REALISTIC VIRTUAL HUMANS FOR VR THERAPY OF BODY IMAGE DISORDERS

DISSERTATION

ZUR ERLANGUNG DES GRADES EINES

DOKTORS DER NATURWISSENSCHAFTEN

DER TECHNISCHEN UNIVERSITÄT DORTMUND
AN DER FAKULTÄT INFORMATIK

VON

STEPHAN WENNINGER

DORTMUND
2025

Tag der mündlichen Prüfung:	12. März 2025
Dekan:	Prof. Dr. Jens Teubner
Gutachter:	Prof. Dr. Mario Botsch
	Prof. Dr.-Ing. Marc Stamminger

ACKNOWLEDGMENTS

First, I would like to thank my supervisor Prof. Dr. Mario Botsch for supporting me throughout all my research and giving me the opportunity to pursue my PhD. Without his tremendous lectures at Bielefeld University, I would not have gained interest in the fields of Computer Graphics and Geometry Processing. He always gave valuable feedback, discussed ideas and new research directions, supported me with his endless knowledge of publications in the field, and initiated many great collaborations with my colleagues in Bielefeld and Dortmund, as well as other research groups.

In this regard, I want to thank Jascha Achenbach, Andrea Bartl, Martin Komaritzan, Erik Wolf, Nina Döllinger, David Mal, and Fabian Kemper for working closely with me to produce the various publications that build the foundation of this thesis. Many thanks also to Prof. Dr. Marc Latoschik and Prof. Dr. Ulrich Schwanecke for additionally supporting my research and giving valuable feedback on our joint publications.

I would like to thank my great colleagues at Bielefeld University and TU Dortmund University. In particular, I want to thank Jascha Achenbach and Thomas Waltemate, whose previous work on reconstructing virtual humans I could build upon. Thank you to Martin Komaritzan and Astrid Pontzen, who accompanied me during all of my PhD time. They were always a big support and gave me the confidence to continue my research.

Last but not least, I want to thank my family, who was and is always there for me. I am grateful to my father, who contributed to my interest in the natural sciences through his relentless math and physics questions after I came home from school. Thank you, mom, for supporting me in every step of my life and teaching me how to stand on my own feet. Thank you to my siblings, who always have an open ear and make me feel welcome at home.

My research was supported by the German Federal Ministry of Education and Research (BMBF) through the project *Virtual Reality Therapy by Stimulation of Modulated Body Perception* (ViTraS) and by the "Stiftung Innovation in der Hochschullehre" through the project *Hybrid Learning Center* (HyLeC).

ABSTRACT

This thesis presents methods for reconstructing and modifying realistic personalized virtual humans to be employed in the context of a VR-based body image disorder therapy system. We start by presenting a method for generating virtual humans from monocular smartphone cameras, thereby lowering the hardware requirements and increasing the availability of personalized virtual humans compared to other methods, which typically depend on elaborate photogrammetry rigs. In a user study, we investigate the perception of the resulting virtual humans by scanning people with both the low-cost smartphone-based method and a standard multi-view stereo photogrammetry rig. Participants then embody and rate both virtual humans in a virtual mirror exposure scenario. The results show, that both virtual humans are perceived similarly, indicating that our smartphone-based method presents a viable alternative to expensive photogrammetry rigs. For employing realistic virtual humans in body image therapy, we present a method for modifiying the body weight of the virtual humans in real-time. Users of the VR-based therapy system then embody a personalized avatar in a virtual mirror exposure scenario and are given active control over the avatar's body shape, enabling researchers to investigate the potential of VR-based therapy and gain insight into possibly occurring body image disorders.

To improve on the purely surface-based body weight modification model, the second part of this thesis focuses on anatomical representations of virtual humans. We present a method for inferring anatomical details from a given skin surface in less than a minute. To this end, we derive a three-layered anatomical model, consisting of a skin, muscle, and skeleton layer, from a commercial high-resolution anatomical model. We then learn a model for predicting body composition, i.e., fat and muscle mass, from a given skin surface and fit the template model to a large database of surface scans while conforming to the estimated body composition. The original high-resolution anatomical structures are transferred to the resulting fit via a triharmonic space warp. Finally, we use the inferred anatomical data to learn an anatomically constrained volumetric human shape model. We enlarge our training data to the full Cartesian product of all skeleton shapes and all soft tissue distributions using physically plausible volumetric deformation transfer. A self-supervised learning technique then produces two separate latent parameter sets, allowing us to sample different soft tissue distributions over the same skeleton shape and vice versa. The resulting anatomical model additionally facilitates fast skeleton inference and semantic localized shape modification.

CONTENTS

1 INTRODUCTION

Digital representations of humans, also called virtual humans or avatars, are increasingly becoming relevant in the fields of entertainment, e-commerce, medicine, sports, virtual reality, and many others. Today, photorealistic virtual humans or stylized humanoid characters are, e.g., commonly used in movies as part of the visual effects pipeline. Modern video games also employ high-fidelity virtual humans with the additional requirement of real-time rendering. In the e-commerce setting, virtual humans are employed in virtual try-on applications as well as made-to-measure garment fabrication. Apart from these applications in entertainment and e-commerce, virtual humans are also employed in medical or sports applications. 3D scanning of humans can for example support the accurate fitting of personalized prosthetics. The fitness industry leverages virtual humans for tracking, analyzing, and visualizing training progress. In this thesis, we will discuss virtual human reconstruction and modification methods in the context of a VR-based therapy system.

The work we will present in this thesis was in large parts developed in the context of the *ViTraS* project, short for *Virtual Reality Therapy by Stimulation of Modulated Body Perception* [DWW+19; ViT24]. This interdisciplinary research project investigates the potential of employing virtual humans in VR therapy settings as a complementary method to classical intervention techniques. In particular, it focuses on VR therapy of body image disorders and specifically centers on obesity and adiposity. The resulting immersive obesity therapy system should allow users to observe their personalized avatar in a virtual mirror setup in VR while their movements are captured via motion tracking and retargeted onto the virtual human, thus replicating the established mirror exposure intervention. To help researchers gain insight into potentially occurring body image disorders, users of the system should be able to modify the body weight of their embodied avatar in real-time. This allows them to (i) match the virtual human's body shape to their current body image, or (ii) explore and discuss possible future body weights, thereby exploiting the unique advantage that VR therapy interventions have over classical forms of therapy. As previous research has shown that virtual body ownership, the feeling of presence, and the emotional response to the virtual environment are increased when embodying realistic personalized virtual humans [WGR+18], this thesis is also concerned with realistic virtual humans as opposed to stylized representations typically found in VR games or Metaverse applications.

The setting of realistic virtual humans in the context of VR therapy of body image disorders defines the main goals of this thesis: (1) Provide a method for generating personalized virtual humans in an adequate time frame and at a reasonably high fidelity, while having minimal hardware requirements in order

to increase the availability of personalized virtual humans for VR user studies. (2) Develop a method for modifying the body weight of the resulting virtual humans, allowing researchers to investigate the effects of modulated body perception in a VR-based obesity therapy system that reproduces classical mirror exposure intervention.

Realistic personalized virtual humans are typically generated by first performing a 3D scan of the person to be reconstructed. A common method for the 3D scanning of people is to record them with multi-view stereo photogrammetry rigs, where multiple cameras record the scanning subject from different angles, producing a dense 3D point cloud through photogrammetric scene reconstruction. We start by describing our custom-built photogrammetry rig and introducing a template fitting method that registers a statistical animatable human template model to the resulting point cloud, thereby generating a virtual human from such a 3D scan in a fully automatic fashion (Chapter 2). This reconstruction method already fulfills parts of the first goal of this thesis: the resulting virtual humans are ready for full-body and facial animation and can easily be integrated into common graphics pipelines and game engines for use in virtual reality. However, performing the required 3D scans by employing an elaborate photogrammetry rig has several disadvantages that limit the broader availability of personalized virtual humans. Since multi-view stereo photogrammetric reconstruction requires a large amount of high-quality DSLR cameras, the hardware costs for building a photogrammetry rig are non-negligible. Furthermore, photogrammetry rigs are stationary and space-consuming and thus cannot be easily incorporated into medical facilities or therapist offices.

Therefore, in Chapter 3, we present a method for generating such realistic virtual humans from two videos recorded on commodity smartphones, drastically reducing the hardware costs compared to previous reconstruction methods. In order to investigate, if generating virtual humans from smartphone videos presents a viable alternative to traditional photogrammetry rigs, Chapter 4 then examines, how people perceive the resulting virtual humans. We conduct a user study where participants are scanned with both a high-cost photogrammetry rig and the low-cost smartphone reconstruction method. Participants then embody, observe, and rate both resulting virtual humans in a virtual mirror exposure scenario.

With the ability to quickly generate realistic personalized virtual humans from photogrammetry rigs or smartphone videos, we then proceed with developing a statistical model of body weight modification in Chapter 5. This completes all components needed for providing personalized realistic modulatable avatars for virtual mirror exposure and VR therapy scenarios. We briefly present the resulting VR-based prototype, where users embody these virtual humans and have active control over the body weight of their avatar in real-time. The presented model is however trained on surface meshes only

and is therefore unable to accurately reason about personal anatomical traits such as body composition.

To tackle this limitation, the second part of this thesis focuses on anatomical volumetric models for representing virtual humans. First, in Chapter 6, we present an approach for fitting a layered anatomical model to a given skin surface. The model consists of three surface meshes with identical topology – a skeleton, muscle, and skin layer – which gives a straightforward way to define volumetric elements between these layers. We learn to infer body composition from a given surface scan, which then informs the volumetric template fitting approach, allowing us to quickly infer anatomical details for a given skin surface in less than a minute.

Finally, Chapter 7 describes a novel approach for learning a statistical model of human skeletal shape and soft tissue distribution. By employing the anatomy inference method of Chapter 6 and using volumetric deformation transfer, we are able to transfer the soft tissue distribution of a given subject onto the skeletal shape of all other subjects of our data set in a physically plausible way. This defines a synthetic Cartesian data set from which we learn an anatomically constrained volumetric human shape model, which facilitates skeleton inference in less than a second and provides localized shape manipulation of skeletal shape and soft tissue distribution.

CONTRIBUTION

To summarize, the main contributions of this thesis are:

- An approach for creating realistic virtual humans from smartphone videos, thereby reducing the hardware requirements and increasing the availability of virtual human reconstruction methods.
- An evaluation of the proposed reconstruction method in terms of user perception, showing that the resulting virtual humans are perceived similarly to virtual humans reconstructed from more elaborate 3D scanning setups.
- A statistical model for body weight modification, which allows users of a VR-based therapy system based on virtual mirror exposure to directly modify the body weight of their embodied avatar in real-time.
- A layered anatomical model which we can efficiently fit to skin surfaces produced by the aforementioned reconstruction methods in an anatomically plausible way, allowing us to estimate anatomical details from surface scans only.
- A statistical anatomical model of human skeleton shape and soft tissue distribution which allows for fast skeleton inference and localized semantic shape modification of virtual humans.

PUBLICATIONS

This thesis is based on the following publications:

- Stephan Wenninger, Jascha Achenbach, Andrea Bartl, Marc Erich Latoschik, and Mario Botsch. "Realistic Virtual Humans from Smartphone Videos". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2020, 29:1–29:11 (Best Paper Award 🏆).
- Andrea Bartl, Stephan Wenninger, Erik Wolf, Mario Botsch, and Marc Erich Latoschik. "Affordable but not Cheap: A Case Study of the Effects of Two 3D-Reconstruction Methods of Virtual Humans". *Frontiers in Virtual Reality* 2 (2021).
- Martin Komaritzan, Stephan Wenninger, and Mario Botsch. "Inside Humans: Creating a Simple Layered Anatomical Model from Human Surface Scans". *Frontiers in Virtual Reality* 2 (2021).
- Nina Döllinger, Erik Wolf, David Mal, Stephan Wenninger, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Resize Me! Exploring the User Experience of Embodied Realistic Modulatable Avatars for Body Image Intervention in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022).
- Stephan Wenninger, Fabian Kemper, Ulrich Schwanecke, and Mario Botsch. "TailorMe: Self-Supervised Learning of an Anatomically Constrained Volumetric Human Shape Model". *Computer Graphics Forum* 43.2 (2024).

Other publications that I contributed to, but that are not directly relevant for this thesis are:

- Erik Wolf, David Mal, Viktor Frohnapfel, Nina Döllinger, Stephan Wenninger, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Plausibility and Perception of Personalized Virtual Humans between Virtual and Augmented Reality". In *Proc. of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2022, pp. 489–498.
- Erik Wolf, Nina Döllinger, David Mal, Stephan Wenninger, Andrea Bartl, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Does Distance Matter? Embodiment and Perception of Personalized Avatars in Relation to the Self-Observation Distance in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022).
- David Mal, Nina Döllinger, Erik Wolf, Stephan Wenninger, Mario Botsch, Carolin Wienrich, and Marc Erich Latoschik. "Am I the Odd One? Exploring (In)Congruencies in the Realism of Avatars and Virtual Others in Virtual Reality". *Frontiers in Virtual Reality* 5 (2024).

- Maria Korosteleva, Timur Levent Kesdogan, Stephan Wenninger, Fabian Kemper, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine. "GarmentCodeData: A Dataset of 3D Made-to-Measure Garments with Sewing Patterns". *Computer Vision – ECCV* (2024).

Additional material such as code and accompanying videos can be found at the following locations:

- The video showing our method for reconstructing virtual humans from smartphone videos can be found at https://www.youtube.com/watch?v=2D3-vn2yFVc.
- The video illustrating the capabilities of our anatomically constrained volumetric human shape model can be found at https://www.youtube.com/watch?v=rrkf_fIhX0Q.
- Our anatomically constrained volumetric human shape model is released at https://github.com/mbotsch/TailorMe.

2 FUNDAMENTALS

This chapter describes fundamental concepts that will be relevant throughout the rest of the thesis. We start by introducing multi-view stereo photogrammetry rigs, which are commonly used for the generation of virtual humans. We describe our custom photogrammetry rig built at TU Dortmund University, which is used to generate 3D point sets of the scanned subjects. Secondly, we will describe the process of generating virtual humans from such a single-shot multi-view stereo scan. Specifically, we will detail our template fitting implementation, which is largely based on previous work but adapted to work in a fully automatic manner.

2.1 PHOTOGRAMMETRY RIGS

The first step of many virtual human generation pipelines is to produce a 3D scan of the person to be reconstructed. The most hardware-, data-, and cost-intensive approaches employ so-called light stages [DHT+00; GFT+11; GLD+19], where multiple cameras take multiple shots under different lighting conditions in order to capture fine-scale geometry and skin reflectance properties of the scanned subjects. Single-shot multi-view stereo reconstruction methods rely on diffuse lighting only [BHP+10; BBB+10] or additionally employ polarized light [RGB+20] to reconstruct highly detailed geometry under a single lighting condition. The captured images are then processed by off-the-shelf commercial photogrammetry software such as Agisoft Metashape [Agi24] or RealityCapture [Rea24], closed source tools such as Apple's photogrammetry toolkit [App24], open source alternatives such as COLMAP [SF16; SZP+16; COL24] or Meshroom [Ali24], or via custom stereo matching algorithms [BHP+10]. Commonly, these software components output either dense point sets or unstructured triangle meshes, to which a template mesh is fitted in order to provide consistent mesh topology over different scans.

We built a custom single-shot multi-view stereo photogrammetry rig at TU Dortmund University in order to generate the 3D scans which serve as input for our virtual human reconstruction pipeline. The photogrammetry rig was built as part of the Hybrid Learning Center (*HyLeC*) project [HyL24]. The design of the scanner booth itself and the selection of the various components was done in collaboration with my colleague Denis Fisseler. I selected the different components such as cameras, USB hubs, and triggerboxes, and implemented the user interface and processing pipeline for controlling the scanner, while Denis Fisseler designed the scanner booth frame, various 3D printed parts for mounting the individual components, and planned the power supply setup.

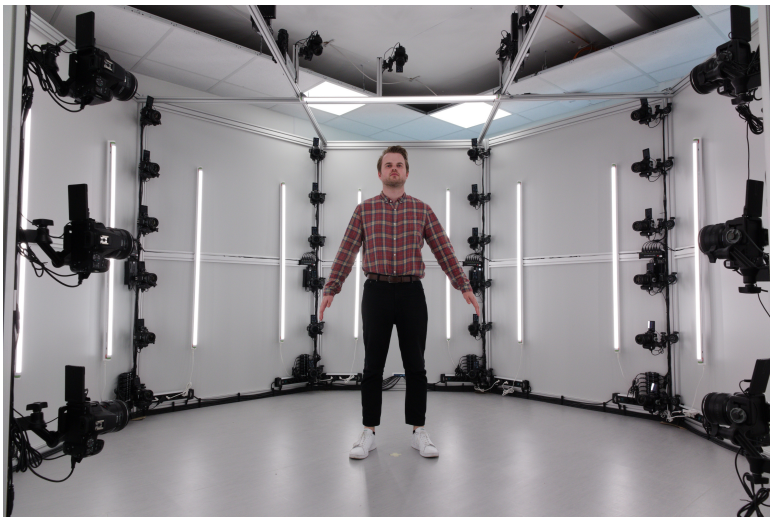

Figure 2.1: The photogrammetry rig built at TU Dortmund University consists of 56 DSLR cameras, which are evenly distributed in the eight corners and the ceiling of the octagonal scanner cabin. High-quality LED tubes ensure uniform lighting of the scanning subject.

The resulting scanner setup should meet various design goals. First, we want a setup which provides static full-body scans from a single shot in order to make data acquisition fast and easy. Second, the photogrammetry rig should require no extensive pre-calibration to allow for any future updates in terms of the number of cameras or their positioning. Third, the scanner and the subsequent virtual human generation should be easy to operate in order to allow non-experts to create virtual humans with our photogrammetry rig. These goals were considered in the design process, resulting in the scanner setup depicted in Figure 2.1. The individual components and settings will be detailed in the following.

The scanner booth has a diameter of ca. 4.4 m and is constructed with an octagonal aluminum profile frame with painted chipboard walls. We evenly distribute 48 Canon EOS 250D cameras equipped with 35 mm fixed wide-angle lenses in an 8×6 pattern in the eight corners of the octagonal frame to capture the scanned person from all angles. To better capture the top of the head, we additionally mounted eight Canon EOS 250D cameras with 50 mm fixed lenses in the ceiling of the scanner booth, yielding a total of 56 cameras with an image resolution of 4020×6024 pixels. All cameras are supplied with power by using AC adapters to allow for continuous operation without the need to change batteries. Concurrently switching on this amount of AC adapters can however trip the circuit breaker. To prevent this, we installed two transformer switching relays that prevent an inrush current but still provide a single access point for powering the whole scanner cabin.

Each of the side walls, as well as the ceiling of the scanner booth, is equipped with two neutral-white 4000 K LED tubes with a high color rendering

index (CIE $R_a > 95$) in order to uniformly light the scanned subjects and accurately reproduce the colors in the scene. All cameras are set to an ISO level of 200, a shutter speed of 1/15 s, and an f-number of $f/8.0$. These settings provide low image noise, sufficiently low motion artifacts and the largest depth of field possible for the amount of light in the scanner booth. We put all cameras into manual focus mode as we found that the automatic focus mode of the cameras does not reliably put the scanning subjects into the depth of field. See Figure 2.2 for an example of a set of images taken with this camera setup.

The cameras are connected via ESPER TriggerBoxes which forward the trigger signal of a pair of remote shutter release controls to all cameras, allowing us to simultaneously activate their shutter release mechanism. Communication between the cameras and the reconstruction PC – a desktop workstation operating Ubuntu 22.04 LTS, equipped with an Intel Core i9-10850K CPU and an Nvidia RTX 3070 GPU – is done using libgphoto2 [GPh24] and USB 2.0, as faster USB connections are not supported by the cameras. Each column of cameras – located in one corner of the octagonal scanner booth – is connected to a USB hub, which is then daisy-chained with the next column of cameras. Special care was taken to not exceed the external USB tier limit of 5 tiers and the maximum number of USB device endpoints per xHCI USB controller (typically 32 or 64). To this end, only four USB hubs at a time are daisy-chained in the aforementioned way. Still, we observed sporadic USB disconnects and reconnects, after which the cameras were no longer responsive. To deal with these disconnects, we monitor the USB connections by registering event handlers through the UNIX *udev* system and whenever a camera disconnects, we call the USB driver's *unbind* and *bind* procedures, thereby turning the respective camera responsive again. We continuously monitor the camera state to check if new images were taken and download the resulting 1.7 GB of image data automatically to the reconstruction PC once a new shot has been taken. The image download takes approximately 45 s due to the limited speed of the USB 2.0 connections.

We download RAW image data in order to apply the same white balance settings to all images before converting them to JPG with libraw [Lib24]. The required white balance parameters are manually determined using the RAW image editing software Darktable [Dar24]. We did not observe substantial differences between using the JPG images and using a lossless alternative such as TIFF when inspecting the quality of the point sets resulting from the subsequent photogrammetry step. As such, we decided to use JPG due to smaller file sizes. After image conversion, we use the image segmentation model Deeplab v3 [CZP+18] to create binary image masks, which mask out all pixels belonging to the background. This constrains feature detection and stereo matching in the subsequent photogrammetry step to pixels belonging to the scanning subject, which in our tests produces less noisy point sets, resolves

Figure 2.2: A set of images created by our photogrammetry rig. 56 cameras are evenly distributed to capture the scanned subject from all angles. These images are passed to a commercial photogrammetry software in the next processing step of our virtual human creation pipeline.

the need for pre-calibrating the dense reconstruction volume, and removes points belonging to the background from the resulting dense point set.

The converted images and the generated binary masks are passed to the commercial photogrammetry software Agisoft Metashape [Agi24]. We do not pre-calibrate the extrinsic and intrinsic parameters of the camera setup, as we empirically found that Metashape produces less noisy point sets and more reliable results without pre-calibration. This also yields a more flexible setup, as it removes the need for re-calibrating the system in case of any future updates of the positioning or number of cameras. We do however copy relevant EXIF metadata from the RAW images to the converted JPG images in order to preserve information about focal length and pixel size. These values are used as initial parameters for the intrinsic camera calibration performed by the photogrammetry software, which models the cameras as standard central projection cameras with Brown's distortion model [Bro71]. Agisoft Metashape then computes the intrinsic and extrinsic camera parameters for every input image and generates a dense point set consisting of about 4 M points. The camera calibration together with the dense point set defines the 3D scan of the scanned person and will be the target of the subsequent template fitting process.

All processing steps are scripted to run automatically, allowing scan operators to easily create virtual humans with our photogrammetry rig. For further ease of use, we developed a custom GUI application that allows to manage a scan session. Scan operators can inspect the state of all cameras, observe the download progress, name specific scans, and start the fully automatic virtual human reconstruction (described in the next section) by a single button click after the image download completes. The backend of the scanner software monitors the cameras for new shots, checks for the aforementioned sporadic USB disconnects, and handles the image data download as well as file system and permission setup. After a short briefing session about the hardware and scanner control GUI, non-experts could successfully use our multi-view stereo photogrammetry rig.

2.2 GENERATION OF VIRTUAL HUMANS

Once a 3D scan of the person to be reconstructed is computed, a common method of creating animatable virtual humans from this static point set data is to fit a template model to the observed data [BTP14; LMR+15; PWH+17; AWL+17]. These template models typically consist of one or several surface meshes with clean mesh topology, an associated UV map for storing color information, an animation rig for full-body animation, and a set of blendshapes for facial animation. Fitting such a template model has several advantages. The explicit representation via surface meshes and an animation rig allows

the resulting virtual humans to easily be incorporated into existing graphics pipelines and game engines like Unity or the Unreal Engine. Rendering one or several virtual humans in real-time is especially relevant for VR scenarios, where low frame rates can lead to nausea and VR sickness. And even though there have been advances in implicit representations, especially for 3D head avatars [GKG+23; KQG+23; TMP+24], there are still modern approaches arguing in favor of explicit representations [BZH+23; SGY+24] or embedding implicit representations around an explicit mesh-based representation [ZBT23; QKS+24]. The common mesh topology over various scans additionally gives a straightforward way to create a statistical model of human body shapes which can serve as a low-dimensional shape prior or as a first dimensionality reduction in various learning tasks [WNT+21; KZB+22].

Human shape modeling has been extensively studied due to its application in various fields such as shape and pose estimation from a single image [BKL+16], body composition estimation [WNT+21], generating synthetic training data for image recognition tasks [WBH+21], or – as in our case – the creation of virtual humans. Representing human body shapes in a low-dimensional space allows for efficient optimization of the model parameters to match the observed data. This can be used either for coarse initialization before further fine-scale registration steps or as a prior model used for regularization during the whole optimization process. When dealing with incomplete data, these models can also be used to fill in any missing data in a statistically plausible way. Most popular models are based on Principal Component Analysis (PCA) of vertex positions [LMR+15; OBB20], following the seminal work of Blanz and Vetter [BV99]. Other approaches directly encode triangle deformations from the template to the registered models [ASK+05] or a decomposition of these triangle deformations [FB12]. Recent work has also used more sophisticated dimensionality reduction techniques such as convolutional neural networks or neural fields [RBS+18; BBP+19; GKG+23]. For an overview of parametric head models, also called morphable models, we refer the reader to the survey of Egger et al. [EST+20].

2.2.1 Animatable Statistical Virtual Human Template Models

Let us now define the components of an animatable statistical human template model more formally. The main skin surface of the template model is given by a triangle mesh with faces $\mathcal{F}$, edges $\mathcal{E}$, and vertices $\mathcal{V} = \{v_1, \ldots, v_V\}$, whose 3D positions are denoted by $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_V\}$. Vertically stacking all 3D positions yields a vector $\mathbf{X} = \left(\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_V^\mathsf{T}\right)^\mathsf{T} \in \mathbb{R}^{3V}$. To facilitate full-body animation, the mesh is tied to a standard skeleton graph – also called an animation rig – defined by a joint hierarchy consisting of J joints, which mimic the actual bone and joint structure of the human body. The pose of

this skeleton can be described by joint angles $\boldsymbol{\theta} \in \mathbb{R}^{3J}$, which define the local per-joint rotations with respect to the initial pose of the skeleton. This initial pose is also called the rest pose or bind pose of the skeleton. Each joint j has a local coordinate system, where the associated local matrix

$$\mathbf{L}_j = \begin{pmatrix} \bar{\mathbf{R}}_j \mathbf{R}_j(\boldsymbol{\theta}) & \mathbf{t}_j \\ \mathbf{0} & 1 \end{pmatrix} \tag{2.1}$$

maps a given point (expressed in homogeneous coordinates) into the parent's coordinate system. Here, $\bar{\mathbf{R}}_j$ is the bind pose rotation of joint j, while $\mathbf{R}_j(\boldsymbol{\theta})$ is the rotation matrix representation of the three angles in $\boldsymbol{\theta}$ which correspond to joint j. The local translational offset towards the parent joint is given by $\mathbf{t}_j$. Bind pose rotations are typically either aligned to the global x, y, z axes or defined such that one of the axes is aligned with the bone axis. To map from joint j into the coordinate system of the root joint, we can iterate through the kinematic chain $\mathcal{K}(j)$ from the root joint towards joint j and concatenate the local transforms, yielding the global transformation matrix

$$\mathbf{G}_j = \prod_{k \in \mathcal{K}(j)} \mathbf{L}_k. \tag{2.2}$$

The process of computing global joint matrices $\mathbf{G}_j$ from given joint angles $\boldsymbol{\theta}$ is called forward kinematics. In order to express a given pose relative to the bind pose of the skeleton, we store the initial global transformation matrices $\bar{\mathbf{G}}_j$ in the bind pose of the model, i.e., $\boldsymbol{\theta} = \mathbf{0}$.

We now need to associate the forward kinematics of the animation rig with an animation of the corresponding surface mesh. This association is given by skinning weights w_{ij}, which define, how much influence a given joint j has on vertex v_i. For each vertex v_i, the corresponding skinning weights are non-negative and sum to one: $\sum_{j=1}^{J} w_{ij} = 1$. The skinned position $\tilde{\mathbf{x}}_i'$ of vertex v_i (in homogeneous coordinates) can be computed by

$$\begin{aligned} \tilde{\mathbf{x}}_i' &= \sum_{j=1}^{J} w_{ij} \mathbf{T}_j \tilde{\mathbf{x}}_i \\ \mathbf{T}_j &= \mathbf{G}_j \, \bar{\mathbf{G}}_j^{-1} \\ \tilde{\mathbf{x}}_i &= \begin{pmatrix} \mathbf{x}_i \\ 1 \end{pmatrix}. \end{aligned} \tag{2.3}$$

This method is called Linear Blend Skinning (LBS), due to the linear blending of the transformation matrices $\mathbf{T}_j$. The amount of joints which influence a given vertex is usually limited to four or eight joints in order to keep this process computationally efficient. Let us denote the function that applies LBS to all vertices $\mathcal{V}$ of the template model by $\text{skin}(\mathbf{X}, \boldsymbol{\theta}) : \mathbb{R}^{3V} \times \mathbb{R}^{3J} \to \mathbb{R}^{V \times 3}$. For brevity of notation in the later formulation of the template fitting algorithm,

we assume that this function outputs the resulting vertex positions as a $V \times 3$ matrix, where the three columns of the matrix correspond to the x, y, z coordinates of the vertex positions. Finding joint angles $\boldsymbol{\theta}$ for a given scan and then applying Linear Blend Skinning to the template model allows us to cover the pose variation present in the 3D scans. The shape variation of the 3D scans is then covered by a statistical human body shape model, described in the following.

Given a set of M registered meshes with the same topology and in the same pose, we collect their respective vertex positions into M vectors $\mathbf{X}_j \in \mathbb{R}^{3V}$. These vectors then define the individual observations from which a PCA-based pose-normalized morphable model can be constructed as follows. Let $\bar{\mathbf{X}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{X}_j$ denote the mean of the registered meshes. Performing PCA of the mean-centered data matrix $\left(\mathbf{X}_1 - \bar{\mathbf{X}}, \ldots, \mathbf{X}_M - \bar{\mathbf{X}}\right) \in \mathbb{R}^{3V \times M}$ and selecting the first $k < M$ components yields the PCA matrix $\mathbf{U} \in \mathbb{R}^{3V \times k}$ as well as the corresponding eigenvalues $\left(\sigma_1^2, \ldots, \sigma_k^2\right)$, sorted in descending order by their magnitude. Here, σ_k denotes the standard deviation of the k^{th} PCA component, and k is typically chosen such that the resulting k components cover a desired percentage p_k of the variance present in the training data, computed by the quotient of the k accumulated eigenvalues and the total variance in the data set: $p_k = \sum_i^k \sigma_i^2 \,/\, \sum_i^M \sigma_i^2$.

This process defines a parametric human body shape model, where the correspondence between the low-dimensional shape parameters $\boldsymbol{\beta} \in \mathbb{R}^k$ and the vertex positions of the resulting mesh is given by $\mathbf{X} = \mathbf{U}\boldsymbol{\beta} + \bar{\mathbf{X}}$. This completes the definition of the components of an animatable statistical template model, which can then be fitted to a given observation by optimizing the pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$. See the discussion by Pishchulin et al. [PWH+17], who describe an iterative human-in-the-loop process for generating the registered meshes needed for training such a model.

2.2.2 Template Fitting

With the components of an animatable statistical virtual human template model defined, we now describe how to use such a model for generating virtual humans from 3D surface scans through a process called template fitting. The template fitting implementation in all following chapters of this thesis is based on the work of Achenbach et al. [AWL+17], who presented a method for generating virtual humans from a 3D scan taken with a similar photogrammetry rig as described in Section 2.1. To provide more details in the face region of the resulting virtual humans, they additionally employed a dedicated face scanner, operated with polarized flash lighting photography (see Figure 2 in [AWL+17]). We omit such a dedicated face scanner in our setup in favor of a simplified data acquisition. For completeness, we will

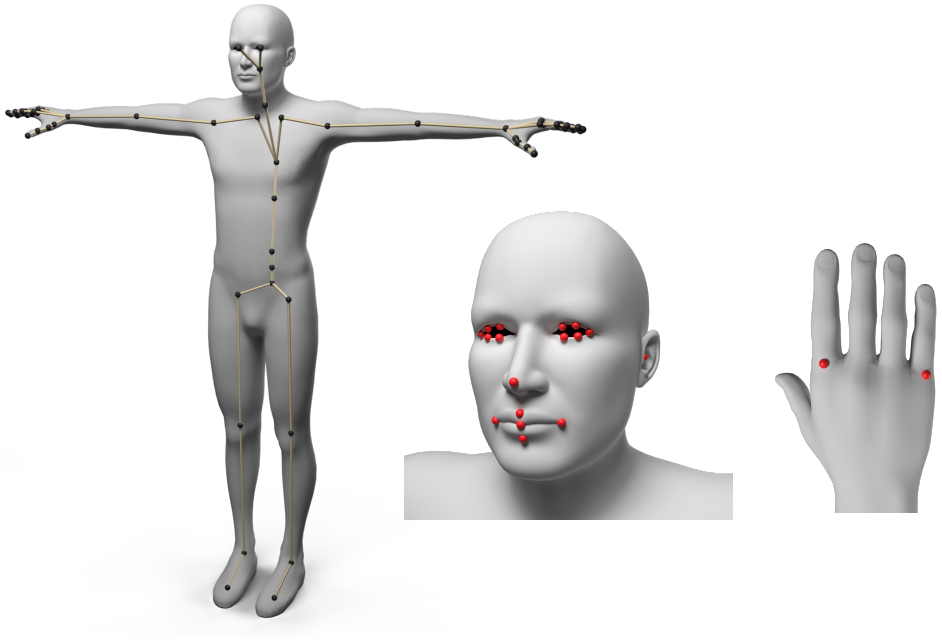

Figure 2.3: Our virtual human template model (left) is derived from the Autodesk CharacterGenerator [Aut24]. It features a main skin mesh, auxiliary meshes for eyes and teeth, and a skeleton graph for full-body animation. The process of fitting this template model to a given 3D scan is guided by a set of point set landmarks, whose counterparts on the template model are pre-selected once (right).

detail all steps of the virtual human reconstruction pipeline proposed by Achenbach et al. [AWL+17], and discuss, how we automized their method to facilitate fully automatic generation of virtual humans from the 3D scans our photogrammetry rig produces.

The result of the photogrammetry step is a point set $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$, where each 3D point $\mathbf{p}_j$ is additionally associated with a normal $\mathbf{n}_j$ and RGB color $\mathbf{c}_j$. In the originally proposed method, the template fitting process is guided by a set of landmarks on the point set, which are manually selected by the user. The counterparts of these point set landmarks are selected once on the template model, which is derived from the Autodesk CharacterGenerator [Aut24] (see Figure 2.3). It consists of one main triangle mesh for the skin surface, whose $V \approx 21\,\mathrm{k}$ vertex positions are again denoted by $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_V\}$. The model additionally provides auxiliary meshes for eyes and teeth. For full-body animation, the template model features an animation rig (controlled by joint angles $\boldsymbol{\theta}$) with corresponding skinning weights, while face animation is facilitated through a set of facial blendshapes.

The training data for a PCA-based statistical shape model was originally generated by registering this template model to a subset of ten A-pose scans from the FAUST database [BRL+14], 111 scans from a database published by Hasler et al. [HSS+09] and 83 synthetic example characters from the Autodesk CharacterGenerator [Aut24]. The registered meshes were then used to create

a parametric shape model in the manner described above. To cover a wider range of body shapes and remove the synthetic training examples, we created a new statistical model based on 1700 scans from the European subset of the CAESAR database [RBD+02], which replaces the original model used by Achenbach et al. [AWL+17].

The proposed template fitting pipeline then starts by optimizing the template model parameters, consisting of the global alignment between the point set and the template model, the joint angles of the animation rig and the shape parameters of the statistical human body shape model. Following this initial registration, a fine-scale deformation process is employed to more closely match the scanner data. After fitting the auxiliary eyes and teeth meshes to the new skin surface, the facial blendshapes are transferred from the template model to the fitted shape. Finally, the corresponding color texture is generated from the input images. The resulting virtual human is ready for animation and can be imported into existing graphics pipelines or game engines like Unity or Unreal Engine.

Initial Registration

The initial registration between the template model and the given point set is done solely based on the specified landmarks $\mathcal{L}$. Alignment, pose, and shape are optimized through an iterative block-coordinate descent scheme, i.e., Achenbach et al. [AWL+17] consecutively optimize these three parameter sets, while keeping the respective other two parameter sets fixed.

The alignment between the point set and the template is defined by the global registration parameters $(s_g, \mathbf{R}_g, \mathbf{t}_g) \in (\mathbb{R} \times SO(3) \times \mathbb{R}^3)$, defining the isotropic scale, global rotation, and translation, i.e., a similarity transformation, which bring the point set and the template into a common coordinate system. For brevity of notation, we can write the similarity transformation defined by $(s_g, \mathbf{R}_g, \mathbf{t}_g)$ as a matrix

$$\mathbf{M}_g = \begin{pmatrix} s_g \mathbf{R}_g & \mathbf{t}_g \\ \mathbf{0} & 1 \end{pmatrix} \in \mathbb{R}^{4\times 4},$$

which applies the similarity transformation to a position given in homogeneous coordinates. The pose of the animatable statistical template model is defined by the joint angles $\boldsymbol{\theta} \in \mathbb{R}^{3J}$, allowing the model to match the approximate A-pose all scanning subjects are instructed to take. The final set of initial registration parameters are the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^k$ of the PCA model, allowing the model to coarsely estimate the observed human body shape. Figure 2.4 visualizes the role of the three parameter sets, whose optimization is detailed in the following.

First, the optimal similarity transformation $(s_g, \mathbf{R}_g, \mathbf{t}_g)$ that minimizes the squared distances between the landmarks on the point set and the mesh is com-

Figure 2.4: Overview of the individual parameter sets that are optimized to fit the template model to a given 3D scan. The alignment between the point set and the template model is defined by the global scale, rotation and translation parameters (left). Joint angles of the animation rig are optimized to match the approximate A-pose that subjects are instructed to take (center-left). The shape parameters of the statistical model are used to coarsely estimate the observed body shape (center-right). Finally, a non-rigid fine-scale deformation process further aligns the template model to the data (right).

puted in a closed-form manner [Hor87]. Second, joint angles of the template model are optimized via Damped Least Squares Inverse Kinematics [ALC+18], incorporating the joint constraint model proposed by Schröder et al. [SMR+14], which prevents anatomically implausible joint angles such as hyperextension of the knees and elbows and additionally constrains the optimized pose to an approximate A-pose. Third, the shape parameters of the 30-dimensional, pose-normalized shape model are optimized to minimize the distance between the template and point set landmarks in a least squares sense. This amounts to solving a linear system of equations due to the linear PCA shape model employed. The three steps are iterated until the relative error, i.e., the absolute difference of the error value in the current and previous iteration normalized by the error value of the previous iteration, falls below 5 %.

After each iteration, since the shape of the template mesh has changed according to the shape parameters $\boldsymbol{\beta}$, the animation rig has to be updated to conform to this new shape. Computing new joint positions is done using mean value coordinates [JSW05], which express the global joint positions as a function of the vertex positions and can be precomputed on the template mesh once. This allows to compute new joint positions for the optimized shape and ensures that the relative positioning of the joints with respect to the vertex positions matches the template model.

Coarse Registration

The initial registration based on the point set landmarks is refined by adding new constraints based on closest point correspondences between the point set and the template (in scan-to-template direction, as discussed by Achenbach et

al. [AZB15]). To speed up the computation of the correspondences and reduce the size of the linear systems involved in the optimization, the point set is first uniformly down-sampled. The amount of down-sampling is proportional to the mean edge length of the template model in order to preserve geometric details that the template model could reproduce given its vertex density. This reduces the size of the point sets from about 4 M to 200 k points. To preserve the point set landmarks during this process, their position, color, and normal are stored and appended back to the point set after down-sampling. Finding the closest points is done in a brute-force manner on the GPU, implemented as an OpenCL compute kernel. For each point on the point set, the closest point on the mesh surface is computed and expressed through barycentric coordinates in the corresponding face. For increased robustness, correspondences are pruned if their distance exceeds 10 cm or if their normals deviate by more than 50°. Let $\mathcal{C} = \{\mathbf{P}_\mathcal{C}, \mathbf{B}_\mathcal{C}, \mathbf{W}_\mathcal{C}\}$ contain the $C \leq N$ correspondences expressed as points $\mathbf{P}_\mathcal{C} \in \mathbb{R}^{C\times 3}$ on the point set $\mathcal{P}$ and a sparse matrix $\mathbf{B}_\mathcal{C} \in \mathbb{R}^{C\times V}$, containing in each row the barycentric coordinates of the closest triangle on the mesh surface. Multiplying $\mathbf{B}_\mathcal{C}$ with vertex positions $\mathbf{X} \in \mathbb{R}^{V\times 3}$ thus yields the corresponding positions on the mesh surface. $\mathbf{W}_\mathcal{C} = \operatorname{diag}(w_1, \ldots, w_C)$ stores normalized per-correspondence weights (i.e., $\sum_i^C w_i = 1$), allowing the optimization to not trust correspondences in regions that are typically not scanned well (such as the hands) and to weight landmarks differently from closest point correspondences. To unify the treatment of landmarks $\mathcal{L}$ and closest point correspondences, we add the corresponding landmark positions, barycentric coordinates, and weights to the correspondence set $\mathcal{C}$.

Again, alignment, pose, and shape parameters are alternatingly optimized until convergence, such that the distance between the template and point set correspondences is minimized in a least squares sense. Note that in the original formulation of Achenbach et al. [AWL+17], since the shape model is trained in T-pose, the point set was unposed from scan pose to T-pose via Linear Blend Skinning in order to optimize the shape parameters. In our formulation, we optimize shape parameters, such that the distance between corresponding points on the resulting *skinned mesh* and the aligned point set is minimized in a least squares sense. Formally, we iteratively minimize the energy function

$$E_\text{init}(\boldsymbol{\beta}, \boldsymbol{\theta}, s_g, \mathbf{R}_g, \mathbf{t}_g) \;=\; E_\text{fit}(\boldsymbol{\beta}, \boldsymbol{\theta}, s_g, \mathbf{R}_g, \mathbf{t}_g) + \lambda_\text{prior} E_\text{prior}(\boldsymbol{\beta})\,, \tag{2.4}$$

in the aforementioned block-coordinate descent optimization scheme. The two components are a fitting term E_fit and a weighted shape prior term E_prior.

The fitting term measures the squared distances between the template and point set correspondences and is computed by

$$E_\text{fit}(\boldsymbol{\beta}, \boldsymbol{\theta}, s_g, \mathbf{R}_g, \mathbf{t}_g) \;=\; \left\| \mathbf{W}_\mathcal{C} \left(\mathbf{B}_\mathcal{C} \operatorname{skin}\left(\mathbf{U}\boldsymbol{\beta} + \bar{\mathbf{X}}, \boldsymbol{\theta}\right) - \boldsymbol{\pi}_g\left(\mathbf{P}_\mathcal{C}, \mathbf{M}_g\right) \right) \right\|_F^2 . \tag{2.5}$$

Here, the current shape parameters $\boldsymbol{\beta}$ yield new vertex positions $\mathbf{X}_{\boldsymbol{\beta}} = \mathbf{U}\boldsymbol{\beta} + \bar{\mathbf{X}}$ in bind pose. These are then skinned to the current pose $\boldsymbol{\theta}$ using the Linear Blend Skinning function $\text{skin}(\mathbf{X}_{\boldsymbol{\beta}}, \boldsymbol{\theta})$. From there, we extract points on the resulting mesh surface through the barycentric coordinates $\mathbf{B}_{\mathcal{C}}$ of the correspondence set $\mathcal{C}$. The function π_g applies the similarity transformation defined by the global registration parameters $(s_g, \mathbf{R}_g, \mathbf{t}_g)$ – expressed here in matrix form by $\mathbf{M}_g$ – to all correspondences on the point set. Correspondences are weighted by $\mathbf{W}_{\mathcal{C}}$, and we compute the weighted squared distances between the corresponding points via the squared Frobenius norm $\|\cdot\|_F^2$.

Overfitting of the shape model is prevented by the Tikhonov regularization term E_{prior}, computed by

$$E_{\text{prior}}(\boldsymbol{\beta}) \;=\; \frac{1}{k}\,\|\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2\,. \tag{2.6}$$

It penalizes the norm of the shape parameters scaled by the inverse of the standard deviation of the respective PCA components, which is achieved by setting $\boldsymbol{\Gamma} = \text{diag}(1/\sigma_1, \ldots, 1/\sigma_k)$. E_{prior} thereby measures how many standard deviations each component of the current shape parameters differs from the mean. We set $\lambda_{\text{prior}} = 10^{-5}$ and iteratively minimize Equation (2.4) until convergence, again measured by the relative error dropping below 5 %. In each iteration, the closest point correspondences between the point set and the template mesh are recomputed. This process yields the final estimation of alignment and pose parameters of the template model, as well as a first coarse shape estimation (see Figure 2.4 (center-right)). In the following fine-scale deformation step, the shape of the template model is further optimized to match the observed data.

Fine-scale Deformation

The in-model registration described in the previous subsection only fits the scanner data coarsely, especially since high-frequency details of the observation cannot be explained by the low-dimensional shape model. Note also that the shape model is trained on a set of 3D scans where the subjects wore minimal clothing, while we typically scan people in casual clothing with our photogrammetry rig, constituting a further domain gap that needs to be accounted for. In order to more closely match the given data, a deformable registration regularized by a surface-based deformation energy is employed to minimize the distance between the template mesh and the point set, while only allowing physically plausible deformations. This process optimizes the positions of all vertices $\mathcal{V}$ of the template model in bind pose, instead of the low-dimensional parameters of the statistical human body shape model

employed in the initial registration phase. The objective function is a weighted sum of individual energy terms and is given by

$$\begin{aligned} E_{\text{fine}}(\mathcal{X}) \;&= \lambda_{\text{cpc}} E_{\text{cpc}}(\mathcal{X}) + \lambda_{\text{lm}} E_{\text{lm}}(\mathcal{X}) + \\ &\quad \lambda_{\text{reg}} E_{\text{reg}}\left(\mathcal{X}, \overline{\mathcal{X}}\right) + \lambda_{\text{shut}} E_{\text{shut}}(\mathcal{X}) \,. \end{aligned} \tag{2.7}$$

It consists of (i) a correspondence term E_{cpc}, which attracts the template mesh to the point set correspondences, (ii) a landmark term, which penalizes the distance of the selected landmarks on the point set and the template mesh, (iii) a regularization term E_{reg} which penalizes deformation from the undeformed state $\overline{\mathcal{X}}$ (resulting from the initial registration phase), and (iv) a corrective term E_{shut}, which keeps vertex pairs from the upper and lower lip close to each other, as subjects are scanned with a neutral face expression.

The correspondence term measures the squared distances between positions on the skinned template model and the aligned point set correspondences and is given by

$$E_{\text{cpc}}(\mathcal{X}) \;=\; \sum_{i=1}^{C} w_i \left\| \text{skin}_{\mathcal{C}}\left(\mathbf{x}_i^{\mathcal{C}}, \boldsymbol{\theta}\right) - \tilde{\mathbf{p}}_i^{\mathcal{C}} \right\|^2 . \tag{2.8}$$

Recall that the closest point correspondences $\mathcal{C}$ between the point set and the mesh surface are defined by points $\mathbf{p}_i^{\mathcal{C}}$ on the point set and positions $\mathbf{x}_i^{\mathcal{C}}$ on the mesh surface, expressed through barycentric coordinates in the closest triangle on the mesh surface. The function $\text{skin}_{\mathcal{C}}\left(\mathbf{x}_i^{\mathcal{C}}, \boldsymbol{\theta}\right)$ applies Linear Blend Skinning to $\mathbf{x}_i^{\mathcal{C}}$ by skinning the three positions of the triangle vertices and interpolating the resulting positions using the barycentric coordinates. $\tilde{\mathbf{p}}_i^{\mathcal{C}}$ denotes the point set position resulting from applying the similarity transform $\left(s_g, \mathbf{R}_g, \mathbf{t}_g\right)$ (computed in the initial registration) to point $\mathbf{p}_i^{\mathcal{C}}$. As in the initial registration, correspondences are weighted by the normalized per-correspondence weights w_i. Note that this again differs from the original formulation of Achenbach et al. [AWL+17], who unposed the point set via Linear Blend Skinning, while we optimize the vertex positions in the rest pose, such that they match the point set correspondences after applying skinning with the optimized pose parameters.

Similar to E_{cpc}, the landmark term E_{lm} is implemented by measuring the squared distance between the selected landmarks on the point set and the corresponding skinned vertex positions:

$$E_{\text{lm}}(\mathcal{X}) \;=\; \sum_{i=1}^{L} w_i^{\text{lm}} \left\| \text{skin}_{\mathcal{L}}\left(\mathbf{x}_i^{\mathcal{L}}, \boldsymbol{\theta}\right) - \tilde{\mathbf{p}}_i^{\mathcal{L}} \right\|^2 . \tag{2.9}$$

Analogously to the notation in the correspondence term (2.8), $\text{skin}_{\mathcal{L}}\left(\mathbf{x}_i^{\mathcal{L}}, \boldsymbol{\theta}\right)$ applies Linear Blend Skinning to the landmark position $\mathbf{x}_i^{\mathcal{L}}$ on the template surface, while $\tilde{\mathbf{p}}_i^{\mathcal{L}}$ denotes the point set landmark position after applying the optimized similarity transform to the point set position $\mathbf{p}_i^{\mathcal{L}}$.

To ensure that the triangles of the template model do not degrade or deform too strongly during the iterative fitting process, the deformation is constrained by the regularization term E_{reg}, which tries to keep the deformation of the mesh locally rigid:

$$E_{\text{reg}}(\mathcal{X}, \bar{\mathcal{X}}) \;=\; \frac{1}{\sum_{e \in \mathcal{E}} w_e A_e} \sum_{e \in \mathcal{E}} w_e A_e \left\| \Delta^e \mathbf{x}(e) - \mathbf{R}_e \Delta^e \bar{\mathbf{x}}(e) \right\|^2 . \tag{2.10}$$

Deformation is measured by the squared deviation of the per-edge Laplacians in the deformed state $\Delta^e \mathbf{x}(e)$ and the undeformed state $\Delta^e \bar{\mathbf{x}}(e)$. The per-edge rotations $\mathbf{R}_e \in SO(3)$ optimally align the two edge Laplacians, thereby cancelling out local rigid transformations. The magnitude of the edge Laplacians is normalized by the associated edge area A_e, given by $1/3$ of the two incident face areas [AZB15]. Additionally, we introduce spatially varying regularization weights w_e, allowing to constrain the deformation of certain regions of the mesh more than others.

Finally, the corrective term E_{shut} penalizes the squared distance between the vertex positions of pairs of vertices on the lower and upper lip:

$$E_{\text{shut}}(\mathcal{X}) \;=\; \frac{1}{S} \sum_{i=1}^{S} \left\| \mathbf{x}_i^u - \mathbf{x}_i^l \right\|^2 . \tag{2.11}$$

This term prevents the mouth of the model from opening during the fitting process, as Achenbach et al. [AWL+17] noted that the mouth is not guaranteed to stay closed, even though subjects are always scanned with a neutral face expression. The $S = 11$ vertex pairs yielding the vertex positions $\{\mathbf{x}_i^u, \mathbf{x}_i^l\}$ on the upper and lower lip respectively are pre-selected once on the template model.

This constitutes all parts of the non-linear energy function (2.7), whose optimization we will detail in the following. The energy term coefficients λ_{cpc}, λ_{lm} and λ_{reg} control (i) the influence of the closest point correspondences $\mathcal{C}$, (ii) the weight of the landmarks $\mathcal{L}$, and (iii) the general surface stiffness, and will change throughout the iterative fitting process. The coefficient of the corrective term E_{shut} is constant and set to $\lambda_{\text{shut}} = 0.5$. Given a fixed set of energy term coefficients $(\lambda_{\text{cpc}}, \lambda_{\text{lm}}, \lambda_{\text{reg}}, \lambda_{\text{shut}})$, the energy function (2.7) is optimized by alternatingly solving for new vertex positions $\mathcal{X}$ and per-edge rotations $\mathbf{R}_e$, resembling the optimization scheme used to minimize the As-Rigid-As-Possible energy [SA07]. This alternating optimization is iterated until convergence, i.e., until the relative error is below 5 %. After solving for new vertex positions, we (i) update the animation rig via mean value coordinates to conform to the new shape, and (ii) recompute the closest point correspondences between the point set and the template model.

In early stages of the fine-scale registration, the template model should only minimally deform in order to gradually attract the mesh surface towards

the point set. To this end, Achenbach et al. [AWL+17] start with a relatively stiff surface by setting $\lambda_{\text{reg}} = 1$. As in the initial registration phase, the first iterations solely rely on the specified landmarks $\mathcal{L}$, achieved by setting $\lambda_{\text{lm}} = 1$ and $\lambda_{\text{cpc}} = 0$. Equation (2.7) is then optimized until convergence, after which λ_{reg} is gradually decreased in order to reduce the surface stiffness. This process is repeated until $\lambda_{\text{reg}} = 10^{-5}$. After these first iterations, the closest point correspondences between the point set and the mesh surface are reliable enough to be incorporated into the optimization. As such, starting from $\lambda_{\text{cpc}} = 1$, $\lambda_{\text{lm}} = 1$, and $\lambda_{\text{reg}} = 10^{-5}$, the landmark weight and surface stiffness are gradually decreased until $\lambda_{\text{reg}} = 10^{-9}$. Every time, λ_{reg} is decreased, the undeformed state $\overline{\mathcal{X}}$ is updated to the current solution $\mathcal{X}$, i.e., the regularization term E_{reg} (2.10) always penalizes the deformation with respect to the solution found in the previous iteration.

During the fitting process, the per-correspondence weights w_i and the regularization weights w_e are used to account for regions that are not scanned well. These typically include the hand region, where missing data can occur due to the fingers occluding each other. Secondly, the eye region is difficult to scan due to the reflective properties of the cornea and the delicate geometric structures of the eyebrows and eyelashes. For the hand region, we weight down correspondences with respect to a vertex weighting defined on the template mesh. We additionally give the hand region a higher regularization weight w_e in order to keep the corresponding faces more stiff. As a result, the hand region is attracted towards the point set, but the shape of the hand stays close to the result from the initial registration phase. For the eye region, we use higher values of w_e to make this pre-selected region of the template mesh more stiff. This yields better results when fitting the auxiliary eyeball meshes back into the eye region, as detailed in the next subsection. The hand and eye region affected by w_i and w_e are highlighted on the template model in Figure 2.5.

After the fitting process, the model needs to be pose-normalized, as employing mean value coordinates to update the joint positions of the animation rig to conform to the new shape can alter the angles between joints. We thus update the bind pose of the resulting model, such that the angles between joints match the angles found in the bind pose of the template model. To ensure that the feet of the model are exactly on the ground after this pose-normalization step, we employ a final corrective non-rigid fitting step. First, the model is rigidly translated, such that the barycenter of a pre-selected set of vertices, which correspond to the soles of the template model, lies on the floor. After this, the vertex positions are optimized to lie on the floor plane by non-rigidly deforming them, while allowing the feet to deform only slightly by employing the regularization energy (2.10).

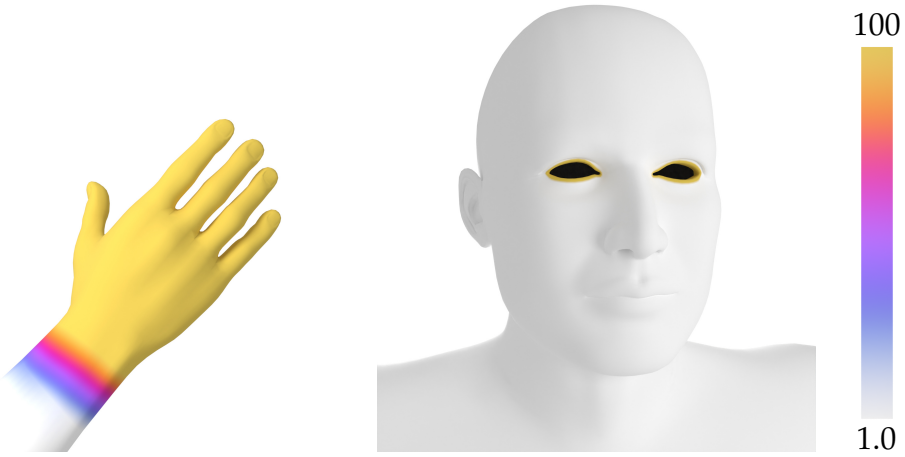

Figure 2.5: Per-correspondence weights w_i and per-edge weights w_e (visualized here) are used to weight down correspondences in the hand region (left) and make the hand and eye region (right) more stiff.

