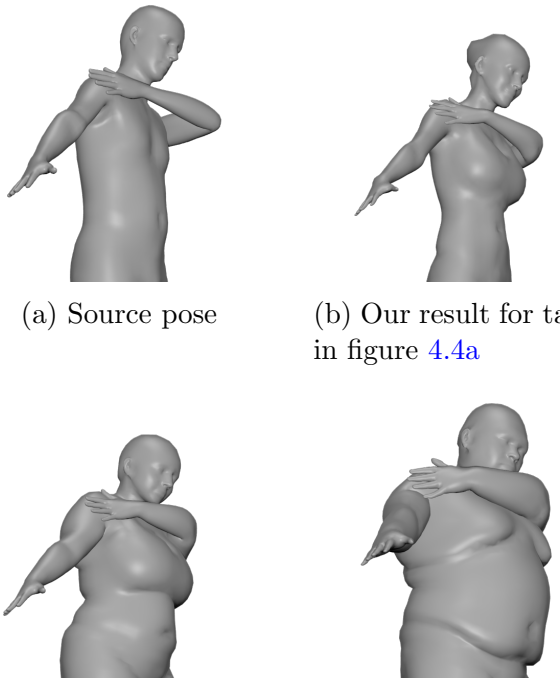


(a) Source (b) Target (c) Identity transfer result viewed from top and front. Left: source's deformed pose. Center: Our result with target 4.9b. Right: The result by an artist.

Figure 4.9: Comparison to an artist performance (courtesy of [Liu et al., 2018]). The results consists in retargetting a source character (a) to a target character (b). (c) front view of our result and a performance of an artist. (d) top view of the same results.

ter.

Figure 4.10 applies our method on a shoulder rubbing pose that is similar to the one used in AuraMesh [Jin et al., 2018, Figure 8]. We see that our method preserves the hand/shoulder contact, even with important changes of morphology. Notice that for a close morphology (Figure 4.10b), the distance between the elbow and the torso does not significantly change in the result. However, for drastically larger target characters (Figures 4.10c and 4.10d) this distance shrinks or even disappears to create new contacts. These pose changes do not alter the contextual meaning but are required to keep the morphology consistent. For the same kind of example, AuraMesh aims at preserving the initial distances observed with the source character, which might result in unnatural positions of the arms to preserve the distance with the torso.



(a) Source pose

(b) Our result for target in figure 4.4a

(c) Our result for target in figure 4.4b

(d) Our result for target in figure 4.3c

Figure 4.10: Comparison to the AuraMesh method [Jin et al., 2018] on a shoulder rubbing pose.

4.5 Conclusion

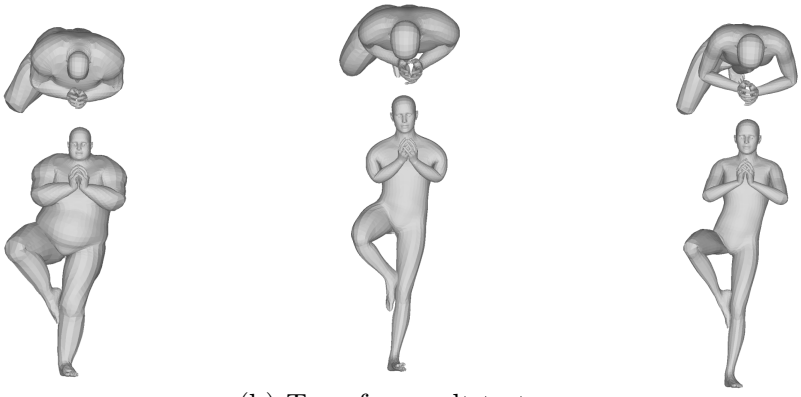
In this chapter we explore the strategy of identity transfer, as an alternative to the widely adopted pose transfer strategy, to address retargeting problems. We focus on preserving self-contacts in the result character in order to preserve the contextual meaning of the pose. Our approach shows that identity transfer allows us to transfer very different morphologies while naturally preserving constraints and adapting the pose. Our method only needs a pre-defined body part segmentation in order to generalize to different shape classes, such as humans or animal models, without changes in the parameters. Finally, our method can be adapted to animations, giving smooth and realistic new motion sequences.

While our experiments validate the identity transfer approach, our method could be further improved to alleviate some limitations, and some interesting questions remain open.

First, our method needs a post-processing step in order to be applied to

animations. This could be alleviated by computing the result of our method at key frames of the animations, and using a space-time constraint solver inspired by [Gleicher, 1998] to generate the missing frames. This would also greatly speed up the computations for animations by limiting the number of frames for which the method needs to be applied. Another limitation is that our method is slow compared to recent methods that can be close to real-time.

The long computation time needed by our method is due to the complexity of our body deformation model. Moreover, this model is imperfect and can lead to artefacts, such as the incorrect body part lengths discussed in Section 4.4.4 or unnatural joint deformations for complex poses and/or morphologies (*e.g.* the shoulders in Figure 4.9). To avoid the difficult task of defining a simpler and more robust body deformation model, a promising direction is to use a deep learning framework in order to learn from real world data how human characters deform. This would greatly speed up our method at inference time, and would allow us to predict more realistic body deformations. We explore this direction in Chapter 5 where we propose a data-driven adaptation of our identity transfer method.



(a) Source pose

(b) Transfer result to target in Figure 4.9a

(c) Result from an artist

Figure 4.11: Identity transfer from a large to a thin character. Our method preserves all contacts present in the source pose, even if they are not meaningful to the pose, such as the contact between the arms and the torso

An interesting question explored in this chapter is how to preserve the contextual meaning of a pose when transferring it to a new morphology, following our assumption that self-contacts are important to the meaning of the pose (see Section 3.3). In this chapter, we focused on preserving contacts in the source pose as well as preventing interpenetrations. Our

results show that this genuinely preserves part of the posture context. In particular, it enables to introduce new contacts that are induced by the target shape, especially when retargeting from a skinny to a large character (see Figure 4.9). However, in some cases contacts should not be preserved as they may not bring any semantic information to the pose. For example in Figure 4.11, the contact between the arms and the torso of the source character is present only because of the larger morphology. When transferring a thinner identity to this pose, preserving this contact in our method resulted in unnatural deformations. Conversely, the artist chose to remove it, which did not change the meaning of the pose in the resulting character. Our method could be adapted to ignore certain contacts specified by the artist to avoid this problem, but it requires an additional manual step to the method. This raises the question of which contacts should be preserved in a transfer between characters and more generally on how to automatically model the contextual meaning of a human posture. We explore this question in Chapter 6, by using human perception to disambiguate cases where contacts are important or not to the meaning of the pose.

Chapter 5

Neural Identity Transfer

5.1 Introduction

In this chapter, we investigate the identity transfer strategy explored in Chapter 4 with a data driven approach. Deep neural networks have been widely applied recently to solve complex nonlinear problems. Particularly, recent works have proposed deep learning architectures to learn how human characters deform, and perform deformation transfer (*e.g.* [Zhou et al., 2020, Wang et al., 2020, Cosmo et al., 2020]). As stated in Section 4.5, this approach greatly alleviates the problem of defining a human body deformation model.

Recent methods cited above usually predict the full deformation of the characters' bodies, making it difficult to correctly generalize to unseen data, such as complex poses far away from the training set. With the identity transfer approach, we propose a deep learning architecture that predicts the identity deformation of a source model already in the correct pose, so that its identity matches that of a target model. This way we predict simpler deformations, and are able to generalize better to unseen poses. The method presented in this chapter thus further validates the advantages of the identity transfer strategy. It also opens interesting future directions, such as using a similar deep learning architecture in order to learn how characters adapt their pose to their morphology.

The architecture of our model consists of an encoder that encodes the identity of the target model into a low-dimensional feature vector, and a decoder that consumes the identity feature vector along with the source model and predicts offsets from the source model that transfer the identity. To encode identity information, we base our losses on the hypotheses discussed in Section 3.2, namely that two characters with the same identity should

be near-isometric, and that body part of a character deform near-rigidly with changes in pose. To structure the latent space, we use a loss that aims to map feature vectors of the same identity to the same location in latent space.

As our identity losses are based on pose independent parameters, we only require identity labels to compute them, which facilitates using real-world datasets. Indeed, 3D models with different characters performing the exact same pose are rare in existing real-world datasets; for a captured motion, different characters will never perform the motion in the exact same way and timing. We thus train our architecture in a weakly supervised way: while we rely on the presence of identity labels for all training data, we only require pose labels for a small subset. To have access to high-quality labelled data, we propose an extension of the FAUST dataset [Bogo et al., 2014] that includes additional poses and identities with full label information. To create this extension, we use the method presented in Chapter 4 to transfer new poses and identities to the FAUST dataset. We then manually curate this dataset to remove failure cases with unnatural deformation. We demonstrate experimentally that having access to full label information, and hence a reconstruction loss, on a small proportion of the training data is sufficient to train our architecture.

Our self supervised losses also allow inference time refinement of our model. Inspired by the few-shot learning of generative models literature (*e.g.* [Zakharov et al., 2019, Arik et al., 2018, Jia et al., 2018]), we observe that fine-tuning our feed forward network at inference time improves the results. This is achieved with a few extra training steps on the inputs using our self-supervised losses. In this strategy, the initial network training can be seen as a meta-learning stage, and the fine-tuning can be interpreted as one-shot learning from a single reference pose/target identity pair, which adapts the network weights further to that specific case. To the best of our knowledge, this idea was not explored by learning-based deformation transfer works until recently, with the simultaneous publication of the method described in this chapter and of the work of [Lombardi et al., 2021]. Not only does the fine-tuning improve our performance quantitatively, but it also allows us to successfully transfer identity for out of training distribution shapes, such as a shape of a simply clothed person with a hat and a backpack, while the training consists merely of minimally dressed body shapes.

We compare our method to deformation transfer results by the recent deep learning approaches Unsupervised Shape and Pose Disentanglement (USPD) [Zhou et al., 2020] and Neural Pose Transfer (NPT) [Wang et al., 2020], and show that geometric detail is better preserved with our method

when applied to poses not observed during training.

In summary, our contributions are:

- our method better generalises to poses not seen during training than the state-of-the-art, achieved by transferring the identity of the target shape to the deformed pose within a deep learning framework;
- our method allows to preserve fine-scale detail linked to the identity of the character and generalizes to characters wearing simple clothing thanks to test time identity transfer refinement with fine-tuning;
- we extend the FAUST dataset [Bogo et al., 2014] to contain more identities and poses with full label information, which can be leveraged for training.

5.2 Method

This section describes our method to adapt the identity transfer strategy to a deep learning framework. We address the problem of deformation transfer between 3D shapes described by triangle meshes with the same topology, *i.e.* all meshes have the same connectivity and vertex to vertex correspondence. We assume a dataset of such shapes *i.e.* meshes $\{\mathcal{M}\}$ where some have ground-truth identity and/or pose labels, and denote them by $\mathcal{M}_p^{id}$, id being the identity label and p the pose label.

5.2.1 Overview

Figure 5.1 provides a visual overview of our approach. Given two meshes, $\mathcal{M}_p$ with an input source pose and $\mathcal{M}^{id}$ with an input target identity, our goal is to generate a third mesh $\tilde{\mathcal{M}}_p^{id}$ representing the shape of the target identity id in the source pose p. We formulate this problem using a deep learning framework and an encoder-decoder architecture.

Our neural architecture implements the identity transfer by predicting the deformation of the source model $\mathcal{M}_p$ so that its identity matches that of the target model $\mathcal{M}^{id}$. This does not require to encode explicitly pose information as the pose is naturally preserved by predicting identity deformations only.

The encoder Enc takes $\mathcal{M}^{id}$ as input and encodes its identity information into a low-dimensional feature vector $\boldsymbol{z}^{id}$ as

$$\boldsymbol{z}^{id} = Enc(\mathcal{M}^{id}). \tag{5.1}$$

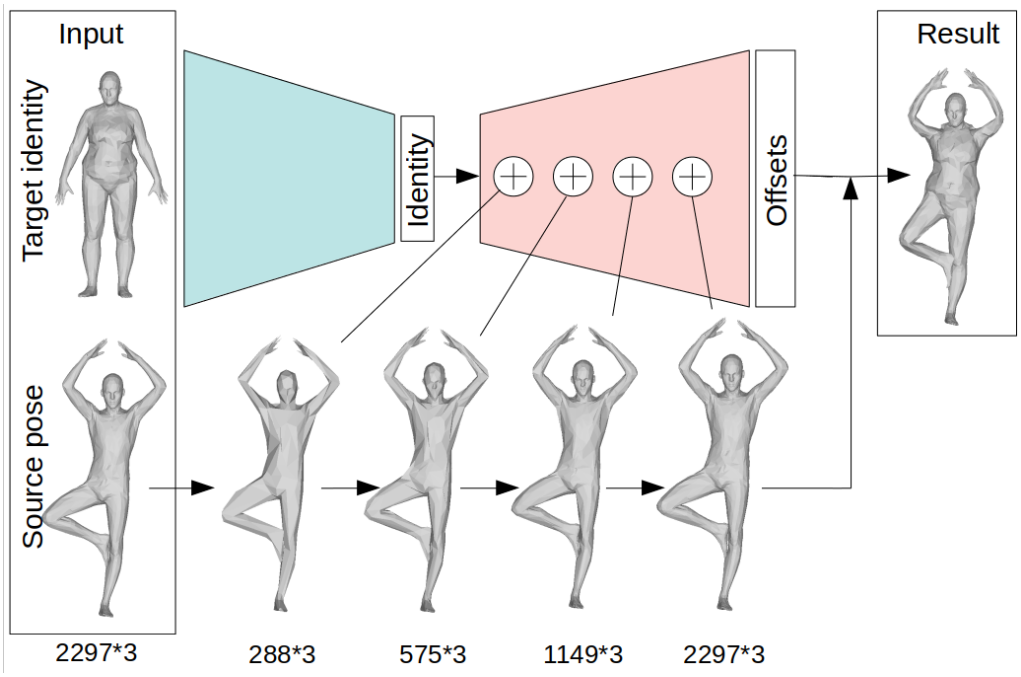


Figure 5.1: Overview of the proposed approach. The encoder (green) generates an identity code for the target identity. We feed this code to the decoder (red) along with the source pose, which is concatenated with the decoder features at all resolution stages. The decoder finally outputs per vertex offsets from the source pose towards the identity transfer result

The decoder Dec takes as input a latent code $\boldsymbol{z}^{id}$ along with $\mathcal{M}_p$ and outputs offsets from $\mathcal{M}_p$ to $\tilde{\mathcal{M}}_p^{id}$ as

$$\tilde{\mathcal{M}}_p^{id} = Dec(\boldsymbol{z}^{id}, \mathcal{M}_p) + \mathcal{M}_p. \tag{5.2}$$

The architectures of both Enc and Dec are based on spiral convolutions at gradually decreasing/increasing mesh resolutions through pooling/unpooling layers as proposed by [Bouritsas et al., 2019]. Note that while the pose information is not explicitly encoded, Dec is anyway conditioned on the pose $\mathcal{M}_p$. This is achieved in practice by concatenating channel-wise at every convolution and unpooling layer of Dec the current vertex features and the 3D coordinates of $\mathcal{M}_p$ at the corresponding mesh resolution.

