Realistic Digital Human Characters: Challenges, Models and Training Algorithms

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Ahmed Osman

aus Kairo, Ägypten

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	03.09.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Michael J. Black
2. Berichterstatter/-in:	Prof. Dr. Gerard Pons-Moll
3. Berichterstatter/-in:	Prof. Dr. Christian Theobalt

Acknowledgments

At the outset of this thesis, I am deeply grateful to the countless people whose guidance, assistance, and unwavering support have made this scholarly pursuit possible. As much as this thesis reflects my own hard work, it also embodies the contributions of a vast network of family members, friends, mentors, colleagues, and institutions who have shaped me as a person and scholar. I owe each of them a tremendous debt of gratitude, and I am humbled to express my appreciation in the following pages.

I would like to express my heartfelt gratitude to my PhD advisor, Michael J. Black, whose guidance and mentorship have been invaluable throughout my doctoral journey. His expertise, insights, and constructive feedback have made me a better researcher and thinker. His unwavering support, encouragement, and patience have sustained me through the highs and lows of graduate school. His mentorship extends far beyond academic matters, as he has also provided me with crucial advice on career choices, work-life balance, and personal development. I feel incredibly fortunate to have had the opportunity to work with Michael, and I will always cherish the lessons and memories that he has imparted to me.

I would like to extend my heartfelt thanks to my fellow colleagues and friends who have made this journey so much more rewarding and enjoyable. Their companionship, encouragement, and camaraderie have been a constant source of inspiration and sup-

port. Whether it was discussing research ideas, sharing setbacks and triumphs, or simply commiserating over the challenges of graduate school life, my friends have been a crucial part of my academic and personal growth. They have also helped me maintain a healthy work-life balance by organizing social events, outings, and celebrations that provided much-needed breaks from the demands of research. In particular I would like to thank the following: Peter Gehler, Sergey Prokudin, Thomas Nestmeyer, Joachim Tesch, Javier Romero, Assem Behl, Betty Mohler, Yan Zhang, Siyu Tang, Lee Millward, Yassine Nemmour, Rocko (Melanie's late dog), Arijit Mallick, Yves Bernaerts, Victoria Fernandez Abrevaya, Tsvetelina Alexiadis, Partha Ghosh, Mohammed Hassan, Nikos Athanasiou, Omri Ben-Dov, Elia Bonetto, Vassilis Choutas, Radek Daněček, Markos Diomataris, Haiwen Feng, Yao Feng, Maria Paola Forte, Yinghao Huang, Marilyn Keller, Muhammed Kocabas, Qianli Ma, Lea Müller, Eric Price, Amir Ahmad, Nadine Rueegg, Nitin Saini, Soubhik Sanyal, Omid Taheri, Shashank Tripathi, Yuliang Xiu, Hongwei Yi, Yufeng Zheng, Joel Janai and Michael Niemeyer, Melanie Feldhofer, Nicole Overbaugh, Senya Polikovsky and Naureen Mahmoud. I feel incredibly fortunate to have been a part of such a vibrant and supportive community, and I will always cherish the memories and bonds that we have forged together.

I would like to express my deep gratitude to my PhD thesis advisory board, consisting of Professor Gerard Pons-Moll and Professor Andreas Geiger, for their invaluable mentorship, feedback, and support. Their insightful comments and critiques helped refine my research questions, methods, and findings. They have also challenged me to think more deeply and critically about the broader implications of my work for the field. Their willingness to provide feedback and guidance at every research stage has been instrumental in shaping my doctoral journey. Beyond their academic expertise, they have also provided me with valuable career advice and networking opportunities that have helped

me navigate the post-PhD landscape. I feel privileged to have had such accomplished scholars and mentors on my advisory board, and I am deeply grateful for their time, energy, and dedication to my success.

Last but certainly not least, I would like to express my deepest gratitude to my family for their unwavering love, support, and encouragement throughout my PhD journey. Their belief in me and my abilities, their patience and understanding during the long hours and weekends spent working, and their unflagging support during the ups and downs of this demanding academic pursuit have been a constant source of strength and inspiration. I feel incredibly fortunate to have such an amazing network of family members who have always been there for me, even when I was thousands of miles away pursuing my academic goals. Their sacrifices and unwavering support have been the bedrock of my academic and personal success, and I am deeply indebted to them for their love and generosity. This achievement would not have been possible without them, and I will always be grateful for their presence in my life.

Abstract

Statistical human body models have proven instrumental in various computer vision and computer graphics tasks. Despite the significant progress in statistical modeling of the human body and its parts, current state-of-the-art models still lack realism. Several unresolved challenges continue to impede their realism. The lack of realism stems from modeling assumptions and training algorithms which became a standard followed practice in constructing body models. Our goal in this thesis is to highlight the limitations of existing practices and propose models and training algorithms that overcome the limitations of existing methods.

The most widely used human body model is the SMPL body model. Despite its wide adoption, SMPL exhibits unrealistic deformations due to learning false long-range correlations from the training data. For instance, bending one elbow results in a bulge appearing in the other elbow. Artifacts of this type are not limited to SMPL but are prevalent in various other models, such as SMPL-X and GHUM. Additionally, Despite the extensive research on body part models for the head and hands, current body parts models, such as FLAME and MANO, can not capture the full range of motion of the head and hands relative to the body. Also, no realistic articulated model of the human foot has been developed despite its crucial role in human locomotion and footwear design. Finally, training current body models, such as SMPL, on small datasets is challenging due

to a large number of parameters, making them easily susceptible to overfitting. To avoid overfitting, an expert must gather an extensive dataset encompassing various subjects and carefully regularize the model during training to avoid overfitting. The necessity of expert-guided model training and the requirement for a substantial training dataset limits the scalability of training robust models by non-experts such as artists with a small collection of 3D scans for a single character. Despite the popular demand for a robust tool for data-efficient learning of articulated characters, such a tool does not exist to date.

The thesis results in two primary contributions, the first focusing on proposing models and the second on proposing training algorithms. We first propose a human body called STAR (Sparse Trained Articulated Human Body Regressor), in Chapter 3, where we introduce a model formulation that results in learning strictly sparse spatial deformations. As a result of the sparse formulation, STAR has significantly fewer parameters than SMPL, and the deformations are more realistic. Secondly, previous body models factor pose-dependent deformations independent of the body shape while, in reality, people with different shapes deform differently. Consequently, we learn shape-dependent pose corrective blendshape that depend on both body pose and BMI. We show that STAR generalizes better than SMPL when both are trained on the same training dataset, despite STAR having 80% fewer parameters. STAR is compatible with the gaming and animation industry standards and is a drop-in replacement for the widely used SMPL body model.

Our second contribution is identifying key limitations in current body parts models for the head and hands, which all surprisingly fail to accurately capture the full range of motion for these body parts relative to the rest of the body. Previous body part models have been trained using isolated 3D scans of individual parts, which do not capture the full range of motion for these body parts relative to the body. In contrast, full-body scans

provide valuable information about the motion of body parts relative to the body. Consequently, we propose a new learning scheme in Chapter 4 where we train an expressive sparse human body model, SUPR (Sparse Unified Part-Based Representation), on a federated training dataset of 1.2 million body, hand, foot, and head scans. As a consequence of the SUPR sparse formulation, we are able to separate the model into a full suite of high-fidelity body part models of the head, hands, and foot. Unlike previous body-part models, the separated body parts can model the body parts' full range of motion.

We further introduce the first articulated human foot model in Chapter 5. Previous attempts at creating such a model have faced challenges in accurately capturing the foot's complex deformations, which are influenced by many factors such as foot shape, pose, and ground contact. To overcome these challenges, we use a custom-built 4D foot scanner that captures dynamic sequences of the foot from all angles, including the foot sole, which is visible through a transparent glass platform. Previous approaches to modeling body deformations have only focused on either body pose or shape, which is inadequate for accurately modeling the deformation of the human foot during ground contact. We address this by introducing a non-linear deformation function that predicts foot deformations based on foot pose, shape, and ground contact. Our foot model is trained on 356 dynamic sequences from 30 subjects, where the capturing protocol explores the full range of motion for the toes, ankles, and foot deformations due to ground contact. Furthermore, we curate additional 7000 high-resolution scans from the ANSUR-II dataset to model foot shape variability. Through a thorough evaluation, we demonstrate the efficacy of our foot model in capturing the full range of motion for the foot, including deformations resulting from ground contact.

Finally, we propose AVATAR (Articulated Virtual Humans Trained By Bayesian Inference From a Single Ccan) in Chapter 6, a novel data-efficient training algorithm, which

can learn subject-specific body models from a single scan. AVATAR is robust to overfitting by posing training as a Bayesian inference problem, where we can incorporate prior distributions and reason about an entire distribution of plausible model parameters, instead of a single point estimate like existing methods. Through extensive evaluation, we show that AVATAR is robust to overfitting given a single training scan, and models trained by AVATAR are able to preserve subject specific deformations, achieve higher visual fidelity and generalization compared to SMPL. AVATAR streamlines character creation for all users, yielding engine-compatible personalized models, which was not possible before.

We make all the models in the thesis publicly available for research purposes. The thesis contributions have all been licensed by industrial vendors.

Contents

List of Figures

Chapter 1

Introduction

1.1 Thesis Statement

Embedding prior knowledge into statistical model's and training algorithms enhances model generalization and learning efficiency.

1.2 Introduction

The human body is defined by its three-dimensional form, which enables movement and interaction with the environment. Through facial expressions and gestures, we are able to communicate our thoughts, emotions, and intentions. Given the importance of the body in human-environment interactions, any artificial intelligence system aiming to simulate human behavior must have the ability to accurately perceive the three-dimensional structure of the body. This necessitates the use of high-fidelity 3D body models that capture variations in body pose, shape, and expression. These models are essential for industries such as gaming, animation, virtual try-on, augmented and virtual reality, and virtual

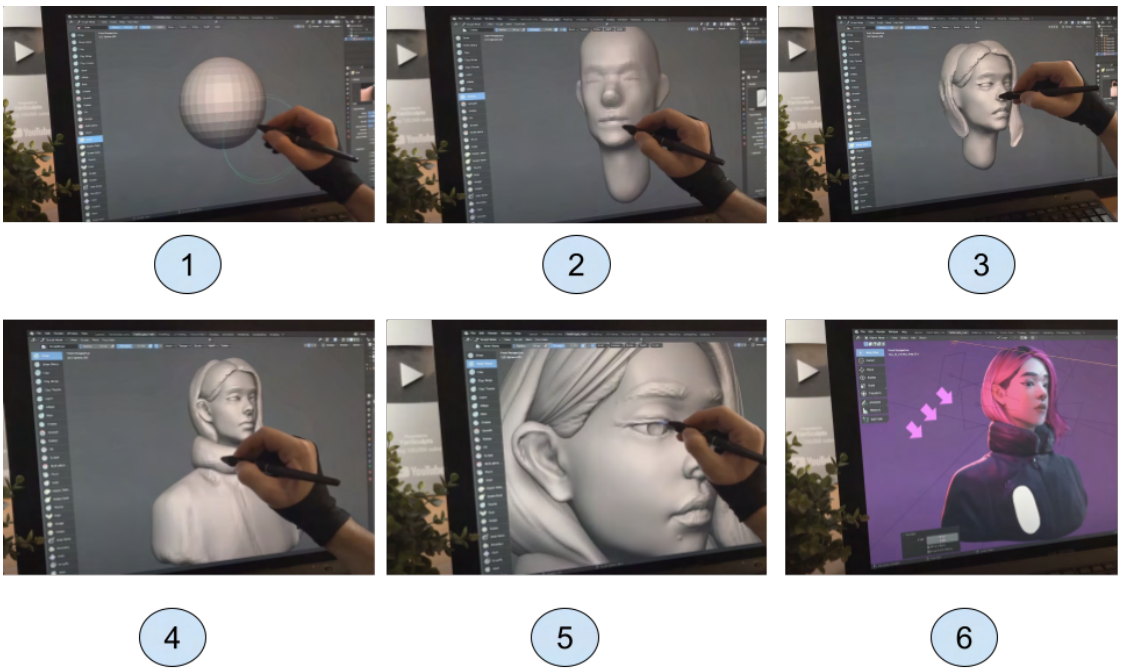


Figure 1.1: **Digital Sculpting:** In *Digital Sculpting*, a professional artist iteratively deforms sculpting a 3D object, such as a sphere, to create a human character. The characters created by an artist have a realistic appearance. In the above figure the artist YanSculpts (https://yansculpts.gumroad.com/) demonstrates digital sculpting for a female character.

telepresence. The central goal of this thesis is to create high-fidelity statistical body models on par with the fidelity produced by professional artists while eliminating the manual labor involved in their construction.

Thesis Goal: Create high-fidelity statistical models and training algorithms for the body and its parts that can be used by artists and animators.

1.2.1 Digital Sculpting

Digital Sculpting is the most commonly used technique by artists and game designers to create realistic digital humans. In *Digital Sculpting* a trained artist creates a 3D model of the body by iteratively sculpting a 3D object (such as a sphere) until reaching a target reference as highlighted in Fig. 1.1. Several professional software tools such as Blender and ZBrush offer artists the tools for digital sculpting. A character created by *Digital Sculpting* captures the coarse body geometry and high-frequency anatomical details; when textured and rendered, it looks realistic. Despite providing the artist with complete creative control over the graphic asset, *Digital Sculpting* is a labor-intensive and time-consuming process that requires specialized expertise. Therefore, it is not a viable solution for generating digital humans at scale. The significant inefficiency of digital sculpting in creating a large number of digital humans has prompted a diverse range of research efforts to find scalable alternatives.

1.2.2 Learned Body Models

Learning statistical models [7] from 3D scans emerged as a practical alternative to *Digital Sculpting*. A statistical body model is trained from a collection of 3D human scans to capture the distribution of body deformations. More formally, a statistical human body model is a parametric function defined by a pose space and a shape space. The pose space captures the position and orientation of the body bones, and the shape space captures variability in identity. Over the past two decades, the vision and graphics communities experienced a proliferation in the number of human body proposed [8, 9, 10, 11]. Existing human body models capture the body's 3D geometry, soft tissue deformations [12, 13], and expressive body models capture the human facial expressions and the hand gestures [14, 15, 16]. Statistical models enabled numerous applications

including reconstructing bodies from images and videos [17, 18, 19], modeling human interactions [20], generating 3D clothed humans [21, 22, 23, 24, 25, 26, 27], or generating humans in scenes [28, 29, 30].

1.3 Problem Statement

Despite the advancements in body modeling research, current models still exhibit drawbacks that limit their realism. To enhance the realism of human body models, several challenges must be addressed. In the following section, we thoroughly discuss the key challenges.

1.3.1 Game Engines Compatibility

The gaming industry is a multi-billion dollar industry [31]. Game engines provide the ecosystem for creating interactive 3D digital experiences. Thousands of developers use game engines such as Unity and Unreal to create interactive 3D worlds populated by 3D characters.

Game engines have numerous scientific applications that are being explored and utilized by researchers across various fields. For example in *Computer Vision*, game engines are used to create synthetic datasets for a wide range of computer vision tasks. These datasets are essential for training and evaluating algorithms such as 3D human pose estimation [32].

In the gaming and animation industry, there are standard conventions for representing articulated digital characters. These conventions are important for ensuring that different tools in the digital production pipeline are able to work together seamlessly. This allows artists to easily transfer their work between different software programs and maintain the

integrity of their digital assets. The industry-wide conventions for representing digital humans also serve to provide artists with full creative control over their graphics assets. By adhering to these standards, artists can make detailed and precise changes to the appearance of the digital characters.

Statistical human body models exists since the early 2000s, but they do not meet the standards of the game and animation industry. As a consequence of the limitations of previous statistical human body models, their utility was limited in the gaming and animation industry. The introduction of the SMPL body model [33] marked a significant milestone, enabling the utilization of statistical human body models within game engines. Since then, SMPL has become the most influential human body model in both academia and industry. Its compatibility with game engines has allowed it to be used across different scientific disciplines.

In recent years, various representations of the human body have been explored, including models utilizing deep architectures [34, 15]. Furthermore, numerous novel representations have been introduced such as the implicit representation [35] and the NeRF-based representation [36]. The latter representations enables unprecedented level of detail and realism which was not possible before. However, current approaches are not compatible with existing game engines. A key challenge we address in this thesis is proposing a novel suite of models for the body and its parts while complying with industry standards to ensure their practical use in real-world applications.

1.3.2 Model Versatility

One of the fundamental concepts in computer graphics is the ability to adapt the model to fit the computational limitation of an application. In a crowd simulation, when characters are far from the camera, there is no need for precise prediction of deformations

for all body joints, such as fingers and toes, because their small on-screen size makes it unnecessary.

The flexibility to adjust the model's computational footprint is a crucial aspect that is often neglected in current statistical models. By varying the fidelity of the model, users such as animators can control the computational footprint of the model. This is essential, as digital characters are often part of a larger software pipeline.

While versatility is a crucial aspect in designing graphic assets, existing statistical human body models lack this versatility. Models like SMPL [33] and SMPL-X [16] have a huge number of parameters. SMPL for example has more than 4 million parameters and over 200 blendshapes which are all necessary to predict how the model deforms with changes in body pose. The large number of parameters is a bottleneck for real-time applications, on devices with a limited computational budget (such as mobile phones).

Unlike current approaches, this thesis introduces a factorized representation of the human body that enables adaptable models, granting users the freedom to select the optimal number of blendshapes for their specific application.

1.3.3 Statistical Models Fidelity

In Section 1.2.1, we described that a significant portion of an artist's time is devoted to creating realistic deformations for a graphic asset. This process often involves using a reference image as a guide to ensure the fidelity of the deformations. The accuracy of the deformations is crucial for the visual fidelity of the model, as users have a strong preconceived notion of how the human body should deform. Any deviation from this expectation can compromise the visual fidelity of the model and make it appear unrealistic.

Despite years of research in statistical modeling of the human body, current models still exhibit artifacts, causing significant challenges for artists and animators. These ar-

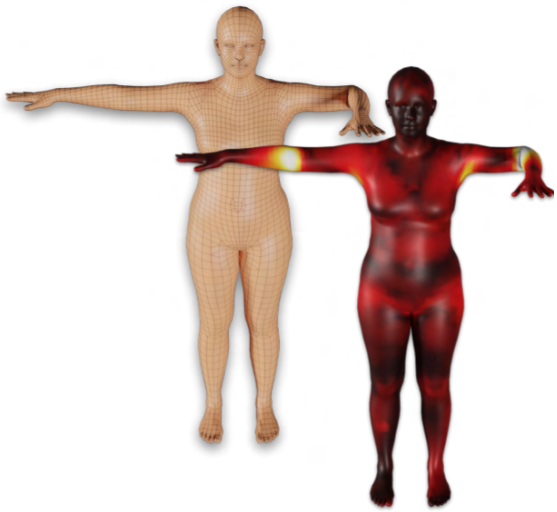


Figure 1.2: **False Long Range Deformations:** Bending the left elbow in SMPL results in a bulge in the other elbow.

tifacts are related to how existing models of the body and its parts deform with changes in pose and shape. In the following section, we will highlight the most common artifacts found in existing models.

Pose Deformations

All current models of the human body use a pose deformation function that predicts how the 3D surface of the body will vary based on changes in the position and orientation of the body joints. Although the specific formulation of the pose deformation function varies from model to model, all existing models suffer from learning false long-range spurious correlations. Models such as SMPL [33], SMPL-X [16], and GHUM [15] exhibit unrealistic deformations in response to changes in body pose due to these false correlations. For example, moving an elbow in SMPL and SMPL-X results in a bulge in the other elbow, as shown in Fig. 1.2. This is implausible, as human body deformations are typically sparse and spatially local, meaning that a joint movement should only affect

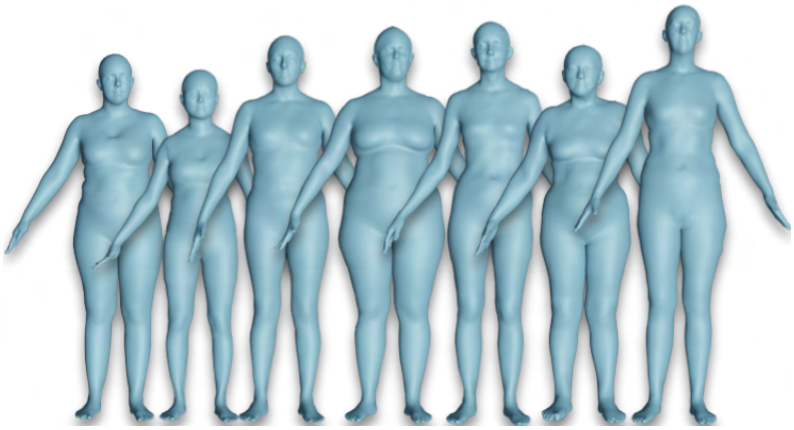


Figure 1.3: **Female Subjects From the CAESAR Dataset:** All female participants wore a sports bra, which biased the female chest's contour.

a small, local subset of the model vertices.

Shape Deformations

The CAESAR database [37] is a crucial resource for the development of statistical models aiming to capture the variability of human body shape. These models are used in a variety of applications, such as virtual clothing fitting and body measurement analysis. One limitation of the CAESAR database is the clothing worn by the female subjects. Specifically, all subjects wore a sports bra, which can have a significant impact on the shape of the chest, as shown in Fig. 1.3. As a result, existing models based on the CAESAR database may not accurately capture the shape of the female chest. The poor modeling of the female chest shape is problematic for applications such as virtual try-on.

The CAESAR database has limited body shape variability, as it does not accurately represent extreme body shapes of high BMI (body mass index) subjects. This limitation reduces the expressiveness of the models learned using the CAESAR datatset [37] for high BMI subjects. The shape space of popular human body models such as SMPL, SMPL-X, and GHUM are all trained on the CAESAR dataset alone, despite its known

limitations. This reliance on a single, limited dataset hinders the accuracy and reliability of these models in modeling a diverse range of body shapes. This limitation is particularly concerning for applications that require an accurate representation of high BMI individuals.

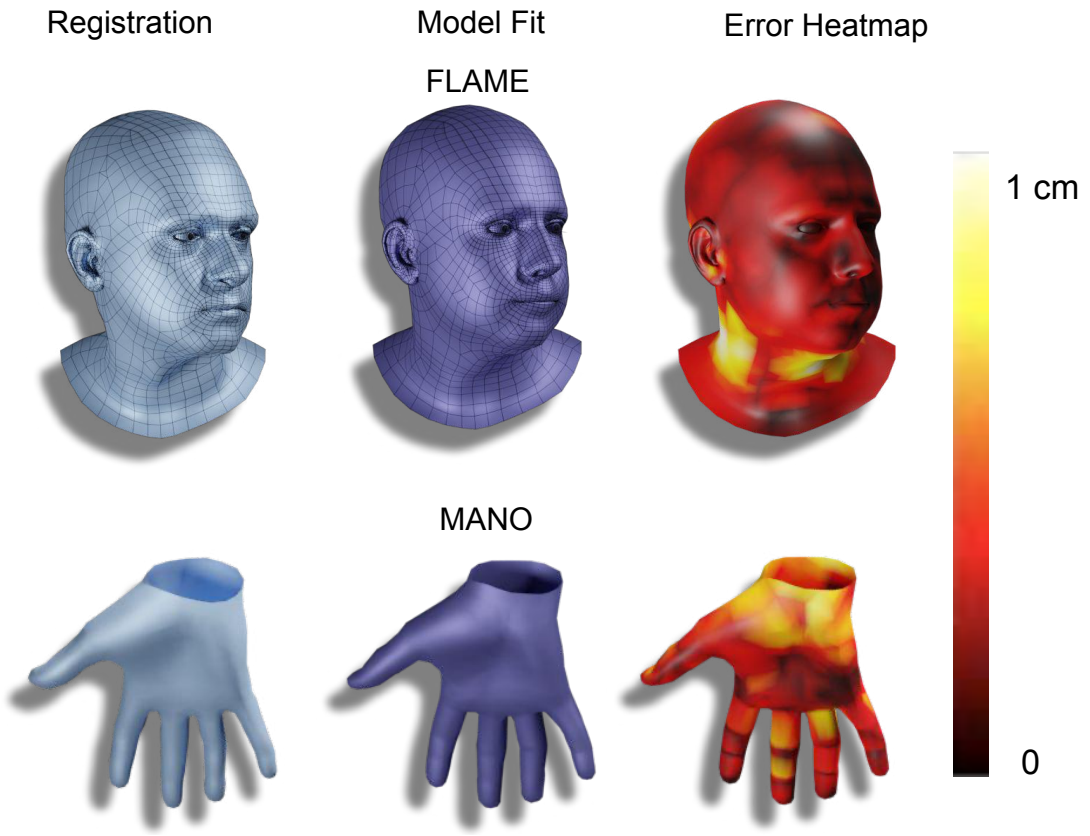


Figure 1.4: **Body Part Models Failure Cases:** Left: Existing body part models such as the FLAME [1] head model and the MANO [2] hand model fail to capture the corresponding body part's shape through the full range of motion. Fitting FLAME to a subject looking left results in significant error in the neck region. Similarly, fitting MANO to hands with a bent wrist, results in significant error at the wrist region.

Body Part Tracking

Current body part models, including those specifically designed for the head and hands, possess certain limitations when it comes to tracking the complete range of motion exhib-

ited by these body parts, as shown in Fig. 1.4. Despite incorporating a representation of the human head along with a neck, these models fail to accurately simulate and emulate the intricate movements and positional adjustments of the head and hands.

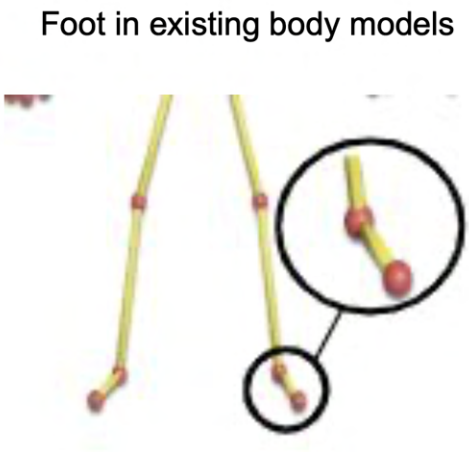


Figure 1.5: **SMPL Joints:** Existing kinematic tree in all body models have a limited number of joints in the foot. For example SMPL uses only two joints in the foot region. The limited number of joints is problematic which is insufficient to model toe articulation.

1.3.4 Foot Articulation

All existing human body models lack an articulated foot, as shown in Fig. 1.5, which is a crucial component in accurately modeling the human body. The SMPL body model, for example, only uses two joints for the foot, which is insufficient to model the full range of motion of the human foot including the ankle and the toes.

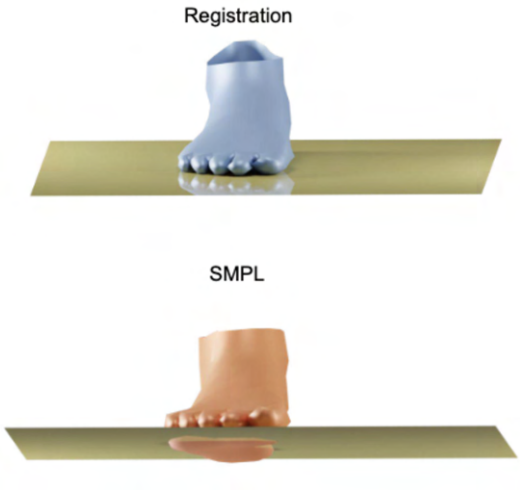


Figure 1.6: **SMPL Foot:** The foot of SMPL fails to model deformations due to ground contact, hence penetrating the ground.

Contact Based Deformation

In contrast to the majority of the body, the human foot undergoes soft tissue deformations. These deformations are influenced by various factors, including the position of the foot, its shape, and the nature of its contact with the surroundings. Existing modeling techniques commonly employed for the body focus on relating deformations to changes in body pose [33] or a combination of body pose and shape [38]. However, the impact of external contact on body deformation, particularly concerning the foot, has been largely overlooked. For example, SMPL will fail to preserve the foot contact deformation as shown in Fig. 1.6. This oversight is significant as accurately modeling the deformations of the human foot requires accounting for their interaction with the ground.

1.3.5 Model Training and Overfitting

Training current body models, such as SMPL, on limited datasets poses significant challenges due to the large number of model parameters, which make the models easily prone to overfitting during the model training. To mitigate these issues, experts are required to curate a comprehensive dataset comprising diverse subjects and employ careful regularization during the model training process. However, this reliance on expert guidance and the need for an extensive training dataset severely hampers the scalability of training reliable models for non-experts, such as artists who may possess only a small collection of 3D scans for a specific character.

1.4 Motivation

In this section, we highlight the various applications that are directly impacted by the thesis's contributions.

Health Care

Obesity is a major public health concern, as it is the biggest risk to human life expectancy. According to a study by Rossner et al. [39], obesity is strongly correlated with heart diseases. This is due to the excess fat accumulation in the body, which puts a strain on the cardiovascular system, leading to an increased risk of heart attacks and strokes.

In addition to heart diseases, obesity has also been associated with an increased risk of COVID-19 severity. A study by Kalligeros et al. [40] found that obese individuals are more likely to experience severe symptoms and complications from COVID-19 compared to those with healthy body weight. This is thought to be due to the underlying metabolic complications of obesity, such as inflammation and impaired immune func-

tion.

The body fat distribution is an important factor in the metabolic complications of obesity. According to Jensen et al. [41], the distribution of fat in the body, particularly in the abdominal region, is a strong predictor of metabolic health. Simple numerical measurements such as body weight or body mass index are not descriptive of the body fat distribution, and may not accurately capture an individual's risk of metabolic complications.

Human body models, which provide dense measurements of the body's surface, can be used to generate a multitude of measurements that better capture fat distribution and improve the prediction of obesity-related health risks [42].

1.4.1 Biomechanics

Biomechanics [43, 44, 45] research focuses on studying human locomotion [46, 47, 48]. This is an important area of study for the footwear industry and sport science. By understanding how the human body moves, researchers can develop footwear that supports and enhances human movement, reducing the risk of injury [49, 50]. Additionally, Biomechanics research is also crucial for the development of robotics. Many robotic systems are modeled after the human body, and a deep understanding of human locomotion is essential for designing robots that can move efficiently and effectively [51].

The human foot has evolved over millions of years to assist and enable bipedal human locomotion. This evolution has allowed us to walk, run, and jump efficiently, using the complex structure of the foot to support our weight and provide stability while moving. Biomechanics and sport science study the foot and its deformation due to ground contact, looking at how the foot adapts to different surfaces and the forces involved in locomotion. An articulated model of the human foot, with learned contact deformation, could provide

a valuable tool for these researchers, allowing them to run simulations and gain a deeper understanding of the foot's function and biomechanics.

1.4.2 Motion Synthesis

One major issue with current motion synthesis techniques is the foot skating problem [52], where the human foot appears to glide or slide across the ground, rather than realistically interacting with it. This results in implausible and unrealistic motion, which can be detrimental for animators and game designers who are trying to create realistic and believable animations.

All existing human body models do not accurately capture the complex interactions between the human foot and the ground. This lack of realism hinders the development of more convincing and natural-looking motions.

To address the foot skating issue, it is important to develop a precise human body model that faithfully represents the dynamics of the human foot and its interactions with the ground. Furthermore, the availability of a comprehensive dataset capturing human foot-ground interactions holds immense potential for training and enhancing the realism of motion synthesis algorithms.

1.4.3 AR & VR

Body and body part models are crucial for augmented reality (AR) and virtual reality (VR) [53]. These models are used to create a realistic representation of the human body, allowing users to better perceive their surroundings and interact with the virtual environments. Models that can capture the full range of motion for the hand and head are particularly important for accurate tracking of user gestures and head pose, which is critical for a plausible user experience.

1.4.4 Footwear Design

Footwear design is a crucial aspect of creating comfortable shoes that fit well and provide support for the foot. Accurate models of the human foot shape variability are essential in order to design shoes that cater to different foot shapes and sizes.

The human foot is a complex structure, with many different bones, muscles, and tendons working together to support the body and allow for movement. The shape and size of the foot can vary greatly between individuals, and it is important for shoe designers to take these differences into account when creating their designs. By using accurate models of the human foot shape variability, shoe designers can create shoes that fit and provide the necessary support for different foot shapes and sizes.

1.5 Contributions

Our main goal is to propose high-fidelity models of the human body that are similar to the model fidelity created by artists, while still being compatible with the game and animation industry standards. Next, we highlight the thesis's key contributions.

1.5.1 Diverse Dataset

Body Scans: A key challenge for creating realistic body models is that existing training datasets have limited pose and shape variability. We address this problem by curating more than $700K$ full body scans for subjects in a diversity of body poses. Unlike standard datasets, the dataset includes extreme body shapes such as bodybuilders and female anorexia nervosa women patients. The poses in the dataset are also diverse and contain poses by professional ballerinas and yoga poses. The dataset was curated to explore the full range of complexity of human body deformations across a wide range of body

shapes.

We address the limitations of the widely used CAESAR dataset for training the model shape space by curating further datasets that better capture the female chest and human body shape variability. To this end, we utilize the SizeUSA dataset [54] which contains an additional $10,000$ human bodies in a standard A-pose. Unlike the CAESAR dataset, the SizeUSA dataset female subjects wore a traditional bra, hence providing more shape variation. Additionally, SizeUSA has a much richer body shape variability. The combined CAESAR and SizeUSA dataset provides valuable information for training human body model shape space, which improves the modeling of the human body shape, particularly high BMI subjects which is crucial for applications as motivated in Section 1.4.

Head Scans: In full body scans, the resolution of the heads is typically low, rendering it inadequate to capture and reconstruct the nuanced subtleties of human facial expressions. To address this issue, we utilize a dedicated head scanner to obtain head scans with the necessary detail and clarity. The data capture protocol involves sequences to explore the human facial deformations due to facial expression space and jaw movement.