Auxiliary Meshes and Blendshapes

The result of the fine-scale registration step and the corrective pose and feet post-processing defines the final geometry of the main skin mesh of the template model. The skin mesh geometry now closely matches the scanned data (Figure 2.4 (right)), but the auxiliary eyes and teeth meshes still need to be registered to the final skin mesh geometry. To this end, Achenbach et al. [AWL+17] use a pre-selected set of vertices in the mouth and eye region respectively to compute an optimal transformation between the template and the deformed shape (inspired by Ichim et al. [IBP15]). For the teeth meshes, the optimal rotation, translation, and anisotropic scaling is computed [Hor87] and applied to transform the auxiliary meshes. Analogously, all eye meshes are transformed by the optimal similarity transformation (i.e., rotation, translation, and isotropic scaling, as to not introduce any shearing) between the template mesh and the deformed shape.

While full-body animation is controlled through the animation rig of the template model, facial animation is facilitated through a set of linear blendshapes [LAR+14]. Since the subjects are only scanned in a neutral face expression in favor of a short data acquisition duration and processing time, the approach of Achenbach et al. [AWL+17] resorts to synthetically generating a suitable set of blendshapes for the resulting virtual human. This is achieved by leveraging the generic blendshapes defined on the template mesh. Every blendshape is interpreted as a deformation from the neutral template state to the respective expression, which is then transferred onto the fitted neutral expression via Deformation Transfer [SP04]. This effectively transfers the generic template expressions onto the deformed shape and defines the set of blendshapes for the resulting virtual human. A limitation of using generic

Figure 2.6: The generic template texture (left) and the personalized texture generated by Agisoft Metashape (center) are combined in several processing steps to deal with artifacts and missing data, yielding the final color texture for the virtual human (right).

template blendshapes is, that the resulting blendshapes are not personalized, i.e., they lack subject-specific details which facial expressions usually exhibit. This could be solved by scanning the subjects while performing all blendshape expressions or a few example expressions [LWP10; MBL22], which would however increase the acquisition time.

Texture Generation

The last step of the virtual human reconstruction pipeline of Achenbach et al. [AWL+17] is to generate a high-quality color texture. For texturing the main skin mesh, Agisoft Metashape's [Agi24] texture generation process is used to create a 4096×4096 texture. The skin mesh is first re-posed from bind pose to the scanning pose $\boldsymbol{\theta}$ and transformed back into Agisoft Metashape's point set coordinate system by the inverse of the similarity transform $(s_g, \mathbf{R}_g, \mathbf{t}_g)$, after which the color information from the input images is blended onto the pre-defined UV layout of the skin mesh. Since hands and eyes are typically not scanned well and the inner mouth region and parts of the armpits are not scanned at all due to occlusions, the resulting texture suffers from some artifacts. Achenbach et al. [AWL+17] apply several post-processing steps to the generated texture in order to alleviate these problems (see Figure 2.6).

First, as the hands are typically not scanned well, the hand geometry is not fitted accurately, and thus the resulting hands cannot be accurately projected back onto the input images. As such, the texture in the hand region contains a lot of incorrect color information. Achenbach et al. [AWL+17] therefore replace the texels in the hand region by a generic texture from a set of synthetic textures from the Autodesk CharacterGenerator [Aut24]. The best matching hand texture in the database is found by computing the color difference between the generated and the synthetic texture in a small patch belonging to the top of the hand, where the texture is free of artifacts. The

hand region of the selected generic texture is then seamlessly cloned into the generated texture by using Poisson Image Editing [PGB03]. This can be done in a straightforward manner, since all resulting virtual humans share their UV layout with the template model.

Secondly, to compute texel colors for the regions belonging to the eyes and teeth of the virtual human, the color values from the generic texture of the template model are copied to the generated texture. The luminance of these regions is then adjusted to match the general luminance in the input images. For this purpose, the mean luminance difference between the generic template texture and the generated texture is computed by converting the textures to the CIELAB color space and averaging the intensities of the respective L-channels in the head region of the texture. The luminance values for eyes and teeth are then modified by adding the luminance difference to the L-channel of the generated texture.

Finally, some areas of the scanned subject are not seen by any cameras during the scanning process due to self-occlusions. When taking scans in A-pose, as in our setting, this is especially the case for the armpits. To fill in any unseen texels, harmonic color interpolation is used, which smoothly propagates the colors of the boundary region to the unseen texels. New texel colors are computed by solving the Laplace equation on the texel grid with Dirichlet constraints on the boundary of the unseen texels. These post-processing steps then result in the final color texture for the virtual human (Figure 2.6 (right)). Figure 2.7 depicts an example of a resulting virtual human, as well as an example expression that is mapped from the template model to the fitted virtual human using Deformation Transfer.

Automatic Landmark Detection

The presented pipeline allows to generate virtual humans from a 3D scan taken with our photogrammetry rig in a semi-automatic fashion. Users need to manually select the facial feature and full-body landmarks on the computed point set. To further reduce the demand on the scan operators, we adapted the template fitting approach to work in a fully automatic manner by automatically providing the required landmarks on the point set. To this end, we apply 2D pose and facial feature detection to the input images and then project the found landmarks back onto the 3D point set by using the camera calibration provided by the photogrammetry reconstruction. We use the hand and pose landmarks provided by OpenPose [CHS+21] and the facial feature detection provided as part of the dlib library [KS14].

We run the respective 2D landmark detectors on all input images and then select those detections, which are best suited for back-projecting the 2D landmarks onto the 3D point cloud. The pose and hand detection of OpenPose [CHS+21] provides up to 67 landmarks (comprised of pixel coordinates

Figure 2.7: Example of a virtual human resulting from our 3D scan and template fitting pipeline (left). After registering the auxiliary meshes of eyes and teeth to the fitted virtual human, transferring the blendshape set defined on the template model (center) to the fitted virtual human (right) is achieved via Deformation Transfer [SP04].

and confidence values): 25 full-body landmarks defining a 2D skeleton and 21 landmarks per hand. Since our input images only depict parts of the scanned subject and do not provide a full-body view (see Figure 2.2), some resulting detections can be unreliable. To address this issue, we first filter the resulting detections by discarding all images, where the 2D skeleton consists of less than 4 bones or the maximum confidence value is lower than 0.5. From the remaining images, we select those images, which depict a "side-view", i.e., images, where the viewing vector is orthogonal to the sagittal plane of the scanning subject, as these are best suited for projecting the hand and ear landmarks back to the point set. We select these images by inspecting the shoulder landmarks, which exhibit a small lateral distance in suitable images. The distinction between the left and right side of the sagittal plane is done based on the confidence values for the left and right finger and ear landmarks.

To detect the required landmarks in the face region, we use the pre-trained facial feature detection implemented in dlib [KS14] and select the required subset (see Figure 2.3 (center)) from the resulting 68 landmarks. Since the facial feature detector is trained on frontal images of human faces, we first find the most frontal image, which depicts the face region of our scanning subject. We select the most frontal image through a combined measure of horizontal and vertical frontality. Horizontal frontality is computed in terms of the symmetry of the landmarks around the center line of the face, where a higher symmetry score indicates higher frontality. Vertical frontality is computed as the ratio between eye height and eye width. The bigger this ratio, the more orthogonal

the viewing vector is to the frontal plane of the human body. From all images we assume the image with the highest sum of these measures to be the most frontal image. To further increase the reliability of the detected landmarks, we run the facial feature detector at various image resolutions and average the resulting 2D pixel coordinates after filtering out detections, which deviate more than two standard deviations from the mean.

The 2D coordinates for all landmarks can then be back-projected onto the point set due to the camera calibration provided by Agisoft Metashape. To this end, for each 2D landmark, we project the point set onto the corresponding image plane, gather all points which deviate less than 15 pixels from the 2D landmark, and from this subset, select the point closest to the camera center. This leaves us with 23 automatically detected point set landmarks which guide the template fitting process. The registration pipeline then proceeds as described, resulting in a method for reconstructing realistic personalized virtual humans in a fully automatic manner.

Results and Limitations

The presented virtual human generation pipeline allows for fast and fully automatic reconstruction of virtual humans from our photogrammetry rig. The whole process from data acquisition to the final virtual human takes about 7 min on a desktop workstation, equipped with an Intel Core i9-10850K CPU and an Nvidia RTX 3070 GPU. See Table 2.1 for timing information about all involved processes.

For optimal results, all scanning subjects are instructed to tie long hair into a knot (such that their ears lie free), remove their glasses, and wear tight clothing and little to no jewelry. Since the quality of the photogrammetry software's point set output depends on matching 2D features in the input images, we

Process	Approximate Time
Body scanning	1/15 s
Image download	45 s
Image conversion	45 s
Mask generation	60 s
Point cloud generation	100 s
Landmark detection	40 s
Template fitting	90 s
Texture generation	40 s
Complete pipeline	7 min

Table 2.1: Timings for all steps involved in our fully automatic virtual human reconstruction pipeline, taking just about 7 min in total.

Figure 2.8: Examples of virtual humans generated with our 3D scanning and template fitting pipeline.

additionally ask subjects to not wear uniformly colored or very dark or white clothing. See Figure 2.8 for examples of virtual humans generated with our 3D scanning and template fitting pipeline. In future work, we would like to be able to place less restrictions on the subjects' clothing and hairstyle by modeling clothing and hair separately from the skin mesh.

During the runtime of the HyLeC project [HyL24], we mainly scanned students of TU Dortmund University. From the various scans taken, we observed a few failure cases, where our pipeline did not successfully reconstruct a virtual human: (1) The facial feature detection sometimes gives incorrect results in the mouth region, especially for subjects with beards (see Figure 2.9 (left)). This can lead to erroneous 3D point set landmarks and thus yields inaccuracies in both the resulting geometry and texture of the virtual human. In cases where the automatic landmark detection did not work, we manually selected the corresponding landmarks and were still able to produce a virtual human in a semi-automatic manner. (2) In rare cases, the image segmentation of Deeplabv3 [CZP+18] fails to provide accurate image masks, especially if subjects wear clothing, that can be misconstrued as natural background (see Figure 2.9 (center)). If the image segmentation fails for several adjacent in-

Figure 2.9: Failure cases of the presented virtual human reconstruction pipeline. The facial feature detection can fail to correctly detect the mouth region. Note that the left mouth and the lip center line is not detected correctly (left). The image segmentation can fail to accurately segment the scanned subject. Here, the pants with a floral pattern are erroneously masked from the input image (center). Lastly, the camera calibration performed by Agisoft Metashape can misalign some of the top cameras. Three of the eight top cameras are not aligned correctly in the depicted example (right).

put images, the incorrectly masked areas are missing from the point set and are ignored for the subsequent texture generation, leading to inaccuracies in both geometry and texture. (3) The automatic camera calibration of Agisoft Metashape sometimes fails to accurately align the top cameras (see Figure 2.9 (right)). This can especially happen for shorter people, where the top cameras mostly capture background information. In these cases, the cameras can be manually re-aligned or disabled in Metashape, after which the rest of the reconstruction pipeline can still produce a resulting virtual human.

Detecting and fixing these failure cases in order to improve the success rate of the fully automatic pipeline should be tackled in future work. The facial feature detection implemented in dlib [KS14] could be evaluated against other facial feature detectors such as Google's MediaPipe [GAK+20]. Capturing the empty scanner booth and performing background subtraction could give additional information to the image segmentation algorithm and allow detecting incorrectly segmented pixels. The automatic camera calibration of Agisoft Metashape could be compared to a manually generated calibration, which would allow to selectively re-align any cameras, which deviate too much from the manual calibration.

3 REALISTIC VIRTUAL HUMANS FROM SMARTPHONE VIDEOS

Figure 3.1: From monocular smartphone videos we generate realistic virtual humans that can readily be used in game engines.

In the previous chapter, we presented a fully automatic pipeline for generating virtual humans. It employs template fitting to closely match the point set data resulting from capturing a subject with a custom-built multi-view stereo photogrammetry rig. The presented pipeline is largely based on previous work [AWL+17] and is able to reconstruct virtual humans which are ready for integration into existing computer graphics pipelines and VR environments. Previous studies have shown that embodying personalized realistic virtual humans in VR environments can improve the sense of virtual body ownership, presence, and emotional response [LRG+17; WGR+18], which hints towards their potential effectiveness in a VR therapy setting. Employing realistic virtual humans in the context of VR studies however adds additional requirements to the employed virtual human reconstruction method.

The chosen reconstruction method should be adequately fast in order to be able to create a personalized virtual human just before the VR exposure. It should be performable by non-experts, and ideally require only a lightweight and non-stationary hardware setup. However, many of the previously proposed approaches, including the one presented in Chapter 2, depend on elaborate RGB camera rigs consisting of multiple dozens to a hundred of interconnected and synchronized camera devices, resulting in complex and expensive setups (e.g., [AWL+17; FRS17]). Approaches which capture the

subjects under various lighting conditions (e.g., [GLD+19; BWS+21]) generate highly realistic virtual representations of the scanned subject, but require an immense amount of recording and processing hardware. Depending on such elaborate 3D scanner setups and potentially requiring manual pre- or post-processing steps limits the availability of personalized realistic virtual humans in VR studies. Lowering the overall complexity of the necessary sensor equipment, reducing the overall costs, and providing a fully automatic pipeline can thus open up many more of the use cases for virtual humans in digital and interactive media applications.

This chapter tackles the described problem by introducing an automated 3D reconstruction method for generating high-quality virtual humans from monocular smartphone cameras. The input of our approach are two video clips: the first video captures the whole body of the scanning subject, while the other video provides detailed close-ups of head and face. The two video clips are processed via optical flow analysis and sharpness estimation in order to select individual frames. From these images, two dense point clouds for the body and head are computed via multi-view reconstruction. Automatically detected landmarks guide the fitting of a virtual human body template to these point clouds, thereby reconstructing the geometry of the scanned subject. A graph-cut stitching approach then reconstructs detailed textures for body and head, which are fused together via Poisson Image Editing. We compare our results to existing low-cost monocular approaches as well as to expensive multi-camera scanning rigs. Our method achieves visually convincing reconstructions that are almost on par with complex camera rigs while surpassing similar low-cost approaches. The generated high-quality avatars are ready to be processed, animated, and rendered by standard XR simulation and game engines such as Unreal or Unity.

Individual Contribution *My main contribution is the development of the processing pipeline from video clips captured with commodity smartphones to photogrammetry data, which constitutes the input to a template fitting approach. I developed the image extraction pipeline, which analyzes the smartphone videos based on optical flow and image sharpness, in order to produce suitable images for photogrammetric reconstruction. To provide point set landmarks, which are required to guide the subsequent template fitting process, I developed the automatic landmark detection based on 2D pose and facial feature detectors. I adopted the template fitting pipeline used in previous work [AWL+17] to handle the less reliable input data, while Jascha Achenbach supported the general integration of the generated photogrammetry data into the template fitting pipeline and additionally incorporated a statistical model of human head shapes into the optimization. Finally, I developed the graph cut based texture generation, which computes a high-quality color texture from the input images. Andrea Bartl integrated the resulting virtual humans into the Unity Game Engine and produced the blendshape retargeting.*

Corresponding Publication *This chapter is based on the following publication*:

Stephan Wenninger, Jascha Achenbach, Andrea Bartl, Marc Erich Latoschik, and Mario Botsch. "Realistic Virtual Humans from Smartphone Videos". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2020, 29:1–29:11. (Best Paper Award 🏆)

3.1 RELATED WORK

As an alternative to reconstructing avatar models, one can record, transmit, and render streams of depth images from RGBD cameras, which creates believable reproductions of recorded users [LBW+15]. However, the quality of reproduction crucially depends on a sufficient resolution in both the spatial, color, and temporal domain of the employed RGBD cameras, which, as of today, still are significantly lower compared to dedicated high-quality sensors. Some performance capture approaches fuse RGBD streams from one or multiple sensors into a volumetric representation, from which a textured mesh is extracted [DDF+17; GXY+17]. These methods are template-free, i.e., they do not include a prior of human performances, and thus allow real-time reconstruction of challenging scenes of people interacting with objects. However, these approaches are restricted to mere reproductions of human performances, whereas full 3D virtual humans allow for more flexibility due to their separation of static geometry and appearance from dynamic animation. Furthermore, template-free approaches need to transmit a lot of data every frame, be it 3D meshes or depth images, which requires a lot of bandwidth to be provided by the employed network. Approaches which employ virtual human template models can transmit the static avatar data once and only need to transmit dynamic low-dimensional pose parameters at each frame. Guo et al. [GLD+19] present a hybrid approach for volumetric relightable performance capture. They record users in a light stage consisting of 90 high-resolution infrared and RGB cameras and 331 programmable lights in order to capture geometry and reflectance properties. In order to compress this data, they generate, parameterize and track a 3D mesh over time, only changing mesh triangulation, once the tracking error becomes too high. This system allows high-fidelity photorealistic performance capture at the cost of a large amount of data and processing power.

A different line of work for generating virtual humans – which we have discussed in Chapter 2 and will also follow here – exploits template models to guide the reconstruction process. See, e.g., Egger et al. [EST+20] and Zollhöfer et al. [ZTG+18] for an overview of parametric face models. Similarly, human body models, such as the SCAPE model [ASK+05], have been used as template

models for *full-body* reconstruction [PWH+17]. Later models from the SMPL family [LMR+15], like SMPL-H [RTB17], SMPL-X [PCG+19], or STAR [OBB20] provide additional features like hand and finger movements, facial expressions, or (sparse) pose-dependent blendshapes.

Template-based performance capture methods employ an actor-specific model for tracking the movements of a person. For instance, Habermann et al. [HXZ+19] generate this model by capturing an RGB video of the actor in a static pose, extracting around 70 frames, reconstructing a textured mesh through photogrammetry, manually embedding a skeleton, and computing rigging weights using Blender. Our approach can act as a fully automatic alternative to their preprocessing stage. Besides providing more geometric details and animation controllers (fingers, facial expressions), it has the advantage that all actor models share the connectivity of the template mesh, allowing for statistical regularization.

The highest quality for avatar reconstructions is achieved using elaborate multi-camera rigs with high-quality image sensors, which often consist of dozens to over a hundred DSLR cameras, as discussed in Section 2.1. Through multi-view stereo, these approaches accurately reconstruct geometry and texture (see, e.g., [PRM+15; LMR+15]). The virtual humans of Feng et al. [FRS17] and Achenbach et al. [AWL+17] are reconstructed from such camera rigs (in 20 and 10 minutes, respectively) and feature skeleton-based body and hand animation as well as blendshape-based facial expressions. However, their complex hardware setup restricts the availability (and hence applicability) of their approaches. Template-based human body models can also be generated from consumer-level RGBD sensors (e.g., [BBL+15]), but the low spatial resolution and limited image quality leads to rather low-quality reconstructions. Malleson et al. [MKK+17] therefore use an RGBD sensor in combination with a stereo RGB camera pair, but their avatars are still of rather low quality, lack facial details, and reconstruct the body in a stylized manner only.

Lowering hardware requirements to the extreme, several learning-based techniques reconstruct 3D body models from a single RGB/RGBD input image or sequence of video frames [BKL+16; KBJ+18; OLP+18; FYR+19; KAB20]. However, these methods optimize parameters of a low-dimensional body model only, without considering fine-scale per-vertex displacements, which inherently limits the accuracy of the shape reconstruction. Moreover, they all do not consider texture reconstruction, which is crucial for realistic avatar appearance. Alldieck et al. [APT+19] reconstruct textured avatars from a single image, by synthesizing normal/displacement maps from a partial texture calculated through DensePose [GNK18] and mapping them onto the SMPL model. Recently, there have been further advances in reconstructing avatars from a single image [AZS22; KSL+22; LZX+23; SAK+24]. Khakhulin et al. [KSL+22] propose a method for reconstructing head avatars from a single in-the-wild photo. They use the DECA model [FFB+21] to retrieve an initial mesh which

is refined using a neural texture to better predict hair geometry. The result is then processed via neural rendering to produce the final image. Similarly, Liao et al. [LZX+23] present an approach for generating full-body avatars from a single image by first optimizing SMPL parameters [LMR+15] and then learning to predict a signed distance function representation in a canonical pose, which is warped and refined to generate a detailed model. While these methods show impressive results given the minimal input, limiting the input to a single image inevitably restricts the faithfulness of the reconstructions.

Alldieck et al. [AMX+18a; AMX+18b] therefore reconstruct a textured and animatable avatar from a monocular RGB video that captures a subject turning 360 degrees in A-pose. Their model is based on SMPL, which is fitted to the subjects silhouettes, extracted by CNN-based semantic segmentation, in a subset of the video frames. The shape is further refined using shape-from-shading techniques, and an albedo texture is generated via a per-texel graph cut optimization with a semantic prior [AMX+18b]. In follow-up work, Alldieck et al. [AMB+19] estimate the SMPL parameters from only 1–8 input images, based on a neural network that incorporates semantic segmentation and estimated 2D landmarks. The texture is again generated via their previously proposed method [AMX+18b]. While their approaches reconstruct full avatars from consumer-level input, we show that our approach leads to higher accuracy and realism. Our approach is inspired by Ichim et al. [IBP15], who generate a quite accurate personalized head model from a smartphone selfie video. From this video, they reconstruct a dense point cloud, to which they fit a parametric template model. We extend their ideas to the challenging case of full-body avatars with detailed hands and faces.

Recently, methods which use video data from a single camera as input for avatar generation became more and more popular. Cao et al. [CSK+22] present a neural network architecture, which is trained from high-resolution multi-view stereo data depicting different facial expressions. From this data, they learn identity, expression, and decoder networks which produce a volumetric avatar representation rendered via ray marching. The input for the decoder can be inferred from a single head scan depicting a neutral face expression taken with a smartphone. The resulting representation can be further personalized and refined via expression recordings, facilitating personalized head avatar generation in about six hours. Zielonka et al. [ZBT23] represent a dynamic head avatar via a neural radiance field embedded in a multi-resolution grid around a tracked FLAME mesh [LBB+17], allowing them to efficiently train their neural radiance field in less than ten minutes. Implicit representations based on (deformable) neural radiance fields have also been explored in the context of full-body avatars [JHB+22; ZHY+22; WCS+22; GJC+23; JCS+23; ZZZ+23; JGK+24]. A common approach is to capture a subject rotating in front of a static camera and then implicitly represent the resulting geometry in a canonical space, which is deformed in order to account for non-rigid

clothing deformation and pose variation from frame to frame. These methods exploit recent advances in neural rendering and implicit scene representation via neural radiance fields, which are well suited for structures with fine details such as hair and clothing, as they are not bound to a fixed mesh topology. However, explicit representations are faster to render and easier to integrate into existing graphics pipelines, which is especially important in the context of employing the resulting virtual humans in the context of VR therapy. Most of the approaches are unable to render virtual humans at adequate frame rates, e.g., HumanNeRF [WCS+22] achieves 0.14 fps, the InstantAvatar approach [JCS+23] reaches 15 fps, and AvatarReX [ZZZ+23] is able to render virtual humans at 25 fps.

Recently, the Gaussian Splatting method [KKL+23] showed promising results in high-fidelity and real-time radiance field rendering. This approach represents a scene by a set of 3D Gaussians, whose attributes (position, covariance, opacity, and color) are optimized to match the input images when using volume splatting to project the Gaussians onto the image plane. After its introduction, this scene representation has also been employed to represent head avatars [CWL+24; XCL+24; SSS+24; XGG+24; QKS+24] as well as full-body virtual humans [HZZ+24; HHL24; WZR+24; MSS24; SWL+24]. These methods commonly tie the positions of the 3D Gaussians to a tracked mesh, e.g., by first optimizing a SMPL [LMR+15] or FLAME [LBB+17] mesh for each frame. The deformation between poses can then be estimated via, e.g., Linear Blend Skinning, thereby exploiting the underlying mesh representation. To model pose-dependent non-rigid motion, a common approach is to train additional neural networks to predict the residual deformation of the Gaussians. Similar to implicit neural field representations, methods based on Gaussian Splatting are well suited for complex geometric structures like hair and clothing. While being faster to render than neural fields, some of the cited approaches still yield frame rates that are too low for use in VR scenarios (43 – 47 fps [WZR+24; MSS24]). Only the SplattingAvatar [SWL+24] and GauHuman [HHL24] approaches yield high enough frame rates when rendering a single avatar in an empty scene (187 – 300 fps).

The discussed reconstruction methods differ significantly in the faithfulness of their resulting models and in their costs, including hardware requirements and the amount of manual intervention needed. High-quality results with few manual interventions is offered by complex multi-camera rigs, like the one used by Achenbach et al. [AWL+17]. In contrast, the approaches of Alldieck et al. [AMX+18a; AMX+18b; AMB+19] require only a single affordable camera, but the quality of their reconstructions is considerably lower than the one achieved by multi-camera rigs. In the following, we describe a method that combines the advantages of both approaches, generating high-quality fully animatable virtual humans from video sequences captured by a consumer-level monocular smartphone camera.

3.2 METHOD

Our avatar generation is inspired by the smartphone-based head scanning of Ichim et al. [IBP15] and builds on our extensions of the previous work on *Fast Generation of Realistic Virtual Humans* by Achenbach et al. [AWL+17], which we will abbreviate as *FGVH* in this chapter. We combine and closely follow these two approaches, but extend them in several important aspects in order to enable full-body avatar reconstructions from simple monocular smartphone videos.

In FGVH [AWL+17], people were scanned using two custom-built single-shot multi-camera rigs: a full-body scanner and a face scanner, consisting of 40 and 8 DSLR cameras, respectively. Given the camera images, a multi-view stereo reconstruction computes two high-quality point clouds for body and face, to which a human body template is fitted using non-rigid (or deformable) registration (see Section 2.2 for an in-depth description of the template fitting method). Since the employed template model features a detailed skeleton for body and hands as well as eyes, teeth, and facial blendshapes, the reconstructed virtual humans are ready for animation in XR simulation and game engines such as Unity or Unreal. The main drawback of FGVH is the expensive, elaborate and stationary hardware setup, an issue shared by several character reconstruction/tracking methods [LMR+15; FRS17; JSS18; GLD+19].

In order to make 3D scanning and avatar generation available to a wider range of people, we considerably lower the hardware requirements and employ a consumer smartphone camera only. We take two video clips of a person, the first one capturing the full body, the second one capturing the head of the subject. From these video clips we automatically select individual frames using optical flow analysis and sharpness estimation (Section 3.2.1), and compute two dense point clouds for the body and head using multi-view stereo reconstruction (Section 3.2.2), thereby resembling the body and face scan of FGVH. The template fitting process presented in FGVH relies on a set of manually picked landmarks on both the body and face point clouds to guide the shape optimization. In contrast, we automate this process (Section 3.2.3) by following the automatic landmark detection described in Section 2.2.2. Given the found landmarks, we then pose and deform a statistical human body template to closely fit the body and face point clouds (Section 3.2.4). When reconstructing the model's texture from the input frames, we cannot rely on standard multi-view reconstruction because of imperfections in our input data. Instead, we employ a graph cut texture stitching approach, which yields visually superior results (Section 3.2.5).

3.2.1 Input Data

Previous works on monocular reconstruction [AMX+18b; AMB+19] facilitate avatar creation from low-cost setups by taking one video that captures the full body of the person. However, we noticed (analogous to FGVH) that a separate head scan improves the quality and detail of the avatar's head region, especially when dealing with lower-resolution data from smartphone videos compared to DSLR camera images. One approach for acquiring a close-up scan of the head would be to simply include it in the video for the full-body scan. However, since we employ a multi-view stereo approach, we rely on the person holding as still as possible during the capture process. Increasing the length of the video by including a detailed scan of the head would imply more motion of the scanning subject and thereby result in a stronger violation of the multi-view stereo assumption.

Instead, we take two videos of the person, the first one capturing the full body in A-pose from a slight distance and the second one capturing the head in a close-up fashion. For the full-body video, the smartphone camera is moved (by a second person) in two circular paths around the scanned subject: The first camera path captures the upper body (head, torso, arms), the second one the lower body (hips, legs, feet). The head scan consists of one circular camera motion around the subject's head and additionally films the top of the head and the region under the chin (Figure 3.2).

Our input videos are shot at 4 k resolution (3840 × 2160) and 30 Hz frequency on a Google Pixel 3. Experiments with other smartphones capable of capturing 4 k videos gave similar results. The full-body video takes about 80 s and the head video about 30 s. The scanned subjects cannot hold perfectly

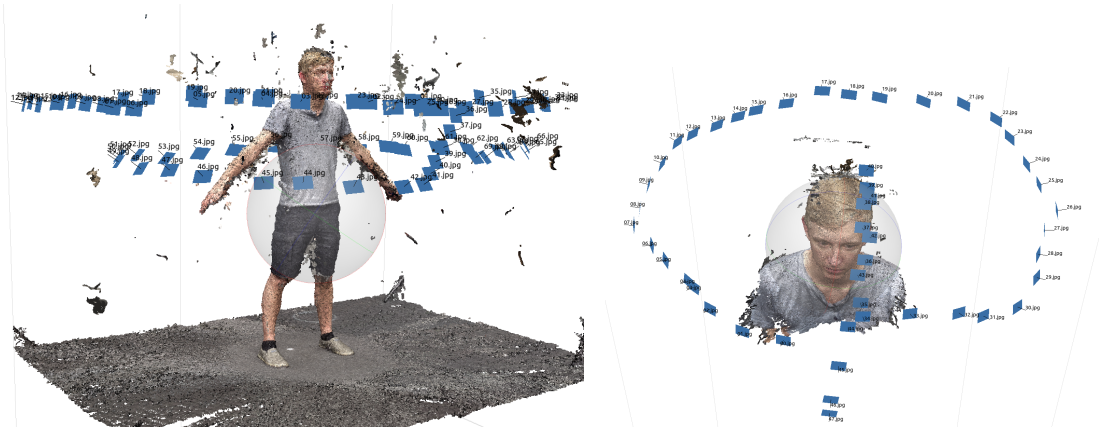

Figure 3.2: Camera locations for the full-body scan, consisting of two orbits around the scanned subject (left), and the head scan, taking a close-up of the head/face region (right).

still for this long, but we found that we could still employ a multi-view stereo approach and produce point clouds of sufficient quality.

To this end we first select I frames of the input video, which are then processed by the multi-view stereo reconstruction (Agisoft Metashape Pro [Agi24] in our case) in order to compute the point clouds for the subsequent template fitting pipeline. Using all frames of the input video would rapidly exceed the capabilities of the photogrammetry software. Our experiments revealed that extracting $I = 75$ images from the full-body video and $I = 50$ images from the head video is a good trade-off between computation time and resulting point cloud quality.

Simply extracting every nth video frame would not account for any non-uniform camera movement by the person performing the scan. To simplify the capturing process while ensuring a uniform coverage of the scanned subject, we instead extract frames based on a uniform inter-frame movement, which we estimate through optical flow analysis using the implementation of Farnebäck [Far03] in OpenCV [Bra00]. This yields a dense 2D flow field $\mathbf{f}_i$ representing the movement between frames i and $i+1$, from which we estimate the *amount* of movement f_i as the average length of the 2D flow vectors in $\mathbf{f}_i$. We treat the resulting inter-frame movements f_i as a noisy 1D signal and smooth it by convolution with a Gaussian kernel ($\sigma = 2$) to compute filtered movement estimates $\tilde{f}_i$. We then iterate through the video and select a new frame once the *accumulated* movement between it and the previously selected frame reaches the threshold $\frac{1}{I}\sum_i \tilde{f}_i$. This defines a set of frames with uniform movement in between them.

We noted, however, that frames selected by the above procedure might be blurry either due to motion blur or the camera being in the process of adjusting the focus. To tackle this problem, we find the sharpest frame in the k-neighborhood $\mathcal{N}_k$ of each selected frame ($k = 5$ in our experiments). Sharpness is estimated as the variance of the Laplacian of the input image [PCC+00] and we select the frame in $\mathcal{N}_k(i)$ which exhibits the highest sharpness value. Finally, we change the orientation of the selected frames according to the EXIF metadata of the video and pass the selected frames $\{\mathbf{I}_1, \ldots, \mathbf{I}_I\}$ to the photogrammetry reconstruction.

3.2.2 Point Cloud Generation

The photogrammetry software Agisoft Metashape [Agi24] proceeds in several steps: First, feature points are detected and matched in between individual input images. Based on these sparse (but reliable) points, the intrinsic and extrinsic camera parameters are computed for each input image. Finally, given the camera calibration, the dense point cloud is computed.

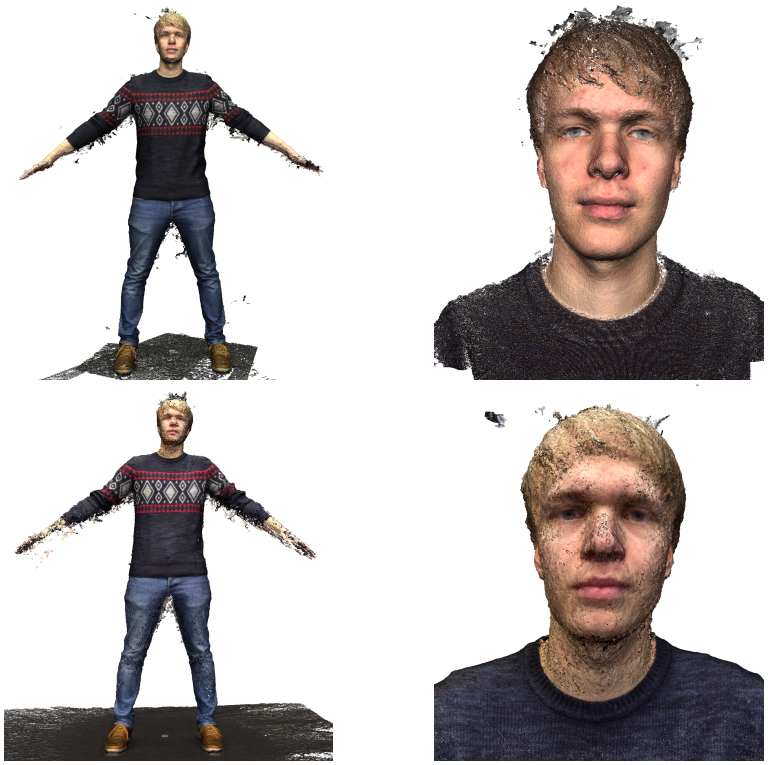

Figure 3.3: Comparison of full-body (left column) and face/head (right column) point clouds between FGVH (top row) and our approach (bottom row). Our point clouds are noisier, less detailed, and more likely to have missing data (e.g., in the arm region).

For the last step, the software allows to restrict the computation of the dense point cloud to an oriented bounding box. This will speed up not only the photogrammetry algorithm, but also all subsequent steps of our pipeline, because the resulting point cloud consists of fewer points. Due to our handheld video input, we cannot rely on a pre-calibrated camera setup and, thus, cannot rely on a constant scanning volume of interest.

However, we know that the camera positions provided by the extrinsic camera calibration enclose the scanned subject, hence we can use them to estimate the bounding box. We first determine an oriented box through PCA of the camera positions, where by design of our camera trajectory (see Figure 3.2) the first two principal directions $\mathbf{e}_1$ and $\mathbf{e}_2$ span the least squares fitting plane through the camera locations, and $\mathbf{e}_3$ corresponds to its normal, i.e., the up-direction. From the extent of the camera box in directions $\mathbf{e}_1$ and $\mathbf{e}_2$ we can estimate the subject's arm span and, since the arm span of humans roughly corresponds to their height, also the height of the bounding box. The bounding box of the head scan is determined in the same manner, making the assumption that the height of person's head roughly corresponds to its width.

After specifying the two bounding boxes, Agisoft Metashape computes dense point clouds from the selected input camera images, leading to a point cloud $\mathcal{P}_B$ for the full-body scan (ca. 2.8 M points) and a point cloud $\mathcal{P}_H$ for the

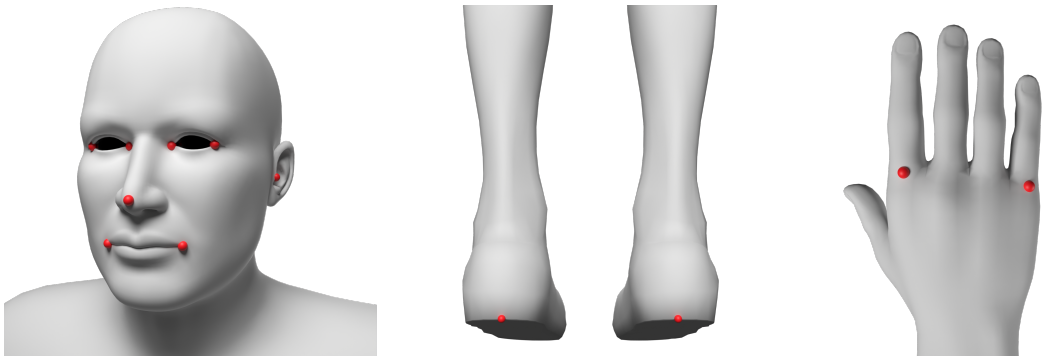

Figure 3.4: We use 15 landmarks on the full-body scan to guide the template fitting process. The location of these landmarks is visualized here on the template mesh.

head scan (ca. 1.6 M points). Due to the lower resolution of our smartphone camera and the inevitable slight motion of the scanned subject during the capture process, our point clouds are more noisy and more likely to have missing data than the point clouds in FGVH (see Figure 3.3 for a comparison).

3.2.3 Landmark Detection

The template fitting procedure (Section 3.2.4) is bootstrapped and guided by feature landmarks on the point clouds $\mathcal{P}_B$ and $\mathcal{P}_H$. While in FGVH landmarks in the reconstructed point clouds are manually selected, we propose a fully automatic landmark detection. To this end, we follow the approach described in Section 2.2.2 and perform 2D landmark estimation using OpenPose [CHS+21] on all input images.

To recall, OpenPose gives us up to 135 landmarks (including confidence values) for each image: 25 full-body landmarks defining a 2D skeleton, 21 landmarks per hand, and 68 facial landmarks. In Section 2.2.2, the 68 facial landmarks were alternatively detected using the facial feature detection implemented in the dlib library [KS14] due to an incompatibility of the OpenPose facial feature detection with the operating system used. Since both methods yield the same landmark set, they can be used interchangeably. The detected landmarks are then filtered in order to deal with unreliable detections.

For the full-body point cloud $\mathcal{P}_B$ we only use a small subset of 15 landmarks (shown in Figure 3.4), since not all of the 135 landmarks can be reliably back-projected from their 2D image location onto the 3D point cloud (using the camera calibration data from the photogrammetry reconstruction). However, this subset turned out to be fully sufficient to guide the full-body template fitting process.

As in Section 2.2.2, we find images that allow for the most robust back-projection from 2D image coordinates onto the 3D point cloud $\mathcal{P}_B$. We choose

Figure 3.5: Result of the automatic landmark detection. We heuristically find the best image for each landmark. Nose, mouth, and eye landmarks are projected from frontal images (left), hand and ear landmarks from lateral images (center) and heel landmarks from dorsal images (right).

the most suitable image based on the following heuristics: Hand and ear landmarks should be back-projected from images orthogonal to the sagittal plane, while heel, nose, mouth, and eye landmarks should be back-projected from images orthogonal to the frontal plane of the human body (see Figure 3.5).

For finding suitable images for the additional heel landmark projection, we look for several characteristics: For one, the left and right heel landmarks have to be located on the left and right side of the image, respectively. However, OpenPose mislabels left and right legs in some cases, so we additionally use the fact that in suitable images the toe landmarks always have to be above the heel landmarks. In images orthogonal to the frontal plane, the heel landmarks should also approximately be located at the same height in the input image.

The landmarks in the selected images are then back-projected onto the point cloud $\mathcal{P}_B$ using the camera calibration provided by the photogrammetry software. The same procedure is repeated for the head scan: We find the most frontal image as described in Section 2.2.2 and project the 68 facial landmarks onto the head point cloud $\mathcal{P}_H$. The resulting point set landmarks are then weighted to account for the fact, that projecting the face contour points (Figure 3.5 (left)) yields unreliable results.

The landmark detection of OpenPose in combination with our filtering and back-projection yields (in a fully automatic manner) 3D landmark positions (15 for the full-body scan, 68 for the head scan), which guide the subsequent template fitting procedure. Note that the back-projection might fail due to missing data in low-quality point clouds. In this rare case, we prompt the user to manually select the corresponding 3D landmark (see Figure 3.14).

3.2.4 Template Fitting

Reconstructing a high-quality avatar mesh from medium-quality scanner data is a challenging problem because of noise, outliers, and holes in the input data. Like FGVH, we exploit prior knowledge (that we are scanning humans) and fit a statistical human body model to the scanner point cloud(s) by optimizing the template's position, orientation, scaling, PCA parameters, and fine-scale per-vertex deformation as detailed in Section 2.2. In this way, the template mesh regularizes the fitting procedure and fills in regions of missing data. Our template fitting approach closely follows the non-rigid registration of FGVH, but extends it at several places in order to deal with our lower-quality data.

We use the same template character from the Autodesk Character Generator [Aut24], which is fully rigged and capable of body, hand, and face animations. The template mesh consists of $V \approx 21\,\text{k}$ vertices with positions $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_V)$. In order to incorporate a statistical prior on human body shapes, we fit this template model to about 1700 human scans from the CAESAR database [RBD+02] and compute a 30-dimensional PCA subspace from the resulting data. This yields a more expressive statistical model – and hence a more robust fitting process – than FGVH, where a 10-dimensional PCA is computed from about 200 scans from mixed sources [BRL+14; HSS+09; Aut24], including synthetic, non-realistic ones [Aut24].

Following FGVH, we uniformly down-sample the two point clouds to twice the vertex density of the template mesh in order to speed up the fitting process, resulting in ca. 150 k points each for the body scan and the head scan. By considering the vertex density of the template mesh, we ensure that geometric details that the template mesh can reproduce, are preserved.

In the first step we fit the template model to the body point cloud $\mathcal{P}_B$ by following the two-step registration scheme detailed in Section 2.2.2. To recall, we first minimize the squared distances between the 15 automatically detected landmarks in the point cloud $\mathcal{P}_B$ and their pre-selected counterparts on the template model by alternatingly (i) computing the optimal scaling s_g, rotation $\mathbf{R}_g$, and translation $\mathbf{t}_g$ [Hor87], (ii) optimizing joint angles $\boldsymbol{\theta}$ through inverse kinematics [ALC+18], and (iii) optimizing the PCA shape parameters $\boldsymbol{\beta}$ (linear least squares problem). After convergence, we further improve alignment, pose, and PCA shape by minimizing, in addition to landmarks, the squared distances between points in $\mathcal{P}_B$ and their closest points on the template mesh.

This defines the initial registration of the template mesh, which in a second step is refined by a fine-scale non-rigid registration to match the point cloud data more closely. Let $\overline{\mathcal{X}} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_V)$ be the vertices resulting from the initial

registration phase. We then perform a fine-scale non-rigid registration by minimizing the non-linear objective function

$$\begin{aligned} E_{\text{body}}(\mathcal{X}) &= \lambda_{\text{cpc}} E_{\text{cpc}}(\mathcal{X}) + \lambda_{\text{lm}} E_{\text{lm}}(\mathcal{X}) + \lambda_{\text{reg}} E_{\text{reg}}\left(\mathcal{X}, \overline{\mathcal{X}}\right), \\ E_{\text{cpc}}(\mathcal{X}) &= \sum_{i=1}^{C} w_i \left\| \text{skin}_{\mathcal{C}}\left(\mathbf{x}_i^{\mathcal{C}}, \boldsymbol{\theta}\right) - \tilde{\mathbf{p}}_i^{\mathcal{C}} \right\|^2, \\ E_{\text{lm}}(\mathcal{X}) &= \sum_{i=1}^{L} w_i^{\text{lm}} \left\| \text{skin}_{\mathcal{L}}\left(\mathbf{x}_i^{\mathcal{L}}, \boldsymbol{\theta}\right) - \tilde{\mathbf{p}}_i^{\mathcal{L}} \right\|^2, \\ E_{\text{reg}}\left(\mathcal{X}, \overline{\mathcal{X}}\right) &= \frac{1}{\sum_{e \in \mathcal{E}} w_e A_e} \sum_{e \in \mathcal{E}} w_e A_e \left\| \Delta^e \mathbf{x}(e) - \mathbf{R}_e \Delta^e \bar{\mathbf{x}}(e) \right\|^2. \end{aligned} \tag{3.1}$$

The data term E_{cpc} penalizes the squared distances between corresponding closest points on the point cloud $\tilde{\mathbf{p}}_i^{\mathcal{C}}$ and skinned points $\text{skin}_{\mathcal{C}}\left(\mathbf{x}_i^{\mathcal{C}}, \boldsymbol{\theta}\right)$ on the template mesh surface (expressed via barycentric coordinates). Using $w_i \in [0, 1]$, we weight down correspondences in the hand and head regions, since the former are typically unreliable and the latter will be replaced by the head scan. Similarly, the landmark term E_{lm} penalizes the squared distance between the 15 automatically detected landmarks in the point cloud and their corresponding vertices on the template mesh. The regularization term E_{reg} penalizes the geometric distortion from the undeformed state $\overline{\mathcal{X}}$ to the deformed state $\mathcal{X}$, by measuring the deviation of the respective edge-Laplacians, aligned by per-edge rotations $\mathbf{R}_e$. In this way, we attract the mesh surface towards the point cloud (E_{cpc} and E_{lm}) while only allowing physically plausible deformations (E_{reg}). For more details about the computation of the individual terms, we refer the reader to Section 2.2.2, specifically to Equations (2.8), (2.9), and (2.10).

This non-linear least squares problem is solved using an alternating optimization of vertex positions $\mathcal{X}$ and per-edge rotations $\mathbf{R}_e$ (repeated block-coordinate descent), where we set λ_{cpc} to 1 and gradually decrease λ_{lm} from 0.1 to 10^{-4} and λ_{reg} from 1 to 10^{-7}. This follows the iterative optimization scheme proposed in FGVH and detailed in Section 2.2.2. However, due to the lower point cloud quality resulting from motion artifacts and the lower-resolution input, we maintain a higher level of regularization and only lower the regularization weight to $\lambda_{\text{reg}} = 10^{-7}$ instead of 10^{-9} as in FGVH.

Having deformed the template model to the full-body scan, we further refine the geometry of the head region by fitting it to the head scan $\mathcal{P}_H$. In order to align the template model to the head scan, we find optimal scaling, rotation, and translation by minimizing squared distances between the detected 68 facial landmarks and their corresponding landmarks on the template model [Hor87]. Afterwards, we further refine scaling, rotation, and translation through ICP [BM92]. In contrast to FGVH and due to our more noisy point clouds, we regularize the head fit by a 30-dimensional statistical head model derived from the publicly available data of [ABG+18]. After this coarse

registration, we add fine-scale geometric detail by performing a non-rigid deformation that minimizes the objective function

$$\begin{aligned} E_{\text{head}}(\mathcal{X}) \; &= \lambda_{\text{cpc}} E_{\text{cpc}}(\mathcal{X}) + \lambda_{\text{lm}} E_{\text{lm}}(\mathcal{X}) + \\ &\quad \lambda_{\text{reg}} E_{\text{reg}}\left(\mathcal{X}, \overline{\mathcal{X}}\right) + \lambda_{\text{shut}} E_{\text{shut}}(\mathcal{X}) \,, \end{aligned} \tag{3.2}$$

where E_{cpc}, E_{lm}, and E_{reg} are the same as before (but restricted to the head region), and E_{shut} ensures that the mouth of the template model stays closed (see Equation (2.11)). The iterative optimization then proceeds in the same way as before. However, this time λ_{reg} is initially weighted by 1 and gradually decreased to 10^{-8}. We again solve the non-linear least squares problem using repeated block-coordinate descent.

After the fine-scale non-rigid registration, we pose-normalize the model and perform a corrective non-rigid registration which puts the feet of the model on the ground (Section 2.2.2). Finally, we add facial details (eyes and teeth) and reconstruct blendshapes. Following FGVH we adjust the template's teeth and eyes by optimizing for scaling, rotation, and translation based on the deformation of the mouth and eye region. To resolve occasional penetrations of the eyes and eyelids, we non-rigidly deform the eyelids to fit the transformed eye geometry. To reconstruct blendshapes, we map all blendshapes from the template mesh to the fitted model using deformation transfer [SP04].

3.2.5 Texture Generation

Given the camera images and the reconstructed avatar mesh, FGVH computes textures for the full-body scan and the face scan using Agisoft Metashape and blends them using Poisson Image Editing [PGB03]. In our case, this approach leads to noticeable artifacts because of inaccuracies in the geometry reconstruction caused by inevitable motion during the capture process, as shown in Figure 3.8. We avoid these problems by computing the texture image through a graph cut optimization [BVZ01].

Using the fitted avatar mesh (Section 3.2.4) and the camera calibration data of Agisoft Metashape (Section 3.2.2), we generate partial textures by rendering the avatar mesh from each camera position. The projection from 3D world coordinates to the respective camera's image plane is modeled as a standard pinhole camera with Brown's distortion model [Bro71], whose parameters are provided by Metashape's intrinsic and extrinsic camera calibration.

This projection is used to generate a partial texture $\mathbf{T}_i$ from each input image $\mathbf{I}_i$ in a two-pass rendering process implemented via OpenGL. In the first pass, all triangles of the resulting avatar mesh are projected onto the image plane of camera $\mathbf{c}_i$ and the resulting depth buffer is stored as $\mathbf{Z}_i$. The second render pass generates the partial texture $\mathbf{T}_i$ by rendering the mesh onto the pre-defined uv-layout of the template character. The fragment shader

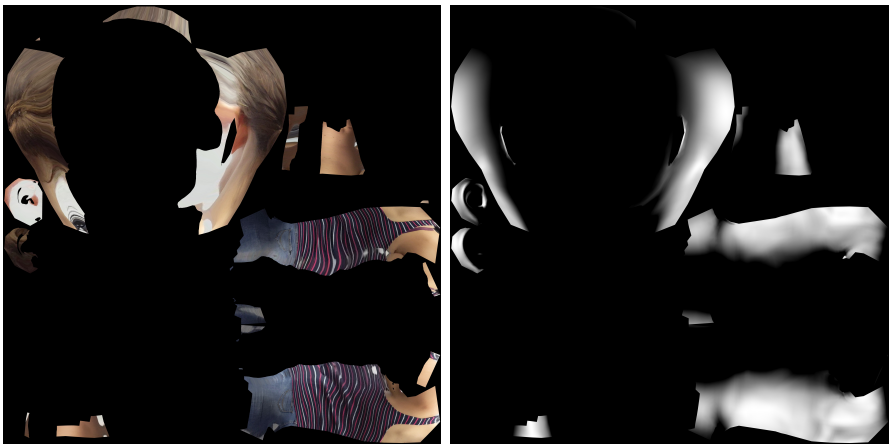

Figure 3.6: Partial texture (left) and visibility map (right).

then discards all fragments that do not pass the depth test against the depth buffer $\mathbf{Z}_i$ from the first rendering pass. This discards all fragments that are not visible from camera $\mathbf{c}_i$. For all remaining fragments, the color value is computed by accessing the camera image $\mathbf{I}_i$ at the texture coordinate $\mathbf{u}_j$ defined by projecting the interpolated surface point $\mathbf{x}'_j$ onto the image plane of camera $\mathbf{c}_i$. We additionally compute the angle α between the surface normal and the viewing ray and discard all fragments where α exceeds a threshold of 45° in order to rule out foreshortening effects. Color information for the remaining fragments is then written to the corresponding texture coordinate at $\mathbf{T}_i$. This rendering procedure results in a partial color texture $\mathbf{T}_i$ and visibility map $\mathbf{V}_i$ (storing $\cos(\alpha)$ for each pixel) for every input image (see Figure 3.6). Since the input images exhibit some overlap in order to facilitate multi-view stereo reconstruction, so do the resulting partial textures, and we are left with the task of generating a complete texture by performing texture stitching.

Stitching the partial textures together could be done by simply performing a "best view" selection, i.e., coloring each texel from the partial texture where the corresponding surface patch was most orthogonal to the viewing vector (see, e.g., [IBP15]). However, because of the inevitable motion during our scanning procedure, the camera calibration resulting from the photogrammetry step is not accurate enough, leading to noisy point clouds. As a consequence, the reconstructed geometry is not accurate enough, and thus the partial textures do not align perfectly. Performing a best view selection would thus lead to noticeable seams between surface patches.

Graph cut methods [BVZ01] have been used to seamlessly stitch together images or textures. We take inspiration from various works [GWO+10; LI07; AMX+18b] and formulate our texture stitching as a combinatorial optimization. Each of the F faces of the reconstructed mesh is to be textured by one of the partial textures $\mathbf{T}_i$. This can be described by an index set $\mathcal{I} = \{l_1, \ldots, l_F\}$ with

Figure 3.7: The patches induced by the best view selection (left) and by our graph cut optimization (right). The latter leads to larger patches and fewer seams.

$l_i \in \{1, \dots, I\}$, which labels each face with a partial texture index. The graph cut optimization then minimizes the error function

$$\begin{aligned} E_{\text{tex}}(\mathcal{I}) &= \sum_{i=1}^{F} \mathrm{D}(f_i, l_i) \; + \sum_{i,j=1}^{F} \mathrm{S}(f_i, f_j, l_i, l_j)\,, \\ \mathrm{D}(f_i, l_i) &= \frac{1}{|\mathcal{U}(f_i)|} \sum_{\mathbf{u} \in \mathcal{U}(f_i)} \left(1 - \mathbf{V}_{l_i}(\mathbf{u})\right)^2, \\ \mathrm{S}(f_i, f_j, l_i, l_j) &= \frac{1}{|\mathcal{U}(f_i, f_j)|} \sum_{\mathbf{u} \in \mathcal{U}(f_i, f_j)} \left\| \mathbf{T}_{l_i}(\mathbf{u}) - \mathbf{T}_{l_j}(\mathbf{u}) \right\|^2, \end{aligned} \tag{3.3}$$

with a data term $D(f_i, l_i)$ and a smoothness term $S(f_i, f_j, l_i, l_j)$. The data term prefers to texture faces from input images where the face normal is parallel to the viewing vector, summing up the visibility map $\mathbf{V}_{l_i}$ for partial texture $\mathbf{T}_{l_i}$ over the set $\mathcal{U}(f_i)$ of texels of face f_i in uv-coordinates. The smoothness term ensures that neighboring faces are textured from images that avoid visible seams, by penalizing color differences on the texels of their shared edge $\mathcal{U}(f_i, f_j) = \mathcal{U}(f_i) \cap \mathcal{U}(f_j)$ in uv-coordinates.

We treat the objective function (3.3) as a multi-label graph cut optimization problem [BVZ01; KZ04; BK04]. This defines a Markov Random Field that we optimize with the implementation provided by Szeliski et al. [SZS+06]. We initialize the optimization with the best view selection, which is a good starting point since it is equal to the minimum of the data term.

The resulting labeling $\mathcal{I}$ defines which patches of the final texture are colored from which partial texture. Figure 3.7 shows the optimized labeling in comparison to the best view selection. Note that bigger parts of the texture are now textured from the same input image, which naturally reduces the amount of visible seams. There are, however, still some luminosity differences at the patch boundaries, which we eliminate by blending the patches using Poisson

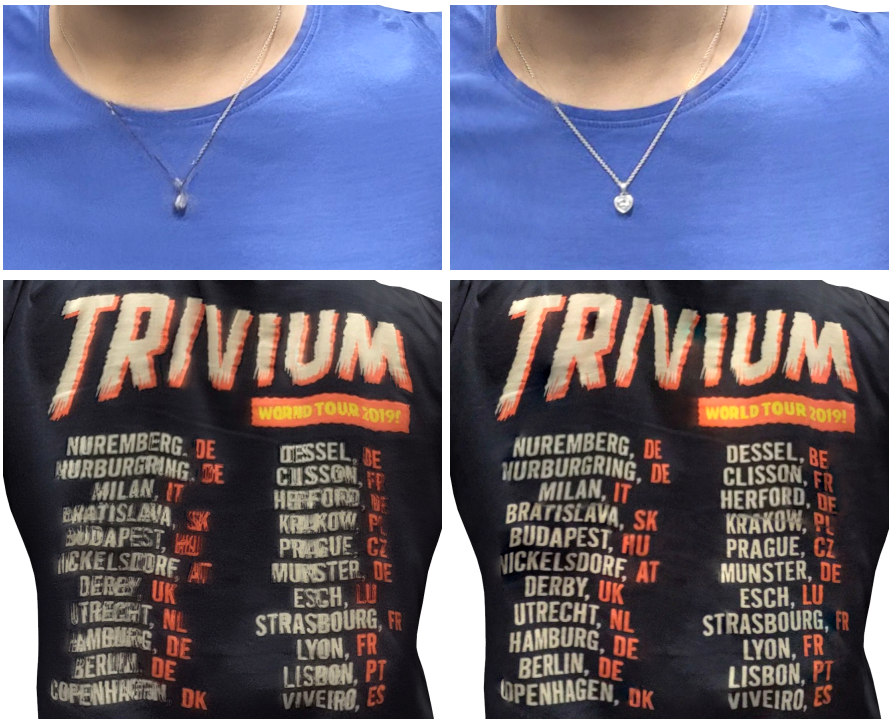

Figure 3.8: Texture generation of Agisoft Metashape (left) and our graph cut optimization (right), the latter yielding more detail on the necklace and the letters on the shirt.

Image Editing [PGB03]. Texture regions belonging to areas on the model that were not seen (e.g., the crotch or armpit region) are automatically filled by harmonic color interpolation.