The two main differences in terms of architecture compared to state-of-the-art encoder-decoder based deformation transfer methods, (*e.g.* [Cosmo et al., 2020, Zhou et al., 2020]) result from the identity transfer strategy. First, rather than encoding pose information, our decoder is conditioned on the input model $\mathcal{M}_p$ that has the desired pose. Second, rather than

predicting 3D vertex information directly, our decoder predicts offsets from $\mathcal{M}_p$. These offsets correspond to the identity deformation applied to the source pose in order to obtain our result.

At inference time, we fine-tune our feed-forward network to improve the results. This strategy, where network training can be seen as a meta-learning stage and fine-tuning as one-shot learning from a single $\mathcal{M}_p, \mathcal{M}^{id}$ pair, is inspired by the few-shots learning of generative models literature (*e.g.* [Zakharov et al., 2019, Arik et al., 2018, Jia et al., 2018]).

5.2.2 Training

The model is trained in a weakly supervised way because labeled 3D models of different characters performing the exact same poses are rare in existing real-world datasets. In particular, while each model is equipped with an identity label, only a small subset of all models is equipped with a pose label. For training, we sample triplets of distinct meshes of the form $(\mathcal{M}_{p_1}^{id_1}, \mathcal{M}_{p_2}^{id_2}, \mathcal{M}_{p_2}^{id_1})$ for fully labeled data, and of the form $(\mathcal{M}_{p_1}^{id_1}, \mathcal{M}_{p_2}^{id_2}, \mathcal{M}_{p_3}^{id_1})$ for data with only identity labels (with unknown pose labels p_1, p_2, p_3). Note that while fully labeled data contains the ground truth of the deformation transfer result $\mathcal{M}_{p_2}^{id_1}$, this information is not available for data with identity labels only.

These triplets are used to train the network based on the following losses

$$\begin{aligned} l_{sup} &= \alpha_{lat} l_{lat} + \alpha_{rec} l_{rec}, \\ l_{weaksup} &= \alpha_{lat} l_{lat} + \alpha_{lap} l_{lap} + \alpha_{rig} l_{rig}, \end{aligned} \tag{5.3}$$

where l_{sup} is the supervised loss used when full label information is available and $l_{weaksup}$ is the weakly supervised loss used when merely identity labels are known. Let $\tilde{\mathcal{M}}_{p_2}^{id_1}$ denote the transfer result predicted by our method for inputs $\mathcal{M}_{p_1}^{id_1}$ as target identity and $\mathcal{M}_{p_2}^{id_2}$ as source pose.

We use three types of losses to train the network. First, a latent loss l_{lat}, which helps structuring the latent space, is used during both full and weak supervision. Second, in case of full supervision, a standard L_2 penalty reconstruction loss l_{rec} is employed. Finally, in case of weak supervision, two self supervised identity losses l_{lap} and l_{rig} are used that measure identity distances based on the identity parameters identified in section 3.2. These losses are weighted using the weights α_{lat}, α_{rec}, α_{lap}, and α_{rig}. Details on these losses follow.

5.2.2.1 Latent Loss

This loss uses the identity label present in our data. Its purpose is to constrain the identity latent space by enforcing models with the same identity label to have a similar latent representation. We define this loss as

$$l_{lat}(\mathcal{M}_{p_1}^{id_1}, \mathcal{M}_{p_2}^{id_1}) = \|Enc(\mathcal{M}_{p_1}^{id_1}) - Enc(\mathcal{M}_{p_2}^{id_1})\|_2^2. \tag{5.4}$$

This loss is evaluated for the two meshes of the input triplet that share the same identity code, namely $\mathcal{M}_{p_1}^{id_1}, \mathcal{M}_{p_2}^{id_1}$ for fully labeled data and $\mathcal{M}_{p_1}^{id_1}, \mathcal{M}_{p_3}^{id_1}$ for data with identity labels only.

5.2.2.2 Reconstruction Loss

When pose labels are available, we use a standard reconstruction loss that measures the vertex-to-vertex L_2 distance between the ground truth $\mathcal{M}_{p_2}^{id_1}$ and the predicted result $\tilde{\mathcal{M}}_{p_2}^{id_1}$ as

$$l_{rec}(\tilde{\mathcal{M}}_{p_2}^{id_1}, \mathcal{M}_{p_2}^{id_1}) = \|\tilde{\mathcal{M}}_{p_2}^{id_1} - \mathcal{M}_{p_2}^{id_1}\|_2^2. \tag{5.5}$$

This strong constraint, that is effective for training, is only required for a small subset of our training data.

5.2.2.3 Identity Losses

When only identity labels are available, we design self supervised losses based on the hypotheses on the identity of 3D human characters introduced in Section 3.2.

The first hypothesis states that two characters with the same identity should be near-isometric. We design a loss that enforces this property, by penalizing differences between isometry descriptors of models with the same identity label. We compute isometry descriptors as the local Laplacian coordinates Δ_{loc} of the vertices of each model, as described in Section 3.2.2.1. Finally, our loss between two meshes $\mathcal{M}_{p_1}^{id}$ and $\mathcal{M}_{p_2}^{id}$ of the same identity is

$$l_{lap}(\mathcal{M}_{p_1}^{id}, \mathcal{M}_{p_2}^{id}) = \|\Delta_{loc_{p_1}}^{id} - \Delta_{loc_{p_2}}^{id}\|_2^2. \tag{5.6}$$

We use this loss between our prediction $\tilde{\mathcal{M}}_{p_2}^{id_1}$ and the input target identity model $\mathcal{M}_{p_1}^{id_1}$ during training.

The second hypothesis is that body parts of a same identity deform near-rigidly between different poses. We use the body part segmentation presented in Section 3.2.2.2, Figure 3.4a. With this hypothesis, we build a loss that penalizes distances between vertices belonging to the same body

part being inconsistent between our prediction and the target identity. The unsupervised rigidity loss between two models $\mathcal{M}_{p_1}^{id}$ and $\mathcal{M}_{p_2}^{id}$ of the same identity is then

$$l_{rig}(\mathcal{M}_{p_1}^{id}, \mathcal{M}_{p_2}^{id}) = \sum_{P \in \mathcal{P}} \sum_{i,j \in P} \|d(\boldsymbol{v}_{i,1}, \boldsymbol{v}_{j,1}) - d(\boldsymbol{v}_{i,2}, \boldsymbol{v}_{j,2})\|_2^2, \tag{5.7}$$

where $\mathcal{P}$ is the set of mesh body parts, $\{\boldsymbol{v}_{i,k}\}$ are the vertices of $\mathcal{M}_{p_k}^{id}$ and $d(.,.)$ is the Euclidean distance. We use this loss between our prediction $\tilde{\mathcal{M}}_{p_2}^{id_1}$ and the input target identity model $\mathcal{M}_{p_1}^{id_1}$ during training.

Note that the two identity losses are evaluated between our prediction and the target identity used for the prediction, and are thus fully unsupervised.

5.2.3 Fine Tuning

We introduce a fine-tuning step that is performed systematically at inference time. At a small additional computational cost, this step allows to improve results, and enables identity transfer to new shapes considerably different from those seen during training, as demonstrated experimentally.

This step acts as an additional adaptation of the weights of our pre-trained network to a specific input. Given a target identity $\mathcal{M}^{id}$ and a source pose $\mathcal{M}_p$, we first generate our result $\tilde{\mathcal{M}}_p^{id}$ using the trained model as described in Eq. 5.2. This result is used as initialisation for further optimisation. We fine-tune our model for a few more iterations, using as input identity the target identity $\mathcal{M}^{id}$, and as input pose the initial inference result $\tilde{\mathcal{M}}_p^{id}$. For these extra training steps, we use a self-supervised loss, combining the Laplacian and the rigidity losses to maintain the target identity, in addition to a regularization loss l_{reg} in the form of a L_2 penalty between the vertices of the initial result $\tilde{\mathcal{M}}_p^{id}$ and those of the final fine-tuned mesh.

$$l_{ft} = \alpha_{lap} l_{lap} + \alpha_{rig} l_{rig} + \alpha_{reg} l_{reg}. \tag{5.8}$$

5.2.4 Implementation Details

For the main model training, we generate training triplets as follows: a triplet $(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3)$ is created for each mesh sample $\mathcal{M}_1$ in the training data. The second and third meshes of the triplet $\mathcal{M}_2$ and $\mathcal{M}_3$ are then randomly sampled within models with a different identity as the first mesh, and a similar identity, respectively. If $\mathcal{M}_1$ comes from the portion of our data with pose labels, we restrict the choice of $\mathcal{M}_2$ to meshes with pose

labels too, in order to be able to select $\mathcal{M}_3$ with the identity label of $\mathcal{M}_1$ and the pose label of $\mathcal{M}_2$. In this case $\mathcal{M}_3$ acts as ground-truth for the transfer of $\mathcal{M}_1$'s identity to $\mathcal{M}_2$'s pose. Every five epochs, we re-sample a tenth of our training triplets chosen at random. This way, while each triplet is likely to be seen multiple times by the model, which helps lowering the loss values for these specific triplets, the re-sampling allows the model to see new triplets, which helps to better capture the variety of the training set.

Our network takes as input a list of 3D points that correspond to the vertices of the input meshes. We preprocess all meshes by aligning them rigidly and down-sampling them to 2297 vertices using a quadratic error criterion following [Ranjan et al., 2018]. This down-sampling balances the computing cost of our losses, while keeping a reasonable level of precision. We propose a simple two step up-sampling for better qualitative visualization. First, we up-sample the meshes to 6890 vertices, by placing the new vertices at the centroid of their neighbours [Ranjan et al., 2018]. Then, we move the new vertices to the local Laplacian coordinates (see Section 5.2.2.3) computed on the unprocessed target identity, while preserving the coordinates of the 2297 vertices predicted by our model.

We use the ADAM optimiser. For the main training, we use a learning rate of 0.001 and a learning rate decay of 0.99 per epoch, and train for 500 epochs. We use batches of size 32. We set the loss weights in Equation 5.3 as $\alpha_{rec} = 10$, $\alpha_{lap} = 1000$, $\alpha_{rig} = 1$ and $\alpha_{lat} = 1000$. For the fine-tuning, we use a learning rate of 0.0001, and fine-tune for 50 iterations. We set the loss weights in Equation 5.8 as $\alpha_{lap} = 10$, $\alpha_{rig} = 1$ and $\alpha_{reg} = 0.1$.

5.3 Evaluation

In this section we evaluate our model's ability to achieve deformation transfer both quantitatively and qualitatively. We perform an ablation study to evaluate the effects of supervising and fine-tuning our model, and compare our method quantitatively to state-of-the-art deformation transfer methods. In particular, we choose to compare to the supervised NPT [Wang et al., 2020] and the unsupervised USPD [Zhou et al., 2020] as they achieve the best results in the literature. While a comparison with [Cosmo et al., 2020] would be interesting, since this method builds on a similar isometry hypothesis, such a comparison appears intractable in practice as a result of the computational cost of their method. We also compare our results to our previous optimization based identity transfer method, presented in Chapter 4. We then present qualitative results of deformation transfer using our

method. We present results of transferring the identity of minimally dressed human characters, extreme morphologies and characters with simple clothing. We apply our method to animations on a frame-by-frame basis, and to an identity morphing scenario, where the pose stays constant while the identity changes.

To evaluate the results numerically, we use input pairs of shapes $\mathcal{M}_p$ and $\mathcal{M}^{id}$ for which the ground truth of the deformation transfer $\mathcal{M}_p^{id}$ is known. As our meshes are in point-to-point correspondences, the error is measured using the mean of the L_2 distances between corresponding vertices of the ground truth $\mathcal{M}_p^{id}$ and the result $\tilde{\mathcal{M}}_p^{id}$ after Procrustes alignment.

5.3.1 Data

5.3.1.1 DFAUST

We train our method using 3D human models from the Dynamic FAUST (DFAUST) dataset [Bogo et al., 2017]. This dataset contains 10 identities performing between 11 and 14 motions, each of them containing a few hundred frames. It is straightforward to label models from this dataset w.r.t. identity, since the identity of a character does not change during a specific motion sequence. However, even if the motions are semantically equivalent across subjects, the poses differ in timing and style, and thus no pose labels are available. We use 41220 models from this dataset for training.

5.3.1.2 ExtFAUST

To obtain labelled data for supervision, we create a new dataset with full identity and pose labels by augmenting the FAUST dataset [Bogo et al., 2014] with additional pseudo-ground-truth. FAUST contains 10 identities performing the same 10 poses each, providing us with 100 meshes with full identity and pose labels. We extend this data by adding meshes with new poses and identities from other datasets, and then applying our deformation transfer method presented in Chapter 4 to transfer every new identity and pose to all pre-existing poses and identities in FAUST. For the new poses and identities added to FAUST, we choose meshes from DFAUST [Bogo et al., 2017], SMPL [Loper et al., 2015], and Adobe's Mixamo [mix,]. We add 11 identities and 17 poses to the original FAUST data, yielding 540 meshes with pose and identity labels after manually removing a few outliers. We refer to the resulting dataset as Extended FAUST (ExtFAUST) in the remainder of this paper. We created a test split by removing all occurrences

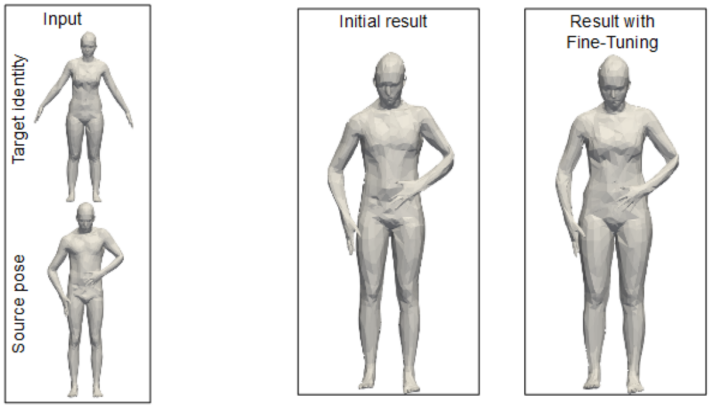


Figure 5.2: Result of identity transfer with our model before and after the fine tuning step

of 4 poses and 4 identities from this dataset. This leaves 369 shapes for training and 171 for testing.

5.3.1.3 Test Sets

We use three different test sets in our evaluations. The *ExtFAUST pose test set* consists of 4 identities in 4 poses, all of which were unseen during training. This allows to evaluate the method's ability to generalize to both new identities and poses. When combining all possible triplets of target identity, source pose, and transfer ground truth, 240 triplets are available for testing. The *ExtFAUST id test set* consists of 4 identities unseen during training in 4 poses that were seen during training. This allows to evaluate the method's ability to generalize to identities on poses that it has been trained on. Similarly, A total of 240 ground truth triplets are available for testing. The *AMASS test set* is used for evaluation w.r.t. the state-of-the-art. It contains 100 triplets generated from motion capture data used in AMASS [Mahmood et al., 2019] combined with random SMPL [Loper et al., 2015] shape parameters.

5.3.2 Ablation Study

To evaluate the influence of the fine-tuning at inference time, we run our method with and without fine-tuning on the *ExtFAUST pose test set*. While the average per vertex error without fine-tuning is 31.51mm, it decreases significantly to 20.19mm when fine-tuning is used.

We further illustrate the impact of fine tuning in Figure 5.2. We observe that before fine tuning, the result of our method has the correct general

Supervision	None	FAUST	ExtFAUST
Mean error (mm)	29.19	24.83	20.19

Table 5.1: Ablation study on supervision.