Hand Scans: In the context of full-body scanning, the hands are poorly reconstructed, corrupted by noise, and with occasionally missing fingers. This problem primarily stems from the substantial disparity between the full body scanner scanning volume and relative size of the human hands. Consequently, the data captured in a full-body scanner is insufficient for the purpose of learning and developing human hand models.

To mitigate this problem and enhance the quality of our hand data, we use a dedicated hand scanner. This specialized scanning setup allows for a more focused and detailed reconstruction of the hands. Our data-capturing protocol incorporates a series of sequences specifically designed to explore the full spectrum of motion exhibited by the

human fingers.

Foot Scans: Similar to the hand and head, we find that the reconstruction of the human foot in full body scanners is notably challenging. Foot scans are often noisy, incomplete, and with the individual toes fused. Additionally, the occlusion caused by the foot's contact with the ground prevents the reconstruction of the foot sole. This limitation also hinders our ability to accurately track movements of individual toes and capture the deformations the foot undergoes upon contact with the ground.

To overcome these obstacles and enable high-resolution capture of the foot, we use a custom-built foot scanning setup designed by 3dMD. This setup allows for comprehensive visibility of the foot from all angles, including the sole through a transparent glass platform.

Furthermore, to accurately model the distribution of human foot shape variability, we leverage over 7,000 high-resolution scans from the ANSUR-II dataset [55]. This dataset is invaluable for creating a high-fidelity shape space that encompasses the wide range of human foot shape variability.

The dataset compiled for this thesis is unparalleled in both its scale and diversity. Similarly, the variety of models and algorithms employed is without precedent, a feat achievable solely due to the extensive scope of data collection. To underscore the magnitude of the data. In Fig 1.7 we compare between the scale of data utilized in training the SUPR model, as introduced in Chapter 5, and the scale of the data used by existing state-of-the-art models.

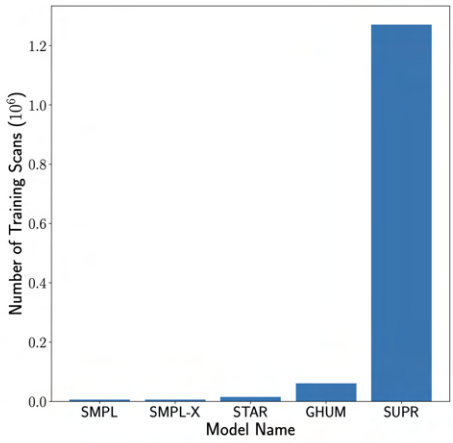


Figure 1.7: **Data Scale:** A comparison between the scale of training datasets for recent human body models. SUPR is trained on a order of magnitude more data compared to the highest number of training scans report in the literature (GHUM 60k).

1.5.2 Sparse Models

A key limitation shared by existing models is the widely used modeling formulation which relates all model joints to all model vertices and results in a range of implausible artifacts. To address this, in Chapter 3 we propose a novel sparse formulation of human body deformations. For each body joint, we learn the sparse set of mesh vertices influenced by that joint movement. A key component for the success of our method is that we learn the set of vertices influenced by the joint movement, instead of using artist predefined regions. As a consequence of the sparse formulation, the model deformation functions are strictly sparse and spatially local.

The sparse spatially local deformation offers a number of advantages that were not possible before with the current widely used fully connected formulation. The sparse deformation allows for a versatile model, where an artist can exclude blendshapes that are not relevant to their application. This allows the end-user to include or exclude model parameters and control their computational footprint. Finally, the sparse formulation

decreases the model parameters, compared to the existing formulation.

A second key limitation is the current body pose deformation formulation is independent of human body shape, which is clearly not realistic. In previous approaches both the shape and pose space were seen as independent. In contrast, we propose conditioning the pose-corrective blendshapes on the subject shape parameters that are correlated with the subject body mass index. Because of the conditioning of the pose space on the shape we are able to capture subject-dependent pose deformations that are due to body shape.

Existing models such as SMPL-X have a large number of pose parameters due to the large number of joints in the hand region, specifically because SMPL-X uses spherical joints for all body joints, which provide redundant degrees of freedom for body parts like the fingers. We address this issue in Chapter 4 by extending the kinematic tree to include a mixture of joint types, including hinge joints with a single axis of rotation for joints like the fingers and knees. Although using different types of joints in the kinematic tree is not a new concept, it is still highly desirable for animators and game designers. For example, to bend a finger, an artist would have to rotate the finger joints across 3 axes in the SMPL-X model, while in our formulation, it would only require a single rotation around a single axis.

1.5.3 Federated Training

Existing body part models, such as the human head and hand, are typically learned in isolation from the human body. This is done by using body part scans that are captured in a limited scanning volume focused on the specific body part. Due to the restricted scanning volume, accurately capturing the movement of a body part in relation to the entire body poses a significant challenge. Training body part models on body part scans only can be problematic, as it does not accurately represent the complex interactions and

movements of the human body.

To address this issue, in Chapter 4 we introduce a training algorithm for learning full-body and body part models jointly. Our algorithm trains a sparse body model on a federated training dataset and then separates the model into individual body parts. This allows us to accurately model the movements and interactions of the body and its parts, providing a more realistic representation of the human body.

1.5.4 Foot Model

The human foot is poorly modeled in existing human body models. Training a foot model is challenging because the human foot is hard to capture. In Chapter 5, we capture the human foot in a custom-built 4D scanner, where the human foot is visible from all views, including the foot sole, which is visible through a glass platform. The scanner features a runaway and is mechanically stable to capture motions such as walking and running, in addition to movements such as jumping. We capture a total of 356 dynamic sequences for 15 male and 15 female subjects perfoming a wide range of motions which explore the human foot's full range of motion including the movement of the ankle and toes, in addition to foot deformation due to ground contact. This is the first dataset that thoroughly captures and models the human foot in motion. To be able to model human foot shape, we train the foot model on $7,000$ scans from the ANSUR-II dataset [55], which features high-resolution scans. The scale and fidelity of the ANSUR-II dataset allows us to learn an expressive shape space of the human foot.

We propose a kinematic tree that contains 12 additional joints in the foot. This allows us to accurately model the pose of the human foot and ankle, as well as the movement of each individual toe. Our model is the first articulated model for the human foot. The inclusion of these additional joints in our kinematic tree allows for a more detailed and

precise representation of the human foot. By modeling individual toe movement, we can capture complex movements of the foot during activities such as walking and running.

Lastly, the existing model formulation only relates the body deformation to body pose, completely ignoring the deformation due to scene contact. The contact-based deformations are critical for application in particular for the human foot. We address this by introducing a novel non-linear formulation that predicts the foot deformations due to ground contact. Our network is conditioned on the foot pose, foot shape, and a ground contact descriptor. We train the network on 356 dynamic sequences of subjects with a diversity of human foot shapes.

1.5.5 Data Efficient Training

Learning from a limited number of scans poses a significant challenge when training articulated virtual humans. To address this issue, in Chapter 6 we propose AVATAR (Articulated Virtual Humans Trained By Bayesian Inference From a Single Scan), a novel algorithm that efficiently learns subject-specific body models.

Present training algorithms for body models featuring a pose-dependent deformations function, like SMPL, are trained by optimizing for a single-point estimate which best fits the training data. Due to the high dimensionality of the pose-dependent deformation function parameters, model training is easily prone to overfitting, especially when the available training dataset is small, particularly if only a single scan is available. We propose an alternative approach to character training. Our key insight is to pose learning the pose-corrective blendshape function parameters as a Bayesian inference problem. Within the Bayesian framework, we can reason about an entire distribution of possible parameters instead of a single-point estimate. Additionally, we can incorporate an informative prior of possible distribution of parameters. The combination of reasoning on an

entire distribution, in addition to the incorporation of a prior, makes AVATAR robust to overfitting even when only a single training scan is available.

1.6 Thesis Outline

The remaining of the thesis is divided into 7 chapters.

In Chapter 2 (Related Work): We review related work in human body modeling, specifically focusing on mesh-based models. Our analysis encompasses prior art in model training and representation.

In Chapter 3 (Sparse Models): We describe our proposed method, STAR, which introduces a state-of-the-art sparse spatially local factorization of the body pose deformation function.

In Chapter 4 (Federated Training): We describe a state-of-the-art federated training approach where we train an expressive human body model, SUPR , on a federated dataset and separate the model into a full suite of body parts models.

In Chapter 5 (Foot Model): We describe SUPR-Foot, an articulated foot model with learned contact deformation.

In Chapter 6 (Data-Efficient Learning): We describe AVATAR, the Bayesian framework for data-efficient learning of personalized human body models from a single scan.

In Chapter 7 (Discussion): We conclude with a discussion that summarizes the thesis contributions and discusses future directions.

Chapter 2

Related Work

Our goal in this thesis is to improve the accuracy of existing models for the body and its parts while maintaining compatibility with existing gaming and animation industry standards. The rest of this chapter is organized as follows: in section 2.1, we formally introduce the gaming and animation industry standards for representing articulated characters, such as the human body. In section 2.2, we conduct a thorough review of the related literature on mesh-based models for the human body. We evaluate prior art on the basis of 1) deformation fidelity, 2) model versatility, and 3) compatibility with existing gaming and animation pipelines. We then review prior art on existing training algorithms for body part models in section 2.3. In section 2.4 we review prior attempts in digitizing the human foot. We conclude with a review on learning algorithms for morphable models, with a particular emphasis on data-efficient training of human body models in section 2.5.

2.1 Gaming and Animation Industry Standards

Linear Blend Skinning (LBS), also known as Skeletal-Subspace Deformation (SSD) [56, 57], is a key formulation used in the gaming and animation industry for representing articulated characters. This technique remains popular to this day due to its simplicity and effectiveness. In LBS, a mesh is rigged with an underlying set of joints that form a kinematic tree. Each mesh vertex is associated with a number of body joints and corresponding skinning weights. The transformations applied to each mesh vertex are a weighted function of the transformations of the associated joints. These skinning weights can be defined by an artist or learned from data.

LBS is widely used in the gaming and animation industry because it is also relatively easy to implement and computationally fast. To date, LBS remains the foundation for many existing body models.

2.1.1 Pose-Corrective blendshapes.

The key widely known drawback of LBS is the loss of mesh volume around the joint regions when articulated. This loss of volume occurs in the areas around joints such as the elbows and knees. The loss of mesh volume around joint regions results in the widely known “ candy wrapper ” effect. This term refers to the appearance of the mesh model after the LBS algorithm has been applied, where the mesh appears to be stretched and thinned around the joint regions, similar to how a candy wrapper might appear after being stretched and twisted. The candy wrapper effect reduces the overall quality of the model. Several methods have been proposed to address the drawback of LBS. Lewis [58] introduces the pose space deformation model (PSD) where LBS is complemented with corrective deformations. The deformations are in the form of corrective offsets added

to the mesh vertices posed with LBS. The corrective deformations are related to the underlying kinematic tree pose. Weighted pose deformation (WPD) [59, 60] adds pose-corrective offsets to the base template mesh in the canonical (rest) pose before posing it with LBS, such that final posed mesh is plausible. Typically, such correctives are artist defined in key poses. Given a new pose, a weighted combination of correctives from nearby key poses is applied. Allen et al. [9] are the first to learn such corrective offsets from 3D scans of human bodies.

The combination of LBS complemented with pose-corrective blendshapes are the gaming and animation industry standard for representing articulated human body models. Despite the widely known drawbacks of LBS, nevertheless LBS remains the defacto representation of articulated characters in industry. The key reason LBS combined with blendshapes remains widely adopted is, LBS is a fully interpretable representation, and an artist can paint the weights with predictable outcome. As a consequence of the fully interpretable representation, LBS offers the artist the full creative control over the graphics asset through the full digital production cycle.

2.2 Statistical Models

There is a long literature on 3D modelling of the human body, constructed either manually or using data-driven methods. We review the most related literature here with a focus on methods that learn bodies from data, pioneered by [9, 10].

The release of the CAESAR dataset of 3D scans [37] enables researchers to train statistical models of body shape [8, 61]. SCAPE [10] is the first model to learn a factored representation of body shape and pose. SCAPE models body deformations due pose and shape as triangle deformations and has been extended in many ways [62, 63, 4, 64, 65,

66]. SCAPE has several downsides, however. It requires a least-squares solver to create a valid mesh, has no explicit joints or skeletal structure, may not maintain limb lengths when posed, and is not compatible with graphics pipelines and game engines.

To address these issues, Loper et al. [33] introduces SMPL, which uses vertex-based corrective offsets. Like SCAPE, SMPL factors the body into shape dependent deformations and pose dependent deformations. The SMPL model is the first statistical human body model, that is learned from 3D scans that is compatible with the gaming and animation pipeline. This is because SMPL adopts an LBS formulation complemented with learned pose-corrective blendshapes which addresses the drawbacks of LBS. SMPL is more accurate than SCAPE when trained on the same data, and to date it reamins the defacto model of the human body. SMPL is also the first model trained using the full CAESAR dataset [37], giving it a realistic shape space; previous methods used a subset of CAESAR or even smaller datasets.

SMPL, similar to many subsequent models, relates all the vertices to all joints. The SMPL model pose-corrective blendshape function is a linear function of the elements of the part rotation matrices. This results in 207 pose blendshapes with each one having a global effect. The SMPL formulation has a number of drawbacks. First SMPL learns false long range spurious correlations from the training data, where bending an elbow in SMPL results in a clearly visible bulge in the other elbow. Secondly, the formulation is not versatile, which does not allow the artist to select a subset of the blendshape required for their application. This results in a constant debt of 207 blendshapes and more than 4.2×10^6 parameters. Despite the many drawbacks of the fully connected formulation introduces in SMPL, it has become an influential formulation for many subsequent body models.

SMPL and SCAPE factors body shape and pose-dependent shape changes, but ignore

correlations between them. Several methods model this with a tensor representation [62, 4]. This allows them to vary muscle deformation with pose depending on the muscularity of the subject.

2.2.1 Sparse Pose-Corrective Blendshapes

In chapter 3 we introduce STAR (Sparse Trained Articulated Human Body Regressor), a sparse formulation of the human body deformations, where each body joint strictly influences a sparse set of the model vertices. This is because human pose deformations are largely local in nature and, therefore, pose-corrective deformations should be similarly local. Kry et al. [67] introduces EigenSkin to learn a localized model of pose deformations. STAR is similar to EigenSkin in that it models localized joint support, but, unlike EigenSkin, we infer the joint support region from posed scan data without requiring a dedicated routine of manually posing joints. Neumann et al. [68], uses sparse Principal Component Analysis (PCA) to learn local and sparse deformations of pose-dependent body deformations but do not learn a function mapping body pose to these deformations. In contrast, STAR learns sparse and local pose deformations that are regressed directly from the body pose. GHUM [15] builds on SMPL and its Rodrigues pose representation but reduces the pose parameters (including face and hands) to a 32-dimensional latent code.

The sparse formulation introduced by STAR is realistic, compared to existing human body models. In addition, the representation is artist-friendly, since an artist can directly relate any vertex deformations to a small number of blendshapes, in contrast to SMPL where all 207 blendshape influences all the vertices. Additionally, since the sparse local formulation allows the artist to exclude joints unnecessary to their applications. STAR remains compatible with the gaming and animation community standards.

The deformations resulting from body pose are influenced by both the pose itself and body shape. Consequently, STAR incorporates the subject's shape into the pose-corrective blendshape function, by conditioning the pose-corrective blendshape on the second principal component, which we show to have a strong correlation with the subject's body mass index (BMI).

2.3 Model training

Prior to this thesis, learning body models and body part models for the head and hand were two separate problems. Expressive body models, such as Frank [14] and SMPL-X [16], are trained by merging a body model with body part models for the head and hands. Frank [14] merges the body of SMPL [33] with the FaceWarehouse [69] face model and an artist-defined hand rig. Due to the fusion of different models learned in isolation, Frank looks unrealistic. SMPL-X [16] learns an expressive body model and fuses the MANO hand model [2] pose blendshapes and the FLAME head model [1] expression space. However, since MANO and FLAME are learned in isolation of the body, they do not capture the full range of motion of the the head and hands. Thus, fusing the parameters results in artifacts at the boundaries. In chapter 4 we address this by introducing a holistic federated training approach for constructing models for the body and its parts. In contrast to the construction of Frank and SMPL-X, we start with a coherent full-body model, named SUPR (A Sparse Unified Part-Based Representation), trained on a federated dataset of body, hand, head, and foot scans, then separate the model into individual body parts. Xu et al. [15] proposes GHUM & GHUML, which are trained on a federated dataset of $60K$ head, hand, and body scans and use a fully connected neural network architecture to predict pose deformations. The GHUM model cannot be separated into body parts as a result of the dense, fully connected formulation

that relates all the vertices to all the joints in the model kinematic tree. In contrast, the SUPR factored representation enables seamless separation of the body into head (SUPR-Head), hand (SUPR-Hand) and foot (SUPR-Foot) models.

2.3.1 Head Models

There are many models of 3D head shape [70, 71, 72], shape and expression [73, 74, 69, 75, 76, 77, 78] or shape, pose and expression [1]. We focus here on models with a full head template, including a neck.

The FLAME head model [1], like SMPL, uses a dense pose-corrective blendshape formulation that relates all vertices to all joints. Xu et al. [15] also propose GHUM-Head, where the template is based on the GHUM head with a retrained pose dependant deformation network (PSD). Both GHUM-Head and FLAME are trained in isolation of the body and do not have sufficient joints to model the full head degrees of freedom. In contrast to the previous methods, SUPR-Head is trained jointly with the body on a federated dataset of head and body meshes, which is critical to model the head full range of motion. It also has more joints than GHUM-Head or FLAME, which we show is crucial to model the head full range of motion.

2.3.2 Hand Models

MANO [2] is widely used and is based on the SMPL formulation where the pose-corrective blendshapes deformations are regularised to be local. The kinematic tree of MANO is based on spherical joints allowing redundant degrees of freedom for the fingers. Xu et al. [15] introduce the GHUM-Hand model where they separate the hands from the template mesh of GHUM and train a hand-specific pose-dependant corrector network (PSD). Both MANO and GHUM-Hand are trained in isolation of the body and

result in implausible deformation around the wrist area. SUPR-Hand is trained jointly with the body and has a wrist joint which is critical to model the hands full range of motion.

2.4 Foot Models

The importance of the foot in various applications has been overlooked by the graphics community due to their lack of significance in animation and gaming. We examine previous research on the measurement and capture of the human foot. The deformations of the human foot are complex and influenced by various factors, and we evaluate recent research on modeling body deformations and their applicability to modeling the human foot deformations. We conclude with a review of research on the dynamic morphology of the foot in motion.

2.4.1 Foot Measurement

Earlier measurement techniques of the human foot are based on traditional anthropometry [79]. A trained expert uses tools such as a caliper to take measurements. This method of measuring the human foot was commonly used in the past, and involves a trained expert using tools such as a caliper to take measurements. This method of measuring the human foot was not without its limitations, however. One of the main limitations of manual foot measurement was that it relies heavily on the expertise of the practitioner. If the individual taking the measurements was not highly trained or experienced in the use of tools such as a caliper, the accuracy of the measurements could be compromised. This lack of consistency in the expertise of practitioners leads to discrepancies in the measurements taken, which can have significant implications in fields such as footwear

design and manufacturing. Another limitation of manual foot measurement was that it can only be used to measure the foot in a static pose. This means that the measurements taken were not representative of the full range of motion and flexibility of the human foot. This leads to footwear that was not well-suited to the dynamic movements of the foot, resulting in discomfort and potentially even injury. Furthermore, manual foot measurement was a time-consuming and labor-intensive process, which could be prohibitively expensive for some applications. This propelled the field to investigate more robust measurement tools for the human foot. Telfer et al. [80] present a review of digital solutions to investigate the human foot shape. The measurements can be broadly divided into laser-based scanning, multi-view camera systems and multi-view camera systems combined with laser projectors. In chapter 5 we introduce an articulated foot model. Our foot model (SUPR-Foot) is trained on 4D scans captured in a multi-view camera system with laser projectors. Ballester et al. [81] shows that digital scanning technologies is substantially more robust compared to manual measurements made by a human expert.

All prior scanning systems, whether manual or digital, measures the foot in a static pose. This means that the foot were not moving or changing position during the scanning process, resulting in a static image of the foot. However, the work of Bopanna et al. [82] introduces the Dynamo system, which is a unique exception to this approach. The Dynamo system is a low cost solution that uses Intel Sense cameras to study the dynamic foot morphology. This means that the system is able to capture the foot in motion, allowing detailed and accurate representation of the foot. This is important because the foot are constantly moving and changing shape, and a static image may not accurately capture these changes. However, the Dynamo system has its limitations. It only reconstructs the outer surface of the foot, and does not capture the foot sole. This is a critical omission, as the foot sole is where many deformations occur and is important for

reconstructing how the foot deforms. Therefore, while the Dynamo system is a valuable tool for studying dynamic foot morphology, it is not a complete solution.

Coudert et al. [83] introduces the foot scanner consisting of 3 pairs of stereoscopic sensors that captures the foot deformation during walking and moving sequences. The system returns high-resolution scans; however, it has a low frame rate and still fails to capture the foot sole.

In contrast to prior work, we use a custom built foot scanner, made by 3dMD. The foot scanner is a mechanically stable structure that is designed to accommodate human subjects weighing up to 150 kg. This is a key feature of the scanner, as it allows the capture of data from a wide range of individuals, performing dynamic motions such as jumping and running. The scanner features a transparent glass platform that allows the detailed reconstruction of the foot, including the toes and foot sole. This is essential for the study of foot deformations, including during the loading phase of a motion. The output foot scans are high resolution, which preserves the full structure of the foot, including the individual toes. This enables the construction of accurate detailed models.

2.4.2 Statistical Foot Models

Statistical 3D foot models are more nascent compared to the body, head, and hands. Conard et al. [84] proposes the first statistical shape model of the human foot, which is a PCA space learned from static foot scans. However, the human foot morphology varies with motion dynamics, and models learned from static scans will not capture the full complexity of the 3D foot deformations. Amstutz et al. [85] learns a low resolution foot model from a foot database of 397 of static foot manually designed by an anatomy specialist. They learn a PCA shape space of the foot using 12 principal components to model the foot. Bopanna [3] captures subjects using the DynaMo system [82] and

registers the scans to sequence where a subject is in motion, and learns a PCA model. In contrast to all prior work, SUPR-Foot is the first articulated model for human foot that supports the full range of motion of the human foot, including the movement of the ankle and the individual toes.

2.4.3 Deformation Modeling

Extensive research has been conducted on modeling body deformations. Numerous studies explores various approaches to predict body deformations based on factors such as body pose or shape. Previous work fails to model deformation due to contact, which is particularly crucial for the human foot. This difficulty primarily arises from the complexity of accurately capturing contact-induced deformations. In contrast to previous work, in chapter 5 we present an approach using a novel neural blendshape function that establishes a correlation between foot sole deformation, foot pose, foot shape, and ground contact. Learning this formulation is only possible due to the custom-built capture setup that is capable of accurately capturing the human foot sole during ground contact.

2.5 Data Efficient Training

Loper et al. [33] introduces the SMPL body model, which has become the most widely adopted human body model to date. The model is predicated on a pose blendshape formulation that applies corrective offsets to counteract the well-documented limitations of linear blend skinning. The training of a SMPL model utilizes an artist-designed prior to guide the skinning weights and joint regressors. Although most of the SMPL parameters reside within the pose blendshape formulation, the training process is based solely on an *L2* regularization term without the use of a specific prior. Consequently, the training of

SMPL requires expert supervision, a substantial training dataset and manual adjustments throughout the model training process to ensure proper regularization of the blendshapes. Despite these efforts and the vastness of the training data, the model still tends to learn incorrect correlations, such as the association of bending one elbow with a bulge in the opposite elbow.

In chapter 3, we introduce the STAR model, which also demands a large dataset comprising multiple subject identities across various body poses to learn its sparse formulation. Similarly, the SUPR model, presented in chapter 4, requires a prohibitively large dataset of 1.2 million registrations for training. The GHUM model, as reported by Xu et al. [15], is trained on a dataset comprising 70K registrations. The sheer scale of the data required and the expertise needed for training these models presents a significant challenge for artists and animators who lack machine learning expertise and access to extensive training datasets. Zhou et al. [86] highlight the difficulty and expertise required to train the SMPL model deformation function and other similar parametric models, and instead proposed a neural architecture to disentangle pose and shape spaces, yet the architecture is still trained on multiple subjects and requires a large training dataset to generalize. Zeitvogel et al. [87, 88] introduce a mixed approach to learning from raw body scans and registrations, which is useful when learning a model from a large dataset of scans.

In chapter 6, we propose AVATAR (Articulated Virtual Humans Trained By Bayesian Inference from a Single Scan), a novel algorithm for learning personalized human body models based on the SMPL representation from a single scan. Traditional training algorithms for body models infer a single point estimate of the model parameters, making them susceptible to overfitting the training dataset. AVATAR, on the other hand, approaches model training as a Bayesian inference problem. It begins with a prior dis-

tribution of the model parameters and seeks to infer a posterior distribution of possible model parameters, thereby enhancing the training process's robustness against overfitting. In contrast to all existing work which focuses on scaling training body models to a large dataset of scans, in AVATAR our focus is on data-efficient learning and inference of high-fidelity personalized mesh-based engine-ready digital characters from a single scan.

Chapter 3

Sparse Body Models

3.1 Introduction

Human body models are widely used to reason about 3D body pose and shape in images and videos. While several models have been proposed [8, 10, 62, 63, 4, 64, 66, 61, 15], SMPL [33] is currently the most widely use in academia and industry. SMPL is trained from thousands of 3D scans of people and captures the statistics of human body shape and pose. Key to SMPL's success is its compact and intuitive parametrization, decomposing the 3D body into pose parameters $\vec{\theta} \in \mathbb{R}^{72}$ corresponding to axis angle rotations of 24 joints and shape $\vec{\beta} \in \mathbb{R}^{10}$ capturing subject identity (the number of shape parameters can be as high as 300 but most research uses only 10). This makes it useful to reason about 3D human body pose and shape given sparse measurements, such as IMU accelerations [89, 90, 91], sparse mocap markers [92, 93] or 2D key points in images and videos [94, 17, 18, 19, 95, 96, 97, 98].

While SMPL is widely used, it suffers from several drawbacks. SMPL augments traditional linear blend skinning (LBS) with pose-dependent corrective offsets that are

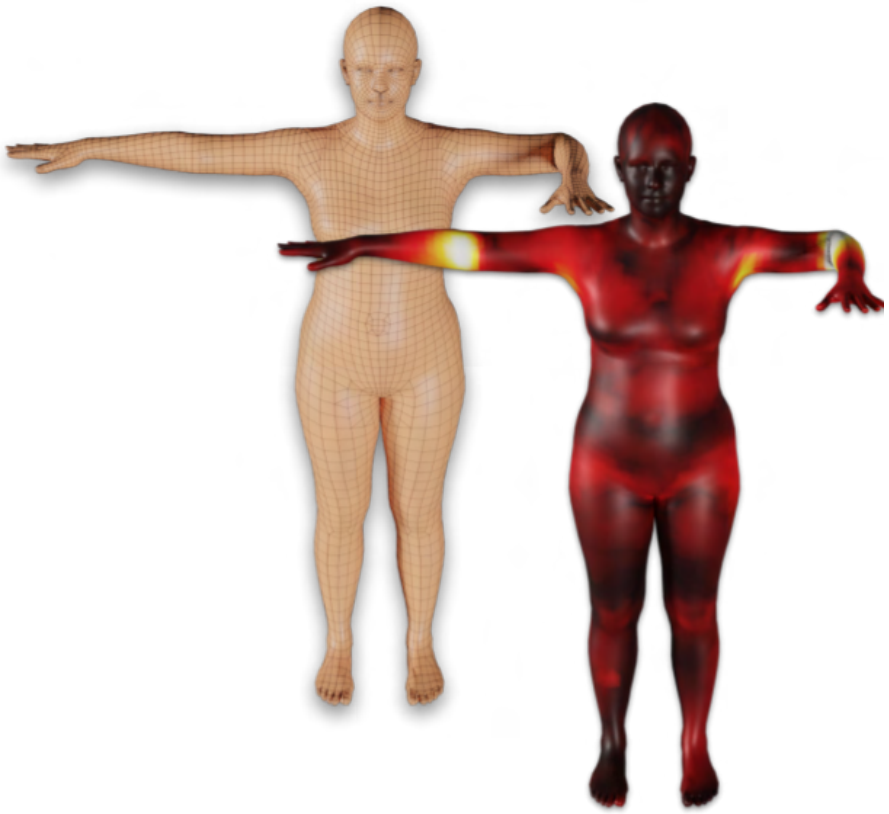


Figure 3.1: **False SMPL Deformations:** Heat maps illustrate the magnitude of the pose-corrective offsets. The spurious long-range correlations learned by the pose-corrective blendshapes SMPL. Bending one elbow results in a visible bulge in the other elbow.

learned from 3D scans. Specifically, SMPL uses a pose-corrective blendshape function $\mathcal{B}_{\mathcal{P}}(\vec{\theta}) : \mathbb{R}^{|\vec{\theta}|} \rightarrow \mathbb{R}^{3N}$, where N is the number of mesh vertices.

The function $\mathcal{B}_P$ predicts corrective offsets for every mesh vertex such that, when the model is posed, the output mesh looks realistic. The function $\mathcal{P}$ can be viewed as a fully connected layer (FC), that relates the corrective offsets of every mesh vertex to the elements of the part rotation matrices of all the body joints. This dense blendshape formulation has several drawbacks. First, it significantly inflates the number of model parameters to $>$ 4.2 million, making SMPL prone to overfitting during training. Even with numerous regularization terms, the model learns spurious correlations in the training

set, as shown in Figure 3.1; moving one elbow causes a bulge in the other elbow.

This is problematic for graphics, model fitting, and deep learning. The dense formulation causes dense spurious gradients to be propagated through the model. A loss on the mesh surface back propagates spurious gradients to geodesically distant joints. The existing formulation of the pose-corrective blendshapes limits the model compactness and visual realism.

To address this, we create a new compact human body model, called **STAR** (Sparse Trained Articulated Regressor), that is more accurate than SMPL yet has sparse and spatially-local blendshapes, such that a joint only influences a sparse set of vertices that are geodesically close to it. The original SMPL paper acknowledges the problem and proposes a model called SMPL-LBS-Sparse that restricts the pose-corrective blendshapes such that a vertex is only influenced by joints with the highest skinning weights. SMPL-LBS-Sparse, however, is less accurate than SMPL.

Our key insight is that the influence of a body joint on the model vertices should be inferred from the training data. The main challenge is formalizing a model and training objective such that we learn meaningful joint support regions that are sparse and spatially local as shown in Figure 3.3. To this end we formalize a differentiable thresholding function based on the Rectified Linear Unit operator, **ReLU**, that learns to predict 0 activations for irrelevant vertices in the model. The output activations are used to mask the output of the joint blendshape regressor to only influence vertices with non-zero activations. This results in a sparse model of pose-dependent deformation.

We go further in improving the model compactness. SMPL uses a Rodrigues representation of the joint angles and has a separate pose-corrective regressor for each element of the matrix, resulting in 9 regressors per joint. We switch to a quaternion representation with only 4 numbers per joint, with no loss in performance. This, in combination with the

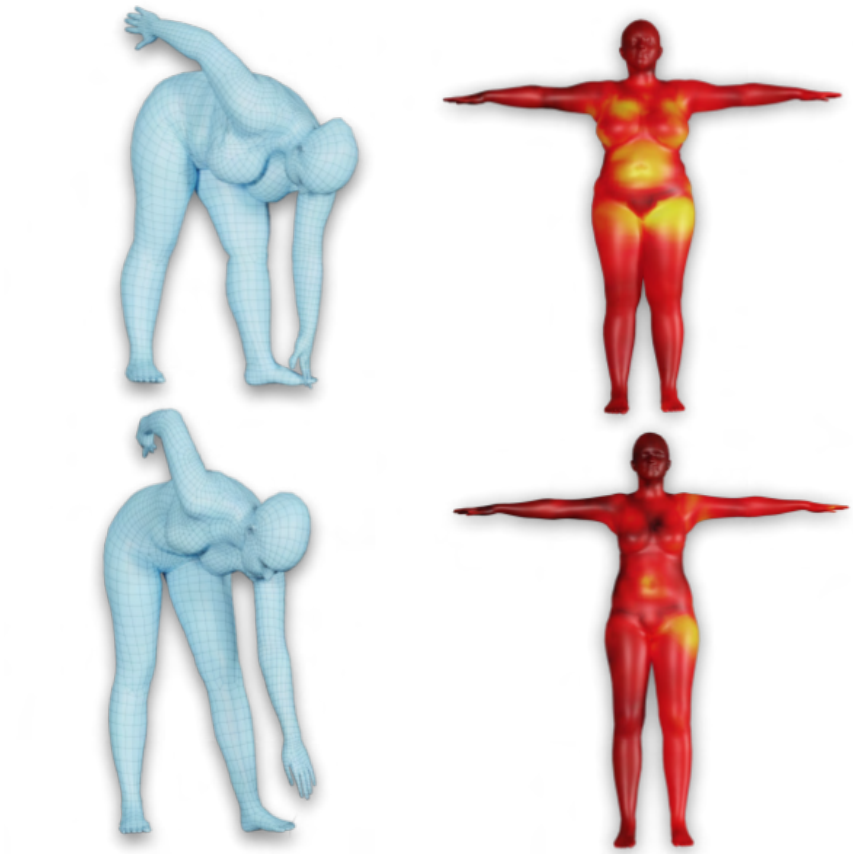


Figure 3.2: **SMPL Deformations Limitations:** Two subject registrations (show in blue) with two different body shapes (High BMI) and (Low BMI). While both are in the same pose, the corrective offsets should be different since body deformations are influenced by both body pose and body shape. The SMPL pose-corrective offsets are the same regardless of body shape.

sparsity, means that STAR has only 20% of the parameters of SMPL. We evaluate STAR by training it on different datasets. When we train STAR on the same data as SMPL, we find that it is more accurate on held-out test data. Note that the use of quaternions is an internal representation change from SMPL and transparent to users who can continue to use the SMPL pose parameters.