This texture generation process is performed for both the full-body scan and the head scan. The head texture is then injected into the full-body texture using Poisson Image Editing in order to cope with illumination differences between the two scans. Since hands and eyes are typically not well scanned, their texture information is taken from the template texture and adapted to the scanned subject using histogram matching in CIELAB space, as proposed by Ichim et al. [IBP15].

As can be observed in Figure 3.8, the textures generated by our graph cut approach have more detail and are sharper compared to the textures generated by Agisoft Metashape.

3.3 RESULTS

Our avatar reconstruction takes about 20 minutes, measured on a desktop PC with 12 $\times$ 3.6 GHz Intel Xeon CPU and an Nvidia GTX 1080 Ti GPU, and consists of the following steps: capturing and transferring the videos (4 min), processing videos and generating point clouds (7 min), landmark detection and template fitting (2 min), and texture generation and merging (7 min). In

the following we provide quantitative comparisons with FGVH, which due to its extensive setup acts as an (approximate) ground truth, as well as qualitative comparisons to the monocular reconstructions of Alldieck et al. [AMX+18a; AMB+19].

In order to quantitatively compare our low-cost reconstruction with the multi-camera reconstruction of FGVH, we scanned and reconstructed 34 people with their method and ours. We had to discard the scan of one person, where the point cloud reconstruction failed due to dark clothing. For the remaining 33 scans we compare the reprojection error, which we compute by rendering for each input image the textured avatar from the corresponding camera location (see Figure 3.9) and computing the root-mean-square error (RMSE) over all rendered pixels in the CIELAB color space. Averaging the RMSE over all input images yields the reprojection error for one avatar reconstruction, which effectively measures reconstruction accuracy in both geometry and texture. Figure 3.10 shows the reprojection errors for all scanned subjects. Not surprisingly, the expensive camera rig of FGVH yields lower errors thanks to more accurate point clouds (cf. Figure 3.3). Although their RMSE ($M = 24.20$, $SD = 2.15$) is 20 % lower than ours ($M = 30.30$, $SD = 4.66$), our hardware costs (about \$600) are only 1 % of theirs (about \$60 000). To evaluate our graph cut based texture stitching, we compare it against Agisoft Metashape's texture generation. The results show that the textures produced by Agisoft Metashape yield a slightly lower RMSE ($M = 29.60$, $SD = 4.47$), but our graph cut optimization yields perceptually superior results (Figure 3.8).

As a purely geometric measure, we compute the modified Hausdorff distance [DJ94], defined by the maximum of the two average per-vertex distances between our reconstructions and the (approximate) ground truth given by

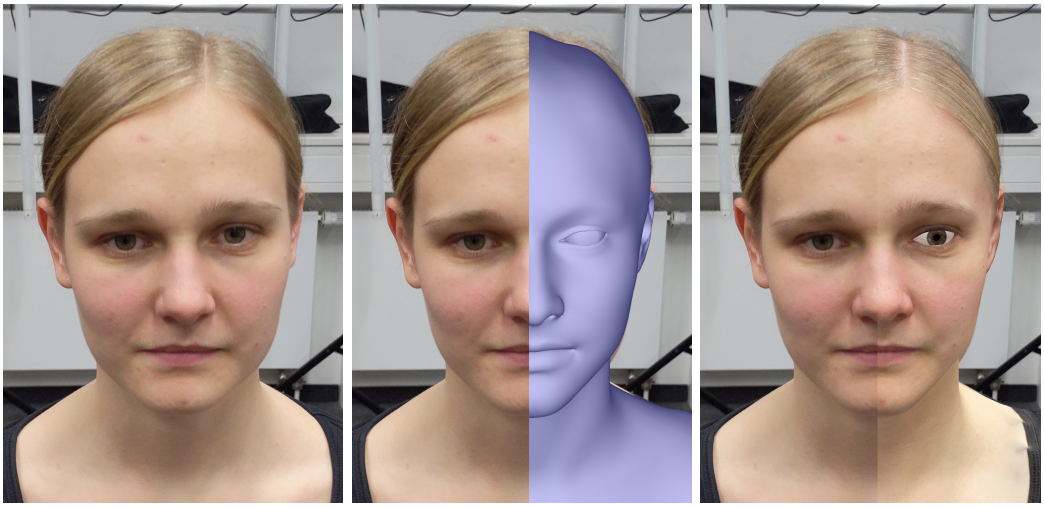

Figure 3.9: We evaluate the reprojection error in image space by rendering (without lighting) the reconstructed avatars from all input camera locations onto the input images.

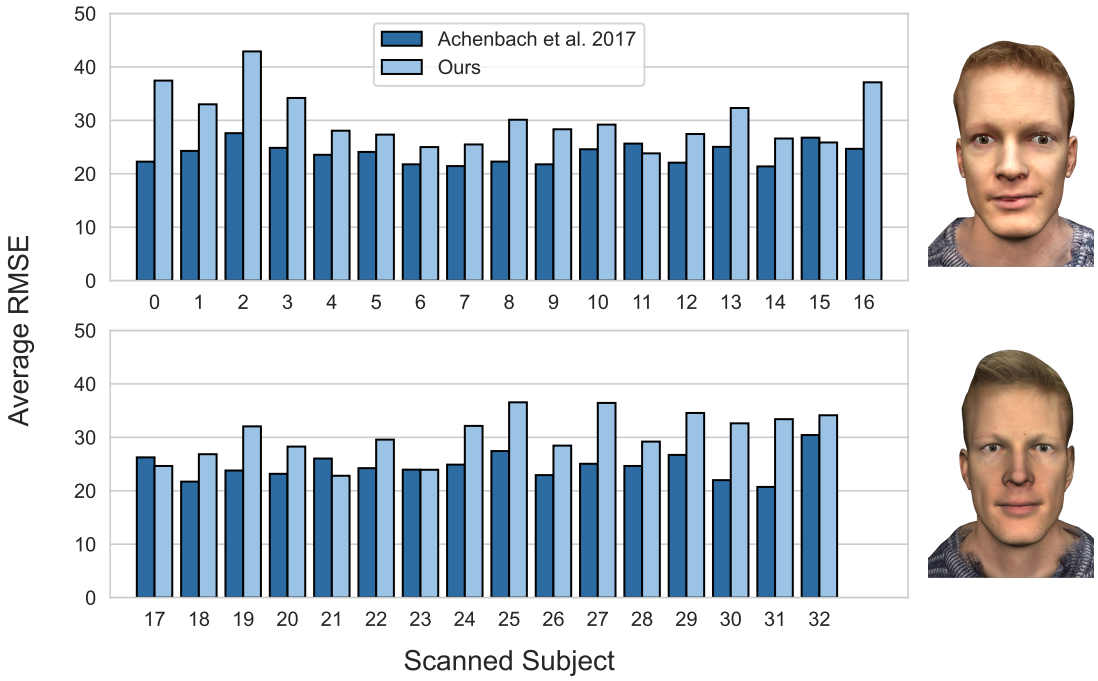

Figure 3.10: Root-mean-square reprojection errors of FGVH and our method over 33 reconstructed avatars. The close-ups on the right show Subject No. 2, for which our method (top) performs the worst compared to FGVH (bottom).

FGVH. Averaging this measurement over all reconstructions yields 7.1 mm, confirming that our avatars are quite accurate despite the low hardware requirements.

We qualitatively compare our method to the monocular avatar generation approaches of Alldieck and colleagues. The first method [AMX+18a] reconstructs avatars from a video of a person turning around 360° in A-pose (taking around 2 h). The second method [AMB+19] requires only eight images of this 360° movement and generates the texture using the stitching technique of [AMX+18b] (taking around 5 min). The input videos/images were taken using the same Google Pixel 3 to provide comparable input data. We used the original implementations provided by the authors, but doubled the default number of pose and shape estimation steps in Alldieck et al. [AMB+19] to achieve better results, as suggested to us by the authors. Figure 3.11 and Figure 3.12 compare avatars reconstructed with Alldieck et al. [AMX+18a], Alldieck et al. [AMB+19], FGVH, and our method, showing that our results are superior to Alldieck et al. and comparable to FGVH. Note that the avatars reconstructed by Alldieck et al. [AMX+18a; AMB+19] lack articulated hands, eyes, teeth, and facial blendshapes.

Our reconstructed avatars provide these facial animation controllers, as demonstrated in Figure 3.13 and the accompanying video. More results and comparisons, including dynamic skeletal and facial animations, can be found in the accompanying video at `https://www.youtube.com/watch?v=2D3-vn2yFVc`.

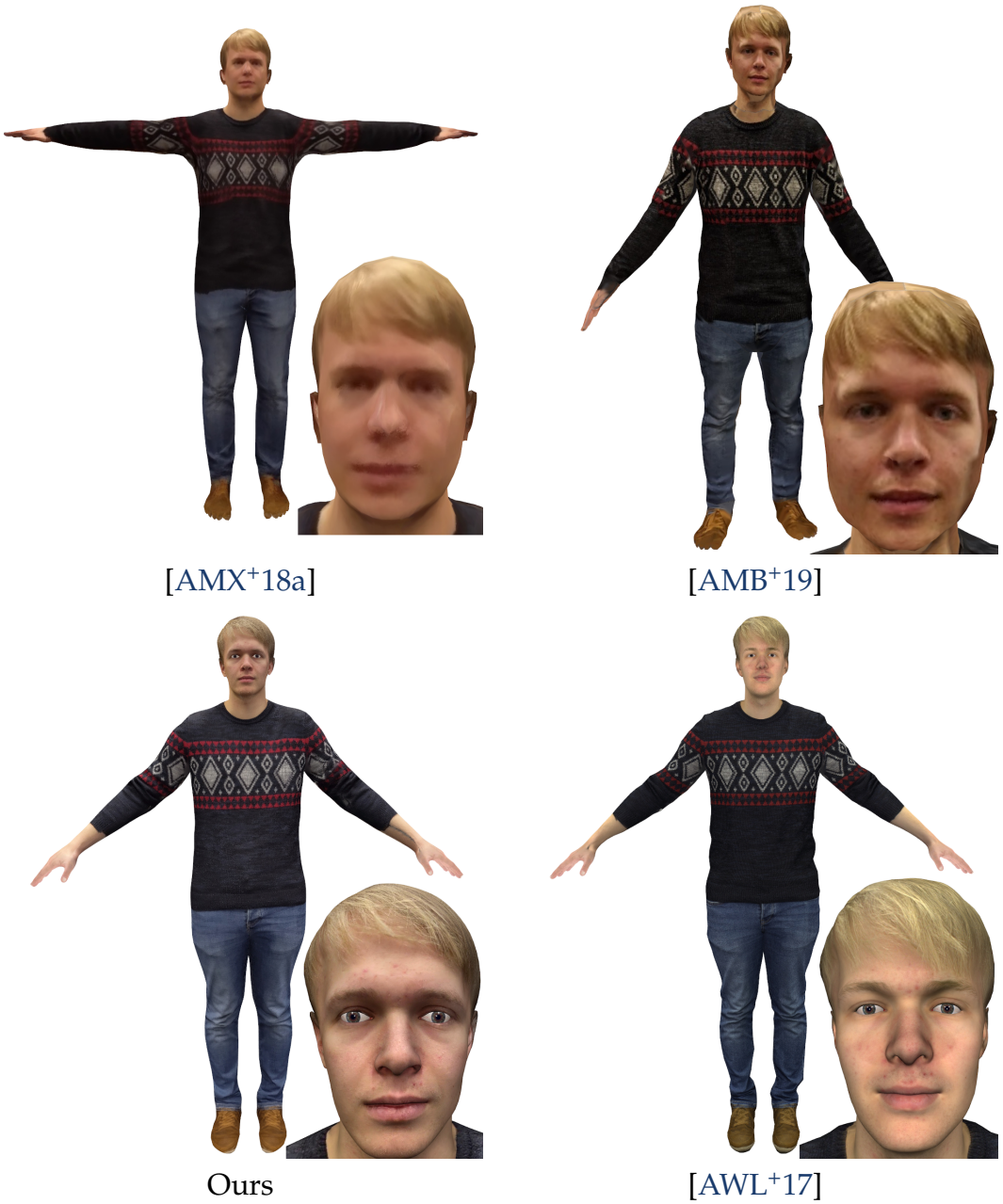

Figure 3.11: Avatars of the same person reconstructed from different methods: [AMX+18a], [AMB+19], ours and [AWL+17]. Note that our reconstruction improves on previous low-cost avatar generation pipelines in both geometry and texture.

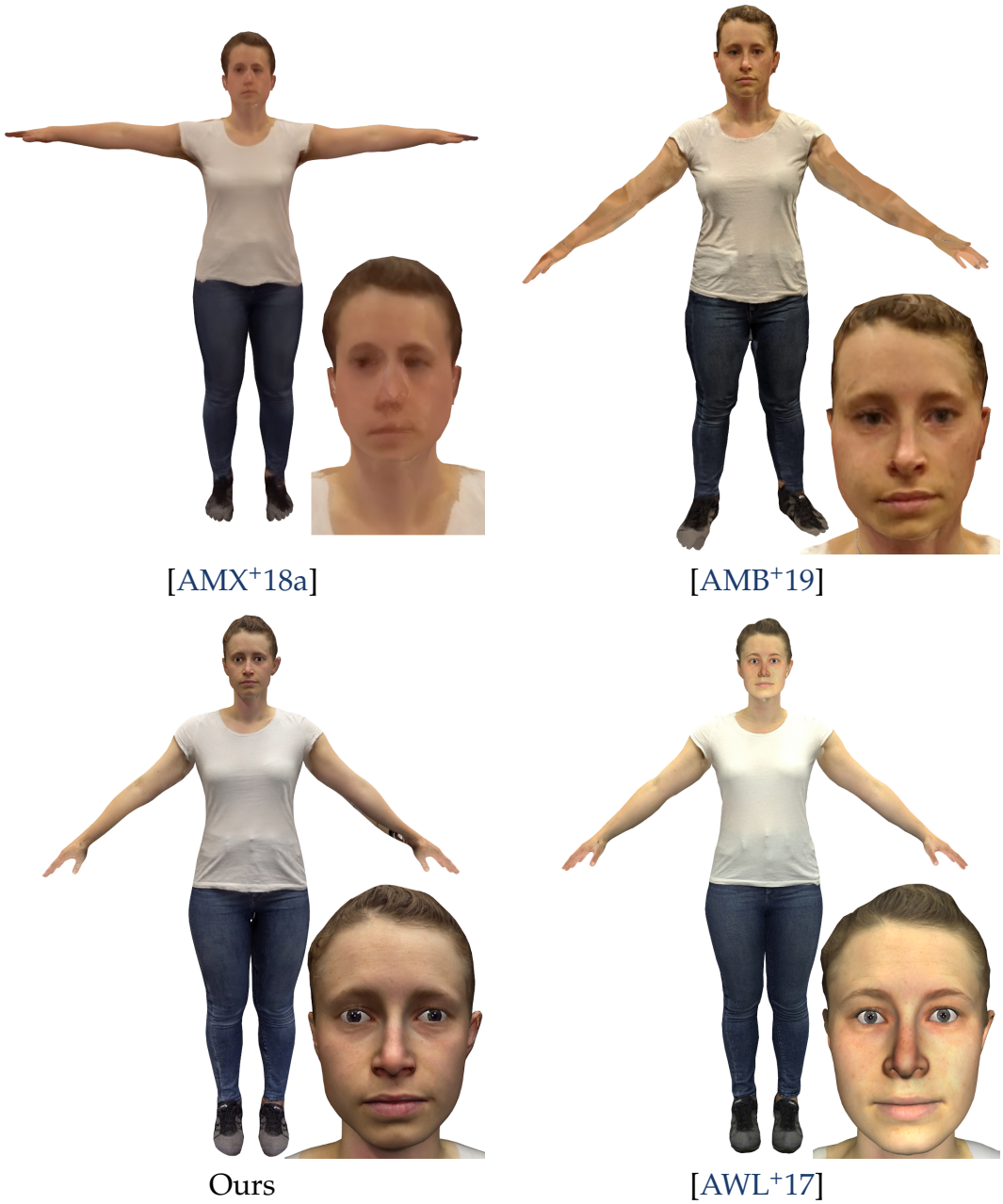

Figure 3.12: Avatars of the same person reconstructed from different methods: [AMX+18a], [AMB+19], ours and [AWL+17]. Note that our reconstruction improves on previous low-cost avatar generation pipelines in both geometry and texture.

Figure 3.13: Our avatars feature eyes, teeth, and facial blendshapes, and can thus be animated out-of-the-box, e.g., through real-time facial motion capturing.

3.4 SUMMARY AND LIMITATIONS

In this chapter we presented a fully automated pipeline for generating high-quality virtual humans from monocular videos, taking just about 20 minutes in total. The input for our pipeline consists of two video clips of the scanning subject taken with a consumer smartphone: one showing the full body and the other providing close-ups of head and face. We extract single frames using optical flow analysis and sharpness estimation, which are then passed to an off-the-shelve photogrammetry software, producing two dense point sets of the scanning subject. Pose and facial feature detectors are employed to automatically detect landmarks in the input images, which are projected onto the point sets. These landmarks guide the fitting of an animatable statistical virtual human template model, which reconstructs the geometry of the scanned person. A graph cut based texture stitching algorithm then produces a high-quality color texture for the resulting virtual human. Comparisons with both hardware-intensive and low-cost approaches show our virtual humans to be almost on par with the former while surpassing the latter. Our avatars are ready to be used in XR applications, as they allow skeletal and facial animation and are compatible with standard engines used in this field. This opens up the ability for the research community to work on high-quality avatars without extensive hardware setups.

Our method still has several limitations, as shown in Figure 3.14. First, the photogrammetry software cannot deal with very dark clothing. Second, the point cloud quality degrades for body parts exhibiting noticeable movement during the capturing process. This is especially true for the arms, leading

Figure 3.14: Limitations of our method. Dark clothing (left) and movement during the capture process (center) is challenging for the multi-view stereo reconstruction. This leads to errors in geometry and texture. The landmark back-projection fails for point clouds with missing data (right).

to a lower accuracy in geometry and texture reconstruction. Third, while the automatic 2D landmark detection worked robustly in all cases, the back-projection to 3D failed for four subjects due to missing data in the point cloud. In these cases, the user was asked to manually select the landmarks. Finally, glasses, hair, and accessories are challenging for all photogrammetric approaches, including ours.

For future work, we want to make our approach more robust to movements by segmenting the extracted video frames either into foreground/background or into semantic parts (e.g., torso, arms, legs, and head), which could potentially improve the quality of the multi-view stereo reconstruction. Furthermore, we plan to exploit the capabilities of smartphone APIs to build a designated application for controlling the capture process and gaining access to the intrinsic camera calibration. Another interesting direction is to scan challenging areas like the arms separately, i.e., to divide the capture process into more than two videos. We recorded the videos for the avatar reconstructions in a controlled indoor environment, where we did not have to deal with hard shadows, unstable lighting conditions or moving objects in the background. Extending the method to also yield high-quality results in these challenging outdoor settings is another direction for future work.

4 COMPARING THE EFFECTS OF TWO AVATAR RECONSTRUCTION METHODS

We have now seen two different reconstruction methods for realistic personalized avatars. The previous chapter presented a low-cost approach for reconstructing virtual humans from two smartphone videos, which drastically lowers the hardware requirements of realistic virtual human reconstruction pipelines, as these typically employ elaborate photogrammetry rigs consisting of multiple DSLR cameras. However, when employing realistic virtual humans in XR environments, and especially in the context of XR therapy, it is important to assess the user acceptance of the virtual humans resulting from the chosen reconstruction method. So far, we have only looked at objective measures such as reprojection error or differences in the resulting geometry when comparing the *low-cost* method to a *high-cost* method such as FGVH [AWL+17]. To investigate, if virtual humans resulting from low-cost methods really present a viable alternative, we have to examine, how people rate the *appearance* of the virtual humans when compared to those resulting from high-cost methods.

The appearance of virtual humans has notable effects on ourselves and our interaction partners (see, e.g., Praetorius and Görlich [PG20] and Ratan et al. [RBL+20]). Previous work found realistic self-avatars used for embodiment to be superior to abstract self-avatars in terms of user acceptance [LRG+17]. Others found personalized realistic-looking self-avatars to be even more superior, enhancing the illusion of virtual body ownership as well as the feeling of presence [WGR+18]. Comparable interesting effects occur for other-avatars (the virtual representations of other users) and virtual agents (embodied entities controlled by artificial intelligence). For example, the appearance of virtual others impacts their perceived trustworthiness [MBB12; SLD+19], approachability [FM21], affinity [SLD+19], and co-presence [BSH+05]. Given the continuous technological advances in the reconstruction of virtual humans, research on their realism is still ongoing [SGH+20]. For example, there is still debate about whether realistic-looking virtual humans are prone to facilitate the uncanny valley effect (e.g., [TG09; KFM+15; WLR15; LLL15b]), which describes the phenomenon that close-to-real looking artificial humans sometimes strike as eerie [MMK12; HM10; HM17].

Today, multiple reconstruction approaches for realistic, lifelike virtual humans exist. They significantly vary in terms of the degree of achievable realism, the technical complexities, and finally, the overall reconstruction costs involved. So far, rather complex and expensive multi-camera rigs achieve the highest quality by using high-quality image sensors (e.g., [FRS17; AWL+17; GLD+19]). However, approaches for reconstructing virtual humans from input data produced by more affordable consumer hardware become more popular

and elaborate. They only require e.g., a single image [APT+19; AZS22; LZX+23; SAK+24], a single smartphone video [IBP15; AMX+18b; AMB+19], or multiple smartphone videos (Chapter 3). Most of these low-cost approaches share the vision to make it possible for everyone to generate a digital alter ego quickly and inexpensively without a complex hardware setup. Such approaches would drastically leverage the possibilities for research, industry, and overall users of embodiment systems. By only utilizing consumer-level hardware, e.g., a $600 smartphone instead of a camera rig costing tens of thousands of dollars, smaller development teams can afford life-like virtual humans, for example, in their games and social VR applications, and users would benefit from a much more personalized experience using their realistic look-alike avatars.

To complement the objective evaluation in Chapter 3, this chapter investigates the subjective perception of low-cost virtual humans. Specifically, we will address the following research questions:

RQ_1 Can low-cost approaches for generating realistic virtual humans keep up with high-cost solutions regarding their perception by users in embodied VR?

RQ_2 Is the quality difference more noticeable for the own virtual body compared to the virtual body of others?

For the investigation of our research questions, we conducted a user study to compare our low- and a high-cost 3D reconstruction approaches for virtual humans. Both produce (i) realistically looking and (ii) ready-to-animate virtual humans (iii) in a time frame that is compliant with common study procedures, i.e., within minutes. The input for the high-cost method is given by a photogrammetry rig including 94 DSLR cameras located at University of Würzburg. The resulting images are processed by the *Fast Generation of Virtual Humans* (FGVH) method [AWL+17]. The second method uses a simple smartphone camera to capture two videos of a person, which are then processed by the *Realistic Virtual Humans From Smartphone Videos* method (Chapter 3). We scanned participants by both methods. Then they embodied the resulting self-avatars in an immersive virtual environment and encountered pre-scanned virtual others of both reconstruction methods. We report on the sense of embodiment for the self-avatars and the perceived similarity, uncanniness, and preference for both the self-avatars and the virtual others. We further look at objective differences between the two methods and investigate whether these differences are more noteworthy for the self-avatar than someone else's body. Our results indicate that the avatars from the low-cost approach are perceived similarly to the avatars from the high-cost approach. This is remarkable since the quality differed significantly on an objective level. The perceived change of the own body was more significant for the low-cost avatars than for the high-cost avatars. The quality differences were more noticeable for the own than for other virtual bodies.

Individual Contribution *My main contribution is the integration of the two virtual human reconstruction methods at the lab of our colleagues at the University of Würzburg, where the user study was conducted. I additionally implemented the objective comparison between the resulting virtual humans. The user study was mainly designed by Andrea Bartl, who also performed the user study and the corresponding statistical evaluation at the University of Würzburg. Andrea Bartl also designed the VR environment, while Erik Wolf provided the avatar embodiment framework, allowing the participants to observe their motions in a virtual mirror inside the VR environment.*

Corresponding Publication *This chapter is based on the following publication*:

> Andrea Bartl, Stephan Wenninger, Erik Wolf, Mario Botsch, and Marc Erich Latoschik. "Affordable but not Cheap: A Case Study of the Effects of Two 3D-Reconstruction Methods of Virtual Humans". *Frontiers in Virtual Reality* 2 (2021).

4.1 RELATED WORK

4.1.1 Perception of Virtual Humans

Virtual humans are part of a great variety of applications. They serve as avatars (representations of real people in digital worlds), virtual trainers, assistants, companions, game characters, and many more. Often, developers strive to make them as realistic as possible. The perceived realism of virtual humans depends on their appearance and their behavior [MT05; SS15]. While we acknowledge the importance of behavioral realism, our work focuses on the appearance of virtual humans. Our appearance and the appearance of others in a virtual environment have notable effects on our perception [HH16; FM21].

The Own Virtual Appearance

When it comes to using virtual humans as avatars, i.e., digital representations of persons in a virtual world, the Proteus effect [YB06; YB07] is a prominent research topic. It describes the phenomenon that the avatar appearance can influence users' attitudes and behavior based on stereotypical beliefs. For example, in previous research, participants who embodied a child associated more child-like attributes with themselves [BGS13], attractive avatars increased intimacy [YB07], strong-looking avatars improved physical performance [KKS+20], and taller avatars led to more confidence [YB07]. Wolf et al. [WMD+21] recently showed that the embodiment of an avatar can potentially alter its body weight perception relating to the user's body weight.

For many VR applications, the Proteus effect is desirable. Users can slip into a body with different size, shape, look, age or gender, enabling experiences one could not easily create in real life. Exploiting this effect potentially even helps to reduce negative attitudes, such as racial bias [PSA+13; BHS16], negative stereotypical beliefs about older people [YB06], or misconceptions of the own body image [DWW+19]. It could also promote positive attitudes and behavior, e.g., motivation to exercise [PKA16]. However, what if the use case requires the users just to be themselves? For example, experiments often assume a user's unbiased evaluation without taking the potential bias of the virtual body into account. Other exemplary scenarios might focus on a person's actual body shape, e.g., virtual try-on rooms, therapy applications, or specific physical training scenarios that prepare people for real-life situations.

In previous work, the self-similarity of the avatar influenced the users' perception in the virtual environment. Personalized realistic-looking avatars enhanced the illusion of body ownership and the feeling of presence in first-person [WGR+18] and third-person [GCH+19] immersive VR. Self-similarity enhanced negative attitude changes when embodying a self-similar but sexualized avatar [FBT13] and impacted body weight perception [TGM+18]. Having a self-similar body in VR promoted creativity [RLE17] and increased presence and social anxiety levels in VR [AKB14]. In a fitness application with a full-body virtual mirror, having an avatar that was self-similar in terms of gender enhanced the illusion of body ownership and increased performance compared to a not self-similar one [LLL15a]. Especially in social VR applications, people very deliberately choose to look or not look like they do in real life [FM21]. Realistic avatar representations used for embodiment have been superior to abstract avatar representations in user acceptance [LRG+17]. Nevertheless, the role of realism in avatars is still in debate. Other work could not reproduce this superiority [LWB+15] and even found realistic avatars to be less accepted than abstract representations [LLL15b]. The context of the experience might be an important factor when it comes to the influence of the own avatar's appearance. The impact seems to be less significant in game-like or overall more stressful scenarios that strongly engage the user in a superordinate task that only marginally focuses on the body (e.g., [LLL15b; LLL15a]). But it might be of greater importance for social scenarios (e.g., [AKB14; FM21]) or experiences where the user and his body is the center of attention.

The Virtual Appearance of Others

In virtual environments, users can also encounter virtual humans as computer-controlled virtual agents or embodied other, real users. Previous work showed that a virtual agent's appearance influenced co-presence [BSH+05]. Nelson et al. [NMJ+20] found that virtual agents' appearance influences users' movement speed and their interpersonal distance to the agents. In social VR applications,

another user's avatar's appearance influences whether and how others approach this user [FM21]. A realistic appearance of a virtual agent impacts its perceived appeal and friendliness [MBB12]. Other previous work looked at the impact of realistic-looking interaction partners on perceived trustworthiness [MBB12; JKK17; SLD+19]. Seymour et al. [SLD+19] found a preference for realistic virtual agents, which also increased the users' place illusion [ZMM19]. Zibrek et al. [ZKM18] investigated the impact of virtual agents' realism in virtual reality games and found complex interactions between the virtual agents' personality and appearance.

A recurring debate about the realism of virtual characters is the uncanny valley effect. Initially described by Mori et al. [MMK12] for human-robot interactions in the 1970s and later transferred to virtual characters, the uncanny valley effect refers to the phenomenon that close-to-real looking artificial humans sometimes strike as eerie. The original work sets human-likeness in correlation with familiarity. It proposes a drop in familiarity when the artificial character looks close to but not entirely like a human. Research on this effect is not at all consensus. Some argue that the uncanny valley effect might only occur under specific circumstances that are yet to be defined [KFM+15]. Some explain that the phenomenon is a wall rather than a valley since people adapt to the technical advances and therefore, the uncanny valley is untraversable [TG09]. Others argue that the key to overcoming the uncanny valley with realistic-looking characters lies in their behavior [SRK17; SLD+19]. And finally, some question the existence of the uncanny valley effect as a whole [WLR15].

In summary, research on the realism of virtual humans has been controversial for decades and is still ongoing. However, it is especially relevant today as methods for creating virtual humans are improving drastically along with the overall evolution of technology, creating new and reviving old research questions [SGH+20].

4.1.2 Creation Methods for Virtual Humans

For the 3D reconstruction of a person, various techniques exist that differ in terms of the degree of achievable realism, the technical complexities, and the overall reconstruction costs involved. As discussed in Section 3.1, hardware requirements for current virtual human reconstruction methods range from immensely involved light stage systems [GLD+19] to single-shot multi-camera photogrammetry rigs [FRS17; AWL+17] to a single RGB(-D) camera [AMX+18a; AMX+18b; AMB+19; LMR+15]. In Section 3.3, we compared the reconstruction fidelity between our low-cost method and the high-cost method of Achenbach et al. [AWL+17] by computing the geometric difference between the resulting avatars and the reprojection error resulting from rendering the textured avatars

back onto their respective input images. The evaluation shows that there still is a difference in both measures but that their low-cost approach can almost reach the same fidelity as the high-cost approach. However, this evaluation only covers purely objective measures. We did not address how the still existing differences affect users' perception of the virtual humans.

Based on the presented literature, we specify our research goal: We build on the purely objective comparison of Section 3.3 and focus on the user perception of the resulting virtual humans. In a user study, we compare a high-cost method to create virtual humans to a low-cost method. The methods differ in their hardware requirements (high-cost vs. low-cost), the input material (multiple images vs. two smartphone videos), and software parameters for tailoring the approach to the specific input material. We investigate whether the differences in the quality of low- and high-cost reconstructions of virtual humans produce differences in the users' perception. The evaluation includes one's reconstructed self-avatars and virtual others, here, computer-controlled reconstructions of other real persons. We compare the users' perception in terms of the similarity of the virtual humans to the original, the sense of embodiment (only for the self-avatars), their uncanniness, and the overall preference for one of the approaches. We also investigate if differences between the high- and low-cost virtual humans are more noticeable for one's self-avatar than for virtual others. Finally, we compare the low- and high-cost virtual humans using objective measures, i.e., the reprojection error and the geometrical error.

4.2 STUDY

To investigate our research questions, we designed a user study that focuses on the perception of the low-cost and high-cost virtual humans. Regarding RQ_1, we compared the subjectively perceived quality of two 3D reconstruction methods for realistic virtual humans. In particular, we compared one method using a high-cost photogrammetry rig containing 94 DSLR cameras with a low-cost method processing two smartphone videos. For this purpose, we scanned participants twice and created one personalized self-avatar with each generation method. In a virtual environment, participants embodied both self-avatars and observed themselves in virtual mirrors. They also encountered and evaluated other virtual humans originating from both scan processes, observing them on virtual monitors. The independent variable for RQ_1 was the reconstruction method (low-cost vs. high-cost) that we investigated for self-avatars and virtual others separately.

To answer RQ_2, participants could adjust the distance between themselves and the mirrors or monitors. The task was to set the distance at which they could no longer tell that one version was better than the other. We assumed

that there would be a difference in the distance that participants set for the mirrors (self) compared to the distance they set for the monitors (other) if the quality discrepancy between methods was more noticeable for one's own or for another virtual body (RQ_2). Therefore, the independent variable for RQ_2 was the virtual human (self vs. other). The study followed a repeated-measures design.

4.2.1 Virtual Humans

High-Cost and Low-Cost Method

Figure 4.1 displays both the high-cost and the low-cost scan processes, including example results for a sample participant. A photogrammetry rig that contains 94 DSLR cameras generates the input for the *high-cost* avatars. In contrast to Achenbach et al. [AWL+17], we did not use a separate face scanner. Instead, 10 of the 94 cameras of the body scanner are zoomed in on the scan subject's face, therefore, capturing more detail in this area. The scanner includes four studio lights with diffuser balls (see Figure 4.1, first row, first picture). For generating the avatars from these images, we follow the method of Achenbach et al. [AWL+17], who combine photogrammetric reconstruction with a template fitting approach. The set of images produced by the camera rig is processed with the commercial software Agisoft Metashape [Agi24], yielding dense point clouds of the scanned subjects. The subsequent template fitting process (detailed in Section 2.2.2) deforms a statistical animatable human template model to match the point cloud data. The employed template model provided by Autodesk Character Generator [Aut24] is fully rigged and also equipped with a set of facial blendshapes, thus making the resulting avatars ready for full-body and facial animation. The pipeline for generating the high-cost avatars operated on a PC containing an Intel Core i7-7700k, a GeForce GTX 1080 Ti, and 4×16 GB DDR4 RAM. The generation took approximately ten minutes per avatar.

To provide the video input for the *low-cost* avatar method, we used a Google Pixel 5 smartphone. We used the camera application OpenCamera because it allows for a non-automatic white balance and exposure. The smartphone captured the videos with 4 k resolution (3840×2160) and 30 fps. We filmed in a room with covered windows, using the installed ceiling lights and eight additional area lights placed on the floor and on tripods around the participant (see Figure 4.1, second row, first picture). The additional lighting is not necessarily required for generating the low-cost avatars. However, it was added to brighten up the resulting low-cost avatars decreasing the brightness difference between the low-cost and the high-cost variant. After taking two videos of each subject, one capturing the whole body and the other capturing the head in a close-up fashion, the videos are processed with the method

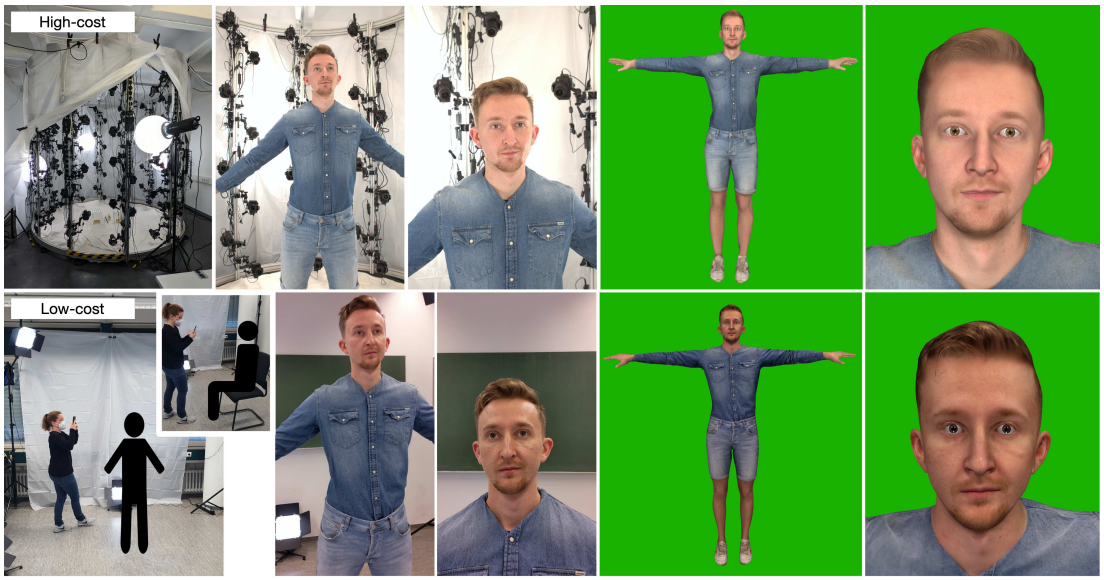

Figure 4.1: The high-cost (top) and the low-cost (bottom) scan process.

described in Chapter 3, which also uses photogrammetric reconstruction and template fitting with the same template model. The pipeline for generating the low-cost avatars operated on a PC containing an Intel Core i7-7820x, a GeForce GTX 1080 Ti, and 6×16 GB DDR4 RAM. The generation took approximately twenty minutes per avatar.

Self-Avatar Animation

The generated low- and high-cost self-avatars were both imported to our Unity application. For the avatar animation, we oriented towards the system architecture introduced by Wolf et al. [WDM+20] and adapted their implementation. During the experiment, the two imported avatars were simultaneously animated in real-time according to the users' movements by using a set of HTC Vive Trackers in conjunction with the VR headset and the VR controllers (see Section 4.2.2 for details about the VR setup). In order to animate the avatars based on the tracking data, we used the calibrated tracking targets of the head, left hand, right hand, pelvis, left foot, and right foot to drive an inverse kinematics (IK) animation approach realized by the Unity plugin FinalIK [Roo24] (version 2.0).

Virtual Others

For the virtual others, we scanned one male and one female person. Figure 4.2 displays both versions of the virtual other. Male participants observed and evaluated the male other, while female participants evaluated the female other. Both virtual others wore identical gray t-shirts and blue jeans. We recruited

Figure 4.2: The female (left) and male (right) virtual others. The left virtual monitor of each pair displays the high-cost version; the right virtual monitor displays the low-cost version.

two persons who do not represent extremes in terms of their appearance. The male other was 1.72 m tall, while the female other was 1.66 m tall. Both persons stated that they do not know any of the students belonging to the study's participant pool. The virtual others were animated using a pre-recorded idle animation. The animation showed a basic idle standing animation including small movements, e.g., slightly moving from one foot to the other. We also added random eye movements and blinking using an existing asset of the Unity Asset Store [Uni24] to increase the virtual others' realism.

4.2.2 Virtual Reality System

We implemented our study system using the game engine Unity [Uni19] (version 2019.4.15f1 LTS) running on Windows 10. The VR hardware explained in the following was integrated with SteamVR [Val24a] (version 1.16.10) and its corresponding Unity plugin (version 2.6.1). As high-immersive VR display system, we used a Valve Index HMD [Val24b], providing the user a resolution of 1440×1600 pixels per eye with a total field of view of 120° running on a refresh rate of 90 Hz. For capturing the user's motions, participants held the two Valve Index controllers in their hands, wore one HTC Vive Tracker 2.0 on a belt around the hips, and one fixed on each shoe's upper side with a Velcro strap. Three SteamVR 2.0 base stations braced the spacious tracking area. The system ran on a high-end, VR-capable PC composed of an Intel Core i7-9700K, an Nvidia RTX2080 Super, and 16 GB RAM.

We determined the motion-to-photon latency of our system by frame-counting [HLP+00]. For this purpose, the graphics card's video signal output was split into two signals using the Aten VanCryst VS192 display port splitter. One signal led to the HMD and the other to the low-latency gaming monitor ASUS ROG SWIFT PG43UQ. A high-speed camera of an iPhone 8 recorded the user's motions and the corresponding reactions on the monitor screen at

240 fps. Counting the recorded frames between the user's motions and the corresponding reactions on the screen, we determined the latency for the HMD and limb movements separately. For HMD and limb movements, we repeated the measurements ten times each. The motion-to-photon latency for the HMD averaged 14.56 ms ($SD = 2.94$) and therefore matched the refresh rate of the HMD closely. The motion-to-photon latency for the limb movements averaged 42.85 ms ($SD = 5.20$) and was considered low enough for real-time avatar animation [WSH+16].

Virtual Environment and Task

The virtual environment consisted of one large virtual room. In the room, two virtual mirrors were mounted on a track system to allow for a direct comparison of the self-avatars and to induce the feeling of embodiment by visuomotor coherence [SSS+10; LW22a]. We told participants that they would see two different mirrors before they saw their self-avatars. The track system was supposed to increase coherence with the users' expectations, making the scenario more plausible [LW22a]. For the evaluation of the virtual other, the mirrors were exchanged with similar-looking, portal-like, virtual monitors (see Figure 4.2). A stencil buffer masks the area inside the monitor to make the virtual others visible only in this area. This setup preserved a stereoscopic view and ensured a spatial distance to the participants. To help the feeling that the virtual others are in a different place and that the monitors were no mirrors, we added textures to the surrounding walls and floor that were different from the main room. Participants received the audio information that they would see two different broadcasts of another person on these two monitors. This information served the purpose of making the scenario more plausible and less intimidating than directly encountering two similar-looking versions of a person in a virtual room that would not react in any way to the user [Sla09; LW22a]. Study participants would encounter these virtual others for the first time. To enable them to evaluate the virtual humans' similarity to the real person, they needed to see reference material first. We displayed a photo of the real person for 10 s before the virtual other appeared on the monitor(s) and asked participants to memorize it. For the self-avatar similarity assessment, we did not show a photo of the person. Instead, we relied on familiarity with the person's own appearance. Mirrors and monitors turned automatically according to the study phases. Figure 4.3 shows the virtual environment throughout the phases of the experiment.

The first two phases of the experiment concentrated on the perception of the virtual humans and the participants' preferences. In Phase 1, participants saw and evaluated the high-cost and the low-cost virtual humans one after another. In Phase 2, they saw both at the same time next to each other. Then they evaluated the left one first. After that, they again saw both at the same

Figure 4.3: The three phases of the VR exposure and the VR questionnaire system.

time and consecutively evaluated the right one. The photo of the real other person was displayed for 10 s before every virtual other observation phase. For a controlled exposure, participants received audio instructions on where to look and what movements to perform. Table 4.1 lists all instructions and the observation duration. In Phase 2, before and after the instructions, the participants got the information which virtual human they will have to rate (left or right). During the self-avatar observation, the participant always embodied the self-avatar to be rated after the observation. Analogously, participants embodied the high-cost self-avatar when viewing the high-cost virtual other and the low-cost self-avatar when viewing the low-cost version of the virtual other.

In Phase 3, participants could adjust the distance between themselves and the mirrors or monitors. The task was to increase the distance until they could no longer tell which virtual human was better. Participants could move the mirrors and monitors using the controllers' touchpads. One controller increased the distance; one decreased it. When moved back and forth, mirrors and monitors automatically rotated on the track system to always face the user. This ensured that the reflections and the virtual others were always visible to the participants.

VR Questionnaire

Participants evaluated the virtual humans directly in VR. The right image in Figure 4.3 shows the VR questionnaires from a third-person perspective. Following the guidelines of Alexandrovsky et al. [APB+20], our VR questionnaire was world-anchored and participants used a controller to operate the questionnaire using a laser pointer. A virtual display presented the VR questionnaire in the virtual environment. It was positioned on the wall left to the user. The integration into the scene's context was supposed to make it more diegetic and thus more plausible [SPR+16]. The virtual display was approximately 1.2 m high and 2 m wide. The user stood approximately 1.5 m away from the

No.	Instructions Phase 1 - Self	Duration
1	Look at your reflection in the mirror. Please remain standing on the marker. You may move your arms and legs freely.	10 s
2	Look at your head in the mirror.	5 s
3	Swing your arms back and forth while looking at your torso.	5 s
4	Let your arms hang relaxed and slowly shift your weight from your left leg to your right leg and back again. Repeat this a few times while looking at your lower body.	5 s
5	Stand relaxed. Wave your dominant hand at your reflection while observing yourself in the mirror.	5 s
No.	**Instructions Phase 1 - Other**	**Duration**
1	Look at the person.	5 s
2	Look at the head of the person.	5 s
3	Look at the torso of the person.	5 s
4	Look at the lower body of the person.	5 s
5	Now look at the whole person again.	5 s

Table 4.1: Instructions that participants received in Phase 1 while they had to inspect the virtual human in the mirror or monitor. In Phase 2, when participants saw both self-avatars or both virtual others at the same time, they received each instruction twice; first for the left mirror, then for the right mirror, e.g., "Look at your head in the left mirror." — 5 s duration — "Look at your head in the right mirror." — 5 s duration.

display. This size and distance allowed the participants to read the questions comfortably without having to move the head. To keep the exposure time with each self-avatar the same for every participant, their embodiment while answering the questions only consisted of visible controllers.

4.2.3 Measurements

Before and after the experiment, participants answered questionnaires on a computer in the experiment room. During the experimental phases, participants answered VR questionnaires. We used German translations of all questions and questionnaires.

Perception of the Virtual Humans

In Phase 1 and 2, participants rated their self-avatar regarding the perceived similarity, their sense of embodiment [KGS12; RL20], and possible uncanny

valley effects [HMP08; HM17]. The questions regarding the virtual other were the same, only omitting the embodiment questions since they did not apply in this condition.

Similarity: For the measurement of perceived similarity, we adapted the item used by Waltemate et al. [WGR+18]. Participants rated their agreement to the statement *"The virtual body looked like me/the person on the image"* on a scale ranging from 1 (*I do not agree at all*) to 7 (*I fully agree*).

Embodiment: For measuring the sense of embodiment, we used the Virtual Embodiment Questionnaire [RLL+17; RL20]. It consists of three subscales with four items each: Body Ownership, Agency, and Change. Participants rate their agreement to each of the twelve statements on a scale ranging from 1 (*I do not agree at all*) to 7 (*I fully agree*). High values indicate a high sense of embodiment.

Uncanny Valley: Regarding the uncanny valley effect, we built three items based on the original uncanny valley questionnaire's subscales of Ho et al. [HMP08] and Ho and MacDorman [HM17]. Participants rated their agreement on the three statements: *"The virtual body looked human."*, *"The virtual body looked eerie."*, *"The virtual body looked beautiful."*. Participants rated their agreement to all statements on a scale ranging from 1 (*I do not agree at all*) to 7 (*I fully agree*).

Preference: At the end of Phase 2, we directly asked participants which self-avatar/virtual other they preferred using the item: *"Which virtual body was better?"* with the answer options *left* or *right*. We asked if they found the left virtual body to be *much worse*, *worse*, *neither worse nor better*, *better* or *much better* than the right virtual body, with a second item. Note that due to the randomization left and right meant different versions for different participants. This was re-coded in the statistical analysis later.

Qualitative Feedback: Between the scan and the experiment, we asked participants how they perceived the two scan processes overall. After the whole experiment, we asked them to write down reasons for their preference regarding the version of the self-avatar and the virtual other.

Distance

In Phase 3, we asked participants to increase the distance between the virtual bodies and themselves until they no longer can say if one of the virtual humans is better than the other. We measured the distance in meters between the HMD and the two mirrors (or monitors in the other-condition). For the self-condition, when the participant moved the mirrors away, the reflection logically also moved away. Therefore, we multiplied the measurement by two to get the actual distance between the participant and the self-avatars. For the other-condition, we added the distance between the virtual other and

the monitor frame (0.5 m) to the distance the participant set. The maximum possible distance between the participant and the monitors was 18 m.

Objective Measures

For comparing the high-cost and the low-cost scans on an objective level, we calculate (i) the reprojection error and (ii) the modified Hausdorff distance [DJ94] between our two reconstruction methods, thereby following the objective evaluation presented in Section 3.3. The reprojection error is computed by projecting (i.e., rendering without lighting) the textured avatar onto each of the cameras as estimated during the avatar generation process. We then calculate the average root-mean-square error (RMSE) between the rendered images and the actual input images in CIELAB color space, giving us a way to measure the reconstruction methods' faithfulness objectively. The modified Hausdorff distance measures the difference in shape between the two reconstruction methods on a purely geometric level.

Control Measures

Before and after the experiment, we used the simulator sickness questionnaire [KLB+93] to measure virtual reality sickness as described by Kim et al. [KPC+18]. The questionnaire includes 16 symptoms of simulator sickness. The participants rated how much they experienced each symptom on a scale ranging from 0 (*none*) to 3 (*severe*). We added three items to check for disturbances in the perceived place and plausibility illusion [Sla09]. At the beginning of each VR question phase, we asked participants how present they felt in the virtual environment. For this, similar to Bouchard et al. [BSR+08] and Waltemate et al. [WGR+18], we used one item, namely *"How present do you feel in the virtual environment right now?"* with a scale ranging from 1 (*not at all*) to 7 (*completely*). At the end of each questionnaire phase, we added two items focusing on the overall plausibility: *"The environment made sense."* and *"The virtual body matched the virtual environment."*. These items served the purpose of measuring the environment's plausibility by checking for any unwanted incoherence in the experience caused by the environment [SBW17; LW22a]. Participants rated their agreement on scales ranging from 1 (*I do not agree at all*) to 7 (*I fully agree*).

Demographics and User Traits

Participants answered a demographic questionnaire including items for age, gender, educational attainment, occupation, language familiarity, problems with telling left from right, visual and hearing impairments, computer game experience, and virtual reality experience. We also asked them if they have been

scanned before. Before the experiment, we measured the participants' height and asked them which of their hands is their dominant one. Additionally, we measured participants' tendency to experience presence using the Immersive Tendency Questionnaire [WS98]. Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

4.2.4 Procedure

Figure 4.4 shows the experimental procedure. Each session took around 90 minutes, divided into 30 minute blocks of (i) scan preparation, performing the two scans, and reconstructing the respective avatars, (ii) answering questionnaires before and after the experiment as well as putting on the VR equipment, and (iii) the VR exposure itself (Phases 1–3). At the beginning of Phase 0, participants received a written introduction and signed consent forms for being scanned, participating in the study, and for COVID-19 related regulations. The video scan to create the low-cost avatars was made first to optimize the schedule. After the two scans, the participant filled in pre-questionnaires while the avatars were generated. Then, the experimenter

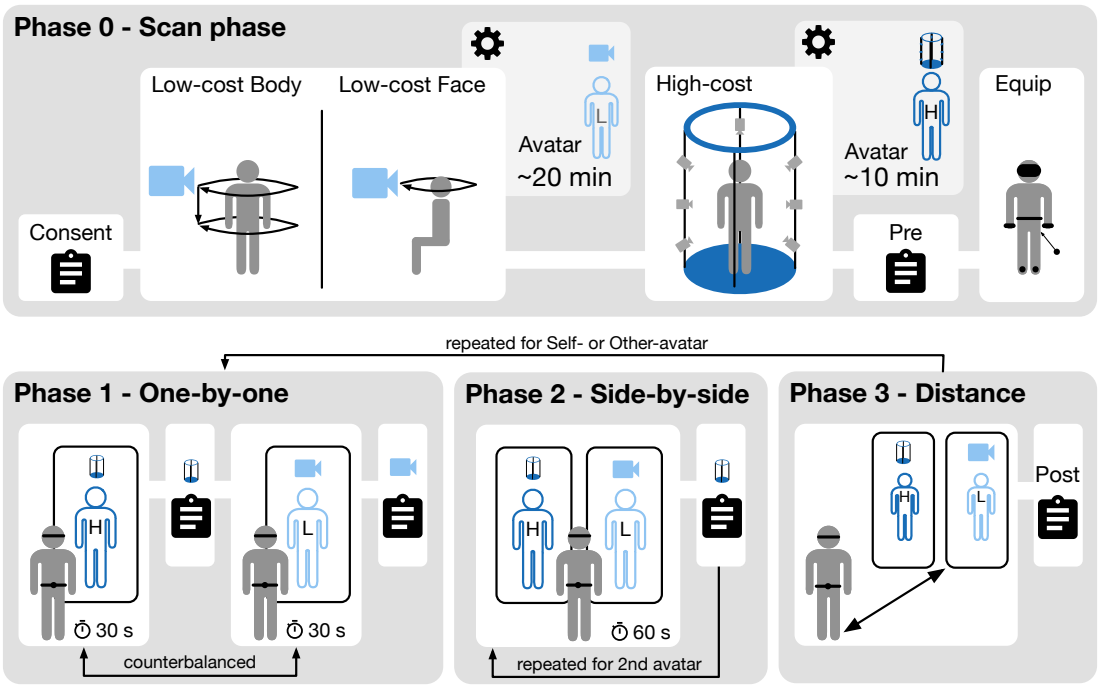

Figure 4.4: The experiment procedure. Phase 0 includes the low- and high-cost scan and avatar creation. Phases 1 to 3 describe the VR exposure. The embodiment in phases 1 and 2 always matched the virtual human to be rated. In Phase 3, participants were embodied with the avatar version they had rated last.

helped the participant to put on the VR equipment and explained how to operate the controllers. After an initial calibration of the avatar, the experiment started. The participants received audio instructions that guided them through the VR exposure phases. The low- and high-cost avatars' rating order and therefore their display in the left or right mirror, was counterbalanced. Each participant went through Phases 1 to 3 twice. Once for the self-avatar, a second time for the virtual other. Half of the participants started with the self-avatar, the other half started with the virtual other. After repeating the phases, participants left the virtual environment and answered the post-questionnaire on a computer in the experiment room.

4.2.5 Participants

A total of $N = 51$ people participated in the study. We had to exclude six participants from the analysis. Three were excluded because the quality of the point cloud of the low-cost scan was insufficient. Another three participants were excluded due to errors in the experimental procedure, e.g., wrong height input when generating the avatars. The mean age of the resulting sample was 21.78 years ($SD = 1.80$), while 75.6 % of the participants stated to be female, and 24.4 % stated to be male. They were all students that received credit points necessary for completing their bachelor's degree. Ten participants had been scanned with the high-cost method before. The sample's VR experience was low, with 84.4 % stating that they have 0 – 5 h of VR experience. Only four participants had no prior VR experience at all.

4.3 RESULTS

The analysis was performed using IBM SPSS Statistics 26. First, we report on the main analysis of the presented user study, including objective measurements. Then we proceed with the results of our control measures. We performed paired t-tests for all within-subjects comparisons and independent t-tests for between-subjects comparisons. Effect sizes are indicated by Cohen's d_z [Coh77].

4.3.1 Perception of the Virtual Humans (RQ_1)

Table 4.2 shows the dependent variables' descriptive data: similarity, uncanniness, and sense of embodiment. Table 4.3 shows the effect sizes of the comparisons.

Similarity: Figure 4.5 shows the results for the perceived similarity. We found no significant difference in the perceived similarity to oneself between

		Phase 1				Phase 2			
		high-cost	low-cost			high-cost	low-cost		
Measurement		$M(SD)$	$M(SD)$	t	p	$M(SD)$	$M(SD)$	t	p
Similarity	self	4.82(1.45)	4.47(1.52)	1.79	.08	4.64(1.45)	4.42(1.52)	0.87	.39
	other	5.22(1.17)	5.51(1.08)	-1.44	.16	5.04(1.30)	5.22(0.97)	-0.85	.40
Human-likeness	self	4.42(1.52)	4.16(1.49)	1.45	.15	4.16(1.41)	3.98(1.34)	1.02	.32
	other	4.93(1.23)	4.91(1.06)	0.11	.92	4.73(1.27)	5.07(0.94)	-1.56	.13
Beauty	self	3.69(1.28)	3.42(1.47)	1.18	.24	3.69(1.51)	3.29(1.41)	1.46	.15
	other	4.38(1.28)	4.67(1.23)	-1.48	.15	4.40(1.39)	4.73(1.01)	-1.39	.17
Eeriness	self	3.98(1.55)	4.36(1.55)	-1.57	.12	4.18(1.81)	4.71(1.63)	-1.76	.09
	other	3.29(1.63)	3.16(1.35)	0.63	.53	3.47(1.78)	3.29(1.36)	0.57	.57
VEQ-Owners.	self	4.09(1.50)	4.06(1.43)	–	–	4.03(1.45)	4.12(1.45)	–	–
VEQ-Agency	self	5.64(1.03)	5.67(0.97)	-0.24	.82	5.43(1.23)	5.42(1.01)	0.08	.94
VEQ-Change	self	3.38(1.53)	3.55(1.47)	-0.82	.42	3.03(1.55)	3.50(1.60)	-2.42	*

Table 4.2: Means, standard deviations, and test statistics for the paired samples t-tests for the perception of the virtual humans. For all t-tests: $df = 44$. $* < .05$.

		Phase 1	Phase 2		Phase 1	Phase 2
Measurement		d_z	d_z		d_z	d_z
Similarity	self	0.27	0.13	other	-0.22	-0.13
Human-likeness	self	0.22	0.15	other	0.02	-0.23
Beauty	self	0.18	0.22	other	-0.22	-0.20
Eeriness	self	-0.24	-0.26	other	0.09	0.09
VEQ-Agency	self	-0.04	-0.07			
VEQ-Change	self	-0.12	-0.36			

Table 4.3: Effect sizes indicated by Cohen's d_z [Coh77] for the perception measures.