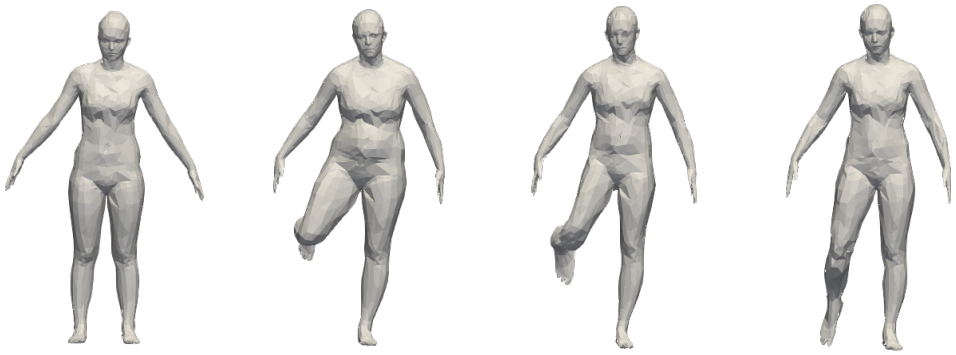


Figure 5.3: Left to right: target identity, source pose, identity transfer result with all ExtFaust used as supervision during training, result with no pose supervision during training

body shape and morphology, but lacks fine scale identity details. However, after the fine tuning step, our result is more precise, and our method is even able to transfer details such as simple clothing and accessories, even though it was trained solely on minimally dressed humans (see Section 5.3.4).

Given the good results obtained using our fine-tuning step, all results discussed in the following are obtained with this refinement unless specified otherwise.

To evaluate the necessity and effectiveness of our supervision scheme (Equation 5.3), we train our model without any full supervision, with only the FAUST data as full supervision (approximately 0.2% of the training data is labeled), and with all the ExtFAUST data as full supervision (approximately 1% of the training data is labeled). Tab. 5.1 reports the errors in mm on the ExtFAUST pose test set. Note that a small percentage of labeled training data allows to improve our results by almost $1cm$ in average per vertex error.

In Figure 5.3, we present a qualitative example of transferring an identity to an unseen pose, using the model without supervision and the model with supervision from ExtFAUST. Notice that without supervision, while the identity of our result is correct, the pose of the source character was not preserved. This is due to the fact that our identity losses are not perfectly disentangled from pose information. Without pose supervision,

pose information from the target identity is therefore also transferred at inference, resulting in *e.g.* the straight right leg of the unsupervised result in the figure.

5.3.3 Comparison to State-of-the-art

To the best of our knowledge none of the existing deformation transfer methods operate in a weakly supervised way and we compare therefore our method to the state-of-the-art supervised method NPT [Wang et al., 2020] and unsupervised method USPD [Zhou et al., 2020]. This results in three methods that make different assumptions on their training supervision. Moreover, while our method and USPD assume full correspondence of the 3D input models, NPT is more general and can handle 3D models without correspondence or fixed topology. These differences make a completely fair comparison difficult. To make the comparison as fair as possible, we train each method in its optimal supervision setting, with the training data presented in the original papers.

We evaluate the errors on our three test sets: one that requires pose generalization from all methods, and two that require pose generalization from some of the methods. For the *ExtFAUST pose test set*, none of the methods have seen during training any of the poses or identities presented at test time. This test set therefore evaluates all method's abilities to generalize to new poses and new identities, and can be considered the hardest test set for all methods. For the *ExtFAUST identity test set*, none of the methods have seen any of the identities presented at test time. However, our method has seen the poses, coupled with other identities, during training. This test set therefore requires NPT and USPD to generalize to new poses, while this is not the case for our method. For the *AMASS test set*, none of the methods have seen any of the identities presented at test time, as we randomly sampled these identities' SMPL parameters. However, NPT and USPD have seen the poses, coupled with other identities, during training. This test set therefore requires our method to generalize to new poses, while this is not the case for NPT and USPD.

Figure 5.4 shows cumulative error plots for each method on each validation set. Note that our method and USPD obtain significantly better results for the first two validation sets, that require a generalization ability to new poses from NPT. This is because NPT does not use correspondence information, and treats points on the 3D human model that are close-by as neighbors and aims to deform them using similar deformations. In cases where different body parts are close-by or in contact in one input pose but not the other, this creates stretching artifacts that explain the high errors.

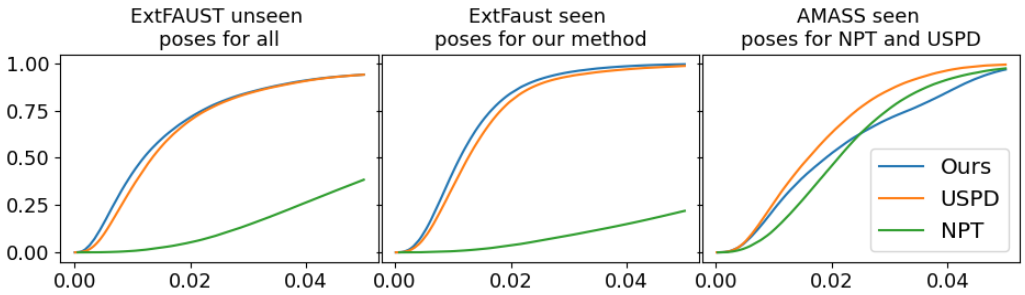


Figure 5.4: Cumulative errors for our method, USPD and NPT on 3 validation sets. The x-axis shows per-vertex errors (m). The y-axis is the proportion of all error values below the corresponding error value

For the AMASS test set, where NPT does not need to generalize to new poses and the results provide a meaningful measure for NPT's performance, their result is better, but our method still outperforms NPT on average and in the fine details.

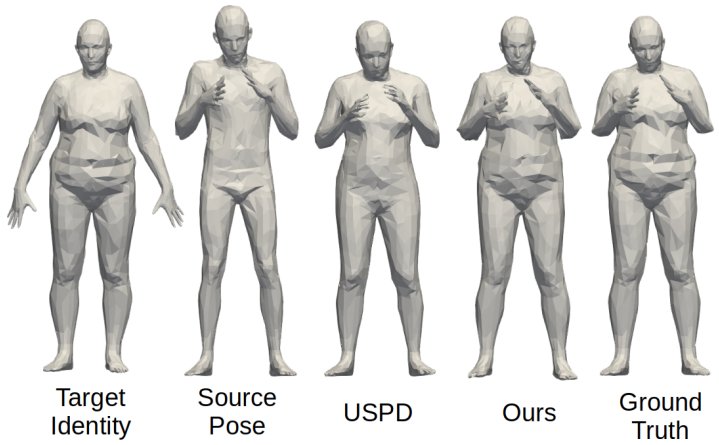


Figure 5.5: Qualitative comparison to USPD.

Our method and USPD perform respectively better than the other when one method has seen the poses during training and not the other. For the *ExtFAUST pose test set* with poses unseen by all methods, both methods have similar performances, but our method gives slightly better results for the low error range, showing that details are better preserved. This can be observed in Figure 5.5, where USPD's result has the correct overall body shape, but the details of the identity are overly smooth, whereas our method better transfers the fine-scale geometric details of the identity. It is also noteworthy to mention that USPD uses approximately 3 times more training data than we do.

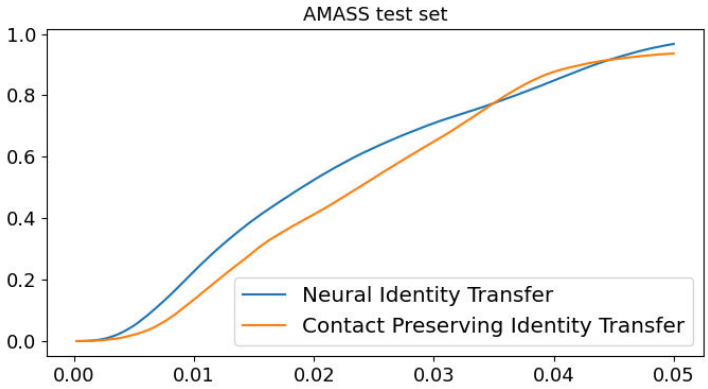


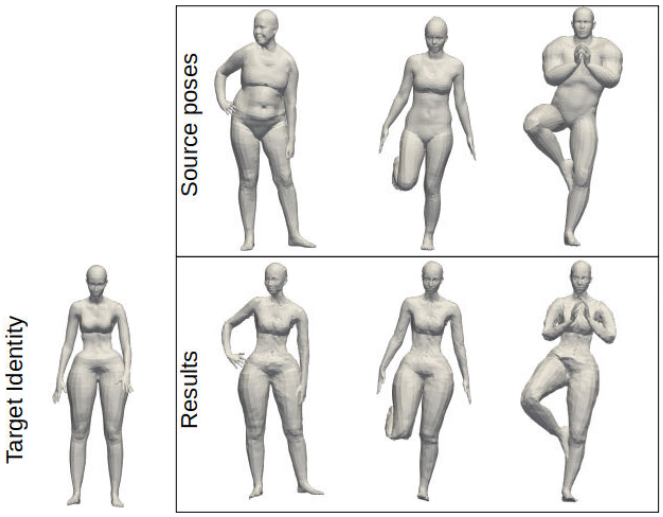
Figure 5.6: Cumulative errors for on the AMASS test set for our optimization based method described in Chapter 4, Contact Preserving Identity Transfer, and the Neural Identity Transfer method. The x-axis shows per-vertex errors (m). The y-axis is the proportion of all error values below the corresponding error value

We also compared our method to our previous optimization based approach described in Chapter 4, Contact Preserving Identity Transfer (CPIT). Figure 5.6 shows the cumulative error plot for these two methods, on the *AMASS test set* where the Neural Identity Transfer (NIT) method didn't see any pose or identity during training. We observe that the NIT method has a slightly better precision in this evaluation set. Note that the NIT method is partially trained on motion retargeting results from CPIT, with the Extended FAUST dataset. However, this concerns a very small subset of the training data, which is simply used for pose supervision. We argue that our method could be also trained with retargeting results from another state of the art method and obtain similar precision. Moreover, the CPIT method needed about 2 minutes of computation per transfer, while the NIT method took 5 seconds per transfer with the fine-tuning step. This validates the advantage of our extension of the identity transfer strategy with a data driven approach.

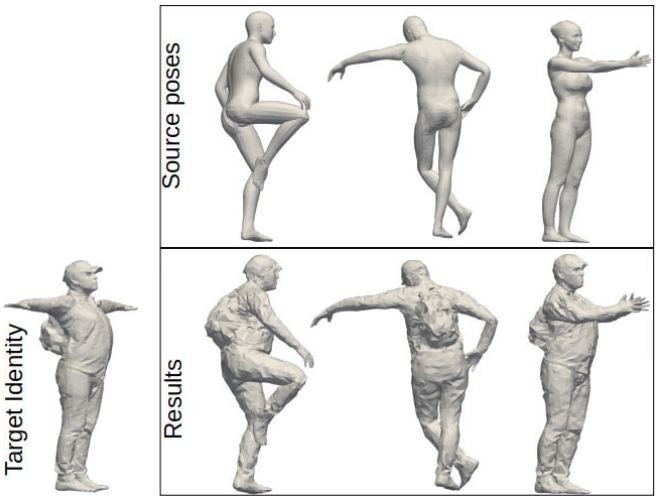
5.3.4 Qualitative Evaluation

Figures 5.7a, 5.7b, 5.9, 5.8 and 5.12 present results upsampled with the method described in Section 5.2.4, for visualization purposes.

Figure 5.7a shows results of transferring an unrealistic identity, unseen at training, to new poses. This figure demonstrates that our method is able to transfer extreme identities, while our method only saw realistic identities



(a) Transferring the identity of an unrealistic character to new poses.



(b) Transferring the identity of a clothed character to new poses.

Figure 5.7: Qualitative results of our method applied to identities far away from the training data

during training. Figure 5.7b shows results of transferring a character with simple clothing and accessories to new poses. Note that during training, no clothed characters or accessories are seen. These results show our method's ability to generalize to data that is far from the distribution of the training data while preserving geometric detail. This property is achieved in large part by the fine-tuning at inference.

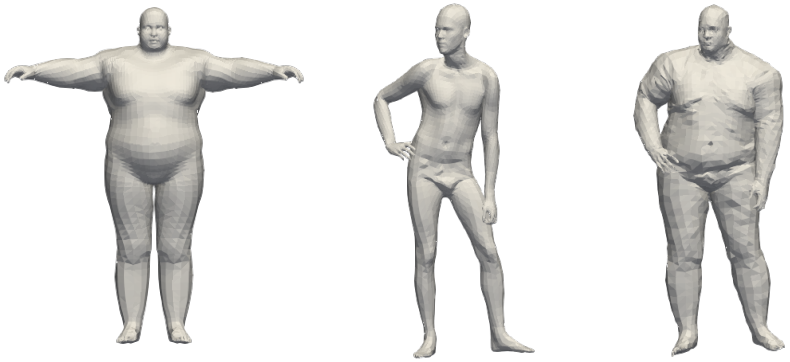


Figure 5.8: Left to right: target identity, source pose, identity transfer result

Since our model is trained with real-world data, it naturally learned to avoid interpenetrations in its results. In Figure 5.8, we show that even when transferring a very large character to an unseen pose with body part in close vicinity, no or limited interpenetrations appear.

To demonstrate the potential of our method, we apply it to two problems arising in automatic content creation.

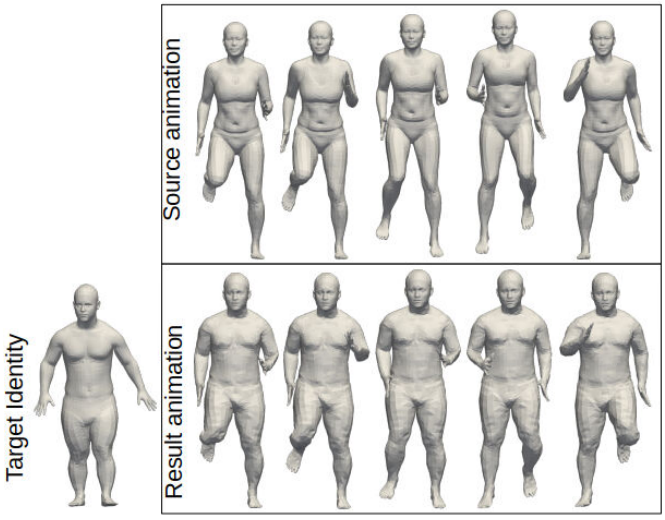


Figure 5.9: Transferring a new identity to an animation

First, Figure 5.9 shows our method applied to solve the motion retargeting problem. Given an input animation and a new identity, we apply our method to the animation on a frame-by-frame basis. In this scenario, our fine tuning step is done by taking as input the full source motion se-

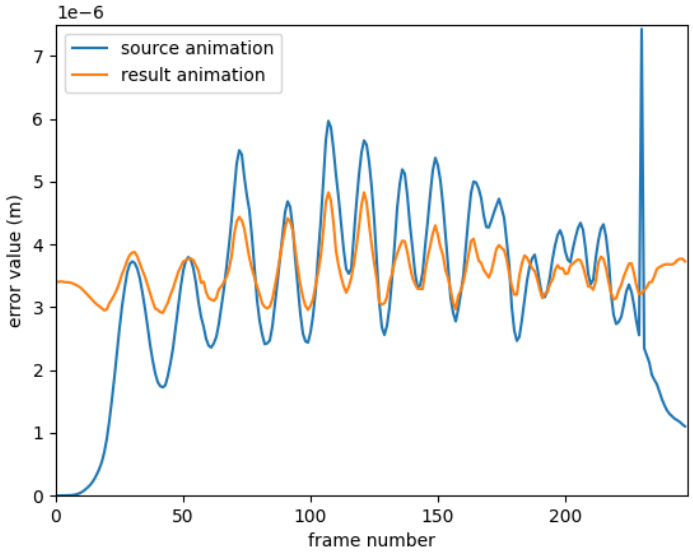


Figure 5.10: Isometry error (Equation 5.6) computed between each frame of the animation in Figure 5.9 and the corresponding identity. Note that our result has a similar identity jitter than the source animation

quence. At each fine-tuning iteration, we thus compute the loss described in Equation 5.8 with the target identity, and with each frame of the source animation treated as source pose successively. Note that although no temporal information is used by our method, the resulting animations are consistent and do not suffer from significant jitter. This is further confirmed by Figure 5.10. In this figure, we observe that the identity jitter computed using Equation 5.6 present in the result animation is well within the bounds of the identity jitter of the source animation. We refer the reader to the supplementary material found at this address for better visualization of these results: https://hal.archives-ouvertes.fr/hal-03440562/file/Neural Human Deformation Transfer.mp4.

