Beyond the Surface:
Statistical Approaches to Internal Anatomy Prediction

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Ing. Marilyn Justine Keller
aus Mulhouse, Frankreich

Tübingen
2024

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

Tag der mündlichen Qualifikation:	29.11.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Michael J. Black
2. Berichterstatter:	Prof. Dr.-Ing. Marc Stamminger
3. Berichterstatter:	Prof. Dr. Thor Besier

To my family

Abstract

The creation of personalized anatomical digital twins is important in the fields of medicine, computer graphics, sports science, and biomechanics. But to observe a subject's anatomy, expensive medical devices (MRI or CT) are required and creating a digital model is often time-consuming and involves manual effort.

Instead, we can leverage the fact that the shape of the body surface is correlated with the internal anatomy; indeed, the external body shape is related to the bone lengths, the angle of skeletal articulation, and the thickness of various soft tissues. In this thesis, we leverage the correlation between body shape and anatomy and aim to infer the internal anatomy solely from the external appearance.

Learning this correlation requires paired observations of people's body shape, and their internal anatomy, which raises three challenges. First, building such datasets requires specific capture modalities. Second, these data must be annotated, i.e. the body shape and anatomical structures must be identified and segmented, which is often a tedious manual task requiring expertise. Third, to learn a model able to capture the correlation between body shape and internal anatomy, the data of people with various shapes and poses has to be put into correspondence. In this thesis, we cover three works that focus on learning this correlation. We show that we can infer the skeleton geometry, the bone location inside the body, and the soft tissue location solely from the external body shape.

First, in the OSSO project, we leverage 2D medical scans to construct a paired dataset of 3D body shapes and corresponding 3D skeleton shapes. This dataset allows us to learn the correlation between body and skeleton shapes, enabling the inference of a custom skeleton based on an individual's body. However, since this learning process is based on static views of subjects in specific poses, we cannot evaluate the accuracy of skeleton inference in different poses. To predict the bone orientation within the body in various poses, we need dynamic data.

To track bones inside the body in motion, we can leverage methods from the biomechanics field. So in the second work, instead of medical imaging, we use a biomechanical skeletal model along with simulation to build a paired dataset of bodies in motion and their corresponding skeletons. In this work, we build such a dataset and learn SKEL, a body shape and skeleton model that includes the locations of anatomical bones from any body shape and in any pose.

After dealing with the skeletal structure, we broaden our focus to include different layers of soft tissues. In the third work, HIT, we leverage segmented medical data to learn to predict the distribution of adipose tissues (fat) and lean tissues (muscle, organs, *etc.*) inside the body.

In conclusion, in this thesis we leverage statistical models and multi-modal data to learn to predict from external body shape: the geometry of the bones, their location and orientation inside the body, as well as the soft tissue distribution inside the body.

Kurzfassung

Die Erstellung von personalisierten anatomischen digitalen Zwillingen ist wichtig für die Bereiche Medizin, Computergrafik, Sportwissenschaft und Biomechanik. Um die Anatomie einer Person zu beobachten, sind jedoch teure medizinische Geräte (MRT oder CT) erforderlich, und die Erstellung eines digitalen Modells ist oft zeitaufwändig und erfordert manuellen Aufwand.

Stattdessen kann die Tatsache zunutze gemacht werden, dass die Form der Körperoberfläche mit der inneren Anatomie korreliert; tatsächlich steht die äußere Körperform in Beziehung zu den Knochenlängen, dem Winkel der Skelettgelenke und der Dicke verschiedener Weichteile. In dieser Dissertation nutzen wir diese Korrelation zwischen Körperform und Anatomie und versuchen, die innere Anatomie allein basierend auf dem äußeren Erscheinungsbild abzuleiten.

Das Erlernen dieser Korrelation erfordert gepaarte Beobachtungen der Körperform und der inneren Anatomie von Personen, was drei Herausforderungen mit sich bringt. Erstens erfordert die Erstellung solcher Datensätze spezifische technische Geräte. Zweitens müssen diese Daten annotiert werden, d. h. die Körperform und die anatomischen Strukturen müssen identifiziert und segmentiert werden, was oft eine langwierige manuelle Aufgabe ist, die Fachwissen erfordert. Drittens: Um ein Modell zu erlernen, das in der Lage ist, die Korrelation zwischen Körperform und innerer Anatomie zu erfassen, müssen die Daten von Personen mit unterschiedlichen Formen und Posen miteinander in Beziehung gesetzt werden. In dieser Arbeit befassen wir uns mit drei Teilen, die sich auf das Erlernen dieser Korrelation konzentrieren. Wir zeigen, dass wir die Skelettgeometrie, die Lage der Knochen im Körper und die Lage der Weichteile allein aus der äußeren Körperform ableiten können.

Im OSSO-Projekt nutzen wir zunächst medizinische 2D-Scans, um einen gepaarten Datensatz von 3D-Körperformen und entsprechenden 3D-Skelettformen zu erstellen. Anhand dieses Datensatzes können wir die Korrelation zwischen Körper- und Skelettformen erlernen, was die Ableitung eines individuellen Skeletts auf der Grundlage des Körpers einer Person ermöglicht. Da dieser Lernprozess jedoch auf statischen Ansichten von Personen in bestimmten Posen basiert, können wir die Genauigkeit der Skelettinferenz in verschiedenen Posen nicht bewerten. Um die Ausrichtung der Knochen innerhalb des Körpers in verschiedenen Posen vorherzusagen, benötigen wir dynamische Daten.

Um die Knochen innerhalb des Körpers in Bewegung zu verfolgen, können wir Methoden aus dem Bereich der Biomechanik nutzen. Deshalb verwenden wir im zweiten Teil statt medizinischer Bildgebung ein biomechanisches Skelettmodell zusammen mit einer Simulation, um einen gepaarten Datensatz von Körpern in Bewegung und ihren entsprechenden Skeletten zu erstellen. In diesem Teil erstellen wir einen solchen Datensatz und lernen SKEL, ein Körperform- und Skelettmodell, das die Positionen der anatomischen Knochen von jeder Körperform und in jeder Pose enthält.

Nachdem wir uns mit der Skelettstruktur befasst haben, erweitern wir unseren Fokus auf die verschiedenen Schichten der Weichteile. Im dritten Teil, HIT, nutzen wir segmentierte medizinische Daten um ein Modell zu trainieren, welches die Verteilung von Fettgewebe (Fett) und magerem Gewebe (Muskeln, Organe, usw.) vorhersagt.

Zusammenfassend befasst sich die Dissertation mit der Nutzung von statistischen Modellen und multimodalen Daten, um aus der äußeren Körperform Vorhersagen über die Geometrie der Knochen, deren Lage und Ausrichtung im Körper sowie die Verteilung der Weichteile im Körper zu erlernen.

Acknowledgments

Five years ago, I dreamt of learning to conduct research, get guidance from skilled mentors, and collaborate with brilliant individuals. I was fortunate to meet Sergi Pujades, who offered me the opportunity to join Perceiving Systems (PS) and introduced me to Michael Black. I am deeply grateful to both of them for providing me with this incredible opportunity. I found at PS many things that I love and value: a passion for science, dedication to excellence, a strong work ethic, respect, kindness, and genuine friendship.

Throughout my PhD, Sergi and Michael were invaluable mentors, guiding me on tracks that led to great papers I am proud of. I owe special thanks to Sergi for his dedicated effort to get the data that made my research possible and for standing by me during those intense deadlines right until the final hour. The camaraderie during these deadlines, where the lab felt like a ship's crew navigating a stormy night, is something I deeply admire about PS.

I was fortunate to have strong support throughout my PhD. I thank Anurag for being my rock during these years and all the friends and colleagues who were there during moments of doubt. Special thanks to Soubhik for his constant presence, Partha for banishing boredom, Omri for his support and humor, and Victoria for her compassionate ear. I also thank my long-lasting friend Florence for leading me on the path into the academic world.

My work would not have been possible without the exceptional technical support I received at PS. I am grateful to Benjamin Pellkofer for the countless times he resolved issues for me almost instantly. I also thank the capture team, especially Markus, for their support during our trial captures and scans. A special thanks to Ahmed for kindly sharing his deep knowledge of body models with me, Arjun for his consistently sound advice, and Nikos for always being ready to help.

I was also lucky to have brilliant desk neighbors whom I will miss dearly: Omid, Yuliang, and Yinghao.

I am grateful to all the sponsors who made my work possible, starting with the CaMeD project and the wonderful collaboration with the Graz team, particularly James and Marcell. Then, IMPRS-IS provided exceptional workshops and valuable opportunities.

I had the privilege of interning at Stanford in the Movement Lab, where I am thankful to Karen Liu for her sharp supervision and to Scott Delp for allowing me to immerse myself in his lab, where I learned so much. I also thank Keenon for his huge help and all my coworkers—Zhaoming, Takara, Tom, and others—for the great conversations and insights

PS has been more than just a workplace; it has been a second family. We cooked,

shared meals, jammed, and went on trips that sometimes turned into memorable adventures. I will always cherish these memories with Partha, Soubhik, Victoria, Shashank, Lea, Omri, and Markos.

And, of course, I thank my parents for teaching me all they knew about the world, always providing me with what I needed to do my best, and trusting me unconditionally. I thank my grandpa - my lifelong coach, my grandma for showing me how to be a woman, and my brother, without whom I would not have grown into the person I am today.

Contents

1 Introduction **1**

2 Background **7**

2.1 The human anatomy . . . 7

2.1.1 Skeleton . . . 7

2.1.2 Soft tissues . . . 9

2.2 Capture modalities . . . 10

2.2.1 External appearance . . . 10

2.2.2 Medical Imaging. . . . 13

2.2.3 Motion . . . 14

2.3 Data processing . . . 19

2.4 Modeling the human body . . . 20

2.4.1 Graphical models . . . 22

2.4.2 Human pose representation . . . 27

2.4.3 Biomechanical models . . . 27

2.4.4 3D anatomical atlases . . . 33

3 Inferring the skeleton geometry **35**

3.1 Introduction . . . 36

3.2 Related Work . . . 39

3.3 Data . . . 40

3.3.1 Skin and skeleton masks from DXA images . . . 41

3.3.2 Computing landmarks from the silhouettes . . . 42

3.3.3 Skin surface from DXA . . . 47

3.3.4 Skeleton model based on Stitched Puppet . . . 47

3.3.5 Skeleton from DXA . . . 48

3.4 Method – OSSO . . . 52

3.4.1 Skeleton statistics and correlation to skin shape . . . 52

3.4.2 Inferring the skeleton from the skin . . . 54

3.4.3 Reposing the inferred skeleton . . . 55

3.5 Experiments . . . 56

3.5.1 DXA to 2D landmarks regressor . . . 56

3.5.2 STAR fits to DXA skin masks . . . 56

3.5.3 Skeleton fits to DXA skeleton masks . . . 60

3.5.4 Skin to 3D landmark regressors . . . 60

3.5.5 Skeleton prediction evaluation on 2D DXA bones masks 60
3.5.6 OSSO vs Anatomy Transfer comparison 65
3.5.7 Generalization to new poses 65
3.6 Conclusion 65

4 Inferring the location and orientation of the bones 73
4.1 Introduction 74
4.2 Related Work 75
4.3 Method overview 77
4.4 The BioAMASS dataset 77
4.4.1 The SMPL body surface model 78
4.4.2 The BSM skeletal model 78
4.4.3 Fitting BSM to SMPL 80
4.5 The SKEL model 82
4.5.1 Establishing the bone locations and orientations 85
4.5.2 Building SKEL: A single rig for skin and bones 87
4.6 Evaluation 90
4.6.1 Evaluating BioAMASS fits 90
4.6.2 Joint regressors 90
4.6.3 SKEL fits to SMPL 93
4.6.4 Qualitative comparisons with OSSO 94
4.6.5 Disentangling body shape and bone lengths 94
4.7 Discussion and Conclusion 96

5 Inferring the soft tissues inside the body 101
5.1 Introduction 103
5.2 Related work 105
5.3 Human Tissue Data 106
5.3.1 MRI segmentation 107
5.3.2 Two-step SMPL fit 110
5.3.3 Human Implicit Tissues (HIT) dataset 112
5.4 HIT method 115
5.4.1 HIT spaces 120
5.4.2 HIT architecture 120
5.4.3 Training, losses and sampling 122
5.5 Experiments 125
5.5.1 Tissues prediction evaluation 125
5.5.2 Comparison with OSSO 130
5.5.3 Generalization to new body shapes and poses 130
5.6 Discussion and Conclusion 131

6 Conclusion and Future Work 135

A OSSO **139**
A.1 OSSO 3D landmark regression errors 139

B HIT **143**
B.1 HIT MLP architecture 143
B.1.1 $\mathcal{B}$ MLP 143
B.1.2 $\mathcal{W}$ MLP 143
B.1.3 $\mathcal{C}$ MLP 143
B.1.4 $\mathcal{T}$ MLP 145

Bibliography **147**

Chapter 1

Introduction

In the realm of science, progress often hinges on the development of models. These theoretical models enable us to explain what we observe and forecast future phenomena. In the case of the human body, creating accurate models involves capturing and measuring the body, which can be done in various ways.

In medicine, patients undergo imaging to capture the internal anatomy. In biomechanics, markers are placed on a subject's body to track motion and measure limb angles. Meanwhile, special effects artists use body scans to create digital replicas of actors. Although these methods serve different purposes, they all rely on capturing different aspects of the body: anatomy, motion, appearance, etc.

While indispensable, these capture modalities are expensive, time-consuming, and require trained operators. Moreover, the captured data comes in various formats, such as 3D point clouds, 3D meshes, and 2D or 3D images. Using this data requires post-processing, such as cleaning and segmentation, as well as an expert to analyze them.

In this thesis, we start with the following observation: the easiest way to capture people is vision. Nowadays, cameras are very accessible and advances in computer vision have made it possible to automatically extract various information from images and videos. Algorithms have been developed to recognize people, count them, estimate their pose in the image, capture their body shape, track their motion, identify their actions, etc.

A limitation of images is that they only capture what is visible. We can attempt to predict what is hidden but this requires priors about the world. If a subject's arm is occluded, there are only so many plausible arm poses, given the way the person stands. If a person faces the camera with a green shirt, the back of the shirt is also probably green. Following this idea, if we have adequate priors about human anatomy, we may predict what is happening inside the body solely from external observations.

There is a logical correlation between how we look and our anatomy. The size of our bones is correlated with the length of our limbs, and our fat percentage is somewhat correlated with our waist size. In this thesis, we aim to answer this specific question: *given the body shape of a person, can we predict the location of the bones and soft tissues inside?*

Applications. Inferring the anatomy has applications in computer graphics, medicine, and biomechanics. One of the goals of computer graphics is to create digital humans that look realistic. Artists need to know human anatomy to draw realistic proportions and reproduce the subtlety of the curves of human limbs. For digital animation, designing the underlying anatomy of a body (its skeleton, muscles, and fat) enables simulating the body's deformation in motion, taking realism to the next level.

In medicine, capturing the inside of the body is a costly process. Medical imaging devices are expensive and require trained operators. Moreover, some imaging techniques, such as X-ray, emit radiation. While low levels of ionizing radiation are part of everyday life, increased exposure through medical use raises the risk of health hazards, such as cancer. Thus, the usage of such techniques should be limited. By inferring the internal anatomy from the body shape, we can reduce the need for such invasive capturing methods and make health care more accessible. Potential applications are estimating the location of the bones inside the skeleton for imaging a specific part of the body or estimating the distribution of fat in the body for monitoring the patient's health.

Biomechanics seeks to understand the mechanical properties of living organisms. Since human motion relies on the relative movement of bones, accurately measuring their location and orientation within the body is essential for analyzing human movement and joint mechanics. Current methods to measure bone location are often marker-based. These systems involve placing markers on the skin, which is time-consuming and has limited precision. Markers attached to the skin can slide relative to the underlying bones and ensuring repeatability between subjects is challenging. The capture systems to track the markers are also expensive. They require numerous cameras and controlled lightning. Being able to accurately predict the location of the bones from videos, just by looking at the body shape, would enable biomechanical analysis outside the lab and make it more accessible.

Challenges. Machine-based inference requires using specific representations, i.e. how to enter "body shape", "bones" or "fat layer" in a computer. In this thesis, we represent the body shape by a 3D mesh, which is the approach used in video games to represent virtual humans. The same representation is used for the bones. As for the fat, we represent it by a mathematical function that, given a point inside the body, outputs whether it is fat or not. Given these representations, we can train machines to predict what is beyond the body surface.

Machine learning also requires data, and capturing people's bones and soft tissues is not trivial. Medical imaging techniques come with various limitations: in some cases, they only yield 2D images, show only specific parts of the body, capture the body in a particular position, and require annotations by an expert to be useful. Consequently, recovering the 3D anatomy from a whole subject in an arbitrary pose is very challenging. In this thesis, we exploit several modalities yielding different data types. Specifically, we work with medical scans such as Dual-energy X-ray Absorptiometry (DXA) and

Magnetic Resonance Imaging (MRI). From these modalities, we can accurately measure the body shape and pose, and the location and shape of the bones and soft tissues. Since medical imaging techniques constrain the body in a specific pose, they are not suitable for capturing motion. Instead, we use marker-based motion capture to estimate the pose of the bones in the body.

Finally, machine learning requires a specific normalization of the data. Learning the anatomy from a collection of 3D images is a very unconstrained problem. Yet, we can rely on some postulates. First, the shape of the human body has common features across the population. For example, most people share the same number of bones. Second, if we capture a subject with two different poses, we know that the shape of the bones should not change across poses. So, modeling the variability of the skeleton shape and how the bones articulate may help normalize the data. Previous works have modeled how the shape of the body varies across the population and poses. The idea that makes this whole thesis possible is leveraging such models to disentangle the influence of the body shape and the body pose in predicting the internal anatomy.

Overview. This thesis is structured as follows. In Chapter 2, we provide a background on human anatomy and review the existing methods to capture and model the human body.

In Chapter 3, we address the problem of inferring the shape of the anatomical skeleton of a person in a static pose, i.e. we predict the inside (bones) from the outside (body shape). Learning such a predictor requires data showing both the skeleton shape and the body shape of different subjects. We thus use 2D medical images called DXA, which yield such pairs. We leverage parametric meshes to lift this data to 3D, and then use this 3D data to learn the correlation between the external body shape and the shape of the bones, as well as the location of the bones. We name the learned predictor OSSO (Obtaining Skeletal Shape from Outside). OSSO can predict the shape and location of the bones with reasonable accuracy, and we can generalize the prediction to unseen poses, assuming that the distance between the surface of the bones and the body surface should remain constant in different poses.

Medical images, like the ones used for training OSSO, only show the skeleton in a static pose. Typically, we get one scan per subject, so it is hard to generalize to unseen poses. Questions arise like: how does the head of the humerus move when the arm is raised, how do the hip bones move when the subject is sitting, etc.? While OSSO can yield a guess, the resulting posed skeleton is hard to validate as we do not have ground truth, and OSSO does not guarantee realistic interfaces between the bones.

Inferring the precise orientation of the bones *in any pose* is the goal of Chapter 4. In that chapter, we start with two postulates. First, the orientation of the bones inside the body is correlated with the visible limb orientation. Second, the orientation of the bones is constrained by the degrees of freedom of the anatomical skeleton. Indeed, human limbs can only articulate in specific ways depending on the skeleton articulations. We use these two postulates to build a dataset of bodies in motion with anatomical skeletons inside. This dataset is made of sequences of bodies in motion, generated from motion capture data. Then, we create an anatomical skeleton model with constrained articulations and align it inside the body on each frame. This gives us, for each frame, the external body's limb orientations, along with the corresponding bone orientations inside. From this dataset, we learn a model that i) captures the dependency between the body shape and the bone orientations and ii) restricts the body motions given the anatomical constraints imposed by the bone orientations. We name this model SKEL and demonstrate its use in predicting the precise orientation of the bones inside a body of any shape or pose.

While in Chapter 3 and Chapter 4 we focus on the bones, the nature of the tissues between the bones and the body surface also influences how the body deforms. In Chapter 5, we address the problem of inferring the location of soft tissues inside the body. Contrary to the body surface and the bones, the soft tissue structure and shape have more variability across the population, making it hard to represent them with a 3D mesh. In consequence, we represent the soft tissues implicitly, as a classification function that, given a point inside the body, outputs whether it is fat, muscle, or bone.

In chapters Chapter 3 and Chapter 4, we learn to predict the anatomy from respectively 2D medical images and 3D motion capture data. However, learning a volume requires volumetric data. So in Chapter 5, we leverage MRI scans, which are 3D medical images. To utilize these images effectively, they must be annotated. Therefore, we train a neural network to classify each voxel of the 3D images into subcutaneous adipose tissue, lean tissues (muscles and organs), or bone tissues. We also measure the 3D shape of the subjects from the MRI scans, constituting a dataset of body shapes paired with their internal tissues. We use this dataset to learn a model that, from the body shape, predicts the location of the tissues inside. We name the learned predictor HIT (Human Internal Tissue) and show that it can predict the location of the soft tissues in the body.

Through these three chapters, this thesis shows that relevant anatomical information can be predicted solely from the body shape. The key to this work is capturing anatomical data that can be used as ground truth, using appropriate representations for these anatomical structures, and leveraging existing human body models as priors. Given a body shape, we can infer the bone shapes, the bone orientations, and the location of lean and adipose tissues inside.

Publications This thesis covers the three following articles:

Marilyn Keller, Silvia Zuffi, Michael J. Black and Sergi Pujades, "OSSO: Obtaining Skeletal Shape from Outside". In Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2022.

Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, Michael J. Black, "From Skin to Skeleton: Towards Biomechanically Accurate 3D Digital Humans". In ACM Transactions on Graphics (TOG). 2022. **Honorable Mention for the Best Paper Award.**

Marilyn Keller, Vaibhav Arora, Abdelmouttaleb Dakri, Shivam Chandhok, Jürgen Machann, Andreas Fritsche, Michael J. Black, and Sergi Pujades, "HIT: Estimating Internal Human Implicit Tissues from the Body Surface". In Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2024.

During my thesis, I also contributed to these publications, which are not covered in this manuscript:

Di Meng, Marilyn Keller, Edmond Boyer, Michael Black, Sergi Pujades, "Learning a Statistical Full Spine Model from Partial Observations". In International Workshop on Shape in Medical Imaging (MICCAI). 2020.

Marilyn Keller, Marcell Krall, James Smith, Hans Clement, Alexander M. Kerner, Andreas Gradischar, Ute Schäfer, Michael J. Black, Sergi Pujades, "Optimizing the 3D Plate Shape for Proximal Humerus Fractures". In Proceedings of the 56th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2023.

Abdelmouttaleb Dakri, Vaibhav Arora, Léo Challier, Marilyn Keller, Michael J. Black, and Sergi Pujades, "On Predicting 3D Bone Locations Inside the Human Body". In Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2024.

Chapter 2

Background

In this chapter, we present some background for our work. First we will take a look at basic human anatomy, then explore the various methods used to capture the human form and process this data. Finally, we will present the related works on human modeling.

2.1 The human anatomy

In this thesis, we only discuss specific parts of human anatomy: the bones and soft tissues. We start by briefly introducing these.

2.1.1 Skeleton

Bone tissues. The human skeleton consists of bones and cartilage.

Bones are hard living tissues that comprise the bulk of our skeleton. They support the body, safeguard vital organs, and constitute the mechanical basis for movement [1].

Cartilage, meanwhile is a semi-rigid form of connective tissue, crucial for parts of the skeleton that demand extra flexibility, such as the areas where ribs connect to the sternum [1].

Bone shape. In humans, like in other species, the shape and size of bones varies among the population. The three main variations are due to age, gender, and geographic region. But the variation also exists between individuals of the same group [2]. As an example, Fig. 2.1 illustrates the variability of the human humerus [3].

The bones can be grouped in different ways. In this thesis we will often refer to the *long bones*, which are the tubular bones in the arms and legs, i.e. the humerus in the forearm, the radius and ulna in the lower arm, the femur in the thigh and the tibia and fibula in the lower leg (Fig. 2.2).

Degrees of freedom. The body moves by muscles acting directly or via tendons on the bones. The limb motions are constrained along specific degrees of freedom and denoted by specific names depending on the nature of the articulation. Thus, we talk about knee

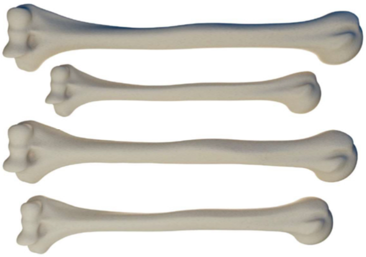


Figure 2.1: Humerus of different subjects. Note the variation in size and twist.

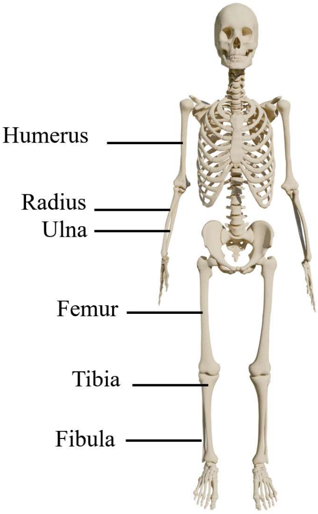


Figure 2.2: The long bones of the human skeleton.

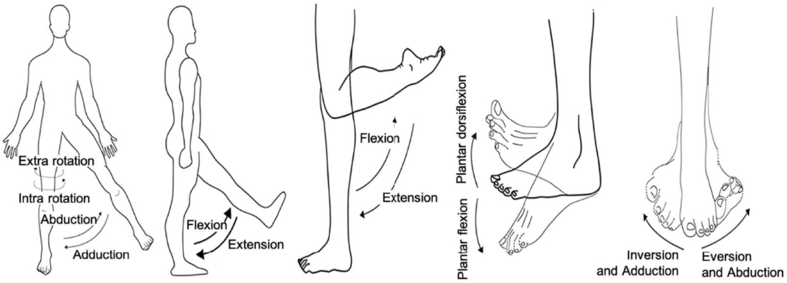


Figure 2.3: The degrees of freedom of the lower limb. (Figure from [4])

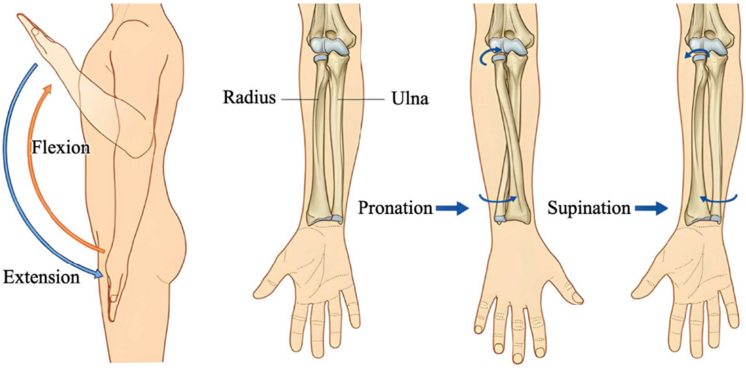


Figure 2.4: Degrees of freedom of the forearm. (Figure from [5])

flexion and **extension** for bending the knee, **abduction** when a limb goes away from the sagittal plane, and **rotation** for the motion of a limb around its axis (Fig. 2.3).

Some articulations like the forearm are more complex than a ball joint, like the rotation of the hand palm towards palm up and down, which is called **pronation** and **supination** (Fig. 2.4)[1].

2.1.2 Soft tissues

In addition to bone tissues, the human body comprises other tissues. In this thesis, we consider three categories of soft tissues that we denote (LT, SAT, IMVAT) and detail

[1] For a more exhaustive list of naming conventions, see Chapter 6 of [2].

below. The muscles and organs are denoted as Lean Tissue (LT). The Subcutaneous Adipose Tissue (SAT) and the Intra-Muscular and Visceral Adipose Tissue (IMVAT) denote the human fat; SAT is located directly under the skin, whereas IMVAT is located inside the muscles and around the organs.

In the medical field, monitoring these tissues is crucial due to their association with various diseases. For instance, an excess of fat with respect to lean tissue is correlated with health risks such as the development of type-II diabetes and cardiovascular diseases [6, 7].

2.2 Capture modalities

Various methods are employed to capture different aspects of the human body. Color cameras and depth sensors are typically used to record the external appearance, offering detailed surface information. Meanwhile, the dynamics of human motion are captured using motion capture systems and inertial sensors, enabling precise tracking of body movements. Those systems can not capture the inside of the body, so for this task, medical imaging techniques such as X-ray, MRI and CT scans are needed. In this section we present those different modalities.

2.2.1 External appearance

The external appearance is usually captured using different camera technologies.

Camera sensors. Color cameras are a straightforward way to capture humans in motion. Recent progress in computer vision has made it possible, from an image or a video, to segment the human body [8], track its motion [9], and even predict its 3D shape [10]. However, due to the 2D nature of images, it is difficult to recover the 3D information accurately, such as distances, speeds, or angles.

Depth sensor. Depth sensors add depth information to the RGB value of a pixel, making it easier to recover the geometry of the observed scene. By leveraging depth measurements, 3D scanners can capture the 3D shape of the body, yielding a textured 3D mesh of the captured subject. Those meshes are usually high resolution and noisy, i.e. their topology is not clean and regular, and they have holes.

Fig. 2.5 shows an example of such a scanner, and Fig. 2.6 its output. With a 3D scanner, like with simple cameras, the capture can be done over time, yielding one detailed textured mesh per frame. Thus, 3D scanners are the best option for precisely capturing the 3D shape of a body.

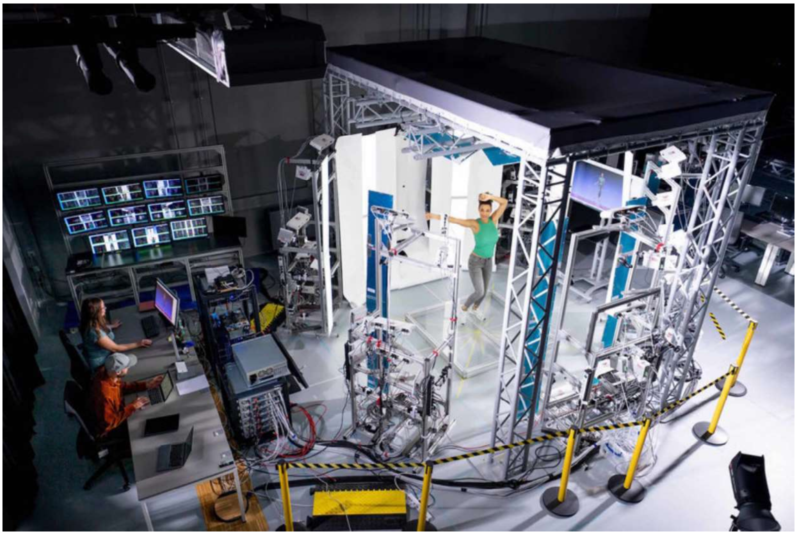


Figure 2.5: The 3dMD 4D scanner at MPI-IS Tübingen. It uses 22 pairs of stereo cameras, 22 color cameras, and speckle-light projectors and can capture the full 3D human body shape at 60 frames per second.

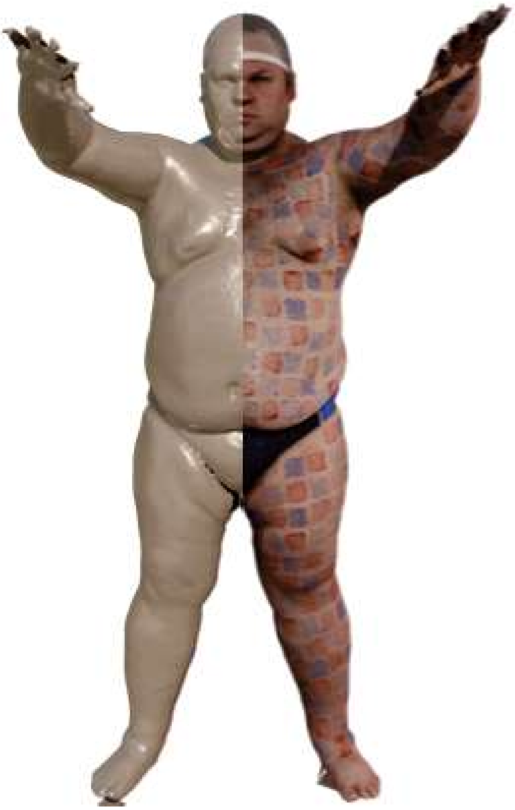


Figure 2.6: Mesh output by a 3D scanner. The left part of the image shows the mesh without texture.

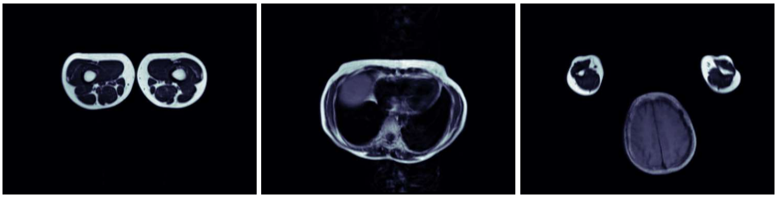


Figure 2.7: Slices from an MRI scan. We can see transversal slices of the legs, trunk, head and arms.

2.2.2 Medical Imaging.

Capturing the inside of the body is challenging as it requires specific technologies gathered under the name of Medical Imaging.

Standard medical imaging techniques are described in detail in [11]. Here we will present them from the perspective of computer vision.

Magnetic Resonance Imaging (MRI). Magnetic Resonance Imaging (MRI) technology enables doctors to examine the interior of the body without relying on ionizing radiation. These systems use strong magnetic fields and offer superior contrast for imaging soft tissues [11].

The resulting MRI scan is a 3D array of float values that need to be interpreted by a medical expert. Often, we visualize specific slices of the MRI as a 2D grayscale image (Fig. 2.7).

X-rays. Bones are hard to distinguish and segment with MRI, so in case of fracture, patients typically undergo radiography to assess the damages and plan the surgery.

Radiography, a widespread form of X-ray imaging, creates two-dimensional images by measuring X-rays' attenuation through the body. X-rays enable us to image the bones clearly. They can then be segmented using traditional image processing techniques. Fig. 2.8 shows an example of a broken arm X-ray.

Computed Tomography. For more complex diagnoses, Computer Tomography (CT) can be used. CT scans are a series of X-ray images taken from different angles around the body. The images are then combined to create a 3D image of the captured zone.

X-ray and CT scans use ionizing radiation. While everyone is naturally exposed to radiation daily from our surroundings, additional exposure, even if minor, can marginally elevate the likelihood of cancer development in the future[2]. For this reason, the use of

[2] https://www.radiologyinfo.org/en/info/safety-xray

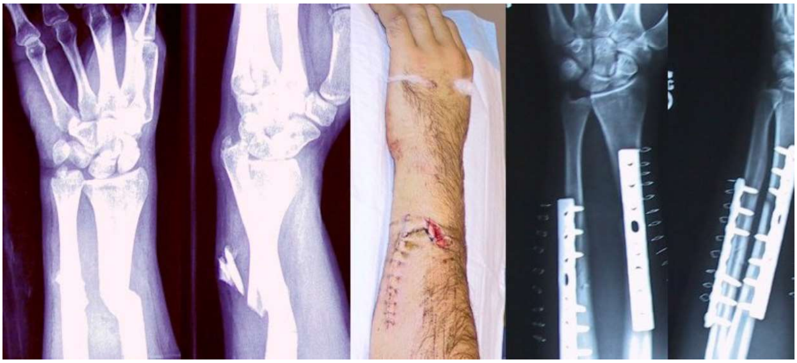


Figure 2.8: Left: Two X-ray views of a broken arm. Middle: A picture of the arm after surgery. Right: Two X-ray views of the arm after surgery. [12]

X-rays is limited to necessary cases, and the capture volume is limited to the body part to diagnose. Fig. 2.9 shows a comparison between an MRI and a CT scan.

DXA. For full-body X-rays, a less ionizing alternative exists. Dual-energy X-ray absorptiometry scans (DXA or DEXA) measure bone density, including the thickness and strength of bones, by transmitting both high and low-energy X-ray beams (a type of ionizing radiation) through the body. This technique is typically used for measuring bone density and body composition.