SMPL disentangles shape due to identity from shape due to pose. This is a strength because it results in a simple model with additive shape functions. It is also a weakness, however, because it cannot capture correlations between body shape and how soft tissue

deforms with pose, as shown in Fig. 3.2. To address this we extend the existing pose-corrective formulation by regressing the correctives using both body pose $\vec{\theta}$ and body shape $\vec{\beta}$. Here we use the second principal component of the of the body shape space, which correlates highly with Body Mass Index (BMI). This change results in more realistic pose-based deformations.

SMPL is used in many fields such as apparel and healthcare because it captures the statistics of human body shape. The SMPL shape space was trained using the CAESAR database, which contains 1700 male and 2107 female subjects. CAESAR bodies, however, are distributed according to the US population in 1990 [37] and do not reflect global body shape statistics today. Additionally, CAESAR's capture protocol dressed all women in the same sports-bra-type top, resulting in a female chest shape that does not reflect the diversity of shapes found in real applications. We show that SMPL trained on CAESAR is not able to capture the variation in the more recent, and more diverse, SizeUSA dataset of 10,000 subjects (2845 male and 6436 female) [54], and vice versa. To address these problems, we train STAR from the combination of CAESAR and SizeUSA scans and show that the complementary information contained in both datasets enables STAR to generalize better to unseen body shapes.

We summarize our contributions by organizing them around impact areas where SMPL is currently used:

1. **Computer vision:** We propose a compact model that is 80% smaller than SMPL. We achieve compactness in two ways: First, we formalize sparse corrective blendshapes and learn the set of vertices influenced by each joint. Second, we use quaternion features for offset regression. While STAR is more compact than SMPL, it generalizes better on held-out test data.

2. **Graphics:** Non-local deformations make animation difficult because changing the

pose of one body part affects other parts. Our local model fixes this problem with SMPL.

3. **Health:** Realistic avatars are important in health research. We increase realism by conditioning the pose-corrective blendshapes on body shape. Bodies with different BMI produce different deformations.

4. **Clothing Industry:** Accurate body shape matters for clothing. We use the largest training set to date to learn body shape and show that previous models are insufficient to capture the diversity of human shape.

The model is a drop-in replacement for SMPL, with the same pose and shape parametrization.

3.2 Model

STAR is a vertex-based LBS model complemented with a learned set of shape and pose-corrective functions. Similar to SMPL, we factor the body shape into the subject's intrinsic shape and pose-dependent deformations. In STAR we define a pose-corrective function for each joint, j, in the kinematic tree. In contrast to SMPL, we condition the pose-corrective deformation function on both body pose $\vec{\theta} \in \mathbb{R}^{|\vec{\theta}|}$ and shape $\vec{\beta} \in \mathbb{R}^{|\vec{\beta}|}$. Additionally, during training, we use a non-linear activation function, $\phi(.)$, that selects the subset of mesh vertices relevant to the joint j. The pose-corrective blendshape function makes predictions for a subset of the mesh vertices. We adopt the same notation used in SMPL [33]. We start with an artist defined template, $\overline{T} \in \mathbb{R}^{3N}$ in the rest pose $\vec{\theta}^*$ (i.e. T-Pose) where $N = 6890$ is the number of mesh vertices. The model kinematic tree contains $K = 24$ joints, corresponding to 23 body joints in addition to a root joint. The template $\overline{T}$ is then deformed by a shape-corrective blendshape function B_S that captures

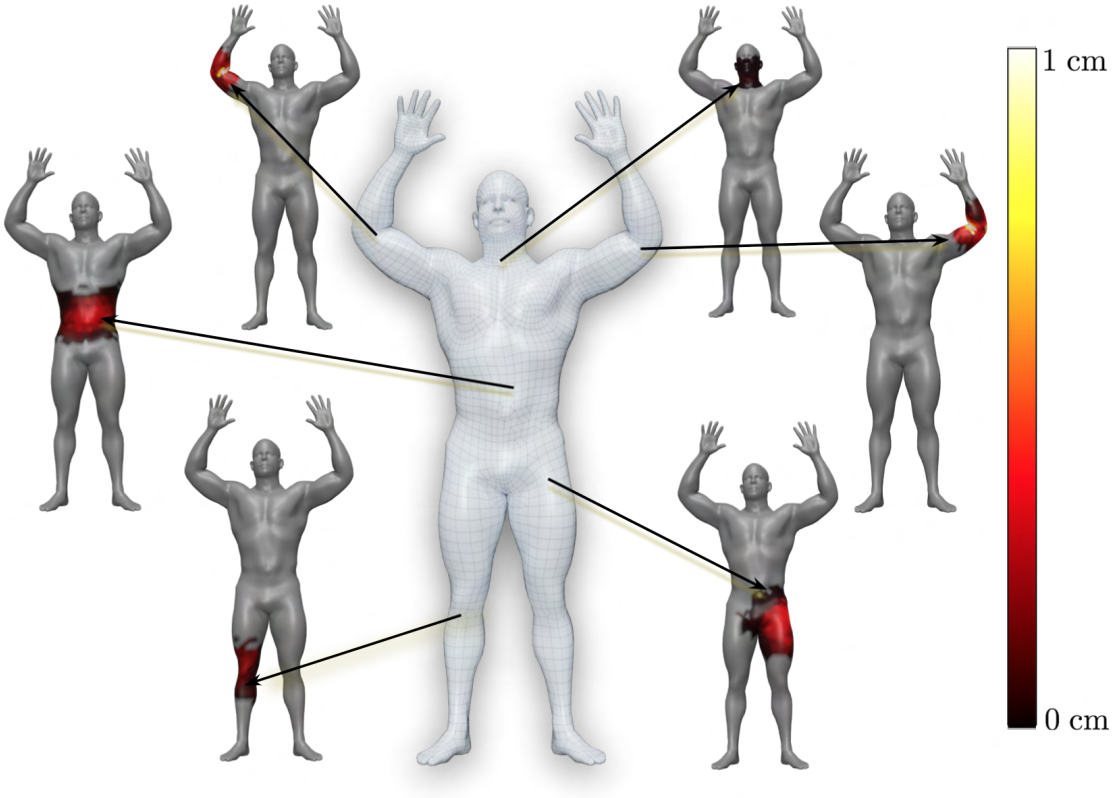


Figure 3.3: **Sparse Local Pose-Correctives:** STAR factors pose-dependent deformation into a set of sparse and spatially-local pose-corrective blendshape functions, where each joint influences only a sparse subset of mesh vertices. The white mesh is STAR fit to a 3D scan of a professional body builder. The arrows point to joints in the STAR kinematic tree and the corresponding predicted corrective offsets for the joint. The heat map encodes the magnitude of the corrective offsets. The joints have no influence on the gray mesh vertices.

the subject's identity and a function B_P that adds corrective offsets such that the mesh looks realistic when posed.

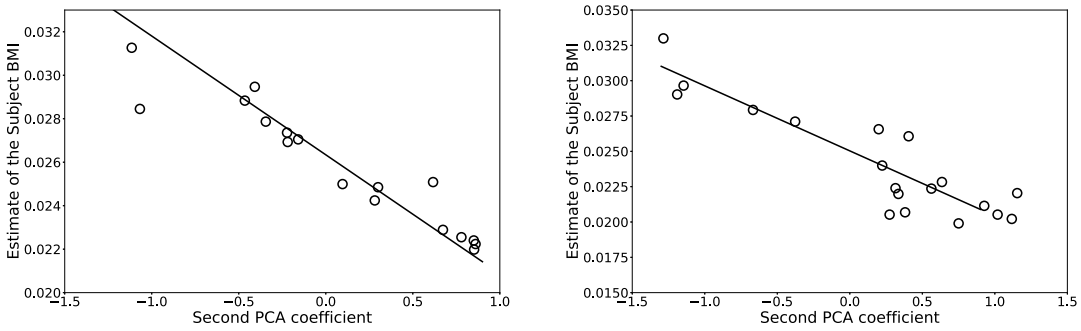
Shape Blendshapes. The shape blendshape function $B_S(\vec{\beta};\mathcal{S}) : \mathbb{R}^{|\vec{\beta}|} \rightarrow \mathbb{R}^{3N}$ maps the identity parameters $\vec{\beta}$ to vertex offsets from the template mesh as

$$B_S(\vec{\beta};\mathcal{S}) = \sum_{n=1}^{|\vec{\beta}|} \beta_n S_n, \tag{3.1}$$

where $\vec{\beta} = [\beta_1, \cdots, \beta_{|\beta|}]$ are the shape coefficients, and $\mathcal{S} = [S_1, \cdots, S_{|\beta|}] \in \mathbb{R}^{3N \times |\vec{\beta}|}$ are the principal components capturing the space of human shape variability. The shape correctives are added to the template:

$$\vec{V}_{shaped} = \overline{T} + B_S(\vec{\beta}; \mathcal{S}), \tag{3.2}$$

where $\vec{V}_{shaped}$ contains the vertices representing the subject's physical attributes and identity.



(a) Male Subjects

(b) Female Subjects

Figure 3.4: **BMI and PCA:** There is a strong linear relationship between the BMI of SMPL training subjects and the second shape principal component, β_2, for both the male and female subjects.

Pose and Shape Corrective Blendshapes. The output of the shape-corrective blendshape function, $\vec{V}_{shaped}$, is further deformed by a pose-corrective function. The pose-corrective function is conditioned on both pose and shape and adds corrective offsets such that, when the mesh is posed with LBS, it looks realistic. We denote the kinematic tree unit quaternion vector as $\vec{q} \in \mathbb{R}^{96}$ (24 joints each represented with 4 parameters). The pose-corrective function is denoted as $B_P(\vec{q}, \beta_2) \in \mathbb{R}^{|\vec{q}| \times 1} \rightarrow \mathbb{R}^{3N}$, where β_2 is the PCA coefficient of the second principal component, which highly correlates with the

body mass index (BMI) as shown in Figure 3.4. The STAR pose-corrective function is factored into a sum of pose-corrective functions:

$$B_P(\vec{q},\beta_2;\mathbf{K},\mathbf{A}) = \sum_{j=1}^{K-1} B_P^j(\vec{q}_{ne(j)},\beta_2;\mathbf{K}_j,A_j), \tag{3.3}$$

where a pose-corrective function is defined for each joint in the kinematic tree excluding the root joint. The per-joint pose-corrective function $B_P^j(\vec{q}_{ne(j)},\beta_2;\mathbf{K}_j,A_j)$ predicts corrective offsets given $\vec{q}_{ne(j)} \subset \vec{q}$, where $\vec{q}_{ne(j)}$ is a set containing the joint j and its direct neighbors in the kinematic tree. This formulation results in spatially local pose-corrective deformation function compared to SMPL. $\mathbf{K}_j \in \mathbb{R}^{3N\times|\vec{q}_{ne(j)}|+1}$ is a linear regressor weight matrix and A_j are the activation weights for each vertex, both of which are learned. Each pose-corrective function, $B_P^j(\vec{q}_{ne(j)},\beta_2)$, is defined as a composition of two functions, an activation function and a pose-corrective regressor.

Activation Function. For each joint, j, we define a learnable set of mesh vertex weights, $A_j = [w_j^1,\cdots,w_j^N] \in \mathbb{R}^N$, where $w_j^i \in \mathbb{R}$ denotes the weight of the i^{th} mesh vertex with respect to the j joint. The weight w_j^i for each vertex i is initialized as the reciprocal of the minimum geodesic distance to the set of vertices around joint j, normalized to the range $[0,1]$. The weights are thresholded by a non-linear activation function, specifically a rectified linear unit (ReLU):

$$\phi(w_j^i) = \begin{cases} 0, & \text{if } w_j^i \leq 0, \\ w_j^i, & \text{otherwise,} \end{cases} \tag{3.4}$$

such that during training, vertices with a $w_j^i \leq 0$ have weight 0. The remaining set of vertices with $w_j^i > 0$ defines the support region of joint j.

Pose-Corrective Regressor. The pose-corrective function for each body joint is defined as $P_j(\vec{q}_{ne(j)}) \in \mathbb{R}^{|\vec{q}_{ne(j)}|+1} \rightarrow \mathbb{R}^{3N}$, which regresses corrective offsets given the joint and its direct neighbors' quaternion values

$$P_j(\vec{q}_{ne(j)}, \beta_2; \mathbf{K}_j) = \mathbf{K}_j((\vec{q}_{ne(j)} - \vec{q}^*_{ne(j)})^T | \beta_2)^T, \tag{3.5}$$

where $\vec{q}^*_{ne(j)}$ is the vector of quaternion values for the set of joints $ne(j)$ in rest pose, and β_2 is concatenated to the quaternion difference vector. $\mathbf{K}_j \in \mathbb{R}^{3N \times |\vec{q}_{ne(j)}|+1}$ is the regression matrix for joint j's pose-correctives offsets. The predicted pose-corrective offsets in Equation (3.5) are masked by the joint activation function:

$$B^j_P(\vec{q}_{ne(j)}; A_j, \mathbf{K}_j) = \phi(A_j) \circ P_j(\vec{q}_{ne(j)}, \beta_2; \mathbf{K}_j), \tag{3.6}$$

where $\vec{X} \circ \vec{Y}$ is the element wise Hadamard product between the vectors $\vec{X}$ and $\vec{Y}$. During training, vertices with zero activation with respect to joint j, will have no corrective offsets added to them. Therefore, when summing the contribution of the individual joint pose-corrective functions in Equation (3.3), each joint only contributes pose-correctives to the vertices for which there is support.

Blend Skinning. Finally, the mesh with the added pose and shape corrective offsets is transformed using a standard skinning function $W(\overline{T}, \vec{J}, \vec{\theta}, \mathcal{W})$ around the joints, $\vec{J} \in \mathbb{R}^{3K}$ and linearly smoothed by a learned set of blend weight parameters $\mathcal{W} \in \mathbb{R}^{6890 \times 24}$. The joint locations are intuitively influenced by the body shape and physical attributes. Similar to SMPL, the joints $\vec{J}(\vec{\beta}; \mathcal{J}, \overline{T}, \mathcal{S}) = \mathcal{J}(\vec{V}_{shaped})$ are regressed from $\vec{V}_{shaped}$ by a sparse function $\mathcal{J} : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{3K}$.

To summarize, STAR is full defined by:

$$M(\vec{\beta},\vec{\theta}) = W(T_p(\vec{\beta},\vec{\theta}),J(\vec{\beta}),\vec{\theta},\mathcal{W}), \tag{3.7}$$

where T_P is defined as:

$$T_p(\vec{\beta},\vec{\theta}) = \overline{T} + B_S(\vec{\beta}) + B_P(\vec{q},\beta_2), \tag{3.8}$$

where $\vec{q}$ is the quaternion representation of pose $\vec{\theta}$. The STAR model is fully parameterized by 72 (i.e. 24 * 3) pose parameters $\vec{\theta}$ in axis-angle representation, and up to 300 shape parameters $\vec{\beta}$.

3.2.1 Model Training

STAR training is similar to SMPL [92]. The key difference is the training of the pose-corrective function in Equation (3.3). STAR's pose-corrective blendshapes are trained to minimize the *vertex-to-vertex* error between the model predictions and the ground-truth registrations. A registration is a tight fit of STAR's mesh to a raw scan. In each iteration, the model parameters (**A**,**K**) are minimized by stochastic gradient descent across a batch of B registrations, denoted as $\vec{R} \in \mathbb{R}^{3N}$. The data term is given by:

$$\mathcal{L}_D = \frac{1}{B}\sum_{i=1}^{B} ||M(\vec{\beta}_i,\vec{\theta}_i) - \vec{R}_i||_2. \tag{3.9}$$

In addition to the data term, we regularize the pose-corrective regression weights (**K**) with an $L2$ norm:

$$\mathcal{L}_B = \lambda_b \sum_{i=1}^{K-1} ||\mathbf{K}_i||_2, \tag{3.10}$$

Iteration	λ_b	λ_c	λ_p	λ_s
1	1	1e-3	50	8e2
2	1e-1	1e-4	50	8e2
3	1e-2	1e-4	50	8e2
4	1e-5	1e-5	50	8e2

Table 3.1: Annealing schedule of the regularization parameters for each training iteration.

where K is the number of joints in STAR and λ_b is a scalar constant. In order to induce sparsity in the activation masks $\phi(\cdot)$, we use an $L1$ penalty

$$\mathcal{L}_A = \lambda_c || \sum_{i=1}^{K-1} \phi_j(A_j) ||_1, \tag{3.11}$$

where λ_c is a scalar constant. Similar to SMPL we use a sparsity regularizer term on the skinning weights $\mathcal{W}$ and regularize the skinning weights to initial artist-defined skinning weights, $\mathcal{W}_{\text{prior}} \in \mathbb{R}^{N \times K}$:

$$\mathcal{L}_W = \lambda_p ||\mathcal{W} - \mathcal{W}_{\text{prior}}||_2 + \lambda_s ||\mathcal{W}||_1, \tag{3.12}$$

where λ_p and λ_s are scalar constants. To summarize the complete training objective is given by

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_B + \mathcal{L}_A + \mathcal{L}_W. \tag{3.13}$$

The objective in Equation (3.13) is minimized with respect to the skinning weights $\mathcal{W}$, pose-corrective regression weights $\mathbf{K}_{1:24}$, and activation weights $A_{1:24}$.

We train the model iteratively. STAR is trained for 4 iterations, in each training iteration we anneal the regularization parameters as outlined in Table 3.1.

Figure 3.5: **STAR Activations:** A sample of the joints activation functions output before training and the bottom row shows the output after training (gray is zero). A joint only predicts deformations for the mesh parts with non-zero activation.

3.3 Experiments

3.3.1 Activation

Key to learning the sparse and spatially local pose-corrective blendshapes are the joint activation functions introduced in Equation (3.4). During training the output of the activation functions becomes more sparse, limiting the number of vertices a joint can influence. Figure 3.5 summarizes a sample of the activation functions output before and after training. As a result of the output of the activation functions becoming more sparse, the number of model parameters decreases. By the end of training, the male model pose blendshapes contains 3.37×10^5 non-zero parameters and the female model contains 3.94×10^5 non-zero parameters. In contrast to SMPL which has a dense pose-corrective

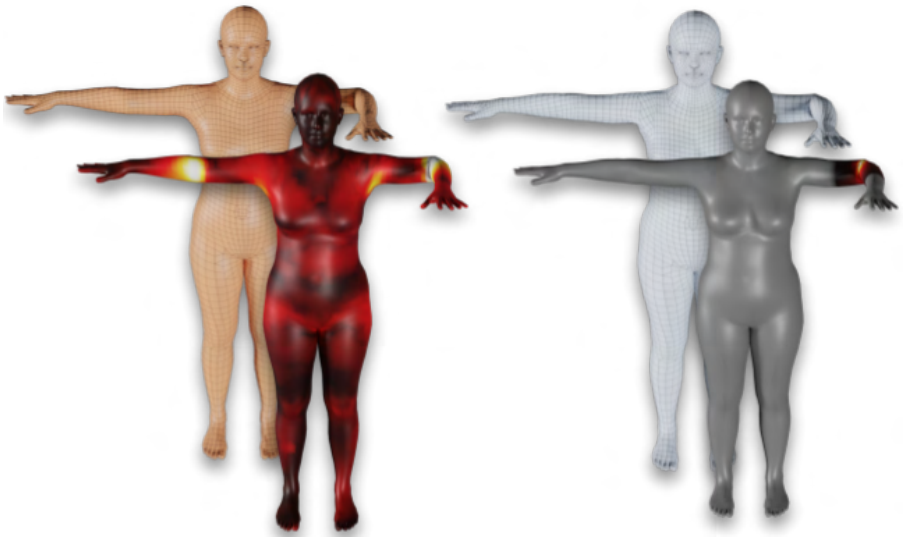


Figure 3.6: **STAR vs SMPL Pose-Dependent Deformations:** SMPL (brown) and STAR (white) in the rest pose except for the left elbow, which is rotated. The heat map visualizes the corrective offsets for each model caused by moving this one joint. Note that unlike STAR, SMPL has spurious long-range displacements.

blendshape formulation with 4.28×10^6 parameters. At test time only the non-zero parameters need to be stored.

Figure 3.6 show a SMPL model bending an elbow resulting in a bulge in the other elbow, as a result of the pose corrective blendshapes learning long range spurious correlations from the training data. In contrast, STAR correctives are spatially local and sparse, this is a result of the learned local sparse pose-corrective blendshape formulation of STAR.

3.3.2 Model Generalization

While the learned activation masks are sparse and spatially local, which is good, it is equally important that the model still generalizes to unseen bodies. To this end, we evaluate the model generalization on held out test subjects. The test set we use contains the publicly available Dyna dataset [99] (the same evaluation set used in evaluating the

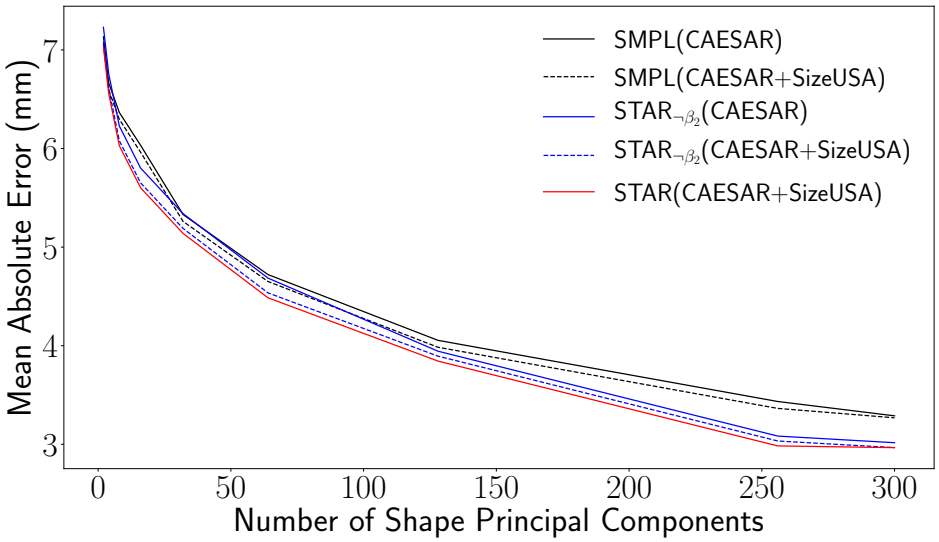


Figure 3.7: **Generalization Accuracy**: Evaluating STAR and SMPL on unseen bodies. STAR$_{\neg\beta_2}$(CAESAR) is STAR trained on CAESAR with pose-correctives depending on pose only (i.e. independent of β_2), STAR$_{\neg\beta_2}$(CAESAR+SizeUSA) is STAR trained on CAESAR and SizeUSA with pose-corrective blendshapes depending on pose only, and STAR(CAEAR+SizeUSA) is STAR trained on CAEASAR and SizeUSA with pose and shape dependent pose-corrective blendshapes.

SMPL model), in addition to the 3DBodyTex dataset [100], which contains static scans for 100 male and 100 female subjects in a diversity of poses. The total test set contains 570 registered meshes of 102 male subjects and 104 female subjects. We fit the models by minimizing the vertex to vertex mean absolute error (v2v), where the pose $\vec{\theta}$ and shape parameters $\vec{\beta}$ are the free optimization variables. We report the mean absolute error in (mm) as a function of the number of used shape coefficients in Figure 3.7. We first evaluate SMPL and STAR when they are both trained using the CAESAR dataset. In this evaluation both models are trained on the exact same pose and shape data. Since they both share the same topology and kinematic tree, differences in the fitting results are solely due to the different formulation of the two models. In Figure 3.7, STAR uniformly generalizes better than SMPL on the unseen test subjects. A sample qualitative

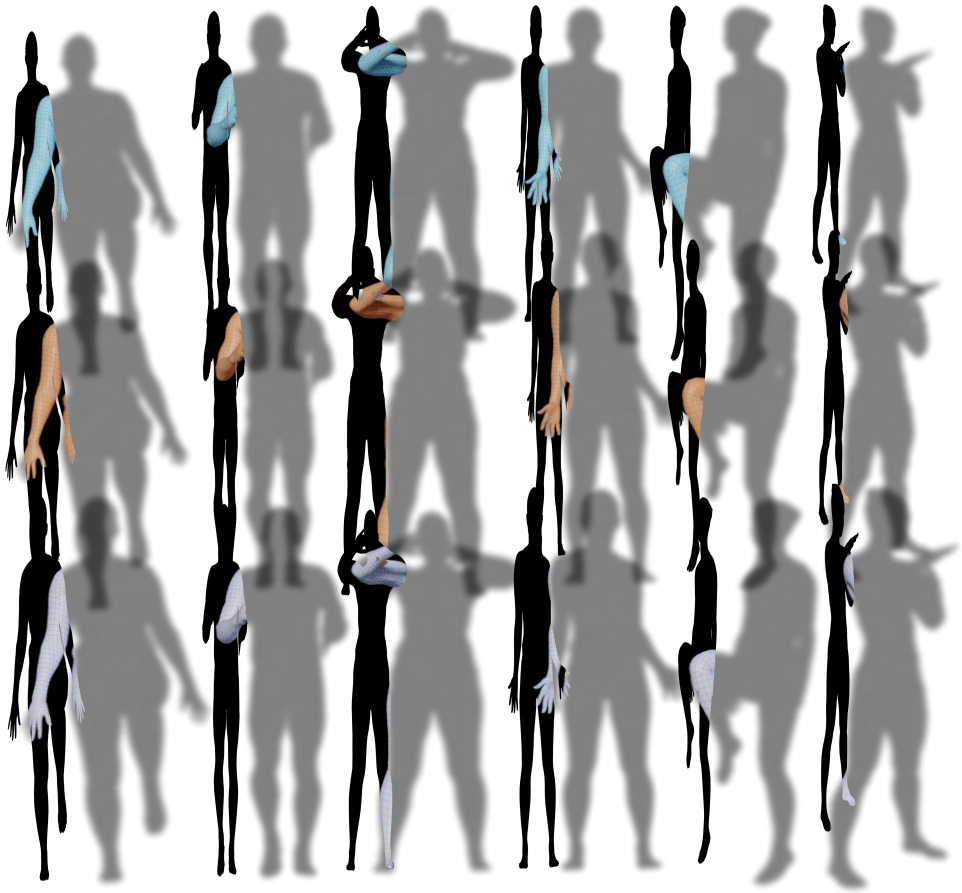


Figure 3.8: **Qualitative Evaluation**: Comparison between SMPL and STAR. The ground truth registrations are shown in blue, the corresponding SMPL model fit meshes are shown in brown and STAR fits are shown in white. Here, both STAR and SMPL are trained on the CAESAR database.

comparison between SMPL and STAR fits is shown in Figure 3.8.

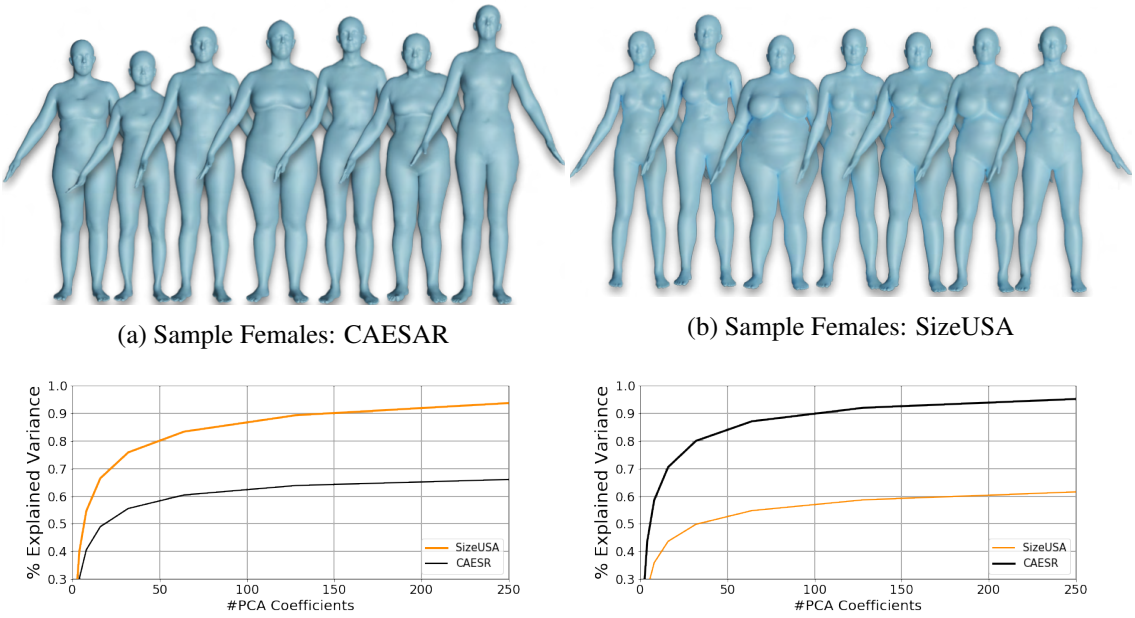
3.3.3 Extended Training Data

The CAESAR dataset is limited in its diversity, consequently limiting model generalization. Consequently, we extend the shape training database to include the SizeUSA database [54]. SizeUSA contains low quality scans of 2845 male and 6434 females with ages varying between 18 to 66+; a sample of the SizeUSA bodies compared to the CAE-

SAR bodies are shown in Figure 3.9a and Figure 3.9b. We evaluate the generalization power of models trained separately on CEASER and SizeUSA. We do so by computing the percentage of explained variance of the SizeUSA subjects given a shape space trained on the CAESAR subjects, and vice versa. The results are shown in Figure 3.9 for the female subjects. The key insight from this experiment is that a shape space trained on a single data set was not sufficient to explain the variance in the other data set. This suggests that training on both datasets should improve the model shape space expressiveness.

We retrain train both STAR and SMPL on the combined CAESAR and SizeUSA datasets and evaluate the model generalization on the held out test set as a function of the number of shape coefficient used as shown in Figure 3.7. Training on both CAESAR and SizeUSA results in both SMPL and STAR generalizing better than when trained only on CAESAR. We further note that STAR still uniformly generalizes better than SMPL when both models are trained on the combined CAESAR and SizeUSA dataset. Importantly STAR is more accurate than SMPL despite the fact that uses many fewer parameters. Finally we extend the pose-corrective blendshapes of STAR to be conditioned on both body pose and body shape and evaluate the model on the held out set. This results in a further improvement in the model generalization accuracy that, while modest, is consistent.

Explained Variance. Figure 3.10 shows the percentage of explained variance for male and female subjects, for shape spaces trained on CAESAR subjects only, on SizeUSA subjects only, or jointly on SizeUSA and CAESAR subjects. Figure 3.10 highlights that a shape space trained on a single dataset is insufficient to explain the variance in body shape for the other data set subjects. This emphasizes that the data is not redundant. Only a shape space trained on both data sets is sufficient to explain the variance in body shapes across both datasets. This observation is consistent for both male and female subjects.



(a) Sample Females: CAESAR

(b) Sample Females: SizeUSA

(c) Shape Space Trained on SizeUSA

(d) Shape Space Trained on CAESAR

Figure 3.9: **Explained Variance:** The percentage of explained variance of SizeUSA and CAESAR subjects when shape space is trained on SizeUSA is shown in Figure 3.9c and when the shape space is trained on CAESAR subjects in Figure 3.9d.

Figure 3.11 highlights the most poorly reconstructed body shapes from both CAESAR and SizeUSA when reconstructed using a shape space trained on the other dataset. The SizeUSA dataset contains extremely obese male subjects, which are poorly reconstructed under a CAESAR shape space, as shown in Figure 3.11c. The CAESAR female shape space is biased to a sport's bra chest shape, hence fails to accurately reconstructs the SizeUSA females chest shapes as shown in Figure 3.11d.