Second, Figure 5.11 shows our method applied to solve the morphing problem. For this result, the identity code of two input characters is linearly interpolated in the latent space before being passed to the decoder. As no target identity exists for the interpolated identities, we do not apply the fine-tuning step for this experiment. Our method is able to interpolate smoothly between identity codes while keeping the pose consistent. In addition to being an interesting application, this result shows that the latent space learned by our method is well structured.

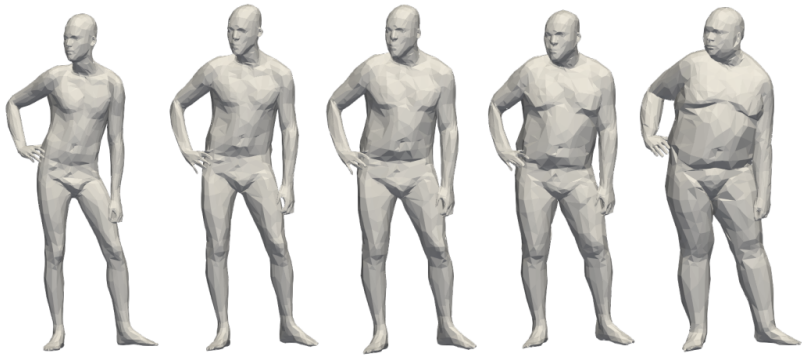


Figure 5.11: Interpolating the identity latent code between the leftmost and rightmost models

5.4 Conclusion

In this chapter, we introduced a neural deformation transfer method that predicts the identity deformation from a source character to a character with the same pose and a new identity. We used geometric properties of meshes to describe identity in a pose invariant way. We introduced a large dataset of human models with full identity and pose labels, which we use in addition to a larger unlabeled dataset to supervise our training. Experiments demonstrate our model's ability to generalize to unseen poses when using around 1% of supervision at training time. A fine tuning step, inspired by the few-shot learning methods, is shown to allow for the transfer of fine-scale geometric details of the identity. The method generalizes well to new identities, and even allows to transfer simple clothing and accessories.

This work successfully adapts the identity transfer strategy explored in Chapter 4 to a deep learning framework. Our experiments showed that this approach better generalizes to unseen poses, which validates the interest of the identity transfer strategy. While the fine-tuning step of the method blocks real-time computations, the method at inference is still much faster than our previous optimization based approach, widening the range of possible applications.

A limitation of our method is the need for supervision, as training our network without supervision results in deformed poses. This is due to the fact that our hypotheses on human characters identity are approximations, as discussed in Section 3.2.2.3. The identity parameters that we derive from these hypotheses are therefore not entirely disentangled from pose. As such, pose information from the target identity is also transferred by our method if no supervision is used to alleviate this limitation. This can

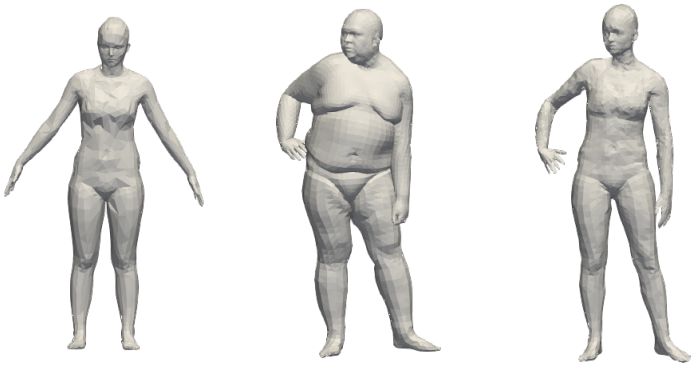


Figure 5.12: Left to right: target identity, source pose, identity transfer result

be seen qualitatively in Figure 5.3, where we can observe that the limbs of the results of unsupervised models tend to deform toward the pose of the target identity. An interesting direction for future work is to explore other identity losses, based on parameters of the identity completely independent from the pose.

While our method is able to naturally handle interpenetrations, as shown in Figure 5.8, it was not explicitly trained to preserve self-contacts such as what was done in Chapter 4. In some cases, this results in important contacts disappearing in the result of the transfer, which changes the meaning of the pose (see Figure 5.12). An interesting future direction would be to adapt our method to make it preserve the contextual meaning of the pose. This would require a better definition of which constraints, such as self-contacts, are important to this contextual meaning. Indeed, while simple hypotheses such as preserving all existing self contacts can give satisfying results such as demonstrated in Chapter 4, failure cases can happen such as the one showed in Figure 4.11. In this goal, we explore which self-contacts are important to the meaning of pose in Chapter 6. This information could then be used in a similar architecture as described in this chapter.

Chapter 6

Impact of Self-Contacts on Perceived Pose Equivalences

6.1 Introduction

In the previous chapters, we presented the identity transfer strategy, an approach to the deformation transfer problem where the source pose is preserved while modifying the identity of the character. An important advantage of this strategy is that since the pose is simply preserved, the result should have an equivalent pose with no significant change. However, small pose adaptations can be necessary to avoid collisions or loss of important constraints. In Section 3.3, we argue that self-contacts, *i.e.* contacts between body surfaces of the same character, are one of these constraints: for example, a pose of a character clapping will be considered incorrect if the hands do not touch. Moreover, we argue that with the identity transfer approach, other constraints such as general orientation and spatial positions of body parts are naturally respected, as the correct pose is considered at the initialization of the algorithm.

However, as illustrated in Figure 3.6, some self-contacts only appear as a consequence of the morphology of the character performing the pose, and do not bring semantic information to the pose. Systematically preserving them during deformation transfer can thus cause artifacts as preserving these unimportant contacts might significantly change the pose (see Figure 4.11).

An interesting question left open by our approach of the previous chapters is therefore how to automatically select which self-contacts must be preserved and which should be ignored. The solution to this problem is not obvious as the importance of these contacts does not follow any precise mathematical rule, and therefore can not be extracted solely from the

given pose. In this Chapter, we make the hypothesis that human observers should be able to determine if two poses are equivalent or not [Harada et al., 2004]. Therefore, we propose the design of a perceptive study aiming to leverage human perception to better understand which contacts are important characteristics of the pose and which are not.

In our study, we first aim to validate our assumption that self-contacts are in general important to the meaning of the pose. Our hypothesis is that observers presented with two similar poses, one with a missing self-contact present in the other, would consider the two poses to be visually different in most cases. We also argue that the importance of self-contacts depends on the body parts involved. Indeed, the studies on pose similarities presented in [Harada et al., 2004, Marinoiu et al., 2016] suggest that different body parts have different impacts for the perceived pose meaning. Those works also highlight that observers tend to give more attention to the positions of the head, wrists and fingertips when evaluating or imitating a pose. Moreover, we observe that self-contacts involving hands are often the goal of the pose, *e.g.* clapping, grabbing an object, or the example in Figure 3.6. Therefore, we make the hypothesis that self-contacts involving the hands are more important to the meaning of the pose than others.

To perform the study, we first selected a variety of poses of human characters presenting self-contacts. We then applied the method of Chapter 4 to transfer several different identities to these poses. For each of these transfers, we generated a first result where all the contacts present in the source were preserved, and variations were each individual self-contact in the source was successively removed by changing the parameters of the method. We then presented human observers with examples composed of the source pose, the target identity, the transfer result with all contacts preserved, and one variation with a contact removed. Observers were asked to select which transfer result best imitated the source pose.

To limit the number combinations of parameters to test, we focused our study to self-contacts between the arms or the hands and the upper part of the characters' body. The results of our study show that observers tend to consider that the pose with the same contacts than the source pose was the best imitation of the source pose in most cases. This tendency appears to be more important for contacts involving hands than contacts involving the arms of the characters. Finally, the results show that the release distance of the contacts, *i.e.* the distance between the body surfaces that were originally in contact in the source, have an impact on perceived pose equivalences; the tendency to chose the pose with the same contacts than the source is stronger when the presented variation has a more important contact release distance. This study thus confirmed our intuition that self-contacts do not

all have the same impact on the pose, and highlighted possible parameters to select important contact.

This study was designed and conducted in collaboration with Badr Ouannas, intern at the Inria Grenoble Morpheo team, Ludovic Hoyet, researcher at the Inria Rennes Mimetic team, Stefanie Wuhrer, Franck Multon and Edmond Boyer.

6.2 Related Work

The Computer Graphics literature has widely investigated human perception of virtual characters. Human observers have proved to be able to recognize the identity of actors in stylized or simplified virtual characters. A lot of interest has been given to which factors impacted the recognition of an actor in a deformed face [Tanaka and Farah, 1993, Zhao et al., 2003, Olivier et al., 2020]. Closer to our problem, recognizing the identity or style of an actor from its body pose and motion was also explored. [Johansson, 1973] represented human motions with a limited number (~ 10) of bright points, and showed that this representation was enough to evoke different kind of motions to observers, such as walking or running, and even to recognize the style of the motion (*e.g.* tired or "wavy"). More recently, [Hoyet et al., 2013] animated two realistic characters (one male one female) from actors' captured motion, and explored which motions better allowed to recognize the actor.

Some works have more specifically explored perception of pose or motion similarity. [Hodgins et al., 1998] showed that observers were better at perceiving pose differences in more realistic models than stick figures. [Harada et al., 2004] designed a quantitative pose similarity metric, and optimized it by comparing their results to pose similarity perceived by observers. They highlighted several parameters that impact perceived pose equivalences, such as weights describing the impact of each body part on the perception of the pose. In particular, they showed that the position of the fingertips has a strong impact on the perception of the pose. [Chen et al., 2009] use this approach to design a similarity metric that accounts for relative similarity, *e.g.* which among a set of examples is more similar to a target. [Tang et al., 2008, Pražák et al., 2009] propose a similar approach applied to human motion similarity. [Tang et al., 2008] also illustrate that simple similarity metrics such as Euclidean distance between joints does not always correlate with human perception of similarity. More recently, Laban Dance Notation [Laban, 1928] has been applied to the study of human poses. Laban Notation serves as a language to evaluate and record human poses and mo-

tion based on qualitative parameters. This notation has been applied to motion generation [Durupinar et al., 2016], and closer to our work perception based pose similarity metrics [Durupinar, 2021]. The latter further confirms that simple similarity metrics do not correlate with human perception.

[Marinoiu et al., 2016] evaluated the capacity of human subjects to imitate the pose of a target character. The subjects were equipped with motion capture markers and an eye tracker. The study showed that subjects gave more attentions to certain joints of the target before imitating the pose, such as the head and the wrists. More recently, [Müller et al., 2021] used a similar approach to annotate a dataset of 2D images of human characters in the wild. They presented 3D models of human characters to subjects tasked to take pictures of themselves imitating the character's pose. Their dataset focuses on poses containing self-contacts, which illustrates that human subjects are able to understand and imitate poses containing self-contacts with acceptable precision.

6.3 Data

In this section, we present the data generation method used to create the human body poses presented to the observers. Our goal is to create deformation transfer examples on poses with self-contacts, and to generate variations of the transfer result with each contact being alternately removed from the result.

6.3.1 Data Generation Method

To create the data used in this study, we first selected a set of poses presenting self-contacts. There is a wide variety of possible surface-to-surface contacts in the human body, and evaluating all possible configurations in a study would require a very large amount of data and number of subjects. To simplify this problem, we chose to focus on poses presenting self-contacts between the hands or the arms and the upper portion of the body (above the thighs, thighs included). We consider that these configurations are the most common in the space of human poses and are therefore the most interesting to study. We selected poses from the FAUST [Bogo et al., 2014] dataset, from Adobe's Mixamo [mix,], and from hand-tuned SMPL parameters [Loper et al., 2015]. We transferred a standard identity, generated with SMPL using the mean shape parameters, to these poses using the method from Chapter 4 (see Figure 6.1).

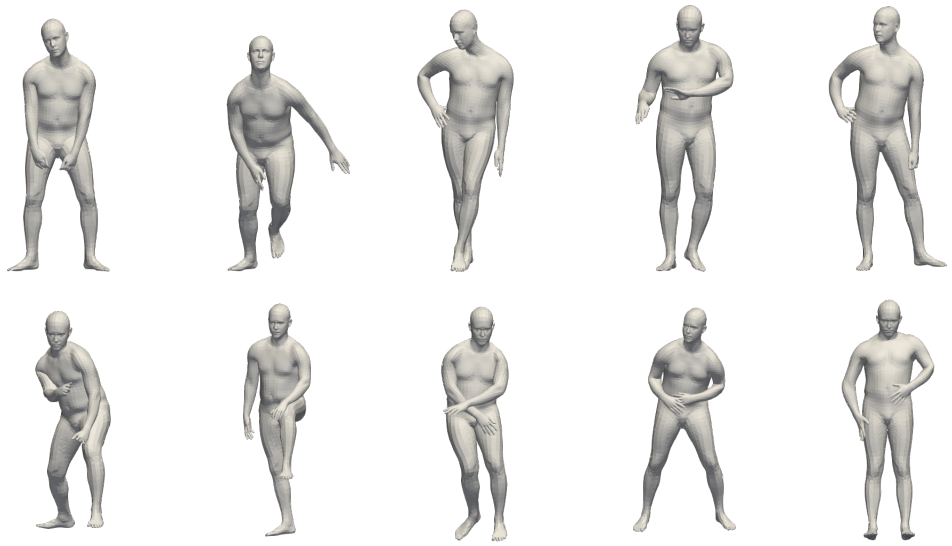


Figure 6.1: Examples of source poses with self-contacts used in the study, applied on the average SMPL identity parameters

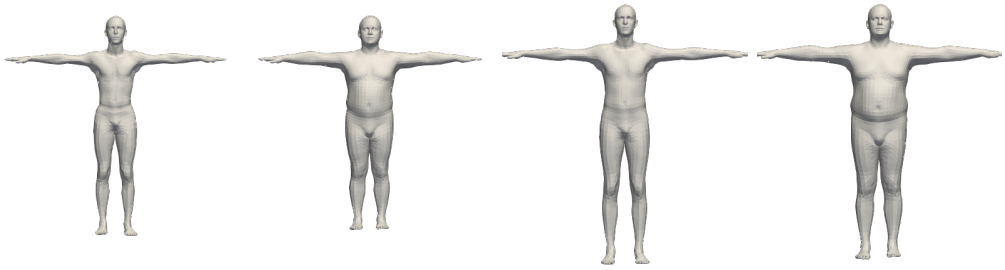


Figure 6.2: Target identities used in the study

For the target identities, we arbitrarily chose to focus our study on male human characters, once again in order to limit the number of examples and subjects needed to perform the study. We consider that conclusions drawn on the importance of self-contacts on this panel should translate on female and neutral body shapes. We generate target identities by sampling models at ± 2 standard deviations of the mean for the two first shape parameters of the SMPL model, height and body weight. We obtain 4 target identities with the following parameters; tall thin, tall big, short thin and short big (see Figure 6.2).