The level of radiation utilized in DEXA scans is much less than that in standard X-ray examinations, and even less than a day's exposure to natural background radiation[3].

Overall, medical imaging devices are most of the time heavy, massive, expensive, and hard to access in many parts of the world. This is in contrast to standard RGB cameras, that nowadays are in every pocket. Here lies one of the main motivations for this thesis: inferring the anatomy from the visual appearance is a key challenge to reduce the need for medical imaging.

2.2.3 Motion

As seen before, estimating 3D measurements from a video is challenging, and medical imaging is mostly done in static poses, constrained by the restrained acquisition volume, the long acquisition time, or radiation emissions. So to capture human motion, specific techniques have been developed.

[3] https://www.radiologyinfo.org/en/info/dexa

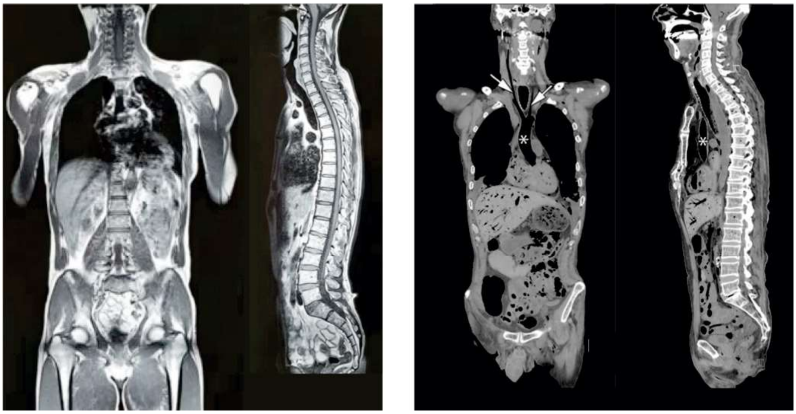


Figure 2.9: Frontal and sagittal slice of an MRI (left) and a CT scan (right). Note the higher contrast of the bones in the CT scan. Figures are from [13] and [14].

Motion Capture, also known as **mocap**, is the task of capturing 3D human motion. It consists of measuring the orientation of the limbs over time. To achieve this, the most accurate method is **marker-based motion capture**. Markers are placed on strategic locations on the body like the wrist, knees, etc. (Fig. 2.10). Those markers are then tracked using infrared cameras, and the orientation of the limbs is recovered using inverse kinematics. Fig. 2.11 shows a mocap setup. This one is equipped with 54 cameras, that are able to accurately track markers at 500 Hz.

Marker-based motion capture is an expensive process, both in terms of equipment and time (Fig. 2.12). The capture system needs to be calibrated, the markers precisely placed on the subject, and eventually, the captured data needs to be cleaned and processed. The result of this process is a 3D skeleton model with adjusted limb scales, and pose varying in time. This moving skeleton can then be used to animate a 3D character or to analyze the motion by measuring the variation of joint locations and angles over time.

To reduce the overhead, several works have attempted **markerless motion capture** [15–17]. For example, OpenCap [15] enables capturing a motion from two smartphones. **Video-based motion capture** techniques rely on recent progress in computer vision on estimating the pose of an individual from a single image or a video. The pose estimation is typically done by running a 2D joint detector on the image [18–20]. Specifically, OpenCap uses OpenPose [18] to detect the subject's 2D joint locations in several camera views, to then reconstruct their 3D locations.

However, existing 2D joint detectors [18–20] have limited accuracy since they are typically trained using manually annotated 2D images. The annotators have to eyeball

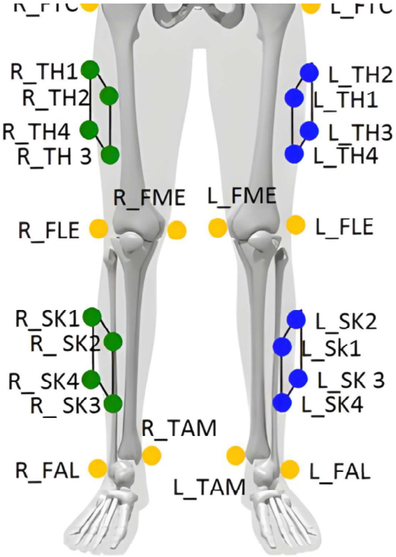


Figure 2.10: Marker set used by the motion capture software Qualisys for gait analysis.

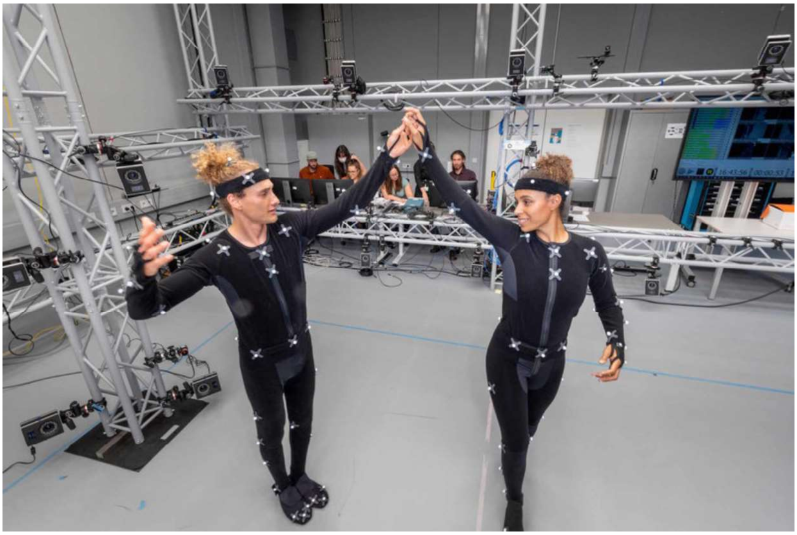


Figure 2.11: Motion capture system.

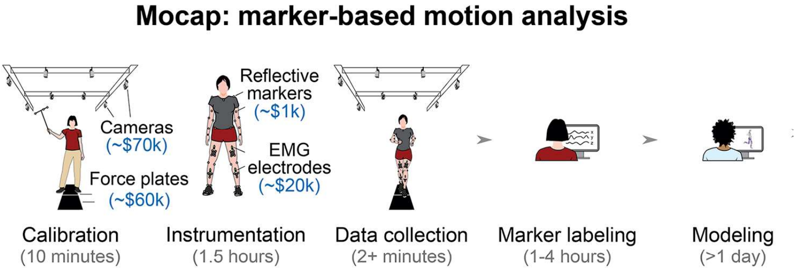


Figure 2.12: Pipeline of acquisition of motion capture data. In the biomechanics field, force plates are often required to enable dynamic simulation. EMG electrodes are used when using Neuromusculoskeletal models (introduced in Sec. 2.4.3). (Figure from [15])

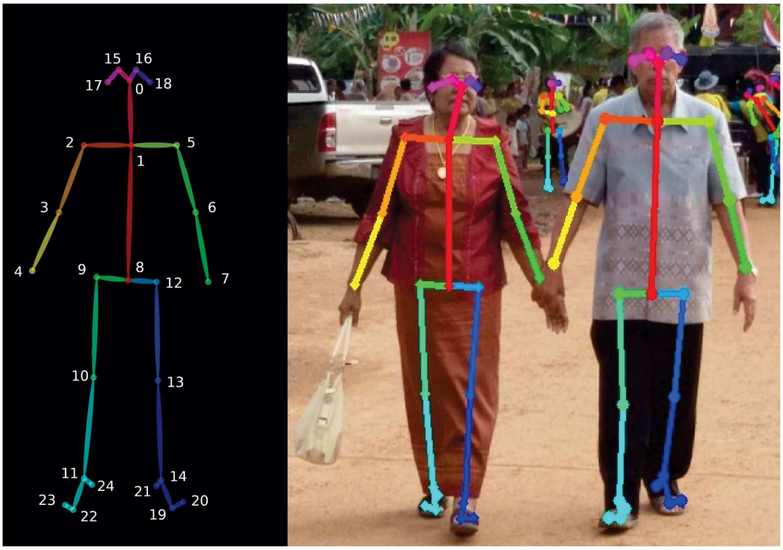


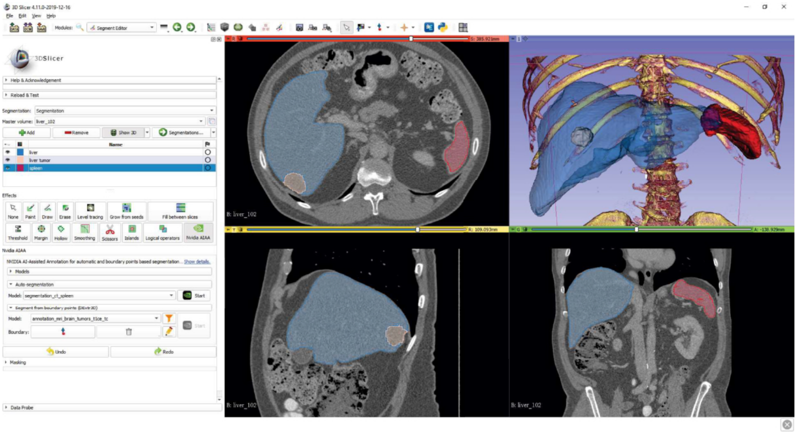
Figure 2.13: OpenPose predicts joint locations from images. These joints correspond to a simplified skeleton and do not match anatomical skeleton joint locations like the humerus and femur head, etc. (Figure from [21])

.

the human joint location on clothed subjects images, which yields approximate joint locations that do not correspond to the human anatomical skeleton. Fig. 2.13 shows an example of the joint locations predicted by OpenPose.

When the joints predicted from images are compared to joint locations computed from motion capture systems [22], the differences are as high as 30 to 50 mm for joints such as the knee. For this reason, in OpenCap, the biomechanical skeleton model is not directly fit to the predicted joint locations. The joint locations are used to infer marker locations (this is done using a neural network trained on real data). Then the skeleton model is fit to these markers.

Another alternative to marker-based motion capture is to use inertial sensors [23]. This setup is lighter but the measurements are less precise; we will not cover them here.



3D Slicer 4.11.0-2019-12-16

Modules: Segment Editor

Segmentation: Segmentation

Master volume: liver_102

NVIDIA AI-Assisted Annotation for automatic and boundary points based segmentation... Show details.

Model: segmentation_ct_spleen

Model: annotation_mri_brain_tumors_t1ce_tc

Figure 2.14: Interface of Slicer3D.

2.3 Data processing

Human data usually needs some processing to be useful. In this thesis, we extensively use segmentation and registration, which we present in this section.

Segmentation. Medical scans usually require the expertise of a medical doctor to be interpreted, and the segmentation protocols often require manual work.

Medical image viewers like Slicer3D[4] (Fig. 2.14) have segmentation tools to assist the segmentation of specific tissues or organs, but this is often done in a semi-automatic way.

Neural networks have shown growing abilities for segmentation, like nnU-Net [24].

In this thesis, we segment different tissues in the body. In Chapter 3, we segment silhouettes on 2D scans using traditional computer vision algorithms like normalization, thresholding, and morphological operations, while in Chapter 5, we use a deep learning approach, nnU-Net, to segment soft tissues in MRIs.

Registration. Capture modalities like 3D scanners yield 3D meshes. But the raw data is hard to work with as it has a lot of triangles, the mesh topology is not regular and the mesh can have holes or self-intersections.

Registration or *alignment* refers to deforming a template to align it to a target. In the scope of 3D bodies and medical imaging, this is a key challenge, as to learn from datasets, data samples must be aligned in a meaningful way. Registration provides a

[4] https://www.slicer.org/

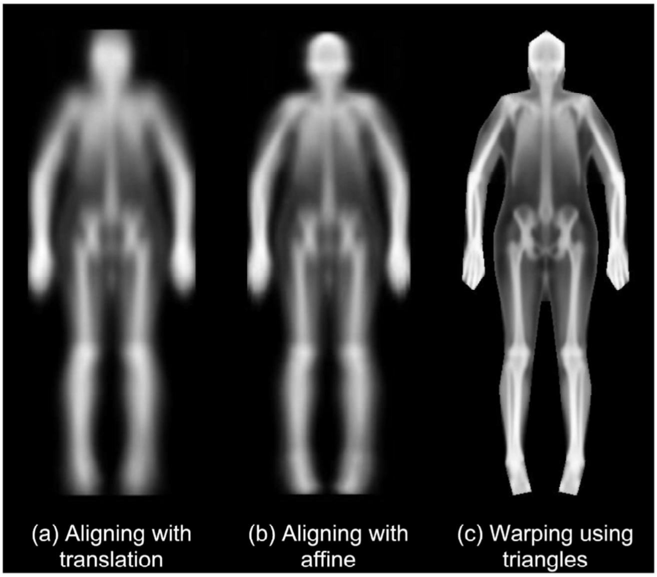


Figure 2.15: Example of the importance of canonicalization. The left image shows an average medical scan computed from several subjects. Because the shape and pose of the subjects vary, the mean image is a useless blurry image. On the right, the subjects' limbs are warped to be all aligned before computing the mean image, and the resulting mean image makes the average skeleton appear. (Figure from [25])

.

way to canonicalize data to make them comparable (see Fig. 2.15 for an example of data canonicalization).

Registration also enables the establishment of correspondence between different meshes. It enables the transfer of information annotated on a template to a target piece of data, such as semantic segmentation, texture, UV maps, etc. See Fig. 2.16 for an example of the SMPL body mesh registered to a body scan (we introduce SMPL in more depth in Sec. 2.4.1).

2.4 Modeling the human body

In this section, we present different models of the human body that have been created in past works. We start from models used for graphical applications and go through

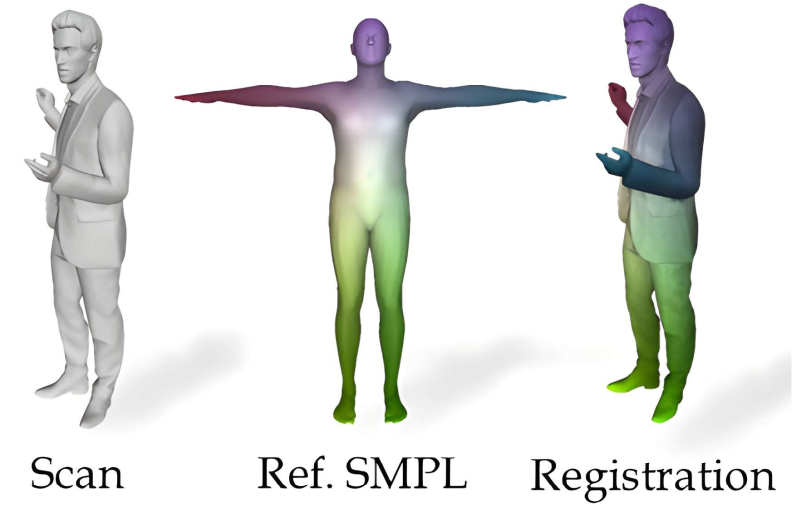
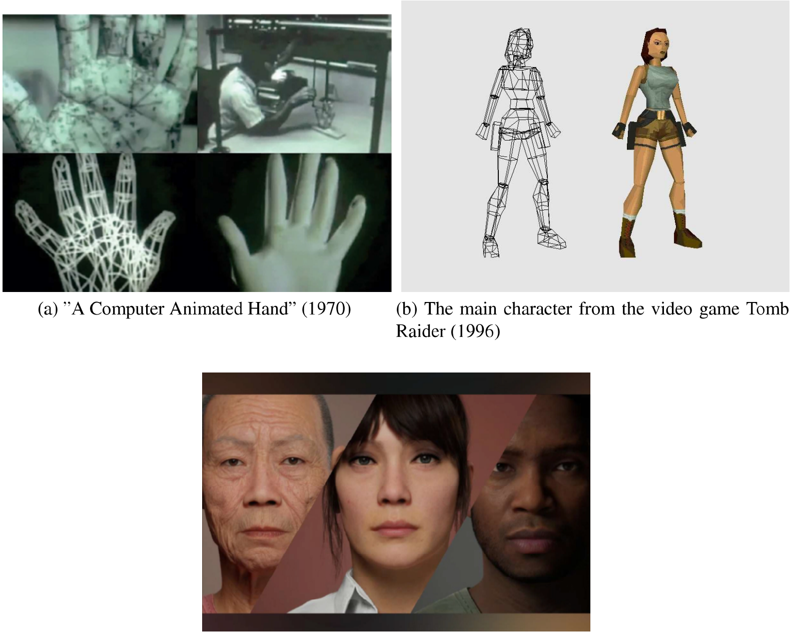


Figure 2.16: A 3D body scan with its SMPL mesh alignment. The mapping yields a correspondence between the SMPL template, which has an average shape and is in T-pose, with the 3D scan, which exhibits geometric details like facial features and clothing details on the surface, and a specific pose (Figure from [26]).



(a) ”A Computer Animated Hand” (1970)

(b) The main character from the video game Tomb Raider (1996)

(c) Modern photorealistic digital human (MetaHuman, 2021)

Figure 2.17: Evolution of computer graphics for human modeling from the 1970s to 2021.

biomechanical models to end with anatomical models.

2.4.1 Graphical models

The evolution of human body modeling in computer graphics has been remarkable. From the first computer-generated 3D humanoids in the early 1970s to today’s photorealistic digital humans, the progress has been significant. Early models were simple wireframes [27], followed by shaded 3D models in the 1970s [28]. The 1990s saw a leap forward with more detailed, textured models in films like Toy Story. Today, the special effects industry can produce nearly indistinguishable virtual humans [29]. Fig. 2.17 illustrates this evolution, showing examples from different eras of computer graphics.

Mesh based models

Traditionally, computer graphics relies on mesh-based models to represent the surface of the human body. A base mesh is rigged to a kinematic tree (Fig. 2.18) to be animated. In the animation community, this kinematic tree is called *skeleton* and is made of ball joints and rigid segments. The mesh is then attached to this skeleton through a process called skinning [30], usually through linear blend skinning (LBS). In LBS, each vertex of the mesh is rigged to the different bones with specific weights. The position of a posed vertex is then computed as a weighted sum of the bone transformations. Fig. 2.18 shows a mesh before and after posing, with the underlying kinematic tree and the skinning weights that rig the thigh vertices to the thigh bone.

Given a mesh rigged to a kinematic tree made of N_j bones, the LBS equation gives us the location of the posed vertices $\mathbf{v}(\theta) \in \mathbb{R}^{N_v \times 3}$ posed with the parameters $\theta \in \mathbb{R}^{N_j,3}$:

$$\mathbf{v}(\theta) = \sum_{i=1}^{N_j} W_i G_i(\theta)\mathbf{T} \tag{2.1}$$

where $\mathbf{v}$ are the vertices of the posed mesh, $\mathbf{T} \in \mathbb{R}^{N_v,3}$ is the template mesh, $G_i \in \mathbb{R}^{N_j \times 3 \times 3}$ is the transformation matrix of the i-th bone, depending on angles θ_i, and W_i is a $N_v \times N_j$ matrix of skinning weights indicating how the N_v vertices of the initial mesh are affected by each rigid transformation i.

This rigging method is straightforward and fast to compute, but has some limitations. For instance, it cannot represent the bulging of the muscles when the limbs are bent. To model those deformations, corrective blend shapes are used [32]. These are pose-dependent additive offsets applied to the mesh to make the posed limbs look more realistic, especially at the articulation level. The skinning equation becomes then:

$$\mathbf{v}(\theta) = \sum_{i=1}^{N_j} W_i G_i(\theta)\mathbf{T} + B_P(\theta) \tag{2.2}$$

where $B_P(\theta) \in N \times 3$ is per vertex corrective blend shapes offsets. These are either crafted by hand or learned from data, and can be applied to the posed mesh or to the template mesh.

Fig. 2.19 shows the difference between the leg flexion with and without corrective blend shapes. Adding corrective blend shape offsets makes the thigh surface more realistic, cancelling out the candy wrap effect visible on the left image.

The mesh representation describes the human body by its surface, through a set of connected triangular faces. Such a mesh can be efficiently rendered into an image, so this representation is widely used in computer graphics and video games.

However, such models are usually not used for medical and biomechanical applications. Indeed, human 3D models often have a simplified kinematic tree that does not model the actual human skeleton. The joints used are mostly ball joints, which offer

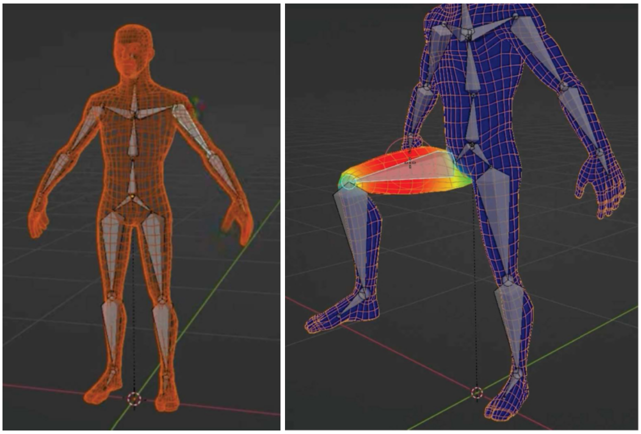


Figure 2.18: Illustration of mesh rigging. Left: A base mesh is created and rigged to a kinematic tree. Right: Each body part is rigged to a bone to be posed. Here, the mesh vertices are colored according to the skinning weights rigging them to the upper leg's bone (Figure from [31])

.

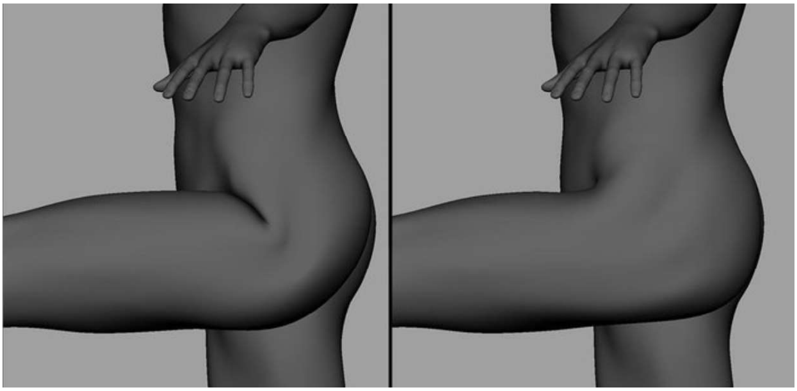


Figure 2.19: Left: Leg extension resulting from linear blend skinning. Right: After applying pose correctives. (Figure from [33])

flexibility for the artist in the posing process. First, this is an over-parametrization of the body motions, as the arm flexion, in reality, is only a single degree of freedom. Second, using ball joints to model articulations like the scapulo-thoracic joint (shoulder), or spine, is quite far from the actual biomechanics of the human skeleton.

This leads to several limitations, first for graphical purposes, using a nonrealistic kinematic tree causes the artist to have more work designing pose-dependent blend shapes that mimic the presence of the human anatomy underneath the surface. Second, this over-parameterization also increases the difficulty of framing the range of plausible poses. To achieve a specific arm pose, those models can rotate the shoulder, the elbow, and the wrist at any angle, while the human skeleton is much more constrained.

Personalizing such models to represent a specific subject is also challenging, as changing the body shape requires editing the mesh vertices, adjusting the kinematic tree, and potentially adjusting the pose-dependent blend shapes to the new subject. Many works have attempted to automate the skeleton transfer between body shapes [34–37] but this task often still requires the artist's intervention.

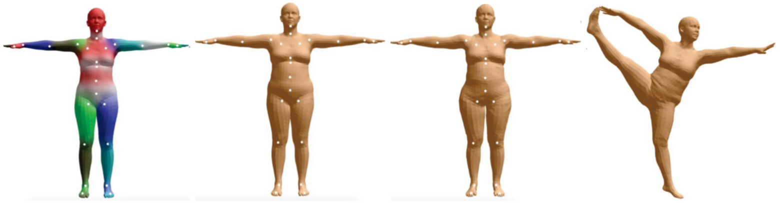
Statistical body shape models.

"Statistical Shape Models (SSMs) are geometric models that describe a collection of semantically similar objects in a very compact way" [38]. SSMs often rely on Principal Component Analysis (PCA) [39]. Instead of storing the vertex locations of each mesh from a collection, the mean shape and the principal components of the shape variations are computed. SSM enable representing a shape and its variability with a few parameters.

This approach has been used to model faces [39], bones [40], organs [41], and the whole human body shape. Successive body models have been proposed, starting from SCAPE [42], which is based on triangle deformation. SMPL [43] is based on vertex deformations, then SMPL-X [44] extends SMPL with face and fingers animation. In 2021, imGHUM is introduced as a full-body model with parametrization learned through variational auto-encoders [45]. Released in 2020, STAR [46] improves the shape expressivity and the pose-dependant blend shapes of SMPL and, more recently, SUPR [47] proposes a better model of feet compression.

To learn those body models, a template human mesh is registered to a collection of 3D scans of human bodies. The advantage of this approach is that the deformation of the soft tissues in different poses can also be learned from those data. Given a pose, learned pose corrective blend shapes can be applied instead of crafting them manually. We refer the reader to [38] for a detailed overview of the use of SSMs for modeling the human anatomy.

SMPL. In this thesis, we model the human body surface using the SMPL model. SMPL was presented by Loper et al. [43] in 2015 and is built as follows. It starts with a neutral template body mesh made of N_v vertices, noted $\mathbf{T} \in \mathbb{R}^{6080 \times 3}$ (Fig. 2.20a).



(a) $\mathbf{T}, \mathbf{W}$ (b) $\mathbf{T} + B_S(\beta)$ (c) $\mathbf{T} + B_S(\beta) + B_P(\theta)$ (d) $SMPL(\beta.\theta)$

Figure 2.20: Illustration of the SMPL parametrization. (a) The template body mesh $\mathbf{T}$ with skinning weights $\mathbf{W}$. (b) The body shape is changed with the shape parameter β. (c) Pose-dependent blend shapes $B_P(\theta)$ are applied. (d) The body is posed with the pose parameter θ.

The shape of the body can be changed with a shape parameter $\beta \in \mathbb{R}^{300}$. Effectively, a per-vertex shape-dependent displacement $B_S(\beta) \in 6080 \times 3$ is computed and added to the template body mesh (Fig. 2.20b). In most of the works presented in this thesis, we only use the first 10 components of the β vector as they are sufficient to model the main variation of the human body shape.

Given the body shape, a learned joint regressor is used to regress a kinematic tree made of 24 joints (white dots on Fig. 2.20b). Each vertex is rigged to this kinematic tree with fixed skinning weights $\mathbf{W} \in \mathbb{R}^{6080 \times 24}$, shown as colors on Fig. 2.20a.

The body surface compresses in various ways depending on the pose. Muscles like the biceps bulge when the arm is bent, the knee flexion causes the leg's soft tissues to compress, *etc*. SMPL can model this thanks to learned corrective blend shapes $B_P(\theta) \in \mathbb{R}^{6080 \times 3}$, which depend on the pose θ and are applied in T-pose (Fig. 2.20c).

The body is then posed with the pose parameter $\theta \in \mathbb{R}^{72}$ using linear blend skinning. The pose parameter contains 3 rotation angles for each of the 24 joints of the kinematic tree. We denote the posed mesh vertices as $SMPL(\beta, \theta)$.

SMPL constitutes a very compact way of representing various bodies in different poses. Indeed, the body shape can be represented with 10 to 300 shape parameters and each frame's pose with only 72 parameters.

This compactness makes SMPL great for optimization and regression tasks. For example given the picture of a subject, inferring their 3D shape is a complex task. Predicting only 72+10 parameters makes this task more tractable.

In Chapter 3, we model the body shape using the more recent model STAR. Compared to SMPL, STAR has localized pose-dependant blend shapes and is trained on many more scans, thus it can model a wider range of body shapes.

In terms of degrees of freedom and anatomical accuracy, SMPL and STAR suffer the same limit as the basic graphical body models discussed in the previous section.

Moreover, statistical body models, as they model the shape variation on a per-vertex basis, are intrinsically linked to the mesh representation. It makes it hard to extend them to shapes that can not be represented well by a mesh, or for which the correspondences between two instances are not well defined.

Implicit representation

A way to overcome the limitations of meshes is to use an implicit representation of the human body. Either a signed distance function (SDF) or a binary occupancy function can be used. The SDF is a function that returns the distance to the surface of the body, with a negative value inside the modeled object and a positive value outside. The binary function returns a boolean value, true inside the object and false outside.

These methods have been applied to model the body surface [45, 48–55], clothed bodies [49, 56–63], and clothing [64, 65].

In contrast to meshes, implicit representations are more computationally expensive to render and the animation is less explicit.

2.4.2 Human pose representation

As mentioned in Sec. 2.1.1, the human body limbs articulate following specific degrees of freedom. A way to abstract those articulations is to represent the body as a kinematic tree, where each limb is represented by a node in the tree and is linked to parent and child body parts with specific degrees of freedom.

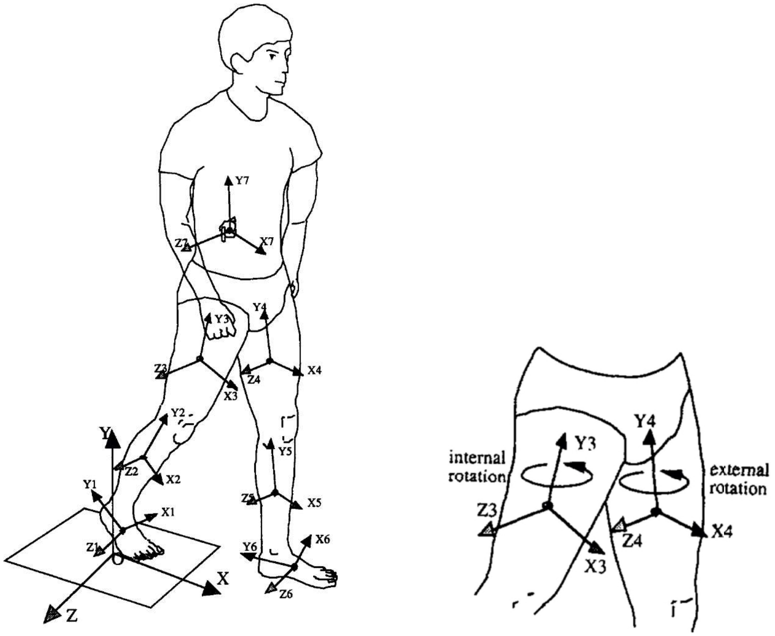
Defining a kinematic tree requires attaching a frame of reference to each limb, and there are many ways to do so. For example, in computer graphics body models and SMPL, the kinematic tree is defined in a rest pose (usually T-pose), so the limb's frame depends on this pose. In this case, the result of a leg extension of 30° will depend on the initial rest pose.

In contrast, in biomechanics, the International Society of Biomechanics (ISB)[5] has defined a standard to qualify the pose of the limbs (Fig. 2.21) [66]. Thus, a leg extension of 30° is defined wrt a specific axis. This standard has been extended to the upper and lower skeleton [67, 68] (Fig. 2.22 and 2.23) and is widely used in the biomechanics community, but less in the Graphics and Computer Vision community.

2.4.3 Biomechanical models

In biomechanics, 3 classes of models are used for modeling human locomotion, as illustrated in Fig. 2.24: (i) in skeletal models, a skeleton motion is driven by torques applied to joints (ii) in musculoskeletal models, the motion is driven by muscle activation, and

[5] The International Society of Biomechanics (ISB) has a dedicated webpage listing all their standards: https://isbweb.org/students/29-standards-documents.



(a) Conventions for *global reference frame* and *segmental local center of mass reference frame*.

(b) The same rotations about *segmental local center of mass reference frames* produce anatomically different motions on the left and right sides of the body.

Figure 2.21: The kinematic tree of the human lower body as defined by the International Society of Biomechanics (Figures from [66]).

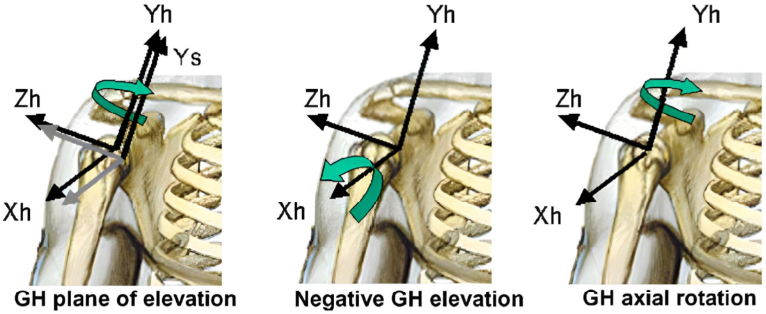


Figure 2.22: Humerus coordinate system and definition of GH (Glenohumeral rotation center) motions. Y_s is the local axis for the scapula coordinate system.

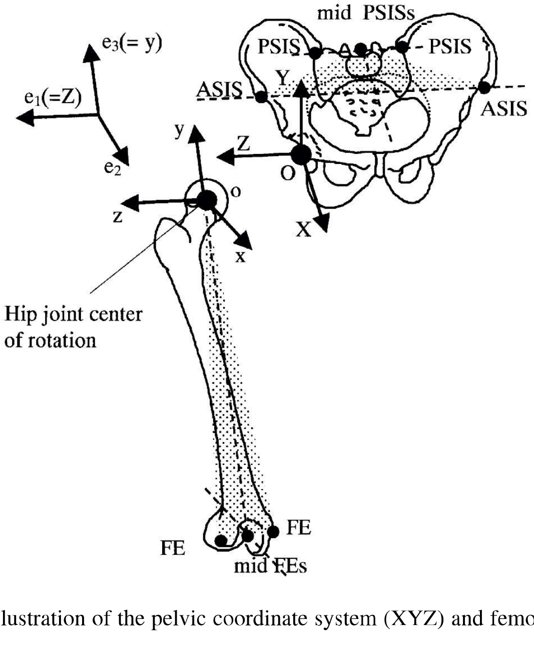


Figure 2.23: Illustration of the pelvic coordinate system (XYZ) and femoral coordinate system (xyz)

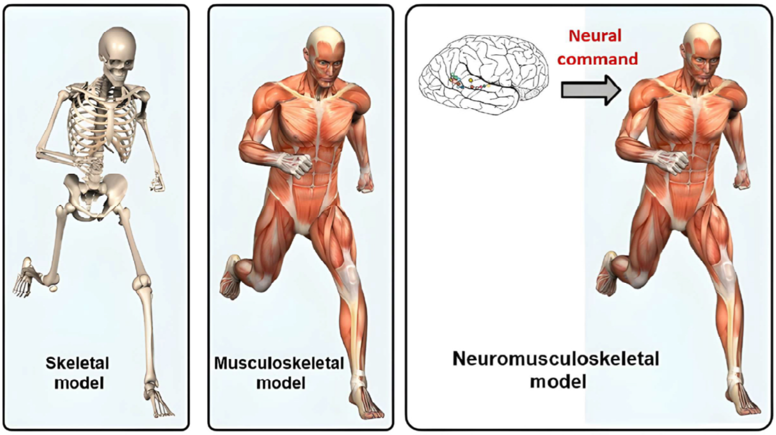


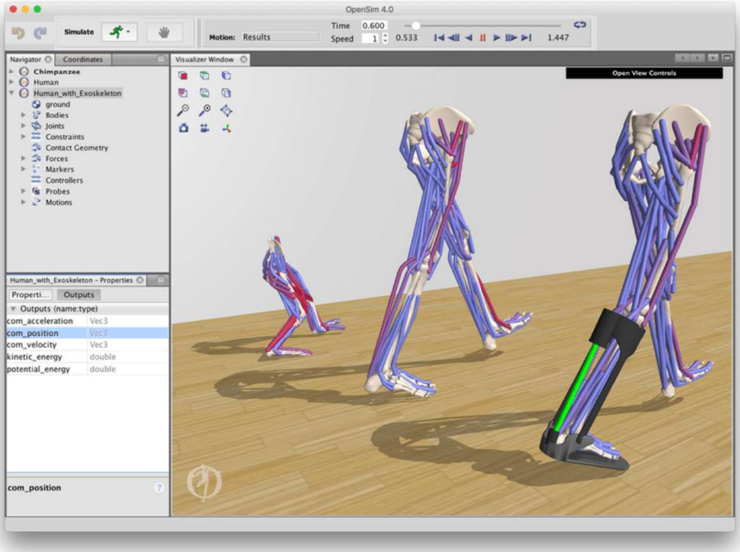
Figure 2.24: Classification of human modeling (figure from [69]).