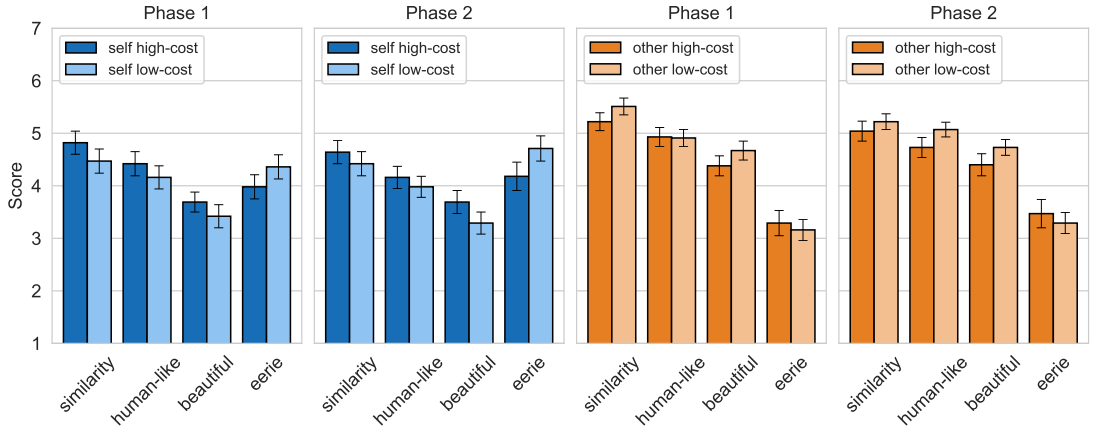

Figure 4.5: Means and standard errors for the measurements of the perceived similarity, human-likeness, beauty, and eeriness of the high- and low-cost self-avatars and virtual others in phases 1 and 2.

the low-cost and the high-cost self-avatar, neither when compared one after the other in Phase 1 nor when compared side-by-side in Phase 2. We also found no significant difference in the perceived similarity to the other person's picture between the low-cost and the high-cost virtual other neither in Phase 1 nor in Phase 2.

Uncanny Valley: Figure 4.5 shows the results for the items human-like, beautiful, and eerie associated with the uncanny valley effect. For the self-avatars, we found no significant difference regarding the perceived human-likeness, beauty, and eeriness of the avatars when evaluated one after the other (Phase 1). We also found no significant difference regarding the perceived human-likeness, beauty, and eeriness of the avatars when evaluated side-by-side (Phase 2). For the virtual others, we also found no significant differences in both phases regarding the perceived human-likeness, beauty, and eeriness.

Sense of Embodiment: We faced problems during the data logging for one of the four items of the subscale Body Ownership. Therefore, we exclude this subscale from the calculation of the comparisons and only report the descriptive statistic derived from the remaining three items. Table 4.2 shows the mean scores calculated with three instead of four items which are almost identical between conditions. Agency did not differ between the high-cost and the low-cost self-avatar in both phases. The perceived change did not differ in Phase 1. It did, however, differ in Phase 2 when participants saw the self-avatars side-by-side. The perceived change of the own body was significantly higher for the low-cost self-avatar than for the high-cost self-avatar. The left diagram in Figure 4.6 shows these results.

Preference: The third diagram in Figure 4.6 shows the participants' preferences for the high- and low-cost self-avatars and virtual others. When asked directly, $n = 27$ participants preferred the high-cost self-avatar and

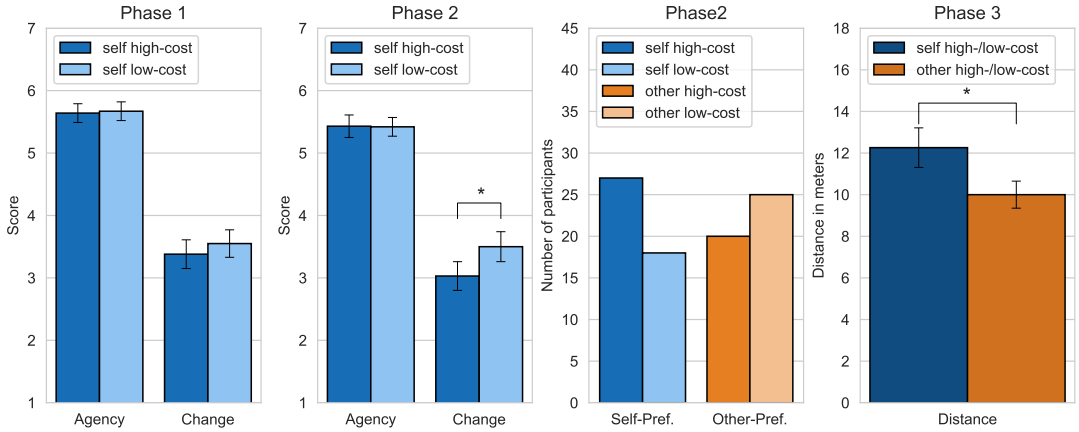

Figure 4.6: From left to right: Means and standard errors for the VEQ subscales Agency and Change for phases 1 and 2. Preference for the low-cost or high-cost self-avatars and virtual others in the number of participants who chose the respective version. Distances in meters at which participants could no longer say that one of the versions was better. $* < .05$

$n = 18$ participants preferred the low-cost self-avatar. On a scale ranging from -2 (*much worse*) to 2 (*much better*), the participants, on average, found the low-cost self-avatar to be only slightly worse than the high-cost self-avatar ($M = -0.42$, $SD = 1.29$). Regarding the virtual others, $n = 20$ preferred the high-cost version and $n = 25$ preferred the low-cost version. On average, they rated the low-cost virtual other to be slightly better than the high-cost virtual other ($M = 0.24$, $SD = 1.15$).

Qualitative Feedback: Participants described the high-cost scan process as interesting, easy, professional, and quick. They stated the number of cameras to be slightly intimidating, futuristic, and strange because they felt observed. As for the low-cost scan, some participants found it strange (especially that a stranger had to film them rather closely), slightly more complicated, more time-consuming, and more exhausting because they had to stand still for a longer time. At the same time, many others described this scan process as easy, interesting, and pleasant. Feedback regarding their preference focused on some main aspects: (1) The face played a vital role in their judgment. Many stated that the bodies of both virtual humans were similarly good in quality. However, artifacts in the face of the one virtual human or a perceived higher similarity made them choose the other version. (2) Participants could rather precisely name artifacts, e.g., messy textures under the arms and inaccuracies in the geometry that deviated from their real body. However, often, they just described an overall feeling that one virtual human was more uncanny or less human-like or more similar to the original. The arguments for the two versions overlapped a lot. However, many of the participants who chose the high-cost avatar as their preference named artifacts on the low-cost avatar as their reason. (3) The lighting and brightness of the virtual human was an

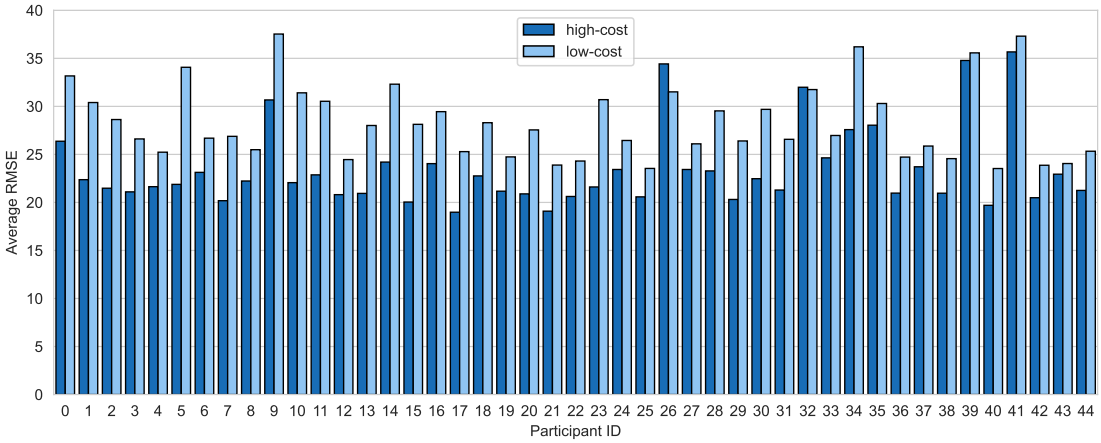

Figure 4.7: Reprojection errors for the high- and low-cost self-avatars of all 45 participants. The reprojection errors were calculated by averaging the root-mean-square error (RMSE) over all input images.

important factor. Some stated that the low-cost version looked more realistic because the lighting looked more natural and that it had more details. Some felt the opposite way, that the high-cost version was illuminated better, was more detailed, and looked more realistic.

4.3.2 Distance (RQ_2)

The right diagram in Figure 4.6 shows the distances that participants set in Phase 3. For the self-avatars, the average distance at which participants could no longer tell which avatar was better was 12.26 m ($SD = 6.36$). For the virtual other, this average distance was 10.00 m ($SD = 4.34$). The distance for the self-avatars was significantly greater than for the virtual other ($t(44) = 2.61$, $p = 0.01$, $d_z = 0.39$).

4.3.3 Objective Measures

Figure 4.7 shows the reprojection error for all participants for both the high-cost and the low-cost self-avatar. On average, the high-cost method's reprojection error was 23.40 ($SD = 4.14$), while the reprojection error of the low-cost method was 28.30 ($SD = 3.84$). A paired samples t-test showed, that the difference was significant ($t(44) = -11.52$, $p < .001$, $d_z = -1.72$). The modified Hausdorff distance [DJ94] between the two reconstructions was, on average, 7.67 mm ($SD = 2.43$). The reprojection errors and the modified Hausdorff distance for the reconstructed avatars in this user study are in the same range as in the objective evaluation conducted in Section 3.3.

		Phase 1				Phase 2			
		high-cost	low-cost			high-cost	low-cost		
Measurement		$M(SD)$	$M(SD)$	t	p	$M(SD)$	$M(SD)$	t	p
Presence	self	5.62(1.17)	5.38(1.35)	1.57	.13	5.44(1.20)	5.47(1.36)	-0.15	.88
	other	5.09(1.35)	5.13(1.25)	-0.39	.70	5.22(1.40)	5.36(1.21)	-1.29	.20
VE made sense	self	5.38(1.34)	5.31(1.38)	0.52	.61	5.50(1.20)	5.38(1.27)	0.93	.36
	other	5.31(1.38)	5.36(1.32)	-0.39	.70	5.33(1.41)	5.29(1.27)	0.36	.72
VE matched	self	5.44(1.14)	5.20(1.16)	1.53	.13	5.51(1.20)	5.16(1.42)	1.91	.06
	other	5.51(1.08)	5.51(0.90)	0	1.00	5.42(1.17)	5.42(1.12)	0	1.00

Table 4.4: Means, standard deviations, and test statistics for the paired samples t-tests for the control measures presence, plausibility of the virtual environment (*VE*), and match of the virtual body to the virtual environment. For all t-tests: $df = 44$.

4.3.4 Control Measurements

The experienced VR sickness before ($M = 7.54$, $SD = 7.98$) and after the experiment ($M = 16.37$, $SD = 12.10$) was low. The observed increase in experienced VR sickness was significant ($t(44) = -5.4$, $p < .001$, $d_z = -0.81$). However, we find this to be uncritical because the values are both low, the application's measured latency was low, the experimenters observed no signs of distress, and the participants did not complain of severe symptoms.

Table 4.4 shows the descriptive data of the control measurements that we took in phases 1 and 2. The subjective experience of presence did not differ between the moment when participants rated the low-cost avatar and when they rated the high-cost avatar, neither when evaluating the self-avatar nor when evaluating the virtual other in both phases. We also found no significant differences regarding the environment's perceived plausibility and the match between the virtual humans and the environment.

4.4 DISCUSSION

This chapter explores the potential of affordable methods for the reconstruction of 3D realistic virtual humans for immersive virtual environments. In a user study, we compared the results of our low-cost method (Chapter 3) to the results of a high-cost method [AWL+17] used as self-avatars and virtual others, while investigating the following two research questions. RQ_1: Can low-cost approaches for generating realistic virtual humans keep up with high-cost solutions regarding the perception of the resulting virtual humans by users

in VR? RQ_2: Are the quality differences more noticeable for the own virtual body than the virtual body of someone else?

For investigating RQ_1, participants evaluated self-avatars and virtual others originating from both reconstruction methods. Users perceived the low-cost virtual humans as similarly human-like, beautiful, and eerie as the high-cost versions for the self-avatars and the virtual others. The perceived similarity between the virtual human and the real counterpart did also not differ between the reconstruction methods. We found no significant differences in perceived similarity, neither when evaluating the self-similarity between participants and the reconstructed virtual humans, nor when evaluating the similarity between the virtual others and pictures of the real persons. The participants' qualitative feedback suggests that the self-avatars' perceived eeriness – independent of the reconstruction method – depended heavily on the virtual humans' face region. A possible explanation is the lack of facial animations. We did not track the users' facial expressions, and therefore, the self-avatars' faces remained static. This rigidity was inconsistent with the otherwise realistic-looking and -moving virtual human. Following the mismatch hypothesis for the uncanny valley effect, which states that inconsistencies in a virtual human's human-like and artificial features may increase negative affinity [KFM+15], this potentially increased the perceived eeriness. The virtual others included basic facial animations and the descriptive data suggests that participants perceived them as less eerie. This is also in line with previous research on the interplay between appearance and behavioral realism, especially regarding the importance of eye movements [GSV+03]. In future work, we plan to track the users' eyes for two reasons. Firstly, this would improve the behavioral realism of the self-avatars. Additional sensors like the Vive face tracker, which entered the market shortly after we conducted our study, would be supplementary improvement options. Secondly, the eye-tracking data could reveal which parts of the virtual humans mostly draw the users' visual attention [DGW+22] and, consequently, impact the evaluation the most. However, our study did not focus on the general perception but on the differences in the perception of the high-cost and low-cost virtual humans.

For the two different self-avatars, we additionally measured the users' sense of embodiment. Participants accepted both self-avatar versions as their virtual body (body ownership) and felt that they were the cause of the self-avatar's actions (agency). In the first phase of the evaluation, when the participants saw the self-avatars consecutively, we also found no significant difference in the embodiment questionnaire's change subscale. However, in the second phase, when participants saw both avatar variations next to each other, the change subscale was significantly higher for the low-cost self-avatars than for the high-cost self-avatars. The subscale change measures the perceived change in the user's body schema [RL20]. According to the questionnaire's authors, the perceived change could be a predecessor of the Proteus effect. When

embodying an avatar that does not look like the user, the perceived change of the users' body would increase with an increased feeling of embodiment. However, a personalized, realistic-looking self-avatar should not create a massive change in the own body schema since it looks (and ideally behaves) like the real body of the user. There are two possible explanations for the increase in perceived change in the second phase: (1) The low-cost self-avatars have more visible inaccuracies than the high-cost self-avatars, e.g., messy textures under the arms. These artifacts on the otherwise very realistic and faithfully reconstructed avatars represent deviations from the users' body, which might cause the increased feeling of change of the own body. (2) These deviations may have also surprised the users and drawn their attention to them. The incoherence with the users' expectations could have created an increased interest and focus on the discrepancies. Latoschik et al. [LKS+19] observed a similar effect when participants interacted with a mixed crowd of virtual characters that drew attention because of their diversity and unexpectedness. However, we did not find significant differences in the feeling of presence, which is usually also partly dependent on the users' attention [SBW17]. The increase in the perceived change only occurred in the second phase, when participants saw the low-cost and high-cost self-avatars next to each other. This direct comparison, and the fact that they saw the self-avatars for the second time at this point, may have further increased the focus on the artifacts. It is possible that the increase in perceived change of the own body only occurs when participants spend a longer time with the virtual body and when they look for discrepancies.

Interestingly, the perception did not differ significantly on most of our measures, even though we found a significant difference in our objective quality measures. The medium may be one possible explanation for this. Despite ongoing technological advances in terms of display quality, today's common consumer HMDs are still limited. We used an HMD with standard resolution (1440×1600) and a wide field of view (120°) that we considered at the upper end of the SteamVR compatible hardware. It would be interesting to see if quality differences between the avatar versions become more apparent using better HMDs, like the HTC Vive Pro 2, that was released after we conducted our study. However, as the user feedback shows, participants were able to spot artifacts quite precisely. Nevertheless, the perceived differences did not manifest themselves in the subjective measurements. This is even more surprising since participants were instructed to really focus on the virtual human. It could mean that other factors, e.g., the movements of the virtual humans, had a stronger influence than the visible artifacts. As a consequence we might want to assume the low-cost smartphone-based version to be an accurate technological match to the available state-of-the-art of VR display devices.

To find out which version was overall preferred, we asked the participants to decide which version of the self-avatars and which version of the virtual others they liked better. Here, the tendency was different between the self-avatars and the virtual others. 60 % of participants preferred the high-cost self-avatars over the low-cost ones. Regarding the virtual others, the result was the exact opposite. Around 56 % of participants preferred the low-cost virtual others over the high-cost ones. This is interesting and supports the overall findings for RQ_1: that the low-cost and high-cost virtual humans are very similar regarding the users' perception.

To sum up our findings regarding RQ_1, we conclude that the low-cost method used in our comparison can indeed keep up with the high-cost method regarding the users' overall perception. The two versions of virtual humans were found to be comparable in terms of their perceived similarity to the original, human-likeness, beauty, and uncanniness. The relatively small effect sizes of the non-significant differences for the self-avatars and the virtual others further support this conclusion.

In our second research question, RQ_2, we focus on the severity of the quality difference for the own body in comparison to the body of a virtual other. Users increased (i) the distance between themselves and their self-avatars and (ii) the distance between themselves and the virtual others until they could no longer tell that one of the virtual humans is better than the other. The distance at which the difference between the low-cost and the high-cost version was no longer noteworthy differed between the self-avatars and the virtual others. For the self-avatars, this point was roughly two meters further away than for the virtual others. This difference implies that smaller discrepancies between the real body and the reconstructed virtual body seem to be more noticeable for one's own body than for another person's body. This is explainable by the familiarity with one's own body, which is usually higher than for someone else's body, in particular if the person is a stranger to you. Our results regarding the participants' preferences also support this assumption. Here, more than half of the participants preferred the low-cost version for the virtual others. However, to further strengthen this finding by correctly representing the interpersonal quality variance of the respective reconstruction methods, a study that evaluates more than two pairs of virtual others would be necessary.

To summarize the results regarding RQ_2, we conclude that the quality difference between the low- and high-cost method plays a more important role for one's own virtual body than for virtual others. In future work, we plan to strengthen this finding by evaluating a more diverse group of virtual others.

The objectively measured quality differences for our sample are similar to those reported in Section 3.3. The reprojection error was significantly higher for the low-cost self-avatars compared to the high-cost self-avatars. However, the severity of the visible artifacts varied a lot within the sample. For

some participants, the reprojection error was even lower for the low-cost self-avatars (subject 26 and 32, see Figure 4.7). We investigated this within-method variance further by scanning the same persons multiple times with both methods. For the resulting virtual humans, we then measured the geometrical variance produced by both methods. This evaluation did, however, not reveal a correlation between the visible artifacts and the geometrical variance.

Although both methods are photogrammetric approaches, they differ in many ways. The low-cost method, for example, uses a stricter regularization to the base model in order to handle uncertainties in the input material. Therefore, the resulting virtual humans' geometry is not as detailed as in the high-cost method. For example, the folds in the clothes are more accurately reconstructed in the geometry of the high-cost version than in the geometry of the low-cost version. The lack of small details in the geometry of the low-cost version is compensated by the texture's great detail instead. Additionally, the low-cost texture contains more baked-in lighting, which gives the impression of detailed geometry even if the underlying geometry is flat, e.g., as in Figure 4.2, where the folds of the clothing are more visible for the low-cost virtual human. Generally, the lighting in the low-cost method is less controlled, since the experimenter walks around the participant. The controlled lighting setup of the high-cost method leads to a more uniform lighting and weaker shadows, allowing for a more faithful lighting in the virtual scene. However, the qualitative feedback shows that the perception of this difference diverges. While some perceived the baked-in lighting as more detailed and more natural, others felt that the more even lighting of the high-cost virtual humans looked more realistic and overall better.

Photogrammetric approaches rely heavily on good quality input material. With the described high-cost setting, it is easier to reach a stable quality of the input photos since many factors are well controlled. Camera positions and lighting conditions stay the same, and experimenters have almost no influence on the outcome since they only trigger the cameras. The low-cost method includes more variable factors that can easily lead to a quality loss in the input video material. For example, the camera may lose focus from time to time, the filming person may make mistakes, the environmental conditions are less controlled, and the subjects have to stand still for a longer period. However, in most cases, the solution to these downsides is straightforward: When the input material is not good enough, repeating the scan process using different camera parameters or different environments, e.g., different lighting conditions or backgrounds, can improve the result. Changing parameters in the complex camera rig proves to be more cumbersome and requires recalibration of the whole system. Therefore, the low-cost variant is not only more affordable but also more viable for a broader range of applications in research and industry.

4.5 SUMMARY AND LIMITATIONS

We compared a high-cost and a low-cost method for 3D reconstruction of virtual humans that differ heavily in their hardware requirements. Both methods use the same photogrammetric reconstruction and template fitting approach with adaptations to the method-specific input material. In a user study, we scanned participants by both methods. Afterwards, they embodied the resulting self-avatars and also encountered virtual others (created with the same methods) in an immersive virtual environment. We found that even though the reconstructions' quality differed on an objective level, the methods did not differ significantly in most of our measurements regarding the users' perception of the virtual humans. Our results further suggest that the quality difference is of greater importance when it comes to one's own virtual body than to a virtual other's body. Based on our findings, we argue that low-cost reconstruction approaches like the method presented in Chapter 3 provide a suitable alternative to high-cost methods, specifically given the current state-of-the-art of available consumer-grade VR displays.

The presented user study has the following limitations: (1) Our sample was predominantly female. Shafer et al. [SCK17] found females to be more prone to VR sickness symptoms, which might partly explain the increase in VR sickness after the experiment. (2) Additionally, the perception of the male and female virtual other differed regarding the perceived uncanniness, which might have resulted from the comparably low number of male participants. For better generalizability of our results, it would be necessary to extend the study by a more balanced sample and more than one female and male pair of virtual others. (3) Our study design is suitable to compare the perception of different versions of virtual humans against each other. However, despite the measurement of the perceived similarity with the real person, we did not include an extensive investigation of the perceived faithfulness of the reconstruction. This was a deliberate decision since it is challenging to find a suitable stimulus for a comparison with reality, e.g., real video material, without changing the medium and therefore impacting the immersion, which in turn can influence the evaluation of a virtual human [WGR+18].

A promising direction for future work is the investigation of causal relations between each method's parameters, their impact on the quality of the reconstruction, and their effect on the users' perception. Our study design can be a helpful basis for conducting these follow-up studies and for guiding the development of similar studies. Ultimately, this allows us to retrieve a set of guidelines for creating and using realistic virtual humans in virtual environments.

5 DATA-DRIVEN BODY WEIGHT MODIFICATION

Having introduced and compared two reconstruction methods for virtual humans in the previous chapters, we will now focus on a crucial feature for employing virtual humans in the context of virtual reality therapy of body image disorders: a model for modifying the body weight of the resulting virtual humans in real-time. In this chapter, we will combine such a body weight modification model with full-body tracking and virtual mirror exposure, thereby following the mirror exposure employed in traditional interventions. This setup allows to immersively expose users to a virtual environment where they can (i) observe a generic or personalized virtual human at different levels of weight or body mass index (BMI), and (ii) directly manipulate the body weight of their personalized avatar to either match their perception of their current body weight or to visualize potential future body weights. This in turn allows researchers to gain insights into possibly occurring body image disturbances. As part of the ViTraS project [DWW+19], we investigate the potential of such a setup in the context of virtual reality therapy of obesity.

Obesity is a complex chronic disease characterized by severe overweight and an above-average percentage of body fat [WHO19]. Its prevalence has more than doubled within recent decades and is expected to rise [VM20; WHO21]. Besides the physical burdens (e.g., an increased risk of several secondary diseases [SBS21]), affected individuals deal with an external or internalized stigmatization that can lead to body image disturbances [MC18; Ros01; TT98]. Body image disturbances are composed of a misperception of body dimensions (body image distortion) and the inability to like, accept, or value one's own body (body image dissatisfaction) and are also associated with a reduced body awareness [TAB+19b; TGK+21]. Various interventions (e.g., cognitive-behavioral therapy supported by mirror exposition or fitness training) have been designed to target persisting disturbances but often only achieve small improvements in the body image [ASW+15]. In recent years, novel methods which employ virtual reality systems for complementing the therapy of body image disturbances have successfully been explored in research with promising results [FGR13; WRG16; RGD+19].

VR-based approaches for supporting body image interventions often use 3D models of human beings [HHM+20; TGK+21], called virtual humans or avatars. VR in general, and the confrontation with embodied avatars in particular, has great potential to influence human perception and behavior [WDH21; RBL+20; YB07]. In the context of body image, avatars have been utilized to expose users of a VR system to generic virtual bodies or body parts varying in size or shape to investigate the principles of body weight perception [Tha19; WDM+20; WMD+21; WFD+22] or to influence the perception or attitude

towards the user's own body [TGK+21]. In the field of computer graphics, recent developments have made progress in all the necessary components of a VR-based body image therapy setup. As detailed in the previous chapters of this thesis, realistic avatars that match a person's real-life appearance can be generated within a short duration and at a low cost. Similarly, there have been advancements in the field of realistic modulation of body dimensions, be it in pictures and videos [ZFL+10; ZJH+18; XTW+20; TSY+21], or in virtual reality [PSR+14; HLZ+20; MMK+21]. However, to the best of our knowledge, no work has yet been presented where users embody their personalized avatar in VR while also having the ability to actively manipulate that avatar's body shape in real-time.

This chapter introduces our statistical model for body weight modification, which is trained on the European subset of the CAESAR database [RBD+02], a collection of 3D scans which are annotated with anthropometric measurements. By correlating the anthropometric measurements with a PCA-based subspace of human body shapes, the resulting model can be evaluated in real-time, giving users the capability to actively modify the body weight of generic virtual humans or personalized avatars. The model is integrated into a VR prototype that allows users to embody a realistic, personalized avatar within a virtual environment, thereby facilitating, e.g., virtual mirror exposure. This makes it possible to further investigate the potential of VR-based interventions in the context of body image therapy.

Individual Contribution *My main contribution is the development of the statistical model of body weight modification and its integration into the VR system. I additionally assisted in integrating the virtual human reconstruction framework at the lab of our colleagues at University of Würzburg. Nina Döllinger and Erik Wolf conceptualized large parts of the design of the virtual reality prototype. They additionally performed the user study which is detailed in the corresponding publication of this chapter. Erik Wolf and David Mal developed the Unity application including the virtual environment and avatar animation system.*

Corresponding Publication *This chapter is based on the following publication*:

> Nina Döllinger, Erik Wolf, David Mal, Stephan Wenninger, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Resize Me! Exploring the User Experience of Embodied Realistic Modulatable Avatars for Body Image Intervention in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022).

The presented body weight modification model has been employed in three further user studies which are not detailed here, but described in the following publications:

> Erik Wolf, Nina Döllinger, David Mal, Stephan Wenninger, Andrea Bartl, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Does Distance Matter? Embodiment and Perception of Personalized Avatars in Relation to the Self-Observation Distance in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022).
>
> Erik Wolf, David Mal, Viktor Frohnapfel, Nina Döllinger, Stephan Wenninger, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Plausibility and Perception of Personalized Virtual Humans between Virtual and Augmented Reality". In *Proc. of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2022, pp. 489–498.
>
> Marie Luisa Fiedler, Erik Wolf, Nina Döllinger, David Mal, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "From Avatars to Agents: Self-Related Cues through Embodiment and Personalization Affect Body Perception in Virtual Reality". *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 30.11 (2024), pp. 7386–7396.

5.1 RELATED WORK

Body image disturbance is characterized by an "excessively negative, distorted, or inaccurate perception of one's own body or parts of it" [WHO19]. It may manifest in body image distortion, the misperception of one's body weight and dimensions that have repeatedly been reported based on underestimations [MMB+08; Val98] or overestimations [TGM+18; DUD+10], or body image dissatisfaction, a negative attitude towards the body that is associated with body image avoidance [WWS18] and reduced body awareness (awareness for bodily signals) [PM11; TAB+19a; TAB+19b; ZSS+13]. While often caused by internalized weight stigma and a fear of being stigmatized by others [MC18], body image disturbance interferes with efforts to stabilize body weight in the long term [Ros01]. Treatments for body image disturbance mainly rely on cognitive-behavioral therapy, typically combining psychoeducation and self-monitoring tasks, mirror exposure, or video feedback [FSL06; ZMS+18; GNH18]. Based on the fundamentals of these established methods, an increasing number of researchers have started to explore VR applications as additional support for attitude and behavior change in general [WDH21] and therapy of body image disturbance [Riv97; FGC+09; FGR13; RGD+19; TGK+21] and obesity in particular [HHM+20; DWW+19].

VR offers the opportunity to immerse oneself in an alternative reality and experience scenarios that are otherwise only achievable via imagination. Endowed with this unique power, mainly the use of avatars has attracted

attention in treating body image disturbance [TGK+21; HHM+20]. Image processing methods for simulating body changes are well established. Using parametric models, it is possible to retouch images to simulate different face or body shapes [ZFL+10; ZJH+18] and even manipulate them in real-time during video playback [XTW+20; TSY+21]. Avatars in VR allow simulating rapid changes in body shape or weight in an immersive environment using life-sized avatars going beyond the presentation of pictures and videos. They enable further general investigation of body weight perception [Tha19; WDM+20; WMD+21]. While some researchers are using multiple generic avatars differing in body weight [NGS+11; PWL+14; KEH+16; PE18; FPM+18], others have developed methods for dynamic body weight modification in VR [APB+00; JSB+08; NZC+18; PSR+14; HLZ+20; MMK+21; NBN+20].

A huge advantage when using advanced body weight modification methods is that the avatar's body weight can be realistically changed to a desired numeric reference value. For this purpose, mainly the body mass index calculated as BMI = Body Weight in kg / (Body Height in m)^2 [WHO00] is used. One example is the work of Thaler et al. [TGM+18], who trained a statistical model to apply realistic BMI-based body weight modification to their generated personalized, photorealistic avatars. But also other factors like muscle mass could be included in such models [MMK+21]. However, while picture and video-retouching methods tend to focus on facial features, the statistical models of weight gain/loss of avatars in VR are usually trained on the whole body [PSR+14] or neglect the head region completely [MMK+21]. For our system, we also learned a statistical model of weight gain/loss for the head region but kept small parts of the face region fixed to preserve the identity of the users when applying the body weight modification.

Besides the shape of the used avatar, application or system-related properties also might alter how we perceive the avatar, and particularly its body weight, in VR. Wolf et al. [WDM+20] presented an overview of potentially influencing factors, noting that while the used display or the observation perspective might unintentionally alter body weight perception [WFD+22; WMF+22], especially the personalization and embodiment of avatars hold potential for application in body image interventions. For example, Thaler et al. [TGM+18] found that the estimator's BMI influences body weight estimations of a realistic and modulatable avatar, but only when the avatar's shape and texture matched the estimator's appearance. This comes along with a recent review by Horne et al. [HHM+20], who identified the personalization of avatars as an important factor when using avatars. For embodiment, Wolf et al. [WMD+21] recently found, for example, that females' own BMI influences body weight estimations of a generic avatar only when embodying it.

5.2 METHOD

To build a statistical model of body weight modification, we follow the approach of Piryankova et al. [PSR+14], who first create a statistical model of body shape using Principal Component Analysis (PCA) and then estimate a linear function from anthropometric measurements to PCA coefficients. For computing the statistical model of human body shape, we use the template fitting process described in Section 2.2.2 to fit our template model to the European subset of the CAESAR scan database [RBD+02]. It consists of $M = 1700$ 3D scans, each annotated with anthropometric measurements such as weight, height, arm span, inseam, waist width, etc. After bringing the scans into dense correspondence via template fitting, we are left with M pose-normalized meshes consisting of V vertices each.

Our approach for data-driven weight gain/loss simulation differs from the method of Piryankova et al. [PSR+14] in the following ways. (1) Instead of encoding body shape as a 3×3 deformation matrix per mesh face [ASK+05], we encode body shape directly via vertex positions. (2) Modeling weight gain/loss as a change in parameters of a statistical parametric shape model [PSR+14; XTW+20] changes face identity during weight modification due to the fact that the learned direction of change is not subject-specific. This leads to undesired effects such as changing the shape of the eye sockets, the pupillary distance or other unrealistic changes in face proportions. To mitigate these effects, we keep vertices in the face region fixed while deforming the rest of the mesh in order to preserve the identity of the participants.

To this end, we define a set $\mathcal{H}$ with cardinality H containing all vertices outside the face region (see Figure 5.1) as well as a selector matrix $\mathbf{H} \in \mathbb{R}^{3H \times 3V}$ which extracts all coordinates belonging to vertices in $\mathcal{H}$. We then build a PCA-based statistical body shape model as described in Section 2.2.1. To recall, let $\mathbf{X}_j = \left(\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_V^\mathsf{T}\right)^\mathsf{T} \in \mathbb{R}^{3V}$ be the vector containing the stacked vertex positions of the j^{th} training mesh and $\bar{\mathbf{X}} = \frac{1}{M} \sum_j \mathbf{X}_j \in \mathbb{R}^{3V}$ be the corresponding mean of all training meshes. Performing PCA of the mean-centered data matrix $\left(\mathbf{H}\left(\mathbf{X}_1 - \bar{\mathbf{X}}\right), \ldots, \mathbf{H}\left(\mathbf{X}_M - \bar{\mathbf{X}}\right)\right) \in \mathbb{R}^{3H \times M}$ and taking the first $k = 30$ components then yields the PCA matrix $\mathbf{U} \in \mathbb{R}^{3H \times k}$, which constitutes our low-dimensional human body shape model.

Our goal now is to use the anthropometric annotations present in the CAESAR dataset to learn a (linear) function from these anthropometric measurements to the low-dimensional coefficients of the human body shape model. In other words, we want to find those directions in the subspace defined by the body shape model, which correlate the most with a given anthropometric measurement. Let $\mathbf{D} = \left(\mathbf{d}_1, \ldots, \mathbf{d}_M\right) \in \mathbb{R}^{k \times M}$ contain the PCA coefficients $\mathbf{d}_j$ of the M training meshes. These are computed by projecting the stacked vertex positions of the training mesh onto the subspace, i.e., $\mathbf{d}_j = \mathbf{U}^\mathsf{T}\mathbf{H}\left(\mathbf{X}_j - \bar{\mathbf{X}}\right)$. If

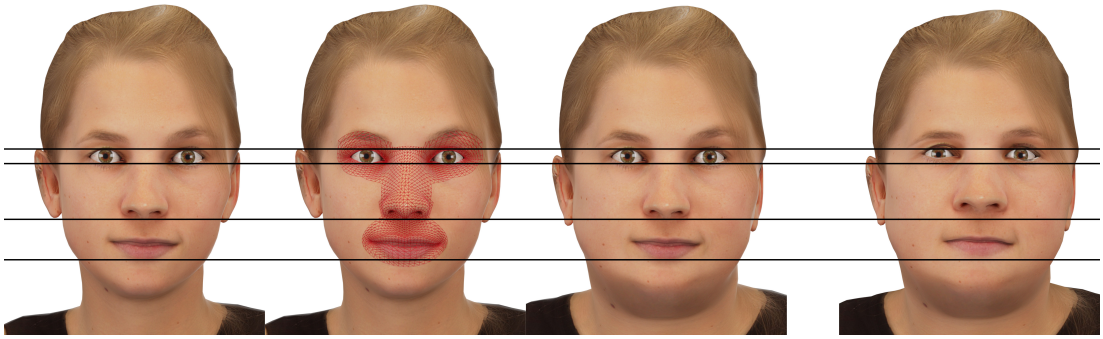

Figure 5.1: The figure illustrates our approach of facial weight gain simulation. When modifying the weight of an avatar (left), part of the face region is fixed (red highlights, center-left). The modified vertices are stitched to the face region in a seamless manner using differential coordinates [Sor05] (center-right). Not keeping these vertices fixed would require recalculating the position of all auxiliary meshes such as eyes and teeth due to the undesired change in facial proportions for nose, mouth and eyes stemming from changing the parameters of the underlying face model (right). For the right image, eyes are copied from the unmodified avatar in order to better highlight the change in shape and position.

we denote by $\mathbf{A} \in \mathbb{R}^{M\times 4}$ the matrix containing the anthropometric measurements weight, height, arm span and inseam of the j^{th} subject in its j^{th} row (following Piryankova et al. [PSR+14]), we can then compute a linear mapping from anthropometric measurements $\mathbf{A}$ to PCA coefficients $\mathbf{D}$ by solving the linear system of equations $\left(\mathbf{A} \mid \mathbf{1}\right)\mathbf{C} = \mathbf{D}^{\mathsf{T}}$ in a least squares sense via normal equations.

New vertex positions for a subject with initial vertex positions $\mathbf{X}$ and a desired change in anthropometric measurements $\Delta\mathbf{a} \in \mathbb{R}^5$ can then be calculated by

$$\tilde{\mathbf{X}} = \mathbf{H}\mathbf{X} + \mathbf{U}\left(\mathbf{C}^{\mathsf{T}}\Delta\mathbf{a}\right) \in \mathbb{R}^{3H}, \tag{5.1}$$

i.e., by first projecting the desired change in measurements into PCA space via the learned linear function and then into vertex position space via the PCA matrix. However, this only updates vertices in $\mathcal{H}$. In order to seamlessly stitch the new vertex positions to the unmodified face region, we compute the Laplacian coordinates (discretized through cotangent weights and Voronoi areas [BKP+10]) of the resulting mesh and then use surface reconstruction from differential coordinates [Sor05]. For the vertices of the face region and its 1-ring neighborhood, the Laplacian is computed based on the unmodified vertex positions $\mathbf{X}$, while for the rest of the vertices, the Laplacian is computed based on the modified vertex positions $\tilde{\mathbf{X}}$. Since the position of vertices of the face region is known and should not change, we treat the position of these vertices as *hard* instead of *soft* constraints as discussed by Botsch and Sorkine [BS08].

Figure 5.2: The figure shows a generated female avatar (BMI = 19.8) with modified body weight corresponding to a BMI range of 16 to 32 in two-point increments.

Consider a subject with weight w, height h, and thus a body mass index of BMI $= w \,/\, h^2$. Setting $\Delta\mathbf{a} = (\Delta w, 0, 0, 0, 0)^\mathsf{T}$ in Equation (5.1) and stitching the modified vertices back to the face region as described then allows modifying the body weight of the user's avatar by Δw while keeping the other anthropometric measurements fixed. Keeping the face region fixed (i) preserves the identity of the user for high values of Δw and (ii) avoids having to recalculate the position of auxiliary meshes of the avatar such as eyes and teeth (Figure 5.1). Results of the described body weight modification method are shown in Figure 5.2. Alternatively, computing the desired change in body weight Δw from a desired change in body mass index ΔBMI can be trivially done by multiplying the desired change in BMI with the (constant) squared height of the user: $\Delta w = \Delta\text{BMI} \cdot h^2$. This can be helpful, as BMI is still used as an intuitive measure of obesity. Note that there are also approaches which argue against the usage of the body mass index and propose to model body composition differently, e.g., via fat and muscle mass [MMK+21].

5.3 VR PROTOTYPE

In our system, the user embodies a personalized avatar from an egocentric perspective while the avatar is animated according to the user's body movements in real-time. Users have active control over the body weight modification model described above through various interaction techniques (for more details, we refer the reader to the study description by Döllinger et al. [DWM+22]). The following sections describe the VR setup, the virtual environments, and the generation and animation of the virtual humans.

5.3.1 VR System

The technical implementation of our VR system is realized using the game engine Unity version 2019.4.15f1 LTS [Uni19]. As VR HMD, we use a Valve Index [Val24b], providing the user a resolution of 1440×1600 pixels per eye with a total field of view of 120° running at a refresh rate of 90 Hz. For motion tracking, we use the two handheld Valve Index controllers, one HTC Vive Tracker 3.0 positioned on a belt at the lower spine, and two HTC Vive Tracker 3.0 on each foot fixed by a Velcro strap. The tracking area was set up using four SteamVR Base Stations 2.0. All VR hardware is integrated using SteamVR in version 1.16.10 [Val24a] and its corresponding Unity plugin in version 2.7.3. In our evaluation, the system was driven by a high-end PC composed of an Intel Core i7-9700K, an Nvidia RTX2080 Super, and 32 GB RAM running Windows 10. The motion-to-photon latency for the body movements was measured as described in Section 4.2.2 and was considered low enough to provide a high feeling of agency towards the avatar [WSH+16], as it averaged 40.9 ms ($SD = 5.4$ ms).

5.3.2 Virtual Environments

We realized two virtual environments. The first environment replicates the real environment, in which the user was located physically during our evaluation, and which is automatically calibrated accurately to overlay the physical environment spatially (see Figure 5.3). Here, all preparatory steps required for exposure are performed and tested (e.g., ground calibration, vision test, equipment adjustments, embodiment calibration). For spatial calibration, we use an implementation of the Kabsch algorithm [MBC+16], based on the positions of the SteamVR base stations in real and virtual environments. Additionally, the

Figure 5.3: The figure depicts a comparison between the real environment where the experiment took place (left) and the replicated virtual environment used for preparation (right). Both environments contain a user, respectively the avatar, performing the embodiment calibration.

Figure 5.4: The images show a participant's personalized avatar standing in front of a mirror within the virtual exposition environment of our concept prototype with a reduced (left), normal (center), or increased (right) body weight.

virtual ground height is calibrated by briefly placing the controller onto the physical ground.

The exposition environment is originally based on an asset taken from the Unity Asset Store that was modified to match our requirements. It is inspired by a typical office of a psychotherapist with a desk and chairs and an exposure area in which the mirror exposure takes place (see Figure 5.4). The exposure area includes a virtual mirror allowing for an allocentric observation of the embodied avatar. We aimed for a realistic and coherent virtual environment to enhance the overall plausibility of the exposure [Sla09; LW22b].

5.3.3 Virtual Human Generation and Animation

The generation of the avatar, which the user embodies inside of the virtual environments, closely follows the virtual human template fitting method of Achenbach et al. [AWL+17] described in Section 2.2. First, the subject is scanned with a custom-built photogrammetry rig located at University of Würzburg. It consists of 94 DSLR cameras, where four studio lights equipped with diffuser balls ensure uniform illumination (see Section 4.2.1). As detailed in the previous chapters of this thesis, the resulting virtual humans are ready for use in VR applications, as they feature a full-body animation rig as well as facial blendshapes for face animations. The statistical body weight modification model described in Section 5.2 can directly be employed to modify the resulting virtual humans, since it is trained on the same template model.

To facilitate avatar animation, the participants' movements are continuously captured using the SteamVR motion tracking devices. For our work, these devices serve as a sufficiently solid and rapid infrared-based tracking solution for the crucial body parts required for animation without aligning different tracking spaces [NLL17]. To calibrate the tracking devices to the user's associated body parts and capture the user's body height, arm length, and current limb orientations, we use a custom-written calibration script that

requires the user to stand in T-pose for a short moment (see Figure 5.3). The calibrated tracking targets of the head, left hand, right hand, pelvis, left foot, and right foot were then used to drive an Inverse Kinematics (IK) [ALC+18] approach realized by the Unity plugin FinalIK version 2.0. FinalIK's integrated VRIK solver continuously calculates the user's body pose according to the provided tracking targets. The tracking pose is automatically adjusted to the determined body height and arm length in order to match the user's body. In the next step, the tracking pose is continuously retargeted to the imported personalized avatar. Potentially occurring inaccuracy in the alignment of the pose or the end-effectors can be compensated by a post-retargeting IK-supported pose optimization step. This leads to high positional conformity between the participant's body and the embodied avatar and avoids sliding feet due to the retargeting process.

5.4 SUMMARY AND LIMITATIONS

We presented a statistical model of body weight modification of virtual humans which allows users of a VR prototype to actively modify the body weight of their personalized embodied avatar in real-time. The presented prototype follows the mirror exposure intervention, which is classically used in therapy of body image disorders, and extends it to exploit the unique capabilities of a VR-based system. The movements of the subject are captured with a lightweight motion tracking setup and retargeted onto the virtual human, which can be observed in a virtual mirror. Additionally, subjects are given active control over their avatar's body weight due to our data-driven body weight modification model. The base of our model is a commercial database of 3D scans annotated with corresponding anthropometric measurements. We then registered a template model to this database and computed a low-dimensional human body shape model. A linear mapping between the anthropometric measurements and the low-dimensional body shape model then allows to map a desired change in anthropometric measurements – body weight in our case – to a change in the learned subspace, which in turn can be mapped to a change in vertex positions. We keep a small part of the face region fixed in order to preserve the user's face identity when modifying body weight. Finally, the modified vertex positions are stitched to the face region in a seamless manner using surface reconstruction from differential coordinates.

Keeping a small part of the face region fixed can however be seen as a small limitation of our work, as it does not completely accurately model weight gain/loss in this region. Studies have shown, that the soft tissue in this area of the face also changes with varying body weight [DCV+06]. Other methods deform the whole face region [PSR+14; ZJH+18; TSY+21] or regularize the deformation of a region similar to ours [XTW+20]. These

methods, however, produce other undesirable effects such as changing eye socket shape or pupillary distance due to the fact that the underlying statistical model produces one direction of change that is applied to all avatars. As the data measured by De Greef et al. [DCV+06] shows, the soft tissue thickness in our fixed region does positively correlate with BMI. However, we note that the correlation for landmarks in our fixed region is smaller than for those outside the fixed region and as such we decided to keep the face region around the eyes, nose, and mouth fixed. As seen in Figure 5.1, this still produces plausible results while avoiding undesirable changes in face identity. Finally, the body weight modification model is trained on surface scans only, preventing us from accurately reasoning about the body composition, i.e., the bone structure and the distribution of muscle and fat tissue. This is a limitation which we will tackle in the next chapters of this thesis.

6 A THREE-LAYERED HUMAN ANATOMY MODEL

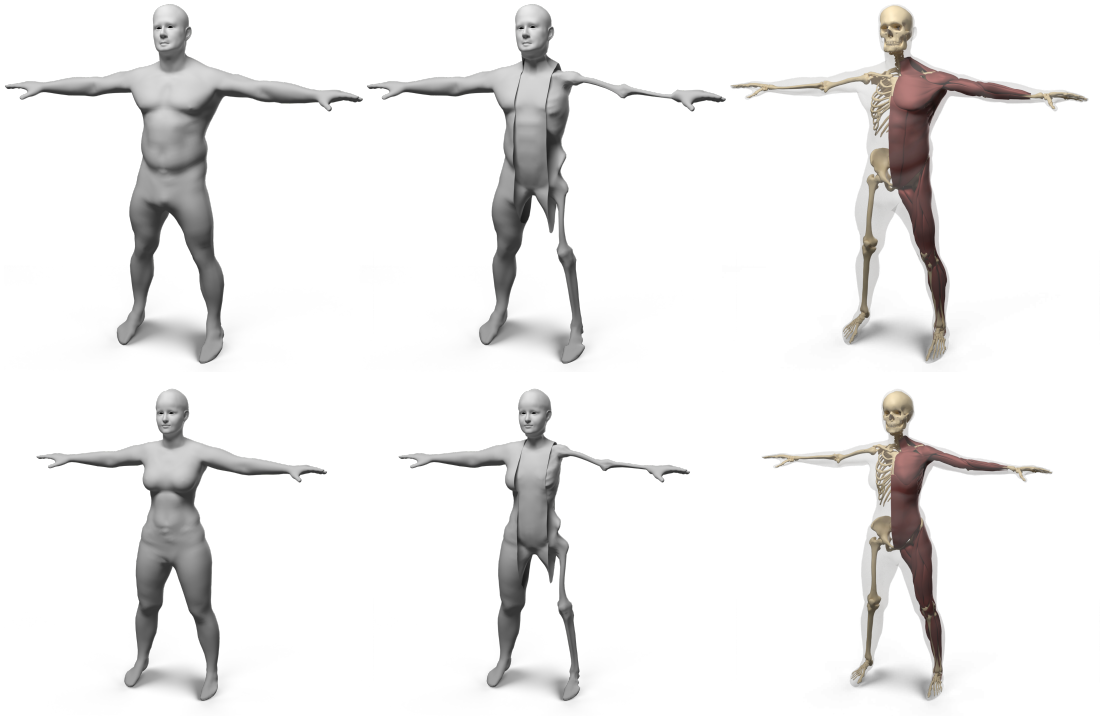

Figure 6.1: Starting with the surface of a human (left), we fit a three-layered model consisting of a skin, muscle, and skeleton surface (middle), which enables physical simulations in a simple and intuitive way. Interior structures, such as individual models of muscles and bones, can also be transferred using our layered model (right).

So far, we have focused on surface-based representations of virtual humans. By bringing 3D scans of various people into dense correspondence via template fitting, we are able to construct surface-based human body shape models (Chapter 2), which can be used as a prior when fitting to noisy or incomplete data (Chapter 3). As seen in the previous chapter, correlating such human body shape models with anthropometric data allows us to learn a model for modifying these virtual humans based on anthropometric measurements such as body weight. This already yields convincing results, but does not take any information about subject-specific anatomical traits such as bone structure or body composition into account. This chapter thus focuses on incorporating anatomical details into surface-based models, in order to provide more information to models that aim to modify and animate virtual humans.

For the purpose of creating convincing animations of surface-based virtual humans, large amounts of 3D scans of human beings have been collected to build sophisticated surface-based shape and pose models [LMR+15; ASK+05; BRP+17]. These models compensate for the fact that they lack anatomical

information by capturing and analyzing surface scans of the same person in various poses, thereby capturing some of the effects that the inner anatomy has on the surface hull. Another way to approach this is to model *volumetric* virtual humans by incorporating (discrete approximations of) their interior anatomical structures. While surface-based models might be sufficient for many applications, for others (e.g., surgery simulation) a volumetric model is an essential prerequisite.

While detailed volumetric models of the human body exist [Ack98; CKH+09; Zyg24], they can be very tedious to work with. Since they usually consist of hundreds of different bones, muscles, organs and tendons, simply creating a volumetric tetrahedral mesh for simulation purposes can be frustratingly difficult. Moreover, those models represent average humans and transferring their volumetric structure and anatomical details to a specific human model (e.g., a scanned person) is not straightforward. Although there are a couple of approaches for transferring the interior anatomy from a volumetric template model into a surface-based virtual human [DLG+13; KIL+16], these methods either deform bone structures in a non-plausible manner [DLG+13] or require a complex numerical optimization [KIL+16].

In this chapter we present a robust and efficient method for transferring an interior anatomy template into a surface mesh in just a couple of seconds. A key component is a simple decomposition of the human body into three layers that are bounded by surface meshes, which share the same triangulation: the *skin surface* defines the outer shape of the human, the *muscle surface* envelopes its individual muscles, and the *skeleton surface* wraps the subject's skeleton (see Figure 6.1 middle). The muscle layer is hence enclosed in between the skeleton and muscle surface, and the subcutaneous fat tissue by the muscle surface and skin surface. This layered template model is derived from the Zygote body model [Zyg24], which provides an accurate representation of both the male and female anatomy.

We propose simple and fast methods for fitting the layered template to surface scans of humans and for transferring the high-resolution anatomical details [Zyg24] into these fitted layers (see Figure 6.1 right). Given a 3D scan of a person, we first fit our surface-based template model to the skin surface. We then estimate body composition via a learned regressor which outputs estimations of fat and muscle mass given the skin surface. Subsequently, we fit our three-layered volumetric template model into the skin surface while conforming to the estimated body composition. Finally, the high-resolution anatomical details are transferred via a space warp based on triharmonic radial basis functions. Our method is robust, efficient, and fully automatic, which we demonstrate on about 1700 scans from the European CAESAR dataset [RBD+02]. We further show a few example applications such as fat growth, fat transfer, and physics-based character animation, which are made feasible or enhanced by our volumetric representation.

Our approach enriches simple surface scans by plausible anatomical details, which are suitable for educational visualizations and volumetric simulations. We note, however, that due to the lack of true volumetric information, it is not a replacement of volumetric imaging techniques in a medical context.

Individual Contribution *The approach for creating volumetric anatomical representations of virtual humans from surface scans was developed together with Martin Komaritzan. I implemented the pre-processing of the different datasets used for training and evaluating the model, enabling statistical analysis by bringing all datasets into dense correspondence via template fitting. My second main contribution is the creation of the regressor that estimates the amount of fat and muscle mass from the skin surface only. Martin Komaritzan created the layered volumetric template from the Zygote model, developed the process for fitting the layered volumetric template into a skin surface, and implemented the space warp which transfers the high-resolution anatomical details from the template model to the fitted model.*

Corresponding Publication *This chapter is based on the following publication*:

> Martin Komaritzan, Stephan Wenninger, and Mario Botsch. "Inside Humans: Creating a Simple Layered Anatomical Model from Human Surface Scans". *Frontiers in Virtual Reality* 2 (2021).

6.1 RELATED WORK

Using a layered volumetric model of a virtual character has been shown to be beneficial compared to a surface-only model in multiple previous works. Deul and Bender [DB13] compute a simple layered model representing a bone, muscle, and fat layer, which they use for a multi-layered skinning approach. Their physically-based character skinning provides dynamic motion effects such as jiggling of fat tissue, collision handling, and volume conservation. Simplistic layered models have also been used to extend the SMPL surface model [LMR+15] in order to support elastic effects in skinning animations [KPP+17; ROC+20]. These approaches combine physics-based simulation based on the finite element method and a data-driven statistical surface model. Compared to these works, our three layers yield an anatomically more accurate representation of the human body, while still being simpler and more efficient than complex irregular tetrahedralizations. Saito et al. [SZK15] show that a layer that envelopes muscles and separates them from subcutaneous fat tissue yields more convincing muscle growth simulations and reduces the number of tetrahedral elements required in their computational model. They also show how to simulate different variations of bone sizes, muscle mass, and fat mass for a virtual character.

When it comes to the generation of realistic personalized anatomical structures from a given skin surface, most previous works focus on the human head: Ichim et al. [IKN+16] register a template skull model to a surface scan of the head in order to build a combined animation model using both physics-based and blendshape-based face animation. In their follow-up work *Phace* [IKK+17], the authors also incorporate facial muscles and a muscle activation model to allow more advanced face animation effects. Gietzen et al. [GBA+19] and Achenbach et al. [ABG+18] use volumetric CT head scans and surface-based head scans in order to learn a combined statistical model of the head surface, the skull surface, and the enclosed soft tissue, which allows them to estimate the head surface from the skull shape and vice versa. More recent work also highlights the advantages of anatomically constrained face models for facial animation and retargeting purposes [CZ24; WSB24a; WSB24b; YZC+24]. Regarding the other parts of the body, Zhu et al. [ZHK15] propose an anatomical model of the upper and lower limbs that can be fit to surface scans and is able to reconstruct motions of the limbs.

There are few approaches for generating an anatomical model of the *complete core* human body (torso, arms, legs) from a given skin surface. In their pioneering work, Dicko et al. [DLG+13] transfer the anatomic details from a template model to various humanoid target models, ranging from realistic body shapes to stylized non-human characters. They transfer the template's anatomy through a harmonic space warp and per-bone affine transformations,

which might, however, distort muscles and bones in an implausible way. Different distributions of subcutaneous fat can be (and have to be) painted manually into a special fat texture. The work of Kadleček et al. [KIL+16] is most closely related to our approach. They reconstruct a personalized anatomically plausible volumetric model from a set of 3D scans of a person in different poses. An inverse physics simulation is used to fit a volumetric anatomical template model based on the commercially available Zygote model [Zyg24] to the set of surface scans, where custom constraints prevent muscles and bones from deforming in an unnatural manner. We discuss the main differences of our approach and Dicko et al. [DLG+13] and Kadleček et al. [KIL+16] in Section 6.3.

Estimating the body composition from surface measures or 3D surface scans (like we do in Section 6.2.3) has been tackled before. There are numerous formulas for computing body fat percentage (BF), or body composition in general, from certain circumferences, skinfold thicknesses, age, gender, height, weight, and density measurements. Prominent examples are the skinfold equations, or the Siri- and Brozek formulas [JP85; Sir56; BGA+63]. These formulas, however, either rely on anthropometric measurements that have to be taken by skilled personnel or on measuring the precise body density via expensive devices, such as BOD PODs [FGM02]. Ng et al. [NHF+16] compute BF based on a 3D body scan of the subject, but their formula is tailored towards body scans and measurements taken with the Fit3D Scanner [Fit24]. Even with the help of the authors we could not successfully apply their formulas to scans taken with different systems, since we could always find examples resulting in obviously wrong (or even negative) BF. Maalin et al. [MMK+21] showed that modeling body composition through body fat alone is an inferior measure for defining the shape of a person compared to a combined model of fat mass and muscle mass. We therefore adapt their data to estimate fat mass and muscle mass from surface scans alone (Section 6.2.3). Incorporating these estimations into the volumetric fitting process allows us to determine how much of the soft tissue layer is described by muscle tissue more plausibly than Kadleček et al. [KIL+16].