To generate our transfer examples, we transferred each identity to each

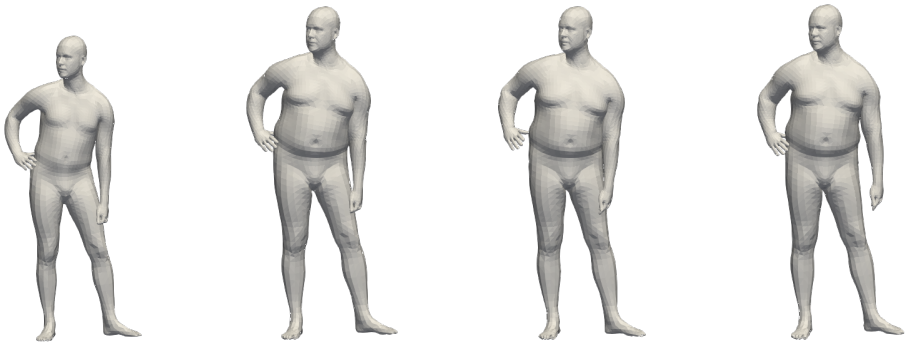


Figure 6.3: Example of generated pose variations. The first model is the source pose. The second is the transfer result to the last identity of Figure 6.2 with the same self-contacts as the source pose (the original pose). The last two models are the pose variations, *i.e.* the same transfer result with respectively one of the two self-contacts present in the source released

pose using the method described in Chapter 4. We first generated the "correct" transfer result by using the method with its original parameters. We then generated the pose variations by alternately selecting each interesting contact to be ignored by the method during the transfer (see Figure 6.3).

6.3.2 Stimuli - Pose Variation Dataset

With the generation method presented in the previous section, we obtained a dataset containing 17 different poses with self-contacts. For each pose we created one transfer result with all contacts preserved, and one variation for each self-contact to be removed while the others are preserved. Each pose had one or two variations in addition to the "correct" transfer result, for a total of 30 variations of poses. We transferred each target identity to all pose variations and obtained a total of 120 examples of transfer with a pose variation to be evaluated in our study.

We annotated each pose variation depending on whether the self-contact removed involved the arm or the hand of the character. Our dataset contains 68 pose variations in the category *arm* and 52 in the category *hand*. We also measured the contact release distance for each pose variation generated with our method, *i.e.* the mean distance between all vertices that were involved in the self-contact removed from the original pose. Our pose variations had a mean release distance of 4.98cm and a standard deviation of 3.42cm. For the *hand* (respectively *arm*) category, the release distances had a mean of 5.23cm (respectively 4.80cm) and standard deviation of 2.99cm

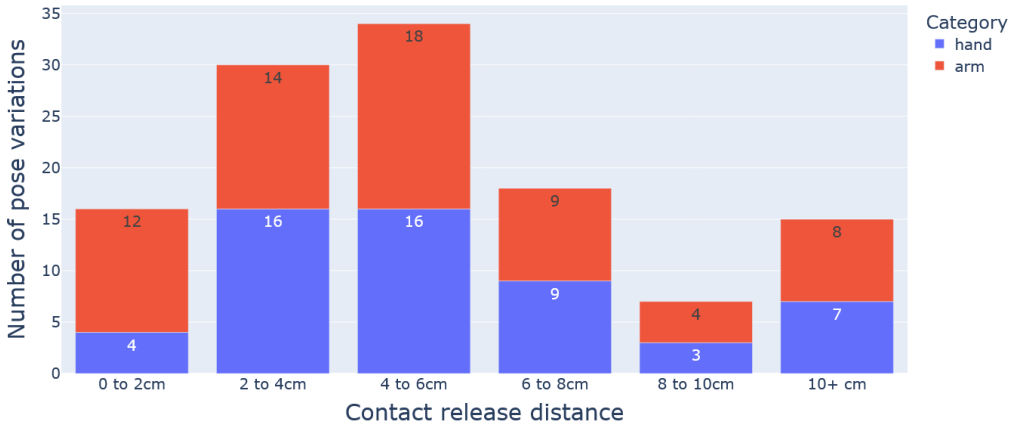


Figure 6.4: Number of pose variations in total and per category, depending on the contact release distance

(respectively 3.70cm). The distribution of release contact distances on our pose variations across categories is illustrated in Figure 6.4.

6.4 Study

The goal of our perceptive study is to better understand in which conditions a self-contact will be important to the meaning of the pose. More specifically, we explore the impact of the involved body parts on the importance of self-contacts. We first aim to validate the assumption that self-contacts are generally an important constraint defining the pose. Based on observations and state-of-the-art studies [Harada et al., 2004, Marinoiu et al., 2016], we make the hypothesis that self-contacts involving the hands are more important to the meaning of the pose than others. Finally, the hypotheses explored in this study are:

H1 Pose variations with released self-contacts will be perceived as different from the source pose in most cases.

H2 A pose variation of the category *hand* (*i.e.* hand self-contact removed) will be perceived as different from the source pose more often than for the category *arm*.

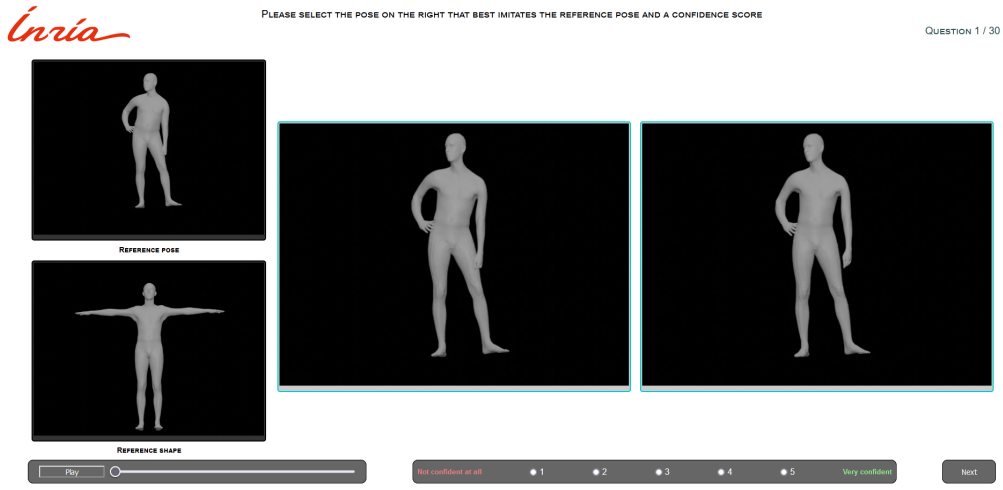


Figure 6.5: Interface presented to users during the study

6.4.1 Population

The study was designed to be shared online to subjects. We shared the study to laboratory staff, and using academic mailing lists. The study was validated by the Inria ethics evaluation comity. Participants gave written and informed consent before starting the study. Eighty-one (52 male, 26 female, and 3 who chose not to specify gender) subjects answered the study, with ages ranging from 19 to 61 years old, with an average of 31 years old. Forty-four participants reported having prior experience with 3D animation. Participants did the study on computer screens, with an average and standard deviation screen size of 21.5 ± 5.8 inches.

6.4.2 Protocol

Before the study, participants were presented with instructions illustrated with an example of question. They were asked to sign an informed consent form before proceeding to the study. They then filled a short demographic questionnaire to gather information on age, gender, experience with 3D animation and screen size.

3D models from the pose dataset presented in Section 6.3.2 were used in this study. We randomly split our variation dataset in 4 subsets of 30 questions. Each participant was presented with only one of these subsets. For each question, observers were tasked to select which of two characters best imitated the pose of a source character displayed in the upper left of the screen. The choices were deformation transfer results of a new identity

(displayed in the screen lower left) to the source pose. One displayed similar self-contacts as the source, and the other had one self-contact released, presented randomly left or right. In the remainder of this chapter, we call these choices the "original pose" and the "pose variation" respectively. The display is illustrated in Figure 6.5. Characters were displayed with no surrounding environment, in a short video (5 seconds) rendered using Blender, rotating around the 3D model in order to present different angles of the pose to the observers. The video could be controlled with a slider and paused, in order to observe a chosen angle.

For each question, observers were asked to report their confidence in their response on a scale from 1 (not confident at all) to 5 (very confident).

6.4.3 Results

In this section, we present descriptive statistics on the results of this study. We also test for statistical significance of the effects of our different stimuli. We tested for potential main effect of a specific factor on our results using one-way Analysis of Variance (ANOVA), and interactions between several factors using n-way ANOVA. When an effect is found for a factor, we explore it further using Tukey post-hoc test to compare pair-wise means. We chose a statistical significance threshold of 5%, meaning that we consider an effect significant if the probability that the difference between means is due to chance is under 5%.

Pose Variation Selection Rate. We first study the rate of pose variation selection, *i.e.* the proportion of observers choosing the pose variation with a contact present in the source pose released. As a reminder, if participants could always detect a difference between the original and the variation poses this rate would be 0%, while in the opposite case if participants could always differentiate them the rate would be close to 50%, (chance level). Observers chose the pose variation in 27.16% of answers. We averaged for each observer the number of cases they chose the pose variation, obtaining an independent pose variation selection rate per observer. We conducted a one-sample t-test in order to determine if the observed selection rate was significantly lower than a random decision that would give a 50% selection rate. The result of this test showed that this difference was indeed significant ($p = 3.159 \times 10^{-29}$).

For each question, we gathered between 17 and 24 answers (mean 20). We averaged these answers to obtain a selection rate for each pose variation. We found a significant main effect of category of self-contacts (*i.e. hand* or *arm*) on variation selection ($p = 0.00006$). This effect was confirmed by the

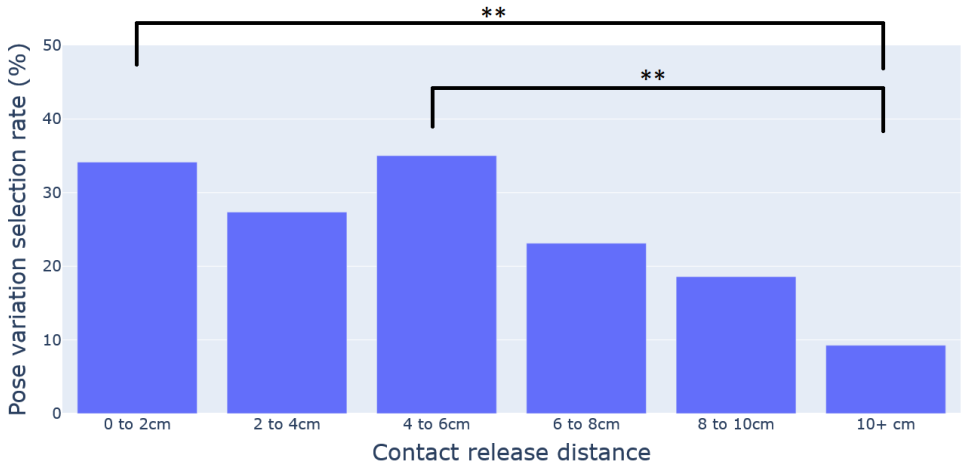


Figure 6.6: Rate at which observers selected the pose variation with a contact present in the source released, depending on the contact release distance. 50% corresponds to the chance level

post-hoc test ($p = 0.0001$). These tests show that observers chose the pose variation significantly less often for the category *hand* (18.00% of answers) than for the category *arm* (34.05% of answers). We also found a main effect of contact release distance, when grouping the questions by intervals of 2cm of release distance as shown in Figure 6.6 ($p = 0.0066$). Post-hoc Tukey test showed that this effect was only significant between release distances inferior to 2cm and superior to 10cm, and between distances in the 4 to 6cm interval and distances superior to 10cm. We found no significant effect of the target identity in the pose variation selection rate ($p = 0.6220$), and no interaction between factors.

Confidence Level. On average for all questions, the average confidence score reported is 3.42 ± 1.29. Similarly to the previous paragraph, we average the confidence score for all answers given to one question, and obtain a mean confidence score per question. We found a significant main effect of category on the average confidence score ($p = 0.0019$). For the category *hand*, observers reported a significantly higher confidence in their answers (3.66 ± 1.29) than for category *arm* (3.24 ± 1.26). This result was validated by the post-hoc Tukey test ($p = 0.019$). We also found a significant main effect of release distance on confidence scores ($p = 8.45 \times 10^{-11}$). The results of the Tukey test showed that the confidence score significantly increased between groups of release distance separated by at least 2cm, except between distances in the 4 to 6cm interval and the 8 to 10cm interval,

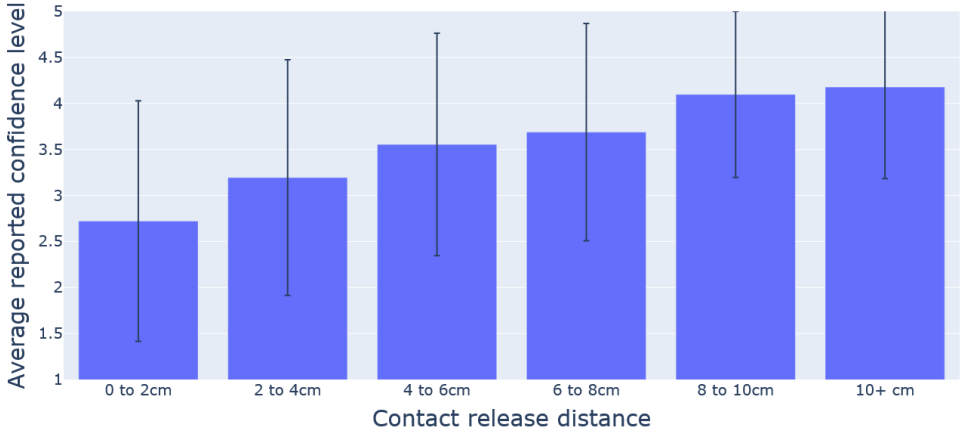


Figure 6.7: Average confidence level reported by observers, depending on the contact release distance

Choice \ Confidence score	1	2	3	4	5
Pose Variation	105	132	161	186	56
Original Pose	165	185	303	574	489

Table 6.1: Number of observers reporting a given confidence score after choosing the original pose (same self-contacts as the source pose) or the pose variation (one self-contact released)

and between distances in the 6 to 8cm interval and distances superior to 10cm (see Figure 6.7). We found no significant effect of the target identity ($p = 0.2797$), and no interaction between factors.