(iii) in neuromusculoskeletal (NMS) models, the movement is controlled by a model of the central nervous system [69]. A comprehensive overview of the development of these models can be found in [70].

Different frameworks have been developed for modeling, simulating, and controlling such models, like OpenSim [71] (Fig. 2.25) or Anybody [72] (Fig. 2.26). These frameworks enable recovering the kinematics and dynamics of a human motion sequence. The input data is usually mocap sequences and ground reaction forces (GRF) measured with a pressure plate. From the mocap sequence, inverse kinematics (IK) is used to recover the sequence of skeleton pose parameters that explain the marker trajectory. This yields joint trajectories over time (location, speed, and acceleration of those joints).

Biomechanical models also specify Body Segment Inertial Parameters (BSIPs), comprising the mass, position of the center of mass, and moments and products of inertia for each segment of the human body. Then, given the joints' acceleration and mass distribution in the body, Newton's laws of motion can be used to compute the torques exerted on each joint, enabling further analysis.

Besides rigid bodies physics, biomechanical simulation frameworks can also model the dynamics of muscle activation [70, 73]. The point of attachment of the muscles to the bones and their paths have to be defined on the model (Fig. 2.27) which makes it possible, through simulation, to recover the muscle activation signal that led to the motion. In the case of a neuromusculoskeletal model, the signal from the central neural system triggering the muscle activation can also be estimated. In this thesis, we use skeletal models without muscle definition.



Time 0.600
Speed 1 0.533 1.447
Motion: Results

Outputs (name:type)

com_acceleration	Vec3
com_position	Vec3
com_velocity	Vec3
kinetic_energy	double
potential_energy	double

com_position

Figure 2.25: The OpenSim framework's interface. Muscle paths are defined as strings along the bones and their color shows the muscle activation.

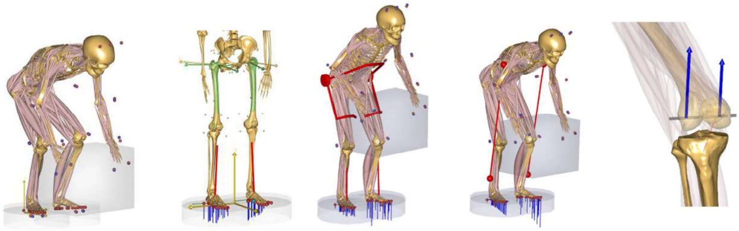


Figure 2.26: Visualizations from the AnyBody framework. We can see the markers defined on the skeleton for fitting it to mocap data, and the measured ground reaction forces applied on the feet.

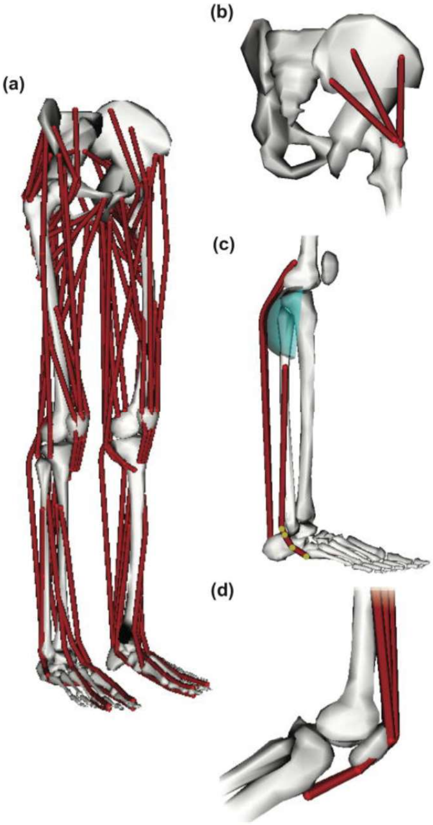


Figure 2.27: In the Rajagopal model [74], muscle paths are defined wrt the bone meshes. Since the model was created for gait analysis, only the lower body muscles are modeled.

Personalization. Personalization of biomechanical skeleton models is key for obtaining accurate simulation outputs for a specific subject. To personalize a biomechanical skeleton model, mocap data are usually used to scale the model's limbs to match the captured data. The kinematic tree segments, the limb meshes, and its BSIPs are then scaled accordingly.

In most cases, the BSIPs estimation for a subject is based on anthropometric tables built from cadaver studies (see [70] for a review). Effectively, the initial skeletal model BSIPs' are scaled according to the limb scales, height and mass of the subject. Note that in simulation frameworks like OpenSim, the actual body shape of the captured subject is not considered, only the limb scales. Moreover, biomechanical models are often not sex specific, so both sex are assumed to have the same mass distribution.

To perform physics simulations, the limb volumes are usually approximated by cuboids or capsules. But for visualization purposes, bone meshes are rigged to the kinematic tree.

Modeling the articulations Many mechanical models of the human joints have been developed, with various complexity. The selection of the appropriate joint model is crucial for valid kinematic and dynamic outcomes.

Most skeletal models use ideal joints, like ball joints to model the shoulder joint and hips, and hinge joints for the knee [70]. Models are usually developed for specific purposes, with the complexity and level of detail adapted to the task. For example, the Rajagopal model [74], illustrated Fig. 2.27, is widely used for gait analysis. While it models the leg's muscles, the torso is rigid and the arm muscles are not modeled.

2.4.4 3D anatomical atlases

To provide a reference for the complete anatomy, 3D anatomical atlases have been created like Zygote[6] or Anatomy 3D Atlas[7]. Such models represent the detailed structures within the human body, such as bones, arteries, nerves, organs, and muscles. They constitute an important educational resource for the medical profession as well as for the general public.

Creating such models is an extensive and meticulous process, often involving segmenting medical scans [75–78]. Therefore the resulting model is subject-specific, is built in a specific standing pose, and can not be reposed.

Personalizing anatomical models

Several works propose generalizing anatomical models to new poses and body shapes.

[6] https://www.zygotebody.com/

[7] https://anatomy3datlas.com/

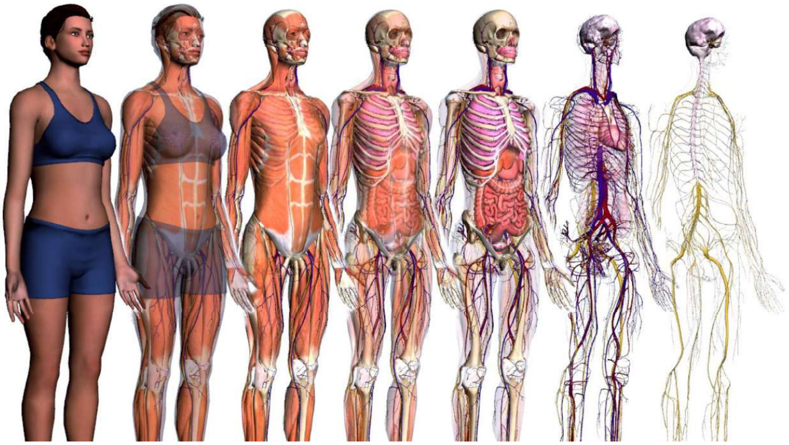


Figure 2.28: The Zygote body model enables the visualization of different layers of the detailed human anatomy.

With Anatomy Transfer [79] (AT), an anatomical template is morphed to match a target body shape. AT's optimization regularizes the different tissues to deform plausibly. The surface of the anatomical model is deformed using Laplacian deformations, while the underlying anatomy is interpolated, except for the bones, which are deformed with affine transformations. The skeletal structure is enforced by defining springs between the bones that maintain coherent articulations.

In the same line of work, Saito et al. [80] simulate the growth of fat, muscle, and bones to match a target body shape. Kadleček et al. [81] propose a physics-based anatomic model that can be adapted to input 3D scans, where, similar to Zhu et al. [82], the bones are deformed using linear blend skinning with bounded bi-harmonic weights.

While the obtained bones are visually plausible, the personalized anatomy yielded by these models is neither learned from data nor validated against it.

Chapter 3

Inferring the skeleton geometry

In this thesis, we aim to infer the internal anatomy from the external body shape. In this chapter, we start by inferring the skeleton geometry. From a subject's 3D body shape, we learn to infer the geometry of their bones.

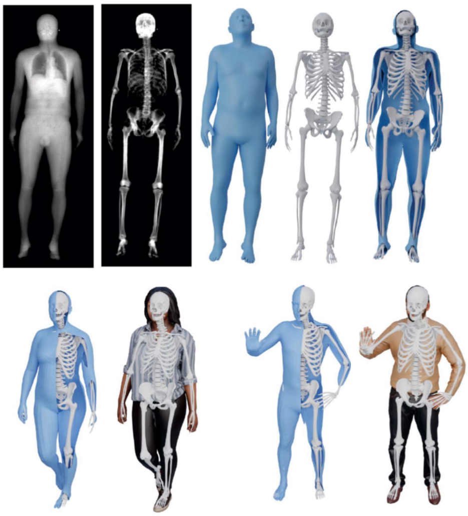


Figure 3.1: From DXA scans we learn to predict the skeleton from the body surface. Left: input DXA soft tissue and bone images; body and skeleton shapes fit to the images; bones predicted from the skin; overlay. Right: predicted OSSO skeletons from RenderPeople [83] scans.

3.1 Introduction

As detailed in Chapter 2, statistical body models enable estimating the 3D human pose and shape (HPS) of subjects from images or videos. While the surface shape is accurate, these models are all based on a "skeleton" that only approximates the kinematic structure of the body using a small number of linear segments with ball joints. These simplified skeletons are useful for animating virtual characters and action recognition, but they are inappropriate for applications in medicine and biomechanics. To be more widely relevant, HPS methods must output a skeleton corresponding to the actual anatomic human skeleton. No statistical body model exists that captures both the detailed outer surface of the body and the anatomic skeletal structure inside. The key problem is the lack of paired data capturing the inside and outside of the body.

In this chapter, we address the problem of inferring the human anatomic skeleton, i.e. the bone shapes and locations, solely from surface observations. That is, we *infer the bones from the skin*. To that end, we learn a statistical model of the skeleton shape and its correlation with the skin surface (Fig. 3.1 top). Given a posed body, our method predicts the skeleton from the body shape, and poses it inside (Fig. 3.1 bottom).

Anatomic body models with skin and bones are important in computer graphics, medicine, and biomechanics, enabling realistic animation of the body anatomy and physical simulation of body motion. Existing state-of-the-art anatomic models [71, 84–86] represent different body parts: skin, muscles, organs, and skeleton. They are mainly developed for sports, health applications, or educational purposes. While very detailed, they don't generalize easily to new subjects. Graphics-oriented anatomic skeletons [79, 81, 82, 87] can deform the individual bones with simple geometric transformations (e.g. scale or affine) and can be fit to new subjects. However, these deformations lack anatomic realism relative to actual skeletons. We argue that this realism can be improved using a data-driven strategy.

In computer vision, 3D statistical shape models of the human body are widely used [42, 43, 46, 88]. These have two elements in common: they model the human external shape, i.e. the skin surface, and are learned from data. Using thousands of 3D scans, these models capture the statistical variability of the human body shape. In this chapter, we use STAR [46] because it has a richer shape representation than SMPL [43]. Such models, however, employ a simplified kinematic skeleton and joints. While they can be readily inferred from data, the idealized skeletal structure means they cannot be used for applications in biomechanics.

To address these issues, we take a data-driven approach and learn a statistical skeleton shape model, as well as the mapping from body shape to this skeleton model. Our method, **OSSO** (Obtaining Skeletal Shape from Outside), takes a STAR model instance of any shape and pose, and estimates its corresponding skeleton. The skeleton can then be animated by reposing the STAR model.

The key problem, however, is obtaining training data that simultaneously gives the inside and outside of the body in 3D. Most imaging technologies that simultaneously capture the inside and outside of the body use ionizing radiation, which is harmful to humans; e.g. Computed Tomography (CT) and X-rays. As a consequence, such data is extremely limited, preventing learning-based methods. Our insight is to use dual-energy X-ray absorptiometry (DXA) data. DXA scans use low-dose X-rays to measure bone mineral density and body fat composition. The radiation level is so low that it is certified to be used on healthy patients for clinical studies, such as the UK Biobank [89]. In a DXA scan, two images are computed by combining two different energy levels: a soft-tissue image I_S and a bone image I_B (Fig. 3.2). In I_S the silhouette of the body can clearly be seen, whereas I_B reveals the structure and shape of the bones.

Unfortunately, DXA does not produce 3D data. Consequently, we fit the STAR body model to the soft-tissue image to obtain an estimate of the outer 3D body surface. We also employ a constrained part-based fitting method to fit bones to the DXA bone image. These then provide pairs of inside and outside data for training. We use 1200 male and 1200 female DXA images from the UK Biobank [89], which we split into training and evaluation sets. From the training set, we learn skeletal shape variation and the mapping from outside to inside. Given a new body shape and pose, OSSO reposes the body to a canonical pose and predicts the skeleton inside. It then reposes the skeleton to the input pose, subject to various anatomic constraints. With ground truth DXA scans, we validate the reposing in lying down poses and show that OSSO outperforms Anatomy Transfer [79]. We also demonstrate the use of OSSO by estimating skeletons for a variety of body shapes and poses.

In this chapter, we make the following contributions: (1) We fit a 3D body model to DXA images to obtain a 3D body shape. (2) We fit a collection of bones to DXA images using a variety of constraints. (3) We learn a statistical (PCA) model of skeleton shape variation, capturing correlations between bones. (4) We learn a mapping from external body shape to internal skeleton shape. (5) Given a 3D body in any pose, we repose the body, predict the skeleton, and repose under physical constraints to obtain a plausible posed skeleton. (6) We demonstrate superior performance to other approaches, validated on DXA imagery. (7) We show varied reposing results for 3D bodies estimated from 3D scans. (8) Inference code is available for research purposes. (9) The paired outer surfaces (skin) and skeleton (bone) meshes are made available as a Biobank Returned Dataset (see the OSSO project page for more info[1]).

In summary, OSSO provides a data-driven approach to enrich 3D human pose and shape estimation with skeletal information. This is a step towards biomechanics in the wild. Methods that estimate models like SMPL or STAR from images and video can immediately use OSSO to estimate the skeletal structure. While many methods provide some sort of visually appealing skeleton, this work is the first to learn and validate such a skeleton based on data of the inside and the outside of the body.

[1] https://osso.is.tue.mpg.de/

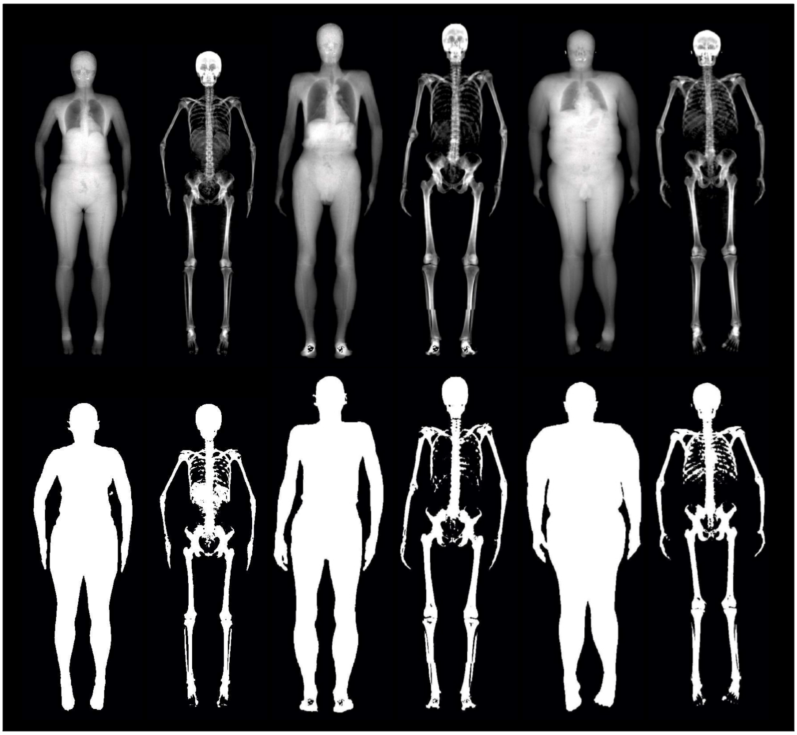


Figure 3.2: Top: Pairs of soft tissue (I_S) and bone (I_B) DXA images. Bottom: Computed skin (M_S) and bone (M_B) masks.

3.2 Related Work

We review work on data-driven skin and bones models, and methods that create personalized anatomic models.

Data-driven skin models. In Sec. 2.4.1, we presented several statistical body models. Learned statistical body models [42, 43, 46, 88, 90]. In this chapter, we use STAR [46] to represent the body surface with two parameter vectors (β_S, θ_S) controlling the shape and pose of the body, respectively.

Data-driven bone models. In medicine, patient-specific 3D bone models are very valuable. Since many scanning modalities are 2D (X-ray), numerous methods address fitting 3D models to 2D images. However, as pointed out by a review of existing methods [91], most models are restricted to individual bones (or groups) and are learned from 3D information. Our method learns a 3D skeleton model from 2D DXA images.

Fitting models to images. The literature of methods fitting 3D body models to 2D images is wide [44, 92–101] and was recently surveyed [102]. However, less work fits such models to X-ray images. Pansiot and Boyer [103] leverage video-based surface motion capture to recover a volumetric representation of the hand from planar X-ray images. In this chapter, we leverage a silhouette term [104] and regressed landmarks to fit models of the body surface and the skeleton to the segmented DXA images.

Personalized anatomic models. Several works have addressed the problem of creating a personalized anatomic model of a subject from external or internal observations.

Gilles et al. [105] propose a morphing algorithm to register a template skeleton to a target skeleton mesh or 3D image. The registration is done by alternating elastic and plastic deformations, and joint position corrections constrained by prior kinematic information. At each step, the deformed individual bones are projected onto a statistical model of the bone to ensure plausible bone shapes. Since the bone shape space is built from synthetically deformed bone shapes and not from actual bone scans, it is unclear how representative the shape space is of the population.

As previously mentioned in Chapter 1, models have also been developed to generate new body shapes by simulating the growth of fat, muscle, on top of bones [80]. A physics-based anatomic model can then be fit to input 3D scans, utilizing linear blend skinning with bounded bi-harmonic weights to deform the bones, as detailed by Kadleček et al.[81] and Zhu et al.[82]. While the obtained bones are visually plausible, these models are neither learned from data nor validated against it.

In Phace [106], two independent face and skull shape models are combined to infer a probabilistic distribution of the face given a skull. This goal is similar to ours, as we want to infer the skeleton shape from the body shape. In contrast, however, we do not have a statistical shape space for the whole skeleton, and thus we learn it.

Wang et al. [76] propose a method to scan a hand with MRI (Magnetic Resonance Imaging) and create an accurate, personalized anatomic model. The model can plausibly extrapolate to new unseen poses with high visual realism. The created model is person-specific and cannot be inferred from skin observations.

Zoss et al. [107] propose a method to track the invisible jaw from the visible skin surface. In their method, they propose a calibration phase to adapt the jaw size to a new subject. OSSO goes further, as the shape of the bones is estimated from the outside in addition to their location.

Bauer et al. [87] infer the skeleton of a subject from RGBD images of the skin. Their skeleton inference method is based on Anatomy Transfer [79] with extra constraints positioning the bones inside the body and avoiding bone intersections. Bones are parametrized with affine transformations, and results are not validated on medical data.

The recent BASH model [108] couples a musculoskeletal biomechanical model to the SCAPE model [42]. BASH generates a skeleton from sparse measurements, but the obtained anatomy is not validated on medical images. Unlike STAR, the SCAPE model does not guarantee constant bone lengths when reposed. The main difference with BASH is that we use a data-driven approach to learn to infer the shape of the skeleton inside the human body, and we validate on medical images.

The most related work to ours is Anatomy Transfer (AT) [79]. In AT, a skeleton is generated from only the external shape of an avatar, without requiring a particular initialization. Given a target body shape, an anatomical template is morphed to match it. The surface of the anatomical model is deformed using Laplacian deformations, and the underlying anatomy is interpolated, except for the bones, which are deformed with affine transformations. The skeletal structure is enforced by defining springs between the bones that keep them coherent. In this chapter, we use a similar approach by leveraging the Stitched Puppet graphical model [109]. While AT generates a plausible anatomy for any kind of humanoid avatar, the generated anatomy is not validated on real data. Our work goes beyond AT by using data to learn the skeletal deformation space and by providing a quantitative evaluation on real DXA images. We consider Anatomy Transfer to be the baseline and compare our predictions to theirs.

Lastly, recent work by Wong et al. [110], shows that the human internal body composition can be predicted from solely body surface measurements. Our work is complementary to theirs, as OSSO predicts the geometry and location of the skeleton inside the body surface.

3.3 Data

A key contribution in this chapter is to create a unique dataset for training and evaluation that contains paired outer surface (skin) and skeleton (bone) meshes $(\mathbf{R}_S, \mathbf{R}_B)$ from DXA

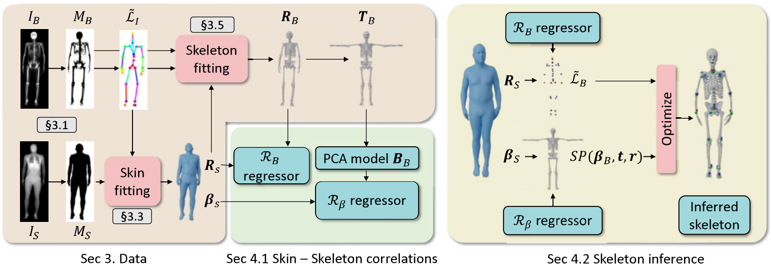


Figure 3.3: Overview of OSSO: learning and inference. From the input DXA images (I_B, I_S) we obtain the skeleton and skin masks (M_B, M_S). From the skeleton mask M_B, we predict 2D landmarks $\mathcal{L}_i$ and use them to register STAR to M_S and obtain $\mathbf{R}_S$ and β_S. We then register our skeleton graphical model to M_B, $\mathcal{L}_i$, and $\mathbf{R}_S$ and obtain $\mathbf{R}_B$ and its unposed version $\mathbf{T}_B$. From the unposed skeletons $\mathbf{T}_B$ we learn a skeleton shape space B_B; with paired $(\mathbf{R}_S, \mathcal{L}_B(\mathbf{R}_B))$ we learn the regressor $\mathcal{R}_B$ and with paired (β_S, β_B) we learn the regressor $\mathcal{R}_\beta$. At test time, from the body surface $(\mathbf{R}_S, \beta_S)$ we regress the skeleton shape $\mathcal{R}_\beta(\beta_S) = \beta_B$ and optimize its pose to match the regressed locations $\mathcal{R}_B(\mathbf{R}_S) = \tilde{\mathcal{L}}_B$.

images. The dataset is made available to the community as a Biobank Returned Dataset[2].

Creating the dataset has several steps: (1) we segment DXA images to get the silhouettes of the body and bones (Sec. 3.3.1), (2) we create synthetic skeleton silhouettes and use them to learn to predict landmarks (Sec. 3.3.2), (3) we register STAR [46] to the skin silhouette images (Sec. 3.3.3), (4) we create a custom skeleton model (Sec. 3.3.4) and register it to real skeleton binary masks (Sec. 3.3.5). Fig. 3.3 shows an overview of the dataset creation procedure.

3.3.1 Skin and skeleton masks from DXA images

From the input images (I_S, I_B), we compute the corresponding skin and skeleton segmentation masks (M_S, M_B).

For the skin mask M_S, we threshold I_S. As some small artifacts remain, mainly due to pixels in the lungs with low-intensity values, we detect the closed contours on the image and fill in small areas. In Fig. 3.2 we show pairs of input I_S and the obtained mask M_S.

The automatic segmentation of bones in DXA images is still an open problem. Often, DXA image regions are obtained by manually annotating keypoints [25], and accurate segmentations are performed manually [111]. Moreover, these methods only focus on a small set of bones. Jamalud et al. [112] use a U-Net to segment body parts from

[2] See https://osso.is.tue.mpg.de/Dataset.html for the instructions.

DXA scans that are relevant for scoliosis classification. Unfortunately, the code is not available.

In this chapter, we use a simple heuristic to automatically segment the bone tissue in the bone images: we assume that the $X\%$ brightest pixels in each I_B image belong to bone tissue. We empirically set $X = 20\%$ for the male DXAs and $X = 17\%$ for the female. As small artifacts remain (earrings, clothing, etc.), we remove small connected components with an area of less than 50 pixels. Note that we do not claim to segment all bone tissues in the DXA images. While our segmentations are coarse, they capture the structure and location of the bones inside the body (as shown in Fig. 3.2); this is what we need to fit a 3D skeleton to them.

3.3.2 Computing landmarks from the silhouettes

Many model-based human pose estimation methods rely on fitting projected 3D joints to 2D landmarks. Landmark detection must be automated as we fit thousands of DXA images. Existing landmark detectors, of course, do not work with DXA imagery. So in this section, we explain how we train a landmark detector for skeleton binary masks M_B.

We start by creating an animatable skeleton model, which we call "Initial model" and denote it K, to then generate a synthetic dataset of DXA silhouettes, and finally train a 2D landmark predictor from DXA skeleton silhouettes.

Initial model creation

To generate synthetic skeleton silhouettes that look similar to real DXA bone masks M_B, we create an articulated skeleton model K, rigged with the STAR body model [46] parameters.

We first generate 21 STAR bodies by sampling the STAR shape space ${B_S}^S$. We consider the mean body, and then, for the $n_\beta = 10$ first components of the STAR shape space, we sample two new body shapes with the shape parameters $\beta = \{-2, 2\}$. Using Anatomy Transfer (AT) [79], we register a template skeleton mesh to each of these body shapes. Effectively we enforce that the skin of the AT mesh matches the STAR mesh.

With the obtained registrations, we define the mean skeleton shape $K(\beta = 0, \theta = 0)$, as the obtained AT skeleton on STAR's mean shape. Then, for each shape space component, we compute the skeleton offsets to the mean skeleton and use these offsets to define an initial skeleton shape space. From these, we compute the shape vectors of K as $B_{S_i} = (\mathbf{T}_{\beta_i=2} - \mathbf{T}_{\beta_i=-2})/4$ for i in $[0, n_\beta]$.

To pose the skeleton, we rig it with the same kinematic tree as STAR. For each skeleton bone, we manually define to which body part it belongs. This is straightforward as the initial template skeleton has the individual bones identified. It is important to note that the created skeleton model $K(\beta, \theta)$ can change its shape and pose using the same shape and pose parameters as STAR.

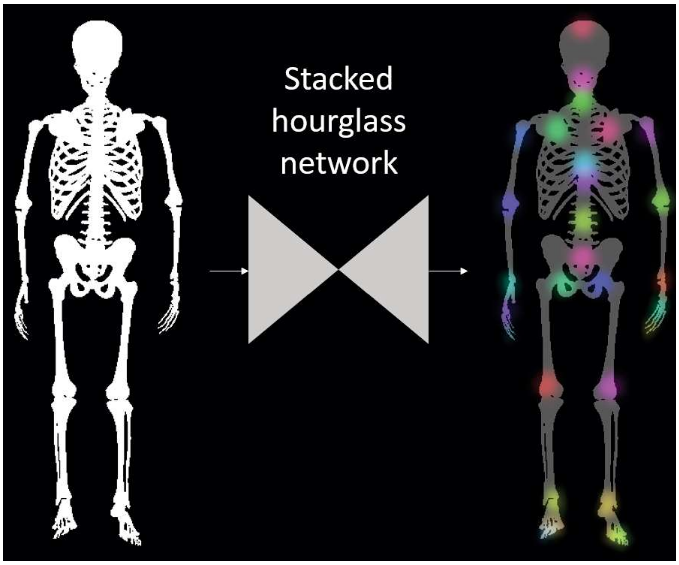


Figure 3.4: From a skeleton mask, a stacked hourglass network predicts the 2D locations of the landmarks $\tilde{\mathcal{L}}_I$.

This initial model has an obvious drawback: the kinematic joint locations are not consistent with the anatomic skeleton articulations. Still, it is sufficient to easily generate plausible synthetic bone masks and the corresponding landmark annotations.

We define 29 landmarks on the skeleton mesh (Fig. 3.5). The first 24 correspond to the closest vertex to the STAR joint locations. Additionally, we select the tip of the head, fingers, and feet. We denote these initial landmarks $\mathcal{L}_I$ or $\mathcal{L}_I(\mathbf{M})$ if we make explicit the mesh $\mathbf{M}$ on which the landmarks are defined.

Generating synthetic DXA masks

We use the skeleton model K to generate synthetic skeleton binary masks $\hat{M}_B$ with their corresponding 2D landmarks, that we denote $\tilde{\mathcal{L}}_I$ to explicitly distinguish them from the 3D landmarks $\mathcal{L}_I$.

We generate synthetic skeleton shapes by uniformly sampling the STAR shape space β in the range $[-2.5, 2.5]^{10}$. As the poses in DXA scans are relatively constrained, i.e. lying down with arms at the side, we manually define a *lying pose* θ_L and sample new angles from a uniform distribution centered at θ_L within a small range.

With the sampled shape and pose parameters, we render the silhouette of the skeleton and the corresponding landmark image. The virtual camera is orthographic to match the DXA scanner camera, and the field of view is set depending on the sample body height to leave a specific margin on the top and bottom of the image. This margin is sampled to match the margin distribution observed on the DXA dataset. A sample of the generated paired data is presented in Fig. 3.6.

This allows us to generate skeletons of different shapes in lying-down poses, which we render as binary images with the projected landmarks, giving us paired training data.

To bridge the domain gap between the synthetic silhouettes and the DXA segmentations, we augment the data by eroding and partially masking the rendered skeleton silhouettes, while keeping the landmarks fixed.

Training a 2D landmarks predictor

From the synthetic silhouettes of the skeleton $\hat{M}_B$, we train the landmark detector using a stacked hourglass network [113] with 8 stacks. The network takes a 256x256 binary silhouette as input and outputs a 29x64x64 tensor, where each channel contains the position for one of the 29 landmarks $\tilde{\mathcal{L}}_I$. The 2D landmark prediction from DXA silhouette is illustrated in Fig. 3.4.

In Fig. 3.7, we show qualitative results of the predicted landmarks on binary masks from real DXA images. In Sec. 3.5.1, we also provide a quantitative evaluation of this network on synthetic data. We visually inspected the predicted 2D landmarks and observe that the silhouette simplification strategy combined with our data augmentation technique yields very good qualitative results on real DXA images.

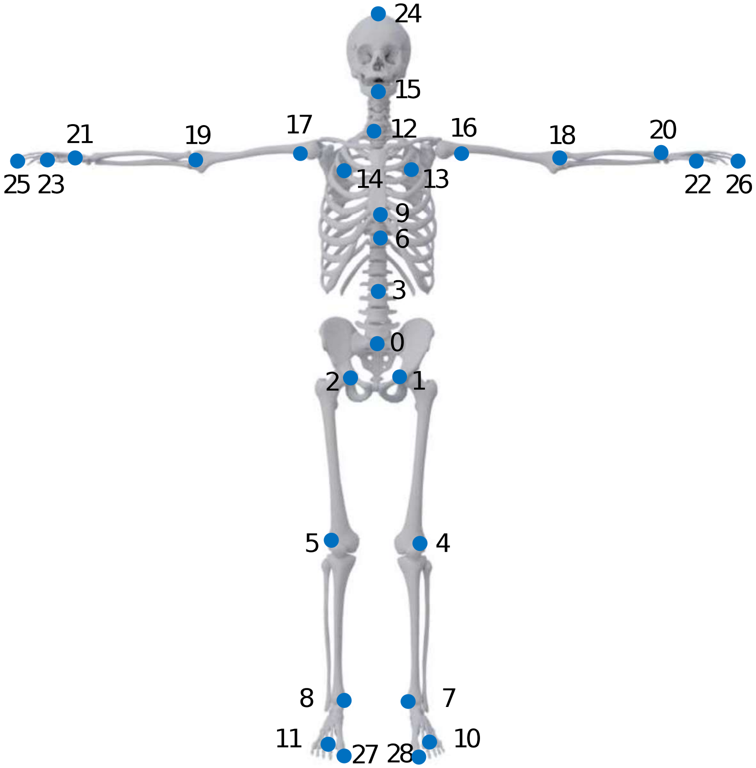


Figure 3.5: Position of the 3D landmarks $\mathcal{L}_I$ on the Stitched Puppet skeleton model P_B. These markers correspond to the location of the STAR 3D joints, plus five additional landmarks.

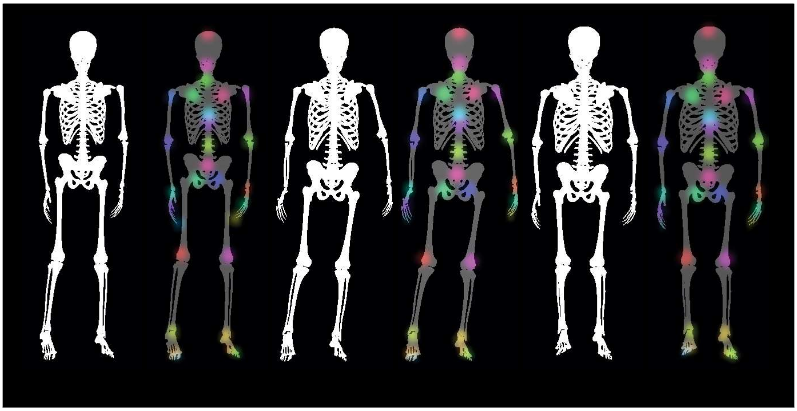


Figure 3.6: Pairs of synthetic skeleton masks (in white) and 2D landmarks $\tilde{\mathcal{L}}_I$ (color-coded) overlayed on the mask (in gray).

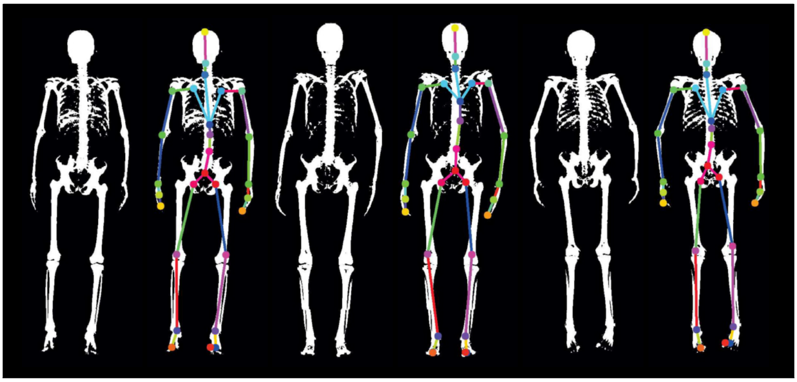


Figure 3.7: Pairs of input and predicted 2D landmarks $\tilde{\mathcal{L}}_I$ on real DXAs. The network learned on synthetic data generalizes well to real data.

3.3.3 Skin surface from DXA

A key step is to estimate the 3D body shape of a subject from their 2D DXA segmentation M_S. There is prior work on fitting a body surface model to DXA images using a silhouette [114, 115]. These methods, however, assume that a 3D scan of the subject is available. Since this is not the case for us, we fit the 3D parametric model STAR [46] to the silhouette and our predicted landmarks from above.

Since the registration is only conditioned by a silhouette and 2D landmarks, we need a good pose prior. Thus, we collected twelve 3D scans of people lying down, computed their STAR poses, learned a distribution of poses, and use this as a pose prior E_θ as in [94]. Moreover, we enforce the hands to stay in the coronal plane with the cost E_h penalizing the distance between the hand and the middle of the thighs.