Similar to the decomposition into fat, muscle, and bone tissue we employ in this chapter, the recent *HIT* method [KAD+24] learns to estimate these three layers from a given skin surface modeled via SMPL [LMR+15]. A set of MRI images is segmented into bone tissue, tissue belonging to muscles and organs, fat tissue, and empty space in a semi-automatic fashion. From this data, the authors then train multiple neural networks, which handle decompression (modeling soft tissue displacement due to contact with the MRI table during scanning), unposing to a common rest pose, deshaping to a canonical template shape space, and finally tissue class prediction. Due to this implicit representation, the final model is then able to predict the soft tissue class given a 3D position as well as SMPL shape and pose parameters.

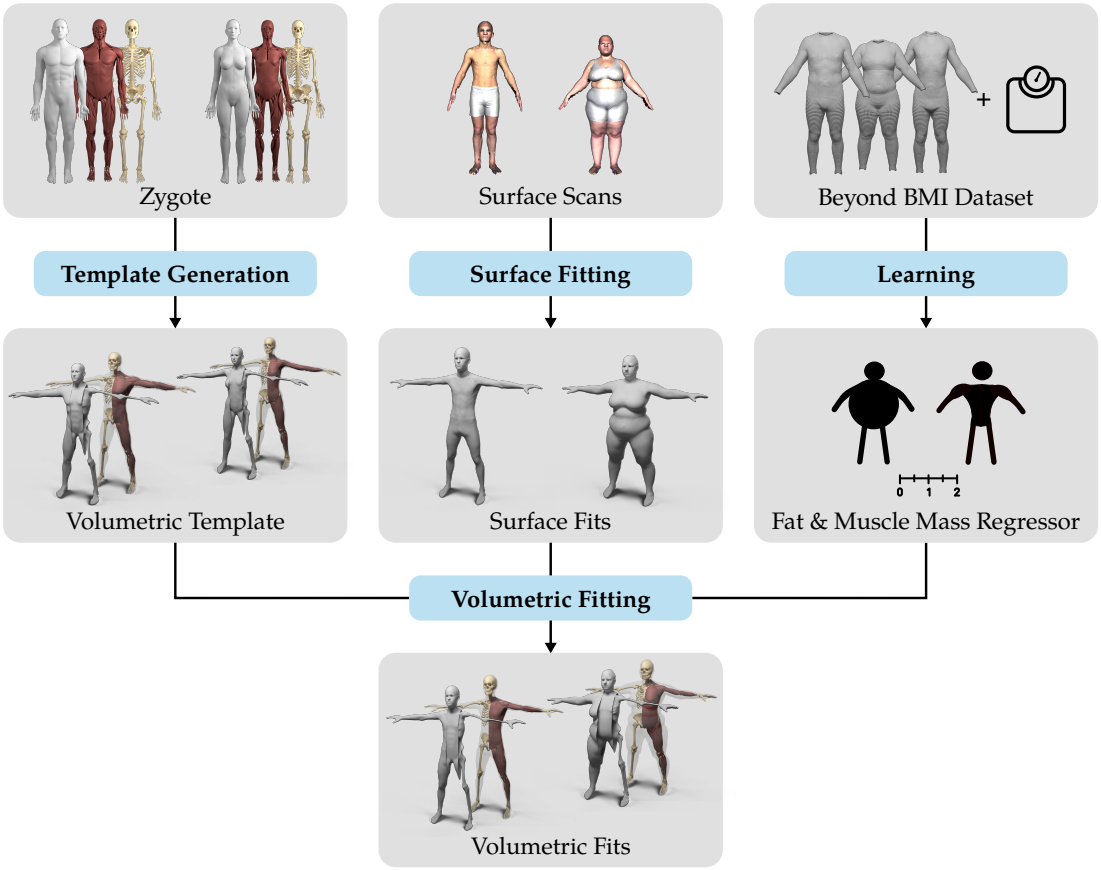

Figure 6.2: Overview of our volumetric template fitting approach. From the Zygote model [Zyg24], we build layered volumetric templates for the male and female anatomy. By adapting the BeyondBMI dataset [MMK+21], we learn a model for estimating fat and muscle mass from a surface model. Given a person's surface scan, we then estimate its fat/muscle mass and use this information to fit the volumetric template (in)to the surface scan, which yields the personalized anatomical model.

6.2 METHOD

Our approach consists of three main contributions: First, the generation of the volumetric three-layered template, described in Section 6.2.2, where we derive the skin, muscle, and skeleton layers from the male and female Zygote model [Zyg24]. Second, an efficient method for fitting this layered model (including all contained anatomical details) (in)to a given human surface scan (Section 6.2.4). Third, the estimation of a person's body composition, i.e., how much of the person's soft tissue is described by muscles and fat (Section 6.2.3). By adapting the BeyondBMI dataset [MMK+21] to our template, we derive this information from the surface scan alone and use it to inform the volumetric template fitting. Figure 6.2 shows an overview of the complete process, starting from the different input data sets, the template model and the muscle/fat regressor, to the final personalized anatomical fit.

6.2.1 Data Preparation

In our approach we make use of several publicly or commercially available datasets for model generation, learning, and evaluation:

- *Zygote:* The Zygote model [Zyg24] provides high-resolution models for the male and female anatomy. We use their skin, muscle, and skeleton models for building our layered template.
- *BeyondBMI:* Maalin et al. [MMK+21] scanned about 400 people and additionally measured their fat mass (FM), muscle mass (MM), and body mass index (BMI) using a medical-grade eight-electrode bioelectrical impedance analysis scale. They provide annotated (synthetic) scans of 100 men and 100 women, each computed by averaging shape and annotations of two randomly chosen subjects. From this data we learn a regressor that estimates fat and muscle mass from the skin surface.
- *Hasler:* The dataset of Hasler et al. [HSS+09] contains scans of 114 subjects in 35 different poses, captured by a 3D laser scanner. The scans are annotated with fat and muscle mass percentage as measured by a consumer-grade impedance spectroscopy body fat scale. We use this dataset to evaluate the regressor learned from the BeyondBMI data.
- *CAESAR:* The European subset of the CAESAR scan database [RBD+02] consists of 3D scans (with about 70 selected landmarks) equipped with annotations (e.g., weight, height, BMI) of about 1700 subjects in a standing pose. We use this data to evaluate our overall fitting procedure.

All these data sources use different model representations, i.e., either different mesh tessellations or even just point clouds. In a preprocessing step we therefore re-topologize the skin surfaces of these datasets to a common triangulation by fitting a surface template using the non-rigid surface-based registration of Achenbach et al. [AWL+17] as described in Section 2.2.

To recap briefly, this approach is based on an animation-ready, statistical template model. Its mesh tessellation, animation skeleton, and skinning weights originate from the Autodesk Character Generator [Aut24], and the variation in human body shape is represented by a PCA-based shape model. We will refer to this template model as the *surface template* in the following. In a preprocessing step we fit the surface template to all input surface scans to achieve a common triangulation and thereby establish dense correspondence. This fitting process is guided by a set of landmarks, which are either specified manually or provided by the dataset. A non-linear optimization then determines alignment (scaling, rotation, translation), body shape (PCA parameters), and pose (inverse kinematics on joint angles) in order to minimize squared distances of user-selected landmarks and automatically determined closest point correspondences in a non-rigid ICP manner [BTP14]. Once the model

parameters are optimized, a fine-scale out-of-model deformation improves the matching accuracy and results in the final template fit. Please refer to Section 2.2 for more details about the template model and the surface-based fitting process.

6.2.2 Generating the Volumetric Template

We use the male and female Zygote body model [Zyg24] as a starting point for our volumetric model. Our volumetric template is defined by the *skeleton surface* $\mathcal{B}$ (for bones), the *muscle surface* $\mathcal{M}$, and the *skin surface* $\mathcal{S}$. The skeleton is enveloped by the skeleton surface, the muscle layer is enclosed between the skeleton surface and the muscle surface, and the (subcutaneous) fat layer is enclosed by the muscle surface and the skin surface. The soft tissue layer is the union of the fat and muscle layers. In our layered model we exclude the head, hands, and toes. These regions will be identical to the skin surface in all layers. See Figure 6.3 for a visualization of the layered template.

The three surfaces $\mathcal{B}$, $\mathcal{M}$, and $\mathcal{S}$ will be constructed to share the same triangulation, providing a straightforward one-to-one correspondence between the i^{th} vertices on each surface, which we denote by $\mathbf{x}_i^{\mathcal{B}}$, $\mathbf{x}_i^{\mathcal{M}}$, and $\mathbf{x}_i^{\mathcal{S}}$, respectively. Each two corresponding triangles $(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_j^{\mathcal{S}}, \mathbf{x}_k^{\mathcal{S}})$ on $\mathcal{S}$ and $(\mathbf{x}_i^{\mathcal{M}}, \mathbf{x}_j^{\mathcal{M}}, \mathbf{x}_k^{\mathcal{M}})$ on $\mathcal{M}$ span a volumetric element of the fat layer. Similarly, the volumetric elements of the muscle layer are spanned by pairs of triangles $(\mathbf{x}_i^{\mathcal{M}}, \mathbf{x}_j^{\mathcal{M}}, \mathbf{x}_k^{\mathcal{M}})$ on $\mathcal{M}$ and $(\mathbf{x}_i^{\mathcal{B}}, \mathbf{x}_j^{\mathcal{B}}, \mathbf{x}_k^{\mathcal{B}})$ on $\mathcal{B}$. We call these elements, built from six vertices of two triangles, *prisms*, and will either use them directly in a simulation or (trivially) split them into three tetrahedra each, resulting in a simple conforming volumetric tessellation.

In the following, we describe how to generate the skeleton surface $\mathcal{B}$ and the muscle surface $\mathcal{M}$. The skin surface $\mathcal{S}$ is generated by fitting the surface-based template of Achenbach et al. [AWL+17] to the skin of the anatomical model [Zyg24], as described in Section 6.2.1.

The Skeleton Surface

The skeleton surface $\mathcal{B}$ should enclose all the bones of the detailed skeleton model, as shown in Figure 6.3 (center). We achieve this by shrink-wrapping the skin surface $\mathcal{S}$ onto the skeletal bones. To avoid problems caused by gaps between bones (e.g., in the rib cage or between tibia and fibula), we first generate a skeleton wrap $\mathcal{W}$, a watertight genus-0 surface that encapsulates the bones, and then shrink-wrap the skin surface to $\mathcal{W}$ instead. The wrap surface $\mathcal{W}$ can easily be generated by a few iterations of shrink-wrapping, remeshing, and smoothing of a bounding sphere in a 3D modeling software like Blender or Maya. This results in a smooth, watertight, and two-manifold

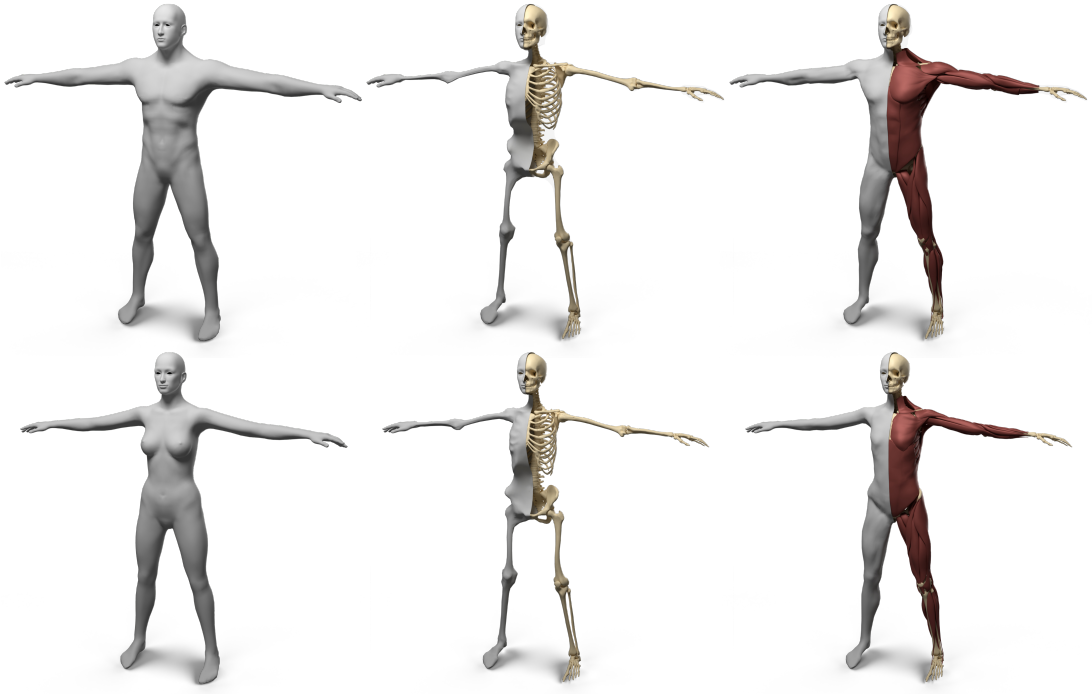

Figure 6.3: Our layered template for both male (top) and female (bottom): the skin surface (left), the skeleton surface enveloping the skeleton (center), and the muscle surface enveloping both muscles and skeleton (right). For the center and right column, the left half shows the enveloping surface while the right half shows the enveloped anatomical details.

surface $\mathcal{W}$ that excludes regions like the interior of the rib cage and small holes like in the pelvis or between ulna and radius.

We generate the skeleton surface $\mathcal{B}$ by starting from the skin surface $\mathcal{S}$, i.e., setting $\mathcal{X} = \mathcal{S}$, and then minimizing a non-linear least squares energy that is composed of a fitting term E_{wrap}, which attracts the surface $\mathcal{X}$ to the bone wrap $\mathcal{W}$, and a regularization term E_{reg}, which prevents $\mathcal{X}$ from deforming in a physically implausible manner from its initial state $\overline{\mathcal{X}} = \mathcal{S}$:

$$\mathcal{B} \;=\; \arg\min_{\mathcal{X}} \; \lambda_{\text{wrap}} E_{\text{wrap}}(\mathcal{X}, \mathcal{W}) + \lambda_{\text{reg}} E_{\text{reg}}\big(\mathcal{X}, \overline{\mathcal{X}}\big)\,. \tag{6.1}$$

The regularization is formulated as a discrete bending energy that penalizes the change of mean curvature, measured as the change of length of the Laplacian:

$$E_{\text{reg}}\big(\mathcal{X}, \overline{\mathcal{X}}\big) = \sum_{\mathbf{x}_i \in \mathcal{X}} A_i \left\| \Delta \mathbf{x}_i - \mathbf{R}_i \Delta \bar{\mathbf{x}}_i \right\|^2, \tag{6.2}$$

where $\mathbf{x}_i$ and $\bar{\mathbf{x}}_i$ denote the vertex positions of the deformed surface $\mathcal{X}$ and the initial surface $\overline{\mathcal{X}}$, respectively. The matrix $\mathbf{R}_i \in SO(3)$ denotes the optimal rotation aligning the vertex Laplacians $\Delta \mathbf{x}_i$ and $\Delta \bar{\mathbf{x}}_i$, which are discretized using the cotangent weights and Voronoi areas A_i [BKP+10].

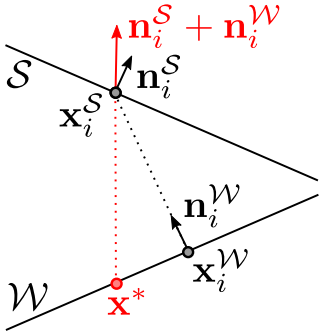

Figure 6.4: An example of minimizing the alignment energy (6.4) for the skin vertex $\mathbf{x}_i^{\mathcal{S}}$. Its closest position on the skeleton wrap is $\mathbf{x}_i^{\mathcal{W}}$, leading to a small minimal angle (between black dotted line and $\mathcal{S}$). The position $\mathbf{x}^*$ maximizes the minimal angle and minimizes the energy (6.4). It can be found by tracing a line from $\mathbf{x}_i^{\mathcal{S}}$ in negative direction of the average normal $\mathbf{n}_i^{\mathcal{S}}$.

The fitting term penalizes the squared distance of vertices $\mathbf{x}_i \in \mathcal{X}$ from their target positions $\mathbf{t}_i \in \mathcal{W}$:

$$E_{\text{wrap}}(\mathcal{X}, \mathcal{W}) = \sum_{\mathbf{x}_i \in \mathcal{X}} w_i A_i \left\| \mathbf{x}_i - \mathbf{t}_i \right\|^2 . \tag{6.3}$$

The target positions $\mathbf{t}_i$ are points (not necessarily vertices) on the skeleton wrap $\mathcal{W}$ of either one of three types: closest point correspondences, fixed correspondences, or collision targets. The type of the target position $\mathbf{t}_i$ then determines the weight w_i, which we empirically set to 0.1 for closest point correspondences, 1 for fixed correspondences, and 100 for collision targets. We define just one target $\mathbf{t}_i$ for each vertex $\mathbf{x}_i$. The default is a closest point correspondence per vertex, which can be overridden by a fixed correspondence, and both of them will be overridden by the collision target in case of a detected collision. Below we explain the three target types.

Closest point correspondences are computed in each step of our iterative minimization and set to the point on the surface of the skeleton wrap $\mathcal{W}$, which is closest to the vertex $\mathbf{x}_i \in \mathcal{X}$, i.e., $\mathbf{t}_i = \arg\min_{\mathbf{y} \in \mathcal{W}} \left\| \mathbf{x}_i - \mathbf{y} \right\|$.

Near complicated regions, like the armpit or the rib cage, the skin has to stretch considerably to deform toward the skeleton wrap. As a consequence, corresponding triangles $(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_j^{\mathcal{S}}, \mathbf{x}_k^{\mathcal{S}})$ on the skin surface $\mathcal{S}$ and $(\mathbf{x}_i^{\mathcal{B}}, \mathbf{x}_j^{\mathcal{B}}, \mathbf{x}_k^{\mathcal{B}})$ on the eventual skeleton surface $\mathcal{B}$ will not be approximately on top of each other, but instead be tangentially shifted. These two triangles span a volumetric element that we call a *prism*. Misaligned triangles will lead to heavily sheared prisms, which can cause artifacts in physical simulations.

We define a per-vertex score penalizing misalignment of corresponding vertices $\mathbf{x}_i^{\mathcal{S}} \in \mathcal{S}$ and $\mathbf{x}_i^{\mathcal{W}} \in \mathcal{W}$ w.r.t. their common averaged normal $\mathbf{n}_i^{\mathcal{S}} + \mathbf{n}_i^{\mathcal{W}}$:

$$E_{\text{align}}\left(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_i^{\mathcal{W}}\right) = \left| \frac{\left(\mathbf{n}_i^{\mathcal{S}} + \mathbf{n}_i^{\mathcal{W}}\right) \cdot \left(\mathbf{x}_i^{\mathcal{S}} - \mathbf{x}_i^{\mathcal{W}}\right)}{\left\|\mathbf{n}_i^{\mathcal{S}} + \mathbf{n}_i^{\mathcal{W}}\right\| \cdot \left\|\mathbf{x}_i^{\mathcal{S}} - \mathbf{x}_i^{\mathcal{W}}\right\|} - 1 \right| . \tag{6.4}$$

A 2D example of this is shown in Figure 6.4.

Fixed correspondences are responsible for reducing these tangential shifts and thereby improving the prism shapes. We determine them for some vertices at the beginning of the fit as explained in the following and keep them fixed

Figure 6.5: Standard non-rigid registration from skin to skeleton (left) results in a bad tangential alignment of corresponding triangles, causing sheared prisms, which we visualize by color-coding the alignment error (6.4). Using fixed correspondences reduces this error (center). Shifting closest point correspondences with bad alignment reduces the error even further (right).

throughout the optimization. Since the alignment error increases faster if the distance between skin surface and skeleton wrap is small, we specify fixed correspondences for vertices on $\mathcal{S}$ that have a distance less than 3 cm to $\mathcal{W}$. For each such vertex we randomly sample points in the geodesic neighborhood of $\mathbf{x}_i^{\mathcal{W}}$ and select the one that minimizes Equation (6.4) as fixed alignment constraint, where we generate normal vectors of sample points using barycentric Phong interpolation. To avoid interference of spatially close fixed correspondences, we add them in order of increasing distance to the skeleton, but only if their distance to all previously selected points is larger than 5 cm. In that way, we get a well distributed set of fixed correspondences, favoring those with a small skin-to-skeleton distance. Figure 6.5 (center) shows that this already reduces the alignment error.

Closest point correspondences can also drag vertices to locations with high alignment error. In each iteration of the non-rigid ICP, we compute $E_{\text{align}}(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_i)$ for each vertex on $\mathcal{S}$ and its counterpart on the current state of $\mathcal{X}$. If this error exceeds a limit of 0.01, which corresponds to an angle deviation of 8° from the optimal angle, we sample the one-ring neighborhood of vertex $\mathbf{x}_i$ on $\mathcal{X}$ and set $\mathbf{x}_i$ to the sample with minimal alignment error and update its closest point correspondence on $\mathcal{W}$. This strategy reduces the alignment error even further, as shown in Figure 6.5 (right).

In the process of moving the surface $\mathcal{X}$ toward $\mathcal{W}$, these two meshes might intersect each other, violating our goal that in the converged state the surface $\mathcal{X}$ (i.e., $\mathcal{B}$, due to Equation (6.1)) should fully enclose $\mathcal{W}$. We therefore detect these collisions during the optimization and resolve them through *collision targets*. We use the exact continuous collision detection of Brochu et al. [BEB12] to detect collisions. In case of a collision, we backtrack the triangles' linear path from the current $\mathcal{X}$ to the initial $\mathcal{S}$ to find the non-colliding state closest to $\mathcal{X}$. This state defines *collision targets* $\mathbf{t}_i$ for colliding vertices $\mathbf{x}_i$, which override the other types of target positions. In case of multiple collision targets $\mathbf{t}_i$ for

the same vertex $\mathbf{x}_i$, we determine all non-colliding states separately and choose the one that is closest to the initial skin surface $\mathcal{S}$.

For the minimization described in Equation (6.1), we use the Projective Dynamics framework of Bouaziz et al. [BDS+12; BML+14], implemented through an adapted local/global solver from the ShapeOp library [DDB+15]. We initialize the optimization by setting $\mathcal{X} = \mathcal{S}$ and $\lambda_{\text{wrap}} = \lambda_{\text{reg}} = 1$. Once the optimization converges, we decrease λ_{reg} by a factor of 0.1, and update the undeformed Laplacians $\Delta\bar{\mathbf{x}}_i$ in Equation (6.2) to the Laplacians $\Delta\mathbf{x}_i$ of the current solution $\mathcal{X}$. This process is iterated until $\lambda_{\text{reg}} = 10^{-7}$, yielding the final skeleton surface $\mathcal{B}$ of the template model (Figure 6.3 (center)).

The Muscle Surface

We generate the muscle surface $\mathcal{M}$ by minimizing the same energy as in Equation (6.1), but using a different method for finding the correspondences $\mathbf{t}_i$ in Equation (6.3), which exploits that we already established a dense correspondence between skin surface $\mathcal{S}$ and skeleton surface $\mathcal{B}$. We do not employ closest point correspondences, but instead set for each vertex $\mathbf{x}_i$ a *fixed* correspondence $\mathbf{t}_i$ to the first intersection of the line from skin vertex $\mathbf{x}_i^{\mathcal{S}}$ to skeleton vertex $\mathbf{x}_i^{\mathcal{B}}$ with the high-resolution muscle model [Zyg24]. If there is no intersection (e.g., at the knee), we set $\mathbf{t}_i = \mathbf{x}_i^{\mathcal{B}}$ and assign a lower weight w_i. When the minimization converges and we decrease λ_{reg}, we project the vertices of the current muscle surface $\mathbf{x}_i^{\mathcal{M}}$ to their corresponding skin-to-skeleton line from $\mathbf{x}_i^{\mathcal{S}}$ to $\mathbf{x}_i^{\mathcal{B}}$. Due to the collision handling, the resulting muscle surface $\mathcal{M}$ will enclose the high-resolution muscle model. To ensure that our volumetric elements always have a non-zero volume, even in regions where there is no muscle between skin and bone, we ensure a minimal offset of 1 mm from to the skeleton mesh. The resulting muscle surface $\mathcal{M}$ is visualized in Figure 6.3 (right). Note that the muscle layer does not exclusively contain muscles: Especially in the abdominal region, a large amount of the muscle layer is filled by organs. We therefore define a *muscle thickness map* that, for each vertex i, stores the accumulated length of the segments of the line $(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_i^{\mathcal{B}})$ that are covered by muscles. This map will be used later in Section 6.2.4.

6.2.3 Estimating Fat Mass and Muscle Mass

Having generated the volumetric layered template, we want to be able to fit it to a given surface scan of a person. To regularize this under-determined problem, we first have to estimate how much of the person's soft tissue is explained by fat mass (FM) and muscle mass (MM), respectively. This is a challenging task since we want to capture a single surface scan of the person only and therefore cannot rely on information provided by additional hardware, such as a DXA

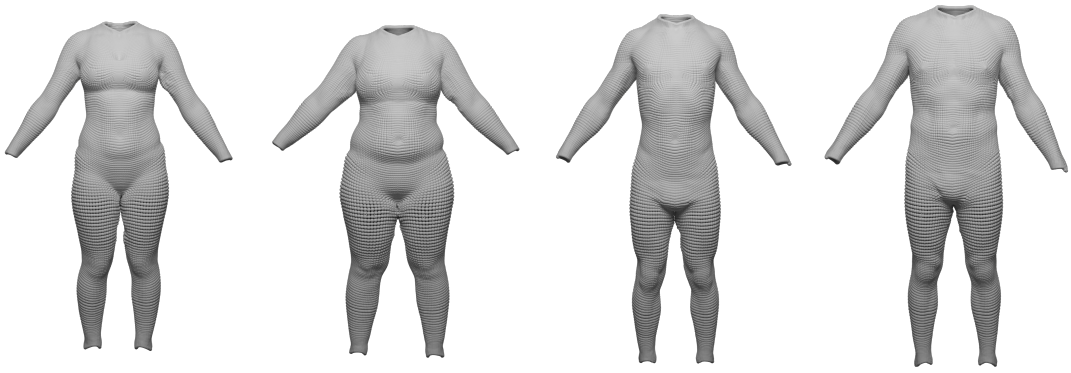

Figure 6.6: Examples for the BeyondBMI dataset provided by Maalin et al. [MMK+21] consisting of scans of 100 men and 100 women, annotated with fat mass, muscle mass, and BMI. The scans lack geometric data for head, hands, and feet and are captured in approximate A-pose (with noticeable variation in pose).

scanner or a body fat scale. Kadleček et al. [KIL+16] handle this problem by describing the person's shape primarily through muscles, i.e., by growing muscles as much as possible and defining the remaining soft tissue volume as fat. This strategy results in adipose persons having considerably more muscle mass than leaner people. Although there is a certain correlation between total body mass (and also BMI) and muscle mass – because the higher weight has a training effect especially on the muscles of the lower limbs [TEM+16] – this general trend is not sufficient to define the body composition of people.

Maalin et al. [MMK+21] measured both FM and MM using a medical-grade eight-electrode bioelectrical impedance analysis scale and acquired a 3D surface scan. From this data, they built a model that can vary the shape of a person based on specified muscle or fat variation, similar to Piryankova et al. [PSR+14] and the body weight modification model described in Chapter 5. Our model should perform the inverse operation, i.e., estimate FM and MM from a given surface scan. We train our model on their BeyondBMI dataset (Section 6.2.1), which consists of scans of 100 men and 100 women captured in an approximate A-pose (see Figure 6.6), each annotated with FM, MM, and BMI.

By applying the surface fitting process described in Section 6.2.1 to the BeyondBMI dataset, we make their scans compatible to our template and unpose their scans to a common T-pose, thereby making any subsequent statistical analysis pose-invariant. After re-excluding the head, hands, and feet of our surface template, we are left with $M = 100$ meshes per sex that consist of $V = 7665$ vertices $\mathbf{x}_i$. We denote the j^{th} training mesh by a $3V$-dimensional vector of stacked vertex coordinates $\mathbf{X}_j = \left(\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_V^\mathsf{T}\right)^\mathsf{T} \in \mathbb{R}^{3V}$ and the mean of the data set by $\bar{\mathbf{X}}$. We then perform PCA of the mean-centered data matrix $\mathbf{X} = \left(\mathbf{X}_1 - \bar{\mathbf{X}}, \ldots, \mathbf{X}_M - \bar{\mathbf{X}}\right) \in \mathbb{R}^{3V \times M}$. Let $\mathbf{U} \in \mathbb{R}^{3V \times k}$ be the basis of the subspace spanned by the first k principal components. Since the data

is now pose-normalized, the dimensionality reduction can focus solely on differences in human body shape. As a result, our model only needs $k = 12$ PCA components to explain 99.5 % of the data variance, while the original BeyondBMI dataset needs $k = 24$ components to cover the same percentage due to noticeable variations in pose during the scanning process (see Figure 6.6). We then perform linear regression to estimate FM and MM from PCA weights, as proposed by Hasler et al. [HSS+09]. Given a skin surface, this allows us to estimate fat and muscle mass by fitting our surface template model, excluding head, hands, and feet, projecting the resulting surface mesh into PCA space, and then applying the learned regressors.

For a first evaluation of this model, we perform a leave-one-out test on the BeyondBMI dataset, i.e., excluding each scan once, building the regressors as described above from the remaining $M - 1$ scans, and measuring the mean absolute error of the predictions. We again use $k = 12$ PCA components, as this covers almost all the variance present in the dataset and gives the linear regression enough degrees of freedom. The leave-one-out evaluation yields a mean absolute error (MAE) of $\text{MAE}_{\text{FM}} = 1.20\,\text{kg}$ ($SD = 0.93$) and $\text{MAE}_{\text{MM}} = 1.01\,\text{kg}$ ($SD = 0.79$) for the female dataset, where the fat mass lies in the range 6.27 – 34.71 kg and the muscle mass in the range 21.59 – 31.63 kg. The linear regression shows an average R^2 score of 0.84, confirming that there is indeed a linear relationship between PCA coordinates and the FM/MM measurements. Performing the leave-one-out test on the male dataset shows similar values: $\text{MAE}_{\text{FM}} = 1.37\,\text{kg}$ ($SD = 1.00$) and $\text{MAE}_{\text{MM}} = 1.46\,\text{kg}$ ($SD = 1.11$), fat mass in the range 3.91 – 27.83 kg, muscle mass in the range 31.51 – 51.20 kg, and an average R^2 score of 0.88.

We compared the linear model to a support vector regression (using scikit-learn [PVG+11] with default parameters and RBF kernels), but in contrast to Hasler et al. [HSS+09] we found that for the BeyondBMI dataset this approach performs considerably worse: $\text{MAE}_{\text{FM}} = 2.98\,\text{kg}$ ($SD = 2.85$) and $\text{MAE}_{\text{MM}} = 1.24\,\text{kg}$ ($SD = 1.02$) with an average R^2 score of 0.64 for the female dataset, and $\text{MAE}_{\text{FM}} = 2.63\,\text{kg}$ ($SD = 2.60$) and $\text{MAE}_{\text{MM}} = 2.48\,\text{kg}$ ($SD = 1.82$) with an average R^2 score of 0.58 for the male dataset. We therefore keep the simpler and better-performing linear regression model.

Whenever we fit the volumetric model to a given body scan, as explained in the next section, we first use the proposed linear regressors to estimate the person's fat mass and muscle mass and use this information to generate the muscle and fat layers in Section 6.2.4.

6.2.4 Fitting the Volumetric Template to Surface Scans

Given a surface scan, we transfer the template anatomy into it through the following steps: First, we fit our *surface* template to the scan, which establishes

one-to-one correspondence with the volumetric template and puts the scan into the same T-pose as the template (Section 6.2.1). After this preprocessing, we deform the *volumetric* template to match the scanned subject. To this end, we adjust global scaling and per-bone local scaling, such that body height and limb lengths of template and scan match. This is followed by a quasi-static deformation of the volumetric template that considers the skin surface $\mathcal{S}$ as hard constraint and yields the skeleton surface $\mathcal{B}$ through energy minimization. Given the skin surface $\mathcal{S}$, the bone surface $\mathcal{B}$, and the estimated fat mass and muscle mass from Section 6.2.3, the muscle surface $\mathcal{M}$ is determined. Having transferred all three layer surfaces to the scan we finally warp the detailed anatomical model to the target.

Global and Local Scaling

Fitting the surface template to the scanner data puts the latter into the same alignment (rotation, translation) and the same pose as the volumetric template. The next step is to correct the mismatch in scale by adjusting body height and limb lengths of the volumetric template.

This scaling does influence all three of the template's surfaces. Since the shape of the skeleton surface $\mathcal{B}$ will be constrained to the result after scaling, we have to scale in a way that keeps bone lengths and bone diameters within a plausible range. The *length* of prominent bones, like the upper arm or the upper leg (humerus and femur), can be approximated by measures on the surface of the model. But finding the correct bone *diameters* is impossible without measurements of the subject's interior. In particular for corpulent or adipose subjects, the subcutaneous fat layer dominates the appearance of the skin surface, preventing us from precisely determining the bone diameters from the surface scan. It has been shown that there is a moderate correlation of bone length and bone diameter [ABY+17; ZM02] and (obviously) a strong correlation of body height and bone length [DSK08]. We therefore perform a *global isotropic* scaling depending on body height (affecting bone lengths and diameters) as well as *local anisotropic* scaling depending on limb lengths (affecting bone lengths only).

The global scaling is determined from the height difference of scan and template and is applied to all vertices of the template model. It therefore scales all bone lengths and bone diameters uniformly. Directly scaling with the height ratio of scan and template, however, can result in bones too thin or too thick for extreme target heights. Thus, we damp the height ratio $r = h_{\text{scan}} \,/\, h_{\text{template}}$ by $r \leftarrow 0.5(r-1)+1$, which means that a person that is 20 % taller than the template will have 10 % thicker bones than the template. This heuristic results in visually plausible bone diameters for all our scanned subjects.

After the global scaling, the local scaling further adjusts the limb lengths of the template to match those of the scan. The (fully rigged) surface-based tem-

Figure 6.7: Scaling the template (opaque) to match the scan (semi-transparent): The preprocessing aligns the scan with the template and puts it into the same pose (left). Body height and limb lengths of the template are then adjusted by a global uniform scaling (center), followed by local scaling for limbs and spine (right).

plate has been fit to both the scan (Section 6.2.1) and the template (Section 6.2.2). This fit provides a simple skeleton graph (used for skinning animation) for both models. We use the length mismatch of the respective skeleton graph segments to determine the required scaling for upper and lower arms, upper and lower legs, feet, and torso. We scale these limbs in their corresponding bone directions (or the spine direction for the torso) using the bone stretching of Kadleček et al. [KIL+16]. As mentioned before, this changes the limb lengths but not the bone diameters.

This two-step scaling process is visualized in Figure 6.7. As a result, the scaled template matches the scan with respect to alignment, pose, body height, and limb lengths. Its layer surfaces, which we denote by $\bar{\mathcal{S}}$, $\overline{\mathcal{M}}$, and $\bar{\mathcal{B}}$, provide a good initialization for the optimization-based fitting described in the following.

Skeleton Fitting

Given the coarse registration of the previous step, we now fit the skin surface $\mathcal{S}$ and skeleton surface $\mathcal{B}$ by minimizing a quasi-static deformation energy. Since the template's skin surface $\mathcal{S}$ should match the (skin) surface of the scan and since both meshes have the same triangulation, we can simply copy the skin vertex positions from the scan to the template and consider them as hard Dirichlet constraints. It therefore remains to determine the vertex positions of the skeleton surface $\mathcal{B}$, such that the soft tissue enclosed between skin surface $\mathcal{S}$ and skeleton surface $\mathcal{B}$ (composed of fat and muscle tissue, which we denote by *flesh*) deforms in a physically plausible manner. This is achieved by minimizing a quasi-static energy consisting of three terms:

$$E(\mathcal{B}) \;=\; \lambda_{\text{reg}} E_{\text{reg}}\big(\mathcal{B}, \bar{\mathcal{B}}\big) + \lambda_{\text{flesh}} E_{\text{flesh}}(\mathcal{B}, \mathcal{S}) + \lambda_{\text{coll}} E_{\text{coll}}(\mathcal{B}, \mathcal{S})\,. \qquad (6.5)$$

The first term is responsible for keeping the skeleton surface (approximately) rigid and uses the same formulation as Equation (6.2), with $\bar{\mathcal{B}}$ and $\mathcal{B}$ denoting the skeleton surface before and after the deformation, respectively. We employ a soft constraint with high weight λ_{reg} instead of deforming bones in a strictly rigid manner [KIL+16], since we noticed that for very thin subjects the skeleton surface might otherwise protrude the skin surface and therefore a certain amount of bone deformation is required. We also do not penalize deviation from rigid or affine transformations as proposed by Dicko et al. [DLG+13], because this penalizes smooth shape deformation in the same way as locally flipped triangles, which we observed to cause artifacts in the skeleton surface. The discrete bending energy of Equation (6.2), with a suitably high regularization weight λ_{reg}, allows for moderate *smooth* deformations and gave better results in our experiment.

The second term prevents strong deformations of the prism elements $p \in \mathbb{P}$, spanned by corresponding triangles $(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_j^{\mathcal{S}}, \mathbf{x}_k^{\mathcal{S}})$ on the skin surface and $(\mathbf{x}_i^{\mathcal{B}}, \mathbf{x}_j^{\mathcal{B}}, \mathbf{x}_k^{\mathcal{B}})$ on the skeleton surface. While we penalize deformation of the top/bottom triangles, we allow changes of prism heights, i.e., anisotropic scaling in the direction from surface to bone. Otherwise, the fat layer cannot grow to bridge the gap from the skeleton surface to the skin surface. This behavior is modeled by the anisotropic strain limiting energy

$$E_{\text{flesh}}(\mathcal{B}, \mathcal{S}) \;=\; \frac{1}{2} \sum_{p \in \mathbb{P}} \left\| \mathbf{F}_p - \mathbf{R}_p \mathbf{B}_p \tilde{\mathbf{S}}_p \mathbf{B}_p^{\mathsf{T}} \right\|_F^2, \tag{6.6}$$

where $\mathbf{F}_p \in \mathbb{R}^{3\times 3}$ is the deformation gradient of the element p, i.e., the linear part of the best affine transformation that maps the undeformed prism $\bar{p}$ to the deformed prism p in the least squares sense. If $\mathbf{E}_p \in \mathbb{R}^{3\times 5}$ denotes the edge direction matrix of the prism p and $\bar{\mathbf{E}}_p$ the respective matrix of $\bar{p}$, then $\mathbf{F}_p = \arg\min_{\mathbf{F}} \left\| \mathbf{E}_p - \mathbf{F}\bar{\mathbf{E}}_p \right\|_F^2$. Polar decomposition $\mathbf{F}_p = \mathbf{R}_p \mathbf{S}_p$ decomposes $\mathbf{F}_p$ into a rotation $\mathbf{R}_p$ and a scale/shear $\mathbf{S}_p$ [SD92]. $\mathbf{B}_p$ is a rotation matrix that aligns the z-axis with the surface normal of the prism's corresponding skin triangle, i.e., the direction in which we allow stretching. The matrix $\tilde{\mathbf{S}}_p = \operatorname{diag}(1, 1, \alpha)$ represents the anisotropic scaling, where $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ allows to tune the amount of stretching in normal direction that should be allowed. We use $\alpha_{\min} = 0.2$ and $\alpha_{\max} = 5.0$ to allow stretching and compression of the element by a factor of five before the energy of this element increases.

Third, we detect all collisions $\mathcal{C}_{\text{coll}}$, defined as vertices of the skeleton surface $\mathcal{B}$ that are outside of the skin surface $\mathcal{S}$. For these colliding vertices we add a collision penalty term

$$E_{\text{coll}}(\mathcal{B}, \mathcal{S}) \;=\; \frac{1}{|\mathcal{C}_{\text{coll}}|} \sum_{\mathbf{x}_i \in \mathcal{C}_{\text{coll}}} w_i \left\| \mathbf{x}_i - \boldsymbol{\pi}_{\mathcal{S}}(\mathbf{x}_i) \right\|^2, \tag{6.7}$$

where $\pi_{\mathcal{S}}(\mathbf{x}_i)$ is the projection of the colliding vertex $\mathbf{x}_i$ to a position 2 mm beneath the closest triangle on the skin surface $\mathcal{S}$. The vertex weight w_i is set to 1 the first time a vertex is colliding, and is increased by 1 each time the minimization was not able to resolve the collision.

The iterative minimization of Equation (6.5) is again implemented via the Projective Dynamics framework. In order to determine the current amount of stretching needed for evaluating the anisotropic strain limiting energy (6.6), we compute the deformation gradient $\mathbf{F}_p$ from the undeformed prism $\bar{p}$ to its current state p, its polar decomposition $\mathbf{F}_p = \mathbf{R}_p\mathbf{S}_p$, and the resulting stretching in skin normal direction by $\alpha = (\mathbf{B}_p^\mathsf{T}\mathbf{S}_p\mathbf{B}_p)_{3,3}$. The amount of stretching is then clamped to the range $[\alpha_{\min}, \alpha_{\max}]$ before computing the anisotropic stretching matrix $\tilde{\mathbf{S}}_p = \operatorname{diag}(1, 1, \alpha)$. We set the energy term coefficients to $\lambda_{\text{reg}} = 0.1$, $\lambda_{\text{flesh}} = 0.01$, and $\lambda_{\text{coll}} = 50$, and iteratively minimize Equation (6.5) until convergence. In the converged state, we then detect collisions and add the corresponding collision constraints (see Equation (6.7)) to the system. Subsequently, minimizing Equation (6.5) and detecting collisions is repeated until no collisions are found in a converged solution. For all of our subjects, the minimization converged within less than 20 iterations.

Muscle Fitting

Having determined the skin surface $\mathcal{S}$ and skeleton surface $\mathcal{B}$, we now fit the muscle surface $\mathcal{M}$ in between $\mathcal{S}$ and $\mathcal{B}$, such that the ratio of fat mass (FM) and muscle mass (MM) resembles the values estimated by our regressors (Section 6.2.3). We proceed in three steps: First, we transfer the template's muscle distribution to the fitted skin and skeleton surfaces, which we call *average muscle layer* in the following. Second, we grow and shrink the muscles as much as anatomically and physically plausible, yielding the *minimum* and *maximum muscle layers*. Third, we find a linear interpolation between these two extremes that matches the predicted fat mass and muscle mass as good as possible.

The average muscle surface is transferred from the *scaled* template $\overline{\mathcal{M}}$ (Section 6.2.4, Figure 6.7) by minimizing an energy consisting of two objectives:

$$E(\mathcal{M}) \;=\; \lambda_{\text{reg}} E_{\text{reg}}\big(\mathcal{M}, \overline{\mathcal{M}}\big) + \lambda_{\text{line}} E_{\text{line}}(\mathcal{M}, \mathcal{B}, \mathcal{S})\,. \tag{6.8}$$

The first term tries to preserve the shape of the scaled template's muscle surface $\overline{\mathcal{M}}$ and is modeled using the regularization energy of Equation (6.2). The second term preserves the template's property that each muscle vertex $\mathbf{x}_i^{\mathcal{M}}$ resides on the line segment from its corresponding skeleton vertex $\mathbf{x}_i^{\mathcal{B}}$ to its skin vertex $\mathbf{x}_i^{\mathcal{S}}$, by penalizing the squared distance from that line:

$$E_{\text{line}}(\mathcal{M}, \mathcal{B}, \mathcal{S}) \;=\; \frac{1}{2} \sum_{\mathbf{x}_i \in \mathcal{M}} \left\| \mathbf{x}_i - \pi_{\text{line}}\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{B}}, \mathbf{x}_i^{\mathcal{S}}\right) \right\|^2, \tag{6.9}$$

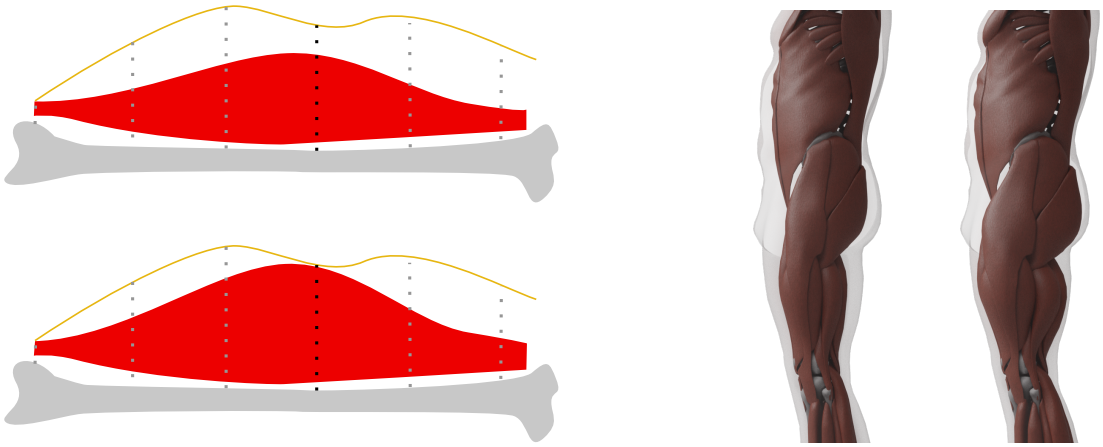

Figure 6.8: Left: When computing the maximum muscle surface, we move muscle vertices toward the skin by an amount proportional to their *muscle potential*, which for each vertex is the length of the dotted line intersected with the muscle. The vertex with the black dotted line defines the maximum allowed stretch in this example. Right: An example of our minimum and maximum muscle layers for the same target. These two surfaces define the lower and upper limit for the muscle mass and vice versa for the fat mass.

where $\pi_{\text{line}}(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{B}}, \mathbf{x}_i^{\mathcal{S}})$ is the projection of $\mathbf{x}_i$ onto the line $(1-\beta)\mathbf{x}_i^{\mathcal{B}} + \beta\mathbf{x}_i^{\mathcal{S}}$, $\beta \in [0,1]$. Minimizing Equation (6.8) leads to flat abdominal muscles as in the template model, which is unrealistic for corpulent or adipose subjects, because the majority of body fat resides in two different fat tissues: the *subcutaneous fat*, which resides between skin and muscle surface, and the *visceral fat*, which accumulates in the abdominal cavity, i.e., under the muscle layer. Since the bulging of the abdomen due to visceral fat causes a bulging of the belly, we inversely want the abdominal muscles in $\mathcal{M}$ to slightly bulge out in case of a belly bulge in the skin surface $\mathcal{S}$. The latter is a combined effect of visceral and subcutaneous fat in the abdominal region. We model this effect by adjusting E_{line} for each vertex $\mathbf{x}_i$ in the abdominal region. Instead of using the full interval $\beta \in [0,1]$, we adjust the lower boundary to $\beta_{\min} = \left\|\bar{\mathbf{x}}_i^{\mathcal{M}} - \bar{\mathbf{x}}_i^{\mathcal{B}}\right\| / \left\|\bar{\mathbf{x}}_i^{\mathcal{S}} - \bar{\mathbf{x}}_i^{\mathcal{B}}\right\|$, i.e., the parameter β where for the (scaled) template the muscle surface intersects the line.

Equation (6.8) is iteratively minimized via the Projective Dynamics framework. We initialize $\mathcal{M}$ with $\overline{\mathcal{M}}$ and set $\lambda_{\text{reg}} = 0.01$, $\lambda_{\text{line}} = 1.0$. When the minimization converges, we update the Laplacians in E_{reg} to those of the current solution and decrease λ_{reg} by a factor of 0.5. This is iterated until the maximal distance of a vertex to its bone-to-skin line (see Equation (6.9)) is less than 0.2 mm. Lastly, we project each vertex onto its corresponding bone-to-skin line to get a perfect alignment.

After transferring the average muscle surface, we grow/shrink muscles as much as possible in order to define the maximum/minimum muscle surfaces (Figure 6.8). Since certain muscle groups may be better developed than others,

we perform the muscle growth/shrinkage separately for the major muscle groups, namely upper legs (including buttocks), lower legs, upper arms, lower arms, chest, abdominal muscles, shoulders, and back. Muscles are built from fibers and grow perpendicular to the fiber direction. In all cases relevant to us, the fibers are approximately perpendicular to the direction from $\mathcal{M}$ to $\mathcal{S}$, thus muscle growth/shrinkage will move vertices $\mathbf{x}_i^{\mathcal{M}}$ along the line from $\mathbf{x}_i^{\mathcal{B}}$ to $\mathbf{x}_i^{\mathcal{S}}$. The amount of vertex movement along these directions is proportional to the muscle thickness map of the template (computed in Section 6.2.2). We determine how much we can grow a muscle before it collides with the skin surface in the thicker parts of the muscle (instead of close to its endpoints where it connects to the bone). Figure 6.8 shows an example, where the leftmost muscle vertex is already close to the skin and would prevent any growth if we took endpoint regions into account. For each muscle group, we also define an upper limit for muscle growth that prevents the muscles from increasing further even if the distance to the skin is still large (e.g., for adipose subjects). To determine the minimal muscle surface, we repeat the process in the opposite direction (towards the skeleton surface). To prevent distortions of the muscle surface, we do not set the new vertex positions directly, but instead use them as target positions $\mathbf{t}_i$ in Equation (6.3) and regularize with Equation (6.2). Figure 6.8 (right) shows an example of minimum/maximum muscle surfaces computed by this procedure.

We determine the final muscle surface $\mathcal{M}$ by linear interpolation between the minimum and maximum muscle surfaces, such that the resulting fat mass FM and muscle mass MM match the values predicted by the regressors (denoted by FM* and MM*) as good as possible. To this end, we have to compute FM and MM from an interpolated muscle surface $\mathcal{M}$. We can compute the volume V_{FL} of the fat layer (between $\mathcal{S}$ and $\mathcal{M}$) and the volume V_{ML} of the muscle layer (between $\mathcal{M}$ and $\mathcal{B}$) and convert these to masses m_{FL} and m_{ML} by multiplying with the (approximate) fat and muscle densities $\rho_{\mathrm{F}} = 0.9\,\mathrm{kg/l}$ and $\rho_{\mathrm{M}} = 1.1\,\mathrm{kg/l}$, respectively.

The resulting masses require some corrections though: First, we have to add the visceral fat (VAT), which is not part of our fat layer but resides in the abdominal cavity. We estimate the VAT mass m_{VAT} by computing the difference of the cavity volumes of the scaled template and of the final fit, thereby assuming a negligible amount of VAT in the template. Second, we subtract the skin mass m_{skin} from the fat layer mass. We assume an average skin thickness of 2 mm, multiply this by the skin's surface area and the density ρ_{F}. Third, our fat layer includes the complete reproductive apparatus in the crotch region. This volume is even larger due to the underwear that was worn during scanning and incorrectly increases the fat layer mass by m_{crotch}. Our corrected fat mass is then

$$\mathrm{FM} = m_{\mathrm{FL}} + m_{\mathrm{VAT}} - m_{\mathrm{skin}} - m_{\mathrm{crotch}}. \tag{6.10}$$

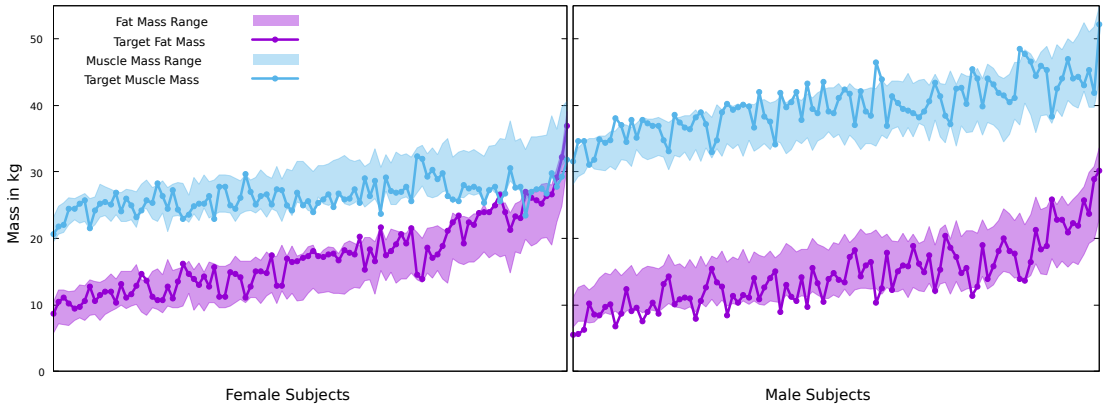

Figure 6.9: True muscle and fat masses for the female and male subjects of the BeyondBMI dataset, plotted on top of the possible ranges defined by our minimum and maximum muscle surfaces. Note that our minimal fat mass is coupled to the maximal muscle mass and vice versa.

We correct the muscle mass by subtracting the mass m_{abd} of the abdominal cavity, which is incorrectly included in the muscle layer. The remaining muscle mass is always too small even when using the maximum muscle surface, due to all muscles not considered in the muscle layer, such as heart, face, and hand muscles or the diaphragm. It is known that the lean body mass roughly scales with the squared body height [HHT+11], which is the basis of the well known body and muscle mass indices. We analogously assume the missing muscle mass to be proportional to the squared height h of the subject, i.e., $m_h = c_h h^2$, with a constant c_h to be determined later. The corrected muscle mass is therefore

$$\text{MM} = m_{\text{ML}} - m_{\text{abd}} + m_h. \tag{6.11}$$

There are other terms like the fat of the head, hands, and toes, which could be added, or the volume of blood vessels and tendons, which could be subtracted. We assume those terms to be negligible.

Since the total volume of the soft tissue layer $V_{\text{ST}} = V_{\text{ML}} + V_{\text{FL}}$ is constant, the muscle layer mass m_{ML} is coupled to the fat layer mass m_{FL} via $m_{\text{ML}} = (V_{\text{ST}} - V_{\text{FL}})\,\rho_{\text{M}}$. We want to compute the fat layer mass such that the resulting FM and MM minimize the least squares error to the values predicted by the regressor: $E = (\text{FM} - \text{FM}^*)^2 + (\text{MM} - \text{MM}^*)^2$. Inserting (6.10) and (6.11) into E, rewriting m_{ML} in terms of m_{FL}, and setting the derivative $\mathrm{d}E/\mathrm{d}m_{\text{FL}} = 0$ yields the optimal fat layer mass

$$m_{\text{FL}} = \frac{\text{FM}^* - m_{\text{VAT}} + m_{\text{skin}} + m_{\text{crotch}} + \rho\,(V_{\text{ST}}\,\rho_{\text{M}} - m_{\text{abd}} + m_h - \text{MM}^*)}{1 + \rho^2}, \tag{6.12}$$

with the density ratio $\rho = \rho_{\text{M}} \,/\, \rho_{\text{F}}$.

The minimum/maximum muscle surface yields a maximum/minimum fat layer mass. The optimized fat layer mass is clamped to this range, thereby

defining the final fat layer mass. We then choose the linear interpolant between the minimum and maximum muscle surface that matches this fat mass, which we find through bisection search.

We did this for the scans of 100 men and 100 women from the BeyondBMI dataset [MMK+21], where we know the true values for FM and MM from measurements, and optimized the value of c_h for this dataset, yielding $c_h = 1.5$ for the male and $c_h = 1.0$ for the female dataset. This is plausible since women generally have a lower muscle mass. For instance, the average muscle mass of the male subjects in the dataset is indeed 50 % higher than the average MM for the female subjects. The mean absolute errors (MAE) for the BeyondBMI dataset are $\mathrm{MAE}_{\mathrm{MM}} = 0.37\,\mathrm{kg}$ ($SD = 0.31$), $\mathrm{MAE}_{\mathrm{FM}} = 0.46\,\mathrm{kg}$ ($SD = 0.38$) for the female subjects and $\mathrm{MAE}_{\mathrm{MM}} = 0.46\,\mathrm{kg}$ ($SD = 0.39$), $\mathrm{MAE}_{\mathrm{FM}} = 0.57\,\mathrm{kg}$ ($SD = 0.48$) for the male subjects. Figure 6.9 shows how well our model can adjust to the target values of muscle and fat mass. All values are inside or at least close to the predicted possible range of minima and maxima. Moreover, in most cases, the muscle/fat mass values for the same person split the two ranges at about an inverse point (e.g., close to maximum muscle and close to minimum fat), resulting in the low errors stated above.

Transferring Original Anatomical Data

After fitting the skin surface $\mathcal{S}$ to the scan and transferring the skeleton surface $\mathcal{B}$ and the muscle surface $\mathcal{M}$ into the scan, the final step is to transform the high-resolution anatomical details (Zygote's bone and muscle models in our case) from the volumetric template to the scanned subject. We implement this in an efficient and robust manner as a mesh-independent space warp $\mathbf{d}\colon \mathbb{R}^3 \to \mathbb{R}^3$ that maps the original template's skin surface $\hat{\mathcal{S}}$, muscle surface $\hat{\mathcal{M}}$, and skeleton surface $\hat{\mathcal{B}}$ (all marked with a hat) to the scanned subject's layer surfaces $\mathcal{S}$, $\mathcal{M}$, and $\mathcal{B}$, respectively. All geometry that is embedded in between these surfaces will smoothly be warped from template to scan.