3.4 Discussion

STAR has 93 pose-corrective blendshapes compared to 207 in SMPL and is 80% smaller than SMPL. It is surprising that it is able to uniformly perform better than SMPL when trained on the same data. This highlights the fact that the local and sparse assumptions

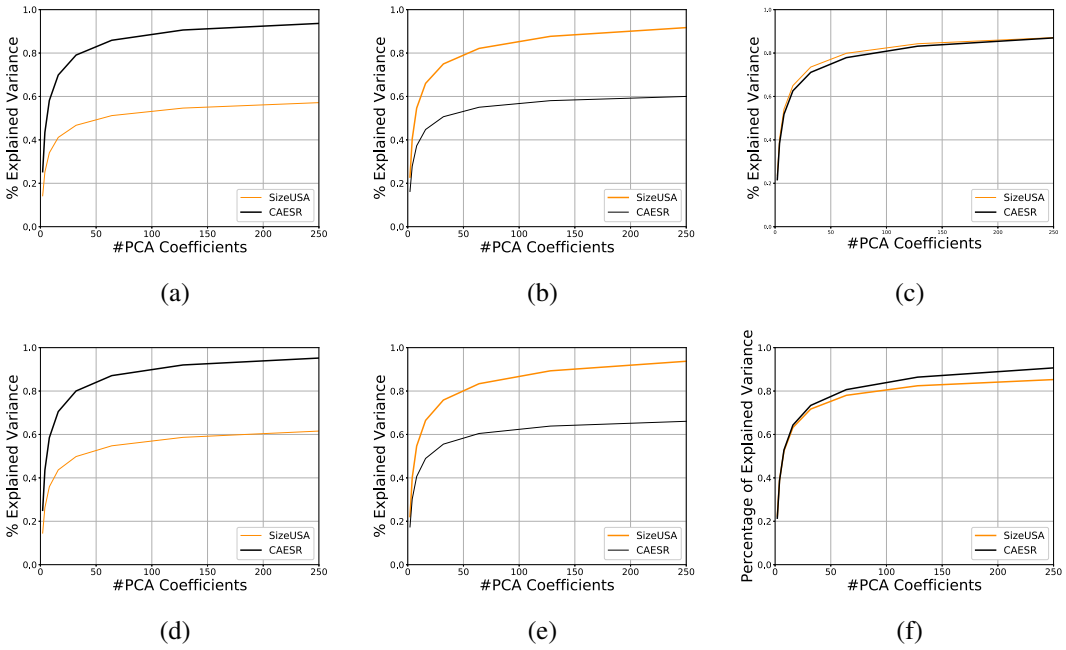
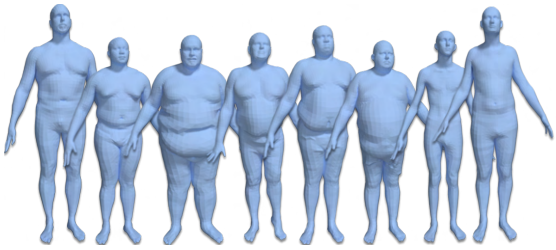


Figure 3.10: **Percentage of explained variance**: Figure highlighting the percentage of explained variance of SizeUSA and CAESAR subjects when reconstructed by a shape space trained on CAESAR subjects (left column), SizeUSA subjects (middle column) and both SizeUSA and CAESAR subjects (right column). Top row is for male subjects and bottom row is female subjects. A shape space trained on either dataset is insufficient to explain the variance in the other dataset; this is consistent for both male and female subjects. Only a shape space trained on the combined male and female subjects was able to adequately explain the variance for both populations.

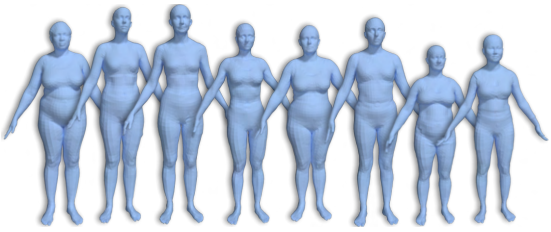
of the pose-corrective blendshapes is indeed realistic a priori knowledge that should be incorporated in any body model. Importantly, having fewer parameters means that STAR is less likely to overfit, even though our non-linear model makes training more difficult.

For SMPL, the authors report that enforcing sparsity of the pose-corrective blendshapes results in worse results than SMPL. We adopt a different approach, where for each body joint we learn the sparse set of model vertices influenced by the joint movement. The key strength of our approach is that it is data driven.

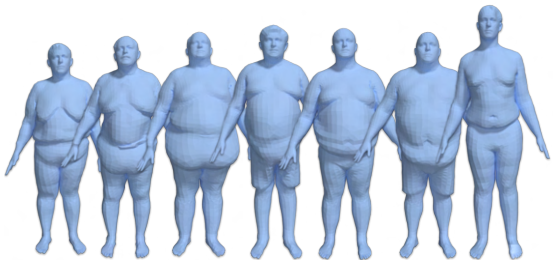
We are able to learn spatially local and sparse joint support regions due to two key implementation details: The initialization of the vertex weight A_j with the normalized



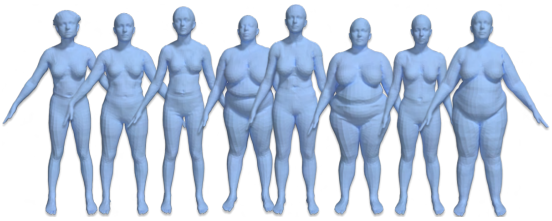
(a) Male CAESAR reconstructed under SizeUSA.



(b) Female CAESAR subjects reconstructed under SizeUSA.



(c) Male SizeUSA subjects reconstructed under CAESAR.



(d) Female SizeUSA reconstructed under CAESAR.

Figure 3.11: **Reconstruction Error**: Subjects with the high reconstruction error. Top row are the most poorly reconstructed subjects in the CAESAR dataset, with a shape space trained on SizeUSA. Bottom row are the most poorly reconstructed SizeUSA subjects under a shape space trained on CAESAR subjects. A CAESAR shape space is biased towards sport bras and fails to capture the female chest shape in SizeUSA. SizeUSA includes more obese subjects that are poorly reconstructed under a CAESAR shape space.

inverse of geodesic distance to a joint. Secondly, the pose-corrective blendshapes for each joint are regressed from local pose information, corresponding to the joint and its direct neighbors in the kinematic tree; this is a richer representation than SMPL. These two factors together with the sparsity inducing $L1$ norm on the activation weights, act as an inductive bias to learn a sparse set of vertices that are geodesically local to a joint.

The sparse pose-correctives formulation reduces the number of parameters and regularizes the model, preventing it from learning spurious long range correlations from the training data. Since each vertex is only influenced by a limited number of joints in the kinematic tree, the gradients propagated through the model are sparse and the derivative of a vertex with respect to a geodesically distant joint is 0, which is not the case in SMPL.

3.5 Conclusion

We have introduced STAR, which has fewer parameters than SMPL yet is more accurate and generalizes better to unseen bodies when trained on the same data. Our key insight is that human pose deformation is local and sparse. While this observation is not new, our formulation is. We define a non-linear (ReLU) activation function for each joint and train the model from data to estimate both the linear corrective pose blendshapes and the activation region on the mesh that these joints influence. We kept what is popular with SMPL while improving on it in every sense. STAR has only 20% of the pose-corrective parameters of SMPL. Our training method and localized model fixes a key problem of SMPL– the spurious, long-range, correlations that result in non-local deformations. Such artifacts make SMPL unappealing for animators. Moreover, we show that, while SMPL is trained from thousands of scans, human bodies are more varied than the CAESAR dataset. More training scans results in a better model. Finally we make pose-corrective blendshapes depend on body shape, producing more realistic deformations. We make STAR available for research with 300 shape principal components. It can be swapped in for SMPL in any existing application since the pose and shape parameterization is the same to the user.

STAR presents an improved formulation over SMPL, but does not represent expressive faces and articulated hands, crucial for depicting emotions and gestures. Employing STAR's framework for developing head and hand models, similar to MANO and FLAME, would result in inheriting their limitations, notably in capturing the full range of motion of the head and hands. How, then, can we create models that retain STAR's strengths but accurately track the range of movements of the head and hands? This question guides our exploration in the following chapter.

Chapter 4

Federated Training

4.1 Introduction

Generative 3D models of the human body and its parts play an important role in understanding human behaviour. Over the past two decades, numerous 3D models of the body [101, 10, 62, 102, 103, 33, 38, 104, 105], face [73, 74, 69, 1, 75, 76, 77, 78] and hands [106, 107, 108, 2, 109, 110] have been proposed. Such models enabled a myriad of applications ranging from reconstructing bodies [17, 18, 19], faces [111, 112, 113], and hands [114, 115] from images and videos, modeling human interactions [20], generating 3D clothed humans [21, 22, 23, 24, 25, 26, 27], or generating humans in scenes [28, 29, 30]. They are also used as priors for fitting models to a wide range of sensory input measurements like motion capture markers [92, 6] or IMUs [116, 89, 90].

Hand [2, 117, 109, 15], head [69, 1, 15] and body [33, 38] models are typically built independently. Heads and hands are captured with a 3D scanner in which a subject remains static, while the face and hands are articulated. This data is unnatural as it does not capture how the body parts move together with the body. As a consequence,

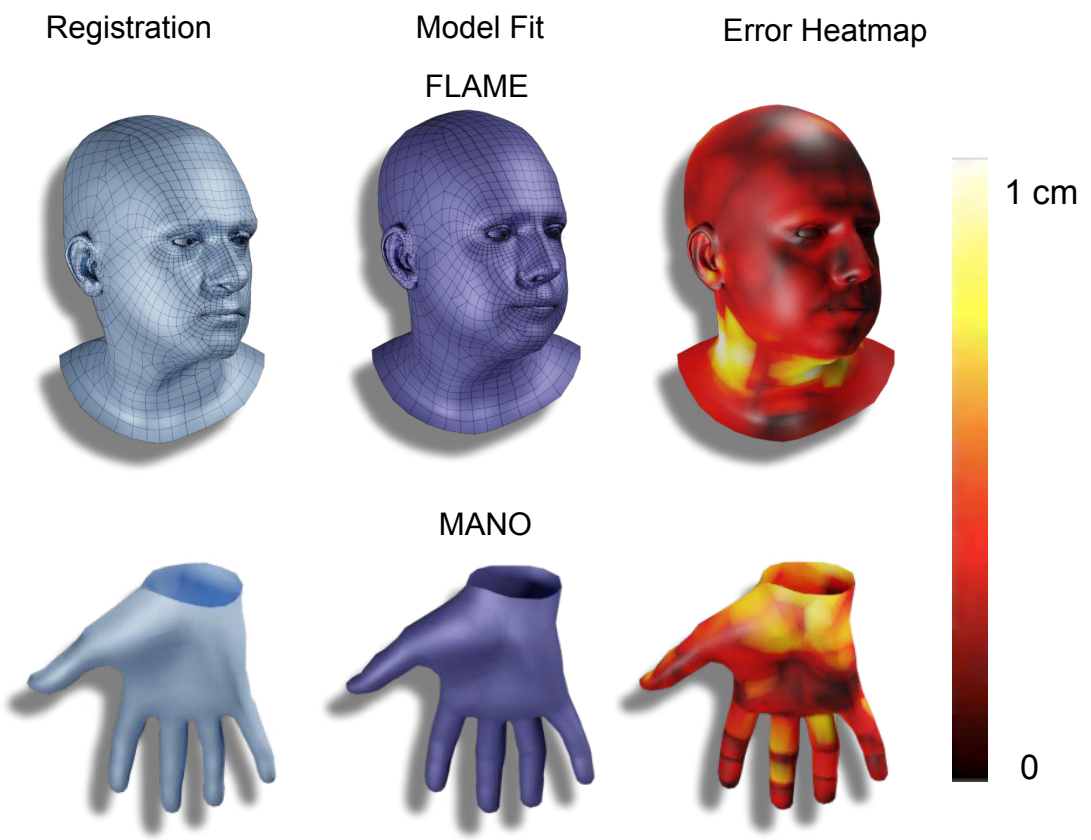


Figure 4.1: **Body Part Models Failure Cases:** Left: Existing body part models such as the FLAME [1] head model and the MANO [2] hand model fail to capture the corresponding body part's shape through the full range of motion. Fitting FLAME to a subject looking left results in significant error in the neck region. Similarly, fitting MANO to hands with a bent wrist, results in significant error at the wrist region.

the construction of head/hand models implicitly assumes a static body, and use simple kinematic trees that fail to model the head/hand full degrees of freedom. For example, in Fig. 4.1 we fit the FLAME head model [1] to a pose where the subject is looking to their left and find that FLAME exhibits a significant error in the neck region. Similarly, we fit the MANO [2] hand model to a hand pose where the the wrist is fully bent downwards. MANO fails to capture the wrist deformation that results from the bent wrist. This is a systematic limitation of existing head/hand models, which can not be addressed by simply training on more data.

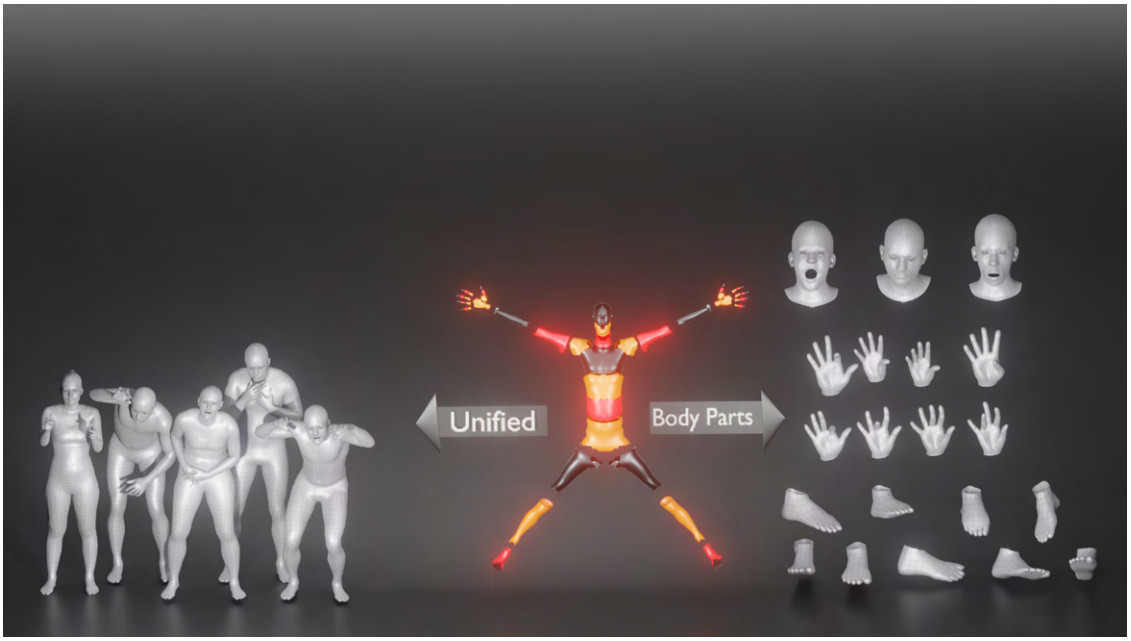


Figure 4.2: **Expressive part-based human body model.** SUPR is a factored representation of the human body that can be separated into a full suite of body part models.

In contrast to the existing approaches, we propose to jointly train the full human body and body part models together. We first train a new full-body model called SUPR , with articulated hands and an expressive head using a federated dataset of body, hand and head scans. This joint learning captures the full range of motion of the body parts along with the associated deformation. Then, given the learned deformations, we separate the body model into body part models. To enable separating SUPR into compact individual body parts we learn a sparse factorization of the pose-corrective blendshapes function as shown in Fig. 6.3. The factored representation of SUPR enables separating SUPR into an entire suite of models: SUPR-Head and SUPR-Hand. A body part model is separated by considering all the joints that influence the set of vertices defined by the body part template mesh. We show that the learned kinematic tree structure for the head/hand contains significantly more joints than commonly used by head/hand models. In contrast to the existing body part models that are learned in isolation of the body, our training algorithm unifies many disparate prior efforts and results in a suite of models that can capture the full range of motion of the head and hands.

The training data contains extreme body shapes such as anorexia patients and body-

builders. All subjects gave informed written consent for participation and the use of their data. Capture protocols were reviewed by the university of Tübingen ethics board.

We quantitatively compare SUPR and the individual body-part models to existing models including SMPL-X, GHUM, MANO, and FLAME. We find that SUPR is more expressive, is more accurate, and generalizes better. In summary, our main contributions are:

1. A unified framework for learning both expressive body models and a suite of high-fidelity body part models.

2. SUPR, a sparse expressive and compact body model that generalizes better than existing expressive human body models.

3. An entire suite of body part models for the head, hand, where the model kinematic tree and pose deformations are learned instead of being artist defined.

4. The Tensorflow and PyTorch implementations of all the models are publicly available for research purposes.

4.2 Federated Training Dataset

SUPR is trained on a federated dataset of 3D scans. In total 4 types of scanners are used: a full body scanner, a hand scanner, a head scanner and a foot scanner. All the scanners are 4D scanners, capturing high resolution dynamic sequences for each body part. We additionally leverage datasets that are either publicly available for research purposes or commercial datasets from private vendors. In this section we describe the scanning setup for each scanner and describe the external datasets.

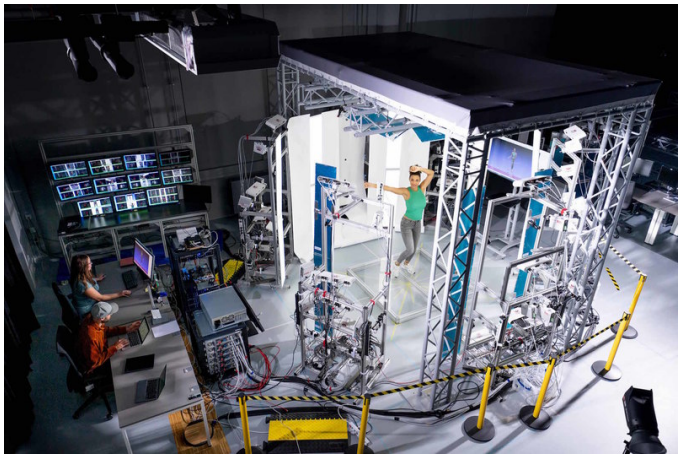


Figure 4.3: **Full Body Scanner** A 4D full body scanner. The system uses 22 pairs of stereo cameras, 22 color cameras, and speckle-light projectors. The speckle patterns allow accurate stereo reconstruction of 3D shape. This speckle pattern alternates at 120fps with large white-light LED panels that provide a smooth nearly uniform illumination. Each frame is a 3D mesh with approximately 150,000 points.

4.2.1 Full Body Scans

Human bodies deform in complex ways as a result of changes in body pose and body shape. To study and model minimally-clothed human body deformations, we use a 4D scanner (shown in Fig. 4.3) that captures the full 3D human body shape at 60 frames per second (fps). The full-body scanner is custom built by 3dMD (Atlanta, GA). The system uses 22 pairs of stereo cameras, 22 color cameras, and speckle-light projectors. The speckle patterns allow accurate stereo reconstruction of 3D shape. This speckle pattern alternates at 120 fps with large white-light LED panels that provide a smooth nearly uniform illumination. The scanner outputs high resolution meshes with approximately 150,000 vertices. The high resolution meshes in addition to the high frame rate (60 fps) enable us to model the subtle deformations of the human body. The full body scanner scanning volume is sufficient to capture poses such as a full leg split by a ballerina, or sitting and lying down poses. Example full-body training scans are shown in Fig. 4.4

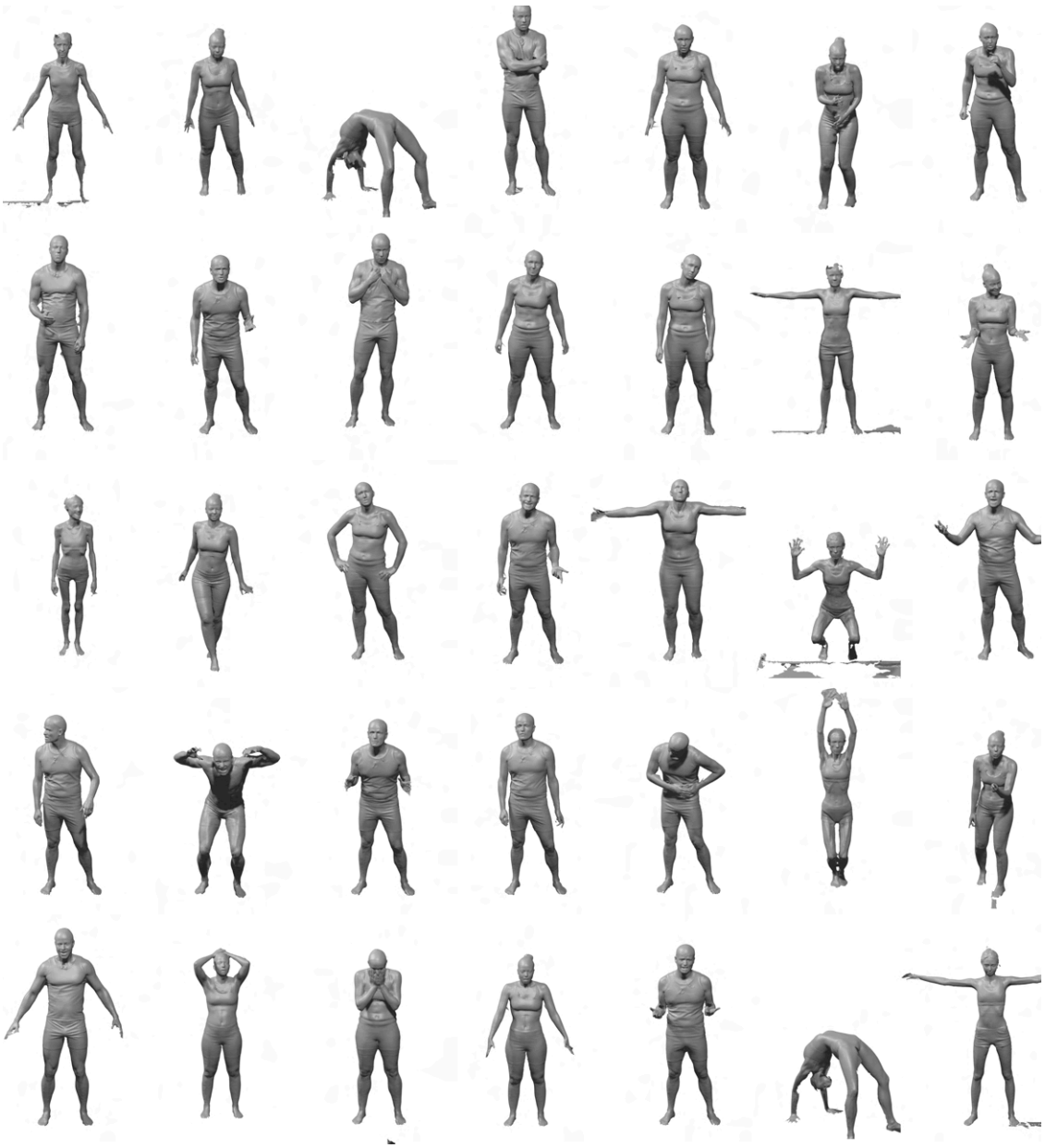


Figure 4.4: **Body Scans:** Example scans captured in the full body scanner. The scans are detailed and high-resolution. Note, however, the hands and the feet are poorly reconstructed, and the head resolution is not sufficient to capture subtle facial expressions.

Figure 4.5: **Head Scanner:** An overview of the head scanner. In contrast to the the full body scanner, the head scanner has a limited scanning volume which is focused on the subject head/neck region. The setup is sufficient for high-resolution capture of the human head including subtle deformation due to facial expression. However, the scanning setup is limited to capture the full range of motion of the head relative to the body.

4.2.2 Head Scans

The human head exhibits a range of highly dynamic deformations. When we refer to the head we mean the face, the back of the head including the scalp and the neck. The human head 3D deformations are due to facial expressions, jaw movement, head movement relative to the neck and body movement relative to the neck (for example when shrugging). We use a dedicated head scanner (shown in Fig. 4.5) to complement the full body 4D scanner. The head scanner has a significantly higher number of cameras focused on the head region compared to the body scanner in Section 4.2.1. The scanning setup enables us to capture the subtle facial expressions. We note, however, that the head scanner has a limited scanning volume making it infeasible to capture the full range of motion of the human head relative to the body.

Similar to the full body scanner, the head scanner is a 4D scanner capturing high-

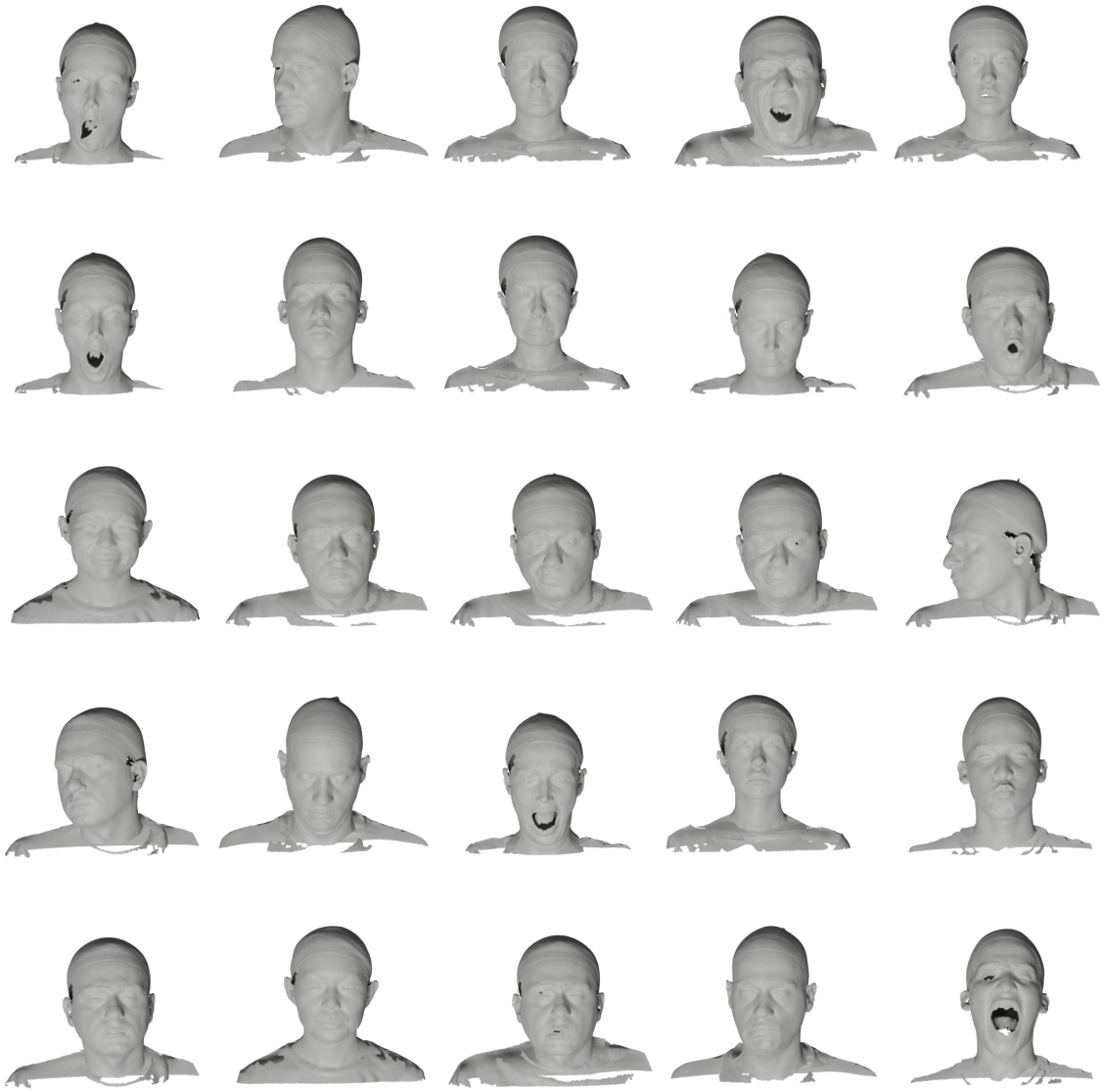


Figure 4.6: **Head Scans** A sample of the head scans used in training SUPR .

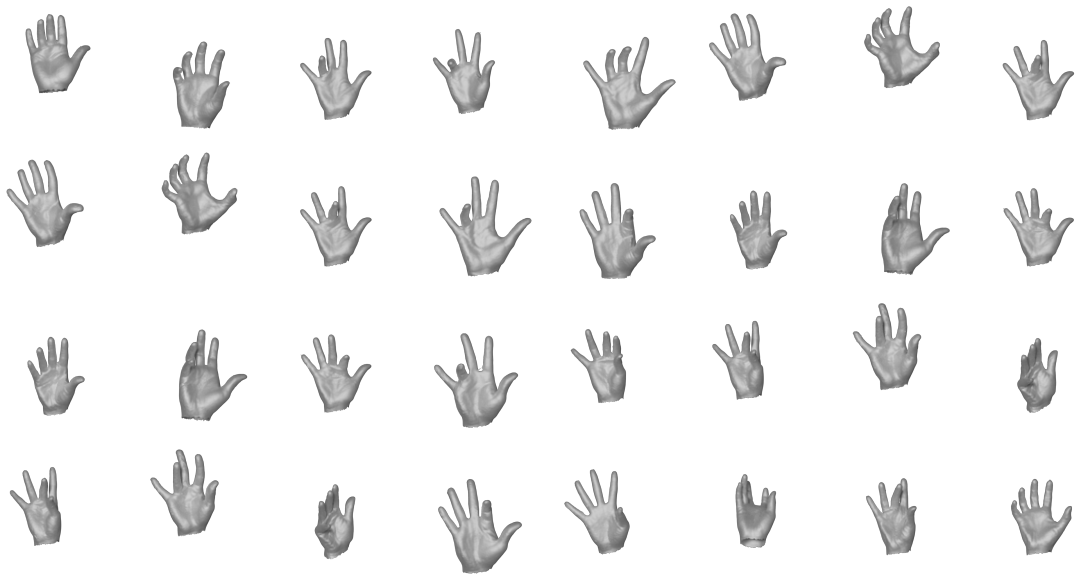


Figure 4.7: **Hand Scans:** A sample of the hand scans used to train SUPR .

resolution dynamic sequences. The scanner employs 6 pairs of stereo cameras to compute shape and geometry with the assistance of custom speckle projectors. It also includes 6 color cameras and white-light panels to capture texture. The data capturing protocol was designed by experts to capture subtle and extreme facial expressions, full movement of the jaw, in addition to neck movement poses such as looking up, down to the left or right. A sample of the head scans are shown in Fig. 4.6

4.2.3 Hand Scans

The reconstructed fingers in full-body scans are typically noisy and poorly reconstructed. To better capture the hands, we use the data from the MANO hand model [2]. These hand scans are used to learn the pose corrective blendshapes due to finger articulation. A sample of the captured hand scans is shown in Fig. 4.7.

4.3 Model

We describe the formulation of SUPR in Section 4.3.1, followed by how we separate SUPR into body parts models in Section 4.3.2.

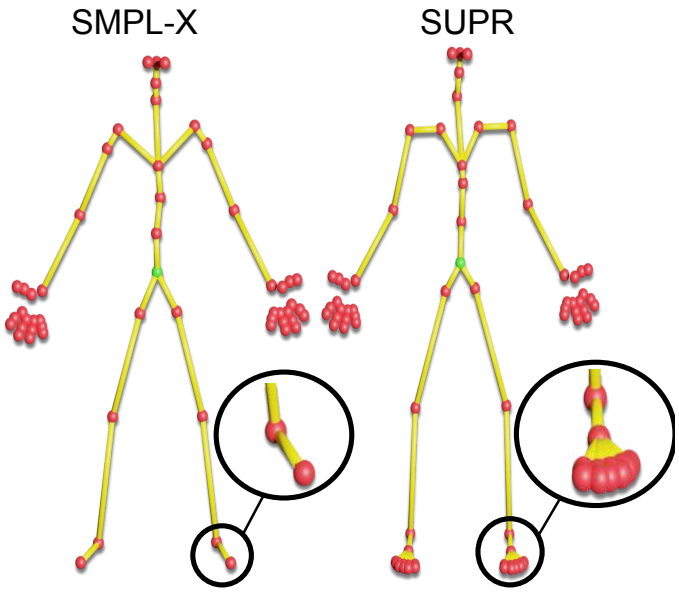


Figure 4.8: **SUPR Kinematic Tree**: The kinematic tree of SUPR . The green sphere is the model root joint, the red spheres are spherical joints.

4.3.1 SUPR

SUPR is a vertex-based 3D model with linear blend skinning (LBS) and learned blendshapes. The blendshapes are decomposed into 3 types: *Shape Blendshapes* to capture the subject identity, *Pose-Corrective Blendshapes* to correct for the widely-known LBS artifacts, and *Expression Blendshapes* to model facial expressions. The SUPR mesh topology and kinematic tree are based on the SMPL-X topology. The template mesh $\overline{T}^{N\times 3}$ contains $N = 10,475$ vertices and $K = 75$ joints. The SUPR kinematic tree is shown in Fig. 4.8. In contrast to existing body models, the SUPR kinematic tree contains sig-

nificantly more joints in the foot, ankle and toes as shown in Fig. 4.8. Following the notation of SMPL, SUPR is defined by a function $M(\vec{\theta},\vec{\beta},\vec{\psi})$, where $\vec{\theta} \in \mathbb{R}^{75\times 3}$ are the pose parameters corresponding to the individual bone rotations, $\vec{\beta} \in \mathbb{R}^{300}$ are the shape parameters corresponding to the subject identity, $\vec{\psi} \in \mathbb{R}^{100}$ are the expression parameters controlling facial expressions. Formally, SUPR is defined as

$$M(\vec{\theta},\vec{\beta},\vec{\psi}) = W(T_p(\vec{\theta},\vec{\beta},\vec{\psi}),J(\vec{\beta}),\vec{\theta};\mathcal{W}), \tag{4.1}$$

where the 3D body, $T_p(\vec{\theta},\vec{\beta},\vec{\psi})$, is transformed around the joints J by the linear-blend-skinning function $W(.)$, parameterized by the skinning weights $\mathcal{W} \in \mathbb{R}^{10475\times 75}$. The cumulative corrective blendshapes term are defined as

$$T_p(\vec{\theta},\vec{\beta},\vec{\psi}) = \overline{T} + B_S(\vec{\beta};\mathcal{S}) + B_P(\vec{\theta};\vec{K},A) + B_E(\vec{\psi};\mathcal{E}), \tag{4.2}$$

where $\overline{T} \in \mathbb{R}^{10475\times 3}$ is the template of the mean body shape, which is deformed by: $B_S(\vec{\beta};\mathcal{S})$, the shape blendshape function capturing a PCA space of body shapes; $B_P(\vec{\theta};\vec{K},A)$, the pose-corrective blendshapes based on STAR formulation introduced in Chapter 3, that address the LBS artifacts; and $B_E(\vec{\psi};\mathcal{E})$, a PCA space of facial expressions.