To fit STAR to the silhouettes, we use the same optimization strategy as in [104] and effectively solve for the STAR shape and pose parameters $(\hat{\beta}_S, \hat{\theta}_S)$ that minimize:

$$\begin{aligned} E_S(\beta_S, \theta_S; M_S, \tilde{\mathcal{L}}_I) = {} & E_{sil}(ST(\beta_S, \theta_S), M_S) \\ & + \lambda_I ||P(\mathbf{J}(\beta_S, \theta_S)) - \tilde{\mathcal{L}}_I|| \\ & + \lambda_\beta ||\beta_S|| + \lambda_\theta E_\theta(\theta_S) + \lambda_h E_h(ST(\beta_S, \theta_S)), \end{aligned} \tag{3.1}$$

where $ST(\beta_S, \theta_S)$ is the STAR mesh and $\mathbf{J}(\beta_S, \theta_S)$ are the STAR joint locations. E_{sil} enforces the projection of the STAR mesh to match the silhouette (as in Eq. 6 of [104])[3], $\tilde{\mathcal{L}}_I$ are the predicted landmarks, and P is the orthographic camera projection function. We denote the obtained mesh $\mathbf{R}_S = ST(\hat{\beta}_S, \hat{\theta}_S)$ (see $\mathbf{R}_S$ and β_S in Fig. 3.3).

This approach works well in general, but can fail in cases like severe scoliosis or limb atrophy. These cases have high silhouette fitting errors, and we use these errors to detect and remove failure cases automatically from the final dataset (see Fig. 3.14 and Fig. 3.15 for examples).

3.3.4 Skeleton model based on Stitched Puppet

Now that we have the skin surface, we need the skeleton inside. To register a 3D skeleton model to the DXA bone mask M_B, we need a model where the individual bones can freely move and deform but can be controlled with connectivity. Our initial skeleton model (Sec. 3.3.2) is not well suited for this task, as it did not let us translate the bones at will since they were rigged to the STAR kinematic tree. Thus, we create a new skeleton model, capitalizing on the *stitched puppet* [109] and the synthetic shape deformation space from *GLoSS* [116].

The *stitched puppet* model, as the name implies, represents an articulated deformable structure, the human body, as a collection of parts that are stitched together at the part interfaces. The model has per-part shape spaces and a pose parametrization in terms of

[3] We use the implementation available at https://github.com/silviazuffi/smalr_online

location of each part center and its global rotation. The *stitched puppet* can be seen as a graphical model, where part parameters are defined at each node, and edge potentials represent stitching costs, that favor the parts to be connected and have smooth skin connections. The original model [109] is fit to 3D scans of people with non-parametric particle belief propagation. In order to define a stitched puppet model given an existing mesh, one needs to define a segmentation of the faces into parts, duplicate the vertices that belong to different adjacent parts, and define stitching potentials that act as springs between the corresponding duplicated vertices.

Starting with the same AT skeleton template as before, we manually define 21 groups of bones that belong to the same anatomic part, and define the interfaces between these parts. In Fig. 3.8 we show the different parts with color codes, their interfaces, as well as the 3D landmarks $\mathcal{L}_B$ defined on the bones.

Also, unlike [109], we do not use a graphical model inference method to register the model to data. We refer to the skeleton mesh as $SP(\beta_B, \mathbf{t}, \mathbf{r})$, where $(\beta_B, \mathbf{t}, \mathbf{r})$ are respectively the shape, translation, and rotation of all the skeleton parts. As we use the same AT skeleton template, the landmarks $\mathcal{L}_I$ are properly defined.

3.3.5 Skeleton from DXA

Now that we have a suitable skeleton model $SP(\beta_B, \mathbf{t}, \mathbf{r})$, we can fit it to the DXA. We use the binary skeleton mask M_B, the estimated landmarks $\tilde{\mathcal{L}}_I$ and the skin registration $\mathbf{R}_S$ and optimize for the skeleton model parameters $(\hat{\beta}_B, \hat{\mathbf{t}}, \hat{\mathbf{r}})$ that minimize:

$$E_B(\beta_B, \mathbf{t}, \mathbf{r}) = E_{data}(\beta_B, \mathbf{t}, \mathbf{r}; M_B, \tilde{\mathcal{L}}_I, \mathbf{R}_S) + E_{prior}(\beta_B, \mathbf{t}, \mathbf{r}), \tag{3.2}$$

where

$$\begin{aligned} E_{data}(\beta_B, \mathbf{t}, \mathbf{r}; M_B, \tilde{\mathcal{L}}_I, \mathbf{R}_S) = & E_{sil}(P(SP(\beta_B, \mathbf{t}, \mathbf{r})), M_B) \\ & + \lambda_I ||P(\mathcal{L}_I(SP(\beta_B, \mathbf{t}, \mathbf{r}))) - \tilde{\mathcal{L}}_I|| \\ & + \lambda_i E_i(SP(\beta_B, \mathbf{t}, \mathbf{r}, \mathbf{R}_S)) \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} E_{prior}(\beta_B, \mathbf{t}, \mathbf{r}) = & \lambda_{shape} ||\beta_B|| + \lambda_{pose} ||\mathbf{r} - \mathbf{r_T}|| \\ & + \lambda_{sti} E_{sti}(SP(\beta_B, \mathbf{t}, \mathbf{r})) \\ & + \lambda_{sy} E_{sy}(SP(\beta_B, \mathbf{t}, \mathbf{r})) \end{aligned} \tag{3.4}$$

where $\mathbf{r_T}$ are the rotations of the bones in a manually defined lying down pose, E_{sti} is the $-\log$ of the stitching potentials described in [109], and E_{sy} forces the symmetric body parts on the right and left to have a similar shape. Note how landmarks $\mathcal{L}_I$ here are now written as a function of the skeleton mesh.

The cost E_i enforces the skeleton to be inside the body $\mathbf{R}_S$ and in contact with the skin in some manually defined regions (knee, tibia, elbow). We decompose E_i as $E_i = E_{in} + E_p + E_{ct}$ and illustrate the intuition of each cost in Fig. 3.9. Below, we detail further each of these 3 costs.

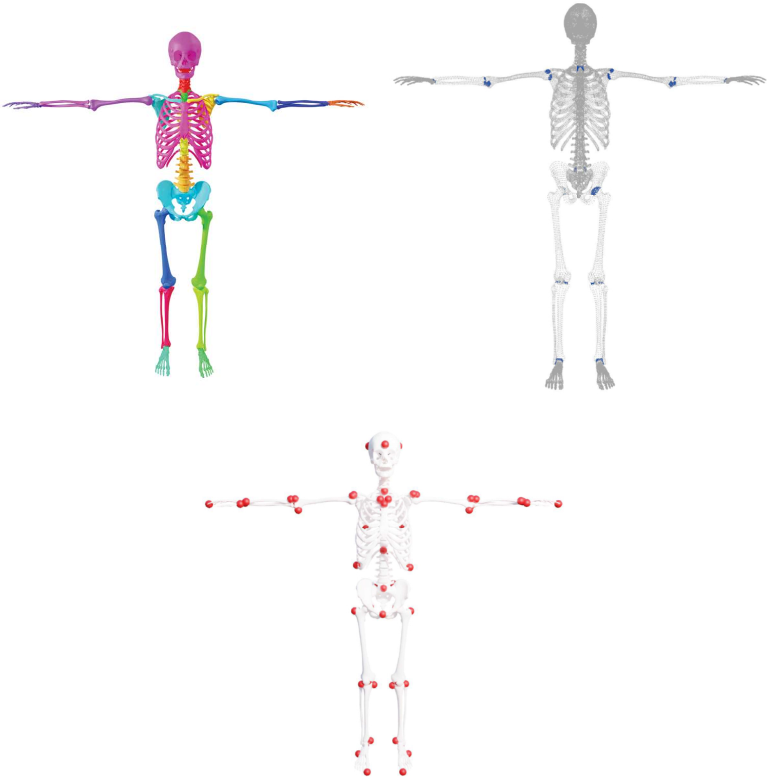


Figure 3.8: Our *stitched puppet* skeleton model, with the different bone groups (top left), the interface point between the groups (top right) and the 3D landmarks $\mathcal{L}_B$ (bottom).

We denote the vertices of SP as v_{sp}, the vertices of ST as v_{st} and z the anterior-posterior axis. v^z denotes the z component of vertex v and v^n the mesh normal at this vertex.

The energy term E_{in} forces the skeleton to be inside the body along the front-back axis.

$$E_{in} = \max(0, D_z(SP(\beta_B, \mathbf{t}, \mathbf{r}), \mathbf{R}_S)) \tag{3.5}$$

where D_z is the distance along z between a SP vertex and the closest skin vertex.

The term E_p forces vertices of the skeleton to be close to specific areas of the skin along the front-back axis. For several manually defined pairs of skeleton vertices and skin area A, we define

$$E_p = v_{sp}^z - \sum_{v_{st} \in A} (v_{st}^z). \tag{3.6}$$

The energy E_{ct} forces the *contact* between some specific vertices of the skeleton and the skin, like the elbow or the fingertips.

We define pairs of skin and skeleton vertices (v_{sp}, v_{st}) and want them to be at a fixed small distance $e = 5mm$. Effectively, E_{ct} is the per vertex distance:

$$E_{ct} = v_{sp} - (v_{st} - e \cdot v_{st}^n). \tag{3.7}$$

After the optimization, we obtain the skeleton's pose and shape parameters $(\hat{\beta}_B, \hat{\mathbf{t}}, \hat{\mathbf{r}})$ and the registered skeleton mesh

$$\mathbf{R}_B = SP(\hat{\beta}_B, \hat{\mathbf{t}}, \hat{\mathbf{r}}). \tag{3.8}$$

Fig. 3.3 shows the registered skeleton $\mathbf{R}_B$ for one subject.

Unposing the skeleton registration. Before we can learn a skeleton shape space, we need to pose-normalize the optimized skeletons $\mathbf{R}_B$. While obtaining the unposed mesh $\mathbf{T}_S$ of a STAR fit $\mathbf{R}_S$ is straightforward - one just zeros the pose parameters, unposing $\mathbf{R}_B$ is ill-posed as one can zero the rotations $\mathbf{r}$, but the translations $\mathbf{t}$ need to be adjusted. To constrain the problem, we make the hypothesis that the 3D offsets between the skin and skeleton do not vary much from one pose to another. From the registrations $(\mathbf{R}_S, \mathbf{R}_B)$ of a low BMI subject, we define 3113 pairs of skin and skeleton indices $\{(sn_p, sk_p)\}$ and define their initial 3D offset $\mathbf{d}_p^0$:

$$\mathbf{d}_p^0 = (\mathbf{R}_S[sn_p] - \mathbf{R}_B[sk_p]) \tag{3.9}$$

This allows us to define a signed distance cost between the unposed meshes: We can then define the cost function E_d to maintain this offset across poses.

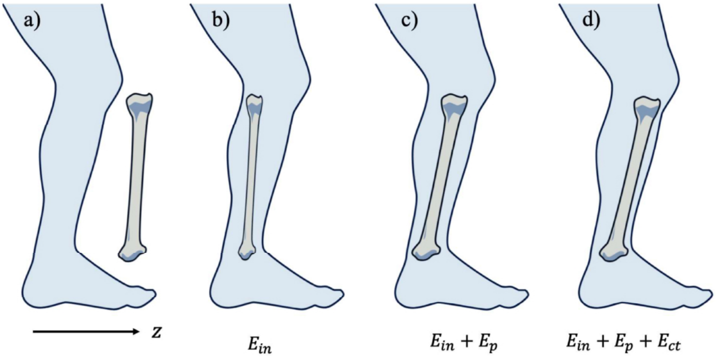


Figure 3.9: We illustrate the intuition behind the costs composing E_i on a profile view of the tibia in the leg. a) The frontal silhouette does not yield any constraint for the bone to be inside the body along the z-axis. b) We use E_{in} to force it to be inside. Forcing it inside is not enough as it could squeeze and collapse; thus, we enforce the bone to be close to the skin surface with E_p (c). In addition, there are regions where the bones are not covered by muscle and fat and should, therefore, lie close to the skin surface. We use E_{ct} to enforce these manually defined areas of contact (d).

$$E_d(\mathbf{T}_S, \mathbf{T}_B) = \sum_p w_p \cdot (\mathbf{T}_S[sn_p] - \mathbf{T}_B[sk_p]) - \mathbf{d}_p^0 \tag{3.10}$$

with

$$w_p = sign(\mathbf{T}_S[sn_p] - \mathbf{T}_B[sk_p]) \cdot N(\mathbf{T}_B[sk_p]) \tag{3.11}$$

where $N(\mathbf{T}_B[sk_p])$ is the normal on the skeleton mesh at vertex sk_p.

In Fig. 3.10 we illustrate the pairs of skin and skeleton vertices that are used for this cost. Our heuristic is that each of these pairs has a fixed distance d_0 that should remain constant independent of the 3D pose.

Thus, to unpose the skeleton, we fix $\beta_B = \hat{\beta}_B$ and find $(\hat{\mathbf{t_u}}, \hat{\mathbf{r_u}})$ that minimize the weighted sum of the two previously introduced losses:

$$E_u(\mathbf{t}, \mathbf{r}) = \lambda_{sti} E_{sti}(SP(\hat{\beta}_B, \mathbf{t}, \mathbf{r})) + \lambda_d E_d(\mathbf{T}_S, SP(\hat{\beta}_B, \mathbf{t}, \mathbf{r})) \tag{3.12}$$

to obtain the unposed skeleton vertices:

$$\mathbf{T}_B = SP(\hat{\beta}_B, \hat{\mathbf{t_u}}, \hat{\mathbf{r_u}}) \tag{3.13}$$

Fig. 3.3 illustrates the unposed skeleton $\mathbf{T}_B$ for a subject.

3.4 Method – OSSO

Now that we have paired meshes $(\mathbf{R}_S, \mathbf{R}_B)$ and unposed $\mathbf{T}_B$ meshes, we learn their correlations and how to predict skeleton landmarks $\tilde{\mathcal{L}}_B$ from the skin vertices $\mathbf{R}_S$ (Sec. 3.4.1). Then, at test time, given an arbitrary STAR body shape in an arbitrary pose, we predict a skeleton mesh (Sec. 3.4.2) and then repose it to match the input skin pose (Sec. 3.4.3). Fig. 3.3 illustrates this pipeline.

3.4.1 Skeleton statistics and correlation to skin shape

With the dataset created in Sec. 3.3, we can now learn the correlation between the skin and the bones. With the unposed skeletons $\mathbf{T}_B$, we first compute a low-dimensional linear subspace B_{SB}, representing the skeleton shape variations using Principal Component Analysis (PCA). We then learn a linear regressor $\mathcal{R}_\beta$ that predicts the skeleton shape space coefficients $\beta_B \in B_{SB}$ from the STAR shape space coefficients β_S computed in Sec. 3.3.3.

To properly constrain the 3D location of the inferred skeleton inside the body, we define a new set of landmarks $\mathcal{L}_B$, composed of three landmarks per bone group. We learn to infer them from the skin with one linear regressor per landmark. The regressor $\mathcal{R}_B$ takes as input the vertices of $\mathbf{R}_S$ and predicts the 3D landmarks on $\mathbf{R}_B$, i.e. $\mathcal{L}_B(\mathbf{R}_B)$.

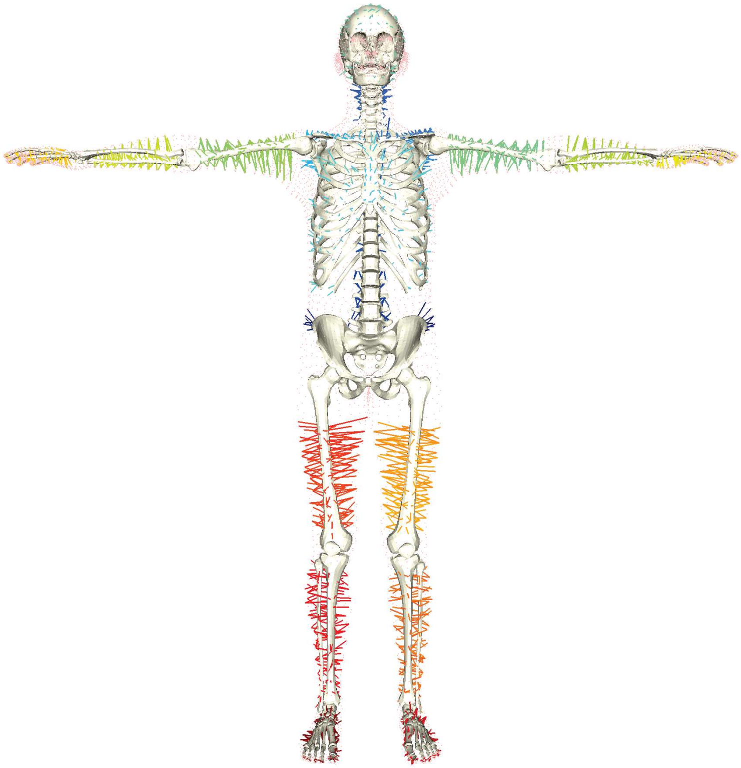


Figure 3.10: Skin to skeleton pairs used in the cost E_d. We color the links in each part with a different color for visualization purposes.

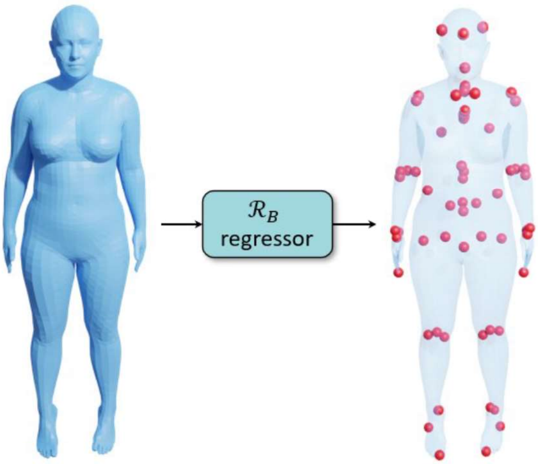


Figure 3.11: Given a skin mesh, the landmark regressor lets us compute the landmark 3D locations as a linear combination of the skin mesh vertices locations.

We formulate the problem as a non-negative least squares problem and solve it with an active set method [117]. This regression is learned in a normalized lying down pose as illustrated in Fig. 3.11.

3.4.2 Inferring the skeleton from the skin

We now have all the elements to predict the skeleton shape from an input body surface in STAR format: $(\mathbf{R}_S, \beta_S)$. Using the learned regressor $\mathcal{R}_\beta$ (Sec. 3.4.1) we predict the subject's skeleton shape $\beta_B = \mathcal{R}_\beta(\beta_S)$ from the body surface shape β_S.

Then, to properly position the skeleton inside the body, we pose the body surface in a normalized lying pose θ_S^L, obtaining $\mathbf{R}_S(\theta_S^L)$ and predict 3D bone landmarks $\mathcal{R}_B(\mathbf{R}_S(\theta_S^L)) = \tilde{\mathcal{L}}_B$.

Let us write $SP(\beta_B, \mathbf{t}, \mathbf{r})$ to refer to the skeleton with shape β_B posed with the *stitched puppet* pose parameters $(\mathbf{t}, \mathbf{r})$. To obtain the bone poses $(\mathbf{t_0}, \mathbf{r_0})$ that match the predicted landmarks, we minimize:

$$E(\mathbf{t},\mathbf{r}) = \lambda_L||\mathcal{L}_B(SP(\beta_B,\mathbf{t},\mathbf{r})) - \tilde{\mathcal{L}}_B|| + \lambda_{ct}E_{ct}(SP(\beta_B,\mathbf{t},\mathbf{r}),\mathbf{R}_S(\theta_S^L)), \tag{3.14}$$

where E_{ct} forces the contact between the skeleton and the skin, as defined in Sec. 3.3.5. The obtained mesh is our skeleton prediction.

3.4.3 Reposing the inferred skeleton

For arbitrary poses, the simple stitching cost between bones can not properly model articulations like the knees and shoulders. A preferable approach would be a biomechanical model with a kinematic tree enforcing a proper definition of the articulations. However, defining a bone shape space for such models is not trivial, so we focus on more anatomical articulations in the next chapter. In this chapter, we model two key articulations in more detail: ball joints and ligaments.

Ball joints like shoulders, elbows, or hips should stay in their sockets. For such joints, we identify sets of vertices on the skeleton that define the joint socket and the insertion bone head. We fit spheres to them, and define an energy E_j that forces the spheres to remain at a similar distance.

Effectively, we define for each articulation, sets of vertices s_i, s_j of the skeleton template representing respectively the joint socket and the corresponding bone head. At each optimization step, we fit spheres with centers S_i, S_j to these vertex sets, and force the sphere pairs to remain at a similar distance during the optimization:

$$E_j(\mathbf{t}, \mathbf{r}; SP_0) = ||S_i(\mathbf{t}, \mathbf{r}) - S_j(\mathbf{t}, \mathbf{r})|| - d_{s0}. \tag{3.15}$$

This cost is not sufficient to model the knee movement, so we define a stitching cost E_l approximating the human knee ligaments. We create pairs of vertices (l_i, l_j) at the bone locations where the ligaments are attached, and define the per-vertex cost E_l as:

$$E_l = ||l_i - l_j|| - d_{l_0}. \tag{3.16}$$

The distances d_{l0} and d_{s0} are defined such that $E_j(\mathbf{t_0}, \mathbf{r_0}; SP_0) = 0$ and $E_l(\mathbf{t_0}, \mathbf{r_0}; SP_0) = 0$.

Note that our articulation models, like all models, are an approximation to the truth and could be further refined for specific needs such as extreme bending poses.

Technically, given an inferred skeleton SP_0

$$SP_0 = SP(\beta_B, \mathbf{t_0}, \mathbf{r_0}) \tag{3.17}$$

inside the corresponding lying down skin mesh ST_0

$$ST_0 = ST(\theta_S^L, \beta_S) \tag{3.18}$$

, and a skin mesh in a specific pose

$$ST_\theta = ST(\theta_S^P, \beta_S) \tag{3.19}$$

, we pose the skeleton inside ST_θ.

We first compute the set of $\mathbf{d_p^0}$ offsets between SP_0 and ST_0 in the lying down pose and then minimize the reposing cost Eq. (3.12) to position the skeleton inside the posed body ST_θ. Then, to enforce more realistic anatomic joints we add:

$$E(\mathbf{t},\mathbf{r}) = \lambda_l E_j(\mathbf{t},\mathbf{r};SP_0) + \lambda_j E_l(\mathbf{t},\mathbf{r};SP_0) \tag{3.20}$$

to Eq. (3.12) and optimize again.

The resulting skeleton SP_θ fits realistically inside the target body ST_θ as shown Fig. 3.21 and Fig. 3.22.

3.5 Experiments

To evaluate our approach, we first evaluate our 2D landmarks regressor (Sec. 3.5.1). We then quantify how accurately the skin registrations $\mathbf{R}_S$ match the skin masks M_S (Sec. 3.5.2) and the skeleton registrations match the skeleton mask M_B (Sec. 3.5.3). We then evaluate how accurately our learned regressors predict the 3D bone landmarks from the skin (Sec. 3.5.4). Finally, we quantitatively and qualitatively evaluate how the projections of the computed bones overlap with the DXA bones masks (Sec. 3.5.5), we compare our prediction to AT (Sec. 3.5.6), and show the inference generalization to new poses (Sec. 3.5.7).

For the following evaluation, we use a training set of 1000 subjects per gender and a test set of 200 subjects from the UK Biobank dataset [89]. We made sure both sets have the same Body Mass Index distribution.

3.5.1 DXA to 2D landmarks regressor

As the original DXA images do not have annotations, we only evaluate the regressed 2D DXA landmarks quantitatively on unseen synthetic data. We evaluate the landmarks predicted by the stacked hourglass network on 100 unseen synthetic skeleton silhouettes. The prediction error is measured in pixels on an image of size 256x256 pixels. The per landmark errors are reported in Table 3.1.

Most errors are on the order of one pixel. The highest prediction errors are for the tip of the middle fingers (L25 and L26) and the toes (L27 and L28). We observe that due to the resizing of the skeleton mask from the original image size (approx 800x800) to the size of the network (256x256), fine structures such as fingers and toes are degraded or lost. This is numerically visible with the standard deviations of the finger markers which are over 1 pixel.

3.5.2 STAR fits to DXA skin masks

We first evaluate how well our skin registrations $\mathbf{R}_S$ overlap with the skin mask M_S (Sec. 3.3.3) on the whole 2400 subjects dataset. We compute the intersection over union metric, as ideally, all segmented skin pixels in M_S should be covered by the projection

	err. (mean ± std)
L0	0.73 ± 0.35
L1	0.95 ± 0.40
L2	0.81 ± 0.38
L3	0.90 ± 0.46
L4	1.14 ± 0.54
L5	1.12 ± 0.60
L6	0.78 ± 0.46
L7	1.16 ± 0.63
L8	1.24 ± 0.68
L9	1.07 ± 0.37
L10	1.18 ± 0.52
L11	1.18 ± 0.62
L12	0.87 ± 0.41
L13	0.87 ± 0.41
L14	1.01 ± 0.43
L15	0.78 ± 0.37
L16	1.01 ± 0.47
L17	0.87 ± 0.50
L18	1.22 ± 0.62
L19	1.01 ± 0.54
L20	1.22 ± 0.69
L21	1.21 ± 0.56
L22	1.08 ± 0.75
L23	1.04 ± 0.69
L24	0.75 ± 0.43
L25	1.87 ± 1.39
L26	1.53 ± 1.02
L27	1.23 ± 0.61
L28	1.23 ± 0.67

Table 3.1: Prediction error in pixels of the predicted 2D landmark $\tilde{\mathcal{L}}_I$ on synthetic skeleton silhouettes. Landmark numbers are visually shown on the mesh in Fig. 3.5.

of the skin registrations $\mathbf{R}_S$. We obtain a mean of 0.94 for the female subjects, 0.95 for the males, with standard deviations below 0.01. The small failure regions are due to soft tissue compression deformations of a lying down person that the STAR model does not capture.

In Fig. 3.12 and 3.13, we also show qualitative results of the STAR fits to the DXA silhouettes. The color-coded images show that the skin registrations faithfully capture the DXA skin silhouettes.

As mentioned in the last paragraph of Sec. 3.3.3, we use the quality of the fit to detect and remove failure cases from our datasets, i.e. subjects whose body shape can not be explained with STAR. In Fig. 3.14 and 3.15, we show some failure cases with low in-

	Male	Female	Male	Female
Method	$\cap_R (\%) \uparrow$		HD (px) $\downarrow$	
$\mathbf{R}_B$	92	94	8.2 ± 2.6	5.6 ± 1.7
OSSO	**88**	**89**	**10.6** ± **3.2**	**9.1** ± **2.3**
AT∼ [79]	84	88	14.4 ± 2.9	11.5 ± 3.1

Table 3.2: Quantitative comparison of OSSO and AT [79]. The $\cap_R$ score standard deviations are all below 2%.

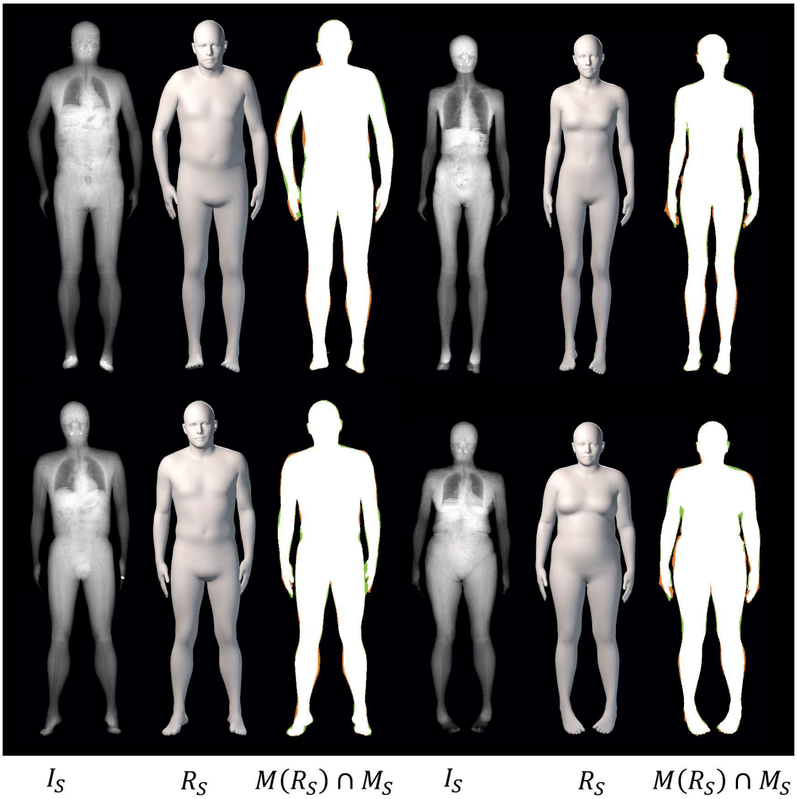


Figure 3.12: Comparison of the aligned STAR models $\mathbf{R}_S$ with the target DXA masks M_S for subjects sampled from the curated dataset. On the left we show males, and on the right females. The masks intersection is color-coded as follow: green: $\mathbf{R}_S$ only, orange: M_S only, white: both.

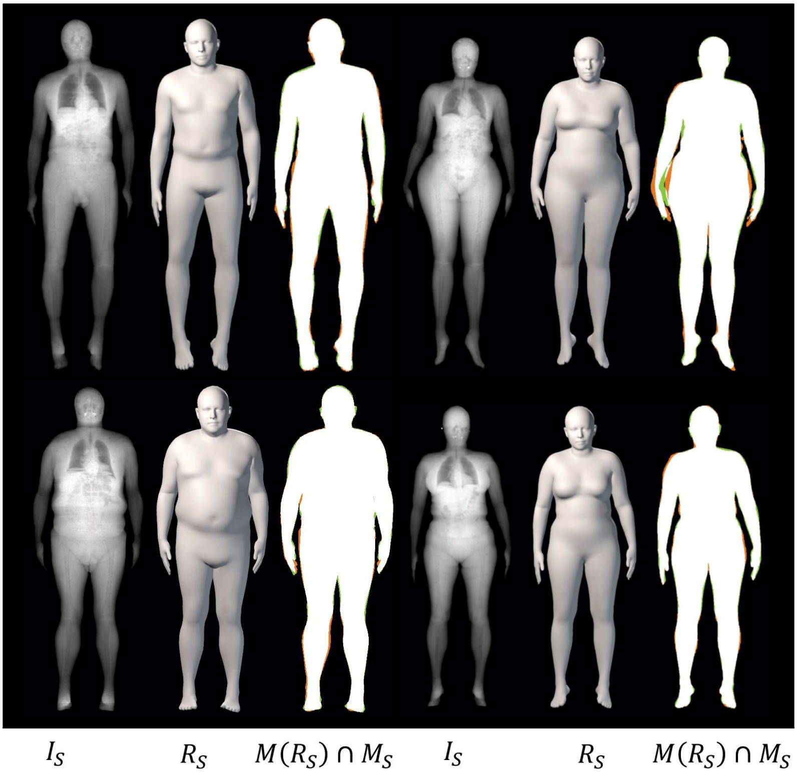


Figure 3.13: Comparison of the aligned STAR models $\mathbf{R}_S$ with the target DXA masks M_S for subjects sampled from the curated dataset. On the left we show males and on the right females. The masks intersection is color-coded as follows: green: $\mathbf{R}_S$ only, orange: M_S only, white: both.

tersection over union values. These examples include subjects with atrophied or swollen limbs, severe scoliosis, or very low BMI. In practice, we used the alignment score to remove outliers of the available DXA scans (about 1%) to constitute a curated dataset containing a training set of 1000 subjects and a test set of 200 subjects for each gender.

3.5.3 Skeleton fits to DXA skeleton masks

Next, we show qualitative results of the skeleton registrations $\mathbf{R}_B$ in Fig. 3.16 and 3.17. Our skeleton model matches the DXA pose and overlaps with the silhouette of the bones. The subjects are the same as in Fig. 3.12 and 3.13.

3.5.4 Skin to 3D landmark regressors

We next evaluate the accuracy of the regressors $\mathcal{L}_B$ (Sec. 3.4.1) by predicting skeleton landmark locations from the body surface on the test set.

For each gender, we train the regressors on 1000 subjects and evaluate on 200 left-out subjects the 3D distance between the landmarks on the aligned skeleton $\mathcal{L}_B(\mathbf{R}_B)$ and the predicted landmarks $\mathcal{R}_B(\mathbf{R}_S) = \tilde{\mathcal{L}_B}$.

Our predictions have a mean distance (MD) below 1 cm: 8.0 ± 6.1 mm for males and 8.4 ± 6.7 mm for females and all individual landmarks results are consistent among male and female ($\pm$1mm). The detailed per landmark errors are presented in Appendix A.1.

The more accurate landmarks correspond to the upper skull (MD $<$ 2 mm) and feet (MD $<$ 4 mm), whereas the least accurate belongs to the hip iliac crest (MD $\approx$ 20 mm). We observe that the supervision of the bone masks M_B is stronger in feet and skull than in the hip iliac crest, which is often not visible (see Fig. 3.2).

3.5.5 Skeleton prediction evaluation on 2D DXA bones masks

Next, we quantify how similar the predicted skeletons are to the subject's skeleton. However, we only have access to 2D DXA bone images (I_B, M_B). In addition, our DXA bone masks M_B are coarse, as some bones, such as the hip bone are not completely segmented. To account for this, we require every bone pixel in M_B to be covered by the skeleton projection, but not the reverse. Given a skeleton $\mathbf{R}_B$ and a bone mask M_B we compute their intersection ratio $\cap_R(\mathbf{R}_B, M_B) = 100|P(\mathbf{R}_B) \cap M_B|/|M_B|$ as a percentage. We also compute the directed Hausdorff Distance (HD) from M_B to $P(\mathbf{R}_B)$ accounting for the maximum pixel-to-pixel distance.

Table 3.2 presents the results on the test set of 200 male and 200 female test subjects held out from any learning. In the first row, we evaluate the skeletons $\mathbf{R}_B$ from Sec. 3.3.5 to validate that they faithfully match the masks M_B. We obtain mean intersection percentages of 92% and 94% and mean HDs of 8.2 and 5.6 pixels for male and females, respectively. OSSO obtains mean intersection percentages of 88% and 89% and mean HDs of 10.6 and 9.1 pixels, while AT obtains mean intersection percentages of 84% and

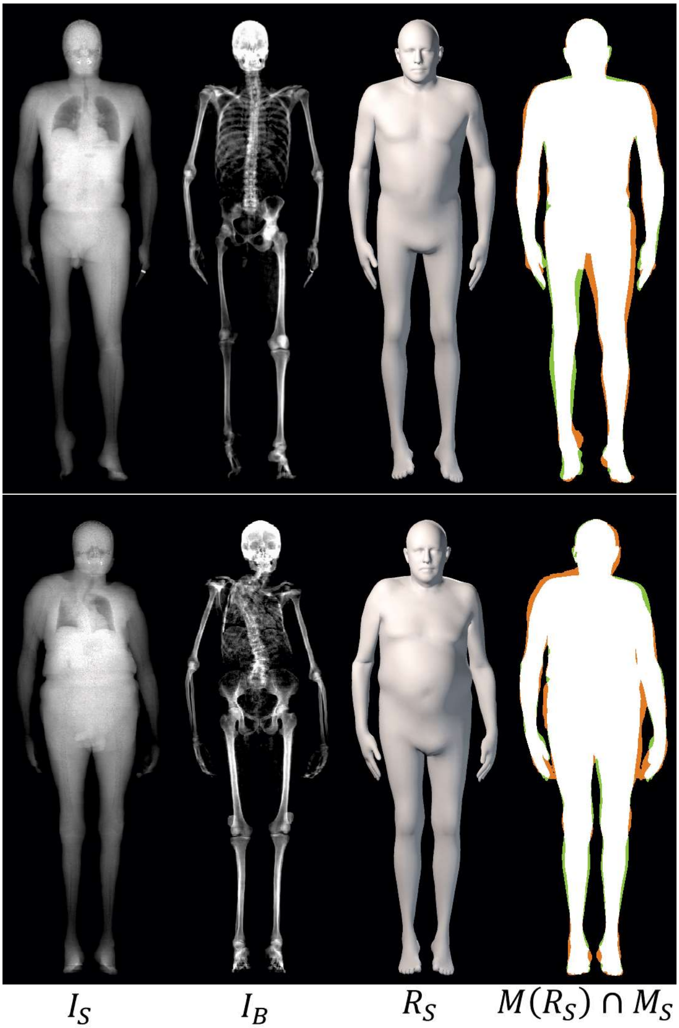


Figure 3.14: Failure cases. For each subject, we show I_S, I_B, the fitted skin mesh $\mathbf{R}_S$, and the intersection of both masks. The masks intersection is color-coded as follow: green: $\mathbf{R}_S$ only, orange: M_S only, white: both. The STAR model can not faithfully capture the shape of these subjects.