Dicko et al. [DLG+13] also employ a space warp for their anatomy transfer, which they discretized by interpolating values $\mathbf{d}_{ijk}$ on a regular 3D grid constructed around the object. Their space warp is computed by interpolating the skin deformation $\hat{\mathcal{S}} \mapsto \mathcal{S}$ on the boundary and being harmonic in the interior (i.e., $\Delta\mathbf{d} = 0$), which requires the solution of a large sparse Poisson system for the coefficients $\mathbf{d}_{ijk}$. We follow the same idea, but use a space warp based on triharmonic radial basis functions (RBFs) [BK05], which have been shown to yield higher quality deformations with lower geometric distortion than many other warps (including FEM-based harmonic warps) [SMB13]. The RBF warp is defined as a sum of n RBF kernels and a linear polynomial:

$$\mathbf{d}(\mathbf{x}) \;=\; \sum_{j=1}^{n} \mathbf{w}_j \varphi_j(\mathbf{x}) + \sum_{k=1}^{4} \mathbf{q}_k \pi_k(\mathbf{x})\,, \tag{6.13}$$

where $\mathbf{w}_j \in \mathbb{R}^3$ is the coefficient of the j^{th} radial basis function defined by $\varphi_j(\mathbf{x}) = \varphi(\|\mathbf{x} - \mathbf{c}_j\|)$ and centered at $\mathbf{c}_j \in \mathbb{R}^3$. As kernel function we use $\varphi(r) = r^3$, leading to highly smooth triharmonic warps ($\Delta^3 \mathbf{d} = 0$). The second term is a linear trivariate polynomial with basis $\{\pi_1, \pi_2, \pi_3, \pi_4\} = \{x, y, z, 1\}$ and coefficients $\mathbf{q}_k \in \mathbb{R}^3$, which ensures linear precision of the warp.

In order to warp the high-resolution *bone model* from the template to the scan, we set up the RBF warp to reproduce the deformation $\hat{\mathcal{B}} \mapsto \mathcal{B}$. To this end, we select 5000 vertices $\hat{\mathbf{x}}_i \in \hat{\mathcal{B}}$ from the template's skeleton surface by farthest point sampling. The corresponding vertices on the scan's skeleton surface are denoted by $\mathbf{x}_i \in \mathcal{B}$. At these vertices $\hat{\mathbf{x}}_i$ the deformation function $\mathbf{d}(\hat{\mathbf{x}}_i)$ should interpolate the displacements $\mathbf{d}_i = \mathbf{x}_i - \hat{\mathbf{x}}_i$. These constraints lead to a dense, symmetric, but indefinite $(n+4) \times (n+4)$ linear system, which we solve for the coefficients $(\mathbf{w}_1, \ldots, \mathbf{w}_n, \mathbf{q}_1, \ldots, \mathbf{q}_4)$ using the LU factorization of Eigen [GJ+24]. For more details, please refer to the work of Sieger et al. [SMB13]. The resulting RBF warp $\mathbf{d}$ then transforms each vertex $\mathbf{x}$ of the high-resolution bone model as $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{d}(\mathbf{x})$. Note that this process can trivially be parallelized over all model vertices, which we implement using OpenMP. For warping the high-resolution *muscle model*, we follow the same procedure, but collect 7000 constraints from the vertices $\hat{\mathbf{x}}_i \in \mathcal{S} \cup \mathcal{M}$ of the skeleton and muscle surfaces, since these enclose the muscle layer.

6.3 RESULTS AND APPLICATIONS

To summarize, generating a personalized anatomical model for a given surface scan of a person consists of the following steps: First, the surface template is registered to the scanner data (triangle mesh or point cloud) with the template fitting method of Achenbach et al. [AWL+17] as described in Section 2.2.2. After manually selecting 10–20 landmarks, this process takes about 50 sec. Fitting the surface template establishes a dense correspondence with the surface of the volumetric template and puts the scan into the same T-pose as the volumetric template. Fitting the volumetric template by transferring the three layer surfaces (Section 6.2.4) takes about 15 sec. Transferring the high-resolution anatomical models of bones and muscles (145k vertices) takes about 4.5 sec for solving the linear system (which is an offline preprocessing) and 0.5 sec for transforming the vertices. Timings were measured on a desktop workstation, equipped with an Intel Core i9 10850K CPU and an Nvidia RTX 3070 GPU.

Dicko et al. [DLG+13] and Kadleček et al. [KIL+16] are the two approaches most closely related to ours. Dicko et al. [DLG+13] also use a space warp for transferring anatomical details, but since they only use the skin surface as constraint, the interior geometry can be strongly distorted. To prevent this, they restrict bones to affine transformations, which can still contain unnatural shearing modes and implausible scaling. Our space warp yields a higher

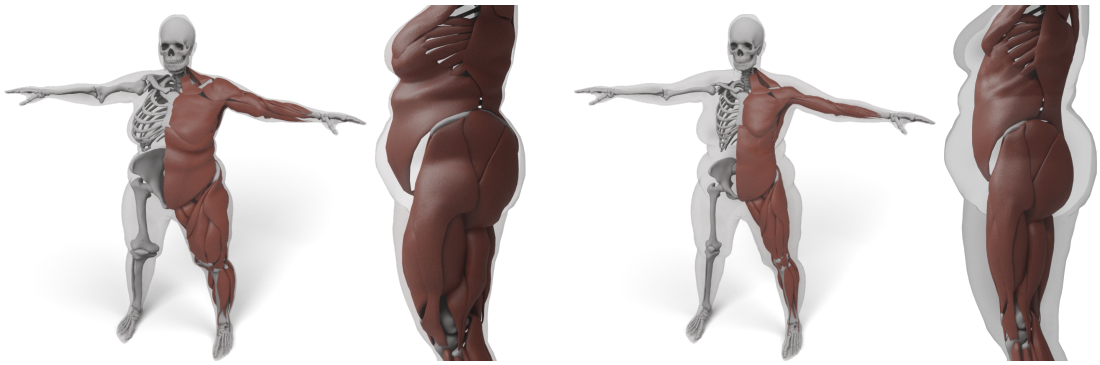

Figure 6.10: Result of transferring the anatomy by using just the skin layer and a harmonic basis (left). Here, both muscles and bones deform too much to fit overweight targets. We use the additional muscle and skeleton layer and a triharmonic basis (right) to prevent unnatural deformations.

smoothness due to the use of C^∞ RBF kernels instead of C^0 trilinear interpolation. In addition, it reduces unnatural distortion of bones and muscles by using three layer surfaces as constraints instead of the skin surface only and by optimizing these layers w.r.t. anatomical distortion. In Figure 6.10, we compare the result of warping the anatomical structures using a harmonic basis and 7000 kernels from only the skin surface to our three-layered, triharmonic warp result. The former shows drastic and unrealistic deformations of both muscles and bones, while our approach solves those issues. Note that additionally restricting the bones to affine transformations like Dicko et al. [DLG+13] would still produce unnaturally thick bones (e.g., the upper leg bone) and muscles.

Compared to Kadleček et al. [KIL+16], we only require a single input scan, since we infer (initial guesses for) joint positions and limb lengths from the full-body PCA of Achenbach et al. [AWL+17]. Putting the scan into T-pose prevents us from having to solve bone geometry and joint angles simultaneously, which makes our approach much faster than theirs (15 sec vs. 30 min). Moreover, our layered model yields a conforming volumetric tessellation with constant and homogeneous per-layer materials, which more effectively prevents bones from penetrating skin or muscles. In their approach the rib cage often intersects the muscle layer for thin subjects, which Kadleček et al. [KIL+16] mention as a limitation of their method. This effect can be seen in Figure 12 (bottom row) of their work. Another difference is that we automatically derive the muscle/fat body composition from the surface scan, which yields more plausible results than growing muscles as much as possible [KIL+16], since the latter leads to more corpulent people always having more muscles. Our model extracts the amount of muscle and fat using data of real humans and can therefore adapt to the variety of human shapes (low FM and high MM, high FM and low MM, and everything in between). Finally, we account for the differences in

male and female anatomy by employing individual anatomical templates and muscle/fat regressors for men and women.

6.3.1 Evaluation on Hasler Dataset

In order to further evaluate the generalization abilities of the linear FM/MM models (Section 6.2.3) to other data sources, we estimate FM and MM for a subset of registered scans from the *Hasler* dataset [HSS+09] and measure the prediction error. We selected scans of 10 men and 10 women, making sure to cover the extremes of the weight, height, fat, and muscle percentage distribution present in the data.

For the female sample, the predictions show a mean absolute error of $\text{MAE}_{\text{FM}} = 0.65\,\text{kg}$ ($SD = 0.44$) and $\text{MAE}_{\text{MM}} = 4.39\,\text{kg}$ ($SD = 1.71$). For the male sample, the model shows a similar error for the MM prediction, but performs worse at predicting FM: $\text{MAE}_{\text{FM}} = 3.32\,\text{kg}$ ($SD = 1.98$) and $\text{MAE}_{\text{MM}} = 4.14\,\text{kg}$ ($SD = 2.74$). Compared to the leave-one-out tests on the BeyondBMI data (Section 6.2.3), the average error increases noticeably, which can partly be explained by differences in the measurement procedure between the two datasets: While Hasler et al. [HSS+09] used a consumer-grade body fat scale, Maalin et al. [MMK+21] used a medical-grade scale, which should lead to more accurate measurements. Nevertheless, these results show that our regressor generalizes well to other data sources, providing a simple and sufficiently accurate method for estimating FM and MM from body scans.

Given the target FM and MM values for a subject as predicted by our regressor, we choose the optimal muscle surface between the minimal and maximal muscle surface as explained in Section 6.2.4. Comparing the final FM and MM of the volumetric model to the ground truth measurements of the *Hasler* dataset we get end-to-end errors of $\text{MAE}_{\text{FM}} = 0.70\,\text{kg}$ ($SD = 0.52$), $\text{MAE}_{\text{MM}} = 4.19\,\text{kg}$ ($SD = 1.39$) (female) and $\text{MAE}_{\text{FM}} = 3.49\,\text{kg}$ ($SD = 2.02$), $\text{MAE}_{\text{MM}} = 3.81\,\text{kg}$ ($SD = 2.56$) (male). This evaluation shows that the additional error induced by fitting the muscle layer is very low.

6.3.2 Evaluation on CAESAR Dataset

In order to demonstrate the flexibility and robustness of our method, we evaluate it by generating anatomical models for all scans of the European subset of the CAESAR data set [RBD+02]. This subset consists of 919 scans of women and 777 scans of men, with a height range of 131 – 218 cm for men and 144 – 195 cm for women (we only considered scans with complete annotation and taken in standing pose). A few examples for men and women can be seen in Figure 6.11 and Figure 6.12.

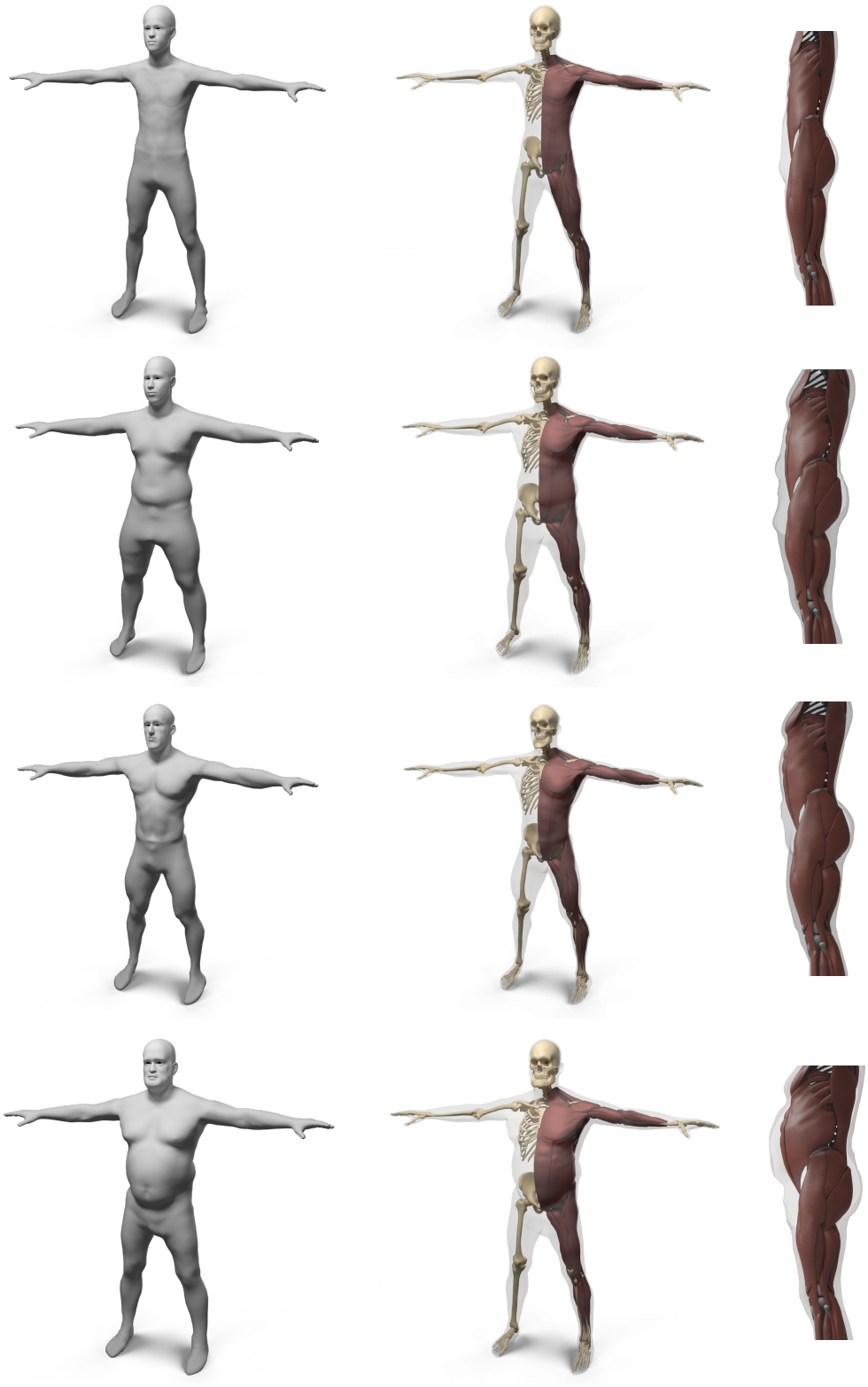

Figure 6.11: Some examples for various male body shape types. For each input surface the transferred muscles and skeleton are shown in front and side view.

Figure 6.12: Some examples for various female body shape types. For each input surface the transferred muscles and skeleton are shown in front and side view.

For the about 1700 CAESAR scans, our muscle and fat mass regressors yield just one slightly negative value for the fat mass of the thinnest male (body weight 48 kg, height 1.72 m, BMI 16.14 kg/m^2). For all other subjects, we get values ranging from 3.5 – 38.9 % body fat ($M = 20.3$ %) for male subjects and 8.0 – 45.3 % ($M = 28.9$ %) for female subjects. The range of predicted muscle masses is 24.9 – 57.8 kg (men) and 20.1 – 37.7 kg (women). When determining the optimal interpolation between the minimum and maximum muscle layer (Section 6.2.4), we meet the estimated target values up to mean errors $\text{MAE}_{\text{FM}} = 1.08$ kg ($SD = 0.90$) and $\text{MAE}_{\text{MM}} = 0.88$ kg ($SD = 0.74$) for the male dataset, and $\text{MAE}_{\text{FM}} = 1.41$ kg ($SD = 1.35$) and $\text{MAE}_{\text{MM}} = 1.15$ kg ($SD = 1.11$) for the female dataset. Note that even the scan with predicted negative FM is reconstructed robustly. In this case, the muscle surface will be the maximum muscle surface, which in general is a suitable estimate for very skinny subjects.

The CAESAR dataset does not include ground truth data for fat and muscle mass of the scanned individuals, which prevents us from directly evaluating our estimations of fat and muscle mass for the CAESAR scans. Thus, in order to further evaluate the plausibility of our estimated body composition, we compare it to known body fat percentiles. Percentiles are used as guidelines in medicine and provide statistical reference values that individual measurements can be compared to. For instance, a 10$^{\text{th}}$ percentile of 20.8 % body fat means that 10 % of the examined population have a body fat percentage less than 20.8 %. Assuming that the European CAESAR dataset is a representative sample of the population, the percentiles we get from our reconstructions of the CAESAR scans should match the percentiles of the European population. We compared the values produced by our fat and muscle mass regressors (Section 6.2.3) to Kyle et al. [KGS+01], who measured body fat using 4-electrode bioelectrical impedance analysis from 2735 male and 2490 female western European adults. Our body fat percentiles on the CAESAR dataset are very well in agreement with their results, as shown in Table 6.1.

	Body Fat Percentile	5$^{\text{th}}$	10$^{\text{th}}$	25$^{\text{th}}$	50$^{\text{th}}$	75$^{\text{th}}$	90$^{\text{th}}$	95$^{\text{th}}$
Male	Our estimate	10.2	12.3	16.0	20.3	24.6	28.1	30.7
	Kyle et al. [KGS+01]	10.9	12.6	15.7	19.2	23.5	27.0	29.2
Female	Our estimate	18.6	21.1	24.7	28.5	33.7	37.4	39.3
	Kyle et al. [KGS+01]	18.5	20.8	23.8	28.1	32.6	37.5	40.5

Table 6.1: Comparison of body fat percentiles resulting from estimating fat and muscle mass for the CAESAR dataset [RBD+02] using our linear regression model (Section 6.2.3) with the data measured by Kyle et al. [KGS+01], showing that our estimates match the distribution found in western European adults.

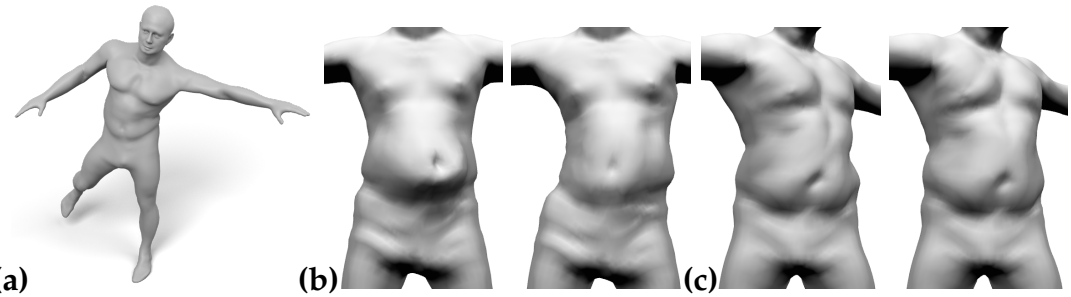

Figure 6.13: Our layered anatomical model can be animated using an extension of Fast Projective Skinning (FPS), as shown in (a). When the character performs a jump to the left (b), our realistic skeleton correctly restricts the dynamic jiggling to the belly region (b-left), while the original FPS deforms the complete torso (b-right). For a static twist of the torso (c), the rib cage of our layered model keeps the chest region rather rigid and concentrates the deformation to the belly (c-left). Without a proper anatomical model, the deformation of FPS is distributed over the complete torso (c-right).

6.3.3 Physics-Based Character Animation

One application of our model is simulation-based character animation [DB13; KB18; KB19], where the transferred volumetric layers can improve the anatomical plausibility. We demonstrate the potential by extending the Fast Projective Skinning (FPS) of Komaritzan and Botsch [KB19]. FPS already uses a simplified volumetric skeleton built from spheres and cylinders, a skeleton surface wrapping this simple skeleton, and one layer of volumetric prism elements spanned between skin and skeleton surface. Whenever the skeleton is posed, the vertices of the skeleton surface are moved, and a projective dynamics simulation of the soft tissue layer updates the skin surface.

We replace their synthetic skeleton by our more realistic version and split their soft tissue layer into our separate muscle and fat layers. This enables us to use different stiffness values for the fat and muscle layers (the latter being three times larger). Moreover, our skeleton features a realistic rib cage, whereas FPS only uses a simplified spine in the torso region. As a result, our extended version of FPS yields more realistic results in particular in the torso and belly region, as shown in Figure 6.13.

6.3.4 Simulation of Fat Growth

Our anatomical model can also be used to simulate an increase of body fat, where its volumetric nature provides advantages over existing surface-based methods.

In their computational bodybuilding approach, Saito et al. [SZK15] also propose a method for growing fat. However, they employ a purely *surface-*

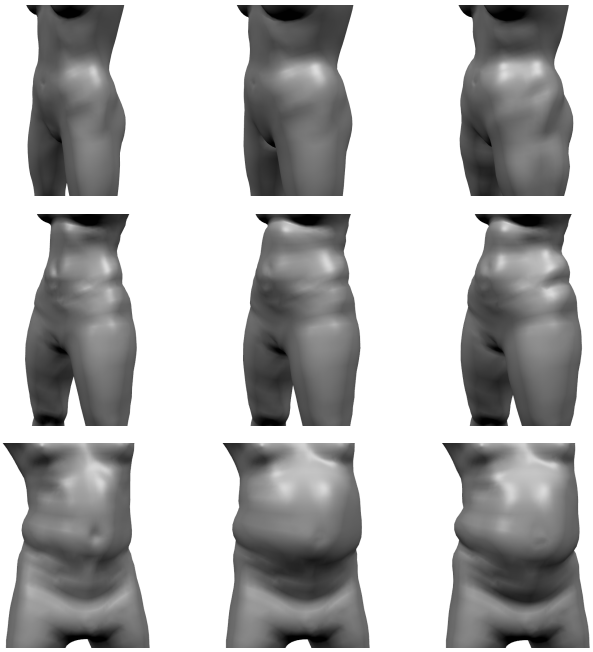

Figure 6.14: Given a reconstructed model (left), the pressure-based fat growth of Saito et al. [SZK15] leads to a more uniform increases in fat volume (center), while our volume-based fat growth increases the initial fat distribution (right).

based approach that conceptually mimics blowing up a rubber balloon. This is modeled by a pressure potential that drives skin vertices outwards in normal direction, regularized by a co-rotated triangle strain energy. The user is required to manually specify a scalar field that defines where and how strong the skin surface should be "blown up", which is used to modulate the per-vertex pressure forces and therefore models, how much each region of the model accumulates fat tissue. Despite the regularization we sometimes noticed artifacts at the boundary of the fat growing region and therefore add another regularization through Equation (6.2). This approach allows the user to tune the amount of subcutaneous fat, but unless a carefully designed growth field is specified, the fat growth looks rather uniform and balloon-like (see Figure 6.14, center-top).

Every person has an individual fat distribution and gaining weight typically intensifies these initial fat depots. We model this behavior by scaling up the local prism volumes of our fat layer. Each fat prism can be split into three tetrahedra, which define volumetric elements $t_j \in \mathcal{T}$ with initial volumes $\bar{V}_j$. A simple uniform scaling $s \cdot \bar{V}_j$ achieves the desired effect that fat increases more in fat-intense regions. The growth simulation is implemented by minimizing the energy

$$E_{\text{grow}}(\mathcal{S}) = \lambda_{\text{vol}} E_{\text{vol}}(\mathcal{S}) + \lambda_{\text{reg}} E_{\text{reg}}\big(\mathcal{S}, \bar{\mathcal{S}}\big) + \lambda_{\text{rest}} E_{\text{rest}}\big(\mathcal{S}, \bar{\mathcal{S}}\big) \tag{6.14}$$

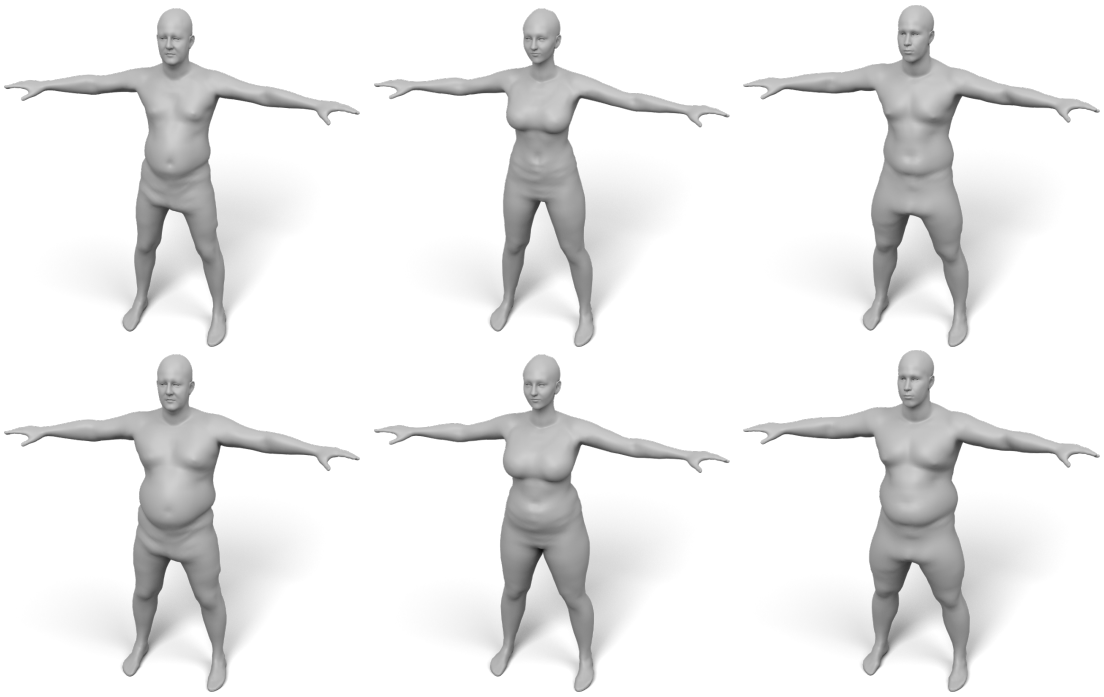

Figure 6.15: Examples of our fat growth simulation with input models shown in the top row and their weight-gained version in the bottom row.

with the Laplacian regularization of Equation (6.2), the displacement regularization

$$E_{\text{rest}}(\mathcal{S}, \bar{\mathcal{S}}) = \sum_{\mathbf{x}_i \in \mathcal{S}} A_i \left\| \mathbf{x}_i - \bar{\mathbf{x}}_i \right\|^2 \tag{6.15}$$

and the volume fitting term

$$E_{\text{vol}}(\mathcal{S}) = \sum_{t_j \in \mathcal{T}} \bar{V}_j \left(\text{vol}(t_j) - s \cdot \bar{V}_j \right)^2, \tag{6.16}$$

where $\bar{\mathcal{S}}$ and $\mathcal{S}$ denote the skin surface before/after the fat growth and s is the global fat scaling factor. Saito et al. [SZK15] argued that *anisotropically* scaling fat tetrahedra in *one* direction does not produce plausible results. However, *isotropically* scaling the volume leaves the minimization more freedom and yields convincing results. Figure 6.14 compares the pressure-based and volumetric fat growth simulations. Figure 6.15 shows some more examples produced by combining both methods.

Our volume-based fat growth has another advantage: If we want to grow fat on a very skinny person, the initial (negligible) fat distribution does not provide enough information about where to grow fat, such that both approaches would do a poor job. But since we can easily fit the volumetric template to several subjects, we can "copy" the distribution of fat prism volumes from another person and "paste" it onto the skinny target, which simply replaces the target volumes in Equation (6.16). This enables to transfer initial fat distributions between different subjects, which is shown in Figure 6.16.

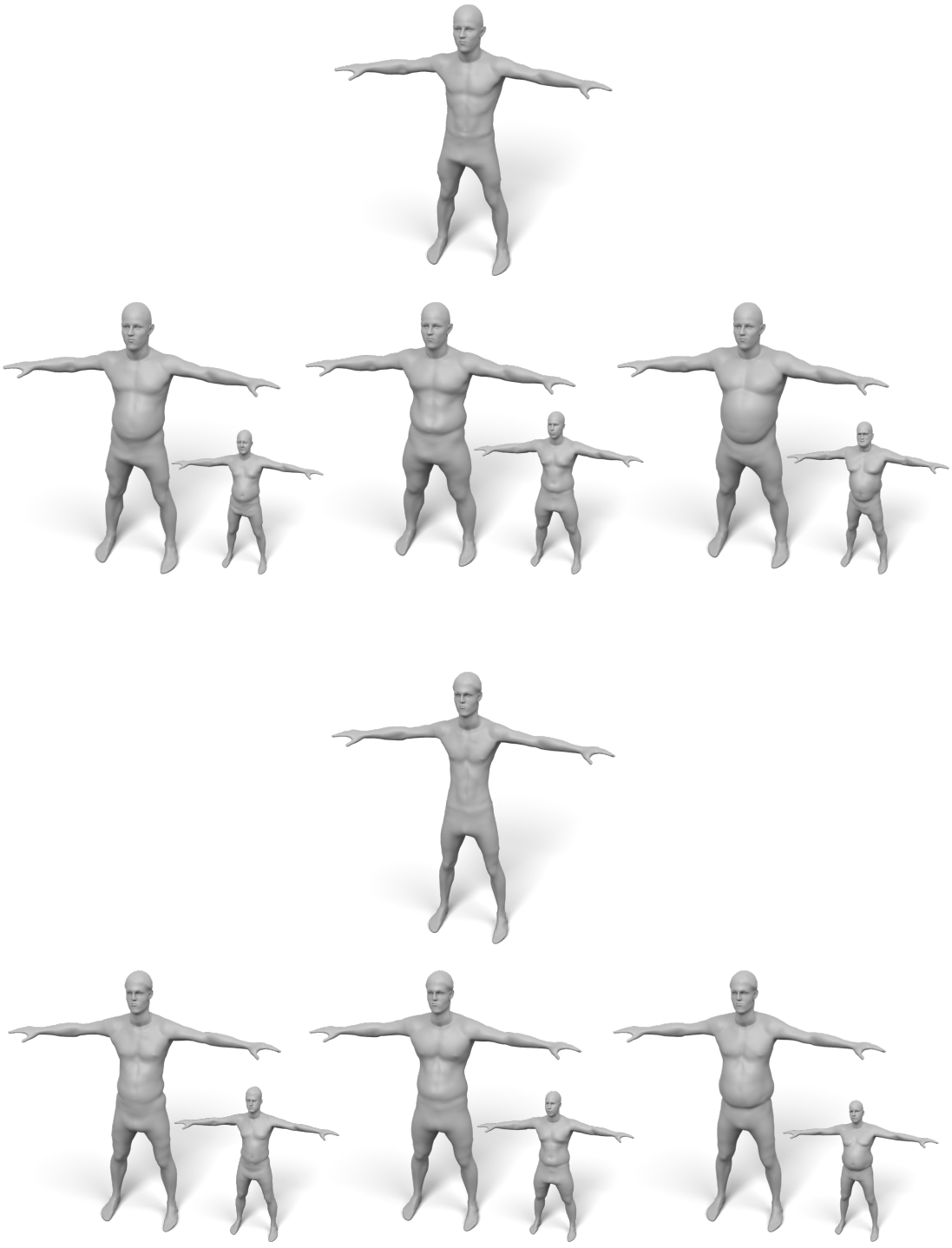

Figure 6.16: Examples of "fat transfer". The two subjects in the first and third row have a very low amount of body fat. Therefore, scaling their fat volumes is not suitable for fat growth. Instead, we copy the fat distributions of other subjects (shown as small insets). This allows us to simulate a similar fat growth behavior for the skinny targets.

6.4 SUMMARY AND LIMITATIONS

We created a simple layered volumetric template of the human anatomy and presented an approach for fitting it to surface scans of men and women of various body shapes and sizes. Our method generates plausible muscle and fat layers by estimating realistic muscle and fat masses from the surface scan alone. In addition to the layered template, we also showed how to transfer internal anatomical structures, such as bones and muscles, using a high-quality space warp. Compared to previous work, our method is fully automatic and considerably faster, enabling the simple generation of personalized anatomical models from surface body scans. Besides educational visualization, we demonstrated the potential of our model for physics-based character animation and anatomically plausible fat growth simulation.

Our approach has some limitations: First, we do not generate individual layers for head, hands and toes, where in particular the head would require special treatment. Combining our layered body model with the multilinear head model of Achenbach et al. [ABG+18] is therefore a promising direction for future work. Second, our regressors for fat and muscle mass could be further optimized by training on more body scans with known body composition. Given more and more accurate training data, as for instance provided by DXA scans or MRI images [KAD+24], we could extend the fat/muscle estimations to individual body parts. Third, we do not model organs, tendons and veins. Those would have to be included in all layers and could be transferred in the same way as high-resolution muscle and bone models. Fourth, our data sources are biased towards Caucasian adults, as both the CAESAR database [RBD+02] and the data provided by Maalin et al. [MMK+21] mostly feature this exact population. Future work should take care to extend this line of research to a more diverse population. Lastly, the fact that the three layers of our model share the same topology/connectivity can also be considered a limitation, since we cannot use different, adaptive mesh resolutions in different layers.

However, as shown in the next chapter of this thesis, the simple structure of our layered anatomical model can be exploited to generate synthetic training data, which allows to perform statistical analysis of human skeleton and soft tissue distributions to generate an anatomically constrained volumetric human shape model.

7 AN ANATOMICALLY CONSTRAINED VOLUMETRIC HUMAN SHAPE MODEL

Figure 7.1: Our anatomically constrained human shape model allows to infer the skeleton from a surface scan. Due to injecting anthropometric measurements into the latent code, our model can then locally manipulate both the skeleton shape and the soft tissue distribution of a person.

Having proposed a method for inferring anatomical details from human surface scans in the previous chapter, we will now present an approach for learning a volumetric human shape model based on the data produced by this method. Human shape modeling has been extensively studied due to its application in various fields, such as shape and pose estimation from multi-view stereo or monocular RGB(-D) input. Starting from simple linear PCA models [ASK+05; LMR+15] to more recent advances in machine learning models [RBS+18; BBP+19], these models are used as the foundation for many downstream tasks such as body composition estimation [WNT+21], the creation of virtual humans (as presented in Chapter 2 and Chapter 3), learning a statistical model of body weight modification (see Chapter 5), or generating synthetic training data for image recognition tasks [WBH+21]. Most of the approaches train on commercially available 3D scan databases such as CAESAR [RBD+02] or 3D Scanstore [3DS24]. These 3D scans naturally provide only what is easily observable from the outside: the silhouette of the scanned subject. However, by setting the focus on modeling the *skin layer* of humans, models that want to learn how to accurately *modify* a given virtual human, suffer from missing anatomical information.

Modifying realistic virtual humans has gained attention due to its promising applicability in VR therapy [MTM+18; HHM+20; TGK+21; WDM+22; WMF+22], which can serve as a complementary intervention technique to classical forms of therapy and forms the context of our research on virtual humans in this thesis. Such VR systems can immersively expose patients with anorexia or obesity to generic or personalized virtual humans at different levels of weight or body mass index (BMI), allowing patients to reflect on and researchers to gain insight into possibly occurring body image disturbances. However,

current models for body weight/BMI modification, such as the one presented in Chapter 5, are typically learned on surface-only models and employ global models such as Principal Component Analysis, leading to shape modification models providing only limited localized control [ACP03; HSS+09; PSR+14]. Participants have stated the request for changing the composition of specific body parts in addition to a global BMI/weight modification [DWM+22].

In this chapter, we present a novel approach for learning a model, which is able to achieve such localized shape manipulation. We leverage recent advances in inferring anatomical structures from surface scans [DLG+13; KIL+16; KZB+22; KWS+23; KAD+24], more specifically, the method proposed in Chapter 6, to register a volumetric anatomical template model to the CAESAR database, resulting in pairs of skeleton and skin meshes. The main contribution of this work is to provide a novel statistical model that clearly separates the distribution of skeleton and soft tissue in its latent space. In order for our model to successfully learn separate parameter sets, we calculate the full Cartesian product of all skeleton shapes and all soft tissue distributions using volumetric deformation transfer [BSP+06], allowing us to transfer the soft tissue distribution of subject i onto the skeleton of subject j.

The resulting data set is then used to train a neural network that learns the common underlying parameters from a person's bone structure and soft tissue distribution. Examples that share either a common skeleton or a common soft tissue distribution can be sampled from the Cartesian data set. These commonalities are learned and encoded with an autoencoder using the SpiralNet++ approach [GCB+19]. To separate the skeleton and soft tissue distribution, a self-supervised learning approach inspired by Barlow Twins [ZJM+21] is used, which reduces redundancy in the underlying distributions. To allow local modification of body regions, measurements are taken on the example Cartesian data set and are additionally injected into the latent code to reduce the correlation of the remaining parameters with these known measurements.

In summary, this chapter presents a novel approach for learning an anatomically constrained volumetric human shape model, which through its learning paradigm disentangles correlations between the skeleton shape space and the soft tissue distributions. The latent code of our model can be sampled to generate various human skeleton shapes with different soft tissue distribution characteristics. The measurement injection into the latent code of our model allows localized shape manipulation: the anatomical structure can be quickly inferred from a 3D scan of a human and then locally modified by the user. This allows to simulate weight gain/loss in different regions of the body. Our model is publicly released at https://github.com/mbotsch/TailorMe to enable further research and development of applications for volumetric anatomical human shape models.

Individual Contribution *The work presented in this chapter was done in collaboration with Fabian Kemper. My main contribution is the implementation of the volumetric deformation transfer used to generate the synthetic data for training our model. I additionally prepared the template data, thereby allowing us to make the model publicly available, due to not relying on the prohibitively licensed Zygote model [Zyg24]. Finally, I implemented the collision avoidance algorithm and worked together with Fabian Kemper on the virtual human modification for clothed scans as well as the comparisons to OSSO [KZB+22] and SKEL [KWS+23]. Fabian Kemper worked on the network architecture and learning algorithm, allowing us to learn separate parameter sets for skeleton and soft tissue. Fabian Kemper additionally implemented the fitting of the model to a given skin surface as well as the model evaluation.*

Corresponding Publication *This chapter is based on the following publication*:

> Stephan Wenninger, Fabian Kemper, Ulrich Schwanecke, and Mario Botsch. "TailorMe: Self-Supervised Learning of an Anatomically Constrained Volumetric Human Shape Model". *Computer Graphics Forum* 43.2 (2024)

7.1 RELATED WORK

7.1.1 Human Shape Models

Data-driven human shape models are ubiquitous and widely studied. Learned from registering a template model to a database of 3D scans, most popular models are based on Principal Component Analysis (PCA) of vertex positions [LMR+15; OBB20]. Pishchulin et al. [PWH+17] discuss best practices and provide a public implementation of the complete pipeline from surface scans to a parametric shape model. Other approaches directly encode triangle deformations from the template to the registered models [ASK+05] or a decomposition of these triangle deformations [FB12]. Since they are based on a database of 3D scans, these methods capture the variation of human body shape only on a surface level. In contrast, our model is trained on additional volumetric information by fitting an anatomically plausible skeleton model into the registered surface scans.

More sophisticated dimensionality reduction techniques have also been applied to human shape models: Ranjan et al. [RBS+18] propose a convolutional mesh autoencoder and introduce a pooling and unpooling operation directly on the mesh surface structure. The Neural 3D Morphable Models (Neural3DMM) network [BBP+19] adjusts the pooling operations and uses a spiral convolutional operator, which has been further refined by Gong et

al. [GCB+19]. Our model uses a similar autoencoder design paired with the self-supervised learning technique Barlow Twins [ZJM+21].

7.1.2 Modifying Virtual Humans

Learning a shape modification model based on anthropometric measurements has been explored in the field of Virtual Reality body image therapy [DWM+22; WDM+22; WMF+22; PSR+14; MTM+18]. The possibility to either passively present a generic virtual human in different weight or BMI variants or letting participants actively change their personalized virtual human can and has been used to gain insights into body image disorders for patients with anorexia or adiposity.

A common approach is to model shape modification by learning linear correlations between a set of anthropometric measurements (e.g., as present in the CAESAR database [RBD+02]) and the low-dimensional shape space [ACP03; HSS+09; PSR+14], as also discussed in Chapter 5. The modified shape can then be computed by mapping the desired measurement changes into the subspace through learned regressors and then projecting the change in subspace coordinates back into vertex space. Commonly used anthropometric measurements, such as arm length and inseam, are highly correlated. The cited methods cannot completely disentangle this correlation in the anthropometric measurements, leading to limited control over local shape manipulation. Our non-linear model learns to separate the correlations between these measurements, thereby enabling more localized shape manipulations.

For surface models, there is some work on creating more local shape space representations. Tena et al. [TDM11] propose a method for automatically segmenting registered head meshes into several components. A shape space is then learned for each component separately and the resulting submeshes are stitched together. *Sparse PCA* combined with spatially-varying regularization weights [NVW+13] has also been shown to result in more localized shape models. For an overview about parametric (head) surface models, including global and local models, we refer the reader to the survey by Egger et al. [EST+20]. These methods could achieve localized shape control, but are only trained on surface meshes.

7.1.3 Anatomical Models

We already introduced some previous works dealing with volumetric anatomical models of virtual humans in Chapter 6. To briefly recap, Achenbach et al. [ABG+18] trained a multilinear model (MLM) to find a lower-dimensional model of skull and corresponding head shape, parameterized by skull shape

and soft tissue distribution. However, the MLM does not completely decouple the two parameter sets, so changing the skin parameters can still affect the skeleton. Our non-linear model better decouples skeleton from skin shape, i.e., when changing skin parameters, the skeleton stays fixed. Anatomy Transfer [DLG+13] is a method for warping an anatomical template model into a target skin surface via a harmonic space warp while constraining bones to only deform via affine transformations. This can however lead to unnaturally scaled or sheared bones.

Saito et al. [SZK15] developed a physics-based simulation of muscle and fat growth on a tetrahedral template mesh including an enveloping muscle layer that separates the tetrahedral mesh representing the subcutaneous fat layer from the rest of the template. Kadleček et al. [KIL+16] present a method for fitting such a physics-based simulation to a set of 3D scans in different poses, to get a personalized anatomical model. Their approach yields visually plausible results but requires a complex numerical optimization strategy taking several minutes and can therefore not be used for interactive VR interventions. In Chapter 6, we presented our own method for reconstructing anatomical details from a surface scan. We will refer to this method as *Inside Humans* in this chapter. Inside Humans follows the approach of Saito et al. [SZK15] by using a multi-layered model to separate skeleton, muscle, and skin surface derived from an anatomical template model [Zyg24]. The model is then fitted to a given skin layer in a multi-stage optimization scheme. Embedding the high-resolution skeleton and muscle meshes from the anatomical template into the resulting layers is done by a triharmonic RBF warp. However, we did not train a statistical model on the resulting shapes. Additionally, the Inside Humans fitting approach is an order of magnitude slower compared to the method presented in this chapter.

There has been a growing series of works concerned with skeleton inference from a given skin surface. The recent work *OSSO* [KZB+22] combines the *STAR* model [OBB20] for human body shapes and a model of skeleton shapes based on the Stitched Puppet Model [ZB15]. By fitting these two models to a set of DXA images, the authors learn to infer skeletal shape from skin shape in PCA space. In the follow-up method *SKEL* [KWS+23] the authors present a parametric biomechanical skeleton and skin model with shared shape and pose parameters and anatomically constrained degrees of freedom. The skeleton model is registered to a subset of the *AMASS* dataset [MGT+19] by optimizing the scale and pose of the bones via a biomechanical optimization framework [WBR+23]. The skeletons inferred with both the OSSO and SKEL approach may however show self-intersections with the given skin mesh. In contrast, our model learns non-linear correlations between skeleton and skin, provides a localized shape modification model, and produces intersection-free pairs of skeleton and skin meshes. Schleicher et al. [SNM+21] introduced the musculoskeletal BASH model, which embeds a skeleton and muscle model

from a biomechanical simulation framework into the surface-based shape and pose model SCAPE [ASK+05]. This allows the authors to visualize the muscle activity from the biomechanical simulation on the skin surface. Shetty et al. [SBJ+23] presented a parametric anatomical model, which, in addition to skin and skeletal shape, handles organ shape. From a set of CT scans that are automatically segmented into skin, bones, and organs, the authors extract corresponding surface meshes. A set of manually annotated landmarks then guides the fitting of template models for skin, skeleton and organs to the extracted surfaces, which are unposed to a common rest pose via a complex optimization step. From the unposed and registered meshes, the authors then train a statistical parametric model using Probabilistic Principal Component Analysis [TB99].

7.2 TRAINING DATA

We start by deriving all the parts of our template model and registering it to surface scans of the CAESAR database, yielding pairs of skeleton and skin meshes (Section 7.2.1). We enlarge this data set by computing the full Cartesian product of skeleton shape and soft tissue distribution via volumetric deformation transfer (Section 7.2.2). The resulting data set then constitutes the training data for our model. See Figure 7.2 for an overview of our method.

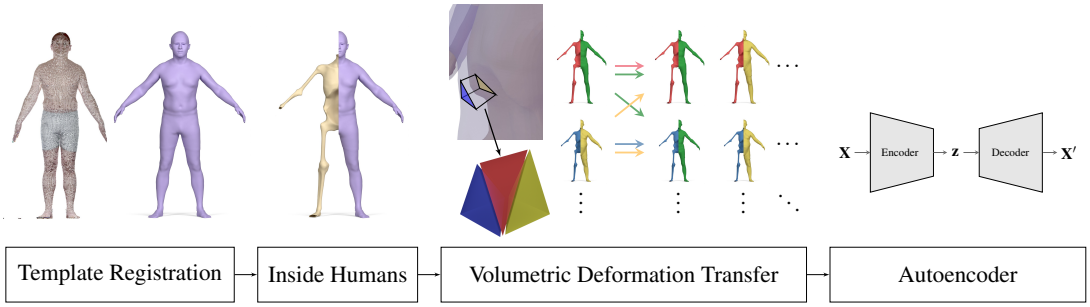

Figure 7.2: Overview of our data processing pipeline. We first fit our template model to the CAESAR database, resulting in registered skin surfaces. We use the Inside Humans method (Chapter 6) to infer anatomical structures from these surface fits. To learn a separated parameter space for skeleton and soft tissue distribution, we generate the Cartesian product of all soft tissue distributions and all skeleton shapes, using volumetric deformation transfer. We train our autoencoder by sampling pairs, which share either a common skeleton or soft tissue distribution from our Cartesian data set.

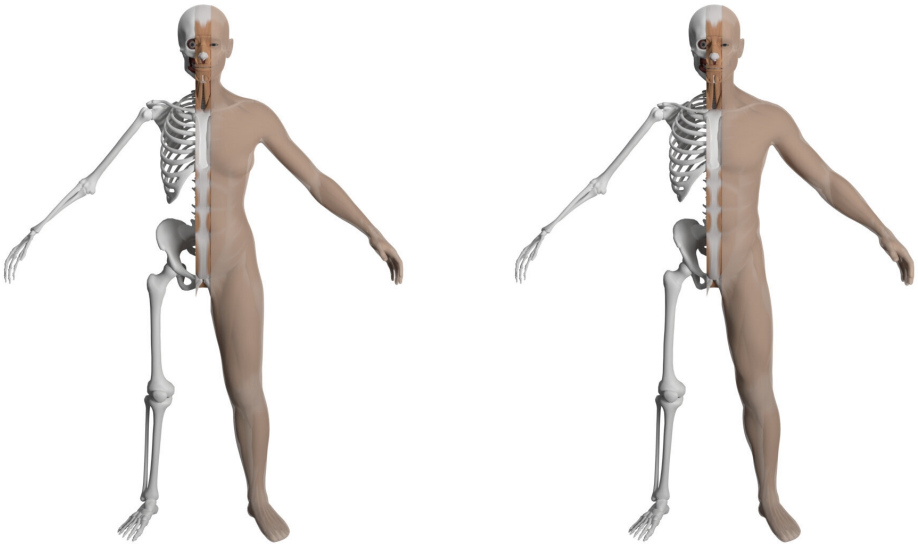

Figure 7.3: Female and male template model. In this work, we derive an additional skeleton layer that wraps the high resolution skeleton mesh and shares the triangulation with the skin layer.

7.2.1 Skin and Skeleton Registration

Existing anatomical models, as provided for example by Zygote [Zyg24] or 3D Scanstore [3DS24], are only available with prohibitive licensing. In order to make our model publicly available, we commissioned a 3D artist to build an anatomical template model. It provides a male and female template, both including meshes for skin, eyes, mouth, teeth, muscle, and skeleton (Figure 7.3). All meshes in the male template are consistently topologized with their counterparts in the female template. With 23752 vertices, our skin mesh has approximately 3.5 times more vertices than the popular SMPL [LMR+15] or STAR [OBB20] models, allowing us to more accurately model skin geometry. We follow the *layered model* approach presented in Chapter 6 and generate a skeleton wrap that envelopes the high-detail skeleton mesh and has the same triangulation as the skin layer. This provides a trivial correspondence between skin and skeletal layers.

Our skin surface input data is derived from the European subset of the CAESAR database [RBD+02], consisting of about 1700 3D scans annotated with 3D landmarks and anthropometric measurements. To bring all scans into uniform topology and pose, we employ the template fitting approach proposed by Achenbach et al. [AWL+17] (see Section 2.2.2), adapted to use the skin surface of our template model. This leaves us with 776 male and 919 female skin meshes denoted by $\mathcal{S}_i$. In the following, all computations are done on the male and female data set separately, due to the anatomical differences especially in the hip and shoulder region.

From the fitted skin meshes, we use the Inside Humans method (Chapter 6) to estimate skeleton layers $\mathcal{B}_i$, resulting in non-intersecting pairs of skeleton and skin meshes $(\mathcal{B}_i, \mathcal{S}_i)$. Since the Inside Humans approach excludes the head, hands, and feet region from the skeleton layer, we inherit this limitation. We denote the set of vertices belonging to these regions by $\mathcal{Z}$. Equipped with this data, we can now enlarge our training data set by computing the Cartesian product of skeleton shape and soft tissue distribution in a physically plausible way.

7.2.2 Volumetric Deformation Transfer

We train our model on the Cartesian product of two shape dimensions: skeleton shape and soft tissue distribution. To this end, we transfer the soft tissue of subject i onto the skeleton of subject j, which we achieve through deformation transfer [SP04; BSP+06].

In the standard formulation of deformation transfer, the deformation gradients are computed from a triangle on $\mathcal{B}_i$ to the corresponding triangle on $\mathcal{S}_i$. These deformations are then applied to $\mathcal{B}_j$ in order to generate $\mathcal{S}_j$. However, as seen in Figure 7.4 (center-right), this formulation can lead to interpenetrations of skin and skeleton. These artifacts can happen because the triangle-based deformation gradients between $\mathcal{B}_i$ and $\mathcal{S}_i$ do not encode any volumetric information of the soft tissue enclosed in between these two surfaces. To alleviate this problem we formulate the soft tissue transfer as a *volumetric* deformation transfer problem.

First, we compute the mean skeleton $\mathcal{B}_\mu$ and mean skin mesh $\mathcal{S}_\mu$ over all training models. Since the two layers share the same triangulation, the

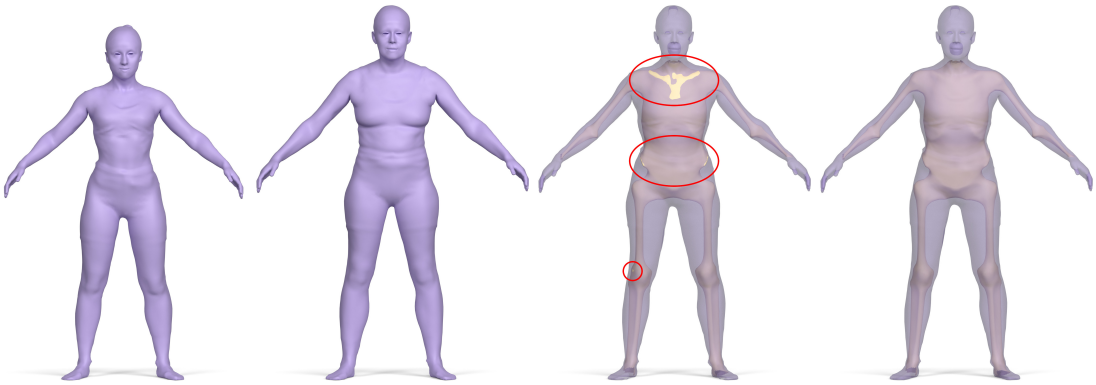

Figure 7.4: Transferring the soft tissue of the left model onto the skeleton of the center-left model using *surface-based* deformation transfer, the skeleton wrap protrudes the skin (center-right). Our volumetric deformation transfer successfully avoids these artifacts (right).

corresponding faces between the skeleton and skin layer span prismatic elements that can trivially be split into three tetrahedra. We denote the resulting tetrahedral mesh enclosed between $\mathcal{B}_\mu$ and $\mathcal{S}_\mu$ (hence representing the mean soft tissue distribution) as $\mathbb{S}_\mu$. The vector $\mathbf{X}_\mu$ containing the stacked vertex positions of $\mathbb{S}_\mu$ is composed of the vertex positions of the bone mesh $\mathcal{B}_\mu$ and the skin mesh $\mathcal{S}_\mu$, denoted by $\mathbf{X}_\mu^{\mathcal{B}}$ and $\mathbf{X}_\mu^{\mathcal{S}}$, respectively.

Transferring the soft tissue layer of subject i onto the skeleton of subject j can then be formulated as a volumetric deformation transfer. The deformation gradients $\mathbf{F}^t \in \mathbb{R}^{3\times3}$ per tetrahedron t encode the deformation from the mean tetrahedral mesh $\mathbb{S}_\mu$ to the tetrahedral mesh $\mathbb{S}_i$ of subject i. From the four vertex positions $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ of tetrahedron t in $\mathbb{S}_i$ we build the edge matrix

$$\mathbf{E}_i^t = \left(\mathbf{x}_1 - \mathbf{x}_4,\, \mathbf{x}_2 - \mathbf{x}_4,\, \mathbf{x}_3 - \mathbf{x}_4\right) \in \mathbb{R}^{3\times3}.$$

The matrix $\mathbf{E}_\mu^t$ is built analogously from the vertices of $\mathbb{S}_\mu$. The deformation gradient of tetrahedron t could then be computed as $\mathbf{F}^t = \mathbf{E}_i^t \left(\mathbf{E}_\mu^t\right)^{-1}$. However, part of the desired deformation is already explained by the deformation of $\mathcal{B}_\mu$ to $\mathcal{B}_j$. To account for this, we express the deformation gradients relative to reference frames on $\mathcal{B}_\mu$ and $\mathcal{B}_j$. Each tetrahedron t can be associated with a triangular face on the skeleton layer $\mathcal{B}_\mu$ and $\mathcal{B}_j$, respectively. These triangles define orthonormal reference frames $\mathbf{R}_\mu^t$ and $\mathbf{R}_j^t$, respectively, which leads to the final formulation for deformation gradients:

$$\mathbf{F}^t = \mathbf{R}_j^t \left(\mathbf{R}_\mu^t\right)^\mathsf{T} \mathbf{E}_i^t \left(\mathbf{E}_\mu^t\right)^{-1}. \tag{7.1}$$

We then solve for vertex positions $\mathbf{X}_j$ conforming to these deformation gradients in a least squares sense, while keeping the vertices of the skeleton layer $\mathcal{B}_j$ and $\mathcal{Z}$ fixed. Formally, we solve the gradient-based mesh deformation system

$$\left(\mathbf{G}^\mathsf{T}\mathbf{D}\,\mathbf{G}\right)\mathbf{X}_j \;=\; \left(\mathbf{G}^\mathsf{T}\mathbf{D}\right)\mathbf{F}, \tag{7.2}$$

with Dirichlet boundary constraints for every vertex belonging to $\mathcal{B}_j \cup \mathcal{Z}$. The matrices $\mathbf{G}^\mathsf{T}\mathbf{D}$ and $\mathbf{G}$ represent the discrete divergence and gradient operators for tetrahedral meshes [BSP+06], and $\mathbf{F}$ vertically stacks the desired deformation gradients $\mathbf{F}^t$. Solving this linear system yields new skin vertices $\mathbf{X}_j^{\mathcal{S}}$. In order to smoothly blend into the boundary region $\mathcal{Z}$, we define per-tetrahedron interpolation weights $w^t \in [0,1]$, which decrease based on the distance to $\mathcal{Z}$. We use w^t to linearly interpolate between the desired deformation gradients $\mathbf{F}^t$ and the deformation gradients computed from $\mathbb{S}_\mu$ to the target subject $\mathbb{S}_j$, thereby ensuring a smooth transition into $\mathcal{Z}$.