We also observed that participants reported different confidence levels depending on their choice. We observe an average confidence score of 2.93 ± 1.22 when the observer chose the pose variation, and 3.60 ± 1.27 when the observer chose the original pose with similar contacts than the source pose. We show in details reported confidence levels of observers depending on their choice in Table 6.1. We observe that observers tend to report higher confidence levels when choosing the original pose over the pose variation. This observation is shown to be significant by a chi-square independence test between the confidence scores and the choice of the observers ($p = 8.95 \times 10^{-31}$). For questions in the category *hand*, when choosing the pose variation (respectively the original pose), observers reported an average confidence of 2.81 ± 1.30 (respectively 3.84 ± 1.21). For the category *arm*, observers reported when choosing the pose variation (respectively the

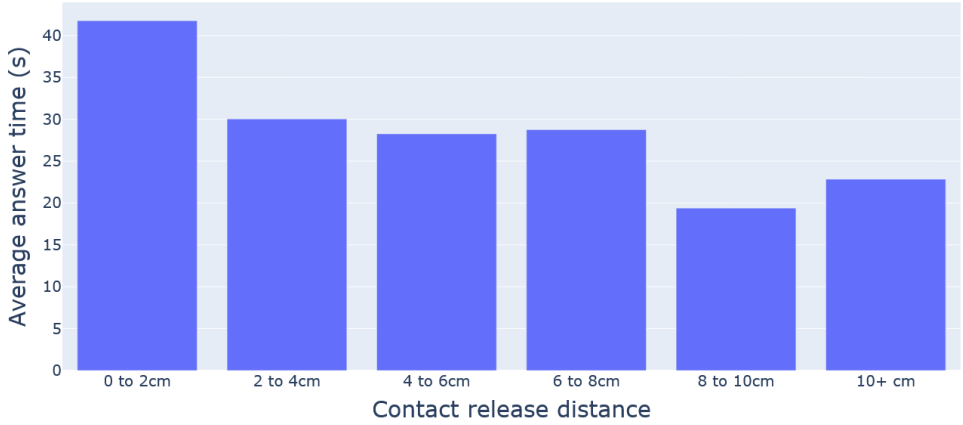


Figure 6.8: Average answer time, depending on the contact release distance

original pose) a confidence level of 2.97 ± 1.19 (respectively 3.37 ± 1.27). In both categories, these higher confidence scores when choosing the original pose are significant (chi-square for category *hand*: $p = 4.58 \times 10^{-20}$, *arm*: $p = 7.38 \times 10^{-9}$). However, we observed that the increase in reported confidence is significantly higher ($p = 0.0002$) for the category *hand* (1.03) than for the category *arm* (0.4).

Time Spent. Participants required on average 15 minutes to finish the study. For each question, the average answer time was 30.16 seconds. We found a main effect of the category on answer time ($p = 0.0158$), confirmed by the post-hoc Tukey test ($p = 0.0158$). Observers thus answered significantly faster for questions in the category *hand* (26.51s) than for questions in the category *arm* (32.92s). We also found a main effect of contact release distance on answer time ($p = 0.0014$). Tukey post-hoc test showed that observers took longer to answers for questions with release distance between 0 and 2cm than for questions in the intervals 2 to 4cm, 4 to 6cm, 8 to 10cm, and $>$ 10cm (see Figure 6.8). We found no effect of the target identity on answer time, and no interaction between factors.

Demographic Data. We found no significant effect of the demographic data collected in the study on the choice of transfer result, or on the confidence level reported.

6.5 Discussion

In this study, we investigated the impact of self-contacts on the perception of pose similarity. We presented observers with characters with a source pose and a target identity, and proposed two new characters with the target identity performing a pose similar to the source pose. One of these transfer results had similar self-contacts than the source pose, while the other had one self-contact released. Observers were tasked to choose which transfer result best imitated the source pose, and to report their confidence in their response on a 1 to 5 scale.

When presented with two possible transfer results, observers chose the variation with a self-contact present in the source pose released in only 27.16% of cases. This selection rate was shown to be significantly lower than a random 50% choice. This tends to confirm our hypothesis **H1**: Pose variations with released self-contacts were perceived as different from the source pose in most cases. Moreover, observers were significantly more confident in their response when choosing the original pose with the same self-contacts than the source pose (average confidence level: 3.60) than when choosing the pose variation (average confidence level: 2.93). Choosing the pose with the same contacts was thus perceived as the right choice with more confidence, further validating our hypothesis.

The category of self-contact of the questions, *i.e.* whether the self-contact removed in the pose variation involved the *hand* or the *arm*, had a clear effect on responses. Observers chose significantly less often the pose variation in the category *hand* (18.00%) than in the category *arm* (34.05%). Observers were also more confident in their answer in the category *hand* (average confidence level 3.66) than in the category *arm* (3.24). This suggests that observers were more confident in their choice, and perceived more easily when there was a difference in the hand contacts than when there was a difference in the arm contacts. Moreover, while observers are in general more confident when choosing the original pose over the pose variation, this confidence increase is significantly higher for the category *hand* (1.03) than for the category *arm* (0.4). This suggests that observers felt the pose with similar contacts was the right choice with more confidence when the contact involved the *hand* over the *arm*. Observers were also significantly faster in their response when the question was in the category *hand* (26.5s) than for the category *arm* (32.92s). This suggests that observers found it easier to chose when there was a difference in hand contacts. These results confirm our hypothesis **H2**; pose variations of the category *hand* were perceived as different from the source pose more often than for the category *arm*.

We also found an effect on our results of the contact release distance,

i.e. the average distance between vertices involved in the self-contact in the pose variation with released self-contact. Observers tended to choose the pose variation less often when the release distance was very high. They reported increasing confidence levels when the release distance augmented by several centimeters, and took significantly longer to answer when the release distance was very low. All these results suggest that observers considered a high release distance as a different pose, and had trouble choosing which proposition best imitated the source pose when the release distance was low. An interesting question would thus be to determine a threshold above which self-contacts are perceived as released by the observers. However as we did not control exactly the release distance in our generated examples, our data is not suitable for this exploration. An interesting future direction would be to duplicate our pose variations by applying increasing release distances to each. By presenting observers with similar poses with a self-contact released at 1, 2, 3, 4, etc. centimeters, we should be able to determine the threshold above which the self-contact is perceived as released, depending on the body parts involved in the contact.

We found no effect of the identity of the target character on our results. As we focused in this study on the effect of the body parts involved in the contact, our study only tested 4 target identities, and only explored transfers from a source pose with an average identity. Another study design focusing on the impact of identity would be necessary to understand which self-contacts are important for a given morphology. As discussed in Section 3.3 and illustrated in Figure 3.6, we argue that some contacts are only present to adapt to the morphology of the characters. To test this hypothesis, an interesting future direction would be to design a similar study where observers are presented with transfer results from an extreme morphology to another, and not only from an average to an extreme morphology. Presented pose variation should also contain new self-contacts absent in the pose in addition to released self-contacts, *e.g.* new contacts between the elbows and the torso for character with a larger torso. Following the observations of Section 3.3, we expect that new self-contacts absent in the source would be more acceptable to observers for bigger morphologies, and conversely that released self-contacts would be more acceptable for thinner morphologies.

6.6 Conclusion

In this Chapter, we explored the importance of self-contacts to the meaning of the pose. We designed a perceptive study to measure the perception of

pose similarity between a target and two poses with different self-contacts. We found that poses with different self-contacts were considered different in the majority of cases. Moreover, we found that the body parts involved in the self-contact had an impact on its importance. Poses with a difference in contacts involving the hands were perceived as different more often than poses with a difference in contacts involving the arms. In deformation transfer applications preserving constraints, such as Chapter 4, preserving all self-contacts involving the hands while allowing contacts involving the arms to change should thus give results that better preserve the meaning of the poses.

Our results confirm that self-contacts are a meaningful property of the pose, and highlight parameters that impact the importance of a specific self-contact. However, further exploration must be conducted to be able to robustly choose which contacts to preserve in a deformation transfer application. An interesting direction for future work is to study the importance of the morphology of characters in self-contact preservation. Another interesting direction that was not explored in our study is the second body part involved in the contact; while we compared self-contacts involving the hand or the arm and another body part, we did not compare the relative importance of contacts involving *e.g.* the torso, the head or the thighs of the characters.

As shown in related studies on perception of pose equivalences, simple similarity metrics often do not correlate with human perception [Tang et al., 2008, Durupinar, 2021]. Moreover, very few studies explored pose equivalences between characters with very different morphologies. Our results combined with further studies would provide tools to build a pose similarity metric between characters with varying morphologies. This metric would greatly help evaluating or designing deformation transfer approaches. It could in particular be used to select constraints to be preserved during transfer in methods such as the ones presented in Chapters 4 and 5.

Chapter 7

Conclusion

7.1 Summary

Automatically creating animations of new virtual characters from existing animations has become an important step of 3D content creation in medias such as video games, animation movies or virtual reality. In this thesis, we proposed new approaches to the deformation transfer problem. We specifically focused on adapting poses to new morphologies, while preserving constraints that define the meaning of the pose. For this purpose, we explored the Identity Transfer strategy, *i.e.* the idea of changing the identity of a character while preserving its pose instead of reposing a character with a new identity. We showed that this strategy can be applied to obtain state-of-the-art transfer results. Moreover, we demonstrated that this approach helps alleviate the difficult problem of preserving pose equivalence in the transfer result, as the method directly considers the source character's pose. In order to better understand pose equivalences, we focused on the impact of self-contacts for the meaning of the pose. We showed that preserving self-contacts present in the source pose during transfer resulted in realistic and similar poses adapted to the new morphology in most cases. We further demonstrated that this was coherent with human perception, as observers considered removing a self-contact from a pose a significant difference in most cases.

In Chapter 4 we proposed an implementation of the Identity Transfer strategy in an optimization-based framework. Our method iteratively deforms the surface of a character in a specific pose in order to match its identity to a target character, while applying minor corrections to the pose in order to avoid collisions and preserve existing self-contacts. We proposed new energy functions based on parameters describing the identity of char-

acters: near-isometry and body part near-rigidity. We showed that this method is able to adapt complex poses to extreme identities. Moreover, this method can be applied with minimal adaptation to a wide variety of shape classes, such as minimally or casually dressed humans, or animals.

In Chapter 5 we presented a data-driven strategy for identity transfer. We proposed a deep encoder-decoder based architecture that predicts the identity deformation between the source character already in the correct pose and the result with the same pose and a new identity. We leveraged identity parameters used in the previous chapter to create self-supervised identity losses. Our architecture can thus be trained in a weakly supervised setting, with as little as 1% pose supervision needed. The self-supervised losses also allowed for inference time fine tuning, which greatly improves the results. We showed that this approach better generalizes to poses unseen at training time than the state-of-the-art, and is able to transfer complex identities far from the training set, such as casually dressed humans.

Finally, in Chapter 6 we explored the impact of self-contacts on pose equivalences perceived by human observers. We designed a perceptive study in which observers were presented with several characters performing similar poses up to a specific self-contacts, and were asked to select which pose best imitated a target pose. Our study showed that observers consider poses with different self-contacts as not equivalent in most cases. Moreover, we showed that the body part involved in the contact had an impact on this effect: self-contacts involving the hands of the characters were considered more important to pose equivalences than others. This study is a first step towards automatically selecting which self-contacts must be preserved for pose equivalences between characters with different morphologies.

7.2 Limitations and Future Works

The deformation transfer methods presented in this thesis (Chapter 4 and 5) require inputs to be meshes in vertex-to-vertex correspondence. In our experiments, we tackled this problem by using inputs from databases fitted to a common template, and fitting this template to the meshes with different topologies. This pre-processing is light and only needs to be applied once per mesh. Moreover it could be made automatic, using some compatible re-meshing methods such as [Kraevoy and Sheffer, 2004, Yang et al., 2018]. This is hence simpler and faster than standard rigging processes traditionally used in computer animation. However, this condition on the input limits the possible applications of our methods. While an additional template fitting step can be used, it is not ideal as it adds complexity and

a new source of errors to the methods. Recently, a lot of state-of-the-art methods have moved past the need for common templates in their inputs, in particular for deformation transfer applications [Wang et al., 2020, Cosmo et al., 2020, Lombardi et al., 2021]. Therefore, an important future direction for our work is to adapt the methods to inputs with arbitrary resolution. The need for the common template in our approaches stems from different constraints that could be solved in this future work:

- **Vertex based architecture**. The first constraint is that our methods compute features at the vertex level. In the optimization based approach of Chapter 4, the deformation is computed for each vertex, which are then moved in the direction of their normal. The mesh fitting problem aims to approximate the geometry of an input mesh to another with a different topology [Yeh et al., 2010]. A similar approach could be applied in our method to compute the deformation of the source that would best approximate the local geometry of the target with no vertex correspondences. In the deep-learning method described in Chapter 5, we compute the network's features at the vertex level using spiral convolutions [Bouritsas et al., 2019]. Other feature computation methods that do not depend on vertices number and ordering could be used in a similar encoder-decoder architecture. [Wang et al., 2020, Cosmo et al., 2020] for example use PointNet [Qi et al., 2017] and can thus theoretically take as inputs point clouds of arbitrary size. Another recent and promising approach is the use of implicit functions describing 3D surfaces. This approach has been used to train deep learning networks with inputs of varying topologies (*e.g.* [Park et al., 2019, Mescheder et al., 2019]), and has already been applied successfully to the deformation transfer problem [Lombardi et al., 2021].

- **Identity parameters defined at vertex level** The second constraint is that losses based on the identity parameters proposed in Section 3.2.2 require the evaluated mesh and the target identity to be in vertex-to-vertex correspondence, *e.g.* same ordering of the neighbours of each vertex for the near-isometry. This problem could be alleviated by using other isometry and body part rigidity descriptors that do not need a common template, such as *e.g.* geodesic distances for the isometry [Cosmo et al., 2020].

In addition to a common template, our methods of Chapters 4 and 5 require inputs to be segmented in rigidly deforming body parts. The

segmentation used in both methods (see Figure 3.4a) was done by hand before applying the method, which is a heavy pre-processing step. However, as it must be done only once for the chosen template, this can be considered as negligible in the long term. In future works removing the need for a common template, automatic body part segmentation methods could be applied such as *e.g.* [Cuzzolin et al., 2007, Varol et al., 2018].

The identity transfer strategy explored in this thesis requires being able to parameterize the identity independently from the pose. We argue that identity should be recognizable across different poses in Section 3.2.1, and propose to use near-isometry and body part rigidity as identity descriptors. However, while these parameters give satisfying results in our methods, they are not perfectly disentangled from the pose of the characters: surface deformations such as muscle contractions or breathing cause non-rigid deformation of body parts, and the near-isometry assumption does not hold close to the joints. This can cause artifacts such as the unrealistic surface deformations discussed in Section 4.5, and is the reason for the need for pose supervision in our deep learning method (Chapter 5). An interesting future direction is thus designing a new identity parameter completely disentangled from pose information, through *e.g.* statistical analysis of human shapes or observers identity recognition. This would allow us to train our neural identity transfer method in a completely self-supervised setting, which would further improve our results on complex poses unseen at training.