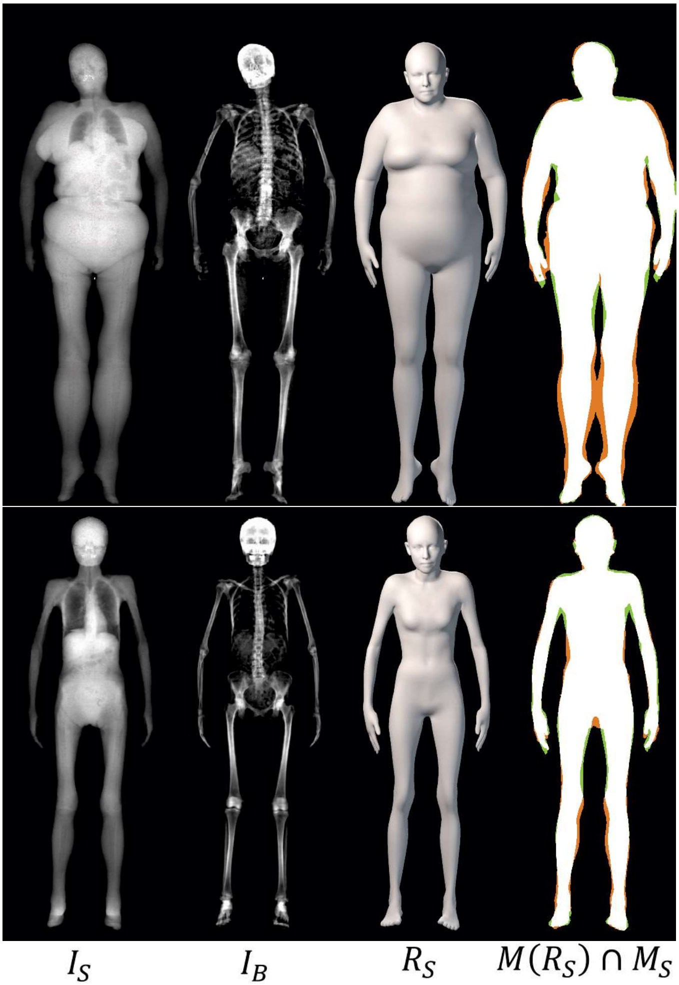


Figure 3.15: Failure cases. For each subject, we show I_S, I_B, the fitted skin mesh $\mathbf{R}_S$ and the intersection of both masks. The masks intersection is color-coded as follow: green: $\mathbf{R}_S$ only, orange: M_S only, white: both. The STAR model can not faithfully capture the shape of these subjects.

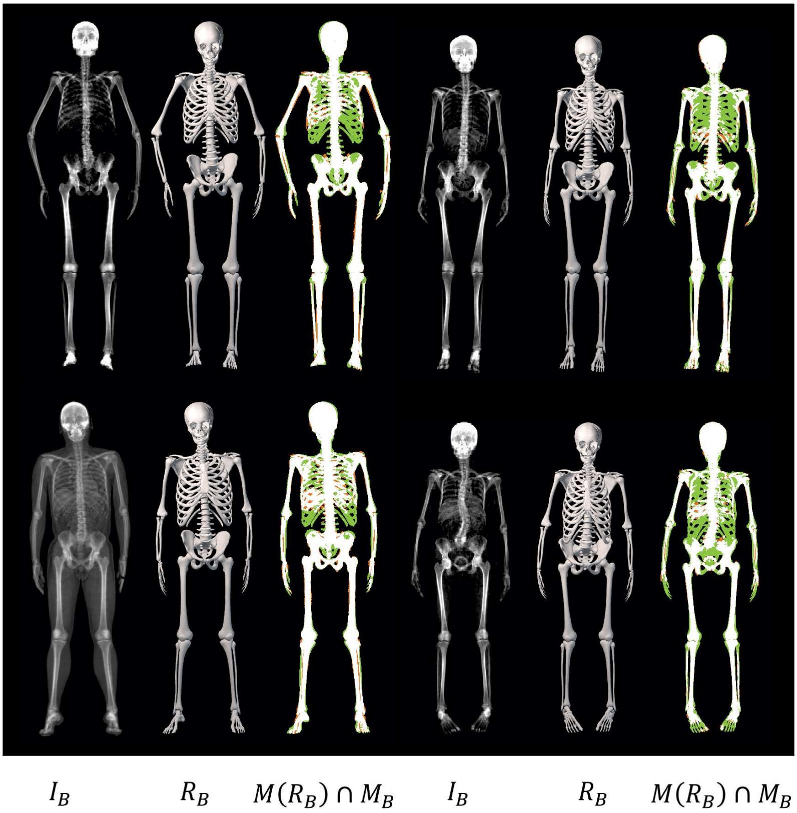


Figure 3.16: Comparison of the registered skeleton $\mathbf{R}_B$ with the target DXA masks M_B for subjects sampled from the training dataset. On the left, we show males, and on the right, females. The masks difference is color-coded as follows: green: $\mathbf{R}_B$ only, orange: M_B only, white: both.

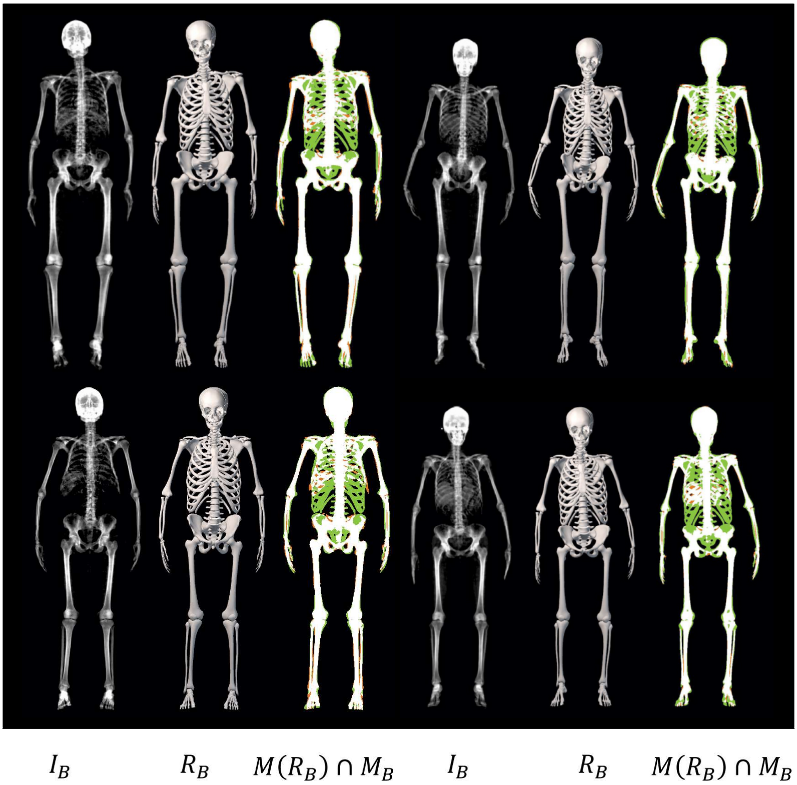


Figure 3.17: Comparison of the registered skeleton $\mathbf{R}_B$ with the target DXA masks M_B for subjects sampled from the training dataset. On the left we show males and on the right females. The masks difference is color-coded as follow: green: $\mathbf{R}_B$ only, orange: M_B only, white: both.

88% and mean HDs of 14.4 and 11.5 pixels for male and females respectively. Consistently, the OSSO predictions have higher mean intersection values and lower HD than those of AT.

The presented metric has a limitation: predicting all the skin volume as bone would obtain a perfect result ($\cap_R = 1$, $HD = 0$).

3.5.6 OSSO vs Anatomy Transfer comparison

From the DXA test set, we select 5 subjects spanning the dataset BMI distribution. From the skin alignment $\mathbf{R}_S$, we infer the skeleton and compare it to the subject's skeleton DXA image. We denote SI_{AT} the skeleton inferred with AT, and SI_{OSSO} the skeleton inferred with OSSO. $M(SI)$ is the mask rendered from the mesh SI. Figure 3.18 shows the comparison between our OSSO predictions and the ones from Anatomy Transfer

As can be seen from the images, our predictions better capture the global shape of the skeletons. Particularly, Anatomy Transfer tends to estimate the location of the hips to be too low with respect to the actual hip location. In comparison, our method predicts a skeleton that is visually closer to the one observed in the DXA images.

3.5.7 Generalization to new poses

Our regressors and model are learned from a limited set of poses, yet OSSO can predict skeletons from STAR bodies in arbitrary poses (Sec. 3.4.3). We show several examples in Fig. 3.1, Fig. 3.19, Fig. 3.21 and Fig. 3.22.

The clothed scans are from RenderPeople [83] and are part of the AGORA dataset [118], which includes high-quality SMPL fits to the scans taking into account clothing. We fit STAR to the SMPL bodies (the templates have the same topology), and then apply OSSO to estimate the posed skeleton. Unfortunately, we cannot quantitatively evaluate the accuracy of the posed skeletons. Although some minor skin interpenetrations remain, the obtained results are visually plausible.

Lateral view. Fig. 3.20 shows side views of the inference result in T-pose for various body shapes. While there is no ground truth with which to evaluate this pose, the results are plausible.

3.6 Conclusion

In this chapter, we addressed the problem of predicting the skeleton mesh of a person from their external body shape with OSSO. We use STAR [46] to represent the skin surface and use a novel method to learn a parametric shape model of the anatomical skeleton using thousands of DXA scans. We learn a mapping from the external body shape to the skeleton and can repose the skeleton inside the body subject to various constraints. We

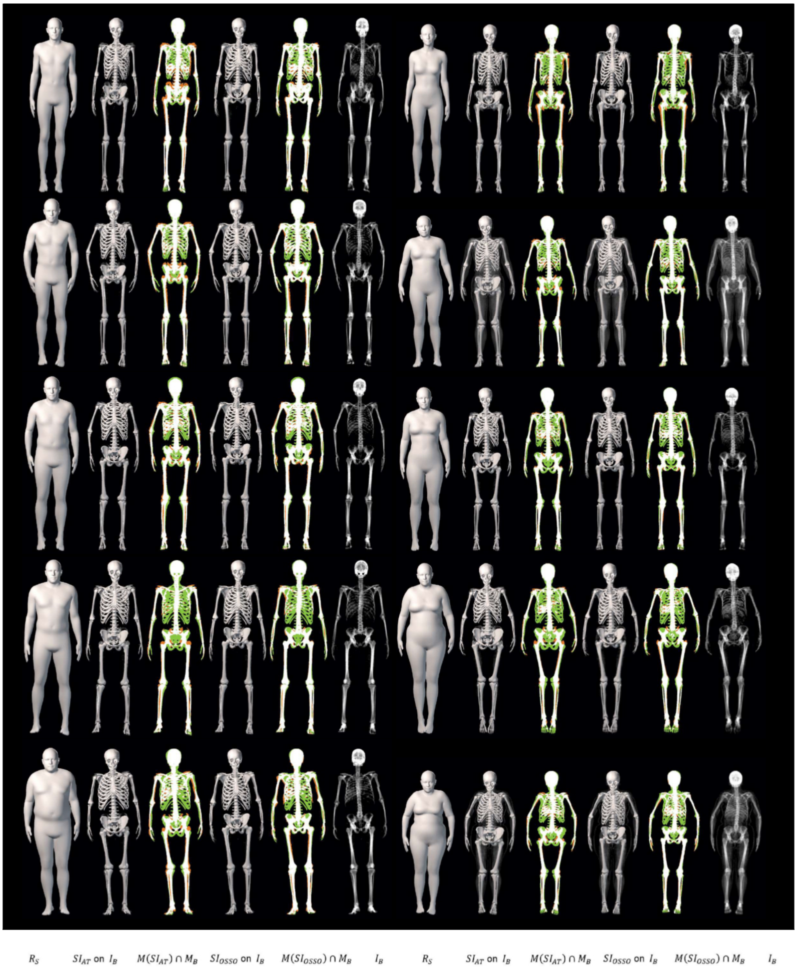


Figure 3.18: For each subject, we show in the order (1) $\mathbf{R}_S$, (2) SI_{AT} superimposed with the ground truth DXA I_B, (3) the overlap of $M(SI_{AT})$ and I_B, (4) SI_{OSSO} superimposed with the ground truth DXA I_B, (5) the difference between $M(SI_{OSSO})$ and M_B, (6) the ground truth DXA I_B

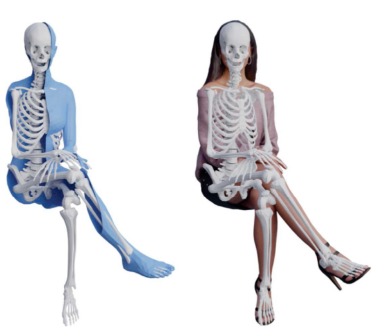
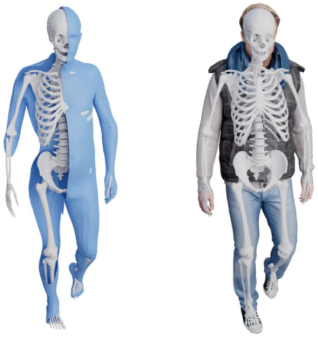



Figure 3.19: Qualitative evaluation of the skeleton inference in arbitrary poses. OSSO yields visually plausible results.

evaluate OSSO using 2D DXA images from the UK Biobank dataset where the skin as well as the structure of the bones are visible. Our skeletal predictions quantitatively outperform the state-of-the-art on silhouette reprojection error. Qualitatively, they are also better aligned with the DXA images.

The main limitations of OSSO are the 3D validation, and the absence of a kinematic tree for the skeleton. We train and validate our method on DXA images that are only 2D, and only capture the body in a lying down pose. Indeed, current medical imaging techniques do not enable capturing a large variety of poses in 3D, making it difficult to get a ground truth skeleton in various poses.

Moreover, the inferred skeleton, based on the stitched puppet, does not have a kinematic tree, leading to two issues. First, we can not guarantee that the skeleton inferred by OSSO has anatomically correct joints. Although we try to enforce plausible bone poses with a loss, the bones can end up out of their socket and not oriented properly. Not enforcing anatomical joint constrains becomes especially problematic if SMPL limb orientations are not anatomically correct. In such cases, our posing method can yield improperly oriented bones, as illustrated in Figure 3.23.

Second, reposing the inferred skeleton requires an optimization, and can lead to inconsistencies between two subsequent frames of a motion sequence.

The main reason we could not build OSSO to have a kinematic tree is that the per-bone shape space changes the size of the bones, and thus the location of the joints. Basically, the skeleton kinematic tree should depend on the skeleton's shape. This dependency is not trivial to learn, especially without skeleton ground truth in various poses. There are two solutions to that. i) Proceeding like SMPL and learning to regress anatomical joint locations from the skeleton vertices, for which we lack data. ii) Defining a scalable

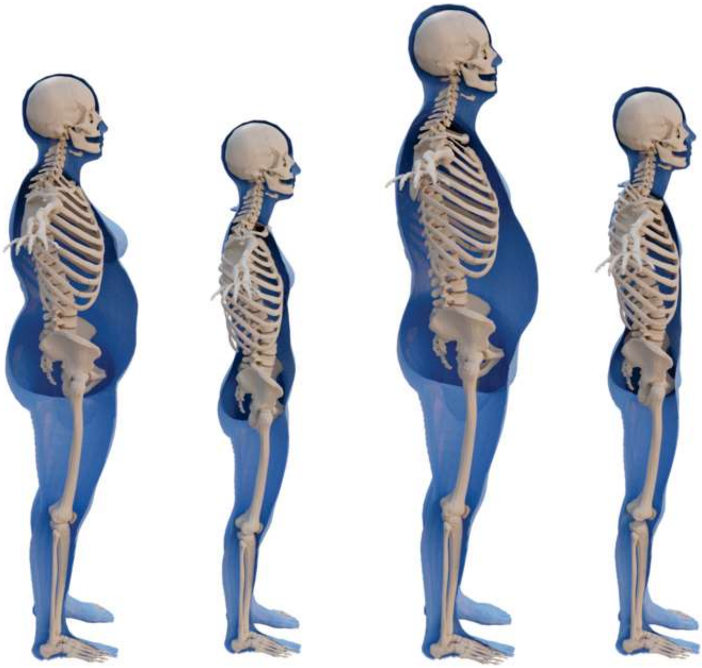


Figure 3.20: Lateral views of skeletons inferred with OSSO.

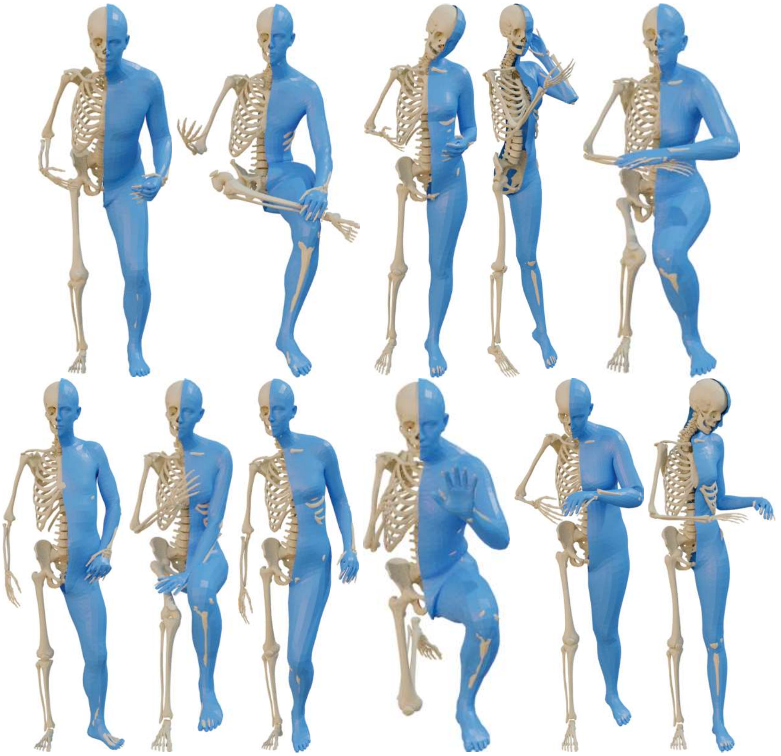


Figure 3.21: Given SMPL bodies aligned to RenderPeople subjects [83, 118], we use OSSO to infer the underlying skeleton.

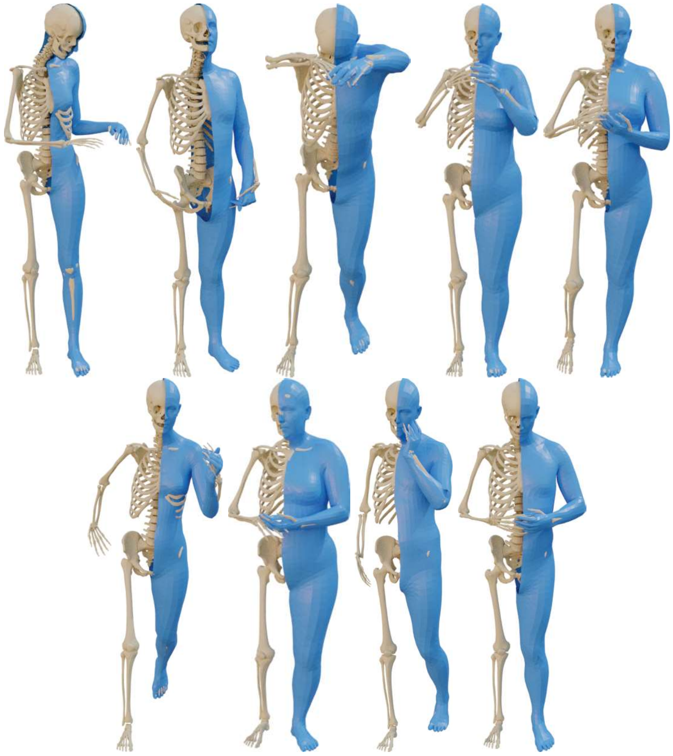


Figure 3.22: Given SMPL bodies aligned to RenderPeople subjects [83, 118], we use OSSO to infer the underlying skeleton.

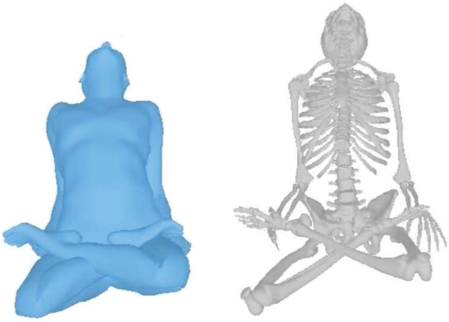


Figure 3.23: Left: SMPL fit to a yoga pose from MOYO [119]. Right: The corresponding OSSO skeleton prediction. Although the SMPL fit looks correct, the OSSO fit reveals that the thighs are not properly oriented. Without a proper kinematic tree for the skeleton constraining the articulations to be anatomically correct, there is no way to recover from the SMPL pose being unrealistic.

kinematic tree, with bones rigged to it, which is the approach followed in the next chapter.

These factors limit the usability of OSSO in biomechanics, as this field requires accurate anatomical joint locations. In the next chapter, we will address some of those limitations.

Chapter 4

Inferring the location and orientation of the bones

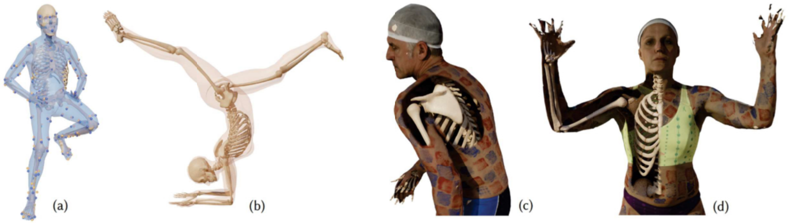


Figure 4.1: (a) We fit our new Biomechanical Skeleton Model, BSM, to SMPL [43] mesh sequences from AMASS [120]. This gives paired data enabling us to learn the mapping from skin to skeleton. (b) We use this to create SKEL, a parametric body model with skin and skeleton meshes, driven by biomechanical pose parameters and incorporating the shape space of SMPL. SKEL is like SMPL but with more realistic degrees of freedom. Fitting SKEL to DFAUST scans [121] results in SKEL's scapula sliding (c) and the forearms twisting appropriately (d).

In Chapter 3, we showed that we can infer the shape of the bones from the shape of the body, but the inferred skeleton could not be validated on 3D ground truth skeletons, and did not have a kinematic tree, making it difficult to repose, and to get precise joint locations and realistic bone orientations. In this chapter, we address these limitation by building a scalable skeleton model with an anatomical kinematic tree, and generate ground truth 3D skeletons in motion by leveraging methods from the biomechanical field. As a result, we learn to infer more precisely the bone location and orientation inside bodies in any pose.

4.1 Introduction

As seen in Sec. 2.2, the methods for capturing the human body vary significantly across the fields of research. While computer vision focuses on estimating 3D humans from images and videos, the biomechanics community focuses on highly accurate marker-based motion capture systems. In this chapter, we take a step towards combining the best of these disciplines, providing new and improved tools to each; see Fig. 4.1.

Specifically, we focus on advances in computer vision that infer the 3D pose and shape of the human body in the form of parametric body models like SMPL [43]. The field has advanced rapidly and the accuracy of markerless video-based 3D motion capture is catching up with marker-based techniques. Unfortunately, the kinematic structure of models such as SMPL is not physically accurate, limiting their applicability in biomechanics.

In the previous chapter, we presented a method to estimate a skeleton mesh in any posed body. However, OSSO suffers from SMPL's limitations and can not guarantee correct bone orientations inside the limbs.

In contrast, the biomechanics field has developed detailed skeletal models to represent the anatomic motion of the knee, spine, shoulder, etc. But the vision and graphics communities are currently not benefiting from these more accurate models of the body and its joints.

To address these issues, we unify the SMPL body model with BSM, a new Biomechanical Skeleton Model. While previous work, including OSSO, has addressed the problem of putting skeletons inside 3D body models [79, 122, 123], such approaches have not addressed the problem of precisely locating the skeleton within a *moving* body. The key challenge is the lack of training data that pairs the posed 3D human body shape with the ground-truth skeleton. We address this by creating a novel dataset called BioAMASS. To create BioAMASS, we take sequences of 3D bodies from the AMASS dataset [120] that cover a wide range of body shapes and challenging poses. We place virtual motion capture markers on the body surface to obtain pseudo-ground-truth skeletons. We then use the recent method "AddBiomechanics" [124] to solve for the BSM skeleton given the virtual markers.

With this paired dataset, we can now address several problems that were previously unsolvable. First, we train a regressor to estimate the 3D anatomical BSM joint locations of the body given a posed SMPL mesh. Note that these locations significantly differ from the joints in SMPL. This learned anatomical joint regressor is useful for generating more relevant training data for 2D or 3D joint detectors, as today, such methods are typically trained from manually labeled joints or projected SMPL joints.

Next, we re-rig the SMPL body model with the BSM biomechanical skeleton, i.e. we use the BSM parameters to drive a SMPL mesh, and we call the resulting model SKEL, which is short for "Skeletal Kinematics Enveloped by a Learned body model". To do so, the skeleton must be properly scaled, located and oriented inside the SMPL body mesh. To that end, we propose a data-driven strategy that places the bones inside the body while ensuring that their orientations are compatible with the anatomic constraints of the limbs. Like SMPL, SKEL provides a body surface but with a skeleton inside that has biomechanical degrees of freedom. For example, the spine in SKEL is modeled by a spline derived from biomechanics. Additionally, shoulders are a complicated structure that is typically crudely approximated in vision and graphics models. SKEL replaces the approximate shoulder of SMPL with a biomechanical shoulder blade [125] that slides along an ellipsoid defined around the thorax. The forearm rotation is another place where standard graphics models like SMPL differ from biomechanics. Instead of a simple rotation around the elbow, SKEL models the motion of the radius and ulna bones to drive forearm pronation and supination.

SKEL has several uses. Specifically, we consider the problem of taking a SMPL body model and computing the correct skeleton inside. Compared to the previous chapter, we now aim to achieve anatomically plausible bone orientations that do not violate the constraints imposed by the degrees of freedom of the human skeleton. To do so, we simply fit SKEL to the posed SMPL mesh by optimizing the SKEL pose to minimize the vertex-to-vertex distance between the meshes. We apply this process to archival datasets such as 3DPW [23] and BEDLAM [126]. This effectively *upgrades* existing computer vision datasets to contain biomechanical ground truth, extending their use to biomechanics. For example, one could evaluate, or learn to directly regress, biomechanical parameters from video.

We evaluate two methods for estimating the skeleton from SMPL: direct regression of BSM joints from SMPL and fitting SKEL to SMPL. Accuracy is defined in terms of 3D joint location error. Since there is no ground truth for this task, we take the joint locations estimated by AddBiomechanics as pseudo-ground truth. We find that both of our methods produce significantly more accurate joint predictions than SMPL. We also provide extensive qualitative experiments that show the articulated structure of SKEL and its use in upgrading existing human motion datasets to support biomechanics.

SKEL can also be used in the other direction. Given an input skeleton mesh obtained after fitting a biomechanical model to mocap data, SKEL can be used to add a plausible skin surface; this is useful for visualization of mocap data. Since there are an infinite number of body shapes that are consistent with a given skeleton, the predicted shape can be constrained, e.g. with the subject's weight.

At the time of publication, SKEL is the first model where the body surface and anatomical skeleton are directly controlled by the same set of shape and pose parameters $(\beta, \mathbf{q})$. The BioAMASS dataset, the code to create it from the AMASS dataset, as well as the SKEL model, are available for research purposes at `https://skel.is.tue.mpg.de`.

4.2 Related Work

As seen in Chapter 2, the accurate representation and animation of human bodies play an essential role in computer graphics, vision, and biomechanics. Here we recall the main approaches to capture and model the human body in motion.

Body models. As presented in Sec. 2.4.1, statistical body shape models [42–44, 46, 47, 88, 90, 127] have a wide range of usage, but their joint locations are not designed to correspond to the anatomical functional joints of the body. For example, their kinematic tree does not match the degrees of freedom of the human anatomic skeleton. The knee and elbow flexion, the spine, the elbow and the arm supination are typically modeled by ball joints, while those functional joints have either a single degree of freedom or are more complex than a pure rotation, such as the forearm, knee, spine, or the shoulder.

Biomechanical skeleton models. In contrast to body models, skeletal models used in biomechanics, like [74, 125, 128], define the degrees of freedom of the human skeleton with a focus on anatomic realism. This is critical for kinematic and kinetic analysis. The size and motion of these skeleton models are computed from optical motion capture data using optimization frameworks like OpenSim [71] or AddBiomechanics [124]. This is the classical approach in biomechanics for measuring the precise location of the functional joints.

Motion capture. While marker-based motion capture (mocap) is the preferred method for analyzing movement, it is expensive, invasive, and time-consuming. It is also hard to reproduce the exact marker placement on different subjects and most methods assume that the markers are rigidly attached to the body, which is not accurate due to soft tissue deformation. MoSh [120, 129] unifies mocap and statistical body models by fitting the parameters of the model to match the marker data. This approach can even mitigate the issues of soft tissue motion.

Traditional mocap, however, typically prevents subjects from wearing regular clothing, complicating capture and limiting its applications. Consequently, many research and commercial solutions for markerless motion capture exist [15–17], which are currently limited by the absence of joints detectors on images, that yield anatomically correct joints (Sec. 2.2.3).

Bones inside bodies. Our goal is to properly place the skeleton inside a parametric body model, providing the best of both worlds. A common approach in previous work uses an anatomic skeleton model and deforms it to register it to a target body mesh [79–82, 105]. This registration is challenging as these skeleton models do not, in contrast to SMPL, have a shape space of deformations. Thus, the applied deformations may create non-plausible anatomies. In contrast, in Chapter 3, OSSO learns to predict the geometry of the bones from a SMPL body mesh, leveraging medical scans. Although this approach gives a plausible skeleton shape that fits inside the subject, the resulting skeleton model can not be easily animated as it does not have a kinematic tree and can not be validated in 3D. For a lying pose, OSSO yields precise skeletal geometry that is close to the ground truth scans, but the reposing of the skeleton requires an optimization process that can lead to biomechanically impossible poses. The BOSS model [122] improves on OSSO by learning a skin-bone-organs model from segmented 3D medical data. While the skin and skeleton model share the same shape space, their kinematic trees used for rigging are different. This does not allow the synchronous posing of both skin and skeleton and an expensive optimization step is required.

Bodies from bones. Going in the other direction, one can infer the body shape given a skeleton. For example, BASH [108] uses the SCAPE body model [42] to envelop a biomechanical skeletal and muscle model [128]. However, the SCAPE model is only scaled to match the limb lengths of the skeleton. Shape accuracy is not critical because their goal is to better visualize muscle activation by displaying it on the human surface.

In contrast to prior work, SKEL provides a correctly scaled skeleton inside any SMPL body model. Any optimization or regression method that estimates SMPL parameters can now be used to produce biomechanical skeletal parameters. SKEL effectively connects parametric shape models with biomechanical skeletons for the first time to enable the integration of these technologies and fields.

4.3 Method overview

Our driving goal is to create SKEL, a model that combines skin and skeleton meshes in which both are synchronously rigged with the same pose parameters $\mathbf{q}$, and can be reshaped by inheriting the SMPL shape space. To create this model, we must know the location of the anatomic joints and bone rotations inside the human body. There is no large-scale medical dataset of subjects in motion where one can extract both the body and skeleton meshes, and static medical scans do not fully constrain the skeleton in motion. For this, we need bodies in motion and leverage the AMASS dataset [120] to address this challenge. In Sec. 4.4 we first present our new custom Biomechanical Skeleton Model, BSM, and describe how to align it inside AMASS sequences of SMPL bodies in motion to obtain the new BioAMASS dataset. Leveraging BioAMASS, Sec. 4.5 shows how we learn the $SKEL(\beta, \mathbf{q})$ model, which inherits the shape space β from SMPL and the pose vector $\mathbf{q}$ from the new BSM biomechanical model. It enables direct animation of the skin and the skeleton meshes using shape and pose parameters, β and $\mathbf{q}$, respectively. Creating SKEL involves two essential steps: learning the bone locations and orientations (Sec. 4.5.1) inside the body, and rigging the skin and bone motions to a common kinematic tree parameterized by $\mathbf{q}$ (Sec. 4.5.2).

4.4 The BioAMASS dataset

The goal of the BioAMASS dataset is to enable the learning of the location and orientation of the 3D bones inside a body surface in motion. To create BioAMASS, we use the SMPL [43] model for the body surface and a new biomechanical skeleton model, BSM, for the bones. We first recall the definition of SMPL function, introduce our biomechanical models, and then describe how we fit BSM to SMPL and create BioAMASS.

4.4.1 The SMPL body surface model

We model the 3D body surface using the SMPL function, which takes as input shape parameters β and pose parameters $\theta \in \mathbb{R}^{72}$, and outputs a 3D mesh with vertices $\mathbf{v} \in \mathbb{R}^{6890\times 3}$. The SMPL model, detailed more extensively in Sec. 2.4.1, includes a joint regressor defined in Eq. 10 of [43]. Given shape parameters β, this regressor lets us compute the location of the 3D joints inside the body mesh in T-pose, thus defining a subject-specific kinematic tree. Each joint pose is then parameterized by three degrees of freedom in an axis-angle representation.

4.4.2 The BSM skeletal model

To model the human skeleton, we create BSM, a custom skeleton model using the OpenSim framework [71]; BSM is described by a file in ".osim" format. BSM consists of 24 rigid groups of bones with joints defined between them as well as a mesh representing the geometry of each bone group. On top of each bone, a set of virtual markers is defined; these markers are used to fit BSM to motion capture sequences.

The BSM is represented by three functions that take scaling and pose parameters as input. Using forward kinematics, these functions output the skeleton joint locations, $BSM^J(\mathbf{s},\mathbf{q})$, the bone meshes vertices, $BSM^v(\mathbf{s},\mathbf{q})$, and the posed marker locations, $BSM^m(\mathbf{s},\mathbf{q},\mathbf{m_0})$. The scale parameter $\mathbf{s} \in \mathbb{R}^{24\times 3}$ scales each of the 24 unposed bones along the axis (x,y,z), while the pose parameters $\mathbf{q} \in \mathbb{R}^{49}$ represent the 49 degrees of freedom of the articulated model. The model markers are defined by designating their 3D coordinates $\mathbf{m_0} \in \mathbb{R}^{N_m \times 3}$ in the corresponding bone reference frame. Each marker is rigidly attached to one bone and, when the bone is scaled with $\mathbf{s}$, the marker location is scaled accordingly. In contrast to SMPL, BSM has a more realistic kinematic tree but lacks a shape space.

Body models like SMPL typically treat every joint as a ball joint with three angular degrees of freedom. In reality, the joints of the body differ significantly from this assumption. Consequently, for BSM we use more realistic models of the spine, shoulder, and forearm. This results in a model with fewer degrees of freedom.

Lower body. For BSM's lower body model, we use the model from Rajagopal et al. [74], which implements the knee flexion model from Walker et al. [130].