In the presented volumetric formulation, the deformation gradients $\mathbf{F}^t$ include information about the volumetric stretching and compression of the tetrahedron t of the soft tissue layer. Since $\mathbb{S}_\mu$ and $\mathbb{S}_i$ do not exhibit inverted

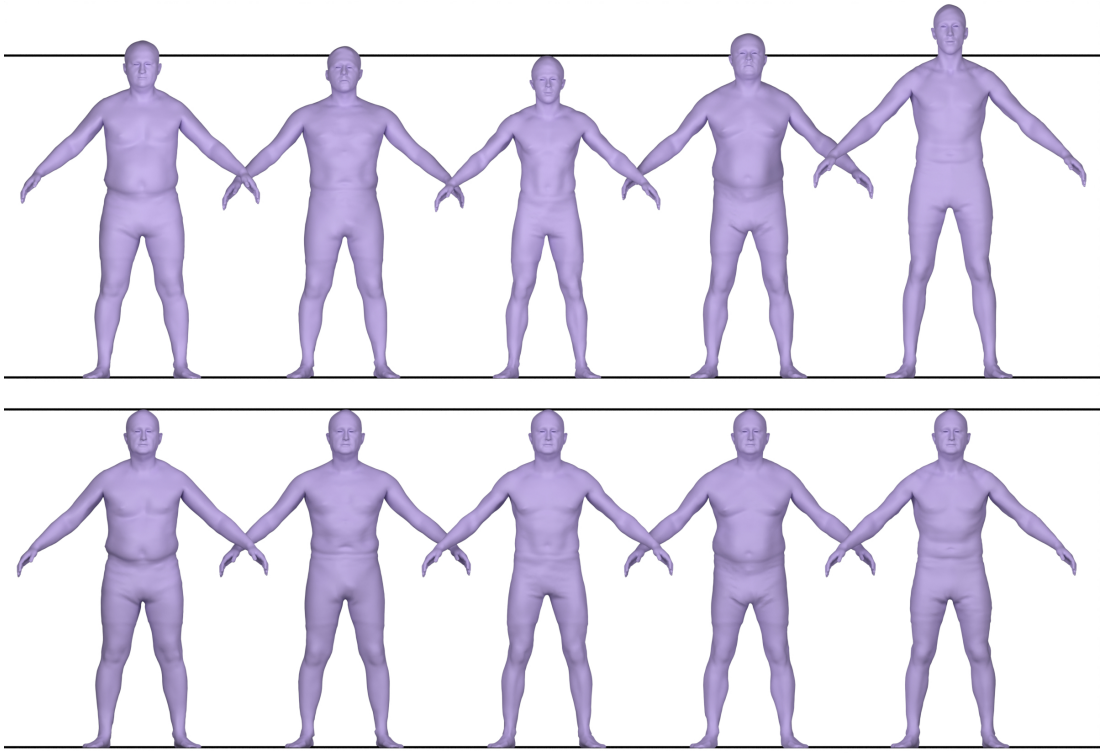

Figure 7.5: Exemplary results of transferring the soft tissue of various people (top row) onto a single target skeleton via volumetric deformation transfer (bottom row). Note that soft tissue characteristics of the top row and skeletal dimensions of the bottom row are faithfully preserved.

elements, the deformation gradients $\mathbf{F}^t$ do not contain any inversions. As such, solving Equation (7.2) avoids self-intersections between skin and skeleton (shown in Figure 7.4, right), since those would require tetrahedra to invert, which in turn would lead to a high deviation from the target deformation gradient. Figure 7.5 shows several examples of transferring the soft tissue distribution of a set of subjects with different height and weight characteristics onto the same skeleton.

7.3 MODEL LEARNING

Our objective is to learn a compact representation of human body shapes. We do so using a specific autoencoder architecture. To enable guided and localized shape manipulation, we inject anthropometric measurements into the autoencoder's latent representation. We measure the length of the torso, arms, and legs on the skeleton, and the circumference of chest, waist, abdomen, and hips on the skin meshes. By injecting normalized values of those measurements into our latent representation, we form an expressive latent code. Our neural network architecture is a convolutional autoencoder with local

mesh convolutions based on SpiralNet++ [GCB+19]. Decoupling the two shape dimensions is accomplished by splitting the latent code into two parameter sets: one for skeleton shape and one for soft tissue distribution (Section 7.3.1). We define a loss function based on the Barlow Twins method [ZJM+21], which allows us to reduce the redundancy in the latent code (Section 7.3.2).

7.3.1 Network Architecture

Our shape compression task is implemented using a convolutional autoencoder. To achieve decoupling of skeleton shape and soft tissue distribution, we encode all samples using the encoder of our network, and split the resulting embeddings $\mathbf{z}$ into two parts $\mathbf{z}^{(\mathcal{B})}$ and $\mathbf{z}^{(\mathcal{S})}$, representing the skeleton and soft tissue distribution. To facilitate semantic control in the latent space, the normalized values of the measurements taken on the original meshes are then appended to the respective part of the latent code. Measurements taken on the skeleton are appended to $\mathbf{z}^{(\mathcal{B})}$, skin measurements are appended to $\mathbf{z}^{(\mathcal{S})}$.

As a first design for the shape compression task, we experimented with utilizing two separate PCA models for the skeleton and soft tissue distribution. The PCA weights then formed the input to our autoencoder, which learned to decouple the two shape dimensions. This is in line with the OSSO approach [KZB+22], where the correlation between skin and skeleton shape is learned by a linear regressor between two PCA subspaces. We found that the resulting model separates the skeleton and soft tissue distribution, but only provides global shape control when modifying semantic parameters in the latent space, due to the global influence of the PCA weights. Figure 7.6 shows an example of the global influence, where modifying arm length also changes the body height.

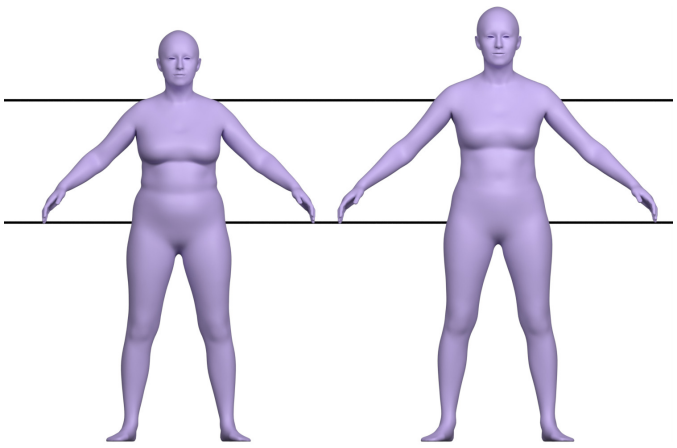

Figure 7.6: Representing skeletons and skins in PCA subspaces separates their parameters, but the global nature of PCA prevents localized changes: Increasing the arm length of the left model also causes the body height to increase (right).

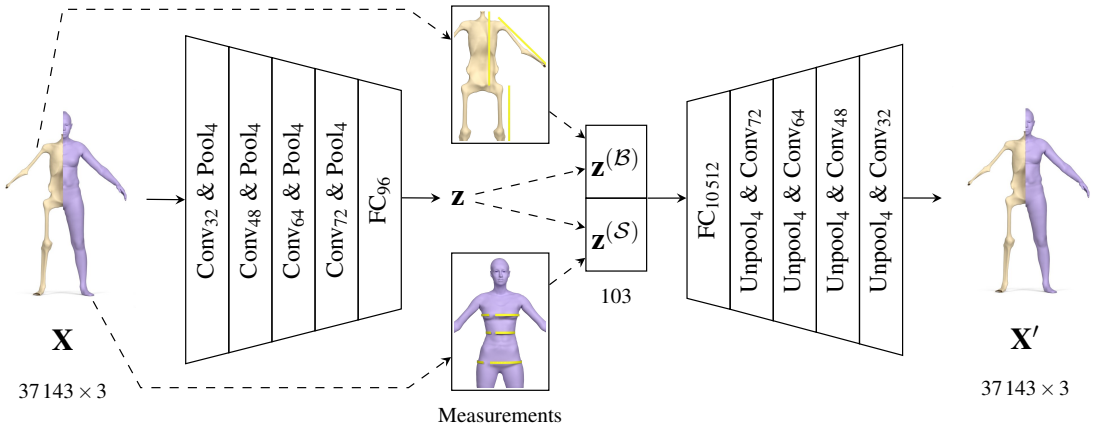

Figure 7.7: Our network architecture is based on four SpiralNet++ [GCB+19] convolution and pooling blocks in the encoder. A final dense layer is connecting the last layer of the encoder to achieve our embeddings $\mathbf{z}$. We divide the embeddings into two parts, one for the skeleton and the other one for the soft tissue distribution. We append the normalized values of the measurements (the lengths of torso, arms and legs and the circumferences of chest, waist, abdomen, and hips) taken on the input mesh via a skip connection to the correspondingc part of the latent code. The decoder is the reversed order of the encoder using four unpooling and convolution blocks.

To mitigate the global effects, we opt for an autoencoder using the SpiralNet++ approach [GCB+19], which utilizes a mesh convolution and pooling operator. This design enables local shape control when modifying the entries in the latent space that correspond to the anthropometric measurements.

The structure of our autoencoder is shown in Figure 7.7. The samples $\mathbf{X} \in \mathbb{R}^{37143\times 3}$ drawn from our Cartesian product data set (Section 7.2.2) consist of the vertex positions $\mathbf{X}^{\mathcal{S}}$ and $\mathbf{X}^{\mathcal{B}}$ of the skin mesh and the skeleton wrap (the latter excluding vertices in $\mathcal{Z}$ belonging to head, hands, and feet). We normalize the vertex positions before using them as input for our autoencoder. Our latent code utilizes 48 parameters for the skeleton and soft tissue distribution each. We take three measurements on the skeleton (torso, arm and leg length), four measurements on the skin (chest, waist, abdomen and hip circumference), and append them to the resulting embeddings via skip connections. This results in a total of 103 parameters in the latent space.

7.3.2 Cross-Correlation Loss

Our encoder creates a latent representation $\mathbf{z}$ for each sample $\mathbf{X}$ in the Cartesian data set. Samples with identical skeleton shape should result in identical skeleton embeddings $\mathbf{z}^{(\mathcal{B})}$, while samples that share soft tissue distribution should result in identical soft tissue embeddings $\mathbf{z}^{(\mathcal{S})}$. We achieve this using a

self-supervised learning approach based on Barlow Twins [ZJM+21], where the loss formulation penalizes dissimilar embeddings for similar input samples.

We extend the concept of pairs in Barlow Twins by using quadruplets, built by all four combinations of skeletal and soft tissue distribution from two samples each for pairs of skeleton and skin meshes. To build such a quadruplet, we randomly select two different indices k, l for skeletons and two different indices for the distribution of soft tissues m, n from our training data set. Let $\mathbf{X}_{km}$ denote the vertex positions resulting from transferring the soft tissue distribution of subject m onto the skeleton of subject k via volumetric deformation transfer (Section 7.2.2). From the chosen indices (k, l, m, n) we create a quadruplet containing the entries $(\mathbf{X}_{km}, \mathbf{X}_{kn}, \mathbf{X}_{lm}, \mathbf{X}_{ln})$, such that each entry shares either its skeleton or its soft tissue distribution with two of the other entries. These quadruplets are processed in batches by our autoencoder. The forward process of the encoder for one quadruplet in a batch is visualized in Figure 7.8.

Following the Barlow Twins method [ZJM+21], we reduce the redundancy in the embeddings of common features – resulting from samples that share either the skeleton shape or the soft tissue distribution – by computing empirical cross-correlation matrices of the embeddings and penalizing their deviation from the identity matrix. The cross-correlation loss is defined as

$$\mathcal{L}_{\mathrm{BT}} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2, \tag{7.3}$$

with

$$C_{ij} = \frac{1}{\mathrm{BS}} \sum_b z_{b,i}^A \cdot z_{b,j}^B. \tag{7.4}$$

The batch size is marked as BS, the batch dimension is denoted by b, and the index dimensions of the network output are represented by i and j. λ is a trainable hyperparameter to weight the importance of off-diagonal entries being close to 0 in the empirical cross-correlation matrices. $\mathbf{z}^{(A)}$ and $\mathbf{z}^{(B)}$ are batches of embeddings, which are selected as described in the following. Note that Equation (7.4) differs from the original definition in that we do not use batch normalization on the embeddings before calculating the entries of the cross-correlation matrices C_{ij}. Our model achieves greater accuracy without batch normalization and enables more efficient manual modifications to the reconstructed meshes.

We rearrange the embeddings in a batch to group all components with similarities on the source data set. These embeddings should have a minimal redundancy when originating from the same distribution. This is indicated when sharing one of their indices in the training data set k, l, m, or n. We can form four cross-correlation matrices for the quadruplets in the batch: $\mathbf{z}_{km}^{(\mathcal{B})} \otimes \mathbf{z}_{kn}^{(\mathcal{B})}$, $\mathbf{z}_{lm}^{(\mathcal{B})} \otimes \mathbf{z}_{ln}^{(\mathcal{B})}$, $\mathbf{z}_{km}^{(\mathcal{S})} \otimes \mathbf{z}_{lm}^{(\mathcal{S})}$, and $\mathbf{z}_{kn}^{(\mathcal{S})} \otimes \mathbf{z}_{ln}^{(\mathcal{S})}$, where $\otimes$ denotes the outer product of the batched embeddings. We calculate the cross-correlation loss

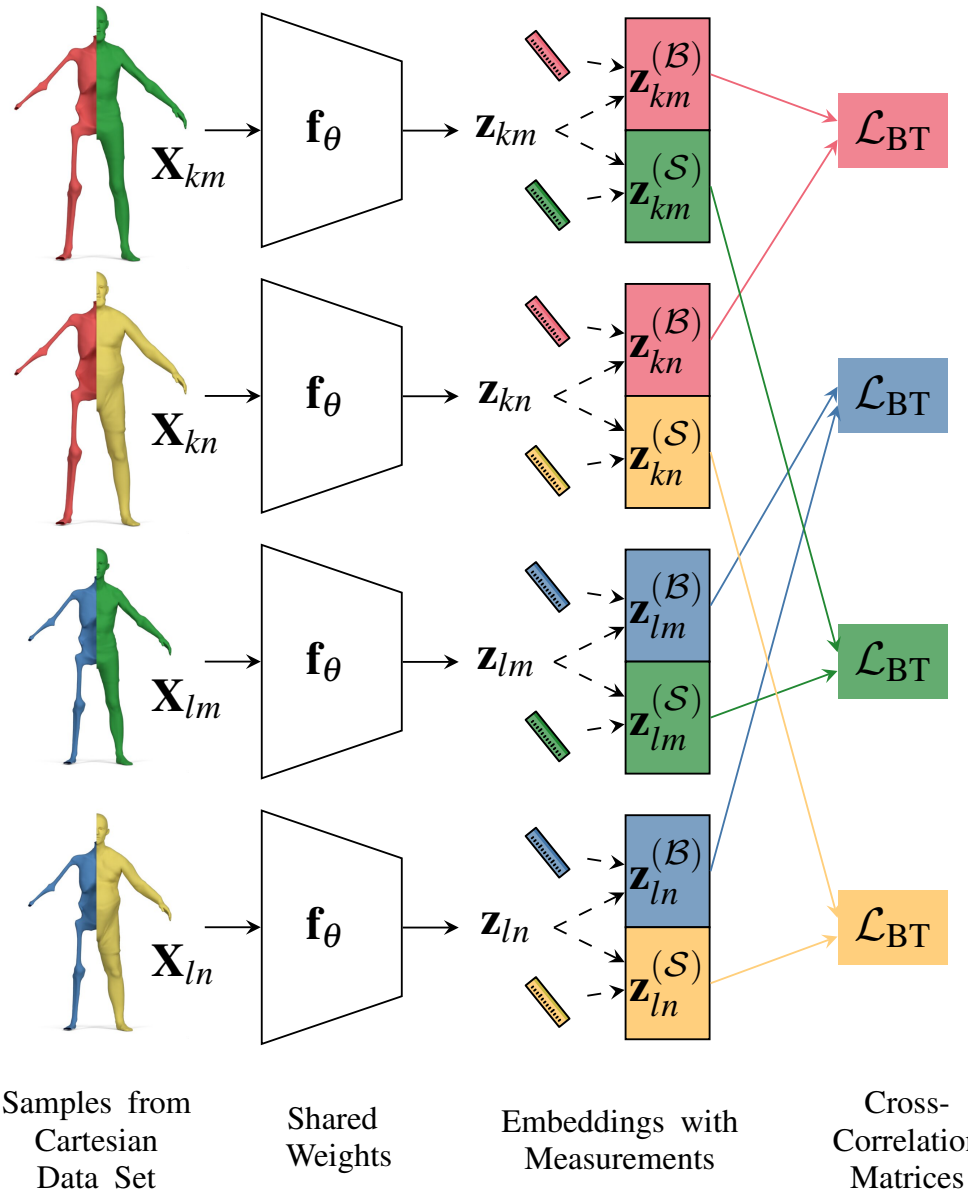

Figure 7.8: Processing a quadruplet of samples in the encoder and dividing the embeddings into two parts: one for the skeleton and one for the soft tissue distribution. After separation, the normalized measurements are appended to the respective embeddings. If the divided embeddings are noted the same way, they are merged for the entire batch. In order to calculate the cross-correlation loss between pairs, the cross-correlation matrices are calculated for the positions colored in the same way, which indicates that they share either the same skeleton shape or the same soft tissue distribution.

$\mathcal{L}_{\text{BT}}$ for each matrix using Equation (7.3) and sum up the total loss for all four matrices of the batch to compute our redundancy loss $\mathcal{L}_Q$. Note that we do not need to minimize the redundancy between skeleton and soft tissue parameters directly to learn a separation.

Let $\mathcal{L}_R$ denote the L^1 reconstruction loss over all samples of the quadruplets in the batch. We train our network to minimize the combined loss function

$$\mathcal{L} = \mathcal{L}_R + \gamma \mathcal{L}_Q, \tag{7.5}$$

where γ is a trainable hyperparameter that balances the importance of reconstruction and redundancy reduction in our loss function.

We train the autoencoder using the Adam Optimizer [KB15]. A randomized hyperparameter search is conducted and the highest performing model in the validation data set was selected. We trained our models over the complete training set, resulting in 12 epochs for female and 27 for males. This model was trained using a learning rate of $1.71 \cdot 10^{-4}$ for females and $1.78 \cdot 10^{-4}$ for male. We used redundancy importance values with $\gamma = 0.52$ for females and $\gamma = 0.42$ for males. The optimal importance hyperparameter for off-diagonal entries to achieve the best performance was $\lambda = 2.3 \cdot 10^{-2}$ for females and $\lambda = 4.3 \cdot 10^{-2}$ for males.

7.4 POST-PROCESSING

After the inference of the decoder, the resulting meshes might show certain artifacts. We observed asymmetries in the face region, which are amplified when modifying the latent code towards the boundary of the learned distribution. This is in part due to the fact that the variance of the face region was not part of the modification in the training data set (Section 7.2.1). To mitigate potential artifacts, we apply three post-processing steps after decoding: (i) the face region is symmetrized, (ii) the resulting skin surface is smoothed and (iii) intersections between the skeleton and skin layer are resolved. As a final step, we embed the high-resolution anatomical skeleton into the skeleton wrap using a triharmonic space warp.

7.4.1 Face Symmetrizing and Smoothing

After the inference of the decoder, we approximately symmetrize the face region by adapting the approach of Mitra et al. [MGP07]. A reflective symmetry plane is defined at the center of the head, based on which corresponding vertex pairs (v_i, v_j) on both sides can be determined. The y-coordinate of these vertex pairs are adjusted to match approximately: $y_i' = \frac{3}{4} y_i + \frac{1}{4} y_j$. Afterwards, one explicit smoothing step [DMS+99] is performed on the skin mesh in order to

reduce high frequency noise, which may occur when applying drastic changes to the latent parameters.

7.4.2 Intersection Avoidance

After decoding from the latent space, the resulting skeleton $\bar{\mathcal{B}}$ with vertices $\mathcal{V}$ might slightly protrude the skin layer $\mathcal{S}$, especially when the target skin measurements in the latent code are set to lower values. We detect protruding triangles and add all vertices belonging to its two-ring neighborhood to the collision set $\mathcal{C}_{\text{coll}}$.

When inferring the skeleton for a skin $\mathcal{S}$ given by a 3D scan (as demonstrated in Section 7.5.5), we want to keep the vertices on $\mathcal{S}$ fixed, as they can be considered ground truth. As such, in order to resolve the detected collisions, we solve for a new skeleton layer $\mathcal{B}$ by minimizing

$$E(\mathcal{B}) \;=\; E_{\text{reg}}\big(\mathcal{B},\bar{\mathcal{B}}\big) + E_{\text{close}}\big(\mathcal{B},\bar{\mathcal{B}}\big) + \lambda_{\text{coll}} E_{\text{coll}}(\mathcal{B},\mathcal{S})\,, \tag{7.6}$$

where E_{reg} is a bending constraint on the skeleton layer:

$$E_{\text{reg}}\big(\mathcal{B},\bar{\mathcal{B}}\big) \;=\; \frac{1}{2}\sum_{\mathbf{x}_i\in\mathcal{B}} A_i \left\|\Delta\mathbf{x}_i - \mathbf{R}_i\Delta\bar{\mathbf{x}}_i\right\|^2. \tag{7.7}$$

$\mathbf{R}_i \in SO(3)$ denotes the rotation matrix optimally aligning the vertex Laplacians between the resolved surface $\mathcal{B}$ and the initial surface $\bar{\mathcal{B}}$. The Laplace operator is discretized using cotangent weights and Voronoi areas A_i [BKP+10],

E_{close} constrains vertices that are not part of the collision set $\mathcal{C}_{\text{coll}}$ to stay close to their original position:

$$E_{\text{close}}\big(\mathcal{B},\bar{\mathcal{B}}\big) \;=\; \frac{1}{|\mathcal{V}\setminus\mathcal{C}_{\text{coll}}|}\sum_{\mathbf{x}_i\notin\mathcal{C}_{\text{coll}}} \left\|\mathbf{x}_i - \bar{\mathbf{x}}_i\right\|^2, \tag{7.8}$$

and E_{coll} defines the collision avoidance term:

$$E_{\text{coll}}(\mathcal{B},\mathcal{S}) \;=\; \frac{1}{|\mathcal{C}_{\text{coll}}|}\sum_{\mathbf{x}_i\in\mathcal{C}_{\text{coll}}} w_i \left\|\mathbf{x}_i - \boldsymbol{\pi}_{\mathcal{S}}(\mathbf{x}_i)\right\|^2, \tag{7.9}$$

where $\boldsymbol{\pi}_{\mathcal{S}}(\mathbf{x}_i)$ projects vertex $\mathbf{x}_i$ to lie 2.5 mm beneath the colliding triangle's plane on the skin $\mathcal{S}$.

We iteratively minimize Equation (7.6) via the projective dynamics solver implemented in the ShapeOp library [DDB+15]. We set the global collision avoidance weight to $\lambda_{\text{coll}} = 50$, the local per-vertex collision avoidance weight to $w_i = 1$, and progressively increase w_i by 1 each iteration in which the collision could not be resolved. After each iteration, the Laplacian of the initial state $\bar{\mathcal{B}}$ in Equation (7.7) is updated to the current solution, thereby making

the skeleton layer slightly less rigid. Following this optimization scheme, we could reliably resolve all between-layer-collisions in our tests.

Note that when modifying the soft tissue distribution over a given skeleton $\mathcal{B}$, we analogously keep the vertices on $\mathcal{B}$ fixed, and solve for a new intersection-free skin layer $\mathcal{S}$.

7.4.3 Embedding High-Resolution Skeleton

Once an intersection-free pair $(\mathcal{B}, \mathcal{S})$ is generated, we embed the high-resolution skeleton mesh (Figure 7.3) by following the approach presented in Chapter 6, i.e., using a space warp based on triharmonic radial basis functions [BK05]. The matrix of the involved linear system depends on the template skeleton $\hat{\mathcal{B}}$ only and hence can be pre-factorized. After generating a new skeleton $\mathcal{B}$, the solution can be inferred by back-substitution, and the space warp can efficiently be evaluated to embed the high-resolution anatomical skeleton.

7.5 RESULTS AND APPLICATIONS

The resulting model allows local shape manipulation based on the injected measurements in the latent space. For a demonstration of the final model, we refer the reader to the accompanying video at `https://www.youtube.com/watch?v=rrkf_fIhX0Q`. In the following, we evaluate the performance of the model on our test data set, and compare our approach to the related approaches of OSSO [KZB+22], SKEL [KWS+23], MLM [ABG+18] and standard PCA approaches [ACP03; PSR+14]. Finally, we demonstrate the modification of 3D scanned realistic virtual humans.

7.5.1 Model Evaluation

To quantitatively evaluate the fit of our trained model defined in Section 7.3, we do not perform the post-processing described in Section 7.4. We use separate data sets for training and model evaluation. The training subset is utilized for model optimization, while a validation subset is used to conduct an automated evaluation and to estimate the model's capability for generalization. To minimize the validation set bias, we utilize a third subset of our data set for testing. The splitting is done such that the skeleton and soft tissue distribution of an individual ends up in only one of these subsets.

For training our model, we use a batch size of 64 samples. We divide our data set into training, validation, and test such that the number of Cartesian pairs in the final data set corresponds to a ratio of $8:1:1$. Let a, b, and c denote the number of samples in the training, validation, and test set, respectively.

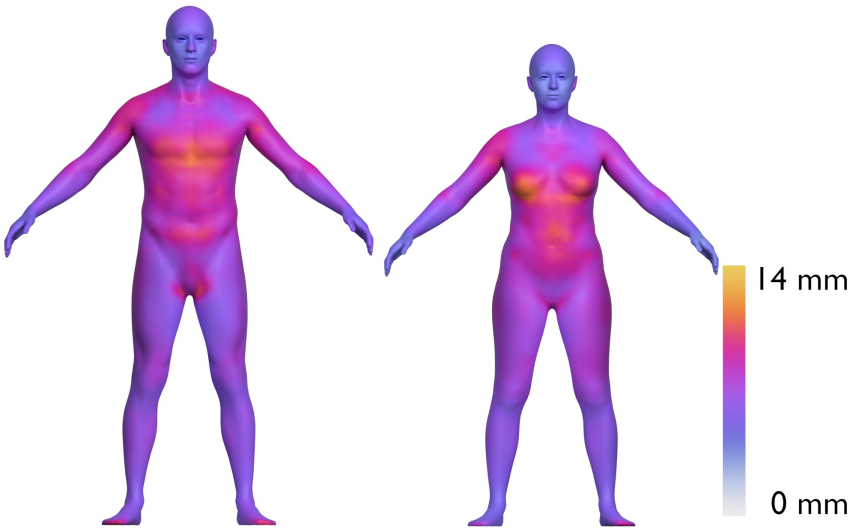

Figure 7.9: Mean Euclidean vertex distance when evaluating over all samples in the male (left) and female (right) test data set. Our model achieves a maximum Euclidean distance of 11.5 mm for males and 13.9 mm for females.

We want the ratio of squared subset samples $a^2 : b^2 : c^2$ to match the target split ratio of $8 : 1 : 1$. This split results in 539 samples for the female and 456 samples for the male training set. The validation and test sets have 190 female and 160 male samples each. Using deformation transfer, we effectively square the training set size to 290 521 female and 207 936 male samples. The validation and test sets contain 36 100 samples for females and 25 600 samples for males each.

Figure 7.9 displays our model's reproduction error for the skin, measured as per-vertex distance averaged over all meshes in the test data set, when using the decoder back propagation method. The vertex distances of the fitted skins are evenly distributed on the limbs and face, but in the chest and abdomen regions the largest average deviation from the target is observed. Overall, our model attains a maximum per-vertex error of 13.9 mm over all samples in the test data set.

The mean absolute error is the L^1 loss for the predicted mesh $\mathbf{X}'$ to the input mesh $\mathbf{X}$ with n vertices, which is defined as

$$\mathcal{L}_1 = \frac{1}{3V} \left\| \mathbf{X} - \mathbf{X}' \right\|_1 = \frac{1}{3V} \sum_{i=1}^{V} \left\| \mathbf{x}_i - \mathbf{x}'_i \right\|_1 . \tag{7.10}$$

When using the encoder and decoder for reproduction of the test data set, we achieve a mean absolute error for the skeleton wrap and the skin of 5.2 mm ($SD = 1.5$) for females and 5.4 mm ($SD = 1.5$) for males. The cross-correlation matrices in Equation (7.4) converge to the identity matrix. The individual mesh measurements positively correlate with each other as the circumferential measurements on the original meshes show a strong connection.

When inferring a skeleton from a given skin $\mathbf{X}^{\mathcal{S}}$, we use the Adam optimizer [KB15] on the latent parameters $\mathbf{z}$ and minimize the L^1 error for the skin arising from decoding $\mathbf{z}$ to the target skin of the sample. For skins, we achieve a mean absolute reproduction error on the test data set of 2.8 mm ($SD = 0.5$) for females and 2.9 mm ($SD = 0.5$) for males. For skeleton wraps, we reach a mean absolute error of 6.3 mm ($SD = 1.4$) for females and 6.9 mm ($SD = 1.7$) for males.

7.5.2 Comparison to OSSO and SKEL

We qualitatively compare our work to the OSSO approach [KZB+22]. This method computes a linear regressor between skin and skeleton PCA shape spaces, after fitting both shape models to a set of DXA Scans. As DXA scans are taken in a lying pose, OSSO first reposes a given skin mesh to this pose, infers skeleton shape there, and finally reposes the given result to the input pose. As seen in Figure 7.10, when compared to our skeleton prediction,

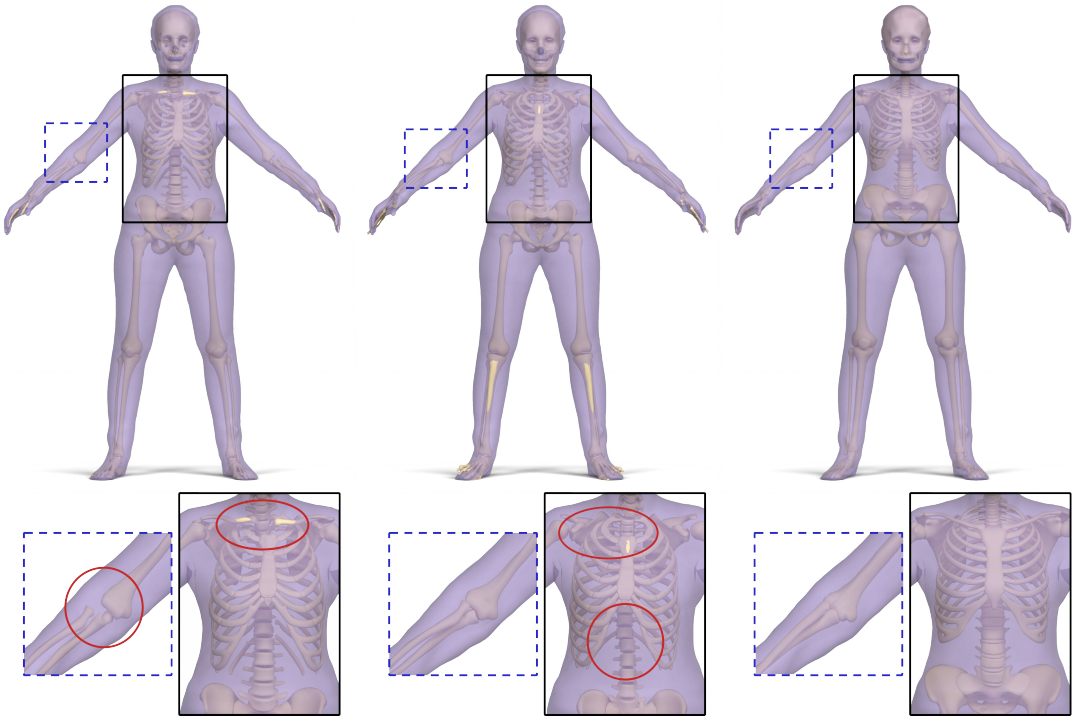

Figure 7.10: Comparison of OSSO [KZB+22] (left) and SKEL [KWS+23] (center) with our approach (right). The OSSO skeleton protrudes through the skin (yellow protrusions circled in red), we resolve these kinds of collisions (right). The rib cage inferred by OSSO is skewed (black inset) and the stitched puppet model results in gaps between bone structures (dashed blue inset). The SKEL skeleton (center) is missing the clavicles (circled in red). The individual spine segments (lumbar, thoracic and cervical) are not properly aligned (black inset).

the skeleton inferred by OSSO exhibits a skewed and unnaturally shifted rib cage as well as large gaps between bone structures, such as the elbow region or in between ribs and spine. We also compare our work to the SKEL approach [KWS+23], where a combined parametric model for biomechanical skeleton and skin shape is presented. The final model is able to infer an animatable biomechanical skeleton from given SMPL parameters as shown in Figure 7.10. We observe that SKEL's template skeleton misses bones for the clavicles and the lower spine unnaturally detaches from the pelvis, when fitting the model to a target skin.

Moreover, both OSSO's and SKEL's skeleton protrudes through the given skin, while our method resolves these intersections (Section 7.4.2). We evaluated the number of skeleton and skin intersections by fitting our model, OSSO, and SKEL to 1697 samples from the European subset of the CAESAR data set [RBD+02]. For OSSO and SKEL, there is no single sample which is free of intersections. Our model produces self-intersections in only 1.30 % of cases, and then only due to the RBF warp in the head region, where hairs are not correctly handled in our method – a limitation we inherit from the Inside Humans approach (Chapter 6).

We quantitatively evaluate the geometric difference between the skeleton fits of our method, OSSO, and SKEL on the 1697 CAESAR samples by calculating for each model the average per-vertex distance and the two-sided Hausdorff distance between the respective skeletons, and then averaging those numbers over all samples. Our model then attains an average per-vertex distance of 0.76 cm to SKEL and 0.56 cm to OSSO, and a Hausdorff distance of 6.20 cm to SKEL and 5.78 cm to OSSO. For comparison, the skeletons inferred by SKEL and OSSO deviate by an average per-vertex distance of 0.64 cm and a Hausdorff distance of 4.71 cm.

7.5.3 Comparison to MLM

We compare our model to the multilinear model (MLM) presented by Achenbach et al. [ABG+18]. The MLM approach requires the computation of a 3D tensor to separate the skeleton and soft tissue dimensions. Given our model with 51 skeleton and 52 soft tissue parameters, the MLM requires a total of 292 M parameters. Our method requires two orders of magnitude fewer parameters (2.1 M) to process the same number of input variables. When applying the MLM approach to our training data, we found that the decoupling process of the two parameter sets is incomplete. This effect can also be observed in the original work. The authors provide a demo application at `https://cg.cs.tu-dortmund.de/publications/2018-multilinear.page`, where changing the first skull parameter causes subsequent changes to the first *soft tissue* parameter to also change the resulting skull shape.

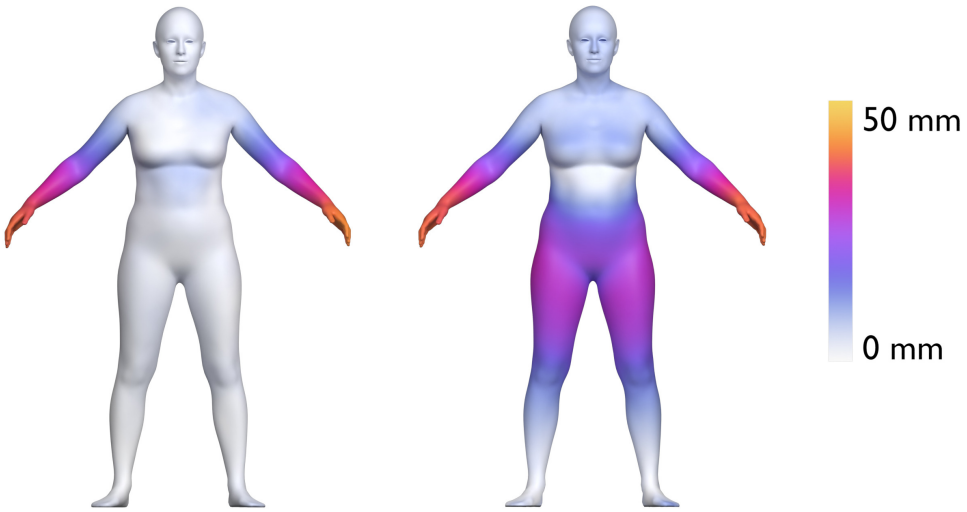

Figure 7.11: Comparison of our localized shape modification (left) with a global PCA approach [PSR+14] (right). Both models were used to shorten arm length by 38 mm. The vertex distance from the original to the modified mesh is color coded. Our model enables more localized shape changes, while the global PCA approach considerably changes the leg region when modifying arm length.

7.5.4 Comparison to Surface PCA

To show the benefits of our local and non-linear autoencoder and mesh convolution design (Section 7.3), we compare our method to the common approach of modeling anthropometric shape manipulation by correlating measurements with a global and linear PCA subspace learned from surface scans [ACP03; PSR+14]. These methods allow global shape manipulation, but provide only limited *local* control. For an example of this effect, we compare the results of shortening arm length with our model and the model proposed by Piryankova et al. [PSR+14]. Figure 7.11 shows that our model provides local control of arm length, whereas the surface based approach results in notable changes in the leg region. To interactively explore the shortcomings of the surface based approach, we invite the reader to experiment with the demo application at `https://bodyvisualizer.com`, and try to change the inseam parameter, while keeping the other measurements fixed. This changes the model's arm length in addition to the desired effect of changing the leg length.

7.5.5 Modifying Virtual Humans

As demonstrated in Figure 7.1, we can also fit our model to surface scans of clothed humans, allowing us to modify virtual humans with our model. To this end, given a registered surface scan conforming to our skin layer topology, we

let our model infer skin and skeleton shape by optimizing the mean absolute error with additional weight decay in order to prevent overfitting.

To determine a fit for the skeleton and soft tissue distribution of a person, gradient descent is performed on the latent parameters $\mathbf{z}$ using the Adam optimizer [KB15] implemented in PyTorch and running on the GPU. We apply a weight decay of $7.5 \cdot 10^{-5}$ to prevent fitting values that are too far outside the learned embeddings. Fitting the skeleton and soft tissue parameters takes less than 100 ms on a desktop PC equipped with an Nvidia RTX 3090 GPU and an Intel Core i9-10850K CPU. The post-processing steps (Section 7.4) add another 700 ms to the total inference time. This is an order of magnitude faster than the Inside Humans approach (Chapter 6), where we reported a total time of approximately 20 s on similar hardware. We measure an inference time of approximately 2 min for the publicly available implementation of OSSO [KZB+22].

In order to modify the 3D scan of a person, we apply the changes made to the latent code as a delta shape manipulation to the scan of the person. By $\mathbf{g}_\theta(\tilde{\mathbf{z}})$ we denote the inference of our decoder for the fitted latent parameters $\tilde{\mathbf{z}}$ to a scan of a person $\mathbf{X}$. For modified latent parameters $\mathbf{z}$ we apply the difference of decoding $\tilde{\mathbf{z}}$ and $\mathbf{z}$ to the scan, resulting in the modified person $\mathbf{X}'$:

$$\mathbf{X}' = \mathbf{X} + (\mathbf{g}_\theta(\mathbf{z}) - \mathbf{g}_\theta(\tilde{\mathbf{z}})) \,. \tag{7.11}$$

To prevent unnatural deformation of the head region, we stitch the original head of the scanned person back onto the resulting mesh using differential coordinates similar to the face region stitching described in Section 5.2.

7.5.6 Experiments with Different Poses

In Figure 7.12, we show examples of fitting our model to poses which differ from our trained A-pose. We follow the OSSO approach [KZB+22] and first unpose a given scan by employing the surface-based template fitting approach of Achenbach et al. [AWL+17], as described in Section 2.2.2. Our anatomical shape model is then fit to the resulting A-pose surface mesh, resulting in an unposed high-resolution skeleton embedding. Since our template skeleton shares its animation rig with the skin surface, we can then use Linear Blend Skinning to repose the resulting skeleton.

Note that after applying the very simplistic Linear Blend Skinning, the resulting skeleton and skin meshes may show self-intersections, as we only resolve these intersections in the A-pose of our template. As Linear Blend Skinning is obviously not an anatomically correct animation technique, more sophisticated animation methods such as proposed in SKEL [KWS+23] or Fast Projective Skinning [KB19] should be incorporated in future work to allow anatomically sound animations based on our model.

Figure 7.12: The results of fitting our model to different scan poses (top row). We first unpose the scan by fitting our human skin surface template. The skeleton inference is then performed in the trained A-pose. Finally, we employ Linear Blend Skinning to repose the results back to the observed scan pose (bottom row). Note that the fingers of the second pose are not faithfully reproduced by the pose estimation of the employed template fitting technique.

7.6 SUMMARY AND LIMITATIONS

This chapter presented a novel approach for learning an anatomically constrained volumetric human shape model. We started by registering an anatomical model of skin and skeletal shape to the European subset of the CAESAR database. This dataset was then extended to the full Cartesian product of skeleton shapes and soft tissue distributions using volumetric deformation transfer, allowing us to transfer the soft tissue distribution of subject i onto the skeleton of subject j in a physically plausible way. To decouple the two shape dimensions, we utilized a Barlow Twins inspired learning approach to train our autoencoder from pairs of skeleton and soft tissue distribution. This learning paradigm enforces similar latent representations of samples in the Cartesian dataset that share either their skeleton shape or soft tissue distribution. The resulting model can be used for shape sampling, e.g., generating various soft tissue distributions on the same skeleton. It provides localized shape manipulation due to the injected measurements in the latent space of our autoencoder, allowing us to modify personalized realistic avatars. Compared to other methods, our model better decouples the skeleton and soft tissue shape dimensions, allows more localized shape manipulation, and provides significantly faster inference time.

Due to the limited availability of such data, our model is not trained on real anatomical data. We do note however, that the data needed for learning our model – different soft tissue distributions on the same skeleton – does not exist as ground truth data. Recent methods have argued that relying on synthetic data alone could also be seen as an advantage and that the trained models can still outperform state-of-the-art methods which are trained on real captured data [WBH+21]. However, evaluating our model on real anatomical data such as CT scans [SBJ+23; SBE+24] or MRI scans [KAD+24] would clearly be desirable in future work.

Our training data lacks information about the bone structure underlying the head, hands, and feet of our subjects. Therefore, our model cannot properly reproduce these areas. Due to this fact, we observe asymmetric structures, especially occurring in the facial region, which are amplified when modifying the latent code. Future work should extend our model in this regard in order to provide a volumetric representation of the complete human body. Similarly, in future work our model can be extended to include other anatomical details such as the muscles. The skeleton, muscle, and soft tissue layers then could be separated by our model applying a triple Barlow twins loss, where pairs of eight are processed in a batch.

As is the case for the method presented in Chapter 6, our training data also lacks diversity. We train our model by fitting to the European subset of the CAESAR database, featuring mostly Caucasian men and women. Incorpo-

rating a more unbiased dataset into the training of volumetric human shape models should be addressed in future work.

Although our resulting meshes are free of self-intersections, this property only holds in the A-pose of our template model. When animating the resulting skeleton and skin, due to our simplistic use of Linear Blend Skinning we cannot guarantee that the skeleton does not protrude the skin. Further investigations into developing an animation method for volumetric virtual humans that avoids self-intersections is an interesting direction for future work.

8 CONCLUSION

This thesis presented surface-based and volumetric models of realistic virtual humans in the context of VR therapy for body image disorders. A large part of our research was embedded into the ViTraS project [ViT24], short for *Virtual Reality Therapy by Stimulation of Modulated Body Perception*. As such, we focused on fully animatable, real-time capable, modulatable, realistic, and personalized virtual humans, ready for virtual mirror exposure as opposed to stylized representations commonly used in VR environments or ultra-high-fidelity virtual humans as used, e.g., in film productions, where real-time rendering is not a concern. To conclude this thesis, we will summarize the main contributions and discuss promising directions for future research on this topic.

We started by introducing our custom-built photogrammetry rig, which is capable of producing high-quality static 3D scans of humans. To generate animatable virtual humans from such scans, we presented a template fitting approach that fits an animatable statistical virtual human body model to the scanned data and generates a high-quality color texture, yielding fully animatable virtual humans ready for use in VR environments. This approach is largely based on previous work but adapted to run in a fully automatic manner. Requiring an elaborate photogrammetry rig for reconstructing virtual humans however limits their availability due to the high hardware costs and stationary scanner setup.

To tackle this problem, we then presented a method for generating realistic virtual humans from smartphone videos. By requiring only two videos taken with commodity smartphones, one depicting the subject's full body and the other featuring a closeup of the subject's head, we greatly reduce the hardware costs compared to existing approaches that reconstruct virtual humans at an adequate fidelity for virtual mirror exposure. From the recorded videos, we extract suitable image frames by means of optical flow analysis and sharpness estimation. The extracted frames are then passed to an off-the-shelve photogrammetry software, yielding two dense 3D point clouds, to which we fit an animatable statistical human template model. To compensate for geometric inaccuracies arising from motion artifacts inherent in the scan process, we texturize the resulting virtual humans by employing a graph cut based texture stitching.

To investigate the perception of the resulting virtual humans, we then conducted a user study, comparing the presented smartphone reconstruction method to virtual humans reconstructed from a 3D scan performed with a high-cost photogrammetry rig. Participants were scanned with both methods and embodied both virtual humans in a VR environment with motion tracking

and virtual mirror exposure. They were then asked to score the similarity, human-likeness, and eeriness of the virtual humans, rate their feeling of virtual body ownership, and state their preference for one of the resulting virtual humans. The results show, that both virtual humans are perceived similarly, as we could find almost no significant differences in the statistical evaluation of the dependent variables. We thus conclude that the presented low-cost method based on video input from commodity smartphones is indeed a viable alternative to high-cost photogrammetry rigs.

To further investigate the potential of virtual humans in the context of VR therapy of body image disorders, we developed a statistical model of body weight modification. The model is learned by first registering a template model to a database of 3D scans annotated with anthropometric measurements such as weight, height, arm span, and inseam. We then build a low-dimensional human body shape model and correlate the resulting subspace with the anthropometric measurements. This allows to map from a change in anthropometrics to a change in body shape via the low-dimensional body shape model. By integrating this statistical model of body weight modification into a VR prototype, we give users the ability to actively control the body weight of their personalized avatar in a virtual mirror exposure setup in real-time. However, the presented model was only trained on surface meshes and is therefore unable to accurately reason about anatomical traits such as body composition.

To address this limitation, we focused on volumetric anatomical representations of virtual humans in the second part of this thesis. First, we presented a layered anatomical model, consisting of a skin, muscle, and skeleton layer with identical topology, enveloping high-resolution muscle and skeleton meshes derived from a high-quality anatomy model. From a data set of 3D scans annotated with data from a medical-grade bioelectrical impedance analysis scale, we learned to infer fat and muscle mass from surface scans. The layered template model is then fit to a given skin surface in a multi-stage optimization scheme while conforming to the inferred body composition. After fitting the layered template model, we employ a space warp based on triharmonic radial basis functions to embed the original high-resolution skeleton and muscle meshes into the fitted model. Given a skin surface conforming to our template topology, we can efficiently generate a personalized anatomical model in a few seconds. We demonstrated the robustness of our method by fitting our template to the European subset of the CAESAR database and showed example applications such as physics-based character animation, fat growth, and fat transfer, which all benefit from – or are made feasible by – our volumetric representation.

Finally, we presented a novel approach for learning an anatomically constrained volumetric model of human skeleton shape and soft tissue variation from the data produced by the anatomy inference method described above. With the use of volumetric deformation transfer, we are able to transfer the

soft tissue distribution of all subjects i onto the skeleton shapes of all subjects j in our database, yielding a Cartesian data set of skeleton shape and soft tissue distribution. Due to the employed self-supervised learning technique, our autoencoder architecture is able to learn separate parameter sets for both input dimensions, providing shape sampling of various soft tissue distributions over the same skeleton shape and vice versa. By additionally concatenating anthropometric measurements to the latent space, we are able to provide semantic localized shape modification. The resulting model is able to infer skeleton shape from a given skin surface in less than a second.

The generation of realistic virtual humans still remains an active field of research. A promising direction for future work on the low-cost generation of virtual humans is to further increase the input quality. The quality of the built-in cameras increases with every new generation of smartphones. Directly capturing images with a dedicated timer-based capture application instead of extracting image frames from videos would additionally improve the input quality due to the higher resolution and less compression artifacts, yielding higher-quality point clouds in the photogrammetry step and improve texture detail in the resulting virtual humans. One big limitation of scanning people with a single camera is the fact that people cannot hold completely still for the duration of the scan. This violates the multi-view stereo assumption, which expects images that capture a scene from different angles at the same moment in time. These unavoidable movements could be compensated for in several ways. One option is to split the set of input images into several consecutive chunks, such that the individual chunks exhibit less movement than the complete set of images, since they cover a smaller period of time. The template fitting method would then be adapted to optimize pose parameters for each chunk separately. Another option could be to semantically segment the input images into various body parts (e.g., arms, legs, torso, and head) and reconstruct separate point clouds for each body part, the assumption being that each individual body part moves approximately rigidly, which can be compensated for by the camera calibration.

Volumetric anatomical representations of virtual humans equally provide opportunity for further research. Due to the limited availability of such data, we did not evaluate our model on real-world data stemming from medical imaging techniques such as DXA, MRI, or CT scans. Therefore, our model remains only anatomically plausible. Future work should (i) include evaluations against ground truth data, and (ii) be trained on a more diverse set of human body shapes. Furthermore, our volumetric representation excludes the head, hands, and feet, which should be incorporated to provide a volumetric representation of the complete body in future work. Our volumetric fitting is informed by estimations of fat and muscle mass. The linear regressor, which provides these estimates, is trained on a small synthetic data set [MMK+21], which could either be extended or replaced by a more sophisticated model of

body composition (see, e.g., [KAD+24]). The presented statistical volumetric human shape model was limited to skeleton shape and soft tissue distribution. Further anatomical structures, such as muscles, should be incorporated in future work, as this would give more fine-grained control to applications which aim to separately change the distribution of muscle and fat tissue. Additionally, we only resolve self-intersections in the A-pose of our template model. Developing an animation method for volumetric virtual humans which upholds this property for any pose is a promising direction of future work.

We believe that further improving both low-cost avatar generation and volumetric anatomical representations of virtual humans will not only prove to be beneficial for further research in VR therapy of body image disorders, but also benefit other applications. After evaluating and improving the anatomical accuracy of our volumetric model, it can provide valuable information for sports science or medical applications, where estimations of muscle mass and bone structure could, e.g., be used for computing forces in biomechanical analysis simulations. Virtual try-on methods would benefit from simulating the interaction of fabric with our predicted soft tissue distribution and skeletal shape, as this would improve the accuracy of cloth fitting and draping simulations.

BIBLIOGRAPHY

[3DS24] 3DScanstore. https://www.3dscanstore.com. 2024.

[ABG+18] Jascha Achenbach, Robert Brylka, Thomas Gietzen, Katja Zum Hebel, Elmar Schömer, Ralf Schulze, Mario Botsch, and Ulrich Schwanecke. "A Multilinear Model for Bidirectional Craniofacial Reconstruction". In *Proc. of Eurographics Workshop on Visual Computing for Biology and Medicine*. 2018, pp. 67–76.

[ABY+17] Anil Didem Aydin Kabakci, Mustafa Buyukmumcu, Mehmet Tugrul Yilmaz, Aynur Emine Cicekcibasi, Duygu Akin Saygin, and Emine Cihan. "An Osteometric Study on Humerus". *International Journal of Morphology* 35.1 (2017), pp. 219–226.

[Ack98] Michael J. Ackerman. "The Visible Human Project". *Proc. of the IEEE* 86.3 (1998), pp. 504–511.

[ACP03] Brett Allen, Brian Curless, and Zoran Popović. "The Space of Human Body Shapes: Reconstruction and Parameterization from Range Scans". *ACM Transactions on Graphics* 22.3 (2003), pp. 587–594.

[Agi24] Agisoft. *Metashape*. https://www.agisoft.com. 2024.

[AKB14] Laura Aymerich-Franch, René F. Kizilcec, and Jeremy N. Bailenson. "The Relationship between Virtual Self Similarity and Social Anxiety". *Frontiers in Human Neuroscience* 8 (2014), p. 944.

[ALC+18] A. Aristidou, J. Lasenby, Y. Chrysanthou, and A. Shamir. "Inverse Kinematics Techniques in Computer Graphics: A Survey". *Computer Graphics Forum* 37.6 (2018), pp. 35–58.

[Ali24] AliceVision. https://alicevision.org/#meshroom. 2024.

[AMB+19] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. "Learning to Reconstruct People in Clothing from a Single RGB Camera". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1175–1186.

[AMX+18a] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. "Video Based Reconstruction of 3D People Models". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8387–8397.

[AMX+18b] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. "Detailed Human Avatars from Monocular Video". In *Proc. of the International Conference on 3D Vision (3DV)*. 2018, pp. 98–109.

[APB+00] Mariano Alcañiz, Conxa Perpiñá, Rosa Baños, José Antonio Lozano, Javier Montesa, Cristina Botella, A. Garcia Palacios, Helena Villa, and J. Alozano. "A New Realistic 3D Body Representation in Virtual Environments for The Treatment of Disturbed Body Image in Eating Disorders". *Cyberpsychology & Behavior* 3.3 (2000), pp. 433–439.

[APB+20] Dmitry Alexandrovsky, Susanne Putze, Michael Bonfert, Sebastian Höffner, Pitt Michelmann, Dirk Wenig, Rainer Malaka, and Jan David Smeddinck. "Examining Design Choices of Questionnaires in VR User Studies". In *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–21.

[App24] Apple. *Introducing Object Capture*. https://developer.apple.com/augmented-reality/object-capture. 2024.

[APT+19] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. "Tex2Shape: Detailed Full Human Body Geometry from a Single Image". In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 2293–2303.

[ASK+05] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. "SCAPE: Shape Completion and Animation of People". *ACM Transactions on Graphics* 24.3 (2005), pp. 408–416.

[ASW+15] Jessica M. Alleva, Paschal Sheeran, Thomas L. Webb, Carolien Martijn, and Eleanor Miles. "A Meta-Analytic Review of Stand-Alone Interventions to Improve Body Image". *PLOS ONE* 10.9 (2015).

[Aut24] Autodesk. *Character Generator*. https://charactergenerator.autodesk.com. 2024.

[AWL+17] Jascha Achenbach, Thomas Waltemate, Marc Erich Latoschik, and Mario Botsch. "Fast Generation of Realistic Virtual Humans". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2017, 12:1–12:10.

[AZB15] Jascha Achenbach, Eduard Zell, and Mario Botsch. "Accurate Face Reconstruction through Anisotropic Fitting and Eye Correction". In *Vision, Modeling, and Visualization*. 2015, pp. 1–8.

[AZS22] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. "Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 1506–1515.

[BBB+10] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. "High-Quality Single-Shot Capture of Facial Geometry". *ACM Transactions on Graphics* 29.4 (2010), 40:1–40:9.

[BBL+15] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. "Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences". In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2015, pp. 2300–2308.

[BBP+19] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. "Neural 3D Morphable Models: Spiral Convolutional Networks for 3D Shape Representation Learning and Generation". In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 7213–7222.

[BDS+12] Sofien Bouaziz, Mario Deuss, Yuliy Schwartzburg, Thibaut Weise, and Mark Pauly. "Shape-Up: Shaping Discrete Geometry with Projections". *Computer Graphics Forum* 31.5 (2012), pp. 1657–1667.

[BEB12] Tyson Brochu, Essex Edwards, and Robert Bridson. "Efficient Geometrically Exact Continuous Collision Detection". *ACM Transactions on Graphics* 31.4 (2012), 96:1–96:7.

[BGA+63] Josef Brožek, Francisco Grande, Joseph T. Anderson, and Ancel Keys. "Densitometric Analysis of Body Composition: Revision of some Quantitative Assumptions". *Annals of the New York Academy of Sciences* 110.1 (1963), pp. 113–140.

[BGS13] Domna Banakou, Raphaela Groten, and Mel Slater. "Illusory Ownership of a Virtual Child Body Causes Overestimation of Object Sizes and Implicit Attitude Changes". *Proc. of the National Academy of Sciences* 110.31 (2013), pp. 12846–12851.

[BHP+10] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. "High Resolution Passive Facial Performance Capture". *ACM Transactions on Graphics* 29.4 (2010), 41:1–41:10.

[BHS16] Domna Banakou, Parasuram D. Hanumanthu, and Mel Slater. "Virtual Embodiment of White People in a Black Virtual Body Leads to a Sustained Reduction in Their Implicit Racial Bias". *Frontiers in Human Neuroscience* 10 (2016), p. 601.

[BK04] Yuri Boykov and Vladimir Kolmogorov. "Experimental Comparison of Min-Cut/Max-Flow Algorithms for An Energy Minimization in Vision". *ACM Transactions on Pattern Analysis and Machine Intelligence* 26.9 (2004), pp. 1124–1137.

[BK05] Mario Botsch and Leif Kobbelt. "Real-Time Shape Editing Using Radial Basis Functions". *Computer Graphics Forum* 24.3 (2005), pp. 611–621.

[BKL+16] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. "Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image". In *Computer Vision – ECCV*. 2016, pp. 561–578.

[BKP+10] Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy. "Polygon Mesh Processing". AK Peters, CRC press, 2010.