Finally, a limitation that our methods share with the state-of-the-art on deformation transfer is the lack of a clear definition of pose equivalences between characters with different morphologies. This results in imperfect heuristics to preserve the meaning of the pose during transfer, such as *e.g.* preserving all self-contacts (Chapter 4) or preserving distances between body parts [Liu et al., 2018, Jin et al., 2018]. This also makes quantitative evaluation of deformation transfer methods difficult. Indeed, these methods are usually evaluated by computing a distance between their result and a ground-truth (*e.g.* Chapter 5, [Zhou et al., 2020, Cosmo et al., 2020]). These ground-truths are either captures of real characters performing the pose (*e.g.* FAUST [Bogo et al., 2014] for [Cosmo et al., 2020]), which can present inter-personal and contextual variations (see Figure 1.4), or result of other deformation transfer approaches (*e.g.* As-Rigid-As-Possible deformations for [Zhou et al., 2020]), which can thus present limitations from these methods. Moreover, measuring equivalence to this ground truth with metrics such as Euclidean distances was shown to be uncorrelated with human perception of pose equivalences [Tang et al., 2008, Durupinar, 2021]. An important future direction for the field of deformation transfer is thus

the exploration of pose equivalences for characters with different morphologies. In Section 3.2.1, we argued that poses with small variations can still be considered as similar (see Figure 1.4). Therefore, instead of trying to determine if two poses are strictly equivalent or not, an interesting direction would be to determine *how similar* the poses are. In Chapter 6, we proposed a first step towards understanding equivalences between poses of different characters, focusing on the importance of self-contacts. Further studies, focusing on self-contacts (see Section 6.6) and other pose equivalence parameters, could be conducted in order to define and evaluate such a pose distance metric.

Bibliography

[mix,] Adobe's mixamo. https://www.mixamo.com/. Accessed: 13-01-2022.

[Aberman et al., 2020a] Aberman, K., Li, P., Lischinski, D., Sorkine-Hornung, O., Cohen-Or, D., and Chen, B. (2020a). Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics*, 39(4):62–1.

[Aberman et al., 2020b] Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., and Chen, B. (2020b). Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics*, 39(4):64–1.

[Abrevaya et al., 2018a] Abrevaya, V. F., Wuhrer, S., and Boyer, E. (2018a). Multilinear autoencoder for 3d face model learning. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 1–9. IEEE.

[Abrevaya et al., 2018b] Abrevaya, V. F., Wuhrer, S., and Boyer, E. (2018b). Spatiotemporal modeling for efficient registration of dynamic 3d faces. In *2018 International Conference on 3D Vision*, pages 371–380. IEEE.

[Akhter et al., 2010] Akhter, I., Sheikh, Y., Khan, S., and Kanade, T. (2010). Trajectory space: A dual representation for nonrigid structure from motion. *Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1442–1456.

[Akhter et al., 2012] Akhter, I., Simon, T., Khan, S., Matthews, I., and Sheikh, Y. (2012). Bilinear spatiotemporal basis models. *Transactions on Graphics*, 31(2):1–12.

[Al-Asqhar et al., 2013] Al-Asqhar, R. A., Komura, T., and Choi, M. G. (2013). Relationship descriptors for interactive motion adaptation. In *ACM SIGGRAPH/SCA*, SCA '13, pages 45–53.

[Al Borno et al., 2018] Al Borno, M., Righetti, L., Black, M. J., Delp, S. L., Fiume, E., and Romero, J. (2018). Robust physics-based motion retargeting with realistic body shapes. In *Computer Graphics Forum*, volume 37, pages 81–92. Wiley Online Library.

[Allen et al., 2003] Allen, B., Curless, B., and Popović, Z. (2003). The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics*, 22(3):587–594.

[Anguelov et al., 2005] Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416.

[Arik et al., 2018] Arik, S. O., Chen, J., Peng, K., Ping, W., and Zhou, Y. (2018). Neural voice cloning with a few samples. In *Conference on Neural Information Processing Systems*.

[Aubry et al., 2011] Aubry, M., Schlickewei, U., and Cremers, D. (2011). Pose-consistent 3d shape segmentation based on a quantum mechanical feature descriptor. In *Joint Pattern Recognition Symposium*, pages 122–131. Springer.

[Aumentado-Armstrong et al., 2021] Aumentado-Armstrong, T., Tsogkas, S., Dickinson, S., and Jepson, A. (2021). Disentangling geometric deformation spaces in generative latent shape models. *arXiv preprint arXiv:2103.00142*.

[Aumentado-Armstrong et al., 2019] Aumentado-Armstrong, T., Tsogkas, S., Jepson, A., and Dickinson, S. (2019). Geometric disentanglement for generative latent shape models. In *International Conference on Computer Vision*, pages 8181–8190.

[Baciu and Iu, 2006] Baciu, G. and Iu, B. K. C. (2006). Motion retargeting in the presence of topological variations. *Computer Animation and Virtual Worlds*, 17(1):41–57.

[Ballan et al., 2012] Ballan, L., Taneja, A., Gall, J., Gool, L. V., and Pollefeys, M. (2012). Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision*, pages 640–653. Springer.

[Baran and Popović, 2007] Baran, I. and Popović, J. (2007). Automatic rigging and animation of 3d characters. *ACM Transactions on graphics*, 26(3):72–es.

[Baran et al., 2009] Baran, I., Vlasic, D., Grinspun, E., and Popović, J. (2009). Semantic deformation transfer. In *ACM SIGGRAPH*, pages 1–6.

[Ben-Chen et al., 2009] Ben-Chen, M., Weber, O., and Gotsman, C. (2009). Spatial deformation transfer. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 67–74.

[Bernardin et al., 2017] Bernardin, A., Hoyet, L., Mucherino, A., Gonçalves, D., and Multon, F. (2017). Normalized euclidean distance matrices for human motion retargeting. In *Proceedings of the Tenth International Conference on Motion in Games*, pages 1–6.

[Bogo et al., 2014] Bogo, F., Romero, J., Loper, M., and Black, M. J. (2014). FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, Piscataway, NJ, USA. IEEE.

[Bogo et al., 2017] Bogo, F., Romero, J., Pons-Moll, G., and Black, M. J. (2017). Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6233–6242.

[Bouaziz et al., 2013] Bouaziz, S., Wang, Y., and Pauly, M. (2013). Online modeling for realtime facial animation. *ACM Transactions on Graphics*, 32(4):1–10.

[Boukhayma et al., 2017] Boukhayma, A., Franco, J.-S., and Boyer, E. (2017). Surface motion capture transfer with gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 184–192.

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

[Chan et al., 2019] Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. (2019). Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942.

[Chen et al., 2009] Chen, C., Zhuang, Y., Xiao, J., and Liang, Z. (2009). Perceptual 3d pose distance estimation by boosting relational geometric features. *Computer Animation and Virtual Worlds*, 20(2-3):267–277.

[Chen et al., 2010] Chen, L., Huang, J., Sun, H., and Bao, H. (2010). Cage-based deformation transfer. *Computers & Graphics*, 34(2):107–118.

[Chen et al., 2013] Chen, Y., Liu, Z., and Zhang, Z. (2013). Tensor-based human body modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–112.

[Chen and Zhang, 2019] Chen, Z. and Zhang, H. (2019). Learning implicit fields for generative shape modeling. In *Conference on Computer Vision and Pattern Recognition*, pages 5939–5948.

[Cheung et al., 2004] Cheung, G. K., Baker, S., Hodgins, J., and Kanade, T. (2004). Markerless human motion transfer. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 373–378. IEEE.

[Choi and Ko, 2000] Choi, K.-J. and Ko, H.-S. (2000). Online motion retargetting. *The Journal of Visualization and Computer Animation*, 11(5):223–235.

[Chua et al., 2003] Chua, P. T., Crivella, R., Daly, B., Hu, N., Schaaf, R., Ventura, D., Camill, T., Hodgins, J., and Pausch, R. (2003). Training for physical tasks in virtual environments: Tai chi. In *IEEE Virtual Reality, 2003. Proceedings.*, pages 87–94. IEEE.

[Cosmo et al., 2020] Cosmo, L., Norelli, A., Halimi, O., Kimmel, R., and Rodolà, E. (2020). Limp: Learning latent shape representations with metric preservation priors. In *European Conference on Computer Vision*, pages 19–35. Springer.

[Cosmo et al., 2019] Cosmo, L., Panine, M., Rampini, A., Ovsjanikov, M., Bronstein, M. M., and Rodola, E. (2019). Isospectralization, or how to hear shape, style, and correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7529–7538.

[Craven and Wahba, 1978] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

[Cuzzolin et al., 2007] Cuzzolin, F., Mateus, D., Boyer, E., and Horaud, R. (2007). Robust spectral 3d-bodypart segmentation along time. In *Workshop on Human Motion*, pages 196–211. Springer.

[De Boor, 1978] De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

[Defferrard et al., 2016] Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*, 29:3844–3852.

[Deng et al., 2020] Deng, B., Lewis, J. P., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., and Tagliasacchi, A. (2020). Nasa neural articulated shape approximation. In *European Conference on Computer Vision*, pages 612–628. Springer.

[Dey et al., 2012] Dey, T. K., Ranjan, P., and Wang, Y. (2012). Eigen deformation of 3d models. *The Visual Computer*, 28(6):585–595.

[Durupinar, 2021] Durupinar, F. (2021). Perception of human motion similarity based on laban movement analysis. In *ACM Symposium on Applied Perception 2021*, pages 1–7.

[Durupinar et al., 2016] Durupinar, F., Kapadia, M., Deutsch, S., Neff, M., and Badler, N. I. (2016). Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Transactions on Graphics*, 36(1):1–16.

[Efros et al., 2003] Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Computer Vision, IEEE International Conference on*, volume 3, pages 726–726. IEEE Computer Society.

[Egerstedt and Martin, 2001] Egerstedt, M. and Martin, C. F. (2001). Optimal trajectory planning and smoothing splines. *Automatica*, 37(7):1057–1064.

[Elad and Kimmel, 2001] Elad, A. and Kimmel, R. (2001). Bending invariant representations for surfaces. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE.

[Esser et al., 2018] Esser, P., Sutter, E., and Ommer, B. (2018). A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866.

[Fribourg et al., 2018] Fribourg, R., Argelaguet, F., Hoyet, L., and Lécuyer, A. (2018). Studying the sense of embodiment in vr shared experiences. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 273–280. IEEE.

[Gao et al., 2018] Gao, L., Yang, J., Qiao, Y.-L., Lai, Y.-K., Rosin, P. L., Xu, W., and Xia, S. (2018). Automatic unpaired shape deformation transfer. In *SIGGRAPH Asia 2018 Technical Papers*, page 237. ACM.

[Gatys et al., 2016] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2414–2423.

[Gkalelis et al., 2009] Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., and Pitas, I. (2009). The i3dpost multi-view and 3d human action/interaction database. In *2009 Conference for Visual Media Production*, pages 159–168. IEEE.

[Gleicher, 1998] Gleicher, M. (1998). Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer Graphics and Interactive Techniques*, pages 33–42. ACM.

[Gong et al., 2019] Gong, S., Chen, L., Bronstein, M., and Zafeiriou, S. (2019). Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

[Harada et al., 2004] Harada, T., Taoka, S., Mori, T., and Sato, T. (2004). Quantitative evaluation method for pose and motion similarity based on human perception. In *4th IEEE/RAS International Conference on Humanoid Robots, 2004.*, volume 1, pages 494–512. IEEE.

[Hasler et al., 2009] Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., and Seidel, H.-P. (2009). A statistical model of human pose and body shape. In *Computer Graphics Forum*, volume 28, pages 337–346. Wiley Online Library.

[Hassan et al., 2021] Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., and Black, M. J. (2021). Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718.

[Hasson et al., 2019] Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M. J., Laptev, I., and Schmid, C. (2019). Learning joint reconstruction of hands and manipulated objects. *Conference on Computer Vision and Pattern Recognition*.

[Hecker et al., 2008] Hecker, C., Raabe, B., Enslow, R. W., DeWeese, J., Maynard, J., and van Prooijen, K. (2008). Real-time motion retargeting to highly varied user-created morphologies. *ACM Transactions on Graphics*, 27(3):1–11.

[Ho et al., 2010] Ho, E. S. L., Komura, T., and Tai, C.-L. (2010). Spatial relationship preserving character motion adaptation. *ACM Transactions on Graphics*, 29(4):33:1–33:8.

[Ho et al., 2014] Ho, E. S. L., Wang, H., and Komura, T. (2014). A multi-resolution approach for adapting close character interaction. In *ACM Virtual Reality Software and Technology*, VRST '14, pages 97–106.

[Hodgins et al., 1998] Hodgins, J. K., O'Brien, J. F., and Tumblin, J. (1998). Perception of human motion with different geometric models. *IEEE Transactions on Visualization and Computer Graphics*, 4(4):307–316.

[Hornung et al., 2007] Hornung, A., Dekkers, E., and Kobbelt, L. (2007). Character animation from 2d pictures and 3d motion data. *ACM Transactions on Graphics*, 26(1):1–es.

[Hoyet et al., 2013] Hoyet, L., Ryall, K., Zibrek, K., Park, H., Lee, J., Hodgins, J., and O'sullivan, C. (2013). Evaluating the distinctiveness and attractiveness of human motions on realistic virtual bodies. *ACM Transactions on Graphics*, 32(6):1–11.

[Huang et al., 2013] Huang, C.-H., Boyer, E., and Ilic, S. (2013). Robust human body shape and pose tracking. In *2013 International Conference on 3D Vision*, pages 287–294. IEEE.

[Huang and Belongie, 2017] Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510.

[Huang et al., 2018] Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, pages 172–189.

[Ionescu et al., 2013] Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2013). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

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

[Jain et al., 2010] Jain, A., Thormählen, T., Seidel, H.-P., and Theobalt, C. (2010). Moviereshape: Tracking and reshaping of humans in videos. *ACM Transactions on Graphics*, 29(6):1–10.

[Jia et al., 2018] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Conference on Neural Information Processing Systems*.

[Jiang et al., 2020] Jiang, B., Zhang, J., Cai, J., and Zheng, J. (2020). Disentangled human body embedding based on deep hierarchical neural network. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2560–2575.

[Jin et al., 2018] Jin, T., Kim, M., and Lee, S.-H. (2018). Aura mesh: Motion retargeting to preserve the spatial relationships between skinned characters. In *Computer Graphics Forum*, volume 37, pages 311–320. Wiley Online Library.

[Johansson, 1973] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211.

[Kadleček et al., 2016] Kadleček, P., Ichim, A.-E., Liu, T., Křivánek, J., and Kavan, L. (2016). Reconstructing personalized anatomical models for physics-based body animation. *ACM Transactions on Graphics (TOG)*, 35(6):1–13.

[Kappel et al., 2021] Kappel, M., Golyanik, V., Elgharib, M., Henningson, J.-O., Seidel, H.-P., Castillo, S., Theobalt, C., and Magnor, M. (2021). High-fidelity neural human motion transfer from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1541–1550.

[Kemelmacher-Shlizerman et al., 2010] Kemelmacher-Shlizerman, I., Sankar, A., Shechtman, E., and Seitz, S. M. (2010). Being john malkovich. In *European Conference on Computer Vision*, pages 341–353. Springer.

[Klokov and Lempitsky, 2017] Klokov, R. and Lempitsky, V. (2017). Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 863–872.

[Komaritzan and Botsch, 2019] Komaritzan, M. and Botsch, M. (2019). Fast projective skinning. In *Motion in Games*, pages 1–10. ACM.