Sparse Pose Blendshapes

In order to separate SUPR into body parts, each joint should strictly influence a subset of the template vertices $\overline{T}$. To this end, we base the pose-corrective blendshapes $B_p(.)$ in Eq. (4.2) on the STAR model discussed in chapter 3. The pose-corrective blendshape function is factored into per-joint pose-corrective blendshape functions

$$B_P(\vec{q},\mathbf{K},\mathbf{A}) = \sum_{j=1}^{K-1} B_P^j(\vec{q}_{ne(j)};\mathbf{K}_j;A_j), \tag{4.3}$$

where the pose-corrective blendshapes are sum of $K-1$ sparse spatially-local pose-corrective blend-shape functions. Each joint-based corrective blendshape $B_P^j(.)$, predicts corrective offsets for a sparse set of the model vertices, defined by the learned joint activation weights $A_j \in \mathbb{R}^{10475}$. Each A_j is a sparse vector defining the sparse set of vertices influenced by the j^{th} joint blendshape $B_p^j(.)$. The joint corrective blendshape function is conditioned on the normalized unit quaternions $\vec{q}_{ne(j)}$ of the j^{th} joint's direct neighbouring joints' pose parameters. We note that the SUPR pose blend-shape formulation in Eq. (4.3) is not conditioned on body shape, unlike STAR, since the additional body-shape blendshape is not sparse and, hence, can not be factored into body parts. Since the skinning weights in Eq. (4.1) and the pose-corrective blend-shape formulation in Eq. (4.3) are sparse, each vertex in the model is related to a small subset of the model joints. This sparse formulation of the pose space is key to separating the model into compact body part models.

4.3.2 Body Part Models

In traditional body part models like FLAME and MANO, the kinematic tree is designed by an artist and the models are learned in isolation of the body. In contrast, here the pose-corrective blendshapes of the hand (SUPR-Hand) and head (SUPR-Head) are trained jointly with the body on a federated dataset. The kinematic tree of each part model is inferred from SUPR rather than being artist defined. To separate a body part, we first define the subset of mesh vertices of the body part $\overline{T}_{bp}$ from the SUPR template $\overline{T}_{bp} \in \overline{T}$. Since the learned SUPR skinning weights and pose-corrective blendshapes are strictly sparse, any subset of the model vertices $\overline{T}_{bp}$ is strictly influenced by a subset of the model joints. More formally, a joint $\vec{j}$ is deemed to influence a body part defined by

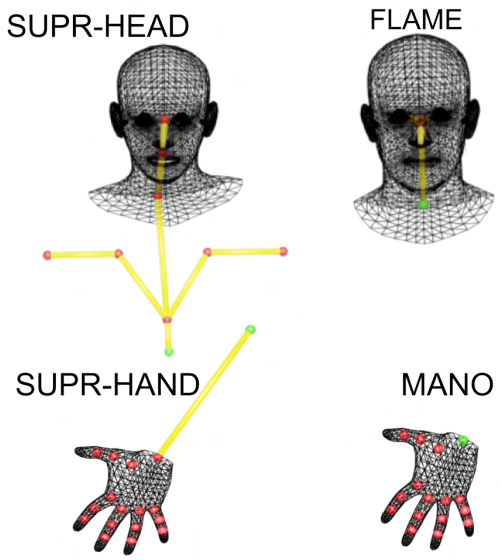


Figure 4.9: **Separated Body Part Models**: The kinematic tree of the separated body parts. The top row compares the kinematic tree of SUPR-Head and Flame. The bottom row compares the kinematic tree of SUPR-Hand and MANO. The green sphere is a model root joint, the red spheres are spherical joints. Note that the SUPR-Head and SUPR-Hand have substantially more joints compared to Flame and MANO.

the template $\overline{T}_{bp}$ if:

$$\mathbb{I}\left(T_{bp}, \vec{j}\right) = \begin{cases} 1 & \text{if } \sum \mathcal{W}\left(\overline{T}_{bp}, \vec{j}\right) \neq 0 \text{ or } \sum A_j\left(\overline{T}_{bp}\right) \neq 0 \\ 0 & \text{othewise,} \end{cases} \tag{4.4}$$

where $\mathbb{I}(.,.)$ is an indicator function, $\mathcal{W}\left(\overline{T}_{bp}, \vec{j}\right)$ is a subset of the SUPR learned skinning weights matrix, where the rows are defined by the vertices of $\overline{T}_{bp}$, the columns correspond to the j^{th} joint, $\vec{j}$, $A_j(\overline{T}_{bp})$ corresponds to the learned activation for the j^{th} joint and the rows defined by vertices $\overline{T}_{bp}$. The indicator function $\mathbb{I}$ returns 1 if a joint $\vec{j}$ has non-zero skinning weights or a non-zero activation for the vertices defined by $\overline{T}_{bp}$. Therefore the set of joints J_{bp} that influences the template $\overline{T}_{bp}$ is defined by:

$$J_{bp} = \left\{ \mathbb{I}(\overline{T}_{bp}, j) = 1 \ \ \forall \ \ j \in \{1, \ldots, K\} \right\}. \tag{4.5}$$

The kinematic tree defined for the body part models in Eq. (4.5) is implicitly defined by the learned skinning weights $\mathcal{W}$ and the per joint activation weights A_j. The resulting kinematic tree of the separated models is shown in Fig. 4.9. Surprisingly, the head is influenced by substantially more joints than in the artist-designed kinematic tree used in FLAME. Similarly, SUPR-Hand has an additional wrist joint compared to MANO. We note here that the additional joints in SUPR-Head and SUPR-Hand are outside the head/hand mesh. The additional joints for the head and the hand are beyond the scanning volume of a body part head/hand scanner. This means that it is not possible to learn the influence of the shoulder and spine joints on the neck from head scans alone.

The skinning weights for a separated body are defined by $\mathcal{W}_{bp} = \mathcal{W}\left(\overline{T}_{bp}, J_{bp}\right)$, where $\mathcal{W}\left(\overline{T}_{bp}, J_{bp}\right)$ is the subset of the SUPR skinning weights defined by the rows corresponding to the vertices of $\overline{T}_{bp}$ and the columns defined by J_{bp}. Similarly, the pose corrective blendshapes are defined by $B_{bp} = B_p\left(\overline{T}_{bp}, J_{bp}\right)$ where $B_p\left(\overline{T}_{bp}, J_{bp}\right)$ corresponds to a subset of SUPR pose blendshapes defined by the vertices of $\overline{T}_{bp}$ and the quaternion features for the set of joints J_{bp}. The skinning weights $\mathcal{W}_{bp}$ and blendshapes B_{bp} are based on the SUPR learned blendshapes and skinning weights, which are trained on a federated dataset that explores each body part's full range of motion relative to the body. We additionally train a joint regressor $\mathcal{J}_{bp}$, to regress the joints $\mathcal{J}_{\text{bp}} : \overline{T}_{bp} \rightarrow J_{bp}$. We learn a local body part shape space $B_S(\vec{\beta}_{bp}; \mathcal{S}_{bp})$, where $\mathcal{S}_{bp}$ represents the body part PCA shape components. For the head, we use the SUPR learned expression space $B_E(\psi; \mathcal{E})$.

4.4 Constrained SUPR

The SUPR kinematic tree introduced in Section 4.3 is based on spherical joints. Each spherical joint j is parameterized by $\vec{\theta}_j \in \mathbb{R}^3$. The spherical joints allow redundant de-

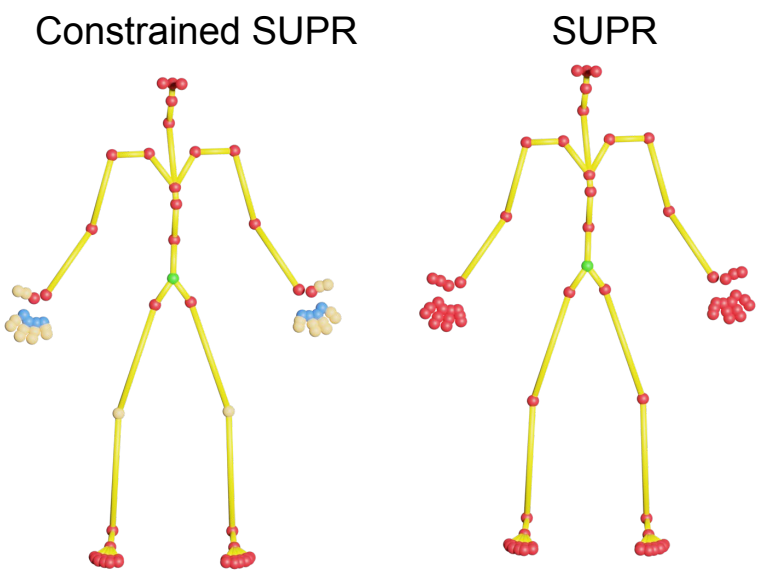


Figure 4.10: **Constrained SUPR Kinematic Tree:** SUPR is based on spherical joints which allow redundant degrees of freedom for body parts such as the fingers. The constrained SUPR kinematic tree contains a mixture of joints: Spherical joints (shown in red), Hinge Joints (shown in beige) and double hinge joints (shown in blue).

grees of freedom for some body parts such as the fingers. For the fingers, for example, the axes of rotation are not bone-aligned. In order to simply bend a finger we have to control 3 axis-angle rotations. This is problematic to use by animators and for architectures that regress hand pose parameters from images. In this section we describe a constrained version of SUPR that uses hinge/double hinge joints in contrast to spherical joints.

4.4.1 Constrained Kinematic Tree Formulation

The kinematic tree of the constrained version of SUPR (shown in Fig. 4.10) uses hinge and double hinge joints. A hinge joint is fully parameterized by an axis of rotation $\vec{a} \in \mathbb{R}^3$ and a pose parameter $\vec{\theta} \in \mathbb{R}$. A double hinge joint is defined by two axes of rotation and pose parameters $\vec{\theta} \in \mathbb{R}^2$. The axes of rotation for the hinge and double hinge joints

are orthogonal to the bone. Therefore, to simply bend a finger in SUPR requires only controlling or regressing one or two scalars. This compact representation is convenient for artists and regression tasks and is more anatomically plausible.

Specifically, this version of SUPR is defined by Eq. (4.6):

$$M(\vec{\theta},\vec{\beta},\vec{\psi}) = W(T_p(\vec{\theta},\vec{\beta},\vec{\psi}),J(\vec{\beta}),AX,\vec{\theta},\mathcal{W}), \tag{4.6}$$

where $AX \in \mathbb{R}^{30\times 3}$ is the axis matrix for the hinge and double hinge joints. The key difference between Eq. (4.1) and Equation (4.6) is the bone transformation rotation matrix. The rotation matrix for a hinge joint is a constrained rotation matrix, which only allows a single degree of freedom with respect to the rotation axis $\vec{a}$. A constrained rotation matrix is defined by:

$$\begin{bmatrix} a_x^2+c_\theta(1-a_x^2) & a_xa_y(1-c_\theta)+a_zs_\theta & a_xa_z(1-c_\theta)-a_ys_\theta \\ a_xa_y(1-c_\theta)-a_zs_\theta & a_y^2+c_\theta(1-a_y^2) & a_ya_z(1-c_\theta)+a_zs_\theta \\ a_xa_z(1-c_\theta)+a_ys_\theta & a_ya_z(1-c_\theta)-a_xs_\theta & a_z^2+c_\theta(1-a_z^2) \end{bmatrix}$$

where a_x, a_y, a_z are the x, y and z coordinates of the axis of rotation $\vec{a}$. c_θ and s_θ are $\cos(\theta)$ and $\sin(\theta)$ correspondingly.

The constrained version of SUPR only limits the bones' degrees of freedom, by constraining the rotation matrices of the corresponding joints.

4.5 Federated Training

The fundamental insight enabling the segmentation of SUPR into high fidielty body part models lies in the STAR formulation of the pose-corrective blendshape. This approach ensures that a singular joint impacts a select group of model vertices, facilitating the

smooth division of the model into distinct body parts. Each part is influenced by a specific subset of the SUPR kinematic tree. Furthermore, training with a federated dataset guarantees that we learn from body part scans that capture the subtle deformations of individual body parts. This is in addition to the information obtained from full-body scans, which captures how each body's parts deform in relation to the entire body. In this section, we provide a detailed explanation of the federated training process for the SUPR pose space.

We train SUPR model parameters to minimize reconstruction error on a federated dataset of hand, head and body scans. The full body dataset meshes are based on the SUPR template topology. For the head dataset, the meshes is based on a topology that is a subset of the SUPR full body mesh topology. Likewise, the hand dataset meshes are based on a topology that is also a subset of the complete SUPR mesh topology. In this section, we simplify our terminology by referring to datasets and models associated with the right hand and left hand collectively as hands. All meshes have been aligned to high-resolution 3D scans [118]. We refer to aligned meshes as registrations.

We train SUPR by minimizing the standard vertex-to-vertex loss between SUPR and the federated dataset of 3D registrations, because there exists a vertex-to-vertex correspondence between all the dataset and the SUPR mesh topology. Our goal is to train the SUPR parameters $\Phi = \{\mathcal{W}, \mathcal{J}, \mathbf{K}, \mathbf{A}\}$ by minimizing the reconstruction error on the federated datasets. We first train $\{\mathcal{J}, \mathcal{W}, \mathbf{K}, \mathbf{A}\}$ using our multi-pose federated dataset.

We refer to head registration as VH_j^i corresponding to the jth registration for the ith subject in the head dataset and the corresponding subset of SUPR (SUPR-Head) is M_{Head}. Similarly, we refer to the hand registrations VA_j^i corresponding to the jth registration for the ith subject in the hands datasets and the corresponding hand part of SUPR (SUPR-Hand) is defined as M_{Hand}. For the body dataset, we refer to a body registration

as V_j^i corresponding to the jth registration for the ith subject in the body registration and the corresponding full SUPR model as M. During training, we estimate three type of parameters: registration specific parameters, subject specific parameters and global model parameters. The registration specific parameters are the SUPR pose parameters $\vec{\theta}_j^i$ corresponding to the SUPR pose parameters for the ith subject jth registration. The subject-specific parameters are the personalized template and personalized joints T^i and J^i corresponding to the jth subject subject specific template and subject specific joints. We train SUPR by iteratively alternating between estimating registration specific parameters, subject specific parameters and the model global parameters.

Estimating Pose Parameters For all training registrations, we first estimate the pose parameters $\vec{\theta}$ for each registration in the training dataset. We minimize an objective function consisting of a data term, E_D which penalizes the squared Euclidean distance between the registration vertices and the corresponding model vertices more formally.

$$E_D = \sum_{j=1}^{N_{Head}} ||\mathbf{VH}_j^i - M_{\text{Head}}(\vec{\theta}_j^i, T^i, J^i)||^2 + \sum_{j=1}^{N_{Hand}} ||\mathbf{VA}_j^i - M_{\text{Hand}}(\vec{\theta}_j^i, T^i, J^i)||^2 + \sum_{j=1}^{N_{Body}} ||\mathbf{V}_j^i - M(\vec{\theta}_j^i, T^i, J^i)||^2$$

where the data term is a federated euclidean loss between the registrations and the corresponding SUPR part of the model. N_{head} is the number of head registrations, N_{hand} is the number of hand registrations and N_{body} is number of full body registrations. The data term is minimized with respect to the SUPR parameters $\vec{\theta}_j^i$.

Estimating Template and Joints For each subject i we further estimate a subject-specific template and a subject-specific joints. To this end we further minimize the data term in Eq. 4.5, with respect to each subject template T^i and joints J^i. To make the estimation well behaved, we define a regularization term by making several assumptions.

A symmetry regularization term, E_Y, penalizes the left-right asymmetry

$$E_Y = \sum_{i=1}^{P_{Head}} \lambda_H \left\| J^i - U\left(J^i\right) \right\|^2 + \left\| T^i - U\left(T^i\right) \right\|^2 + \sum_{i=1}^{P_{Body}} \lambda_B \left\| J^i - U\left(J^i\right) \right\|^2 + \left\| T^i - U\left(T^i\right) \right\|^2$$

where $\lambda_H = 10$ and $\lambda_B = 4$, P_{Head} is the total number of subjects in the head dataset, P_Body is the total number of subjects in the full body dataset and where $U(T)$ finds a mirror image of vertices T, by flipping across the sagittal plane and swapping symmetric vertices. This term encourages symmetric template meshes and, more importantly, symmetric joint locations. We note here that the symmetry regularization term is only used for the head and body, and not for the hands. The final objective of estimating the template and joints is defined by:

$$E_T = E_D + E_Y \tag{4.7}$$

where we minimize the objective with respect to the subjects template T^i and joints J^i.

Skinning Weights The skinning weights, $\mathcal{W}$, are further refined by minimizing the federated data term in Eq. 4.5. In addition to the data term, we use a regularization term defined by:

$$E_W = \lambda_p ||\mathcal{W} - \mathcal{W}_{\text{prior}}||_2 + \lambda_s ||\mathcal{W}||_1, \tag{4.8}$$

where $\mathcal{W}_{\text{prior}}$ is an artist prior. The regularization term regularizes the skinning weights towards an artist defined prior in addition to L1 sparsity inducing loss. The sparsity of the skinning weights is crucial such that each joints only influences a sparse set of the model vertices. The complete objective to train the skinning weights is defined by:

$$E = E_D + E_W \tag{4.9}$$

where we the optimization free variable is the model skinning weights $\mathcal{W}$.

Pose Corrective Blendshapes Finally, the pose-corrective blendshapes are based on STAR introduced in chapter 3. We further minimize the federated euclidean loss between the model and the corresponding registrations. In addition to the data term, we use a reguarlization with the activation function:

$$E_A = \lambda_c || \sum_{i=1}^{K-1} \phi_j(\vec{A}_j)||_1 \tag{4.10}$$

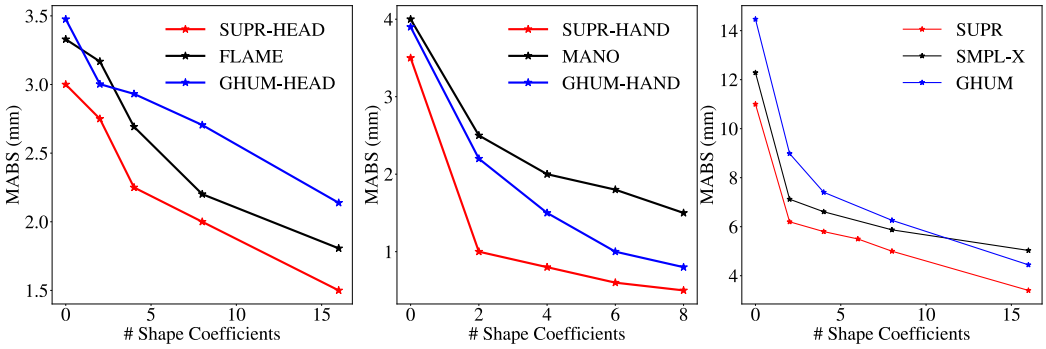
where λ_c is $1e-6$, K is the total number of joints, $\phi(.)$ is the ReLU function, A are the per-joint activation. The full objective function is defined by:

$$E = E_D + E_A \tag{4.11}$$

where the optimization free variables are the pose corrective blendshapes activation A and the model pose-corrective blendshapes $\mathbf{K}$.

4.6 Experiments

Our goal is to evaluate the generalization of SUPR and the separated head and hand models to unseen test subjects. We first evaluate the full SUPR body model against existing state of the art expressive human body models SMPL-X and GHUM (Section 4.6.1), then we evaluate the separated SUPR-Head model against existing head models FLAME and GHUM-Head (Section 4.6.3), and compare the hand model to GHUM-Hand and MANO



(a) Head Evaluation (b) Hand Evaluation (c) Body Evaluation

Figure 4.11: **Quantitative Evaluation:** Evaluating the generalization of SUPR and the separated head and hand models from SUPR against: GHUM-HEAD and FLAME for the head (Fig. 4.11a), GHUM-HAND and MANO (Fig. 4.11b) and GHUM (Fig. 4.11c). We report the *vertex-to-vertex* error (*mm*) as a function of the number of the shape coefficients used when fitting each model to the test set.

(Section 4.6.2).

4.6.1 Full-Body Evaluation

We use the publicly available 3DBodyTex dataset [100], which includes 100 male and 100 female subjects. We register the GHUM template and the SMPL-X template to all the scans; note SMPL-X and SUPR share the same mesh topology. We visually inspected all registered meshes for quality control. Given registered meshes, we fit each model by minimizing the vertex-to-vertex loss (*v2v*) between the model surface and the corresponding registration. The free optimization parameters for all models are the pose parameters $\vec{\theta}$ and the shape parameters $\vec{\beta}$. Note that, for fair comparison with GHUM, we only report errors for up to 16 shape components for all models since this is the maximum in the GHUM release. SUPR includes 300 shape components and using all of those would reduce the errors significantly. We follow the 3DBodyTex evaluation protocol and exclude the face and the hands when reporting the mean absolute error

(a) 3DBodyTex scans.

(b) Model Fits

Figure 4.12: **Body Qualitative Evaluation:** We evaluate SUPR on the 3DBodyTex dataset in Fig. 4.12a against GHUM, SMPL-X and SUPR using 16 shape components. The corresponding model fits are shown in Fig. 4.12b

(*mabs*). We report the mean absolute error of each model on both male and female registrations. For the GHUM model, we use the PCA-based shape and expression space. We report the model generalization error in Fig. 4.11c and show a qualitative sample of the model fits in Fig. 4.12b. SUPR uniformly exhibits a lower error than SMPL-X and

GHUM.

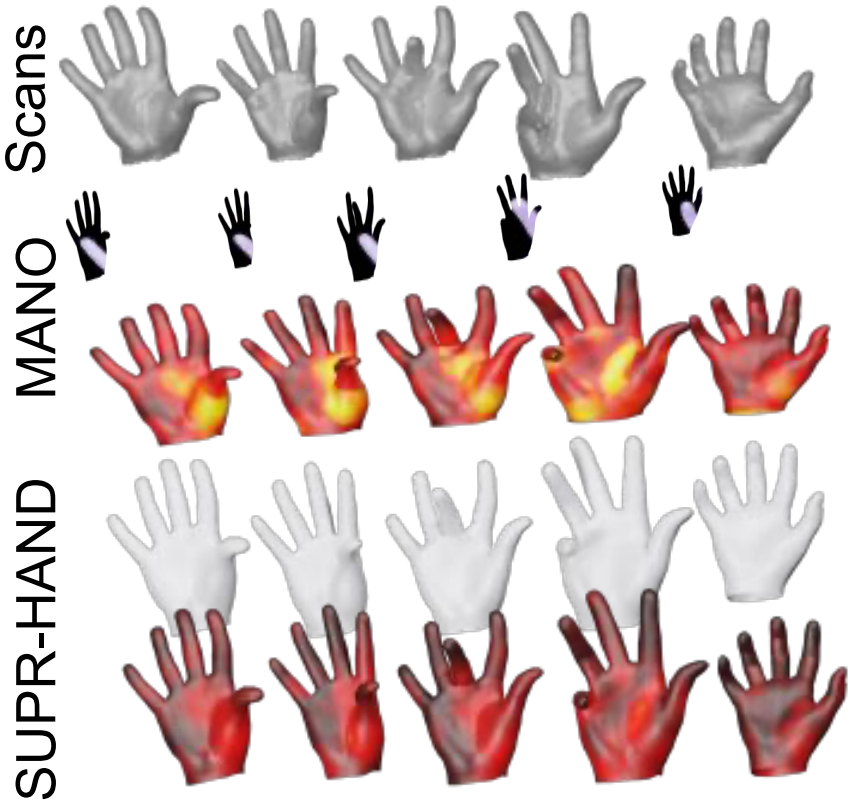


Figure 4.13: **Hand Qualitative Evaluation:** Evaluation of SUPR-Hand against MANO using 8 shape components.

4.6.2 Hand Evaluation

We use the publicly available MANO test set [2]. Since both SUPR-Hand and MANO share the same topology, we used the MANO test registrations provided by the authors to evaluate both models. To evaluate GHUM-Hand, we register the model to the MANO test set. We fit all models to the corresponding registrations using a standard *v2v* loss. For GHUM-Hand, we fit the model only to the selected hand vertices. The optimization free variables are the model pose and shape parameters. Fig. 4.11b shows generalization as a function of the number of shape parameters, where SUPR-Hand uniformly exhibits a

lower error compared to both MANO and GHUM-Hand. A sample qualitative evaluation of MANO and SUPR-Hand is shown in Fig. 4.13. In addition to a lower overall fitting error, SUPR-Hand has a lower error around the wrist region than MANO.

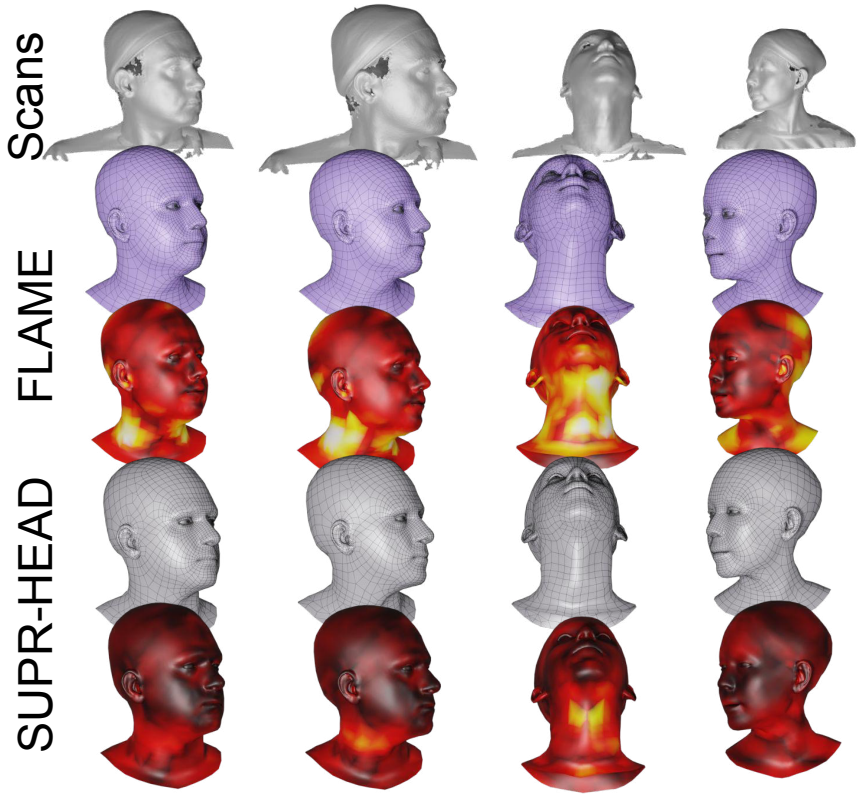


Figure 4.14: **Head Qualitative Evaluation:** We evaluate SUPR-Head against FLAME using 16 shape components

4.6.3 Head Evaluation

The head evaluation test set contains a total of 3 male and 3 female subjects, with sequences containing extreme facial expression, jaw movement and neck movement. As for the full body, we register the GHUM-Head model and the FLAME template to the test scans, and use these registered meshes for evaluation. For the GHUM-Head model, we use the linear PCA expression and shape space. We evaluate all models using a

standard *v2v* objective, where the optimization free variables are the model pose, shape parameters, and expression parameters. We use 16 expression parameters when fitting all models. For GHUM-Head we exclude the internal head geometry (corresponding to a tongue-like structure) when reporting the *v2v* error. Fig. 4.11a shows the model generalization as a function of the number of shape components. We show a sample of the model fits in Fig. 4.14. FLAME fails to capture head-to-neck rotations plausibly, despite each featuring a full head mesh including a neck. This is clearly highlighted by the systematic error around the neck region in Fig. 4.14. In contrast, SUPR-Head captures the head deformations and the neck deformations plausibly and uniformly generalizes better.

Model Name	Sparse Pose Deformations	Federated Training Data	Articulated Hands	Expressive Head	Part Based	Game Engine Compatiability	Publicly Available
SCAPE	✗	✗	✗	✗	✗	✗	✔
Stitched Puppet	✔	✗	✗	✗	✔	✗	✔
SMPL	✗	✗	✗	✗	✗	✔	✔
SMPL-H	✗	✔	✔	✗	✗	✔	✔
Frank	-	✗	✔	✔	✔	✔	✔
SMPL-X	✗	✔	✔	✔	✗	✔	✔
GHUM	✗	✔	✔	✔	✗	✗	✔
STAR	✔	✗	✗	✗	✗	✔	✔
BLSM	✗	✗	✗	✗	✗	✔	✗
SUPR	✔	✔	✔	✔	✔	✔	✔

Figure 4.15: A comparison between SUPR and existing body models.

4.7 Model Comparison

SUPR is trained on a federated dataset of head, body and head registrations. As a consequence of the sparse factorization of the pose space, we are able to separate the model into body part models. A comparison between SUPR and existing body models is shown in Fig 4.15.

Model	# Pose	# Joints	# Blendshapes
SUPR	225	75	296
SMPL-X [16]	165	55	486
GHUM [15]	124	63	-

Table 4.1: **Body Models Comparison:** Comparing existing expressive human body models according to the number of pose parameters, number of joints and number of pose corrective blendshapes.

4.7.1 SUPR

SUPR is a compact model that is compatible with the existing gaming and animation industry standards. The number of parameters of SUPR compared to existing expressive human body models is summarised in Table 4.1.

Comparison with SMPL-X: SUPR has 30% fewer pose-corrective blendshapes, despite having significantly more joints compared to SMPL-X. This is because of the Quaternion-based representation, which is significantly more compact compared to the Rodrigues representation used by SMPL-X. However, despite SUPR 's compactness, it uniformly generalizes better than SMPL-X. The shape space of SMPL-X is trained on the CAESAR dataset [37], while SUPR is trained on $15,000$ registrations from both CAESAR and SizeUSA [54]. The pose space of SMPL-X is trained on 2000 full body registrations. In contrast, SUPR 's pose space is trained on a federated dataset of 1.2 million registrations of head, hand and body registrations.

SMPL-X's pose blendshape formulation is based on SMPL. As a result, SMPL-X suffers from the same drawbacks of SMPL, namely SMPL-X also learns false long range spurious correlations; e.g. bending one elbow results in a bulge in the other elbow.

Comparison with GHUM: The GHUM model [15] pose-space deformation function (PSD) is modeled by a neural network, which is not compatible with the gaming and animation industry standards. SUPR 's learned blendshapes are linearly related to the model pose parameters, and hence the formulation is fully compatible with the gaming and animation industry standards. While both SUPR and GHUM are trained on a federated dataset, and the GHUM authors propose a separated suite of models (GHUM-Head and GHUM-Hand), there are key important differences. The GHUM shape space is trained only on the CAESAR data (5K subjects), while the SUPR shape space is trained on both CAESAR and SizeUSA, for a combined total of 15K registrations. On the other hand, the pose space of GHUM is trained on a dataset of 60K head, hand and body registrations, while the SUPR pose space is trained on 1.2 million body, head and hand registrations.

The GHUM PSD formulation is a dense non-linear formulation, where all the joints are related to all the vertices using a VAE [119]. As a result the body pose-space formulation of GHUM can not be separated into compact body parts. To define separate body part models, the GHUM authors segment the mesh and re-train the PSD function of the separated parts. The proposed head and hand models for GHUM fail to capture the full degrees of freedom of the head. SUPR and the separated head/hand models are jointly trained once. In contrast to GHUM, the SUPR pose-space formulation is strictly sparse, where each joint only influences a sparse set of the model vertices. As a result, SUPR can be separated into a suite of compact models. The learned kinematic tree of SUPR-Head has significantly more joints (neck and shoulders).

4.7.2 SUPR-Head

The SUPR-Head has a pose, shape and expression space. We train 3 head models: female, male and a gender neutral model. The pose blendshape function is a subset of the

learned SUPR pose corrective blendshapes, which are also sparse and spatially local. A comparison between and existing full head models is shown in Table 4.2.

Model	# Pose	# Joints	# Blendshapes
SUPR-Head	29	10	40
FLAME [16]	12	4	36
GHUM-Head [15]	23	10	-

Table 4.2: **Head Models Comparison:** Comparing existing head models models according to the number of pose parameters, number of joints and number of pose corrective blendshapes.

4.7.3 SUPR-Hand

We train a single gender-neutral SUPR-Hand model. SUPR-Hand has a pose and shape space. A comparison between SUPR-Hand and existing hand models is shown in Table 4.3. In comparison to MANO, SUPR-Hand has an additional wrist joint, which is necessary to model the hand deformations as a result of the wrist movement. A comparison between SUPR-Hand and existing hand models is shown in Table 4.3.

Model	# Pose	# Joints	# Blendshapes
SUPR-Hand	102	32	120
MANO [16]	90	30	270
GHUM-Hand [15]	18	36	-

Table 4.3: **Hand Models Comparison:** Comparing existing hand models according to the number of pose parameters, number of joints and number of pose corrective blendshapes.

4.8 Conclusion

We present a novel training algorithm for jointly learning high-fidelity expressive full-body and body parts models. We highlight a critical drawback in existing body part models such as FLAME and MANO, which fail to model the full range of motion of the head/hand. We identify that the issue stems from the current practice in which body parts are modeled with a simplified kinematic tree in isolation from the body. Alternatively, we propose a holistic approach where the body and body parts are jointly trained on a federated dataset that contains the body parts' full range of motion relative to the body. We train SUPR with a federated dataset of 1.2 million scans of the body, hands, and head. The sparse formulation of SUPR enables separating the model into an entire suite of body-part models. Surprisingly, we show that the head and hand models are influenced by significantly more joints than commonly used in existing models. We thoroughly compare SUPR and the separated models against SMPL-X, GHUM, MANO and FLAME and show that the models uniformly generalize better and have a significantly lower error when fitting test data. The pose-corrective blendshapes of SUPR and the separated body part models are linearly related to the kinematic tree pose parameters, therefore our new formulation is fully compatible with the existing animation and gaming industry standards. A Tensorflow and PyTorch implementation of SUPR and the separated head (SUPR-Head) and hand (SUPR-Hand) is publicly available for research purposes. SUPR is compatible with the gaming and animation industry standards.