Spine. We extend the original OpenSim framework with a new custom joint that we call "constant curvature", to model the spine bending with a constant length. Our BSM model's spine comprises 3 such joints, enabling lumbar, thoracic, and cervical bending, as illustrated in Fig. 4.10a. Given the parent joint location J_{i-1} and a spine curve of length l, the child joint J_i will move on a curve of constant arc length and curvature, parameterized by one termination angle $\mathbf{q}_i = [q_x, q_z, q_y] \in \mathbb{R}^3$, represented as Euler-angles in XZY. The child joint location is $J_i = R(\mathbf{q}_i) \cdot (J_{i-1} - J_i) + \mathbf{t}^{\text{spine}}(\mathbf{q}_i)$, where

$$\mathbf{t}^{\text{spine}}(\alpha) = \begin{cases} x = & (r - \cos(\alpha)) * \frac{-sin(q_z)}{sin(\alpha)} \\ y = & r \cdot \sin(\alpha) \\ z = & (r - \cos(\alpha)) * \frac{cos(q_z) * sin(q_x)}{sin(\alpha)} \end{cases} \tag{4.1}$$

$$\text{with } \alpha = \arcsin \sqrt{sin(q_z)^2 + (cos(q_z) * sin(q_x))^2}, \text{ and } r = \frac{l}{\alpha}.$$

Shoulder blades. In BSM we follow the scapulothoracic model from [125] and parameterize the shoulder blade joint such that it slides along an ellipsoid defined around the thorax, making the scapula slide along the ribs. The three degrees of freedom are linked to scapula abduction, elevation, and upward rotation as illustrated in Fig. 4.2.

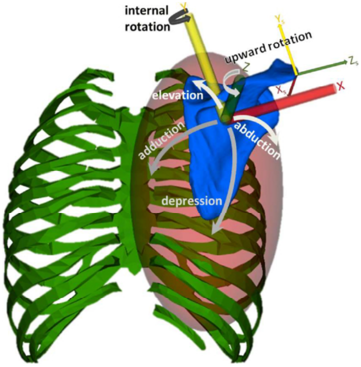
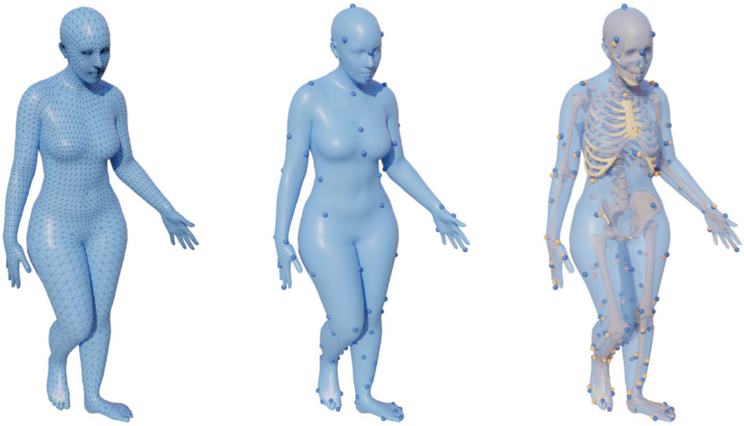


Figure 4.2: The shoulder blade model used in BSM (figure from [125]).

Forearm. The forearm pronation and supination are modeled by a single degree of freedom; which is distinct from the elbow flexion, wrist flexion, and wrist deviation [74]. The forearm is made of two bones: the radius and the ulna. The ulna is linked to the humerus through a hinge joint, enabling the elbow flexion. During the forearm pronation and supination, the hand rotates while the ulna stays fixed. We model this by rotating the radius along the axis defined by the ulna's parent joint location and the radius extremity as illustrated in Fig. 4.10c.



(a) SMPL (b) Synthetic markers (c) BSM alignment

Figure 4.3: Creation of the paired skeleton and body dataset. Given a SMPL motion sequence (a), we generate synthetic markers (b), and fit a biomechanical model to the markers using AddBiomechanics [124] (c).

4.4.3 Fitting BSM to SMPL

To leverage the SMPL body meshes in AMASS, we define a mocap marker set on the SMPL mesh and obtain synthetic sequences of markers. We use these as input to fit our BSM skeleton using AddBiomechanics [124], a recent biomechanical optimization framework. Fig. 4.3 illustrates this pipeline.

Establishing marker correspondences with BSM

To fit BSM to SMPL, we need to define corresponding markers on both models. Theoretically, we could define each skin vertex of SMPL to be a marker attached to BSM. But OpenSim rigidly attaches markers to the bones, hence we define a set of markers that are influenced mainly by one bone and not subject to significant soft tissue deformation.

Specifically, we define 54 *bony markers* that are close to the bones, as typically done in motion capture. Each marker is defined on BSM and SMPL by examining tight SMPL fits to 3D scans and identifying specific SMPL vertices. Figure 4.4 shows all the *bony markers* on SMPL in orange.

Although this marker set follows the rigidity assumption, it is too sparse in some areas to properly constrain the location of the bones. So we introduce an additional 52 *soft*

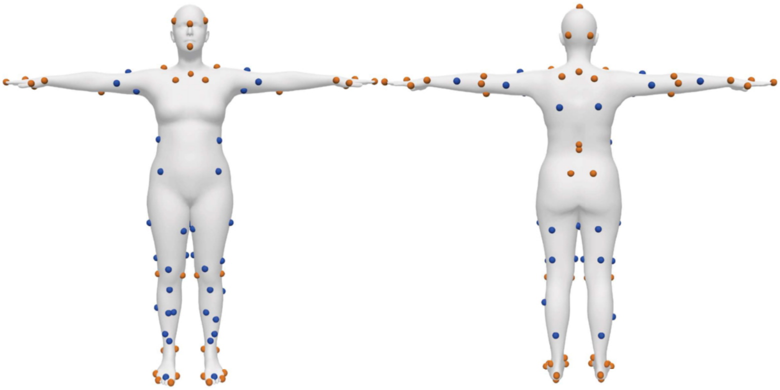


Figure 4.4: The markers defined on SMPL: bony in orange, soft in blue.

markers, located on soft body parts (blue in Fig. 4.4). To define a new BSM marker, it needs to be positioned on the BSM skeleton template. While this can be achieved quite precisely for *bony markers*, it is harder to estimate at what distance to the bones *soft markers* should lie. Moreover, this distance varies significantly for different body shapes (e.g. due to adipose tissue). Initializing markers close to the bones for subjects with more adipose tissue can lead to AddBiomechanics over-stretching the bones to fit the SMPL markers.

To address this marker offset issue, we propose a method to automatically define markers on BSM with personalized offsets depending on the body shape. We leverage the OSSO model [123], which predicts the location and shape of the skeleton inside SMPL. In contrast to BSM, OSSO models the geometry of the skeleton with respect to the body shape and, as it was trained on medical scans, it learned the offset between the bones and the skin. We can thus use it to compute where skin markers should be located with respect to the bone surface, given a body shape. We first compute the relationship between the OSSO and BSM bones. Precisely, we register each OSSO bone mesh to the corresponding BSM bone mesh and effectively obtain all OSSO bones in the reference frames of the BSM bones. This relationship only needs to be computed once. Then, for each AMASS subject, we use OSSO to obtain their skeleton mesh. We use the lying down pose in which OSSO is trained to obtain the best possible OSSO prediction. Now, given a marker location on the SMPL mesh and the computed OSSO bone mesh inside the body, we parameterize the marker location using the closest triangle on the OSSO bone mesh (Fig. 4.5a). This allows us to transfer the marker location onto the OSSO bone mesh and, consequently to the corresponding template BSM bone (Fig. 4.5b).

We use this method to generate a BSM model for each subject, with personalized markers $\mathbf{m_0}(\beta)$, thus avoiding over-stretching the bones during the AddBiomechanics optimization, as shown Fig. 4.5c. We experimented with different marker sets, adjusting their number and placement, to obtain the best possible fits from AddBiomechanics; i.e. minimizing the marker errors and yielding a satisfactory fit visually.

Fitting BSM to motion data

With corresponding markers defined on both SMPL and BSM, we can fit the BSM skeleton to any SMPL mesh. Given a sequence of N_f frames and N_m target 3D marker locations per frame, $\mathbf{m}_k^T$ $(k \in \{1, \ldots, N_m\})$, extracted from the sequence of SMPL meshes, we use AddBiomechanics [124] to obtain the BSM scale parameters $\mathbf{s}$ and the N_f poses $\{\mathbf{q}_f\}$. We optimize a bi-level objective, to find the best $\mathbf{s}$ such that inverse-kinematics with these scales yields poses $\{\mathbf{q}_f\}$ with minimal distance to the N_m target markers:

$$\arg\min_{\mathbf{s},\delta} \left(\Big(\sum_{f=1}^{N_f} \arg\min_{\mathbf{q}_f} \sum_{k=1}^{N_m} \lambda_k (\mathrm{BSM}^m(\mathbf{s}, \mathbf{q}_f, \mathbf{m_0}(\beta) + \delta)_k - \mathbf{m}_k^T) \Big) + \lambda_p P(\mathbf{s}, \beta) \right), \quad (4.2)$$

where $\delta \in \mathbb{R}^{N_m \times 3}$ is a 3D per marker offset. The weighting factor $\lambda_k \in \mathbb{R}$ is set to a low value for *soft markers* and a high value for *bony markers* to allow larger fitting errors due to secondary soft tissue motions.

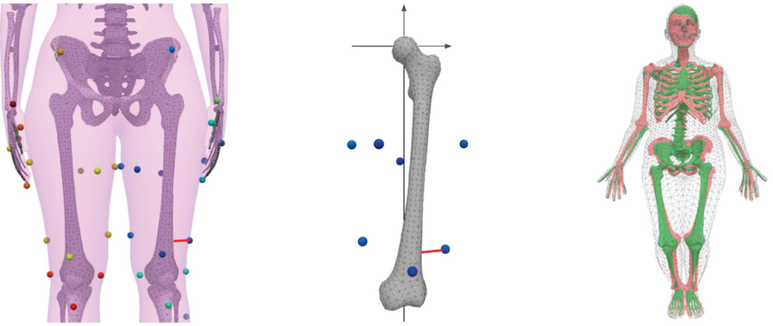
The prior P regularizes the scale of the bones, given the subject's height, weight, and biological sex as in [124]. We automatically estimate the height and weight of each subject from their SMPL shape parameters β, by assuming that the body has a uniform density [10, 131] and thus re-parameterize this prior term as $P(\mathbf{s}, \beta)$.

Despite the scale prior, using a generic marker set can lead to AddBiomechanics over-stretching the bones for heavy subjects. Defining personalized marker locations $\mathbf{m_0}(\beta)$ on the skeleton template as described in the previous section helps further regularize the bone scales (Fig. 4.5c).

We apply this optimization process to a subset of AMASS consisting of 113 subjects and 2198 motion sequences, amounting to over 9 hours of motion data. The paired SMPL meshes and BSM skeletons form the BioAMASS dataset. For each subject p there is a SMPL body shape β_p and the scaled personalized BSM model $\mathbf{s}_p$. Further, for each motion frame f it includes the bone angles $\mathbf{q}_f$ as well as the bone joint locations $\mathbf{J}_f^o$. Figure 4.6 shows examples of the BioAMASS dataset.

4.5 The SKEL model

Now that we have a dataset of BSM skeletons inside SMPL, we can now learn a corrected kinematic tree for SMPL, that matches i) the proper bone orientations and ii) the anatomical joint locations. Effectively, we aim to parameterize the 3D body model with the biomechanical skeleton. To that end, we develop SKEL, which is designed to be



(a) Markers wrt OSSO

(b) Markers in the BSM bone frame

(c) AddBiomechanics fit result

Figure 4.5: **(a)** The OSSO skeleton is aligned to the subject's SMPL mesh. **(b)** We deduce the personalized markers location $\mathbf{m_0}(\beta)$ on the BSM bone template. **(c)** On high BMI subjects, a shape-agnostic marker definition for all subjects yields over-stretched bones (**red**). Using personalized marker locations $\mathbf{m_0}(\beta)$ defined using OSSO prevents this over-stretching (**green**).

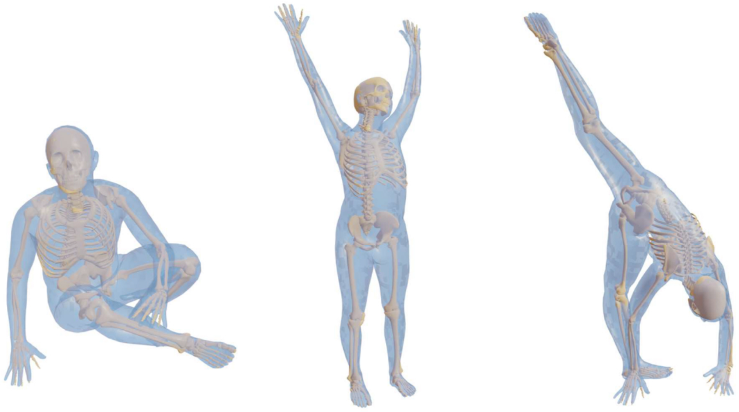


Figure 4.6: BioAMASS: examples of BSM fits to AMASS poses.

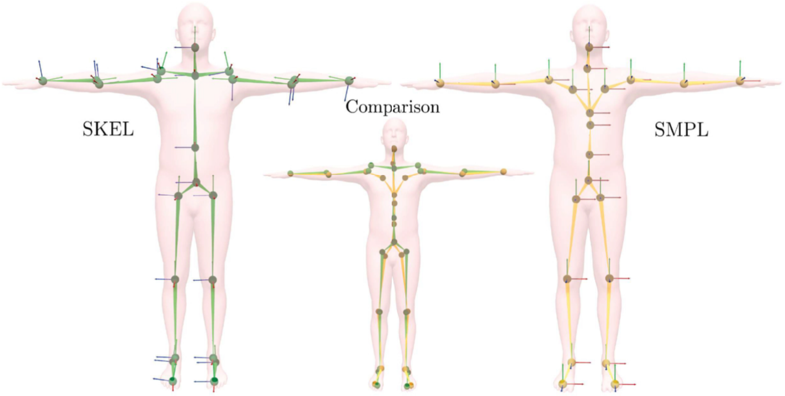


Figure 4.7: **Left:** SKEL kinematic tree with learned anatomical joint locations. **Right:** SMPL's kinematic tree. **Middle:** the superposition of both. In contrast to SMPL, which has axis-aligned rotation axes, SKEL's rotation axes are bone-aligned.

compatible with SMPL and posed like BSM. This allows us to leverage SMPL's learned shape space as well as all the existing datasets where SMPL bodies are estimated from different modalities. To create SKEL, we must put the corresponding SMPL limbs and BSM bones together in the same reference frame. To that end, in Sec. 4.5.1, we learn to regress the anatomical joint locations from SMPL's mesh using the BioAMASS dataset. We also learn the relation between the orientation of SMPL's limbs and the orientation of the underlying bones. These orientations, together with the anatomical joint locations, define a kinematic tree inside SMPL's T-pose. Then, in Sec. 4.5.2 we describe how we rig the SMPL model and the BSM joint to this learned kinematic tree.

Note that, in SMPL, all joint orientations are defined in a global T-pose space with an axis-aligned frame of reference for each joint as illustrated in Fig. 4.7 right. This means that SMPL assumes, for example, that the elbow rotation axis is aligned with the world y-axis, independent of the orientation of the humerus. The overparametrized nature of SMPL allows plausible arm articulation by combining several axis rotations. But BSM, with its reduced degrees of freedom for the rotations, requires the local frame on which the rotation is applied to be precisely aligned with the anatomy in order to obtain a proper anatomic rigging. In addition, the location and orientation of the humerus and ulna bones have to be coherent with the rotation axis. In SMPL this coherence does not exist: the joint reference frames are not aligned with the articulation axis. As shown in Fig. 4.7, the elbow frame is not aligned with the segment defining the humerus position. Hence, we first learn to predict the location of the joints inside SMPL and, with these, we learn to properly orient the bones inside a SMPL body mesh.

4.5.1 Establishing the bone locations and orientations

Anatomical joint locations. Given paired SMPL body meshes and their corresponding BSM anatomic joint locations, we learn a function that predicts the joints from the body surface. We proceed similarly to Loper et al. [43] by learning a joint regressor $\mathcal{J}$ that takes as input the SMPL mesh vertices $\mathbf{v}^{\text{smpl}} \in \mathbb{R}^{6890\times 3}$ and predicts the new anatomic joints $\mathbf{J}^{o} \in \mathbb{R}^{24\times 3}$. We follow the same methodology as used in Sec. 3.4.1, by formulating a non-negative least squares problem for each joint i, and solving it with an active set method [117]. We train these regressors from the posed vertices and joints of the BioAMASS dataset.

Figure 4.7 shows the new regressed kinematic tree in green. Notice that the hip joint locations, corresponding to the femur heads, are more anatomically correct than the ones in SMPL. The comparison also shows significant differences at the shoulders, as well as more subtle, but important, differences for the other joints.

Bone orientations. We aim to find the orientation of the bones inside the SMPL T-pose mesh, i.e. find the rotation R_i to apply to the i-th BSM bone template mesh, to position it inside the SMPL T-pose mesh. In BSM, the rest position of each individual bone template is centered at the origin and oriented along the canonical axis x, y, z. In the following, we refer to the "bone axis" as the axis passing through the bone's proximal and distal ends.

In contrast to BSM, in SMPL T-pose, the bones should be positioned and oriented between pairs of regressed anatomical joints. This brings two challenges: (i) the rotation of the bone around its bone axis is not known, and (ii) as the regressed joint location depends on β, the orientation of the bones also varies with β.

To solve those two issues, we split the bone rotation $R_i(\beta)$ into a learned base rotation R_i^{base} and a shape-dependant rotation $R_i^{\beta}(\beta)$:

$$R_i(\beta) = R_i^{\beta}(\beta) \cdot R_i^{base}, \tag{4.3}$$

where R_i^{base} is learned to define the bone's orientation around its bone axis, ensuring that bones are properly orientated wrt their parent bone. $R_i^{\beta}(\beta)$ is computed dynamically to align the bone to the segment defined by its parent and child joints, so that the bone stays in its socket regardless of the shape of the subject.

First, we learn R_i^{base} from BioAMASS. For each bone i, we can define a corresponding SMPL joint and limb. For example, the right humerus bone corresponds to the 17th joint and the right upper arm of SMPL. Thus, for each frame f of our dataset, we obtain the bone BSM rotation $R_{i,f}^{B}$ and its SMPL rotation $R_{i,f}^{S}$.

R_i^{base} is the rotation that the bone needs to undergo so that when chained to the SMPL rotation, the corresponding BSM rotation is obtained. So, for each bone i we learn its

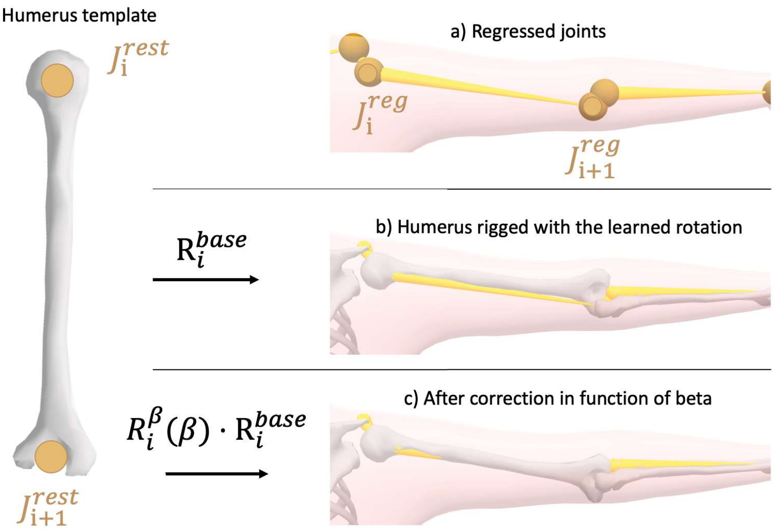

Humerus template
J_i^{rest}
J_{i+1}^{rest}
a) Regressed joints
J_i^{reg}
J_{i+1}^{reg}
R_i^{base}
b) Humerus rigged with the learned rotation
$R_i^\beta(\beta) \cdot R_i^{base}$
c) After correction in function of beta

Figure 4.8: Left: Humerus template in the rest pose. We want to find its transformation to position it inside SMPL's T-posed arm. On the right we show, in order: a) the anatomical bone joints J_i^{reg} regressed from SMPL skin vertices (pink). b) We translate the bone to J_i^{reg} and orient it with R_i^{base}. This provides a rough alignment, rotating the bone properly around its bone axis. c) We then compute and apply the personalized rotation $R_i^\beta(\beta)$ to perfectly align the bone with the limb segment. Notice how the ulnar head now properly fits in the humerus distal end.

base rotation R_i^{base} by minimizing:

$$\sum_{f=1}^{N_F} (R_{i,f}^B - R_{i,f}^S R_i^{base})^2 \tag{4.4}$$

over the N_F frames of the dataset.

This rotation properly orients the bone around its bone axis. But, as shown in Fig. 4.8, this rotation alone does not guarantee that the bones are aligned between their T-pose parent and child joints.

Thus we explicitly compute $R_i^\beta(\beta)$, a shape-dependent corrective rotation that aligns the bone segment $(R_i^{base}\,(J_{i+1}^{rest} - J_i^{rest}))$ with $(J_{i+1}^{reg}(\beta) - J_i^{reg}(\beta))$, where J_i^{rest} is the location of joint i in the bone rest pose and $J_i^{reg}(\beta)$ is the shape-dependent regressed joint.

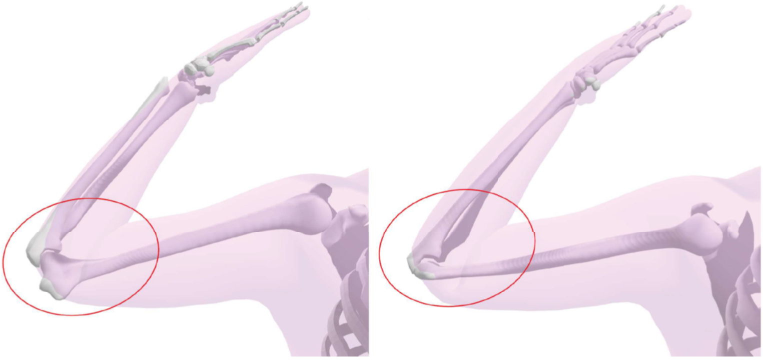


Figure 4.9: Left: Rigging the skeleton to the regressed joints and posing them using SMPL parameters θ can yield unrealistic articulations. We see that the humerus posed with the SMPL upper arm transformation does not yield the correct humerus orientation. Right: BSM fit for the same frame.

The rotation axis of $R_i^{\beta}(\beta)$ is computed from the cross-product of the segments. As shown in Fig. 4.8, this effectively ensures a proper fit of the bone geometry into the regressed joint location.

It is worth noting that computing a direct rotation between the rest bone and the regressed segment $(J_{i+1}^{reg}(\beta) - J_i^{reg}(\beta))$ leaves a degree of rotation open: the rotation around the bone axis. With the proposed approach, we obtain an anatomically coherent placement of the skeleton. Thanks to BioAMASS, a consensus orientation R_i^{base} is found, which is then specialized per subject with $R_i^{\beta}(\beta)$. Effectively, the per-joint R_i^{base} is learned from the dataset once and kept fixed, and each per-joint $R_i^{\beta}(\beta)$ is recomputed when the shape parameters change.

4.5.2 Building SKEL: A single rig for skin and bones

As we saw in Fig. 4.7, the SMPL kinematic tree is not suited to rig the skeletal structure, as its joints do not match the anatomic ones. Moreover, because of its over-parameterization, applying SMPL's transformation to the bones can yield unrealistic bone orientations, as shown in Fig. 4.9.

Consequently, we re-rig SMPL with new anatomic degrees of freedom using the learned bone locations and orientations (Sec. 4.5.1).

The SKEL function. The SKEL function takes as input a vector of SMPL shape parameters β, and the $\mathbf{q} \in \mathbb{R}^{49}$ pose parameters of BSM. SKEL outputs $(\mathbf{v}^{\text{skin}}, \mathbf{v}^{\text{skel}}, \mathbf{J}^{o})$ where $\mathbf{v}^{\text{skin}}$ are the body surface vertices, $\mathbf{v}^{\text{skel}}$ the skeleton mesh vertices, and $\mathbf{J}^{o}$ the learned anatomic joint locations.

Skin. SKEL builds on the additive approach of SMPL, starting with a mean template mesh $\mathbf{T} \in \mathbb{R}^{6890\times 3}$ and adding the learned displacements $B_S(\beta) + B_P(\mathbf{q})$, where B_S is the shaping function presented in Sec. 2.4.1 and $B_P(\mathbf{q})$ are pose dependent displacements. The posed SKEL body vertices $\mathbf{v}^{\text{skin}}$ are then computed with the following linear blend skinning equation:

$$\mathbf{v}^{\text{skin}}(\beta, \mathbf{q}) = \left[\sum_{i=1}^{N_j^o} W_i^{\text{skin}} G_i^{\text{skin}}(\mathbf{q}, \beta) \right] (\mathbf{T} + B_S(\beta) + B_P(\mathbf{q})) \tag{4.5}$$

where $G_i^{\text{skin}}(\mathbf{q}, \beta)$ is a rigid transformation that will be defined in Eq. (4.6). It translates and rotates the vertices associated with the i-th limb depending on the pose parameter $\mathbf{q}$. W_i^{skin} is a 6890×24 matrix of skinning weights indicating how the vertices of the SMPL mesh are affected by each rigid transformation i. Those weights are inherited from SMPL, by defining a corresponding SMPL joint for each of the $N_j^o = 24$ joints of SKEL.

To define the transformations G_i^{skin} we use the composition of rigid transformations $T(\mathbf{R}, \mathbf{t})$ defined by a rotation matrix $\mathbf{R}$ and a translation $\mathbf{t}$, as well as per-joint local transformations $G_k^B(\mathbf{q}_k, \beta)$, which are pure rotations for most joints, and a combination of rotation and translation for the spine and shoulder blades as explained in Sec. 4.4.2. The global transformation to apply to the skin vertices is computed as:

$$G_i^{\text{skin}}(\mathbf{q}, \beta) = \prod_{k=0}^{i} T(R_k(\beta), \mathbf{J}_k^o(\beta))\; G_k^B(\mathbf{q}_k, \beta)\; T(R_k(\beta), 0)^{-1}\; T(0, \mathbf{J}_k^o(\beta))^{-1}. \tag{4.6}$$

The green term transforms the i-th limb vertices back to the unposed bone space, by centering it on its joint location $T(0, \mathbf{J}_k^o(\beta))^{-1}$, and undoing the T-pose bone rotation $T(R_k(\beta), 0)^{-1}$. Then, the joint-specific transformation $G_k^B(\mathbf{q}_k, \beta)$ is applied. Finally, the bone vertices are posed back to SMPL's posed space by applying the rotation $R_k(\beta)$ and the translation $\mathbf{J}_k^o(\beta)$. $\mathbf{J}_k^o(\beta)$ is the k-th joint location in T-pose ($\mathbf{q} = 0$) as defined in Eq. (4.7). The leading product enforces the kinematic tree structure.

The pose-dependent deformations of SKEL are inherited from SMPL. For the degrees of freedom of SKEL prone to cause a candy wrapper effect on the body mesh, we define a corresponding degree of freedom of SMPL and transfer the pose-dependent deformations $B_P(\mathbf{q}_i)$. For SKEL's joints that do not have an equivalent joint in SMPL, we default to linear blend skinning with no pose correctives.

It is important to note that these blend shapes are approximate because certain degrees of freedom, like knee flexion, have slightly different axes and neutral positions in SKEL compared to SMPL. While this transfer is not optimal and creates artifacts in extreme poses (see the video at https://skel.is.tue.mpg.de/), SKEL can match SMPL meshes with an average vertex-to-vertex error below 3 cm; see Fig. 4.13. We leave the learning of SKEL-specific pose-dependent deformations using BioAMASS for future work.

Joints. The locations of SKEL's unposed joints are regressed from the unposed skin vertices $\mathbf{v}^{\text{skin}}(\beta, \mathbf{q} = 0)$ with the learned anatomical joint regressor $\mathcal{J}$, to get the unposed joints $\mathbf{J}^o(\beta)$. These joints are then posed with the parameter $\mathbf{q}$, like the skin vertices, by applying the rigid transformations G_i^{skin}:

$$\mathbf{J}^o(\beta, \mathbf{q}) = \left[\sum_{i=1}^{N_j^o} W_i^J G_i^{skin}(\mathbf{q}, \beta) \right] \mathbf{J}^o(\beta) \tag{4.7}$$

only with different weights W_i^J that ensure that the proper joint is affected by the transformation. Note that for SKEL we use a simplified hinge joint at the knee.

Skeleton. To obtain the shaped and posed skeleton mesh, a similar equation is used. We name the initial skeleton template mesh $\mathbf{T}^o$ in which every bone mesh is axis-aligned and has its parent joint at the world's origin (Fig. 4.5b middle shows the unposed template femur). This mesh is scaled using $s(\mathbf{J}^o(\beta))$, a per-bone 3D scaling factor defined by the regressed joint locations, namely the limb lengths they define. Then, the scaled vertices are posed to obtain the posed skeleton vertices

$$\mathbf{v}^{\text{skel}}(\mathbf{q}, \beta) = \left[\sum_{i=1}^{N_j^o} W_i^{\text{skel}} G_i^{\text{skel}}(\mathbf{q}, \beta) \right] (s(\mathbf{J}^o(\beta)) \circ \mathbf{T}^o) \tag{4.8}$$

where W_i^{skel} are boolean per-bone weights, except for the spine and rib cage where the weights are interpolated to be 0 at the bottom of the spine section and 1 at the top. The skeleton vertex transformations are computed with

$$G_i^{\text{skel}}(\mathbf{q}, \beta) = \prod_{k=0}^{i} T(R_k(\beta), \mathbf{J}_k^o(\beta)) G_k^B(\mathbf{q}_k, \beta) \tag{4.9}$$

in which the unposed bone mesh is transformed by the joint transformation $G_k^B(\mathbf{q}_k, \beta)$, then oriented with $R_k(\beta)$ to be aligned with the limb's skin and translated to its T-pose joint $\mathbf{J}_k^o(\beta)$.

Finally, we define the range of possible angles for specific degrees of freedom like the shoulder blades, knee, arms, and spine motions. Figure 4.10 illustrates SKEL's degrees of freedom for the spine, shoulder blades, and arm pronation. Note that the deformation of the body surface (pink) is driven by the BSM pose, thus combining the SMPL surface model with an anatomical skeleton.

4.6 Evaluation

In this section, we evaluate the fit accuracy of the BioAMASS dataset, the learned anatomical joint regressors, and the skeleton meshes obtained by fitting SKEL to SMPL meshes.

4.6.1 Evaluating BioAMASS fits

In Sec. 4.4 we simulate optical motion capture markers on SMPL sequences and fit the BSM biomechanical skeleton to them. We evaluate these fits by computing the Mean Absolute Error (MAE) between the target and the fitted markers. In Tab. 4.1, for each subset of AMASS, we report the average error of bony and soft markers across all frames. For comparison, these distances are similar to the body shape reconstruction error from markers reported in [129] and significantly more accurate than the held-out marker error [129].

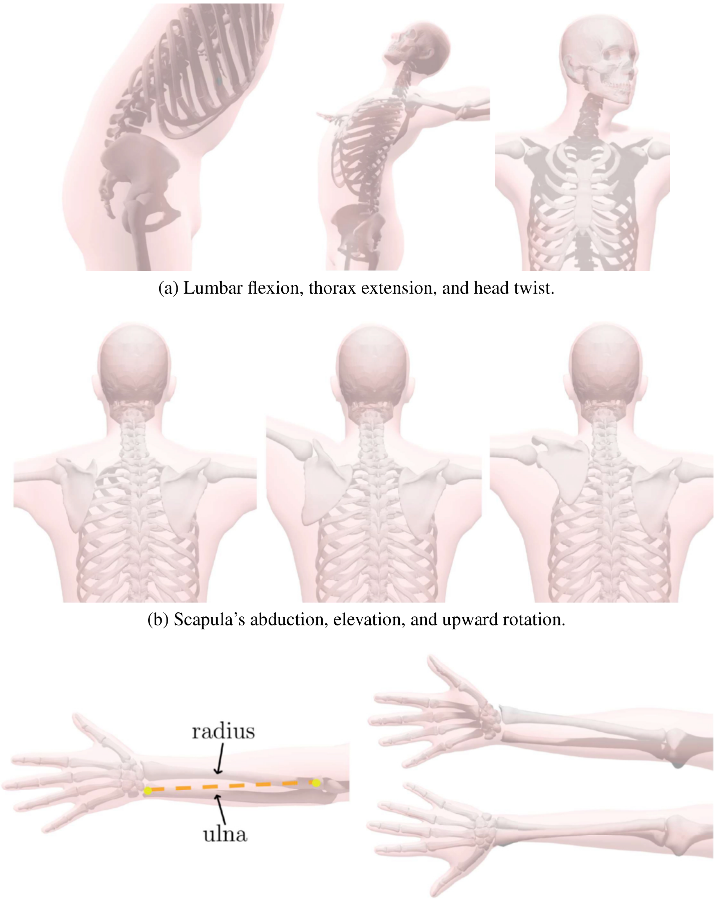
4.6.2 Joint regressors

We evaluate the regressors learned in Sec. 4.5.1 on unseen body meshes by comparing the regressed values with the reference BSM alignment. We train our anatomical joint regressors on the CMU [132] and MPI_Limits [133] datasets, which are part of AMASS [120]. CMU contains good variation in body shape, while MPI_Limits contains extreme poses. Once trained, we evaluate our regressor on the DFAUST dataset [121], with various motion sequences for 10 subjects with diverse BMIs; DFAUST contains precise SMPL fits to 3D scan sequences.

For each frame of the DFAUST dataset, BioAMASS provides the anatomical joint locations $\mathbf{J}^o$ that we consider ground truth. Then, from the frame's SMPL mesh, we use our learned joint regressor to regress the anatomical joint location $\mathbf{J}^{reg}$. In Fig. 4.11, we report the per joint regression errors $|\mathbf{J}_i^{reg} - \mathbf{J}_i^o|$, which are below a centimeter for most

Table 4.1: Marker fitting error of the BSM model on the AMASS dataset.

MAE in (cm)	Bony markers	Soft markers
DFAUST	1.54	2.00
CMU	1.70	2.37
MPI_Limits	1.70	2.37



(a) Lumbar flexion, thorax extension, and head twist.

(b) Scapula's abduction, elevation, and upward rotation.

(c) Left: axis of rotation of the radius (lateral view). Right: forearm supination and pronation (top view).

Figure 4.10: Illustration of SKEL's degrees of freedom. The bone and body surface meshes are controlled by the same kinematic tree.

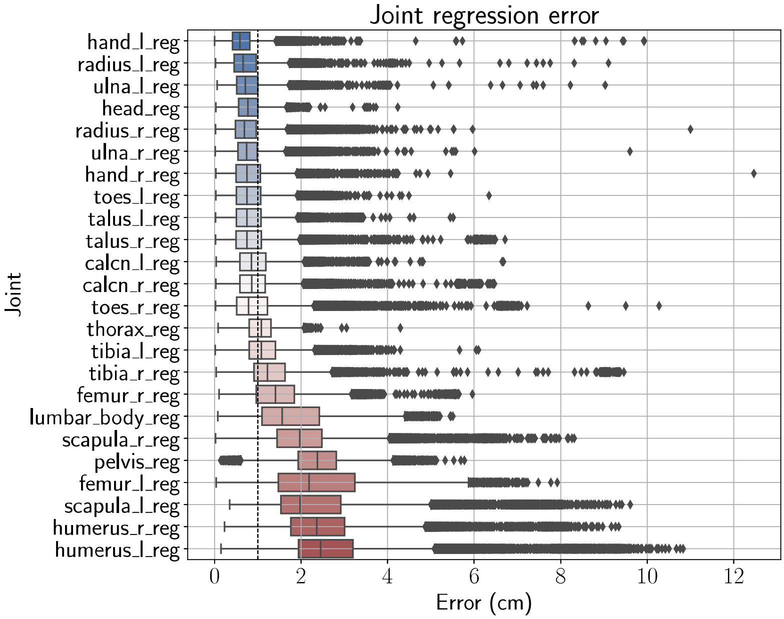


Figure 4.11: Anatomical joints regression error over the female DFAUST dataset.