[BM92] Paul J. Besl and Neil D. McKay. "A Method for Registration of 3-D Shapes". *ACM Transactions on Pattern Analysis and Machine Intelligence* 14.2 (1992), pp. 239–256.

[BML+14] Sofien Bouaziz, Sebastian Martin, Tiantian Liu, Ladislav Kavan, and Mark Pauly. "Projective Dynamics: Fusing Constraint Projections for Fast Simulation". *ACM Transactions on Graphics* 33.4 (2014), 154:1–154:11.

[Bra00] Gary Bradski. "The OpenCV Library". *Dr. Dobb's Journal of Software Tools* 25 (2000), pp. 120–125.

[BRL+14] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. "FAUST: Dataset and Evaluation for 3D Mesh Registration". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3794–3801.

[Bro71] Duane C. Brown. "Close-Range Camera Calibration". *Photogrammetric Engineering* 37.8 (1971), pp. 855–866.

[BRP+17] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. "Dynamic FAUST: Registering Human Bodies in Motion". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6233–6242.

[BS08] Mario Botsch and Olga Sorkine. "On Linear Variational Surface Deformation Methods". *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 14.1 (2008), pp. 213–230.

[BSH+05] Jeremy N Bailenson, Kim Swinth, Crystal Hoyt, Susan Persky, Alex Dimov, and Jim Blascovich. "The Independent and Interactive Effects of Embodied-Agent Appearance and Behavior on Self-Report, Cognitive, and Behavioral Markers of Copresence in Immersive Virtual Environments". *Presence: Teleoperators and Virtual Environments* 14.4 (2005), pp. 379–393.

[BSP+06] Mario Botsch, Robert Sumner, Mark Pauly, and Markus Gross. "Deformation Transfer for Detail-Preserving Surface Editing". *Vision, Modeling, and Visualization* (2006), pp. 357–364.

[BSR+08] Stéphane Bouchard, Julie St-Jacques, Geneviève Robillard, and Patrice Renaud. "Anxiety Increases the Feeling of Presence in Virtual Reality". *Presence: Teleoperators and Virtual Environments* 17.4 (2008), pp. 376–391.

[BTP14] Sofien Bouaziz, Andrea Tagliasacchi, and Mark Pauly. "Dynamic 2D/3D Registration". In *Eurographics Tutorials*. 2014, pp. 1–17.

[BV99] Volker Blanz and Thomas Vetter. "A morphable model for the synthesis of 3D faces". In *Proc. of the ACM on Computer Graphics and Interactive Techniques*. 1999, pp. 187–194.

[BVZ01] Yuri Boykov, Olga Veksler, and Ramin Zabih. "Fast Approximate Energy Minimization via Graph Cuts". *ACM Transactions on Pattern Analysis and Machine Intelligence* 23.11 (2001), pp. 1222–1239.

[BWS+21] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. "Driving-Signal Aware Full-Body Avatars". *ACM Transactions on Graphics* 40.4 (2021), 143:1–143:17.

[BWW+21] Andrea Bartl, Stephan Wenninger, Erik Wolf, Mario Botsch, and Marc Erich Latoschik. "Affordable but not Cheap: A Case Study of the Effects of Two 3D-Reconstruction Methods of Virtual Humans". *Frontiers in Virtual Reality* 2 (2021).

[BZH+23] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. "FLARE: Fast Learning of Animatable and Relightable Mesh Avatars". *ACM Transactions on Graphics* 42.6 (2023), 204:1–204:15.

[CHS+21] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Real-Time Multi-Person 2D Pose Estimation Using Part Affinity Fields". *ACM Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2021), pp. 172–186.

[CKH+09] Andreas Christ, Wolfgang Kainz, Eckhart G. Hahn, Katharina Honegger, Marcel Zefferer, Esra Neufeld, Wolfgang Rascher, Rolf Janka, Werner Bautz, Ji Chen, Berthold Kiefer, Peter Schmitt, Hans-Peter Hollenbach, Jianxiang Shen, Michael Oberle, Dominik Szczerba, Anthony Kam, Joshua W. Guag, and Niels Kuster. "The Virtual Family – Development of Surface-Based Anatomical Models of Two Adults and Two Children for Dosimetric Simulations". *Physics in Medicine & Biology* 55.2 (2009), pp. 23–38.

[Coh77] Jacob Cohen. "Statistical Power Analysis for the Behavioral Sciences". Academic Press, 1977.

[COL24] COLMAP. https://colmap.github.io. 2024.

[CSK+22] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. "Authentic Volumetric Avatars from a Phone Scan". *ACM Transactions on Graphics* 41.4 (2022), 163:1–163:19.

[CWL+24] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. "MonoGaussianAvatar: Monocular Gaussian Point-based Head Avatar". In *Proc. of the ACM SIGGRAPH Conference*. 2024, 58:1–58:9.

[CZ24] Prashanth Chandran and Gaspard Zoss. "Anatomically Constrained Implicit Face Models". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 2220–2229.

[CZP+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In *Computer Vision – ECCV*. 2018, pp. 833–851.

[Dar24] Darktable. https://www.darktable.org. 2024.

[DB13] Crispin Deul and Jan Bender. "Physically-Based Character Skinning". In *Proc. of Virtual Reality Interactions and Physical Simulations*. 2013, pp. 25–34.

[DCV+06] Sven De Greef, Peter Claes, Dirk Vandermeulen, Wouter Mollemans, Paul Suetens, and Guy Willems. "Large-Scale in-vivo Caucasian Facial Soft Tissue Thickness Database for Craniofacial Reconstruction". *Forensic Science International* 159 (2006), pp. 126–146.

[DDB+15] Mario Deuss, Anders Holden Deleuran, Sofien Bouaziz, Bailin Deng, Daniel Piker, and Mark Pauly. "ShapeOp – A Robust and Extensible Geometric Modelling Paradigm". In *Proc. of Design Modelling Symposium*. 2015, pp. 505–515.

[DDF+17] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. "Motion2fusion: Real-Time Volumetric Performance Capture". *ACM Transactions on Graphics* 36.6 (2017), 246:1–246:16.

[DGW+22] Nina Döllinger, Christopher Göttfert, Erik Wolf, David Mal, Marc Erich Latoschik, and Carolin Wienrich. "Analyzing Eye Tracking Data in Mirror Exposure". In *Proc. of Mensch Und Computer*. 2022, pp. 513–517.

[DHT+00] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. "Acquiring the reflectance field of a human face". In *Proc. of the ACM on Computer Graphics and Interactive Techniques*. 2000, pp. 145–156.

[DJ94] Marie-Pierre Dubuisson and Anil K. Jain. "A Modified Hausdorff Distance for Object Matching". In *Proc. of the International Conference on Pattern Recognition*. 1994, pp. 566–568.

[DLG+13] Ali-Hamadi Dicko, Tiantian Liu, Benjamin Gilles, Ladislav Kavan, François Faure, Olivier Palombi, and Marie-Paule Cani. "Anatomy Transfer". *ACM Transactions on Graphics* 32.6 (2013), 188:1–188:8.

[DMS+99] Mathieu Desbrun, Mark Meyer, Peter Schröder, and Alan H. Barr. "Implicit Fairing of Irregular Meshes Using Diffusion and Curvature Flow". In *Proc. of the ACM on Computer Graphics and Interactive Techniques*. SIGGRAPH '99. 1999, pp. 317–324.

[DSK08] Manisha R. Dayal, Maryna Steyn, and Kevin L. Kuykendall. "Stature Estimation from Bones of South African Whites". *South African Journal of Science* 104.3-4 (2008), pp. 124–128.

[DUD+10] Aurélie Docteur, Isabel Urdapilleta, Cécile Defrance, and Jocelyne Raison. "Body Perception and Satisfaction in Obese, Severely Obese, and Normal Weight Female Patients". *Obesity* 18.7 (2010), pp. 1464–1465.

[DWM+22] Nina Döllinger, Erik Wolf, David Mal, Stephan Wenninger, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Resize Me! Exploring the User Experience of Embodied Realistic Modulatable Avatars for Body Image Intervention in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022).

[DWW+19] Nina Döllinger, Carolin Wienrich, Erik Wolf, and Marc Erich Latoschik. "ViTraS – Virtual Reality Therapy by Stimulation of Modulated Body Image – Project Outline". In *Mensch und Computer – Workshopband*. 2019, pp. 606–611.

[EST+20] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. "3D Morphable Face Models – Past, Present and Future". *ACM Transactions on Graphics* 39.5 (2020), 157:1–157:38.

[Far03] Gunnar Farnebäck. "Two-Frame Motion Estimation Based on Polynomial Expansion". In *Image Analysis*. 2003, pp. 363–370.

[FB12] Oren Freifeld and Michael J. Black. "Lie Bodies: A Manifold Representation of 3D Human Shape". In *Computer Vision – ECCV*. 2012, pp. 1–14.

[FBT13] Jesse Fox, Jeremy N. Bailenson, and Liz Tricase. "The Embodiment of Sexualized Virtual Selves: The Proteus Effect and Experiences of Self-Objectification via Avatars". *Computers in Human Behavior* 29.3 (2013), pp. 930–938.

[FFB+21] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. "Learning an Animatable Detailed 3D Face Model from In-the-Wild Images". *ACM Transactions on Graphics* 40.4 (2021), 88:1–88:13.

[FGC+09] Marta Ferrer-Garcia, José Gutiérrez-Maldonado, Alejandra Caqueo-Urizar, and Elena Moreno. "The Validity of Virtual Environments for Eliciting Emotional Responses in Patients with Eating Disorders and in Controls". *Behavior Modification* 33.6 (2009), pp. 830–854.

[FGM02] David A. Fields, Michael I. Goran, and Megan A. McCrory. "Body-Composition Assessment via Air-Displacement Plethysmography in Adults and Children: A Review". *The American Journal of Clinical Nutrition* 75.3 (2002), pp. 453–467.

[FGR13] Marta Ferrer-Garcia, José Gutiérrez-Maldonado, and Giuseppe Riva. "Virtual Reality Based Treatments in Eating Disorders and Obesity: A Review". *Journal of Contemporary Psychotherapy* 43.4 (2013), pp. 207–221.

[Fit24] Fit3D. *Fit3D Scanner Systems*. https://www.fit3d.com. 2024.

[FM21] Guo Freeman and Divine Maloney. "Body, Avatar, and Me: The Presentation and Perception of Self in Social Virtual Reality". In *Proc. of the ACM on Human-Computer Interaction*. Vol. 4. 2021, pp. 1–27.

[FPM+18] Marta Ferrer-Garcia, Bruno Porras-Garcia, Manuel Moreno, Paola Bertomeu, and José Gutiérrez-Maldonado. "Embodiment in different size virtual bodies produces changes in women's body image distortion and dissatisfaction". *Annual Review of CyberTherapy and Telemedicine* 16 (2018), pp. 111–117.

[FRS17] Andrew Feng, Evan Suma Rosenberg, and Ari Shapiro. "Just-in-Time, Viable, 3-D Avatars from Scans". *Computer Animation and Virtual Worlds* 28 (2017), pp. 3–4.

[FSL06] Clare Farrell, Roz Shafran, and Michelle Lee. "Empirically Evaluated Treatments for Body Image Disturbance: A Review". *European Eating Disorders Review* 14.5 (2006), pp. 289–300.

[FWD+24] Marie Luisa Fiedler, Erik Wolf, Nina Döllinger, David Mal, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "From Avatars to Agents: Self-Related Cues through Embodiment and Personalization Affect Body Perception in Virtual Reality". *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 30.11 (2024), pp. 7386–7396.

[FYR+19] Xianyong Fang, Jikui Yang, Jie Rao, Linbo Wang, and Zhigang Deng. "Single RGB-D Fitting: Total Human Modeling with an RGB-D Shot". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2019, 24:1–24:11.

[GAK+20] Ivan Grischenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. "Attention Mesh: High-Fidelity Face Mesh Prediction in Real-Time". *CVPR Workshop on Computer Vision for Augmented and Virtual Reality* (2020).

[GBA+19] Thomas Gietzen, Robert Brylka, Jascha Achenbach, Katja Zum Hebel, Elmar Schömer, Mario Botsch, Ulrich Schwanecke, and Ralf Schulze. "A Method for Automatic Forensic Facial Reconstruction Based on Dense Statistics of Soft Tissue Thickness". *PLOS ONE* 14.1 (2019).

[GCB+19] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. "SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator". In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019, pp. 4141–4148.

[GCH+19] Geoffrey Gorisse, Olivier Christmann, Samory Houzangbe, and Simon Richir. "From Robot to Virtual Doppelganger: Impact of Visual Fidelity of Avatars Controlled in Third-Person Perspective on Embodiment and Behavior in Immersive Virtual Environments". *Frontiers in Robotics and AI* 6 (2019), p. 8.

[GFT+11] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. "Multiview Face Capture Using Polarized Spherical Gradient Illumination". *ACM Transactions on Graphics* 30.6 (2011), pp. 1–10.

[GJ+24] Gaël Guennebaud, Benoît Jacob, et al. *Eigen v3*. https://eigen.tuxfamily.org. 2024.

[GJC+23] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. "Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 12858–12868.

[GKG+23] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. "Learning Neural Parametric Head Models". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 21003–21012.

[GLD+19] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. "The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting". *ACM Transactions on Graphics* 38.6 (2019), 217:1–217:19.

[GNH18] Trevor C. Griffen, Eva Naumann, and Tom Hildebrandt. "Mirror Exposure Therapy for Body Image Disturbances and Eating Disorders: A Review". *Clinical Psychology Review* 65 (2018), pp. 163–174.

[GNK18] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. "DensePose: Dense Human Pose Estimation in the Wild". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7297–7306.

[GPh24] GPhoto. http://www.gphoto.org. 2024.

[GSV+03] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. "The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment". In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. 2003, pp. 529–536.

[GWO+10] Ran Gal, Yonatan Wexler, Eyal Ofek, Hugues Hoppe, and Daniel Cohen-Or. "Seamless Montage for Texturing Models". *Computer Graphics Forum* 29.2 (2010), pp. 479–486.

[GXY+17] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. "Real-Time Geometry, Albedo, and Motion Reconstruction Using a Single RGB-D Camera". *ACM Transactions on Graphics* 36.4 (2017), 44:1–44:13.

[HH16] Irwin Hudson and Jonathan Hurter. "Avatar Types Matter: Review of Avatar Literature for Performance Purposes". In *International Conference on Virtual, Augmented and Mixed Reality*. 2016, pp. 14–21.

[HHL24] Shoukang Hu, Tao Hu, and Ziwei Liu. "GauHuman: Articulated Gaussian Splatting from Monocular Human Videos". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 20418–20431.

[HHM+20] Maria Horne, Andrew J. Hill, Trevor Murells, Hassan Ugail, Rajeswaran Chinnadorai, and Maryann L. Hardy. "Using avatars in weight management settings: A systematic review". *Internet Interventions* 19 (2020).

[HHT+11] Steven B. Heymsfield, Moonseong Heo, Diana Thomas, and Angelo Pietrobelli. "Scaling of Body Composition to Height: Relevance to Height-Normalized Indexes". *The American Journal of Clinical Nutrition* 93.4 (2011), pp. 736–740.

[HLP+00] Ding He, Fuhu Liu, Dave Pape, Greg Dawe, and Dan Sandin. "Video-Based Measurement of System Latency". In *International Immersive Projection Technology Workshop*. 2000.

[HLZ+20] Geoffrey M. Hudson, Yao Lu, Xiaoke Zhang, James Hahn, Johannah E. Zabal, Finza Latif, and John Philbeck. "The Development of a BMI-Guided Shape Morphing Technique and the Effects of an Individualized Figure Rating Scale on Self-Perception of Body Size". *European Journal of Investigation in Health, Psychology and Education* 10.2 (2020), pp. 579–594.

[HM10] Chin-Chang Ho and Karl F. MacDorman. "Revisiting the Uncanny Valley Theory: Developing and Validating an Alternative to the Godspeed Indices". *Computers in Human Behavior* 26.6 (2010), pp. 1508–1518.

[HM17] Chin-Chang Ho and Karl F. MacDorman. "Measuring the Uncanny Valley Effect". *International Journal of Social Robotics* 9.1 (2017), pp. 129–139.

[HMP08] Chin-Chang Ho, Karl F. MacDorman, and Z. A. Dwi Pramono. "Human Emotion and the Uncanny Valley: A GLM, MDS, and Isomap Analysis of Robot Video Ratings". In *Proc. of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2008, pp. 169–176.

[Hor87] Berthold K. P. Horn. "Closed-Form Solution of Absolute Orientation Using Unit Quaternions". *Journal of the Optical Society of America A* 4.4 (1987), pp. 629–642.

[HSS+09] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. "A Statistical Model of Human Pose and Body Shape". *Computer Graphics Forum* 28.2 (2009), pp. 337–346.

[HXZ+19] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. "LiveCap: Real-Time Human Performance Capture from Monocular Video". *ACM Transactions on Graphics* 38.2 (2019), 14:1–14:17.

[HyL24] Hybrid Learning Center (HyLeC). https://hylec.tu-dortmund.de. 2024.

[HZZ+24] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. "Gaussian-Avatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 634–644.

[IBP15] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. "Dynamic 3D Avatar Creation from Hand-held Video Input". *ACM Transactions on Graphics* 34.4 (2015), 45:1–45:14.

[IKK+17] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. "Phace: Physics-Based Face Modeling and Animation". *ACM Transactions on Graphics* 36.4 (2017), 153:1–153:14.

[IKN+16] Alexandru-Eugen Ichim, Ladislav Kavan, Merlin Nimier-David, and Mark Pauly. "Building and Animating User-Specific Volumetric Face Rigs". In *Proc. of the ACM SIGGRAPH/Eurographics symposium on Computer animation*. 2016, pp. 107–117.

[JCS+23] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. "InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 16922–16932.

[JGK+24] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. "MultiPly: Reconstruction of Multiple People from Monocular Video in the Wild". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 109–118.

[JHB+22] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. "SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5605–5615.

[JKK17] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. "Effects of Avatar and Background Types on Users' Co-Presence and Trust for Mixed Reality-Based Teleconference Systems". In *Proceedings the 30th Conference on Computer Animation and Social Agents*. 2017, pp. 27–36.

[JP85] Andrew S. Jackson and Michael L. Pollock. "Practical Assessment of Body Composition". *The Physician and Sportsmedicine* 13.5 (1985), pp. 76–90.

[JSB+08] Alexandra M. Johnstone, Arthur D. Stewart, P. J. Benson, Maria Kalafati, L. Rectenwald, and Graham Horgan. "Assessment of Body Image in Obesity Using a Digital Morphing Technique". *Journal of Human Nutrition and Dietetics* 21.3 (2008), pp. 256–267.

[JSS18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. "Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8320–8329.

[JSW05] Tao Ju, Scott Schaefer, and Joe Warren. "Mean Value Coordinates for Closed Triangular Meshes". *ACM Transactions on Graphics* 24.3 (2005), pp. 561–566.

[KAB20] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. "VIBE: Video Inference for Human Body Pose and Shape Estimation". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5253–5263.

[KAD+24] Marilyn Keller, Vaibhav Arora, Abdelmouttaleb Dakri, Shivam Chandhok, Jürgen Machann, Andreas Fritsche, Michael J. Black, and Sergi Pujades. "HIT: Estimating Internal Human Implicit

Tissues from the Body Surface". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 3480–3490.

[KB15] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In *International Conference on Learning Representations (ICLR)*. 2015.

[KB18] Martin Komaritzan and Mario Botsch. "Projective Skinning". *Proc. of the ACM on Computer Graphics and Interactive Techniques* 1.1 (2018), 12:1–12:19.

[KB19] Martin Komaritzan and Mario Botsch. "Fast Projective Skinning". In *Proc. of the ACM SIGGRAPH Conference on Motion, Interaction and Games*. 2019, 22:1–22:10.

[KBJ+18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. "End-to-End Recovery of Human Shape and Pose". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7122–7131.

[KEH+16] Anouk Keizer, Annemarie van Elburg, Rossa Helms, and H. Chris Dijkerman. "A Virtual Reality Full Body Illusion Improves Body Image Disturbance in Anorexia Nervosa". *PLOS ONE* 11.10 (2016).

[KFM+15] Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. "A Review of Empirical Evidence on Different Uncanny Valley Hypotheses: Support for Perceptual Mismatch as One Road to the Valley of Eeriness". *Frontiers in Psychology* 6 (2015), p. 390.

[KGS+01] Ursula G Kyle, Laurence Genton, Daniel O Slosman, and Claude Pichard. "Fat-Free and Fat Mass Percentiles in 5225 Healthy Subjects Aged 15 to 98 Years". *Nutrition* 17.7-8 (2001), pp. 534–541.

[KGS12] Konstantina Kilteni, Raphaela Groten, and Mel Slater. "The Sense of Embodiment in Virtual Reality". *Presence: Teleoperators and Virtual Environments* 21.4 (2012), pp. 373–387.

[KIL+16] Petr Kadleček, Alexandru-Eugen Ichim, Tiantian Liu, Jaroslav Křivánek, and Ladislav Kavan. "Reconstructing Personalized Anatomical Models for Physics-based Body Animation". *ACM Transactions on Graphics* 35.6 (2016), 213:1–213:13.

[KKL+23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". *ACM Transactions on Graphics* 42.4 (2023), 139:1–139:14.

[KKS+20] Martin Kocur, Melanie Kloss, Valentin Schwind, Christian Wolff, and Niels Henze. "Flexing Muscles in Virtual Reality: Effects of Avatars' Muscular Appearance on Physical Performance". In *Proc. of the Annual Symposium on Computer-Human Interaction in Play*. 2020, pp. 193–205.

[KKW+24] Maria Korosteleva, Timur Levent Kesdogan, Stephan Wenninger, Fabian Kemper, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine. "GarmentCodeData: A Dataset of 3D Made-to-Measure Garments with Sewing Patterns". *Computer Vision – ECCV* (2024).

[KLB+93] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness". *The International Journal of Aviation Psychology* 3.3 (1993), pp. 203–220.

[KPC+18] Hyun K. Kim, Jaehyun Park, Yeongcheol Choi, and Mungyeong Choe. "Virtual Reality Sickness Questionnaire (VRSQ): Motion Sickness Measurement Index in a Virtual Reality Environment". *Applied Ergonomics* 69 (2018), pp. 66–73.

[KPP+17] Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Seungbae Bang, Jinwook Kim, Michael J. Black, and Sung-Hee Lee. "Data-driven Physics for Human Soft Tissue Animation". *ACM Transactions on Graphics* 36.4 (2017), 54:1–54:12.

[KQG+23] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. "NeRSemble: Multi-View Radiance Field Reconstruction of Human Heads". *ACM Transactions on Graphics* 42.4 (2023), 161:1–161:14.

[KS14] Vahid Kazemi and Josephine Sullivan. "One Millisecond Face Alignment with an Ensemble of Regression Trees". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1867–1874.

[KSL+22] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. "Realistic One-Shot Mesh-Based Head Avatars". In *Computer Vision – ECCV*. 2022, pp. 345–362.

[KWB21] Martin Komaritzan, Stephan Wenninger, and Mario Botsch. "Inside Humans: Creating a Simple Layered Anatomical Model from Human Surface Scans". *Frontiers in Virtual Reality* 2 (2021).

[KWS+23] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, and Michael J. Black. "From Skin to Skeleton: Towards Biomechanically Accurate 3D Digital Humans". *ACM Transactions on Graphics* 42.6 (2023), 253:1–253:12.

[KZ04] Vladimir Kolmogorov and Ramin Zabih. "What Energy Functions can be Minimized via Graph Cuts?" *ACM Transactions on Pattern Analysis and Machine Intelligence* 26.2 (2004), pp. 147–159.

[KZB+22] Marilyn Keller, Silvia Zuffi, Michael J. Black, and Sergi Pujades. "OSSO: Obtaining Skeletal Shape from Outside". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20460–20469.

[LAR+14] J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. "Practice and Theory of Blendshape Facial Models". In *Eurographics State of the Art Reports*. 2014.

[LBB+17] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. "Learning a Model of Facial Shape and Expression from 4D Scans". *ACM Transactions on Graphics* 36.6 (2017), 194:1–194:17.

[LBW+15] Yunpeng Liu, Stephan Beck, Renfang Wang, Jin Li, Huixia Xu, Shijie Yao, Xiaopeng Tong, and Bernd Froehlich. "Hybrid Lossless-Lossy Compression for Real-Time Depth-Sensor Streams in 3D Telepresence Applications". In *Advances in Multimedia Information Processing – PCM 2015*. 2015, pp. 442–452.

[LI07] Victor S. Lempitsky and Denis V. Ivanov. "Seamless Mosaicing of Image-Based Texture Maps". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–6.

[Lib24] Libraw. https://www.libraw.org. 2024.

[LKS+19] Marc Erich Latoschik, Florian Kern, Jan-Philipp Stauffert, Andrea Bartl, Mario Botsch, and Jean-Luc Lugrin. "Not Alone Here?! Scalability and User Experience of Embodied Ambient Crowds in Distributed Social Virtual Reality". *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 25.5 (2019), pp. 2134–2144.

[LLL15a] Jean-Luc Lugrin, Maximilian Landeck, and Marc Erich Latoschik. "Avatar Embodiment Realism and Virtual Fitness Training". In *Proc. of the IEEE Conference on Virtual Reality (VR)*. 2015, pp. 225–226.

[LLL15b] Jean-Luc Lugrin, Johanna Latt, and Marc Erich Latoschik. "Avatar Anthropomorphism and Illusion of Body Ownership in VR". In *Proc. of the IEEE Conference on Virtual Reality (VR)*. 2015, pp. 229–230.

[LMR+15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. "SMPL: A Skinned Multi-Person Linear Model". *ACM Transactions on Graphics* 34.6 (2015), 248:1–248:16.

[LRG+17] Marc Erich Latoschik, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. "The Effect of Avatar Realism in Immersive Social Virtual Realities". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2017, 39:1–39:10.

[LW22a] Marc Erich Latoschik and Carolin Wienrich. "Coherence and Plausibility, not Presence?! Pivotal Conditions for XR Experiences and Effects, a Novel Model". *Frontiers in Virtual Reality* 3 (2022).

[LW22b] Marc Erich Latoschik and Carolin Wienrich. "Congruence and Plausibility, Not Presence: Pivotal Conditions for XR Experiences and Effects, a Novel Approach". *Frontiers in Virtual Reality* 3 (2022).

[LWB+15] Jean-Luc Lugrin, Maximilian Wiedemann, Daniel Bieberstein, and Marc Erich Latoschik. "Influence of Avatar Realism on Stressful Situation in VR". In *Proc. of the IEEE Conference on Virtual Reality (VR)*. 2015, pp. 227–228.

[LWP10] Hao Li, Thibaut Weise, and Mark Pauly. "Example-Based Facial Rigging". *ACM Transactions on Graphics* 29.4 (2010), 32:1–32:6.

[LZX+23] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. "High-Fidelity Clothed Avatar Reconstruction from a Single Image". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 8662–8672.

[MBB12] Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. "Render Me Real? Investigating the Effect of Render Style on the Perception of Animated Virtual Humans". *ACM Transactions on Graphics* 31.4 (2012), pp. 1–11.

[MBC+16] Matthias Müller, Jan Bender, Nuttapong Chentanez, and Miles Macklin. "A Robust Method to Extract the Rotational Part of Deformations". In *Proc. of the International Conference on Motion in Games*. 2016, pp. 55–60.

[MBL22] Timo Menzel, Mario Botsch, and Marc Erich Latoschik. "Automated Blendshape Personalization for Faithful Face Animations Using Commodity Smartphones". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2022, 22:1–22:9.

[MC18] Angela Meadows and Rachel M. Calogero. "Studies on Weight Stigma and Body Image in Higher-Weight Individuals". In *Body Image, Eating, and Weight*. Springer International Publishing, 2018, pp. 381–400.

[MDW+24] David Mal, Nina Döllinger, Erik Wolf, Stephan Wenninger, Mario Botsch, Carolin Wienrich, and Marc Erich Latoschik. "Am I the Odd One? Exploring (In)Congruencies in the Realism of Avatars and Virtual Others in Virtual Reality". *Frontiers in Virtual Reality* 5 (2024).

[MGP07] Niloy J. Mitra, Leonidas Guibas, and Mark Pauly. "Symmetrization". *ACM Transactions on Graphics* 3 (2007), 63:1–63:8.

[MGT+19] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. "AMASS: Archive of Motion Capture As Surface Shapes". In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 5442–5451.

[MKK+17] Charles Malleson, Maggie Kosek, Martin Klaudiny, Ivan Huerta, Jean-Charles Bazin, Alexander Sorkine-Hornung, Mark Mine, and Kenny Mitchell. "Rapid One-Shot Acquisition of Dynamic VR Avatars". In *Proc. of the IEEE Conference on Virtual Reality (VR)*. 2017, pp. 131–140.

[MMB+08] Katerina Maximova, John J. McGrath, Tracie Barnett, Jennifer O'Loughlin, Gilles Paradis, and Martin Lambert. "Do You See What I See? Weight Status Misperception and Exposure to Obesity Among Children and Adolescents". *International Journal of Obesity* 32.6 (2008), pp. 1008–1015.

[MMK+21] Nadia Maalin, Sophie Mohamed, Robin S. S. Kramer, Piers L. Cornelissen, Daniel Martin, and Martin J. Tovée. "Beyond BMI for Self-Estimates of Body Size and Shape: A New Method for Developing Stimuli Correctly Calibrated for Body Composition". *Behavior Research Methods* 53.3 (2021), pp. 1308–1321.

[MMK12] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. "The Uncanny Valley [From the Field]". *IEEE Robotics & Automation Magazine* 19.2 (2012), pp. 98–100.

[MSS24] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. "Expressive Whole-Body 3D Gaussian Avatar". In *Computer Vision – ECCV*. 2024.

[MT05] Nadia Magnenat-Thalmann and Daniel Thalmann. "Virtual Humans: Thirty Years of Research, What Next?" *The Visual Computer* 21.12 (2005), pp. 997–1015.

[MTM+18] Simone Claire Mölbert, Anne Thaler, Betty J Mohler, Stephan Streuber, Javier Romero, Michael J Black, Stephan Zipfel, H-O Karnath, and Katrin Elisabeth Giel. "Assessing Body Image in Anorexia Nervosa Using Biometric Self-Avatars in Virtual Reality: Attitudinal Components Rather Than Visual Body Size Estimation Are Distorted". *Psychological Medicine* 48.4 (2018), pp. 642–653.

[NBN+20] Solène Neyret, Anna I Bellido Rivas, Xavi Navarro, and Mel Slater. "Which Body Would You Like to Have? The Impact of Embodied Perspective on Body Perception and Body Evaluation in Immersive Virtual Reality". *Frontiers in Robotics and AI* 7 (2020).

[NGS+11] Jean-Marie Normand, Elias Giannopoulos, Bernhard Spanlang, and Mel Slater. "Multisensory Stimulation Can Induce an Illusion of Larger Belly Size in Immersive Virtual Reality". *PLOS ONE* 6.1 (2011).

[NHF+16] Bennett K. Ng, Benjamin J. Hinton, Bo Fan, Alka M. Kanaya, and John A. Shepherd. "Clinical Anthropometrics and Body Composition from 3D Whole-Body Surface Scans". *European Journal of Clinical Nutrition* 70.11 (2016), pp. 1265–1270.

[NLL17] Diederick C. Niehorster, Li Li, and Markus Lappe. "The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research". *i-Perception* 8.3 (2017).

[NMJ+20] Michael G. Nelson, Angshuman Mazumdar, Saad Jamal, Yingjie Chen, and Christos Mousas. "Walking in a Crowd Full of Virtual Characters: Effects of Virtual Character Appearance on Human Movement Behavior". In *International Symposium on Visual Computing*. 2020, pp. 617–629.

[NVW+13] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. "Sparse Localized Deformation Components". *ACM Transactions on Graphics* 32.6 (2013), 179:1–179:10.

[NZC+18] Chontira Nimcharoen, Stefanie Zollmann, Jonny Collins, and Holger Regenbrecht. "Is That Me? – Embodiment and Body Perception with an Augmented Reality Mirror". In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 2018, pp. 158–163.

[OBB20] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. "STAR: A Sparse Trained Articulated Human Body Regressor". In *Computer Vision – ECCV*. 2020, pp. 598–613.

[OLP+18] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. "Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation". In *Proc. of the International Conference on 3D Vision (3DV)*. 2018, pp. 484–494.

[PCC+00] José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martínez, and Joaquín Fernández-Valdivia. "Diatom Autofocusing in Brightfield Microscopy: A Comparative Study". In *Proc. of International Conference on Pattern Recognition*. 2000, pp. 3318–3321.

[PCG+19] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10975–10985.

[PE18] Catherine Preston and H. Henrik Ehrsson. "Implicit and Explicit Changes in Body Satisfaction Evoked by Body Size Illusions: Implications for Eating Disorder Vulnerability in Women". *PLOS ONE* 13.6 (2018).

[PG20] Anna Samira Praetorius and Daniel Görlich. "How Avatars Influence User Behavior: A Review on the Proteus Effect in Virtual Environments and Video Games". In *International Conference on the Foundations of Digital Games*. 2020, pp. 1–9.

[PGB03] Patrick Pérez, Michel Gangnet, and Andrew Blake. "Poisson Image Editing". *ACM Transactions on Graphics* 22.3 (2003), pp. 313–318.

[PKA16] Jorge Peña, Subuhi Khan, and Cassandra Alexopoulos. "I Am What I See: How Avatar and Opponent Agent Body Size Affects Physical Activity Among Men Playing Exergames". *Journal of Computer-Mediated Communication* 21.3 (2016), pp. 195–209.

[PM11] Christine M. Peat and Jennifer J. Muehlenkamp. "Self-Objectification, Disordered Eating, and Depression: A Test of Mediational Pathways". *Psychology of Women Quarterly* 35.3 (2011), pp. 441–450.

[PRM+15] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. "Dyna: A Model of Dynamic Human Shape in Motion". *ACM Transactions on Graphics* 34.4 (2015).

[PSA+13] Tabitha C. Peck, Sofia Seinfeld, Salvatore M. Aglioti, and Mel Slater. "Putting Yourself in the Skin of a Black Avatar Reduces Implicit Racial Bias". *Consciousness and Cognition* 22.3 (2013), pp. 779–787.

[PSR+14] Ivelina V. Piryankova, Jeanine K. Stefanucci, Javier Romero, Stephan De La Rosa, Michael J. Black, and Betty J. Mohler. "Can I Recognize My Body's Weight? The Influence of Shape and Texture on the Perception of Self". *ACM Transactions on Applied Perception* 11.3 (2014), 13:1–13:18.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830.

[PWH+17] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. "Building Statistical Shape Spaces for 3D Human Modeling". *Pattern Recognition* 67 (2017), pp. 276–286.

[PWL+14] Ivelina V. Piryankova, Hong Yu Wong, Sally A. Linkenauger, Catherine Stinson, Matthew R. Longo, Heinrich H. Bülthoff, and Betty J. Mohler. "Owning an Overweight or Underweight Body: Distinguishing the Physical, Experienced and Virtual Body". *PLOS ONE* 9.8 (2014).

[QKS+24] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. "GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 20299–20309.

[RBD+02] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, and Scott Fleming. *Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report. Volume 1: Summary*. Tech. rep. Sytronics Inc, 2002.

[RBL+20] Rabindra Ratan, David Beyea, Benjamin J Li, and Luis Graciano. "Avatar Characteristics Induce Users' Behavioral Conformity with Small-to-Medium Effect Sizes: A Meta-Analysis of the Proteus Effect". *Media Psychology* 23.5 (2020), pp. 651–675.

[RBS+18] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. "Generating 3D Faces Using Convolutional Mesh Autoencoders". In *Computer Vision – ECCV*. 2018, pp. 725–741.

[Rea24] Capturing Reality. https://www.capturingreality.com. 2024.

[RGB+20] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. "Single-Shot High-Quality Facial Geometry and Skin Appearance Capture". *ACM Transactions on Graphics* 39.4 (2020), 81:1–81:12.

[RGD+19] Giuseppe Riva, José Gutiérrez-Maldonado, Antonios Dakanalis, and Marta Ferrer-García. "Virtual Reality in the Assessment and Treatment of Weight-Related Disorders". In *Virtual Reality for Psychological and Neurocognitive Interventions*. Springer, New York, NY, 2019, pp. 163–193.

[Riv97] Giuseppe Riva. "The Virtual Environment for Body-Image Modification (VEBIM): Development and Preliminary Evaluation". *Presence: Teleoperators and Virtual Environments* 6.1 (1997), pp. 106–117.

[RL20] Daniel Roth and Marc Erich Latoschik. "Construction of the Virtual Embodiment Questionnaire (VEQ)". *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 26.12 (2020), pp. 3546–3556.

[RLE17] Alwin de Rooij, Sarah van der Land, and Shelly van Erp. "The Creative Proteus Effect: How Self-Similarity, Embodiment, and Priming of Creative Stereotypes with Avatars Influences Creative Ideation". In *Proc. of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 2017, pp. 232–236.

[RLL+17] Daniel Roth, Jean-Luc Lugrin, Marc Erich Latoschik, and Stephan Huber. "Alpha IVBO – Construction of a Scale to Measure the Illusion of Virtual Body Ownership". In *Proc. of CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2017, pp. 2875–2883.

[ROC+20] Cristian Romero, Miguel A Otaduy, Dan Casas, and Jesus Perez. "Modeling and Estimation of Nonlinear Skin Mechanics for Animated Avatars". *Computer Graphics Forum* 39.2 (2020), pp. 77–88.

[Roo24] Rootmotion. *FinalIK*. http://root-motion.com. 2024.

[Ros01] James C. Rosen. "Improving Body Image in Obesity". In *Body Image, Eating Disorders, and Obesity: An Integrative Guide for Assessment and Treatment*. American Psychological Association, 2001, pp. 425–440.

[RTB17] Javier Romero, Dimitrios Tzionas, and Michael J. Black. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". *ACM Transactions on Graphics* 36.6 (2017), 245:1–245:17.

[SA07] Olga Sorkine and Marc Alexa. "As-Rigid-As-Possible Surface Modeling". In *Proc. of the Eurographics Symposium on Geometry Processing (SGP)*. 2007, pp. 109–116.

[SAK+24] Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. "DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1439–1449.

[SBE+24] Karthik Shetty, Annette Birkhold, Bernhard Egger, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, and Andreas Maier. "HOOREX: Higher Order Optimizers for 3D Recovery from X-Ray Images". In *Machine Learning for Multimodal Healthcare Data*. 2024, pp. 115–124.

[SBJ+23] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Bernhard Egger, Markus Kowarschik, and Andreas Maier. "BOSS: Bones, organs and skin shape model". *Computers in Biology and Medicine* 165 (2023).

[SBS21] Norbert Stefan, Andreas L. Birkenfeld, and Matthias B. Schulze. "Global Pandemics Interconnected – Obesity, Impaired Metabolic Health and COVID-19". *Nature Reviews Endocrinology* 17.3 (2021), pp. 135–149.

[SBW17] Richard Skarbez, Frederick P. Brooks Jr., and Mary C. Whitton. "A Survey of Presence and Related Concepts". *ACM Computing Surveys (CSUR)* 50.6 (2017), 96:1–96:39.

[SCK17] Daniel M. Shafer, Corey P. Carbonara, and Michael F. Korpi. "Modern Virtual Reality Technology: Cybersickness, Sense of Presence, and Gender". *Media Psychology Review* 11.2 (2017).

[SD92] Ken Shoemake and Tom Duff. "Matrix Animation and Polar Decomposition". In *Proc. of the Conference on Graphics Interface*. 1992, pp. 258–264.

[SF16] Johannes Lutz Schönberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4104–4113.

[SGH+20] Mel Slater, Cristina Gonzalez-Liencres, Patrick Haggard, Charlotte Vinkers, Rebecca Gregory-Clarke, Steve Jelley, Zillah Watson, Graham Breen, Raz Schwarz, William Steptoe, Dalila Szostak, Shivashankar Halan, Deborah Fox, and Jeremy Silver. "The Ethics of Realism in Virtual and Augmented Reality". *Frontiers in Virtual Reality* 1 (2020).

[SGY+24] Soubhik Sanyal, Partha Ghosh, Jinlong Yang, Michael J. Black, Justus Thies, and Timo Bolkart. "SCULPT: Shape-Conditioned Unpaired Learning of Pose-dependent Clothed and Textured Human Meshes". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[Sir56] William E. Siri. *Body Composition from Fluid Spaces and Density: Analysis of Methods*. Tech. rep. UCRL-3349. Lawrence Berkeley National Laboratory, 1956, pp. 1–33.

[Sla09] Mel Slater. "Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments". *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1535 (2009), pp. 3549–3557.

[SLD+19] Mike Seymour, Yuan Lingyao, Allan Dennis, and Kai Riemer. "Crossing the Uncanny Valley? Understanding Affinity, Trustworthiness, and Preference for More Realistic Virtual Humans in Immersive Environments". In *Proc. of the 52nd Hawaii International Conference on System Sciences*. 2019, pp. 1748–1758.

[SMB13] Daniel Sieger, Stefan Menzel, and Mario Botsch. "High Quality Mesh Morphing Using Triharmonic Radial Basis Functions". In *Proc. of the 21st International Meshing Roundtable*. 2013, pp. 1–15.

[SMR+14] Matthias Schröder, Jonathan Maycock, Helge Ritter, and Mario Botsch. "Real-Time Hand Tracking Using Synergistic Inverse Kinematics". In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 5447–5454.

[SNM+21] Robert Schleicher, Marlies Nitschke, Jana Martschinke, Marc Stamminger, Björn Eskofier, Jochen Klucken, and Anne Koelewijn. "BASH: Biomechanical Animated Skinned Human for Visualization of Kinematics and Muscle Activity". In *Proc. of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 2021, pp. 25–36.

[Sor05] Olga Sorkine. "Laplacian Mesh Processing". In *Eurographics 2005 - State of the Art Reports*. 2005, pp. 53–70.

[SP04] Robert W. Sumner and Jovan Popović. "Deformation Transfer for Triangle Meshes". *ACM Transactions on Graphics* 23.3 (2004), pp. 399–405.

[SPR+16] Paola Salomoni, Catia Prandi, Marco Roccetti, Lorenzo Casanova, and Luca Marchetti. "Assessing the Efficacy of a Diegetic Game Interface with Oculus Rift". In *Proc. of the 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. 2016, pp. 387–392.

[SRK17] Mike Seymour, Kai Riemer, and Judy Kay. "Interactive Realistic Digital Avatars-Revisiting the Uncanny Valley". In *Proc. of the 50th Hawaii International Conference on System Sciences*. 2017, pp. 547–556.

[SS15] Anthony Steed and Ralph Schroeder. "Collaboration in Immersive and Non-Immersive Virtual Environments". In *Immersed in Media*. Springer, 2015, pp. 263–282.

[SSS+10] Mel Slater, Bernhard Spanlang, Maria V. Sanchez-Vives, and Olaf Blanke. "First Person Experience of Body Transfer in Virtual Reality". *PLOS ONE* 5.5 (2010).

[SSS+24] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. "Relightable Gaussian Codec Avatars". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 130–141.

[SWL+24] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. "SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1606–1616.

[SZK15] Shunsuke Saito, Zi-Ye Zhou, and Ladislav Kavan. "Computational Bodybuilding: Anatomically-Based Modeling of Human Bodies". *ACM Transactions on Graphics* 34.4 (2015), 41:1–41:12.

[SZP+16] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. "Pixelwise View Selection for Unstructured Multi-View Stereo". In *Computer Vision – ECCV*. 2016, pp. 501–518.

[SZS+06] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. "A Comparative Study of Energy Minimization Methods for Markov Random Fields". In *Computer Vision – ECCV*. 2006, pp. 16–29.

[TAB+19a] Jennifer Todd, Jane E Aspell, David Barron, and Viren Swami. "An Exploration of The Associations Between Facets of Interoceptive Awareness and Body Image in Adolescents". *Body Image* 31 (2019), pp. 171–180.

[TAB+19b] Jennifer Todd, Jane E. Aspell, David Barron, and Viren Swami. "Multiple Dimensions of Interoceptive Awareness are Associated with Facets of Body Image in British Adults". *Body Image* 29 (2019), pp. 6–16.

[TB99] Michael E. Tipping and Christopher M. Bishop. "Probabilistic Principal Component Analysis". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3 (1999), pp. 611–622.

[TDM11] J. Rafael Tena, Fernando De la Torre, and Iain Matthews. "Interactive Region-Based Linear 3D Face Models". *ACM Transactions on Graphics* 30.4 (2011), 76:1–76:10.

[TEM+16] David J. Tomlinson, Robert M. Erskine, Christopher I. Morse, Keith Winwood, and Gladys Onambélé-Pearson. "The Impact of Obesity on Skeletal Muscle Strength and Structure Through Adolescence to Old Age". *Biogerontology* 17.3 (2016), pp. 467–483.

[TG09] Angela Tinwell and Mark Grimshaw. "Bridging the Uncanny: An Impossible Traverse?" In *Proc. of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*. 2009, pp. 66–73.

[TGK+21] Collin Turbyne, Abe Goedhart, Pelle de Koning, Frederike Schirmbeck, and Damiaan Denys. "Systematic Review and Meta-Analysis of Virtual Reality in Mental Healthcare: Effects of Full Body Illusions on Body Image Disturbance". *Frontiers in Virtual Reality* 2 (2021).

[TGM+18] Anne Thaler, Michael N. Geuss, Simone C. Mölbert, Katrin E. Giel, Stephan Streuber, Javier Romero, Michael J. Black, and Betty J. Mohler. "Body Size Estimation of Self and Others in Females Varying in BMI". *PLOS ONE* 13.2 (2018).

[Tha19] Anne Thaler. "The Role of Visual Cues in Body Size Estimation". In *MPI Series in Biological Cybernetics*. 56. Logos Verlag Berlin GmbH, 2019.

[TMP+24] Kartik Teotia, B. R. Mallikarjun, Xingang Pan, Hyeongwoo Kim, Pablo Garrido, Mohamed Elgharib, and Christian Theobalt. "HQ-3DAvatar: High-quality Implicit 3D Head Avatar". *ACM Transactions on Graphics* 43.3 (2024), 27:1–27:24.

[TSY+21] Xiangjun Tang, WenXin Sun, Yong-Liang Yang, and Xiaogang Jin. "Parametric Reshaping of Portraits in Videos". In *Proc. of the 29th ACM International Conference on Multimedia*. 2021, pp. 4689–4697.

[TT98] J. Kevin Thompson and Stacey Tantleff-Dunn. "MINI-REVIEW Assessment of Body Image Disturbance in Obesity". *Obesity Research* 6.5 (1998), pp. 375–377.

[Uni19] Unity Technologies. *Unity*. https://unity3d.com. 2019.

[Uni24] Unity Technologies. *Asset Store*. https://assetstore.unity.com/packages/tools/animation/realistic-eye-movements-29168. 2024.

[Val24a] Valve. *SteamVR*. https://store.steampowered.com/steamvr. 2024.

[Val24b] Valve Corporation. *Index*. https://store.steampowered.com/valveindex. 2024.

[Val98] Giovanni G. Valtolina. "Body-Size Estimation by Obese Subjects". *Perceptual and Motor Skills* 86.3 (1998), pp. 1363–1374.

[ViT24] ViTraS. https://hci.uni-wuerzburg.de/projects/vitras/. 2024.

[VM20] Ollin Venegas and Raman Mehrzad. "Prevalence and Trends in Obesity in the United States and Affluent Countries". In *Obesity*. Elsevier, 2020. Chap. 3, pp. 19–41.

[WAB+20] Stephan Wenninger, Jascha Achenbach, Andrea Bartl, Marc Erich Latoschik, and Mario Botsch. "Realistic Virtual Humans from Smartphone Videos". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2020, 29:1–29:11.

[WBH+21] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. "Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone". In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 3681–3691.

[WBR+23] Keenon Werling, Nicholas A. Bianco, Michael Raitor, Jon Stingel, Jennifer L. Hicks, Steven H. Collins, Scott L. Delp, and C. Karen Liu. "AddBiomechanics: Automating Model Scaling, Inverse Kinematics, and Inverse Dynamics from Human Motion Data Through Sequential Optimization". *PLOS ONE* 18.11 (2023).

[WCS+22] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. "HumanNeRF: Free-Viewpoint Rendering of Moving People from Monocular Video". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16210–16220.

[WDH21] Carolin Wienrich, Nina Döllinger, and Rebecca Hein. "Behavioral Framework of Immersive Technologies (BehaveFIT): How and Why Virtual Reality can Support Behavioral Change Processes". *Frontiers in Virtual Reality* 2 (2021).

[WDM+20] Erik Wolf, Nina Döllinger, David Mal, Carolin Wienrich, Mario Botsch, and Marc Erich Latoschik. "Body Weight Perception of Females Using Photorealistic Avatars in Virtual and Augmented Reality". In *Proc. of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2020, pp. 462–473.

[WDM+22] Erik Wolf, Nina Döllinger, David Mal, Stephan Wenninger, Andrea Bartl, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Does Distance Matter? Embodiment and Perception of Personalized Avatars in Relation to the Self-Observation Distance in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022).

[WFD+22] Erik Wolf, Marie Luisa Fiedler, Nina Döllinger, Carolin Wienrich, and Marc Erich Latoschik. "Exploring Presence, Avatar Embodiment, and Body Perception with a Holographic Augmented Reality Mirror". In *Proc. of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2022, pp. 350–359.

[WGR+18] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. "The Impact of Avatar Personalization and Immersion on Virtual Body Ownership, Presence, and Emotional Response". *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 24.4 (2018), pp. 1643–1652.

[WHO00] World Health Organization. *Obesity: Preventing and Managing the Global Epidemic: Report of a WHO Consultation*. 2000.

[WHO19] World Health Organization. *International Classification of Diseases, Eleventh Revision (ICD-11)*. 2019.

[WHO21] World Health Organization. *Obesity and Overweight*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. 2021.

[WKS+24] Stephan Wenninger, Fabian Kemper, Ulrich Schwanecke, and Mario Botsch. "TailorMe: Self-Supervised Learning of an Anatomically Constrained Volumetric Human Shape Model". *Computer Graphics Forum* 43.2 (2024).

[WLR15] Shensheng Wang, Scott O. Lilienfeld, and Philippe Rochat. "The Uncanny Valley: Existence and Explanations". *Review of General Psychology* 19.4 (2015), pp. 393–407.

[WMD+21] Erik Wolf, Nathalie Merdan, Nina Döllinger, David Mal, Carolin Wienrich, Mario Botsch, and Marc Erich Latoschik. "The Embodiment of Photorealistic Avatars Influences Female Body Weight Perception in Virtual Reality". In *Proc. of the IEEE Conference on Virtual Reality (VR)*. 2021, pp. 65–74.

[WMF+22] Erik Wolf, David Mal, Viktor Frohnapfel, Nina Döllinger, Stephan Wenninger, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. "Plausibility and Perception of Personalized Virtual Humans between Virtual and Augmented Reality". In *Proc. of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2022, pp. 489–498.

[WNT+21] Michael C. Wong, Bennett K. Ng, Isaac Tian, Sima Sobhiyeh, Ian Pagano, Marcelline Dechenaud, Samantha F. Kennedy, Yong E. Liu, Nisa N. Kelly, Dominic Chow, Andrea K. Garber, Gertraud Maskarinec, Sergi Pujades, Michael J. Black, Brian Curless, Steven B. Heymsfield, and John A. Shepherd. "A Pose-Independent Method for Accurate and Precise Body Composition from 3D Optical Scans". *Obesity* 29.11 (2021), pp. 1835–1847.

[WRG16] Brenda K. Wiederhold, Giuseppe Riva, and José Gutiérrez-Maldonado. "Virtual Reality in the Assessment and Treatment of Weight-Related Disorders". *Cyberpsychology, Behavior, and Social Networking* 19.2 (2016), pp. 67–73.

[WS98] Bob G. Witmer and Michael J. Singer. "Measuring Presence in Virtual Environments: A Presence Questionnaire". *Presence: Teleoperators and Virtual Environments* 7.3 (1998), pp. 225–240.

[WSB24a] Nicolas Wagner, Ulrich Schwanecke, and Mario Botsch. "SparseSoftDECA – Efficient high-resolution physics-based facial animation from sparse landmarks". *Computers & Graphics* 119 (2024).

[WSB24b] Nicolas Wagner, Ulrich Schwanecke, and Mario Botsch. "AnaConDaR: Anatomically-Constrained Data-Adaptive Facial Retargeting". *Computers & Graphics* 122 (2024).

[WSH+16] Thomas Waltemate, Irene Senna, Felix Hülsmann, Marieke Rohde, Stefan Kopp, Marc Ernst, and Mario Botsch. "The Impact of Latency on Perceptual Judgments and Motor Performance in Closed-Loop Interaction in Virtual Reality". In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*. 2016, pp. 27–35.

[WWS18] D. Catherine Walker, Emily K. White, and Vamshek J. Srinivasan. "A Meta-Analysis of The Relationships Between Body Checking, Body Image Avoidance, Body Image Dissatisfaction, Mood, and Disordered Eating". *International Journal of Eating Disorders* 51.8 (2018), pp. 745–770.

[WZR+24] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander Schwing, and Shenlong Wang. "GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 2059–2069.

[XCL+24] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. "Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[XGG+24] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. "FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian Embedding". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1802–1812.

[XTW+20] Qinjie Xiao, Xiangjun Tang, You Wu, Leyang Jin, Yong-Liang Yang, and Xiaogang Jin. "Deep Shapely Portraits". In *Proc. of the 28th ACM International Conference on Multimedia*. 2020, pp. 1800–1808.

[YB06] Nick Yee and Jeremy N. Bailenson. "Walk a Mile in Digital Shoes: The Impact of Embodied Perspective-Taking on the Reduction of Negative Stereotyping in Immersive Virtual Environments". *Presence: Teleoperators and Virtual Environments* 24 (2006), pp. 147–156.

[YB07] Nick Yee and Jeremy N. Bailenson. "The Proteus Effect: The Effect of Transformed Self-Representation on Behavior". *Human Communication Research* 33.3 (2007), pp. 271–290.

[YZC+24] Lingchen Yang, Gaspard Zoss, Prashanth Chandran, Markus Gross, Barbara Solenthaler, Eftychios Sifakis, and Derek Bradley. "Learning a Generalized Physical Face Model From Data". *ACM Transactions on Graphics* 43.4 (2024), 94:1–94:14.

[ZB15] Silvia Zuffi and Michael J. Black. "The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3537–3546.

[ZBT23] Wojciech Zielonka, Timo Bolkart, and Justus Thies. "Instant Volumetric Head Avatars". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 4574–4584.

[ZFL+10] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. "Parametric Reshaping of Human Bodies in Images". *ACM Transactions on Graphics* 29.4 (2010), 126:1–126:10.

[ZHK15] Lifeng Zhu, Xiaoyan Hu, and Ladislav Kavan. "Adaptable Anatomical Models for Realistic Bone Motion Reconstruction". *Computer Graphics Forum* 34.2 (2015), pp. 459–471.

[ZHY+22] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. "Structured Local Radiance Fields for Human Avatar Modeling". In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 15893–15903.

[ZJH+18] Haiming Zhao, Xiaogang Jin, Xiaojian Huang, Menglei Chai, and Kun Zhou. "Parametric Reshaping of Portrait Images for Weight-Change". *IEEE Computer Graphics and Applications* 38.1 (2018), pp. 77–90.

[ZJM+21] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In *Proc. of the 38th International Conference on Machine Learning*. Vol. 139. 2021, pp. 12310–12320.

[ZKM18] Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. "The Effect of Realistic Appearance of Virtual Characters in Immersive Environments – Does the Character's Personality Play a Role?" *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 24.4 (2018), pp. 1681–1690.

[ZM02] Taner Ziylan and Khalil Awadh Murshid. "An Analysis of Anatolian Human Femur Anthropometry". *Turkish Journal of Medical Sciences* 32.3 (2002), pp. 231–235.

[ZMM19] Katja Zibrek, Seán Martin, and Rachel McDonnell. "Is Photorealism Important for Perception of Expressive Virtual Humans in Virtual Reality?" *ACM Transactions on Applied Perception* 16 (2019), pp. 1–19.

[ZMS+18] Katrin Ziser, Simone Claire Mölbert, Felicitas Stuber, Katrin Elisabeth Giel, Stephan Zipfel, and Florian Junne. "Effectiveness of Body Image Directed Interventions in Patients with Anorexia Nervosa: A Systematic Review". *International Journal of Eating Disorders* 51.10 (2018), pp. 1121–1127.

[ZSS+13] Tatiana Zanetti, Paolo Santonastaso, Eleonora Sgaravatti, Daniela Degortes, and Angela Favaro. "Clinical and Temperamental Correlates of Body Image Disturbance in Eating Disorders". *European Eating Disorders Review* 21.1 (2013), pp. 32–37.

[ZTG+18] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. "State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications". *Computer Graphics Forum* 37.2 (2018), pp. 523–550.

[Zyg24] Zygote. https://www.zygote.com. 2024.

[ZZZ+23] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. "AvatarReX: Real-Time Expressive Full-Body Avatars". *ACM Transactions on Graphics* 42.4 (2023), 158:1–158:19.