This chapter presents a comprehensive suite of models dedicated to the body, head, and hands. Notably, SUPR advances the modeling of the foot by incorporating a greater number of joints, thereby capturing its entire range of motion. However, the challenge of training SUPR for the foot lies in the poor foot reconstructions obtained from body scans. Moreover, since SUPR's deformation are derived from STAR, it is limited to modeling

deformations related to body pose, which, although sufficient for modeling general body deformations, fall short in accurately modelling the deformations of the foot caused by ground contact. The quest for enhancing the foot model's fidelity requires exploring novel data sources and developing formulations that specifically address contact-based foot deformations. What strategies might we employ to further enhance modeling the foot? Insights into this question are offered in the next chapter.

Chapter 5

Human Foot Model

5.1 Introduction

The human foot is a complex structure containing muscles, soft tissue and a quarter of the bones in the skeleton [120, 121]. The evolution of the foot over a period exceeding 1 million years was crucial for enabling locomotion activities associated with an upright posture, including walking, running, and jumping. [122]. In all existing human body models [10, 38, 33, 16, 15, 14] the foot kinematic tree is modeled with significantly fewer joints than in the human foot as shown in Fig. 5.1. In comparison to the human foot the existing kinematic tree can not model the full range of motion of the human foot bones, such as toe articulation. The movement of the toes is critical for human locomotion and balancing. The simplistic modeling of the human foot in existing models limits their application in Biomechanicas and Physics-based modeling of human locomotion. Displicines such as Biomechanicas research require a more faithful modeling of the human foot.

The human foot deformations are distinctly different from the rest of the human body.

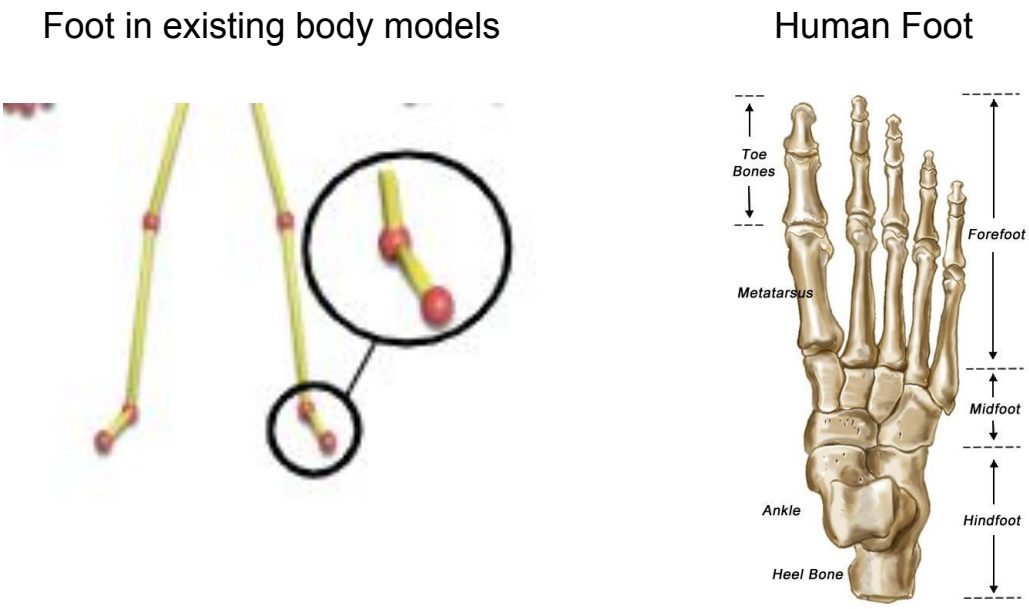


Figure 5.1: **Human Foot Kinematic Tree:** The human foot is a complex structure containing joints, bones, muscles and soft tissue. Each human foot contains more than 30 joints, 26 bones and more than 100 muscles (as shown on the right), however existing body models such as SMPL and SMPL-X use only two joints for the foot (as shown on the left).

This is because, the human foot is in frequent contact with the ground as we walk and move. The deformations due to contact are correlated with foot shape (over weight subjects have more soft-tissue on the foot, which deforms with scene contact), foot pose and scene contact. All existing body deformations are related to the bone rotations/body pose (such as SMPL [33] and SMPL-X [16]), or body pose, shape (such as in STAR [38]). The existing formulations completely ignore the body interaction with the scene. This simplified approach is typically satisfactory for the remaining regions of the body, but it is suboptimal for the human foot, which maintains nearly continuous contact with the surrounding environment.

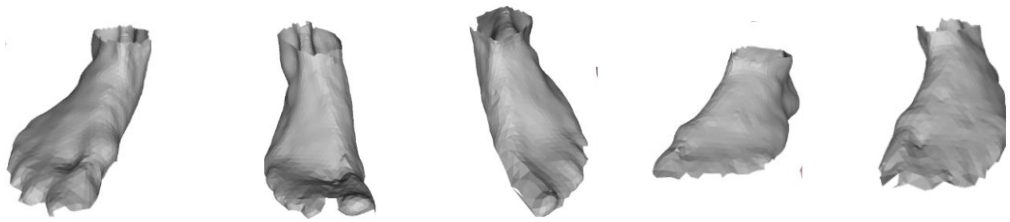
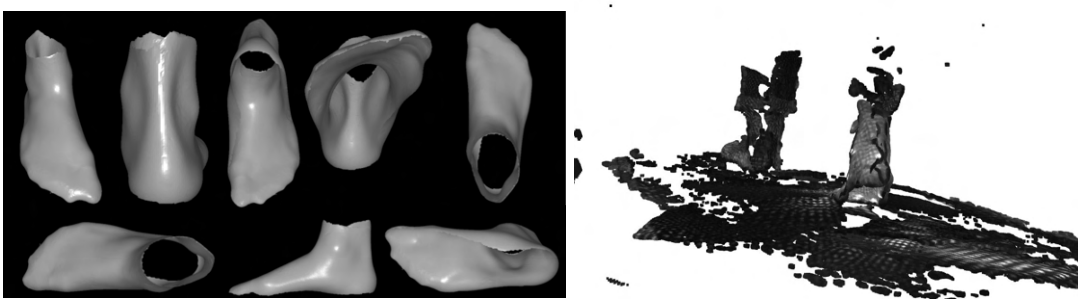


Figure 5.2: **Foot in Full Body Scans:** Human foot is typically poorly reconstructed in a full body scanner. The foot region is low resolution compared to the rest of the body, due to the limited number of cameras focused on the foot. The individual toes are often merged in the scans and the scans are often corrupted by noise and missing toes. The foot sole is not captured, since it is invisible to the cameras.



(a) Foot model. (b) A Scan from the DynaMo system.

Figure 5.3: **DynaMo System:** Bopanna et al. [3] foot model based on Principal Component Analysis shown in Fig. 5.3a. The model is trained on dynamic foot scans captured by the DynaMo system in Fig. 5.3b. The scans and the model do not contain toes or a foot sole.

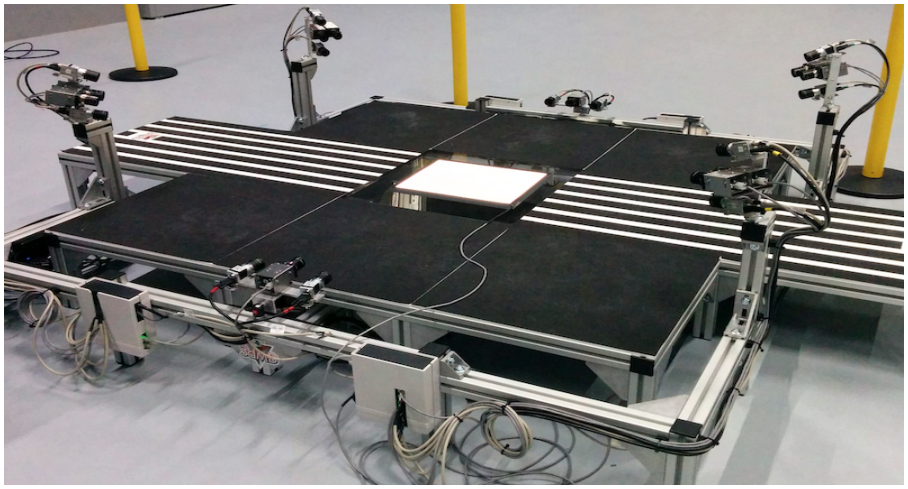
5.2 Problem Statement

There are two primary challenges that must be addressed when modeling the human foot. First, there is the issue of under articulation, which affects the ability to accurately represent the intricate movements of the foot. Second, accurately modeling the contact deformations between the human foot and the surrounding scene presents another significant challenge. A major obstacle in achieving more precise human body models is the lack of available data. In human body scans, the foot is often poorly reconstructed compared to the rest of the body, as depicted in Fig. 5.2. Capturing the human foot in motion is particularly challenging with current scanning solutions, as shown in Fig. 5.3, where the scans are noisy and the foot sole is poorly reconstructed. Previous setups for capturing feet only allow for static poses, thereby limiting the representation of foot shape variations. In full body scans, the foot sole is typically occluded and inadequately reconstructed due to the limited number of cameras focused on the foot.

5.3 Foot Scanner

To enable capturing the full range of the human foot deformations, we use a custom built scanner designed for the foot. The scanner is designed to be mechanically stable to capture dynamic poses such as walking, running or jumping. The output scans are high resolution and can capture the movement of the toes. The scanner floor is a transparent glass platform (which can support subjects up to 150 kg), which enables us to capture the foot sole deformation due to ground contact.

An overview of the foot scanner is shown in Figure 5.4. The scanner setup features a runway for the subjects to run or walk. In Figure 5.4b, we show raw scanner images, where the foot is visible from all views, including the foot sole. The scanner uses 10



(a) Scanner setup

(b) Raw scanner images

Figure 5.4: **Overview of the Foot Scanner:** A 3dMD foot scanner using 10 pairs of stereo cameras (Fig. 5.4a), including dedicated cameras capturing the bottom of the foot through a transparent glass platform(Fig. 5.4b). The scanner features a runway to capture dynamic sequences such as walking.

pairs of stereo cameras, including dedicated cameras capturing the bottom of the foot. The frame rate of the scanner is 10 fps. The output scans contain on average 30,000 points.

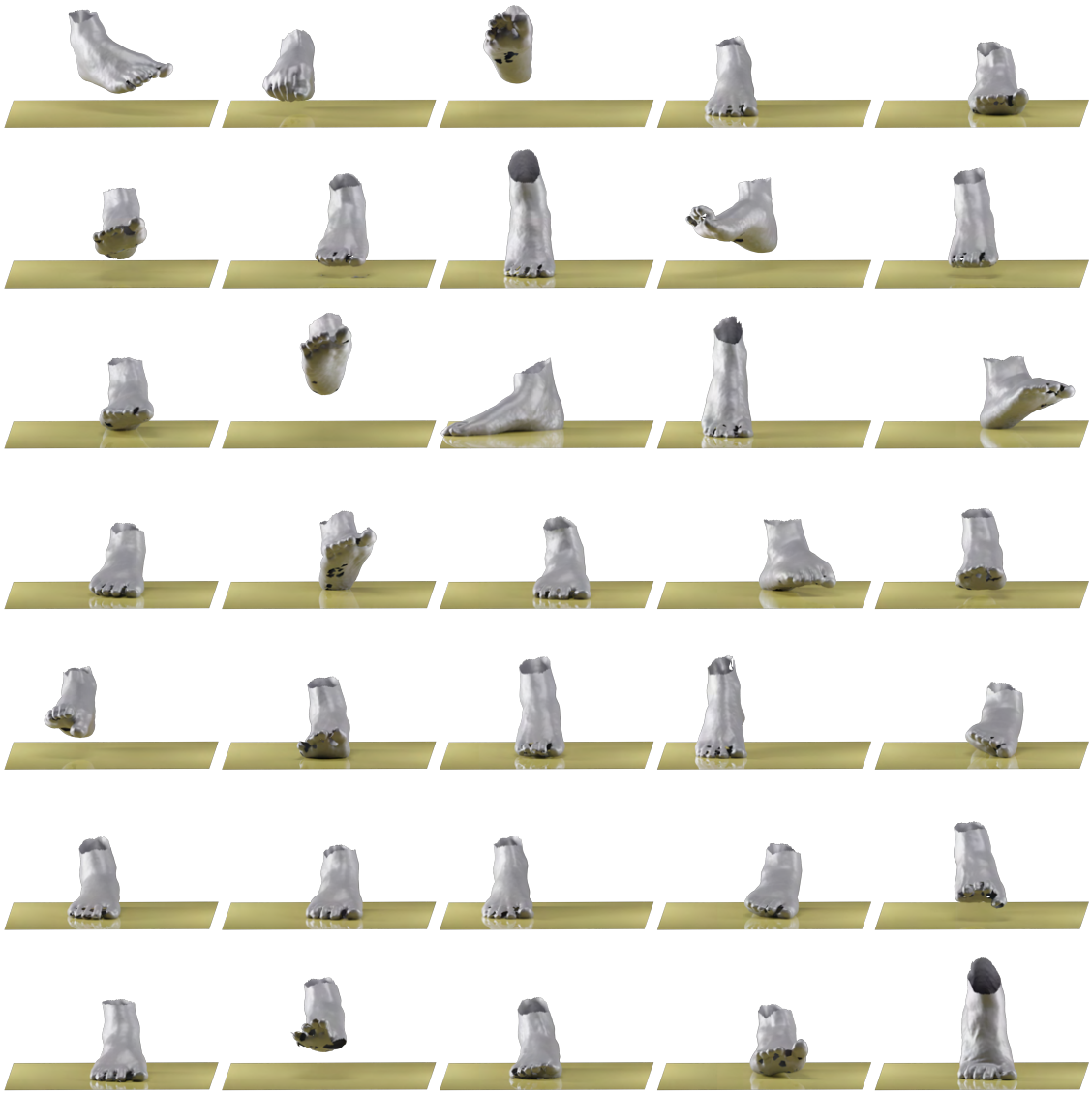
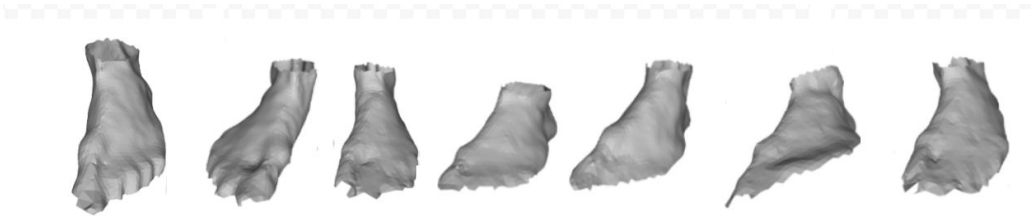


Figure 5.5: **Foot Scans:** A sample of the foot scans. The foot is fully reconstructed including the toes and the foot sole.

Data Capture Protocol

We capture a total of 30 subjects, 15 female and 15 male subjects with a total of 70,000 scans. The data capture protocol is designed by experts to explore the space of human foot deformations. The capture protocol is divided into two main parts: 1) Non-Contact

sequences, and 2) Contact Sequences. In the non-contact sequences, the subject foot is not in contact with the glass platform. The data capture protocol for such sequences is designed to explore the degrees of freedom of the toes and the ankle. In the contact sequences, the subject's foot is partially or in full contact with the glass platform. The contact sequences include motions such as walking/running and jumping. In total we capture 356 dynamic sequences. An overview of the captured scans is shown in Figure 5.5



(a) Cropped Foot from full body scans

(b) High resolution Foot Scans

Figure 5.6: **Scans Comparison:** Comparing reconstructed Foot from a full body scanner (Fig. 5.6a) with curated high resolution foot scans (Fig. 5.6b). We curate a total of $7,000$ high-resolution foot scans. The curated scans have 10x the resolution of foot scans captured in a body scanner and preserve the individual toe geometry.

Foot Shape Scans

The 30 subjects captured in the dynamic foot scanner do not represent the diversity of human foot shape. Accurate modeling of the human foot shape is crucial for the footwear industry. The Foot in the CAESAR and SizeUSA scans, shown in Figure 5.6a, are noisy, missing, and are not good enough to learn a statistical shape model. To accurately model the diversity of the human foot shapes, we acquired an additional $7,000$ high resolution foot scans from the ANSUR-II database collected by the United States army [55]. Figure 5.6 compares the curated high resolution foot scans in comparison to CAESAR and SizeUSA foot scans. In contrast to CAESAR and SizeUSA, the curated dataset of foot scans is significantly less noisy, with, on average, 10x the resolution of a foot scans from CAESAR/SizeUSA. The high resolution foot scans preserve the 3D geometry of the individual toes. We use this data in learning the the local shape space of SUPR-Foot.

5.4 Model Formulation

SUPR-Foot is a vertex-based 3D model with linear blend skinning (LBS) and learned blendshapes. The blendshapes are decomposed into 3 types: *Shape blendshapes* to capture the subject identity, *Pose-Corrective blendshapes* to correct for the widely-known LBS artifacts. To obtain an initial foot model, we further include a dataset of foot scans not in contact with the ground to SUPR introduced in Chapter 4, then separated the foot model. As result, the SUPR-Foot mesh topology and kinematic tree are based on Foot of the SUPR model topology. The template mesh contains $N = 267$ vertices and $K = 12$ joints. The SUPR-Foot kinematic tree is shown in Figure 5.7. We note, in contrast to existing body models like SMPL and SMPL-X, the SUPR-Foot kinematic tree contains significantly more joints in the foot, ankle and toes. Following the notation of SUPR,

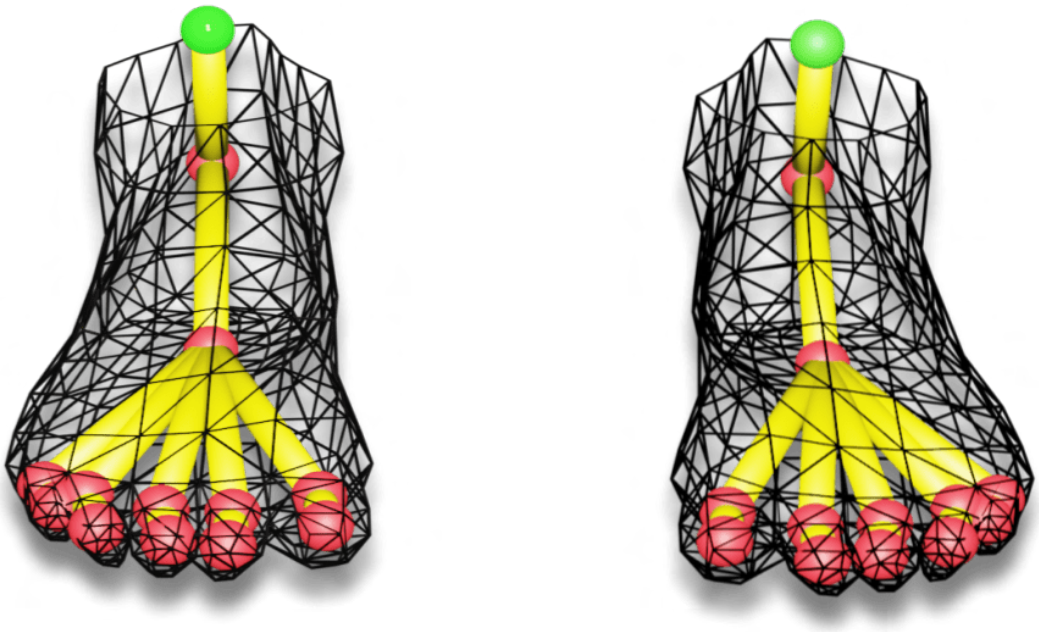


Figure 5.7: **Foot Kinematic Tree:** The kinematic tree of the Foot Model SUPR-Foot model for the right and left foot. The green sphere is the model root joint, and the red spheres are spherical joints.

SUPR-Foot is defined by a function $M(\vec{\theta},\vec{\beta})$, where $\vec{\theta} \in \mathbb{R}^{12\times 3}$ are the pose parameters corresponding to the individual bone rotations, $\vec{\beta} \in \mathbb{R}^{100}$ are the shape parameters. Formally, SUPR-Foot is defined as

$$M(\vec{\theta},\vec{\beta}) = W(T_p(\vec{\theta},\vec{\beta}),J(\vec{\beta}),\vec{\theta};\mathcal{W}), \tag{5.1}$$

where the template mesh, $T_p(\vec{\theta},\vec{\beta})$, is transformed around the joints J by the linear-blend-skinning function $W(.)$, parameterized by the skinning weights $\mathcal{W} \in \mathbb{R}^{267\times 12}$. The cumulative corrective blendshapes term is defined as

$$T_p(\vec{\theta},\vec{\beta}) = \overline{T} + B_S(\vec{\beta};\mathcal{S}) + B_P(\vec{\theta};\mathcal{P}) \tag{5.2}$$

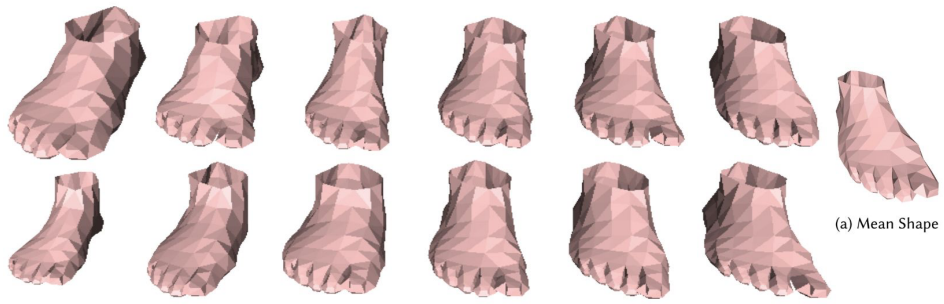


Figure 5.8: **Foot Shape Space**: Visualizing the first 6 principal components of the foot shape space learned from high resolution 3D scans. Upper row is 4 standard deviations from the mean, and the bottom row is -4 standard deviation from the mean. The first principal components (starting from the left) capture variations in the overall foot shape, while later principle components capture variations in the toe appearance

While the separated foot from SUPR have sufficient joints to capture the foot full range of articulation, the formulation still only relates the deformation of the foot to body pose, which is insufficient to model the foot deformation due to ground contact. We train the separated foot, SUPR-Foot shape space on the ANSUR-II dataset. The shape space of SUPR-Foot is shown in Fig. 5.8.

5.4.1 Foot deformation Network

The foot body part model, separated from SUPR , is defined by the pose parameters $\vec{\theta}_{foot} \in \vec{\theta}$, corresponding to the ankle and toe pose parameters in addition to $\vec{\beta}_{foot}$, the PCA coefficients of the local foot shape space. We extend the pose blendshapes in Eq. (5.2) to include a deep corrective deformation term for the foot vertices defined by $\overline{T}_{foot} \in \overline{T}$. With a slight abuse of notation, we will refer to the deformation function $T_p(\vec{\theta},\vec{\beta})$ in Eq. (5.2) as T_p for simplicity. The foot deformation function is defined by:

$$T'_p(\vec{\theta},\vec{\beta},\vec{c}) = T_p + B_F(\vec{\theta}_{foot},\vec{\beta}_{foot},\vec{c};\mathcal{F}), \tag{5.3}$$

where $B_F(.)$ is a multilayer perceptron-based deformation function parameterized by $\mathcal{F}$, conditioned on the foot pose parameters $\vec{\theta}_{foot}$, foot shape parameters $\vec{\beta}_{foot}$ and foot contact state $\vec{c}$. The foot contact state variable is a binary vector $\vec{c} \in \{0,1\}^{267}$ defining the contact state of each vertex in the foot template mesh, a vertex is represented by a 1 if it is in contact with the ground, and 0 otherwise.

Implementation details. The foot contact deformation network is based on an encoder-decoder architecture. The input feature, $\vec{f} \in \mathbb{R}^{320}$, to the encoder is a concatenated feature of the foot pose, shape and contact vector. The foot pose is represented with a normalised unit quaternion representation, shape is encoded with the first two PCA coefficients of the local foot shape space. The input feature $\vec{f}$ is encoded into a latent vector $\vec{z} \in \mathbb{R}^{16}$ using fully connected layers with a leaky LReLU as an activation function with a slope of 0.1 for negative values. The latent embedding $\vec{z}$ is decoded to predict deformations for each vertex using fully connected layers with LReLU activation.

We train a deformation network for each foot separately. Below we describe the network for the right foot. We first introduce the notation we use:

- B_P: is the linear pose corrective blendshape.
- B_C: are the predicted deformations for the foot related to pose, contact and foot shape.
- $\vec{c}$: is a binary vector of which vertices are in contact with the glass platform.
- $\vec{z}$: is a latent code vector.
- $\vec{\theta}$: are the foot pose parameters.
- $\vec{\beta}$: are the foot shape parameters.
- $\vec{f}$: is a concatenated vector of the pose, shape and contact vector.

- LReLU: leaky rectified linear units with a slope of 0.1 for negative values.

- FC_m: fully connected layer with output dimension m.

Feature Representation

The input $\vec{f}$ to the network is a concatenated feature representation of the foot pose, foot shape and contact. The foot pose representation is based on normalized unit quaternion representation defined by:

$$F(\vec{\theta}) = Q(\vec{\theta}) - Q(\vec{\theta}^*) \tag{5.4}$$

where $Q(.) : \mathbb{R}^3 \rightarrow \mathbb{R}^4$ is a function computing the quaterion representation of the input axis angle rotation, θ^* is the foot in the rest pose. The feature representation in Equation (5.4) will evaluate to 0 when the foot is in the rest pose. The foot template mesh $\overline{T}_{foot} \in \mathbb{R}^{267\times 3}$ is a high dimensional representation to represent the foot shape. We represent the foot shape using the first two principal components which roughly correspond to the foot length and foot volume. We experimented with different numbers of coefficients, and the first two principal components result in the lowest generalization error on the validation set. The state of foot contact with the scene is represented using $\vec{c}$. More formally the input feature to our network:

$$\{F(\vec{\theta}), \vec{\beta}_1, \vec{\beta}_2, \vec{c}\} \xrightarrow{\text{concat}} \vec{f} \in \mathbb{R}^{320}, \tag{5.5}$$

where $\vec{\beta}_1, \vec{\beta}_2$ are the first two PCA components and the *concat* operator is a standard vector concatenation operator.

5.4.2 Network Architecture

The architecture is an encoder-decoder fully-connected network, with non-linear activations based on LReLUs. Encoder:

$$\begin{aligned}\vec{f} \in \mathbb{R}^{320} &\rightarrow FC_{256} \rightarrow \\ &\rightarrow FC_{128} \rightarrow FC_{64} \rightarrow \\ &\rightarrow FC_{32} \rightarrow \vec{z} \in \mathbb{R}^{16}\end{aligned}$$

The dimensionality of the latent code $\vec{z}$ was chosen by grid search. We experimented with dimensionality 64, 32 and 16. A latent code with dimensionality 16 result in the lowest generalization error of the validation set. The decoder is described by:

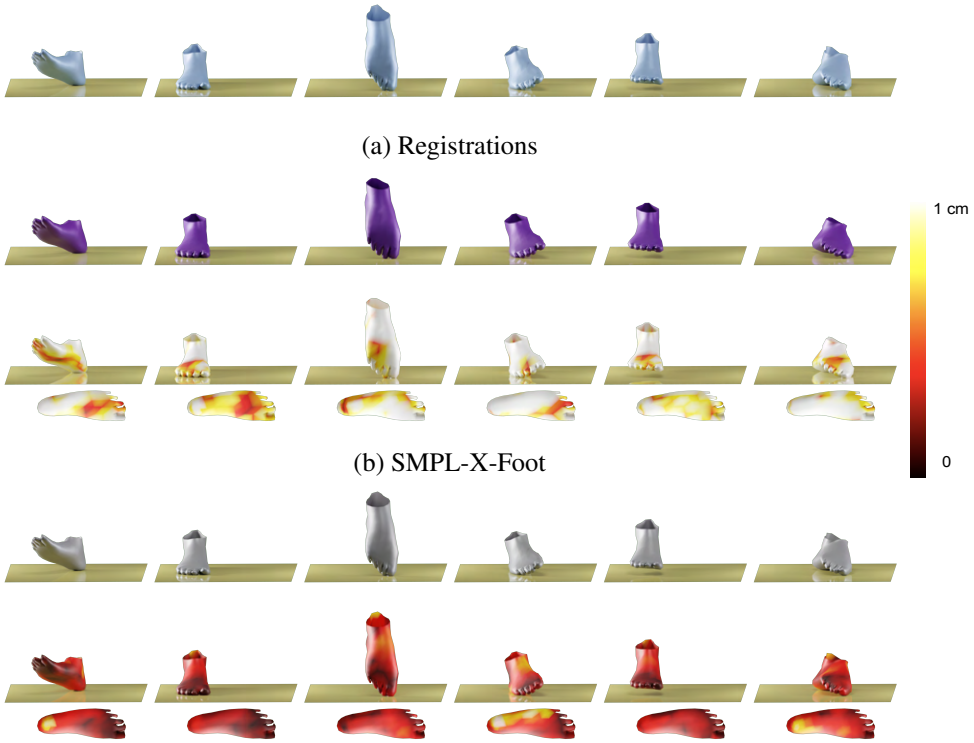
$$\begin{aligned}\vec{z} \in \mathbb{R}^{16} &\rightarrow FC_{32} \rightarrow \\ &\rightarrow FC_{64} \rightarrow FC_{128} \rightarrow FC_{266} \rightarrow B_C\end{aligned}$$

where B_C is added to the linear blendshape B_P as shown in Equation (5.3).

5.5 Evaluation

5.5.1 Model Generalization

We evaluate SUPR-Foot generalization on a test set of held-out subjects. The test set contains 120 registrations for 5 subjects that explore the foot's full range of motion, such as ankle and toe movements. We extract the foot from the SMPL-X body model as a



(a) Registrations

(b) SMPL-X-Foot

(c) SUPR-Foot

Figure 5.9: **Foot Evaluation:** Evaluating SUPR-Foot against SMPL-X-Foot.

baseline and refer to it as SMPL-X-Foot. We register the SUPR-Foot template to the test scans and fit the SUPR-Foot and SMPL-X-Foot to the registrations using a standard *v2v* objective. For SUPR-Foot, the optimization free variables are the model pose and shape parameters, while for SMPL-X-Foot the optimization free variables are the foot joints and the SMPL-X shape parameters. We report the models' generalization as a function of the number of shape components in Fig. 5.10. A sample of the model fits is shown in Fig. 5.9. SUPR-Foot better captures the degrees of freedom of the foot, such as moving the ankle, curling the toes, and contact deformations.

We evaluate SUPR-Foot against SMPL-X-Foot on a held out test set of contact and non-contact foot scans. We further break down the evaluation in Figure 5.11. We report the model mean absolute error as a function of the number of shape components used on non-contact frames in Figure 5.11a and contact frames in Figure 5.11b.