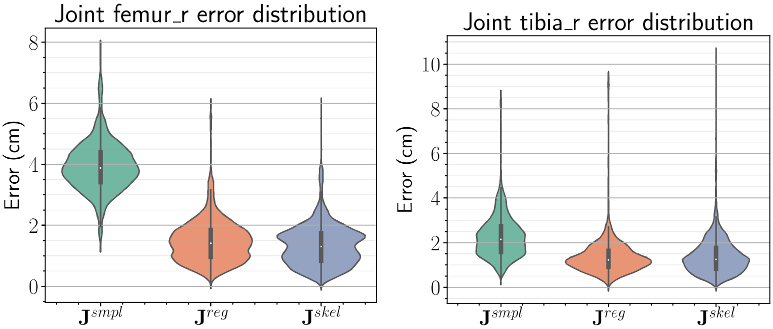


Figure 4.12: On DFAUST female subjects, we predict the joint locations and show the Euclidean distance errors wrt the “ground truth” BSM joint location for the right femur (left) and right tibia (right). We compare three methods: $\mathbf{J}^{smpl}$: using the joints directly from the SMPL fit to the DFAUST bodies. $\mathbf{J}^{reg}$: joint regressed from the SMPL mesh using our learned anatomical joint regressor. $\mathbf{J}^{skel}$: anatomical joints obtained by fitting SKEL to the SMPL mesh.

joints. Some joints, such as the humerus, have higher errors. We inspected the outlying frames and observed some failure cases of the AddBiomechanics fits for the shoulder joints, which explains the higher values. In these cases, the regressed anatomical joints are more plausible than those obtained with BSM, as shown in the project’s website video[1].

Further, we evaluate the femur and tibia joint location given by different methods as shown in Fig. 4.12. We consider $\mathbf{J}^{o}$ as the ground truth joint locations and compute the 3D Euclidean distance error of the joints given by SMPL, $\mathbf{J}^{smpl}$, the anatomical joints we regress from SMPL, $\mathbf{J}^{reg}$, and the anatomical joints, $\mathbf{J}^{skel}$, obtained by fitting SKEL to the SMPL mesh. As expected, the SMPL joints have higher errors than the learned anatomical ones.

4.6.3 SKEL fits to SMPL

Since SKEL has the same surface mesh topology and shape parameters β as SMPL, it can be directly fit to existing SMPL meshes by optimizing its pose parameters to minimize the vertex-to-vertex error.

To quantitatively evaluate how similar SKEL shapes are to SMPL, we consider motion sequences from the DFAUST dataset and their SMPL fits with 10 shape parameters. We fit SKEL to each of these SMPL meshes by optimizing its pose parameters $\mathbf{q}$. To

[1] https://skel.is.tue.mpg.de/

evaluate the mesh fits, we compute the mean absolute difference (MAD) between SKEL skin vertices and the target SMPL vertices, and then average over all the frames. For males, we find an average difference of 1.1 cm and an average max difference of 2.5 cm, while for females we obtain an average mean difference of 0.9 cm and max of 1.9 cm. A visualization of these differences on the SMPL body mesh is shown in Fig. 4.13. The larger differences can be explained by the approximate pose-dependant blend shapes inherited from SMPL, which could be retrained in future work.

Fitting SKEL to SMPL provides joint locations with similar accuracy as the regressed ones, as reported on Fig. 4.12. Let us note that direct joint regression is faster than estimating the SKEL model fit. Applications that require the joint locations but not the skeleton pose parameters, and for which time is critical, should prefer the direct regression approach.

Upgrading SMPL datasets with SKEL. Since SKEL is compatible with SMPL, we can fit SKEL to SMPL meshes from the 3DPW dataset [23] and the synthetic BEDLAM [126] dataset (Fig. 4.14). The whole sequences are shown in the video on the SKEL project website[2]. This effectively upgrades these datasets to include anatomical joint locations and biomechanical pose parameters.

4.6.4 Qualitative comparisons with OSSO

SKEL fits to SMPL also yield anatomically correct orientations of the bones. To illustrate this, we compare the SKEL predictions to OSSO skeletons (Chapter 3) on the MOYO dataset [119]. The SKEL skeletons yield more anatomically correct joint locations and biomechanically relevant bone angles, as visible in Fig. 4.15; see, for example, the knee orientation as well as arm supination. See the video at https://skel.is.tue.mpg.de/ for more examples.

We also compare the SKEL fits and OSSO fits to SMPL meshes of the Total Capture dataset [134]. The results in Fig. 4.16 illustrate that SKEL provides better bone locations and orientations. This is particularly visible in regions such as the wrist and knee, where in SKEL, the femur is correctly oriented with respect to the tibia, and the ulna is rotated to stay connected to the wrist.

4.6.5 Disentangling body shape and bone lengths

Since our skeleton mesh is entirely defined by the joint segment lengths, we can modify the body shape of a person, while maintaining their skeletal identity. This can be helpful for generating a plausible skin mesh from a given biomechanical skeleton. As illustrated in Fig. 4.17, we optimize the SKEL shape parameters β to fit a subject's limb lengths with different target weights. This results in body meshes with different body shapes but the same bone lengths.

[2] https://skel.is.tue.mpg.de/

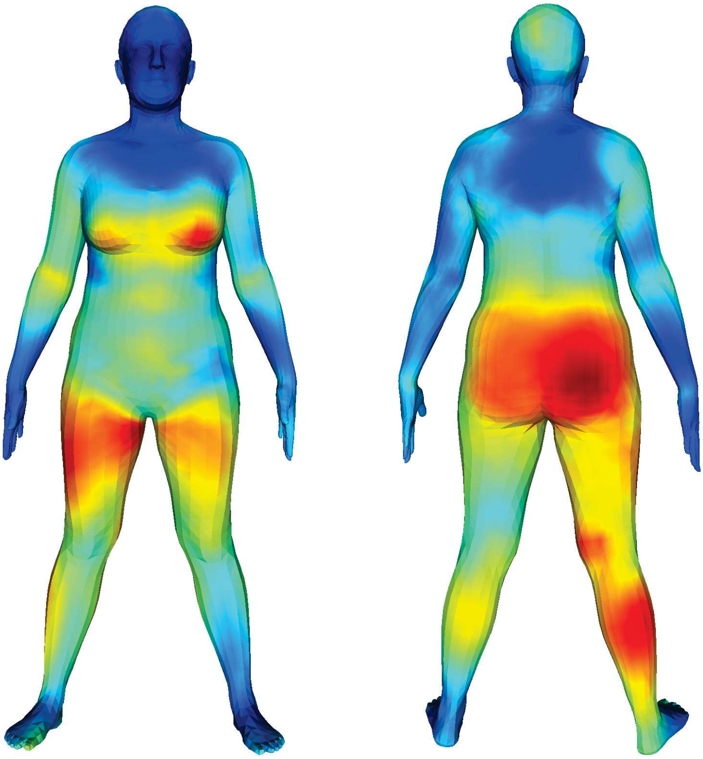


Figure 4.13: Average per-vertex distance between SKEL and SMPL fit to the females of the DFAUST dataset. Blue: 0 cm, Red 2cm.

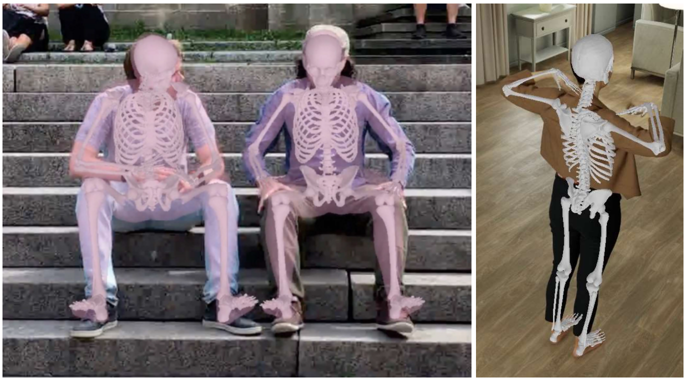


Figure 4.14: SKEL can be fit to existing SMPL datasets to upgrade them with biomechanical pose parameters. Left: SKEL skeleton mesh on a frame of 3DPW [23]. Right: SKEL skeleton mesh on a frame of BEDLAM [126].

4.7 Discussion and Conclusion

In this chapter, we described SKEL, a new parametric 3D human body shape model driven by anatomically sound parameters, providing consistent skin and bone geometries. SKEL is learned from BioAMASS, a new dataset of skeletons inside SMPL meshes in diverse AMASS poses. We build BioAMASS by optimizing BSM, a new biomechanically accurate skeleton model, to fit inside SMPL mesh sequences. Using this paired internal and external data, we then learn a regressor from SMPL mesh vertices to the anatomic joint locations and bone orientations. SKEL inherits the shape space from SMPL and the anatomic kinematic parameters from BSM. From the point of view of vision and graphics, the new model can be used in place of SMPL and it has fewer and more anatomically sound pose parameters (46 for SKEL vs 72 for SMPL). Moreover, our learned regressor enables regressing more accurate anatomic joints from video compared to current approaches solely based on SMPL joints. From the biomechanics point of view, SKEL provides a shape space, which is advantageous to adapt the model to varied body shapes without overstretching certain bones. In addition, it provides an animatable model that can take BSM poses and add a SMPL skin for visualization.

Compared to the previous chapter, we now predict a skeleton inside the skin that has anatomical joint locations and bone orientations, and we can repose the body and skeleton mesh simultaneously.

BioAMASS accuracy limitation. Although the skeletal structures and joint locations computed by AddBiomechanics are anatomically plausible, they should not be consid-

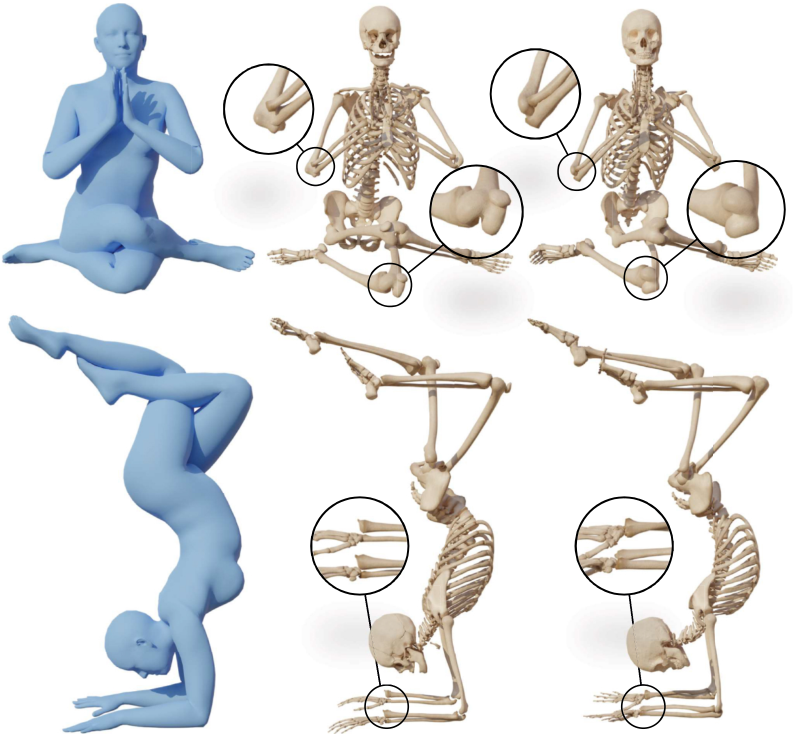


Figure 4.15: Qualitative comparison between the OSSO and SKEL skeletons fitted to MOYO SMPL meshes [119]. From left to right: Input SMPL mesh, OSSO skeleton, SKEL skeleton. First row: Due to the anatomic degrees of freedom of SKEL, the humerus and femur orientation are properly recovered, while OSSO fails. Second row: OSSO does not model the forearm supination: the radius is not properly rotated with respect to the ulna. The forearm bones have an anatomically correct orientation inside SKEL.

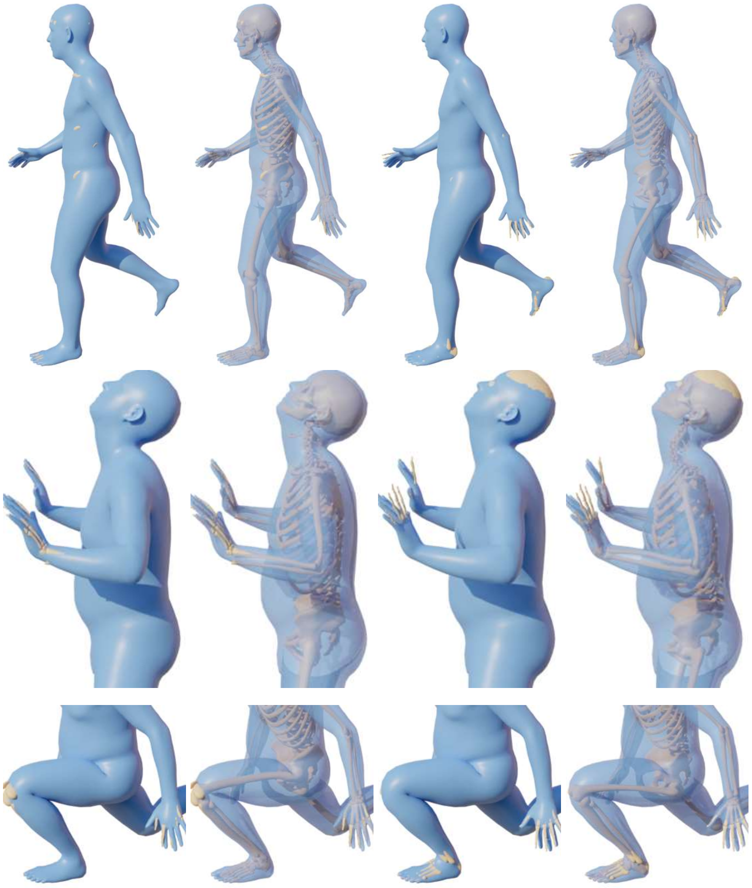


Figure 4.16: Given SMPL meshes from the Total Capture dataset [134] (in blue) we obtain OSSO [123] (left two columns) and the aligned SKEL skeleton (right two columns). SKEL provides a more anatomically correct skeleton, particularly at the joint level bone orientation, such as the knee and elbow.

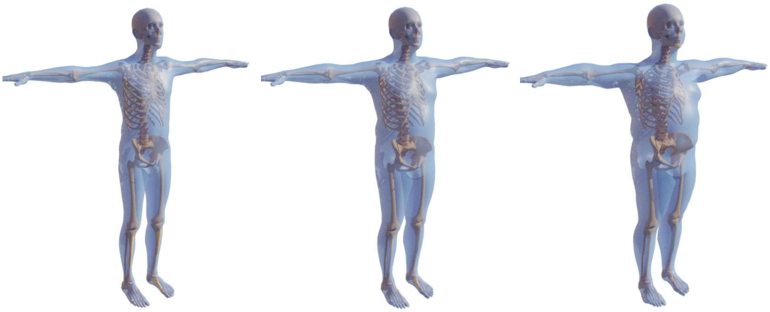


Figure 4.17: Given an input skeleton, and a target weight, SKEL can generate plausible skins while preserving the bone lengths. From left to right, we set the weight to be 70, 100, and 130 kg.

ered as *actual ground truth*, but rather as pseudo-ground truth. Obtaining actual ground-truth bone measurements of people in motion is not technically feasible. Thus, we rely on marker-based motion capture to obtain estimates of bone motion; this is the current “gold standard” in biomechanics. We also inherit the accuracy limits of this method, especially for the humerus head prediction, as shown in Fig. 4.11. A key next step is to use SKEL in diagnosing disease and injury and to compare this with traditional motion capture methods. This is necessary to validate the clinical relevance of the model and methods. It is worth noting that the learning and rigging pipeline described in Sec. 4.5 are, in fact, independent of the biomechanical model. If a new biomechanical model is clinically validated, one can rerun our approach with it to obtain an improved dataset and model.

Conclusion. In summary, SKEL effectively connects data-driven parametric body shape models with biomechanical skeletons to enable the integration of these technologies and fields, paving the way towards a new generation of body models and methods that combine the best of both worlds.

Chapter 5

Inferring the soft tissues inside the body

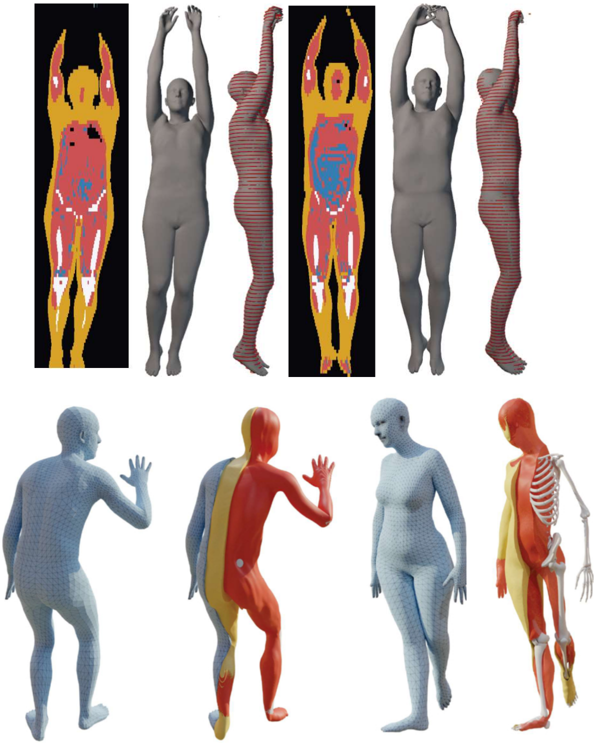


Figure 5.1: Top half: From volumetric human MRI scans, we learn to segment human internal tissues: subcutaneous adipose tissue (yellow), intra-muscular and visceral adipose tissue (blue), lean tissue (red), and long bones (white). We segment the MRI to extract a point cloud of the human body surface (red rings) to which we fit a human body model (SMPL, gray mesh). From this internal and external paired data, we learn Human Implicit Tissues (HIT), an implicit volumetric model that predicts the type and location of internal tissue. Bottom half: input body (blue mesh) and predicted tissues: subcutaneous adipose tissue (yellow) and lean tissue (red). We use OSSO [123] to infer the bones.

In Chapters 3 and 4, we infer the skeleton's geometry, along with bone locations and orientations, from the body shape. In this chapter, we extend this approach to infer the 3D locations of two additional anatomical tissues: subcutaneous adipose tissue (fat) and lean tissue (muscles and organs). Additionally, we predict again the location of the long bones; however, unlike in the previous chapters, we now utilize ground truth 3D bone volumes.

5.1 Introduction

Creating personalized anatomical digital twins of humans is key in fields such as medicine, sports science, biomechanics, and computer graphics.

In recent years, researchers have shown that the shape of the human body surface is related to the internal body composition [135–138], leading the way towards fast and non-invasive methods for early screening of body-composition-related pathologies.

In addition, we showed in Chapter 3 and 4 that predicting internal anatomical structures from the outer surface is also possible, paving the way towards the automatic creation of digital twins solely from body surface observations.

In this chapter, we extend the work presented in the previous chapters by introducing the prediction of several tissues inside the body. We focus on three key body tissues: subcutaneous adipose tissue (SAT), i.e. fat under the skin; lean tissue (LT), i.e. muscles and organs, and long bones, i.e. femur, tibia, fibula, humerus, ulna, radius, and hips. From a medical perspective, monitoring these tissues is essential: an excess of fat with respect to lean tissue is correlated with health risks such as the development of type-II diabetes and cardiovascular disease [6, 7]. From a biomechanics perspective, these tissues have different physical properties and dynamic behaviors, i.e. lean tissue is stiffer, adipose tissue is more elastic, whereas bones are rigid. These differences affect, for example, marker-based motion capture (mocap) systems [129], as markers on soft tissue exhibit artifacts [139]. Thus, having a good estimate of the tissue distribution could improve mocap accuracy and enable the simulation of soft-tissue compression in the apparel industry. In computer graphics, several methods assume [140, 141] or optimize [142] a *soft tissue layer* attached to a rigid structure to simulate physical interactions of the avatars in a virtual world. Also, artificial muscle systems [80] are widely used in character animation but these are complex to design by hand. Having a good estimate of the tissue distribution could improve the anatomic realism of these computational models.

To the best of our knowledge, the precise 3D prediction of these layers inside the body, *given only the outer body surface*, is a novel problem that has not been tackled in the previous literature. Specifically, our goal is to provide a prediction of the internal structures within a body model like SMPL for arbitrary body shapes and to be able to repose the predicted tissues. See Fig. 5.1 for a visualization of the resulting 3D representation.

Three main challenges must be overcome to learn a model that predicts soft tissues in the body from its surface. First, like in the two previous chapters, one needs paired observations of the inside and the outside of the body. Again, while medical scanners can capture the raw data, datasets are scarce and usually need to be annotated (segmented). Another challenge arising when working with soft tissues is that scanners that can see inside a body, such as MRI, require the subject to be lying down. This position introduces significant shape deformations due to the displacement of the soft tissues through contact with the scanning table. The last challenge is to design a neural network that can be effectively trained to extract the relevant information from the body's surface to infer the inner tissues. Our approach, *Human Implicit Tissues (HIT)*, addresses these challenges.

To obtain paired *inside* and *outside* data, we acquired a dataset of full-body MRI scans (260 female and 182 male). We start with a small subset (40 female and 40 male) for which we compute initial segmentations of lean and adipose tissues [143]. We curate them and enrich them with manual segmentations of the long bones and then train a nnU-Net [24] to segment all tissues in the entire dataset. These segmented volumes provide the distributions of the tissues inside the body.

To represent the outer body surface, we use the SMPL body model [43], which lets us model the dependency of the tissue locations inside the body on the pose and shape of a subject. However, unlike the surface of the body, an explicit mesh is not appropriate to represent the inner tissues since their topology can significantly vary between subjects. Implicit functions are particularly well suited to model the occupancy in a given volume [144] and recent work has extensively explored their use in modeling the body surface, clothed bodies and clothing itself, but not for modeling internal body structures. In our approach, given a point inside a body, we predict its tissue class; that is, we formulate the problem as a multi-tissue classification problem. Inspired by recent work on modeling clothed humans and neural rendering [56, 57], we combine implicit and explicit models and learn to map a 3D point inside a SMPL body into a canonical space. This allows us to learn the multi-tissue classification function in the canonical space. The decomposition of the problem into canonicalization and tissue classification offers the advantage of allowing generalization to unseen poses and body shapes. Yet, one more problem remains. Since full-body MRI scans are performed in a prone pose, the bodies exhibit significant deformation, which SMPL can not model, as SMPL was learned from upright scans of people. We capture these deformations by optimizing the SMPL mesh vertices to fit the body surface extracted from the MRI tightly. These tight fits allow us to quantify the geometric changes between the SMPL model mesh vertices and their deformed version. Our neural network can thus learn the 3D volumetric displacement of internal soft tissue caused by lying down. This allows us to uncompress the surface and internal structures from lying down to an upright position.

In summary, HIT provides a novel representation of the human body that connects the outer surface to the inner structure. It employs a hybrid of implicit and explicit shape representation and effectively extends the SMPL body model to infer internal structures that can be reshaped and reposed. The key contributions of HIT are: (a) we formulate the new problem of estimating the 3D structure of human internal tissues from surface observations as a multi-tissue classification problem; (b) we contribute a new dataset containing the volumetric tissue location inside the body extracted from real MRI scans, as well as the corresponding SMPL meshes representing the body surface; (c) we learn a volumetric deformation field accounting for the compression between a standing body shape and its counterpart lying prone on an MRI table; (d) we propose a neural implicit formulation to represent the tissue locations inside the body and show that this generalizes to new subjects and new poses; (e) we evaluate the proposed model on the created dataset. The HIT dataset and learned models are available for academic research at https://hit.is.tue.mpg.de.

5.2 Related work

Motivated by prior work on the prediction of body composition from 3D scans [110, 135, 137], silhouettes [136], or images [138], we go further to predict the location of subcutaneous adipose tissue, lean tissue, and the long bones, solely from the external body surface. In this section, we recall the relevant related work for this chapter.

Anatomic models. Early models [145, 146] use the Visible Human data [147], consisting of high-quality images from a cadaver, to build an anatomic model that can be animated. Other works address the creation of detailed personalized anatomic models from data of the hand [148] or the combination of multiple scans of the body [149] into one full-body avatar. Many other personalized anatomic human models have been created, with a focus on physical simulation of the tissues [76, 80–82, 106], pedagogic purposes [79, 108], or biomechanics [150]. The recent statistical model BOSS [122] includes the skeleton and several organs, but, unlike HIT, it does not model lean and adipose tissues.

Several methods create avatars with **soft tissue** deformation, enabling physics simulation. These typically model the soft tissue as a continuum layer coupled to an articulated skeleton [80, 140–142, 151]. This layer can be manually defined [80], estimated [151], obtained with an actual scan [140, 142], inferred from skin motion observations [141], or estimated using contact sensors [152]. None of these are validated against clinical data. Some other works have also addressed the modeling of deformation of the hands [153] and feet [47, 154] due to contact with the world.

Anatomy inference. Most internal anatomical structures cannot be inferred from skin observations alone, but some can, such as estimating the skull or jaw from the face shape [106, 107]. Anatomy Transfer [79] deforms an anatomical template model to be consistent with a new body surface. Similarly, Bauer et al. [87] leverage [79] to infer the skeleton inside a body from an RGBD image. Guo et al. [41] estimate the deformation of organs as a patient moves, but the organs' initial shapes are obtained by a scan of the patient. Anatomy Completor [155] can complete the shape of missing organs from the shape of the neighboring ones and OSSO (Chapter 3) can infer the skeletal bones from the skin surface. Only the last three works [41, 123, 155] evaluate on clinical data.

Recently, SKEL (Chapter 4) goes further and parameterizes the SMPL body model with a biomechanical skeleton that can be inferred from the body surface.

Datasets. Training and evaluation data, i.e. segmented full-body volumetric images, are key for solving this problem. While databases with medical scans [89, 156] exist, their per-pixel automatic segmentation into tissues is not straightforward. To create our paired dataset, we use an MRI protocol [157] and an automatic method [143] to obtain initial segmentations that we manually curate and enrich.

Human implicit shape models. Implicit shape representations have a long history and have recently become more popular due to the use of neural networks to learn occupancy or signed distance fields. Here, we focus on methods that model deformable volumes like the body surface [45, 48–55, 158], clothed bodies [49, 56–63], and clothing [64, 65]. Implicit shape representations enable efficient inside/outside tests, allowing the models to take into account the surrounding scene [48, 65], as well as supporting arbitrary topologies. Implicit functions for representing human bodies mainly use three approaches to encode the input query point: part-based, relative, and global. *Part-based* approaches [48, 52, 54] learn the occupancy in each part's canonical space whereas *relative encoding* approaches encode a point's occupancy with respect to joint locations [159], sparse skin vertices [64], or detected keypoints [160]. HIT uses a *global* approach, which learns occupancy in a canonical pose (namely a "star" pose). SCANimate [49], Meta Avatar [58] and ARAH [50] learn a subject-specific avatar in a canonical space and train a neural network to predict the skinning weights of any point in space. This learned inverse LBS (Linear Blend Skinning) lets them transform a point to the canonical pose space before querying the occupancy. In gDNA [56], a multi-subject occupancy model is learned in a canonical pose. A root-finding algorithm [161] enables unposing points, and a displacement field that maps shaped points to the canonical space is learned. We leverage a pretrained SMPL occupancy network to generalize to new shapes and poses, and add a new module to model the body compression and pose-dependent deformations.

A few works jointly model two surfaces, e.g. hand-object interaction [162] or multiple clothes [65], by adding interpenetration losses. Our multi-tissue classification formulation naturally avoids reasoning about interpenetration.

5.3 Human Tissue Data

A crucial requirement for learning the relationship between the body's inner tissues and the body surface is a structured dataset of paired observations. We use MRI scans of human subjects that we segment into several tissues. Each pixel of the MRI volume is classified as Bone Tissue (BT), Lean Tissue (LT), Intra-Muscular and Visceral Adipose Tissue (IMVAT), Subcutaneous Adipose Tissue (SAT), or Empty (E) (see Fig. 5.2 with segmentation examples). From the segmented volume, we extract the subject's body

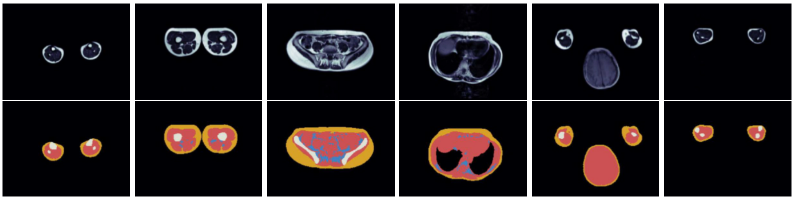


Figure 5.2: First row: input MRI images. Second row: segmentation results from the nnU-Net. Tissues color-code: bone (white), lean (red), subcutaneous adipose (yellow), intra-muscular and visceral adipose (blue), empty (black). From left to right: calf, thighs, hips, chest, head and arms, forearms.

surface as a point cloud and fit the SMPL [43] body model to it. We also compute tight fits of the SMPL body mesh that capture the flattened body shape extracted from the MRI (see Fig. 5.6). In this way, we create a dataset of paired observations of the inner body tissues together with the human body surface.

5.3.1 MRI segmentation

MRI scans dataset. We work with 442 scans (260 female, 182 male) acquired with a 1.5 T scanner (Magnetom Sonata, Siemens Healthcare) following a standardized protocol for whole body adipose tissue topography mapping [157]. All subjects gave prior informed written consent and the study was approved by the local ethics board. Each scan has around 110 slices, slightly varying depending on the subject's height. The slice resolution is 256×192, with an approximate voxel size of $2 \times 2 \times 10$ mm.

Tissue definitions. Given an input MRI image (slice), our goal in this section is to classify the tissue type of each pixel. For the Bone Tissue (BT) we focus on the long bones: femur, tibia, fibula, humerus, ulna, radius, and hips. We do not segment smaller bones, such as vertebrae, ribs, or phalanges, as, with the limited resolution of the MRI images, it is difficult to identify them in the images consistently. The muscles and organs are segmented as Lean Tissue (LT). The Subcutaneous Adipose Tissue (SAT) and the Intra-Muscular and Visceral Adipose Tissue (IMVAT) denote the human fat; SAT is located directly under the skin, whereas IMVAT is located inside the muscles and around the organs. MRI pixels where no tissue is detected are classified as Empty (E). Empty areas include the background outside the body, the lungs, skull cortical bone, and other cavities inside the body.

Segmentation Process. To segment the whole MRI dataset into tissues, we use a *human-in-the-loop* approach similar to SAM [163]. We leverage initial automatic segmentations [143] and manual annotations to train and refine a nnU-Net [24] with the help of human supervision.

nnU-Net [24] is an Auto-ML framework that automatically configures a U-Net and adapts the training procedure to the input data. For segmentation tasks in the medical domain, it has been shown to be state-of-the-art in many benchmarks and it has been used by others in the creation of datasets [164, 165]. We use it to train two networks, $\mathbf{W_{bones}}$ and $\mathbf{W_{all}}$, which we present next. For each model, we use the default settings configured by the Auto-ML framework.

To segment our dataset, we start by manually annotating the long bones (femur, tibia, fibula, humerus, ulna, radius and hips) in a small subset of the dataset (1105 slice images from 10 subjects). Then we train a segmentation model $\mathbf{W_{bones}}$, to segment the bones in the MRI images. Fig. 5.3 presents examples of predicted bone annotations. The input to the network is a single-channel DICOM MRI that contains normalized MRI intensities and the output is a pixel-wise labeled mask with labels $\mathbf{L_{bones}} \in [0, 1]$. We empirically validate that 1K images were enough to obtain a good generalization to left-out subjects (DICE score: mean 0.91/median 0.95).

In parallel, we use an automatic approach [143] that segments MRI images into adipose tissue (AT), empty (E), lean tissue (LT), as well as Visceral Adipose Tissue (VAT), i.e. fat around organs only in the abdominal region (see Fig. 5.4). The segmentations from this method are generally good, but the method uses empirically defined constants that do not generalize well across subjects. Most failures come from the sequential approach of the automatic method, first detecting anatomic landmarks and then segmenting the tissues. A landmark detection error often leads to some missing parts in the segmentation. Typical errors at this stage are shown in Fig. 5.5.

Screening the full dataset ($\sim 442 \times 110$ slices) is impractical, so we focus on a gender-balanced subset (80 subjects, ~ 8900 images) and curate the generated segmentation artifacts. For the gender-balanced subset, we also infer the bone masks with $\mathbf{W_{bones}}$ and merge them with the curated segmentations to obtain one multi-tissue mask per image. We then post-process the merged segmentation masks in order to remove small artifacts and split erroneous adipose tissue (AT) segmentations from [143] into subcutaneous adipose tissue (SAT) and intra-muscular and visceral adipose tissue (IMVAT).

The obtained merged and post-processed segmentations were visually inspected and failure cases were corrected (~ 150 images from the total of ~ 8900). Then, with the curated data, we train a new nnU-Net model, $\mathbf{W_{all}}$, that takes an MRI image as input and predicts a label for each pixel corresponding to one of the 5 tissue types (BT, LT, SAT, IMVAT, E). This network effectively replaces the previous network $\mathbf{W_{bones}}$ and the automatic method [143] which are not used anymore. We quantitatively evaluate the new segmentation predictions from $\mathbf{W_{all}}$ on a held-out test set, obtaining a mean/median DICE score of $0.92/0.98$;

To obtain the final segmentations, we use $\mathbf{W_{all}}$ to infer the segmentation masks for all 442 subjects. Fig. 5.12 and Fig. 5.13 show examples of final segmentation masks for a single female and male subject respectively. One final visual inspection of the

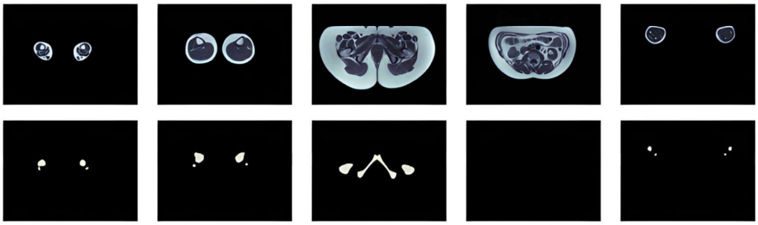


Figure 5.3: Examples of bone predictions from the model trained on manually annotated long bones.

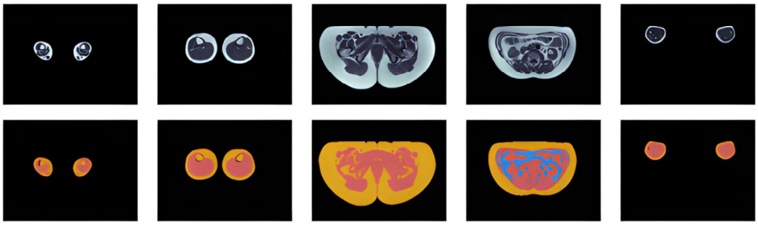


Figure 5.4: Examples of tissue predictions from the Würslin et al. [143] method.

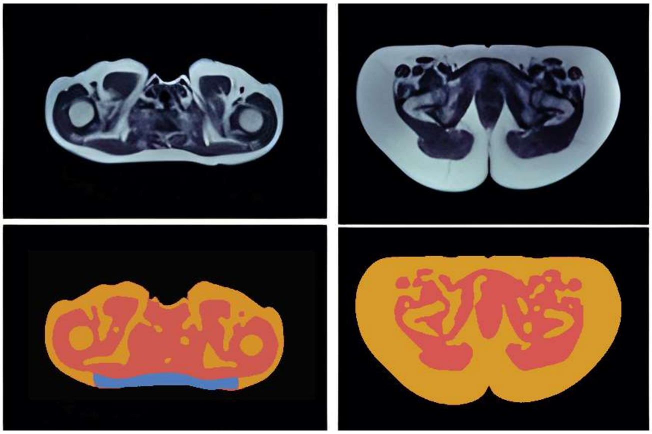


Figure 5.5: Typical errors from the Würslin [143] method. Left: subcutaneous adipose tissue (SAT) here is inaccurately labeled as visceral adipose tissue (VAT) in blue. Right: visceral adipose tissue (VAT) is labeled as subcutaneous adipose tissue (SAT) in yellow.

obtained segmentations was performed to validate the full dataset of segmentations. In the remainder of this chapter, these segmentations are treated as the ground truth internal tissues.