[Kovnatsky et al., 2013] Kovnatsky, A., Bronstein, M. M., Bronstein, A. M., Glashoff, K., and Kimmel, R. (2013). Coupled quasi-harmonic bases. In *Computer Graphics Forum*, volume 32, pages 439–448. Wiley Online Library.

[Kraevoy and Sheffer, 2004] Kraevoy, V. and Sheffer, A. (2004). Cross-parameterization and compatible remeshing of 3d models. *ACM Transactions on Graphics*, 23(3):861–869.

[Kulpa and Multon, 2005] Kulpa, R. and Multon, F. (2005). Fast inverse kinematics and kinetics solver for human-like figures. In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, pages 38–43. IEEE.

[Kulpa et al., 2005] Kulpa, R., Multon, F., and Arnaldi, B. (2005). Morphology-independent representation of motions for interactive human-like animation. In *Eurographics*.

[Laban, 1928] Laban, R. (1928). Tanzkomposition und schrifttanz.

[Le Callennec and Boulic, 2006] Le Callennec, B. and Boulic, R. (2006). Robust kinematic constraint detection for motion data. In *ACM SIGGRAPH/SCA*, SCA '06, pages 281–290.

[Le Naour et al., 2019] Le Naour, T., Courty, N., and Gibet, S. (2019). Skeletal mesh animation driven by few positional constraints. *Computer Animation and Virtual Worlds*, 30(3-4):e1900.

[Lee and Shin, 1999] Lee, J. and Shin, S. Y. (1999). A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer Graphics and Interactive Techniques*, pages 39–48.

[Lévy, 2006] Lévy, B. (2006). Laplace-beltrami eigenfunctions towards an algorithm that" understands" geometry. In *IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*, pages 13–13. IEEE.

[Lifkooee et al., 2019] Lifkooee, M. Z., Liu, C., Liang, Y., Zhu, Y., and Li, X. (2019). Real-time avatar pose transfer and motion generation using locally encoded laplacian offsets. *Journal of Computer Science and Technology*, 34(2):256–271.

[Liu et al., 2019] Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., and Gao, S. (2019). Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913.

[Liu et al., 2018] Liu, Z., Mucherino, A., Hoyet, L., and Multon, F. (2018). Surface based motion retargeting by preserving spatial relationship. In *Motion in Games*, MIG '18, pages 7:1–7:11. ACM.

[Lombardi et al., 2021] Lombardi, S., Yang, B., Fan, T., Bao, H., Zhang, G., Pollefeys, M., and Cui, Z. (2021). Latenthuman: Shape-and-pose disentangled latent representation for human bodies. In *2021 International Conference on 3D Vision*, pages 278–288. IEEE.

[Loper et al., 2015] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). Smpl: A skinned multi-person linear model. *ACM transactions on graphics*, 34(6):1–16.

[Lyard and Magnenat-Thalmann, 2008] Lyard, E. and Magnenat-Thalmann, N. (2008). Motion adaptation based on character shape. *Computer Animation and Virtual Worlds*, 19(3-4):189–198.

[Magnenat-Thalmann et al., 1988] Magnenat-Thalmann, N., Laperrire, R., and Thalmann, D. (1988). Joint-dependent local deformations for hand animation and object grasping. In *In Proceedings on Graphics interface'88*. Citeseer.

[Mahmood et al., 2019] Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451.

[Marin et al., 2020] Marin, R., Rampini, A., Castellani, U., Rodola, E., Ovsjanikov, M., and Melzi, S. (2020). Instant recovery of shape from spectrum via latent space connections. In *2020 International Conference on 3D Vision*, pages 120–129. IEEE.

[Marinoiu et al., 2016] Marinoiu, E., Papava, D., and Sminchisescu, C. (2016). Pictorial human spaces: A computational study on the human perception of 3d articulated poses. *International Journal of Computer Vision*, 119(2):194–215.

[McAdams et al., 2011] McAdams, A., Zhu, Y., Selle, A., Empey, M., Tamstorf, R., Teran, J., and Sifakis, E. (2011). Efficient elasticity for character skinning with contact and collisions. In *ACM SIGGRAPH*, pages 1–12.

[Mescheder et al., 2019] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. (2019). Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470.

[Mihajlovic et al., 2021] Mihajlovic, M., Zhang, Y., Black, M. J., and Tang, S. (2021). Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471.

[Molla et al., 2018] Molla, E., Debarba, H. G., and Boulic, R. (2018). Egocentric mapping of body surface constraints. *IEEE Transactions on Visualization and Computer Graphics*, 24(7):2089–2102.

[Monzani et al., 2000] Monzani, J.-S., Baerlocher, P., Boulic, R., and Thalmann, D. (2000). Using an intermediate skeleton and inverse kinematics for motion retargeting. In *Computer Graphics Forum*, volume 19, pages 11–19. Wiley Online Library.

[Müller et al., 2021] Müller, L., Osman, A. A. A., Tang, S., Huang, C.-H. P., and Black, M. J. (2021). On self-contact and human pose. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

[Neophytou and Hilton, 2013] Neophytou, A. and Hilton, A. (2013). Shape and pose space deformation for subject specific animation. In *2013 International Conference on 3D Vision*, pages 334–341. IEEE.

[Olivier et al., 2020] Olivier, N., Hoyet, L., Danieau, F., Argelaguet, F., Avril, Q., Lecuyer, A., Guillotel, P., and Multon, F. (2020). The impact of stylization on face recognition. In *ACM Symposium on Applied Perception 2020*, pages 1–9.

[Osman et al., 2020] Osman, A. A., Bolkart, T., and Black, M. J. (2020). Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613. Springer.

[Park et al., 2019] Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174.

[Pavlakos et al., 2019] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 10975–10985.

[Pishchulin et al., 2017] Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., and Schiele, B. (2017). Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286.

[Pons-Moll et al., 2015] Pons-Moll, G., Romero, J., Mahmood, N., and Black, M. J. (2015). Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34(4):1–14.

[Pons-Moll and Rosenhahn, 2011] Pons-Moll, G. and Rosenhahn, B. (2011). Model-based pose estimation. In *Visual Analysis of Humans*, pages 139–170. Springer.

[Popović and Witkin, 1999] Popović, Z. and Witkin, A. (1999). Physically based motion transformation. In *Proceedings of the 26th annual conference on Computer Graphics and Interactive Techniques*, pages 11–20.

[Pražák et al., 2009] Pražák, M., McDonnell, R., Kavan, L., and O'SULLIVAN, C. (2009). A perception based metric for comparing human locomotion. *Eurographics Ireland*.

[Prilepin, 2020] Prilepin, E. (2017-2020). csaps cubic spline approximation. https://csaps.readthedocs.io/en/latest/index.html.

[Pumarola et al., 2019] Pumarola, A., Sanchez-Riera, J., Choi, G., Sanfeliu, A., and Moreno-Noguer, F. (2019). 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2242–2251.

[Qi et al., 2017] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 652–660.

[Ranjan et al., 2018] Ranjan, A., Bolkart, T., Sanyal, S., and Black, M. J. (2018). Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision*, pages 704–720.

[Rhodin et al., 2014] Rhodin, H., Tompkin, J., In Kim, K., Varanasi, K., Seidel, H.-P., and Theobalt, C. (2014). Interactive motion mapping for real-time character control. In *Computer Graphics Forum*, volume 33, pages 273–282. Wiley Online Library.

[Rhodin et al., 2015] Rhodin, H., Tompkin, J., Kim, K. I., De Aguiar, E., Pfister, H., Seidel, H.-P., and Theobalt, C. (2015). Generalizing wave gestures from sparse examples for real-time character control. *ACM Transactions on Graphics*, 34(6):1–12.

[Robinette et al., 1999] Robinette, K., Daanen, H., and Paquet, E. (1999). The caesar project: a 3-d surface anthropometry survey. In *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*, pages 380–386.

[Rong et al., 2008] Rong, G., Cao, Y., and Guo, X. (2008). Spectral mesh deformation. *The Visual Computer*, 24(7):787–796.

[Seo et al., 2003] Seo, H., Cordier, F., and Magnenat-Thalmann, N. (2003). Synthesizing animatable body models with parameterized shape modifications. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer Animation*, pages 120–125. Citeseer.

[Shi et al., 2020] Shi, M., Aberman, K., Aristidou, A., Komura, T., Lischinski, D., Cohen-Or, D., and Chen, B. (2020). Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics*, 40(1):1–15.

[Shi et al., 2007] Shi, X., Zhou, K., Tong, Y., Desbrun, M., Bao, H., and Guo, B. (2007). Mesh puppetry: cascading optimization of mesh deformation with inverse kinematics. In *ACM SIGGRAPH 2007 papers*, pages 81–es.

[Sigal et al., 2010] Sigal, L., Balan, A. O., and Black, M. J. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of Computer Vision*, 87(1-2):4.

[Sorkine and Alexa, 2007] Sorkine, O. and Alexa, M. (2007). As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116.

[Sridhar et al., 2016] Sridhar, S., Mueller, F., Zollhöfer, M., Casas, D., Oulasvirta, A., and Theobalt, C. (2016). Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*, pages 294–310. Springer.

[Su et al., 2015] Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–953.

[Sumner and Popović, 2004] Sumner, R. W. and Popović, J. (2004). Deformation transfer for triangle meshes. In *ACM Transactions on Graphics*, volume 23, pages 399–405. ACM.

[Tan et al., 2018] Tan, Q., Gao, L., Lai, Y.-K., and Xia, S. (2018). Variational autoencoders for deforming 3d mesh models. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5841–5850.

[Tanaka and Farah, 1993] Tanaka, J. W. and Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly journal of experimental psychology*, 46(2):225–245.

[Tang et al., 2008] Tang, J. K., Leung, H., Komura, T., and Shum, H. P. (2008). Emulating human perception of motion similarity. *Computer Animation and Virtual Worlds*, 19(3-4):211–221.

[Teschner et al., 2003] Teschner, M., Heidelberger, B., Müller, M., Pomerantes, D., and Gross, M. H. (2003). Optimized spatial hashing for collision detection of deformable objects. In *Vmv*, volume 3, pages 47–54.

[Teschner et al., 2005] Teschner, M., Kimmerle, S., Heidelberger, B., Zachmann, G., Raghupathi, L., Fuhrmann, A., Cani, M.-P., Faure, F., Magnenat-Thalmann, N., Strasser, W., et al. (2005). Collision detection for deformable objects. In *Computer Graphics Forum*, volume 24, pages 61–81. Wiley Online Library.

[Tretschk et al., 2020] Tretschk, E., Tewari, A., Zollhöfer, M., Golyanik, V., and Theobalt, C. (2020). Demea: Deep mesh autoencoders for non-rigidly deforming objects. In *European Conference on Computer Vision*, pages 601–617. Springer.

[Tzionas et al., 2016] Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., and Gall, J. (2016). Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193.

[Van der Maaten and Hinton, 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

[Varol et al., 2018] Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., and Schmid, C. (2018). Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision*, pages 20–36.

[Verma et al., 2018] Verma, N., Boyer, E., and Verbeek, J. (2018). Feastnet: Feature-steered graph convolutions for 3d shape analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2598–2606.

[Villegas et al., 2021] Villegas, R., Ceylan, D., Hertzmann, A., Yang, J., and Saito, J. (2021). Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9720–9729.

[Villegas et al., 2018] Villegas, R., Yang, J., Ceylan, D., and Lee, H. (2018). Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8639–8648.

[von Marcard et al., 2018] von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., and Pons-Moll, G. (2018). Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision*, pages 601–617.

[Wang et al., 2020] Wang, J., Wen, C., Fu, Y., Lin, H., Zou, T., Xue, X., and Zhang, Y. (2020). Neural pose transfer by spatially adaptive instance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5831–5839.

[Wuhrer et al., 2012] Wuhrer, S., Shu, C., and Xi, P. (2012). Posture-invariant statistical shape analysis using laplace operator. *Computers & Graphics*, 36(5):410–416.

[Xu et al., 2011] Xu, F., Liu, Y., Stoll, C., Tompkin, J., Bharaj, G., Dai, Q., Seidel, H.-P., Kautz, J., and Theobalt, C. (2011). Video-based characters: creating new human performances from a multi-view video database. In *ACM SIGGRAPH*, pages 1–10.

[Yang et al., 2016] Yang, J., Franco, J.-S., Hétroy-Wheeler, F., and Wuhrer, S. (2016). Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, pages 439–454. Springer.

[Yang et al., 2018] Yang, Y., Fu, X.-M., Chai, S., Xiao, S.-W., and Liu, L. (2018). Volume-enhanced compatible remeshing of 3d models. *IEEE Transactions on Visualization and Computer Graphics*, 25(10):2999–3010.

[Yeh et al., 2010] Yeh, I.-C., Lin, C.-H., Sorkine, O., and Lee, T.-Y. (2010). Template-based 3d model fitting using dual-domain relaxation. *IEEE Transactions on Visualization and Computer Graphics*, 17(8):1178–1190.

[Yin et al., 2015] Yin, M., Li, G., Lu, H., Ouyang, Y., Zhang, Z., and Xian, C. (2015). Spectral pose transfer. *Computer Aided Geometric Design*, 35:82–94.

[Zakharov et al., 2019] Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468.

[Zayer et al., 2005] Zayer, R., Rössl, C., Karni, Z., and Seidel, H.-P. (2005). Harmonic guidance for surface deformation. In *Computer Graphics Forum*, volume 24, pages 601–609. Citeseer.

[Zhang and Chen, 2001] Zhang, C. and Chen, T. (2001). Efficient feature extraction for 2d/3d objects in mesh representation. In *International Conference on Image Processing*, volume 3, pages 935–938. IEEE.

[Zhang et al., 2010] Zhang, H., Van Kaick, O., and Dyer, R. (2010). Spectral mesh processing. In *Computer Graphics Forum*, volume 29, pages 1865–1894. Wiley Online Library.

[Zhao et al., 2003] Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458.

[Zhao et al., 2009] Zhao, Y., Liu, X., Xiao, C., and Peng, Q. (2009). A unified shape editing framework based on tetrahedral control mesh. *Computer Animation and Virtual Worlds*, 20(2-3):301–310.

[Zhao et al., 2011] Zhao, Y., Pan, B., and Peng, Q. (2011). Robust deformation transfer via dual domain. In *2011 12th International Conference on Computer-Aided Design and Computer Graphics*, pages 302–305. IEEE.

[Zhou et al., 2020] Zhou, K., Bhatnagar, B. L., and Pons-Moll, G. (2020). Unsupervised shape and pose disentanglement for 3d meshes. In *European Conference on Computer Vision*, pages 341–357. Springer.

[Zhou et al., 2010a] Zhou, K., Xu, W., Tong, Y., and Desbrun, M. (2010a). Deformation transfer to multi-component objects. In *Computer Graphics Forum*, volume 29, pages 319–325. Wiley Online Library.

[Zhou et al., 2010b] Zhou, S., Fu, H., Liu, L., Cohen-Or, D., and Han, X. (2010b). Parametric reshaping of human bodies in images. *ACM transactions on graphics*, 29(4):1–10.

[Zhou et al., 2016] Zhou, X., Sun, X., Zhang, W., Liang, S., and Wei, Y. (2016). Deep kinematic pose regression. In *European Conference on Computer Vision*, pages 186–201. Springer.

[Zuffi et al., 2017] Zuffi, S., Kanazawa, A., Jacobs, D. W., and Black, M. J. (2017). 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6365–6373.