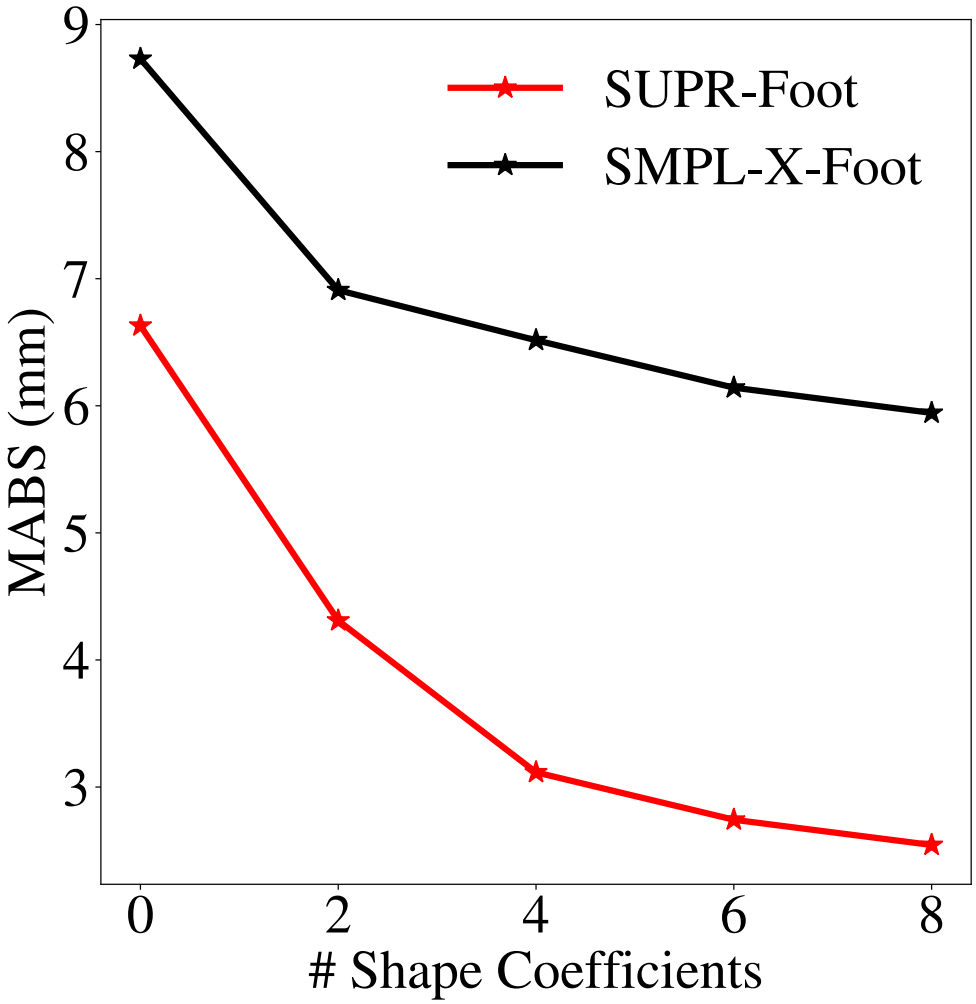
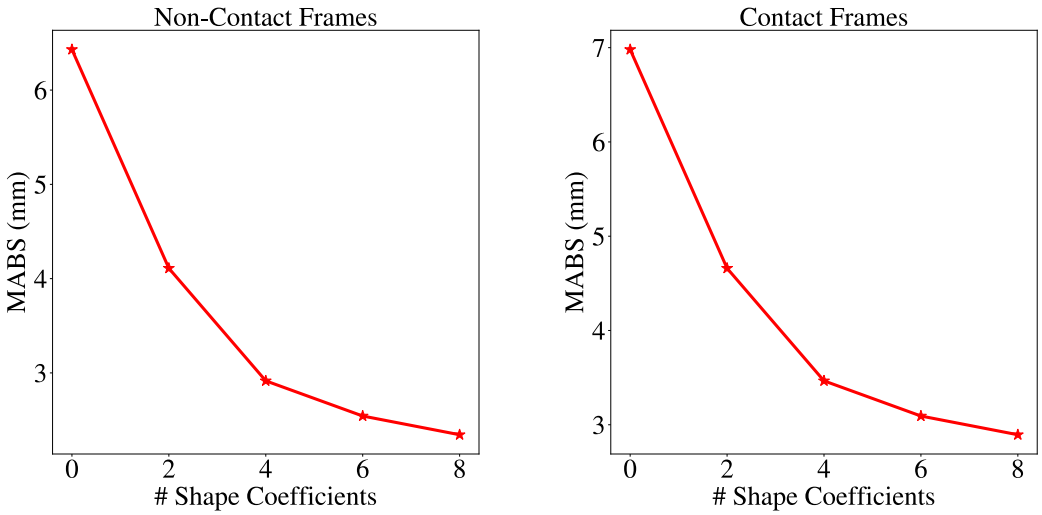


Figure 5.10: **Foot Model Generalization:** Evaluating the Generalization of SUPR-Foot against the SMPL-X Foot on a held out test set of dynamic human foot registrations.



(a) Non-Contact (b) Contact

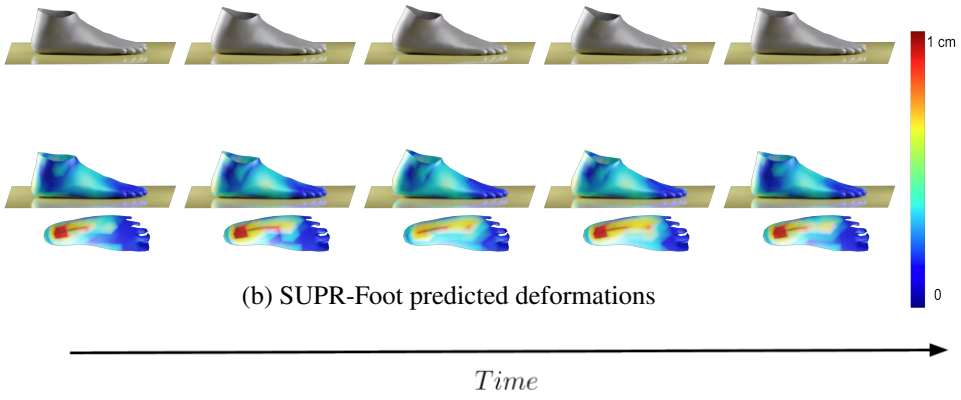
Figure 5.11: **Foot Qunatative Evaluation:** Evaluating SUPR-Foot on frames where the foot was not in contact with the glass platform shown in Figure 5.11a, and frames where the foot was partially or fully in contact with the glass platform in Figure 5.11b.

Model	Non-Contact v2v (mm) ↓	With-Contact v2v (mm) ↓
SUPR-Foot $_{lbs}$	5.235 ±0.126	6.691 ±1.369
SUPR-Foot $_{lbs+l}$	4.587 ±0.589	5.364 ±1.279
SUPR-Foot $_{lbs+l+f(\theta)}$	2.982 ±0.859	4.129 ±1.883
SUPR-Foot $_{lbs+l+f(\theta,\beta)}$	2.910 ±0.728	3.934 ±1.819
SUPR-Foot (ours)	2.753 ±0.821	3.122 ±1.462

Table 5.1: Foot Deformation Ablation Study. SUPR-Foot $_{lbs}$ corresponds to model with linear blend skinning, no additive correctives used. SUPR-Foot $_{lbs+l}$ corresponds to lbs in addition to the linear correctives, SUPR-Foot $_{lbs+l+f(\theta)}$ adds the non-linear deformation where the network is conditioned on pose only, SUPR-Foot $_{lbs+l+f(\theta,\vec{\beta})}$ the network conditioned on pose and shape information, while SUPR-Foot is the full model.

Deformation function: A key contribution of our work is introducing a novel deformation function that relates the foot deformations to the foot pose, shape and ground contact. We illustrate the influence of each term on the model generalization by ablating the foot deformation network. We retain variations of the deformation network from scratch and refit each model to the test set. We report the model *v2v* error in Table. 5.1. The result clearly shows the vertex to vertex error decreasing on the held out test set when adding each term in the foot deformation function across both the contact and non-contact frames.

(a) Raw Scanner Images



(b) SUPR-Foot predicted deformations

Figure 5.12: **Dynamic Evaluation**: Evaluating the SUPR-Foot predicted deformations on a dynamic sequence where the subject leans backward and forward, effectively shifting their center of mass.

5.5.2 Dynamic Evaluation

We further evaluate the foot deformation network on a dynamic sequence shown in Fig. 5.12. Fig. 5.12a shows raw scanner footage of a subject performing a body rocking

movement, where they lean forward then backward effectively changing the body center of mass. We visualise the corresponding SUPR-Foot fits and a heat map of the magnitude of predicted deformations in Fig. 5.12b. When the subject is leaning backward and the center of mass is directly above the ankle, the soft tissue at heel region of the foot deforms due to contact. The SUPR-Foot network predicts significant deformations localised around the heel region compared to the rest of the foot. However, when the subject leans forward the center of mass is above the toes, consequently the soft tissue at the heel is less compressed. The SUPR-Foot predicted deformations shift from the heel towards the front of the foot.

5.6 Conclusion

The human foot plays a vital role in numerous applications and industries. In this chapter, we introduced SUPR-Foot, a novel articulated model of the human foot with learned contact deformations. The development of SUPR-Foot was made possible through several significant contributions. Firstly, we used a custom-built Foot scanner specifically designed for capturing the foot, especially during moments of contact with the surrounding scene. Furthermore, by training a contact deformation network using the captured data, we were able to effectively model the foot's deformations during contact. SUPR-Foot is publicly available for research purposes.

Throughout this thesis, we have presented a comprehensive array of body and body-part-specific models designed for use by artists and animators. The development of these models requires extensive datasets and the expertise of a specialist in their training. Modifying these models to adapt to new datasets requires substantial proficiency in machine learning and computer graphics, expertise that may exceed the capabilities of most artists

and animators. This highlights a significant gap: How can artists, with limited resources, train a model using their own datasets? Addressing this question is the focus of the following chapter.

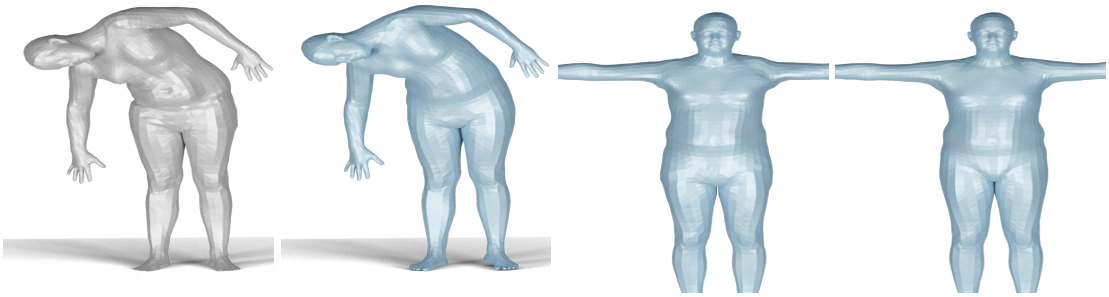
Chapter 6

AVATAR

“ That’s what we storytellers do. We restore order with imagination. We instill hope again and again and again.”

Walt Disney

Game engines have revolutionized storytelling [123, 124], a potent medium for expressing intricate ideas and emotions [125, 126, 127, 128]. By leveraging their sophisticated rendering [129] and interactivity features, these engines craft immersive narratives that significantly enhance audience engagement. Realistic digital characters [130, 131, 132] are a key pillar in these narratives because they deepen the experience of storytelling, providing audiences with relatable figures to connect with. Currently, the industry standard for generating realistic digital character is based on a labor-intensive digital sculpting process [133, 134, 135, 136, 137, 138]. Frequently, this entails numerous months of time consuming labor by skilled artists for sculpting the character deformations corresponding to the character’s body pose and anatomical details. The sculpting framework, although granting artists the complete creative control over the character, it is labor-intensive, not scalable, and remains limited by the artist’s level of expertise.



(a) Registration (b) SMPL fit (c) Pose deformations (d) Shape deformations

Figure 6.1: **SMPL Deformations:** We fit SMPL to a registration(Fig. 6.1a) from Hasler et al. [4], and show the SMPL fit (Fig. 6.1b), and the corresponding predicted SMPL pose dependent deformation (Fig. 6.1c) and the predicted SMPL shape dependent deformation (Fig. 6.1d). SMPL fails to model the deformations in the abdominal, chest and hips regions of the subject.

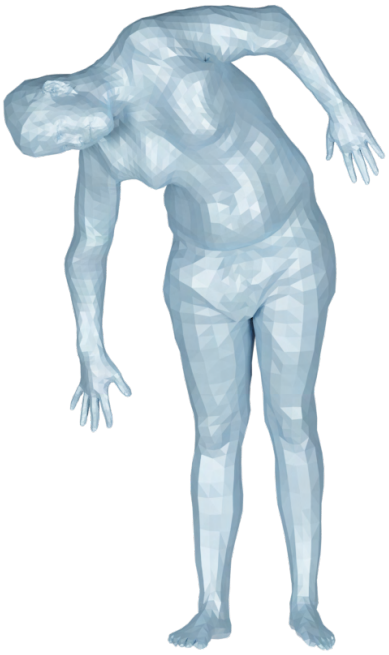
Statistical body models trained on human scans emerged as a scalable alternative to constructing virtual humans [10, 4, 33, 104, 14, 16, 139, 38, 140]. Although numerous models have been proposed, SMPL [33] is the most widely used human body model by the computer vision and graphics communities. A substantial body of literature and the corresponding open source tools built on SMPL exist, each streamlining numerous time-consuming tasks for artists and animators, such as animating SMPL by textual prompts [141, 142, 143, 144], estimating the pose and shape parameters of SMPL from images and videos [145, 146, 147, 21, 95, 148, 149, 150, 151, 18, 152, 153, 154, 155, 156, 157], automated placement of SMPL in 3D scenes [158, 159, 160, 161, 162, 163], SMPL-based motion capture datasets [6, 164, 165, 166] and automated construction of 3D scenes conditioned on the parameters of SMPL [167, 168]. Additionally, the SMPL formulation is fully compatible with the standards of the gaming and animation industry, where multiple plug-ins exist to insert SMPL into game engines such as Unity, Unreal, and Blender. This synergy of the available SMPL-based tools and game engine compatibility has the potential to empower artists and animators to focus more on the nuanced and creative aspects of characters and game design. However, there are significant draw-

backs in the deformations predicted by SMPL, which restricts its practical utility.

SMPL uses two key functions, to predict deformations related to the subject body pose and body shape. The SMPL predicted deformation fails to preserve the subject identity and pose deformations. In Fig. 6.1, we fit SMPL with 10 shape parameters, to a 3D registration from the publicly available Hasler et al. [4] dataset and visualize the corresponding SMPL fit and predicted deformations. The SMPL fit (Fig. 6.1b) captures the overall coarse body geometry of the groundtruth registration (Fig. 6.1a), yet the deformations are smooth and fail to preserve the subject's identity and the rich deformations in the subject's hip, chest, and abdominal regions.

The smooth deformation is because SMPL is trained on a multiple identity training dataset, and hence it learns average smooth deformations of the training subjects. Retraining SMPL on a single subject scan can improve the model's deformation realism. However, SMPL has a large number of parameters (4.2×10^6), which makes it easily prone to overfitting to small scale datasets. The large number of parameters is due to the pose corrective blendshape function which predicts the pose-dependent deformations of the model. In Fig. 6.2 we retrain SMPL on a single training registration for the subject in Fig. 6.1, and evaluate on the held out registration shown in Fig. 6.1a. The corresponding SMPL fit in Fig. 6.2a better captures the subject's identity compared to SMPL, however, the model suffers from clearly visible artifacts on the elbow, hip and knees. Currently, there are no existing methods to create engine-ready personalized mesh based body models that is data efficient and can be used by users with no background in machine learning such as artists. This is precisely the gap we address.

For a single subject, obtaining a large number of scans is challenging, as most online stores have a limited number of scans for a single subject [5], which makes training SMPL based models challenging. To this end, we introduce AVATAR (**A**rticulated



(a) Retrained SMPL

(b) AVATAR

Figure 6.2: **Retraining SMPL:** We fit SMPL to a registration(Fig. 6.1a) from Hasler et al. [4], and show the SMPL fit (Fig. 6.2a), and the corresponding AVATAR fit 6.2b.

Figure 6.3: **AVATAR:** is a data efficient training algorithm to learn personalized human body models from a single scan. Given a single scan downloaded from an online store [5] and registered to the SMPL mesh, we are able to learn a game engine ready human body model (on the right). The learned model can be seamlessly inserted in Blender using the publicly available SMPL Blender plug-in.

Virtual hum**A**ns **T**rained by B**A**yesian infe**R**ence from a Single Scan), a data efficient training algorithms for learning personalized models based on SMPL from a single scan. Our key insight is all existing training algorithms optimize for a single point estimate of the model parameters, which best explains the training data, which makes learning prone to overfitting. Instead, we formulate character learning as a Bayesian inference problem, where we use a prior distribution of possible model parameters and reason about a distribution of possible parameters which fits the training data, instead of single point estimate. Additionally, to learn the detailed subject shape, we perform a parameter-free optimization to optimize for a subject specific body shape and joint location which are reguarlized to a symmetric prior to reduce the risk of overfitting. We train a model using AVATAR trained on a single scan and show the corresponding fit in Fig. 6.2b, compared to SMPL fit in Fig. 6.2a, the AVATAR based model captures the subject identity and rich deformation on the chest, abdominal region, and hips. AVATAR is an automated algorithm that does not require fine-tuning or user intervention and can be used by artists without requiring a background in machine learning to train a model from a single scan, which can be imported into a game engine, as shown in Fig. 6.3.

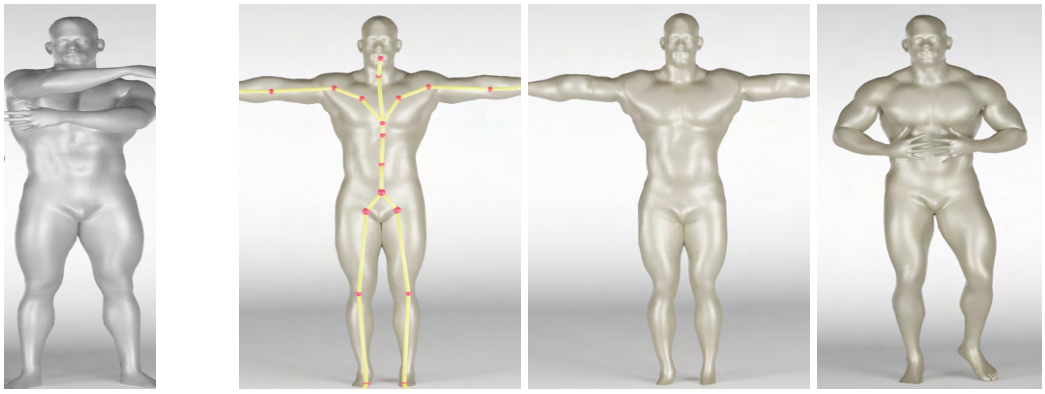
We note the characters trained in this chapter are based on the SMPL formulation with all its widely known drawbacks, as discussed in Chapter 3. However, the principles introduced in this chapter are independent of a specific representation and can be used to train models based on any linear blend skinning formulation such as SMPL-X, STAR and SUPR.

We evaluate the characters trained using AVATAR, and show that AVATAR characters can generalize better than SMPL on a held out test set given a single training registration. Characters trained with AVATAR better capture intricate subject-specific deformations influenced by both body shape and pose. We evaluate AVATAR characters against retrained personalized SMPL models, and highlight that SMPL is prone to overfitting. Furthermore, we animate an AVATAR character using a motion capture sequence from the AMASS dataset [6], and highlight the AVATAR deformation fidelity compared to SMPL. To summarize AVATAR key contributions:

1. We introduce a training algorithm to learn SMPL based personalized models, from a single scan.

2. We pose model training as a Bayesian inference problem, which is key to robustify the training even from a single scan.

The rest of the chapter is organized as follows: Sec. 6.1 we describe the AVATAR training pipeline, Sec. 6.2 we evaluate AVATAR characters generalization, we describe the potential negative impact of our work in Sec. 6.3, Sec. 6.4 describes future work. We finally conclude with a discussion in Sec. 6.5.



(a) Registration (b) Template Joints (c) Blendshapes (d) AVATAR

Figure 6.4: **AVATAR Pipeline:** Given a single training registration (shown in Fig. 6.4a), we first estimate personalized subject template mesh and joints (shown in Fig. 6.4b). Given the subject-specific template and joints we infer a distribution of pose corrective deformations. The pose deformations capture subject-specific deformations such as muscle bulges for the bodybuilder (shown in Fig. 6.4c). The template and corrective blendshapes are then rotated around the personalized joints to predict the final mesh shape (shown in Fig. 6.4d).

6.1 Method

AVATAR is a training algorithm for any vertex-based linear blend skinning model (LBS). While we base our formulation on the SMPL body template topology and kinematic tree, the AVATAR framework is applicable to any mesh-based LBS model.

6.1.1 Model

We start with a low-resolution template mesh $\overline{T} \in \mathbb{R}^{N\times 3}$ with $N = 6,890$ vertices. Similar to SMPL, the model kinematic tree contains $J = 24$ joints. The model is fully parameterized by pose parameters $\vec{\theta} \in \mathbb{R}^{24\times 3}$. To address the widely known drawbacks of standard LBS, we use a pose-corrective blendshape function which is added to the template mesh $\overline{T}$ such that when posed with a standard skinning function $W(.)$, it looks realistic. More

formally:

$$T_p(\vec{\theta}) = \overline{T} + B_P(\vec{\theta}; \mathcal{P}), \tag{6.1}$$

where $\mathcal{P} \in \mathbb{R}^{6890\times3\times207}$ is the pose corrective blendshapes which regresses corrective offsets related to the model joint rotations. The pose corrective blendshape includes 4.2×10^6 parameters.

The template and pose blendshapes in Eq. (6.1) is transformed around the model joints J in the kinematic tree using standard LBS:

$$M(\vec{\theta}) = W(T_p(\vec{\theta}), J, \mathcal{W}), \tag{6.2}$$

where the linear blend skinning function $W(.)$ rotates the template $\overline{T}$ and the cumulative sum of blendshapes term $T_p(\vec{\theta})$ around the 3D model joints J, linearly smoothed with the skinning weights $\mathcal{W} \in \mathbb{R}^{6890\times24}$. We note in contrast to generic human body models such as SMPL, in AVATAR the joints J and the template $\overline{T}$ are subject specific and are inferred from the character training registration as shown in Fig. 6.4.

6.1.2 Model Training

Our goal in this section is to learn the model variables from a single registration. The model trainable variables are: the personalized subject specific template $\overline{T}$, the subject specific joints J and subject specific pose corrective blendshapes $\mathcal{P}$. The full AVATAR training pipeline is summarized in Fig. 6.4.

AVATAR trains a model by minimizing the vertex-vertex loss between the model and the corresponding training registration. AVATAR pose parameters, template and joints are trained by minimising a standard vertex-to-vertex loss between the AVATAR model

in Eq. (6.3) and the groundtruth registration:

$$E_D(\vec{\theta}) = ||V - M(\theta, T, J)||^2, \tag{6.3}$$

where $V \in \mathbb{R}^{6890\times 3}$ is the training registration. We minimise Eq. (6.3) relative to the model pose parameter $\vec{\theta}$. Given the pose parameter we estimate the personalized template $\overline{T}$ and J by minimising Eq. (6.4):

$$E = E_D + E_R \tag{6.4}$$

$$E_R = \sum_j ||J - U(J)||^2 + ||T - U(T)||^2 \tag{6.5}$$

where the data term is additionally regularized by a symmetrical prior term E_R over the joints and the template as shown in Eq. (6.5), where U(.) is a function that mirrors the template vertices and joints across the Y-Z plane. This term encourages symmetric template meshes and symmetric joint locations.

6.1.3 Training the Pose Blendshapes

Given the character pose parameter $\vec{\theta}$, personalized template $\overline{T}$ and joints J we estimate the pose corrective blendshape term, by minimizing the data term in Eq. (6.3), to obtain corrective residuals $Y \in \mathbb{R}^{6890\times 3}$ the residuals between the model $M(\theta, T, J)$ and the corresponding registration. Similar to SMPL, we use the Rodrigues feature $X \in \mathbb{R}^{23\times 9}$ as the feature representation of the model's kinematic tree. The training data used to infer the pose corrective blendshapes distributions is $\mathcal{D} = \{Y, X\}$, such that:

$$Y = X\mathcal{P}. \tag{6.6}$$

We infer a probability distribution over the pose corrective blendshapes $\mathcal{P}$ using Bayesian linear regression [7]. In a Bayesian inference framework, we model our beliefs using a prior distribution, and given a stream of data we update our prior beliefs to obtain a posterior distribution. The prior distribution over the pose corrective blendshapes are defined by:

$$P(\mathcal{P}) = \mathcal{N}(\mathcal{P}|\mathcal{P}_0, V_0) \tag{6.7}$$

where $\mathcal{N}$ is a multivariate Gaussian distribution parameterized by a mean $\mathcal{P}_0$ and covariance matrix V_0. The likelihood term of the data conditioned on the pose corrective blendshapes is given by:

$$P(Y|X, \mathcal{P}, \mu, \sigma^2) = \mathcal{N}(Y|X\mathcal{P}, \sigma^2), \tag{6.8}$$

where σ^2 is the observation noise (the 4D scanner noise). The posterior distribution is given by Eq. (6.9):

$$P(\mathcal{P}|\mathcal{P}_N, V_N) \propto \mathcal{N}(\mathcal{P}|\mathcal{P}_0, V_0)\mathcal{N}(Y|X\mathcal{P}, \sigma^2), \tag{6.9}$$

where the mean of the $\mathcal{P}_N$ is given by:

$$\mathcal{P}_N = V_N V_0^{-1}\mathcal{P}_0 + \frac{1}{\sigma^2} V_N X^T Y, \tag{6.10}$$

$$V_N^{-1} = V_0^{-1} + \frac{1}{\sigma^2} X^T X, \tag{6.11}$$

and the covariance V_N is given by

$$V_N = \sigma^2(\sigma^2 V_0^{-1} + X^T X)^{-1}, \tag{6.12}$$

The posterior predictive distribution of the model vertices is derived by marginalizing over the probability distribution of all possible $\mathcal{P}$ where the mean and co-variance are defined by:

$$p(y|X,\mathcal{D},\sigma^2) = \int \mathcal{N}(y|x^T\mathcal{P},\sigma^2)\mathcal{N}(\mathcal{P}|\mathcal{P}_n,V_n)d\mathcal{P} \tag{6.13}$$

$$= \mathcal{N}(y|\mathcal{P}_N^T X,\sigma_N^2(X)) \tag{6.14}$$

$$\sigma_N^2(X) = \sigma^2 + X^T V_N X \tag{6.15}$$

We note that the mean of the posterior predictive distribution is still linearly related to the Rodrigues feature representation X of the model's kinematic tree, which is critical for the model to be compatible with the gaming and animation industry standards.

6.1.4 Gaussian Motivation

The motivation to use the multivariate Gaussian distribution for the prior in Eq. (6.7) and the likelihood Eq. (6.8) is that the posterior distribution has a closed-form analytic solution that is also a multivariate Gaussian distribution. The posterior is a Gaussian since both the prior and the likelihood are Gaussians and the residuals Y are linearly related to the feature X as shown in Eq. (6.6). The closed-form analytic solution for the posterior simplifies the inference step for the distribution of the pose corrective blendshapes.

6.2 Experiments

Our objective in the evaluation is to assess the fidelity of the AVATAR characters and its robustness to overfitting when trained using a single scan.

Dataset We use 9 subjects from the publicly available 3D scan dataset Hasler et al. [4].

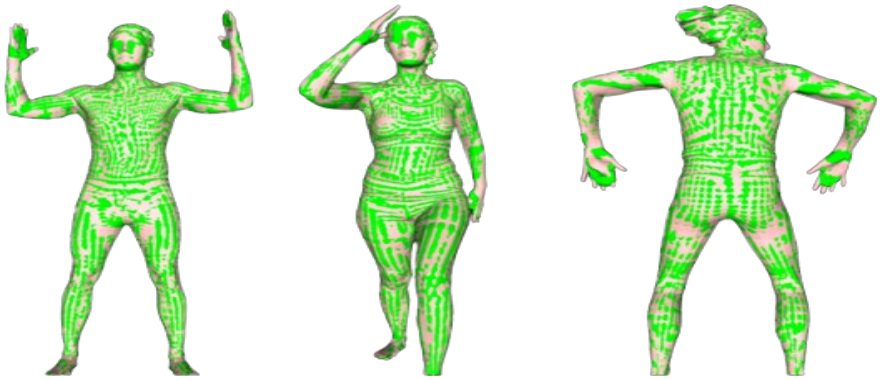


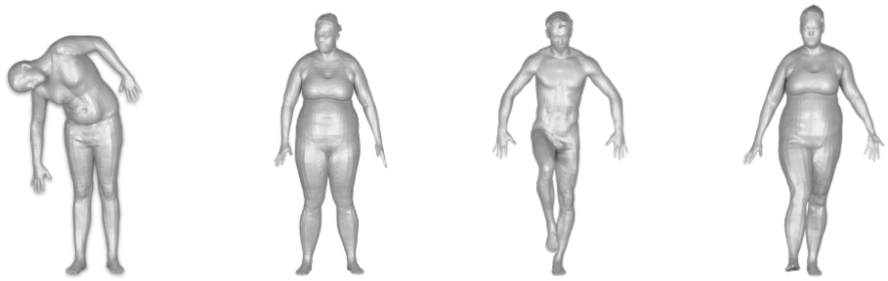
Figure 6.5: **Registration:** Sample registration for three different subject, where the SMPL model template mesh (in pink) is tightly fit to a raw 3D scan in green.

For each subject, we use a single scan for the model training, and the remaining scans are held out for evaluation. We register all scans to the SMPL mesh template as shown in Fig. 6.5.

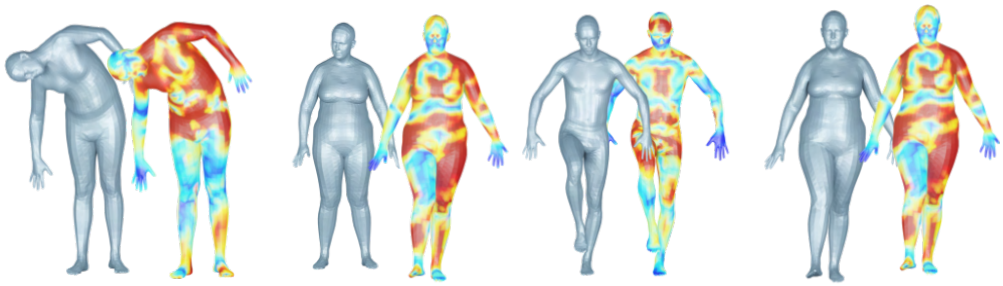
Baselines The first baseline is gendered SMPL using 10 shape components. The second baseline is a retrained SMPL model. We retrain a personalized SMPL model, which we refer to as SMPL^R on a single training registration.

6.2.1 Characters Generalization

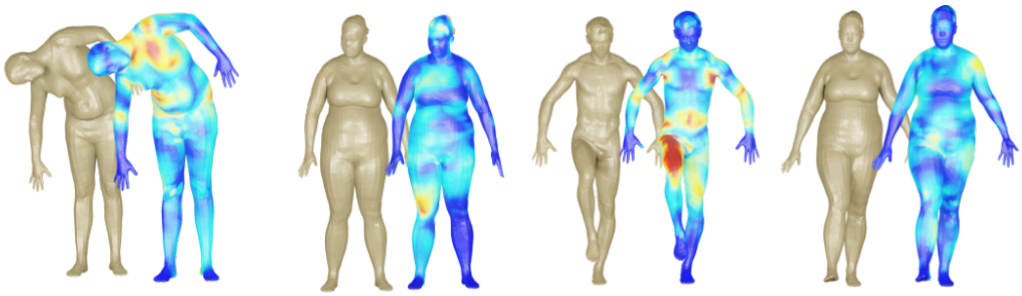
All the model parameters for SMPL , SMPL^R and AVATAR are trained solely from a single train registration. Then we evaluate all models on the held out test set using a standard vertex-to-vertex ($v2v$) objective and report the corresponding mean absolute error(mabs) in Tab. 6.1. Characters trained using AVATAR uniformly have a lower *mabs* error compared to all baselines. In Fig. 6.6, we show a qualitative comparison on the held out test set. We note that for SMPL (Fig. 6.6b) the fits capture the overall coarse body geometry, yet are still overly smooth and fail to capture the subject identity or deformations related to the body pose. The characters trained by our proposed method AVATAR (Fig. 6.6c) are able to generalize better, preserve subject identity, and the deformations



(a) Held out Registration



(b) SMPL



(c) AVATAR

Figure 6.6: **Qualitative Evaluation:** Qualitative comparison between characters trained by AVATAR and baseline methods. Given a held out test set (shown in Fig. 6.6a), we fit the publicly available gendered SMPL model with 10 shape components (Fig. 6.6b), and, and personalized characters models trained on a single scan using AVATAR(Fig. 6.6c).

ID	SMPL [33]	SMPL^R	AVATAR (Ours)
S1	5.49	8.41	**4.90**
S2	4.68	8.85	**4.09**
S3	4.35	8.82	**3.30**
S4	5.38	6.99	**2.89**
S5	5.23	8.83	**4.43**
S6	5.88	8.65	**3.60**
S7	4.14	7.50	**3.24**
S8	5.77	7.74	**2.81**
S9	5.23	10.32	**5.06**
All	4.99	8.46	**3.81**

Table 6.1: The per subject mean absolute error (mabs) - in mm - on the held out test set of 10 different subjects. S1 denotes subject with ID = 1. Models trained using AVATAR uniformly have the lowest error across all subjects.

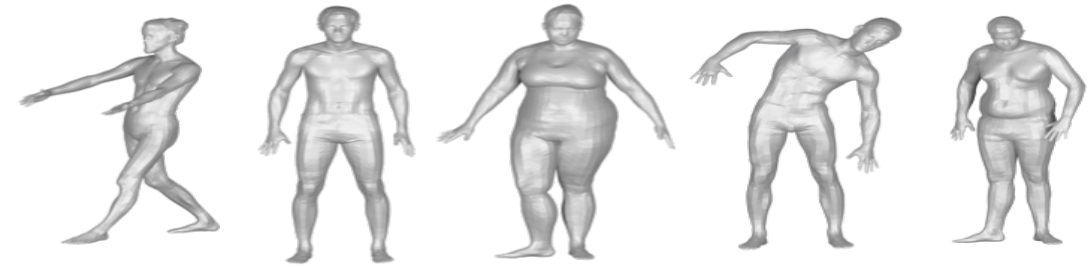
related to the subject body pose.

6.2.2 Personalized Shape

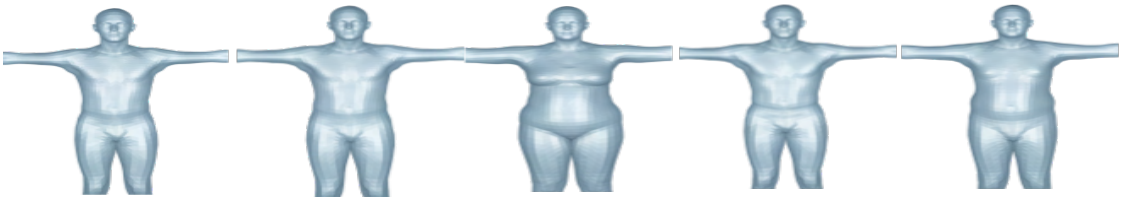
AVATAR learns a personalized shape of the subject, by optimizing the base template. The personalized shape captures subject-specific details which cannot be captured with the SMPL shape space. In Fig. 6.7, we show the estimated subject shape for a sample training registrations from the Hasler et al. [4] dataset. Using 10 shape components, SMPL model does not adequately maintain the subject's identity nor the high-frequency anatomical details. On the other hand, the shape estimates produced by the AVATAR personalized shape capture the subject shape, identity and rich anatomical details.

6.2.3 Character Ablation

The AVATAR characters feature two main types of personalized deformations: those associated with the subject's body shape (personalized templates and joints) and those corresponding to deformations related to the subject's body pose (personalized pose de-



(a) Training registrations



(b) SMPL with 10 components shape estimate.

(c) AVATAR personalized shape estimate.

Figure 6.7: **Shape Estimation:** For each training subject in Fig. 6.7a, we show the estimated subject shapes SMPL with 10 components (Fig. 6.7b) and AVATAR personalized shape (Fig. 6.7c)
.

formations). In Tab. 6.2 we perform an ablation on each of the components of AVATAR and report the mean absolute error on the held out test set. To ablate the deformations related to the subject shape, we retrain characters without estimating the personalized template and joints, and for the deformation related to subject pose, we do not update the distribution of, hence using the prior distribution mean. Each of the ablated models was retrained from scratch and we report the *mabs* error on the held out test set. The best performing AVATAR characters correspond to the subjects trained with personalized anatomical and subject specific deformations.

Case	Personalised Shape	Personalised Pose	mabs (mm)
1	+	+	3.81
2	-	+	10.76
3	+	-	4.31
4	-	-	10.93

Table 6.2: Ablation of AVATAR characters for the pose prior and shape priors used during the characters training.

6.2.4 Motion Capture Evaluation

We evaluate a bodybuilder character trained using AVATAR on a climbing motion capture sequence from AMASS [6]. We animate the SMPL body fit to the bodybuilder and show the results in Fig. 6.8a, and similarly we show the AVATAR character in Fig. 6.8b. The AVATAR model is able to preserve the character muscularity and the results are significantly more plausible and convincing. In the given scenario, it becomes evident that as the subject's legs are raised in the direction of the abdominal region, there is a significant engagement of the abdominal muscles. This engagement is prominently noticeable on the AVATAR character. On the other hand, the SMPL model mesh exhibits a bulge that appears quite unrealistic, especially when considering a subject with an athletic build. This implausibility in the SMPL model's depiction stands in contrast to the realistic muscular activation observed in the AVATAR character.

6.3 Negative Impact

AVATAR streamlines the construction of virtual humans from 3D scans. We note that no prior methods exist to date for learning engine ready human body models. Its implementation may have negative impact for 3D artists, and potentially jeopardizing multiple jobs. However, the primary aim of AVATAR is not to replace 3D artists, but to serve as an AI assistant that enhances their capabilities. AVATAR is designed to be a tool that

(a) SMPL (b) AVATAR

Figure 6.8: **Motion Capture Evaluation:** We animate the SMPL model (Fig. 6.8a) and the AVATAR trained character (Fig. 6.8b) by a climbing sequence from the AMASS dataset [6]. Unlike the SMPL mesh, the AVATAR mesh demonstrates greater plausibility, particularly in capturing muscle bulges in the abdominal and chest areas as the subject climbs.

augments artists, rather than replacing them. The characters created through AVATAR are fully interpretable to artists, allowing for subsequent refinement and customization.

(a) Registration

(b) AVATAR personalized shape

(c) Top view: AVATAR personalized shape

Figure 6.9: **Symmetric Training Scans:** We estimate subject specific personalized templates for the subject in Fig. 6.9a. The estimated template is shown in Fig. 6.9b and Fig. 6.9c.

This ensures that artists retain full creative control over their work while being more productive.

6.4 Limitations and Future Work

Symmetric Training Poses AVATAR factorizes deformations into pose and shape deformations. Shape deformations are symmetric deformations. If a subject is in a perfect symmetric body pose as shown in Fig. 6.9a, as a consequence the symmetric deformations related to the subject pose will be explained by the base template as shape deformation. As a result, the base template will contain deformations not related to the subject body shape as shown in Fig. 6.9b and Fig. 6.9c, where the template contains bulges

clearly seen around the model elbow.

Registration Currently, our method involves the registration of raw scans to a template mesh, a step that requires expert oversight to ensure an accurate fit. This dependence on expert input presents a challenge to the scalability and efficiency of our algorithm. Consequently, a promising area for future research lies in developing techniques that reduce or eliminate the necessity of registering scans to a template mesh. Such advances would significantly improve the efficiency and scalability of the construction of personalized body models.

Expressive AVATARs The AVATAR formulation is currently limited to learning the personalized shape and pose deformation and does not consider the facial expressions of the subject. A promising future direction is to extend the AVATAR formulation to expressive humans, to include the development of a personalized expression space and a specialized pose deformation space for the hands, adding further depth and expressiveness to the models.

6.5 Conclusion

We introduce AVATAR, a data-efficient algorithm for creating high-fidelity virtual characters. We tackle two critical challenges associated with statistical human body models. Current models, such as SMPL, are trained on multiple identity datasets with diverse body shapes and poses. However, because of this setting, SMPL fails to capture the rich subject-specific pose and shape deformations. When fitting SMPL to a held-out test registration, the deformations are overly smooth and unrealistic. Existing body models tend to have a large number of parameters and are susceptible to overfitting. To address this,

we propose a Bayesian learning training algorithm for personalized body models. We learn subject-specific anatomical details, including joint locations, detailed body geometry, and subject-specific pose deformations. Rather than inferring a single-point estimate of the model parameters, we derive a full probability distribution of possible parameters to enhance the model's robustness against overfitting. Our results demonstrate that with only a single training registration, an AVATAR character can generalize better than the widely used SMPL body model. No prior work has focused on developing scalable data-efficient digital character. In this chapter, we introduce the concepts that make this possible.

Chapter 7

Conclusion

The primary objective of this thesis was to highlight the limitations of existing statistical models and training algorithms for the body and its parts and to propose alternatives that are compatible with the existing gaming and animation industry standards. To this end, we introduce a comprehensive set of models and algorithms. Our key hypothesis was that the limitation of current models can be addressed by enforcing prior-domain knowledge or alternatively leveraging large training datasets. This results in a full suite of models and training algorithms, which can readily be integrated in an artists workflow.

7.1 Thesis Contributions

This thesis presents four key deliverables, each advancing the state-of-the-art in modeling and training models of the human body and its individual parts.

STAR In Chapter 3, we highlight the drawbacks of a commonly used representation that relates all body joints to all model vertices, which results in learning false long-range spurious correlations from the training data. We address this by introducing STAR, a

sparse representation of the human body deformations, which results in learning strictly sparse pose deformations. Our key hypothesis in STAR was to incorporate our prior knowledge that a single joint movement influences only a subset of the model vertices. We adopt a learning approach to infer that the set of vertices influenced by each joint's movement. Additionally, we further condition the pose deformation function on the subject body shape. Both innovations in the STAR's corrective blend shape formulation are grounded in domain-specific priors related to the deformation patterns of the human body.

SUPR In Chapter 4, we show that existing models for the head and hands cannot model the full range of motion of their corresponding body parts. We address this by introducing a federated training algorithm for the body and its parts, which enables learning a full suite of models. Unlike existing approaches, we start with a sparse expressive human body model, SUPR, and train the model on a federated dataset of body, head, and hand registrations, then separate the model into individual parts. The separated body parts included significantly more joints compared to existing body part models, which are critical to modeling the full range of motion of the body part. The key to the success of SUPR is the use of a federated training dataset of head, hand, and body registrations that allows us to learn the influence of each body joint on the separated body part models.

SUPR-Foot In Chapter 5, we introduce the first articulated model of the human foot. Existing human-body models only use two joints to model foot articulation, which is insufficient to capture the full range of motion of the foot. Due to the frequent ground contact of the foot, we propose a novel neural deformation model based on the pose, shape, and ground contact of the foot. The construction of the foot and the learning of the contact-based neural deformation function is made possible because of the large

registration training dataset captured by a dedicated 4D foot scanner.

AVATAR In Chapter 6, we address the problem of data-efficient learning of personalized human body models. Artists typically have access to very few scans and would like to learn a personalized subject specific model that can preserve the subject identity. All existing models are trained on a large dataset and will fail to prepare the subject-specific identity and pose deformations. AVATAR enables learning models from as few as a single registration. Key to AVATAR success is a Bayesian formulation which incorporates a prior distribution on the model pose corrective blendshapes.

In this thesis, we advance the field of human body modeling by using two principal frameworks: those that integrate priors, namely STAR and AVATAR, and those powered by extensive datasets, such as SUPR and SUPR-Foot. This dichotomy highlights the diverse approaches towards enhancing the state-of-the-art in human body models. Methods incorporating prior knowledge leverage existing knowledge to improve the realism of existing models, while data-driven approaches utilize large datasets to improve accuracy and adaptability, broadening the potential applications of these models. Consequently, this work sets the stage for future research to further explore the two principals introduced to further advance realism of existing human body models.

7.2 Limitations

3D Registration Through the thesis, we use 3D Registrations to train models for the body and its parts. We use registration as the primary method to establish ground truth. This is because raw 3D scans, which are the starting point for generating these models, are an unstructured data format that often contains missing parts and noise. As a result, a registration process to fit the scans to a common template mesh and eliminate any errors

or inconsistencies.

In Chapter 3, we use the SizeUSA dataset registrations, which allow the creation of a more comprehensive shape space, particularly for accurately modeling the female chest shape in a traditional bra. In Chapter 4, the federated dataset of body, head, and hand data enables the development of a full suite of body and body part models capable of tracking the full range of motion of each human body part. In Chapter 5, the use of foot registrations enables the creation of a novel foot model with contact deformations.

The registration process involves a model-free optimization in which the mesh is deformed to fit a scan. This process is designed to correct for any discrepancies between the two, ensuring that the resulting model is as accurate as possible. This is a computationally expensive process, as it requires computing an AABB (Axis Aligned Bounding Box) data structure in each optimization iteration to calculate the point-to-plan distance between each scan point and the mesh triangles. It is important to manually inspect each registration to ensure that it is free of artifacts and can be used as a reliable source of ground truth.

As a consequence of the manual labor, expertise required, and computational cost, this makes 3D registrations a very time-consuming process and, if not done carefully, will eventually introduce systematic biases in the training data, which will in turn result in artifacts in the resulting model. 3D registration remains a key bottleneck for training models at scale.

In future work, it is worth considering training a model directly from scanner images rather than relying on 3D registration. This approach would eliminate the need for raw 3D scans and potentially bypass some of the challenges associated with 3D registration. One potential benefit of this approach is that the scanner images contain both RGB cameras and scattered patterns from a dedicated projector, which can provide a more detailed

Figure 7.1: **Modeling Detail:** A raw scan shown on the right (in gray) and the corresponding SMPL fit on the left (shown in blue). SMPL fails to capture the rich detail in the scan.

and accurate representation of the 3D geometry.

However, training a model directly from scanner images presents its own set of challenges. One issue is that the camera views are often close up views, meaning that only patches of the geometry are visible. This can make it difficult to accurately train a model and may require additional computational resources or expertise. Despite these challenges, training a model directly from scanner images has the potential to significantly improve the efficiency and accuracy of model training.

Modeling High Frequency Details All the models we propose in the thesis are based on a low-resolution mesh. A low-resolution mesh is desirable for computational efficiency; it will fail to capture details such as wrinkles and muscle bulges, as shown in Fig. 7.1. The lack of detail significantly compromises the model's visual realism. Increasing the model's mesh resolution will enable modeling more high-frequency details; nevertheless, this will in turn significantly inflate the model's computational foot print. Artists typically represent details using displacement maps. Displacement maps

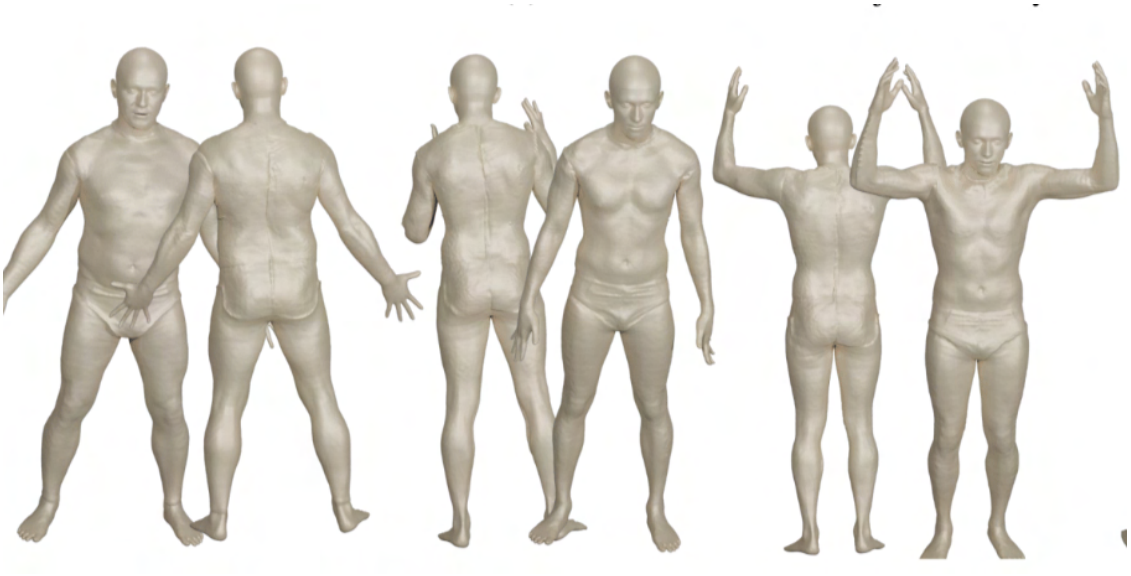


Figure 7.2: **Seams Artifacts:** SMPL enhanced with a displacement map generated by a variational auto-encoder presents a detailed model, yet with noticeable artifacts along the UV seams.

are based on UV map representation. In a UV map representation, the mesh is unwrapped on a 2D image where there is a 1-1 correspondence between each pixel and a point on the mesh surface. Displacements maps are gray-scale images where each pixel encodes a displacement offset such that at run-time the displacement shader subdivides the mesh into a high-resolution mesh and samples the displacement map for detail offsets that when added the mesh looks realistic. The combination of a low-resolution model and displacement maps is advantageous because only a low-resolution model needs to be trained and the displacement map can be computed relative to the low-resolution model.

Numerous generative models for 2D images exists, which we can use to learn generative models of displacements maps. Training a generative model of displacement maps conditioned on a low-resolution model pose and shape parameters will result in capturing more details as shown in Fig. 7.2. However, a key drawback of existing generative architecture of displacements maps is that they consistently result in clearly visible arti-

facts on the model seams, as shown in Fig. 7.2. Future work should focus on developing generative models for UV maps, which preserve the 3D surface properties along the UV seams.

Topology Agnostic Modeling The models we introduce adhere to a consistent mesh topology; the STAR model is based on the SMPL mesh, while the SUPR model utilizes the SMPL-X mesh. However, the choice of the mesh topology is largely influenced by the specific requirements of artists, who may prioritize different levels of detail across various body parts, such as preferring more vertices on the hands than on the head, or vice versa. However, our reliance on a fixed template topology is a significant constraint within the artist's creative workflow. Future research should explore the development of models and training algorithms that are indifferent to the template topology, thus giving artists the flexibility to dictate the desired structure of the model template. One potential approach can include the use of implicit representation [35], which is a continuous representation of the model surface using a neural network. This implicit surface can then be discretized at run-time into a template mesh, tailored to the artist's specific mesh requirements, thus eliminating a notable bottleneck in the artist's creative process.

7.3 Neural Models

Recent advancements and innovations in the field of computer graphics have led to the introduction of various 3D representations. Among these are implicit representation [35], Neural Radiance Fields [36], and Gaussian splatting [169], each contributing to the expansion of the field's capabilities. This phase of rapid evolution has also seen the emergence of foundational models capable of generating content of unprecedented fidelity. Notably, the development of technologies such as OpenAI's Sora has facilitated

the production of high-quality videos, while tools like DALL-E have revolutionized the creation of realistic images, all from simple user-generated text prompts. The advent of these novel representations and architectural advancements has significantly broadened the horizons for achieving highly realistic and intricately detailed visual content. How might future research endeavor to integrate the latest advances in 3D modeling into the workflows of artists?

Artists' primary requirement is to have complete creative authority over graphic assets, entailing control at the pixel level for the asset's visual appearance. For instance, Linear Blend Skinning (LBS), established in the 1990s, remains the favored technique for crafting articulated human figures, despite its acknowledged flaws and the existence of more accurate methods. This ongoing preference is significantly attributed to the ease with which artists can intuitively adjust skinning weights, alter joints as necessary, and observe the immediate impact on the model's geometry. Such a comprehensible and manipulable framework is crucial for the technology's widespread acceptance. The demand for creative control is a constant that transcends the boundaries of specific technologies or methodologies.

" I believe in creative control.

No matter what anyone makes, they should have control over it."

David Lynch, American filmmaker

NeRFs and Gaussian splatting advances the capability to capture the realistic appearance of scenes, though they offer limited control over graphic assets. For example, these representations do not allow for the seamless removal of objects from a scene or the addition of new objects. This limitation stems from the fact that the representations are not compositional; they lack native segmentation within the 3D scene, making it impossible to modify specific elements independently. Furthermore, there is no facility

to relight the scene, which restricts the ability to adjust lighting conditions post-capture, thereby diminishing the flexibility and realism that can be achieved in the final visual output.

Radiance fields provide an excellent solution for capturing snapshots of reality, but the future of 3D modeling and scene reconstruction lies in addressing their limitations through the solution to inverse problems. By converting a radiance field into a native graphic asset, including the segmentation of the scene, recovery of textures, and assignment of materials, we can significantly enhance the utility and adaptability of radiance fields for practical applications. This process involves solving complex inverse problems, raising the critical question of how we can train models to effectively tackle such challenges. The development of models capable of solving these inverse problems will pave the way for more dynamic, editable, and realistic representations of 3D environments, expanding their applicability in various fields including virtual reality, film production, and video game development. To date, the only representations that guarantee the artists the full creative authority on an articulated graphic assets are based on the representation presented in this thesis.

Bibliography

[1] 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, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[2] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, November 2017.

[3] Abhishektha Boppana and Allison P Anderson. Dynamic foot morphology explained through 4d scanning and shape modeling. *Journal of Biomechanics*, 122:110465, 2021.

[4] 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):337–346, 2009.

[5] 3DScan store. `https://www.3dscanstore.com/`, 2022.

[6] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019.

[7] Christopher M Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[8] Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 22(3):587–594, July 2003.

[9] Brett Allen, Brian Curless, and Zoran Popović. Articulated body deformation from range scan data. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 21(3):612–619, July 2002.

[10] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of PEople. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 24(3):408–416, 2005.

[11] Dragomir Angeulov. *Learning models of shape from 3D range data*. PhD thesis, Stanford University, 2005.

[12] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):120:1–120:14, July 2015.

[13] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6233–6242, 2017.

[14] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018.

[15] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6184–6193, 2020.

[16] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, October 2019.

[20] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7214–7223, 2020.

[21] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Pro-*

ceedings of the IEEE/CVF International Conference on Computer Vision, pages 2293–2303, 2019.

[22] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model of people in clothing. In *International Conference on Computer Vision (ICCV)*, pages 853–862, 2017.

[23] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6468–6477, 2020.

[24] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5484–5493, 2017.

[25] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM TOG*, 36(4):73:1–73:15.

[26] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *International Conference on Computer Vision (ICCV)*, pages 5419–5429, 2019.

[27] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7363–7373, 2020.

[28] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. In *AAAI*, pages 12749–12756, 2020.

[29] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6194–6204, 2020.

[30] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *IEEE 3DV*, 2020.

[31] Das statistik-portal. https://de.statista.com/statistik/kategorien/kategorie/video-gaming-esports/, 2022.

[32] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[34] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018.

[35] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[37] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002.

[38] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613, 2020.

[39] Stephan Rössner. Obesity: the disease of the twenty-first century. *International journal of obesity*, 26(4):S2–S4, 2002.

[40] Markos Kalligeros, Fadi Shehadeh, Evangelia K Mylona, Gregorio Benitez, Curt G Beckwith, Philip A Chan, and Eleftherios Mylonakis. Association of obesity with disease severity among patients with coronavirus disease 2019. *Obesity*, 28(7):1200–1204, 2020.

[41] Michael D Jensen. Role of body fat distribution and the metabolic complications of obesity. *The Journal of Clinical Endocrinology & Metabolism*, 93(11_supplement_1):s57–s63, 2008.

[42] Beata Jabłonowska-Lietz, Małgorzata Wrzosek, Marta Włodarczyk, and Grażyna Nowicka. New indexes of body fat distribution, visceral adiposity index, body

adiposity index, waist-to-height ratio, and metabolic disturbances in the obese. *Polish Heart Journal (Kardiologia Polska)*, 75(11):1185–1191, 2017.

[43] D Gordon E Robertson, Graham E Caldwell, Joseph Hamill, Gary Kamen, and Saunders Whittlesey. *Research methods in biomechanics*. Human kinetics, 2013.

[44] James Hay. *The biomechanics of sports techniques*. Prentice-Hall, 1978.

[45] Tom F Novacheck. The biomechanics of running. *Gait & posture*, 7(1):77–95, 1998.

[46] Kenji Masumoto and John A Mercer. Biomechanics of human locomotion in water: an electomyographic analysis. *Exercise and sport sciences reviews*, 36(3):160–169, 2008.

[47] Francesca Sylos-Labini, Francesco Lacquaniti, and Yuri P Ivanenko. Human locomotion under reduced gravity conditions: biomechanical and neurophysiological considerations. *BioMed research international*, 2014, 2014.

[48] Yildirim Hurmuzlu and Cagatay Basdogan. On the measurement of dynamic stability of human locomotion. *Journal of biomechanical engineering*, 116(1):30–36, 1994.

[49] Barbara Heil. Running shoe design and selection related to lower limb biomechanics. *Physiotherapy*, 78(6):406–412, 1992.

[50] EC Frederick. Kinematically mediated effects of sport shoe design: a review. *Journal of sports sciences*, 4(3):169–184, 1986.

[51] A Morecki and K Kdzior. Biomechanical aspects in robotics. In *Theory and Practice of Robots and Manipulators*, pages 17–22. Springer, 1985.

[52] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 459–468, 2020.

[53] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, pages 180–200. Springer, 2022.

[54] SizeUSA. SizeUSA dataset. http://http://www.sizeusa.com/, 2020.

[55] Claire C Gordon, Cynthia L Blackwell, Bruce Bradtmiller, Joseph L Parham, Patricia Barrientos, Stephen P Paquette, Brian D Corner, Jeremy M Carson, Joseph C Venezia, Belva M Rockwell, et al. 2012 anthropometric survey of us army personnel: Methods and summary statistics. Technical report, ARMY NATICK SOLDIER RESEARCH DEVELOPMENT AND ENGINEERING CENTER MA, 2014.

[56] Nadia Magnenat-Thalmann, Richard Laperrire, and Daniel Thalmann. Joint-dependent local deformations for hand animation and object grasping. In *In Proceedings on Graphics interface'88*. Citeseer, 1988.

[57] Nadia Magnenat-Thalmann and Daniel Thalmann. Human body deformations using joint-dependent local operators and finite-element theory. Technical report, EPFL, 1990.

[58] J. P. Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *Pro-*

ceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '00, pages 165–172, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[59] Tsuneya Kurihara and Natsuki Miyata. Modeling deformable human hands from medical images. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 355–363. Eurographics Association, 2004.

[60] Taehyun Rhee, John P Lewis, and Ulrich Neumann. Real-time weighted pose-space deformation on the gpu. In *Computer Graphics Forum*, volume 25, pages 439–448. Wiley Online Library, 2006.

[61] Hyewon Seo, Frederic Cordier, and Nadia Magnenat-Thalmann. Synthesizing animatable body models with parameterized shape modifications. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '03, pages 120–125, 2003.

[62] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–112, 2013.

[63] Oren Freifeld and Michael J. Black. Lie bodies: A manifold representation of 3D human shape. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part I, LNCS 7572, pages 1–14. Springer-Verlag, October 2012.

[64] David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conf. on Computer Vision (ECCV)*, pages 242–255, 2012.

[65] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276 – 286, 2017.

[66] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Trans. Graph.*, 34(4):120:1–120:14, July 2015.

[67] Paul G Kry, Doug L James, and Dinesh K Pai. Eigenskin: real time large deformation character skinning in hardware. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 153–159. ACM, 2002.

[68] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. Sparse localized deformation components. *ACM Transactions on Graphics (TOG)*, 32(6):1–10, 2013.

[69] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[70] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3D faces. In *Siggraph*, volume 99, pages 187–194, 1999.

[71] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3D morphable models. *Int. J. Comput. Vis.*, 126(2-4):233–254, 2018.

[72] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In

2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pages 296–301. Ieee, 2009.

[73] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. pages 1–6, 2008.

[74] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *Eur. Conf. Comput. Vis.*, pages 297–312, 2014.

[75] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3410–3419, 2020.

[76] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *Eur. Conf. Comput. Vis.*, pages 725–741, 2018.

[77] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 601–610, 2020.

[78] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. *ACM TOG*, 24(3):426–433, 2005.

[79] Bogdan Sarghie, Mariana Costea, and Dumitru Liute. Anthropometric study of the foot using 3D scanning method and statistical analysis. In *Proceedings of the*

International Symposium in Knitting and Apparel, Isai, Romania, volume 2122, 2013.

[80] Scott Telfer and James Woodburn. The use of 3d surface scanning for the measurement and assessment of the human foot. *Journal of foot and ankle research*, 3(1):1–9, 2010.

[81] Alfredo Ballester, Ana Piérola, Eduardo Parrilla, Mateo Izquierdo, Jordi Uriel, Beatriz Nácher, Vicent Ortiz, Juan C Gonzalez, A Page, and Sandra Alemany. Fast, portable and low-cost 3D foot digitizers: Validity and reliability of measurements. *Proceedings of 3DBODY. TECH*, pages 11–12, 2017.

[82] Abhishektha Boppana and Allison P Anderson. Dynamo: Dynamic body shape and motion capture with intel realsense cameras. *Journal of Open Source Software*, 4(41):1466, 2019.

[83] Térence Coudert, Pierre Vacher, Cathy Smits, and Marc Van der Zande. A method to obtain 3d foot shape deformation during the gait cycle. In *9th International Symposium on the 3D analysis of Human Movement: 28-30th June 2006 Valenciennes, France*, 2006.

[84] Bryan P Conrad, Michael Amos, Irene Sintini, Brian Robert Polasek, and Peter Laz. Statistical shape modelling describes anatomic variation in the foot. *Footwear Science*, 11(sup1):S203–S205, 2019.

[85] Edmée Amstutz, Tomoaki Teshima, Makoto Kimura, Masaaki Mochimaru, and Hideo Saito. Pca based 3D shape reconstruction of human foot using multiple viewpoint cameras. In *International Conference on Computer Vision Systems*, pages 161–170. Springer, 2008.

[86] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *European Conference on Computer Vision (ECCV)*, August 2020.

[87] Samuel Zeitvogel, Johannes Dornheim, and Astrid Laubenheimer. Joint optimization for multi-person shape models from markerless 3d-scans. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020.

[88] Samuel ZEITVOGEL and Astrid LAUBENHEIMER. An open-source articulated multi-person shape model training and inference pipeline.

[89] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision (3DV)*, pages 421–430, 2017.

[90] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, November 2018. Two first authors contributed equally.

[91] Timo von Marcard, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3D human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, pages 349–360, 2017.

[92] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion

and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, November 2014.

[93] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019.

[94] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018.

[95] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[96] Nadine Rueegg, Christoph Lassner, Michael J. Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, February 2020.

[97] J Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human body shape and pose prediction. *Proceedings of the BMVC, London, UK*, pages 4–7, 2017.

[98] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018.

[99] Dyna dataset. http://dyna.is.tue.mpg.de/, 2015. Accessed: 2015-05-15.

[100] Alexandre Saint, Eman Ahmed, Kseniya Cherenkova, Gleb Gusev, Djamila Aouada, Bjorn Ottersten, et al. 3dbodytex: Textured 3D body dataset. In *2018 International Conference on 3D Vision (3DV)*, pages 495–504. IEEE, 2018.

[101] Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM TOG*, 22(3):587–594, 2003.

[102] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. A statistical model of human pose and body shape. *Comput. Graph. Forum*, 28(2):337–346, 2009.

[103] David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conference on Computer Vision*, volume 7577, pages 242–255, 2012.

[104] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3D human modeling. *Pattern Recognition*, 67:276–286, 2017.

[105] Haoyang Wang, Riza Alp Guler, Iasonas Kokkinos, George Papandreou, and Stefanos Zafeiriou. BLSM: A bone-level skinned model of the human mesh. In *Eur. Conf. Comput. Vis.*, pages 1–17, 2020.

[106] Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. Learning an efficient model of hand shape variation from depth images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2540–2548, 2015.

[107] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. Single image 3D hand reconstruction with mesh convolutions. In *Brit. Mach. Vis. Conf.*, page 45, 2019.

[108] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3D tracking of hand articulations using kinect. In *Brit. Mach. Vis. Conf.*, pages 1–11, 2011.

[109] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K. Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM TOG*, 39(6):219:1–219:14, 2020.

[110] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM TOG*, 35(6):222:1–222:11, 2016.

[111] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *Eur. Conf. Comput. Vis.*, pages 534–551, 2018.

[112] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. FML: Face Model Learning from Videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10812–10822, 2019.

[113] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7763–7772, 2019.

[114] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3D hand shape and

pose from images in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10843–10852, 2019.

[115] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11807–11816, 2019.

[116] Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and IMUs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1533–1547, 2016.

[117] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision (ECCV)*, 2020.

[118] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In *European conference on computer vision*, pages 242–255. Springer, 2012.

[119] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[120] Krystal D'Costa. What makes the human foot unique?, Oct 2018.

[121] Ryan K. Card. Anatomy, bony pelvis and lower limb, foot muscles, Jun 2021.

[122] Dominic James Farris, Luke A Kelly, Andrew G Cresswell, and Glen A Lichtwark. The functional importance of human foot muscles for bipedal locomotion. *Proceedings of the National Academy of Sciences*, 116(5):1645–1650, 2019.

[123] Craig A Lindley. The gameplay gestalt, narrative, and interactive storytelling. In *CGDC Conf.*, 2002.

[124] David Thue, Vadim Bulitko, Marcia Spetch, and Eric Wasylishen. Interactive storytelling: A player modelling approach. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 3, pages 43–48, 2007.

[125] John Truby. *The anatomy of story: 22 steps to becoming a master storyteller*. Farrar, Straus and Giroux, 2008.

[126] Klaus Fog, Christian Budtz, and Baris Yakaboylu. *Storytelling*. Springer, 2005.

[127] Jason Ohler. The world of digital storytelling. *Educational leadership*, 63(4):44–47, 2006.

[128] Shilo T McClean. *Digital storytelling: The narrative power of visual effects in film*. Mit Press, 2007.

[129] Brian Karis. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013.

[130] MetaHuman creator. https://www.unrealengine.com/en-US/metahuman, 2022.

[131] Marc Cavazza, Fred Charles, and Steven J Mead. Interacting with virtual characters in interactive storytelling. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 318–325, 2002.

[132] Michael Seymour, Kai Riemer, and Judy Kay. Interactive realistic digital avatars-revisiting the uncanny valley. 2017.

[133] Scott Spencer. *Zbrush digital sculpting human Anatomy*. John Wiley & Sons, 2010.

[134] Jason Patnode. *Character Modeling with Maya and ZBrush: Professional polygonal modeling techniques*. CRC Press, 2012.

[135] Eric Keller. *Introducing ZBrush*. John Wiley & Sons, 2011.

[136] Ronald N Perry and Sarah F Frisken. Kizamu: A system for sculpting digital characters. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 47–56, 2001.

[137] Amit Zoran and Joseph A Paradiso. Freed: a freehand digital sculpting tool. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2613–2616, 2013.

[138] Mike De la Flor and Bridgette Mongeon. *Digital sculpting with Mudbox: essential tools and techniques for artists*. CRC Press, 2012.

[139] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)*, 34(4):1–14, 2015.

[140] Ahmed A A Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human body model. In *European Conference on Computer Vision (ECCV)*, 2022.

[141] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023.

[142] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

[143] Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. Story-to-motion: Synthesizing infinite and controllable character animation from long text. In *SIGGRAPH Asia 2023 Technical Communications*, pages 1–4. 2023.

[144] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *Proc. International Conference on Computer Vision (ICCV)*, October 2023.

[145] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[146] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *arXiv preprint arXiv:1712.06584*, 2017.

[147] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019.

[148] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 803–812, 2019.

[149] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[150] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11107–11117, Piscataway, NJ, October 2021. IEEE.

[151] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, September 2020.

[152] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3D representations. In *NeurIPS*, 2021.

[153] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022.

[154] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location and pose. In *CVPR*, 2022.

[155] Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3D humans and environments in TV shows. In *ECCV*, 2022.

[156] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa,

and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

[157] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and motion estimation from in-the-wild videos. In *3DV*, 2024.

[158] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Int. Conf. Comput. Vis.*, pages 2282–2292, 2019.

[159] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3D scenes. In *International Conference on 3D Vision (3DV)*, March 2024.

[160] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. Synthesizing diverse human motions in 3D indoor scenes. In *International conference on computer vision (ICCV)*, 2023.

[161] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, , and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European conference on computer vision (ECCV)*, October 2022.

[162] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022.

[163] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3D scenes. In *International Conference on Computer Vision (ICCV)*, October 2021.

[164] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *Eur. Conf. Comput. Vis.*, pages 581–600, 2020.

[165] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023.

[166] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3D tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022.

[167] Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3D objects and their poses from human interactions alone? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023.

[168] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12965–12976, June 2023.

[169] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.