5.3.2 Two-step SMPL fit

The subjects of our dataset are lying down during the MRI scan, which causes the body shape to flatten. This skin compression is highly subject-specific, depending on their body composition, and SMPL is not able to model it. To obtain SMPL fits that faithfully capture the shape of the subject as well as the compression, we use a two-step process. Fig. 5.6 illustrates the obtained fits, and below we detail further how to obtain these fits.

Initial SMPL fit. First, for each subject i, we extract the outer body contour from the segmented MRI images and, using the metric units of the volumetric MRI, we create a 3D point-cloud that we denote ST_i for *skin* (see Fig. 5.8 left). Then, we compute a first approximation of the subject's shape and pose parameters (β_i^1, θ_i^1) that minimize the distance between the point cloud and body surface. As the subject's MRI poses are similar, we define a reference pose, θ_{MRI}, which we use to initialize the fitting of all subjects and we regularize the estimated pose so that it does not differ too much from the reference pose. We use the female or male version of SMPL according to the subject's sex and we denote these fitted SMPL meshes $\mathbf{S}_i^1$ (see Fig. 5.8 middle).

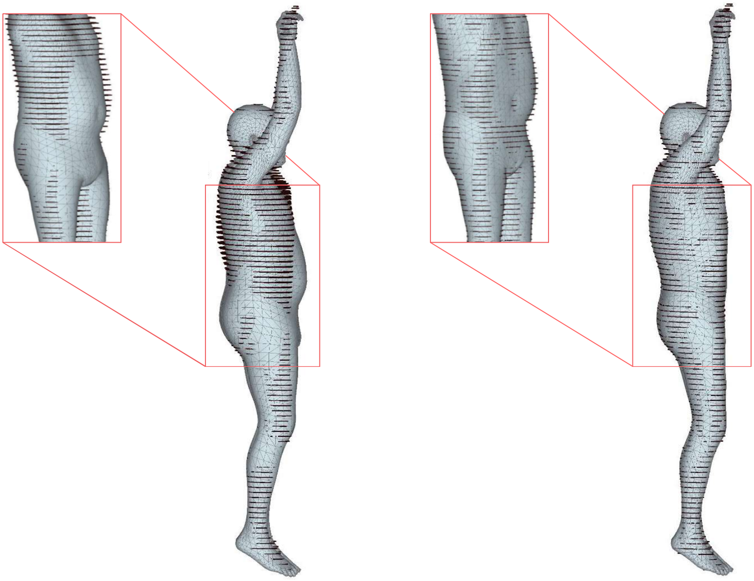


Figure 5.6: SMPL fits to MRI point clouds. The SMPL fit $\mathbf{S}_i$ (left) does not capture the flattened shape, whereas the tight fit $\mathbf{f}_i$ (right) does.

Next, we optimize the SMPL mesh vertices to deform and match the point-cloud ST_i. Inspired by the literature in the context of clothing capture [166, 167] we compute meshes in SMPL parametrization that tightly fit the segmented skin point clouds, ST_i. We optimize the new vertex locations, bound with Laplacian regularisation [167], and denote the resulting *free form meshes* $\mathbf{f}_i^1$. In Fig. 5.8 we show an example of an input MRI point-cloud ST_i and the obtained meshes $\mathbf{S}_i^1$ and $\mathbf{f}_i^1$. It is worth noting, that while visually similar, the volumes of the meshes $\mathbf{S}_i^1$ and $\mathbf{f}_i^1$ can be very different. Fig. 5.7 shows the volume disagreement between $\mathbf{S}_i^1$ and $\mathbf{f}_i^1$ for the 442 subjects. It shows that the SMPL fit $\mathbf{S}_i^1$ tends to under estimate the body volume, yielding a shape vector β biased toward a lower body volume.

Constant volume SMPL fit. To get a shape vector for each subject that represent properly their body volume, we need the SMPL fit to have the same body volume as the subject. To ensure this, we start by computing the volume of the mesh $\mathbf{f}_i^1$, which we denote V_i^0. As this mesh is a tight fit to the point cloud, it yields a good estimation of the actual volume of the subject. Next, from the point-cloud ST_i, we now only consider the subset of points that are less affected by the table compression, i.e. those for which the normal vector is pointing in the same direction as the normal vector of the table, $\mathbf{n}_T$. Effectively, we weight the contribution of each point $\mathbf{s}_p \in ST_i$ with $w(\mathbf{s}_p, \mathbf{n}_p) = \sigma(\mathbf{n}_p \cdot \mathbf{n}_T)$ where σ is the sigmoid function. In Fig. 5.10 we show the effect of this weight on a SMPL mesh. Then, we compute a new SMPL mesh $\mathbf{S}_i$ and its parameters (β_i, θ_i) that match the weighted vertices with the additional volume constraint $||V_i^0 - V_i||_{L2}$, where V_i is the volume of $\mathbf{S}_i$. This enforces $\mathbf{S}_i$ to have a consistent volume with the MRI observation. In Fig. 5.9 we show the difference between the computed parameters β_i^1 and β_i, showing that the volume-preserving constraint effectively affects the computed body shape. The last step is to compute a deformed mesh $\mathbf{f}_i$ that is consistent with the new SMPL mesh $\mathbf{S}_i$. To this end, we recompute the free-vertex optimization starting from $\mathbf{S}_i$ to obtain a new tight fit $\mathbf{f}_i$. In Fig. 5.11 we show further results of the obtained SMPL meshes. These meshes allow us to compute the compression displacements $\mathbf{d}_{\text{comp}}$ between the $\mathbf{S}_i$ and $\mathbf{f}_i$ vertices. An animated illustration of this displacement can be seen in the video on the HIT project page[1].

As SMPL can not model the stomach compression observed in the dataset, this two-step approach is crucial to get SMPL β values for each subject that actually match their body volume.

5.3.3 Human Implicit Tissues (HIT) dataset

The segmented MRI, along with the SMPL fits, constitute the Human Implicit Tissues (HIT) dataset. This new dataset contains, for each subject i: a) the volumetric image segmented into BT, LT, SAT, IMVAT, and E, b) the image MRI center and per-pixel spacing

[1] https://hit.is.tue.mpg.de/

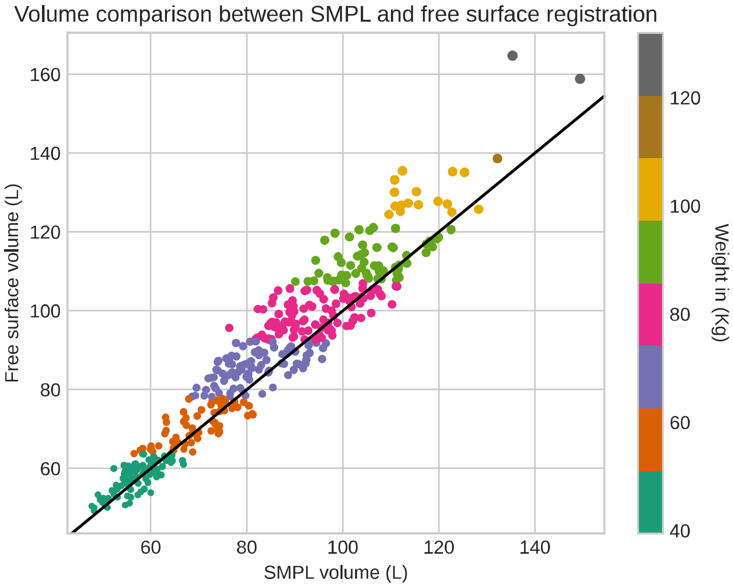


Figure 5.7: Volume differences between the naive SMPL body fit and the free-vertices version.

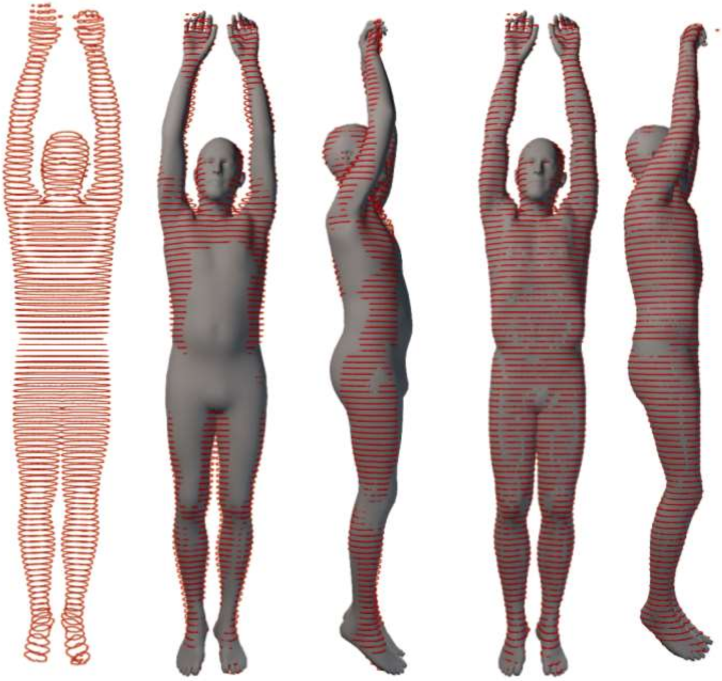


Figure 5.8: Initial fit to the MRI skin point-cloud. Left: point-cloud ST_i extracted from the MRI. Middle: SMPL model fit $\mathbf{S}_i^1$. Right: Free-vertex fit $\mathbf{f}_i^1$.

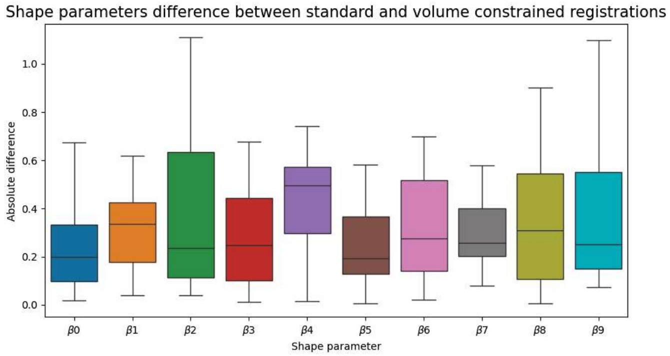


Figure 5.9: Boxplot of the shape coefficients difference between the $\mathbf{S}_i^1$ and the volume preserving $\mathbf{S}_i$ for the 442 subjects.

to transform indices from the volumetric image into 3D metric locations, and c) the skin point cloud ST_i, the fitted SMPL model mesh $\mathbf{S}_i$ represented by its parameters (θ_i, β_i), as well as the SMPL tight fit $\mathbf{f}_i$. From b) we compute the compression displacements $\mathbf{d}_{\text{comp}} \in \mathbb{R}^{N_v}$, between the $\mathbf{S}_i$ and $\mathbf{f}_i$ vertices, where $N_v = 6890$ are the number of SMPL vertices. Note that the original MRI images are not included. This dataset is available for academic research at https://hit.is.tue.mpg.de/.

Final three-layer representation. The Intra-Muscular and Visceral Adipose Tissue (IMVAT) segmented in the MRI images is sparsely located around the muscles and abdominal organs (blue in Figs. 5.1 and 5.2). As its precise 3D location highly varies among people, we leave the precise localization of IMVAT for future work and infer three layers of tissue: Bone Tissue (BT), Lean Tissue and Intra-Muscular and Visceral Adipose Tissue (LT + IMVAT) and Subcutaneous Adipose Tissue (SAT). That is, in the remainder of this chapter, we merge IMVAT with the surrounding LT structures and refer to them together as LT. In Fig. 5.12 and 5.13, we show examples of segmented slices from our HIT dataset.

5.4 HIT method

Problem statement. We formalize the inference of the tissues inside the body as a 4-tissue classification problem (BT, LT (+ IMVAT), SAT, E). HIT learns an implicit function that takes as input SMPL shape and pose parameters (β, θ) and a 3D point $\mathbf{x}$, and outputs the tissue class at that point.

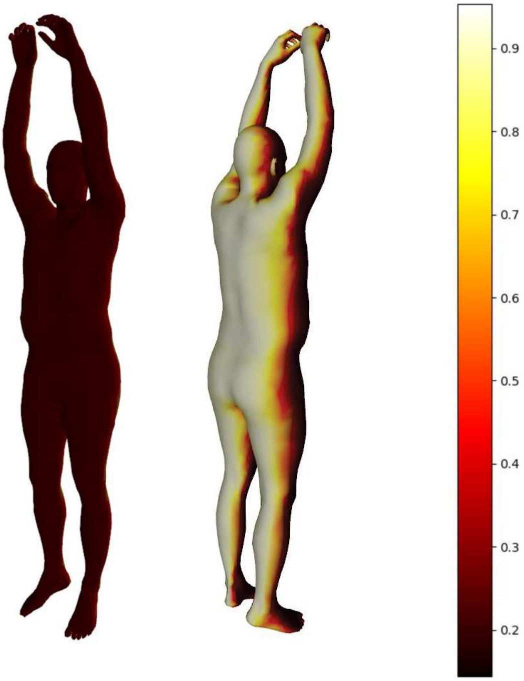


Figure 5.10: SMPL mesh vertices color-coded with the computed weights $w(\mathbf{s}_p, \mathbf{n}_p)$. Vertices affected by the compression have a low weight, whereas vertices far from the MRI table have a high weight. These are used to compute $\mathbf{S}_i$.

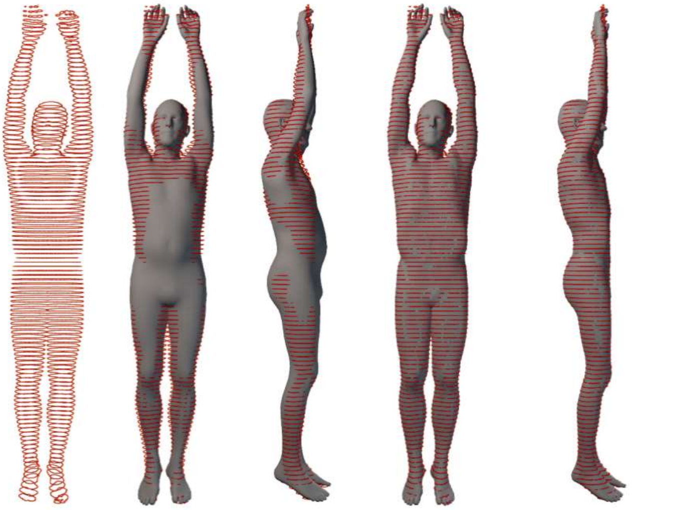


Figure 5.11: Examples of the obtained fit results. Left: point-cloud ST_i extracted from the MRI. Middle: Volume preserving SMPL model fit $\mathbf{S}_i$. Right: Free-vertex fit $\mathbf{f}_i$.

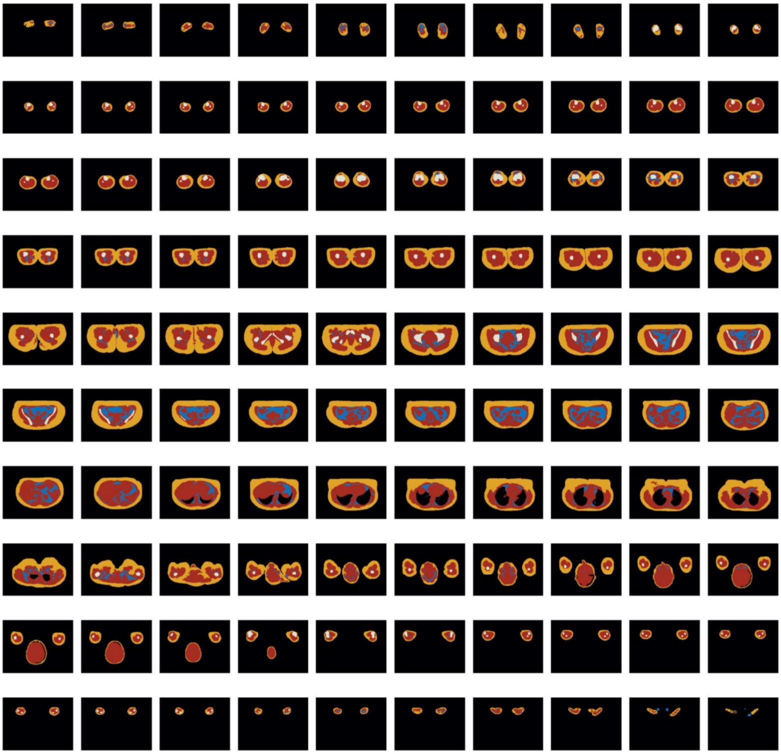


Figure 5.12: Example segmentation masks of a female subject.

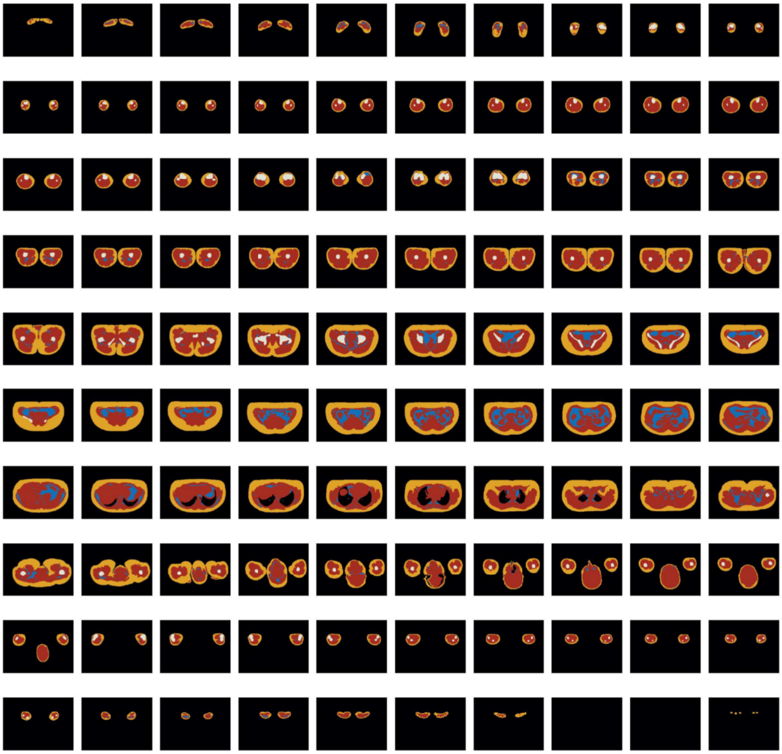


Figure 5.13: Example segmentation masks of a male subject.

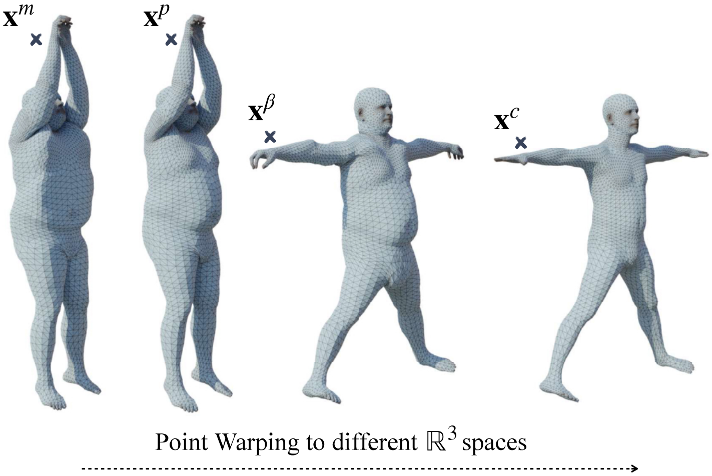


Figure 5.14: HIT defines four $\mathbb{R}^3$ spaces. A point $\mathbf{x}^m$ in the original MRI space corresponds to $\mathbf{x}^p$ in the posed space, $\mathbf{x}^\beta$ in the shaped space, and $\mathbf{x}^c$ in the canonical space.

5.4.1 HIT spaces

To learn the tissue occupancy, HIT warps the data from the input MRI space into a canonical space. These warps are defined between four spaces, illustrated in Fig. 5.14. The *canonical space* is where the SMPL template mesh, $\mathbf{T}$, in a "star" pose, is defined. Points in the canonical space are indexed by $\mathbf{x}^c$. The *shaped space* is where additive offsets, $\mathbf{d}_\beta \in \mathbb{R}^{N_v \times 3}$, controlled by the shape β of the subject, are applied to the template. We denote points there as $\mathbf{x}^\beta = \mathbf{x}^c + \mathbf{d}_\beta$. The shaped points can then be posed through linear blend skinning into the *posed space* $\mathbf{x}^p = \text{LBS}(\mathbf{x}^\beta, \mathbf{w}, \theta)$, where $\mathbf{w} \in \mathbb{R}^{N_p}$ is a vector of blend-weights, $N_p = 24$ is the number of SMPL body parts, and θ represents the pose parameters. Finally, to model the MRI table compression on the body, we define volumetric offsets $\mathbf{d}^x_{\text{comp}} \in \mathbb{R}^3$ and denote points in the *original MRI space* $\mathbf{x}^m = \mathbf{x}^p + \mathbf{d}^x_{\text{comp}}$. Note that points in all spaces live in $\mathbb{R}^3$ and can be inside, outside, or on the SMPL surface.

5.4.2 HIT architecture

Our architecture is composed of 4 building blocks. Three modules enable warping an MRI point into the canonical space $\mathbf{x}^c = (\mathcal{S} \circ \mathcal{U} \circ \mathcal{D})(\mathbf{x}^m, \beta, \theta)$ by Decompressing ($\mathcal{D}$), Unposing ($\mathcal{U}$) and Deshaping ($\mathcal{S}$). The warping architectures are illustrated in Fig. 5.15.

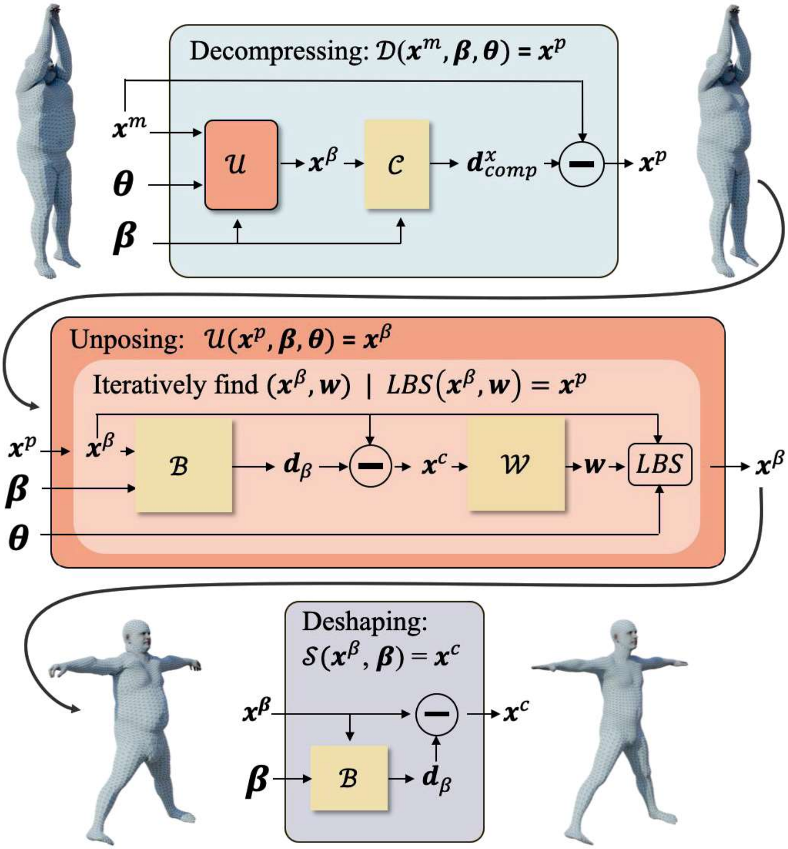

Decompressing: $\mathcal{D}(\boldsymbol{x}^m, \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{x}^p$
$\boldsymbol{x}^m$
$\boldsymbol{\theta}$
$\boldsymbol{\beta}$
$\mathcal{U}$
$\boldsymbol{x}^\beta$
$\mathcal{C}$
$\boldsymbol{d}^x_{comp}$
$\boldsymbol{x}^p$
Unposing: $\mathcal{U}(\boldsymbol{x}^p, \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{x}^\beta$
Iteratively find $(\boldsymbol{x}^\beta, \boldsymbol{w}) \mid LBS(\boldsymbol{x}^\beta, \boldsymbol{w}) = \boldsymbol{x}^p$
$\boldsymbol{x}^p$
$\boldsymbol{x}^\beta$
$\boldsymbol{\beta}$
$\boldsymbol{\theta}$
$\mathcal{B}$
$\boldsymbol{d}_\beta$
$\boldsymbol{x}^c$
$\mathcal{W}$
$\boldsymbol{w}$
LBS
$\boldsymbol{x}^\beta$
Deshaping:
$\mathcal{S}(\boldsymbol{x}^\beta, \boldsymbol{\beta}) = \boldsymbol{x}^c$
$\boldsymbol{x}^\beta$
$\boldsymbol{\beta}$
$\mathcal{B}$
$\boldsymbol{d}_\beta$
$\boldsymbol{x}^c$

Figure 5.15: HIT modules ($\mathcal{D}$,$\mathcal{U}$,$\mathcal{S}$) and networks ($\mathcal{C}$,$\mathcal{B}$,$\mathcal{W}$) to warp points between spaces.

Once the warped point is in the canonical space, the network $\mathcal{T}(\mathbf{x}^c)$ predicts its tissue class.

Deshaping module. Given a shape β, the Deshaping module $\mathcal{S}$ transforms shaped points into canonical points, i.e. $\mathcal{S}(\mathbf{x}^\beta, \beta) = \mathbf{x}^c$. In the module, the function $\mathcal{B}$ predicts the offsets $\mathcal{B}(\mathbf{x}^\beta, \beta) = \mathbf{d}_\beta$, which are subtracted from $\mathbf{x}^\beta$ to obtain $\mathbf{x}^c$ (see Fig. 5.15 bottom diagram).

Unposing module. Given shape and pose parameters, the Unposing module $\mathcal{U}$ warps points from the *posed space* into the *shaped space*: $\mathcal{U}(\mathbf{x}^p, \beta, \theta) = \mathbf{x}^\beta$. Similar to Chen et al. [56], this module uses two MLPs: $\mathcal{B}(\mathbf{x}^\beta, \beta) = \mathbf{d}_\beta$ defined in the previous paragraph and the function $\mathcal{W}(\mathbf{x}^c) = \mathbf{w}$ which predicts the skinning weights $\mathbf{w}$ of a point in the canonical space (see Fig. 5.15 middle diagram). Unposing points inside or outside a posed SMPL mesh is challenging because the skinning weights are only defined on the SMPL surface. Chen et al. use a root-finding algorithm [168] that finds candidate points $\{\mathbf{x}_i^\beta\}$ for a given posed point $\mathbf{x}^p$. Then, their SMPL occupancy prediction is used to decide on the best candidate. However, this is not applicable in our multi-tissue case; i.e. if a point has two roots, one in LT and one in BT, there is no way to know which one is correct. To overcome this limitation, we initialize the root finding with skinning weights $\mathbf{w}$, fetched from the closest SMPL vertex. This initialization allows the iterative algorithm to converge to the skinning weights that properly unpose the point.

Decompression module. To model the body deformation displacement induced by the MRI table, we learn a Decompression module $\mathcal{D}$ that maps points in the MRI space to the posed space: $\mathcal{D}(\mathbf{x}^m, \beta, \theta) = \mathbf{x}^p$. We do so by learning to predict the volumetric compression displacements $\mathbf{d}^x_{\text{comp}}$, which generalize the computed $\mathbf{d}_{\text{comp}}$ on the SMPL surface. However, learning the volumetric body decompression is challenging: a 3D point $\mathbf{x}^m$ can represent a different anatomic region for two different subjects. Thus, instead of learning to predict displacements in the MRI space, we first unpose $\mathbf{x}^m$ into the *shaped space* $\mathbf{x}^\beta$ and predict $\mathcal{C}(\mathbf{x}^\beta, \beta) = \mathbf{d}^x_{\text{comp}}$ so that $\mathbf{x}^m + \mathbf{d}^x_{\text{comp}} = \mathbf{x}^p$ (see Fig. 5.15 top diagram). The *shaped space* has a natural shape consistency, which helps predict the compression.

Multi-Tissue network. Once points are in the *canonical space*, HIT uses an MLP to predict the point tissue class $\mathcal{T}(\mathbf{x}^c) = \{\text{E}, \text{LT}, \text{SAT}, \text{BT}\}$.

Appendix B.1 provides the implementation details of each of these networks.

5.4.3 Training, losses and sampling

Training. To train HIT, we proceed in 3 steps. First, $\mathcal{B}$ and $\mathcal{W}$ are pre-trained by randomly sampling shaped and posed SMPL bodies. In parallel, $\mathcal{C}$ is pre-trained using the computed $\mathbf{d}_{\text{comp}}$ (see Sec. 5.3). Then, the weights of $\mathcal{C}$ are frozen, and $\mathcal{B}$, $\mathcal{W}$, $\mathcal{T}$ and $\mathcal{M}$ are jointly trained on the HIT dataset. We denote the network's trainable weights as ψ_* and use the subscript $*$ to refer to the network name, i.e. $\psi_{\mathcal{B}}$ are the weights of the network $\mathcal{B}$. To train our architecture, we minimize the following losses.

Deshaping loss. Let $\mathbf{x}_v^\beta$ be a vertex of the shaped mesh $SMPL(\beta)$, and $\mathbf{x}_v^c$ be the corresponding vertex on the SMPL template mesh. To train the weights $\psi_{\mathcal{B}}$, we enforce the predicted displacement to match the SMPL's β offset at the body surface level by minimizing

$$l_s(\psi_{\mathcal{B}}) = \text{MSE}_v(\mathcal{B}(\mathbf{x}_v^\beta, \beta) - (\mathbf{x}_v^c - \mathbf{x}_v^\beta)), \tag{5.1}$$

where MSE_v is the mean square error over sampled points.

Skinning weight loss. To train the $\psi_{\mathcal{W}}$ weights we enforce the predicted skinning weights to be consistent with the SMPL ones by minimizing

$$l_w(\psi_{\mathcal{W}}) = \text{MSE}_v(\mathcal{W}(\mathbf{x}_v^c) - \mathbf{w}_v), \tag{5.2}$$

where $\mathbf{w}_v \in \mathbb{R}^{N_p}$ denotes the SMPL's skinning weights of the SMPL template vertex $\mathbf{x}_v^c$.

Decompression loss. To train $\mathcal{C}$ to predict a displacement that is similar to the one between $\mathbf{x}_v^m$ and $\mathbf{x}_v^p$, we minimize

$$l_c(\psi_{\mathcal{C}}) = \text{MSE}_v(\mathcal{C}(\mathcal{U}(\mathbf{x}_v^m, \beta)), \mathbf{x}_v^p - \mathbf{x}_v^m). \tag{5.3}$$

Multi-tissue loss. Given a point sampled inside the compressed body $\mathbf{x}_k^m$ and canonicalized to $\mathbf{x}_k^c$, we train $\mathcal{T}$ to predict the correct tissue label. This is done by optimizing the weighted cross-entropy loss between the tissue predictions and the training data, where weights are inversely proportional to the tissue sample size.

Sampling strategy. The MRI scans have a discrete volumetric representation. To avoid aliasing artifacts, we train HIT with points $\mathbf{x}_k^m$ sampled at the center of voxels. Note that this only holds for querying the ground truth volume; once the implicit function is learned, one can sample arbitrary locations inside the body and extract smooth tissue volumes.

Uniformly sampling the MRI voxel centers means that the canonical space is not uniformly sampled, e.g. the space between the legs is wider in the canonical space. Thus, we also uniformly sample points outside the SMPL template mesh in the canonical space and classify them as E .

As the MRI resolution is low for hands, we force these parts to always be predicted as LT by uniformly sampling points in the canonical space inside the hands' bounding boxes.

	Female						Male					
	LT		SAT		BT		LT		SAT		BT	
	D.S. ↑	Δ% ↓	D.S. ↑	Δ% ↓	D.S. ↑	Δ% ↓	D.S. ↑	Δ% ↓	D.S. ↑	Δ% ↓	D.S. ↑	Δ% ↓
Chance	51.4	4.9	40.2	6.7	3.9	0.7	60.3	5.1	31.8	6.5	4.1	0.8
HIT	77.8	4.0	57.7	9.4	45.5	0.6	81.0	5.1	54.7	6.5	52.2	0.7

Table 5.1: Quantitative evaluation for females and males on the three tissues (LT, SAT, BT). DICE Score (D.S.) - higher is better and Δ % is the relative difference in the tissue percentage prediction (in percent) - lower is better.

Multi-Tissue mesh extraction. A classical approach to visualize the occupancy of an implicit shape is to extract the mesh at a given level set of the implicit surface. Modeling occupancy as a multi-class classification problem has the advantage that we can extract meshes for each class, such that there is no class overlap in the predictions. However, for a given class, class scores do not directly yield the continuous occupancy function that is required for level-set extraction.

Thus, given one tissue, our goal is to define a function to apply to the per-class probabilities with the following properties:

1. If class k has the highest probability, the function should yield an occupancy value > 0.5.

2. If class k does not have the highest probability, the function should yield an occupancy value < 0.5.

3. If class k has the highest probability, but is equal to another class, the function should yield an occupancy value $= 0.5$. This case defines the boundary between two or more tissues, which will be extracted by the level-set method.

For a point in the canonical space $\mathbf{x}^c$, let $\{p_i\}_{i\in[1,C]}$ with $p_i \in [0,1]$ be the probabilities of each of the $C = 4$ classes ($\sum_{i=0}^{C} p_i = 1$). For a tissue k, we define the function l_k as:

$$l_k(\{p_i\}) = \frac{p_k}{p_k + \underset{\forall j!=k}{max(p_j)}} \tag{5.4}$$

which fulfills the desired properties. We can then pass $l_k(\{p_i\})$ to a marching cube algorithm to extract the k-th tissue mesh surface, and get tissue volumes that match the predicted occupancy.

The volumes displayed in Fig. 5.1, 5.21, and 5.23, as well as the volumes in Fig. 5.18 and 5.19 were extracted using this technique.

5.5 Experiments

To evaluate HIT we split the data into 80% train, 10% validation and 10% test sets (females 201/25/26, males 136/17/16). Furthermore, we train separate models for males and females as the literature reports significant differences in their body composition [169].

5.5.1 Tissues prediction evaluation

Since we address a novel problem, to the best of our knowledge, no prior work can be directly used for comparison: e.g. OSSO [123] solely predicts the bone structures. To have a numeric calibration for the multi-tissue problem, we propose a *Chance* baseline, which, for each queried point, predicts the tissues [E, LT, SAT, BT] with probabilities 0.03, 0.52, 0.41, 0.04 and 0.04, 0.60, 0.32, 0.04 for females and males respectively. These values follow the average percentage of each tissue in the training set.

Metric evaluation. To quantitatively evaluate the HIT architecture, we report mean Dice scores [170] for each predicted tissue on the test set. Additionally, we compute the relative error of the predicted tissue percentage by computing $\Delta = \mid V_{pred} - V_{GT} \mid / V_B * 100$, where $V_{pred}, V_{GT}, V_B \in \mathbb{R}$ are the volumes of the predicted tissues, ground-truth tissues, and whole body, respectively.

Table 5.1 shows that the HIT Dice scores are significantly better than the Chance baseline.

Slices evaluation. Additionally, 5.16 and Fig. 5.17 present qualitative results of the tissue predictions on transverse planes and in 3D. We show that the predicted inner tissues are consistent across the body and exhibit plausible compression. In addition, most errors arise at the tissue interfaces.

In Fig. 5.16 right column lines 9 to 11, we see that empty tissue is predicted inside the thigh. We conjecture this happens due to the root finding algorithm initialization, which rigs a query point to the closest SMPL skin vertex. In the cases where the SMPL mesh fit to the MRI contour exhibits self-penetration at the thigh, which can happen when the legs are compressed together, the points inside the intersection get rigged to the wrong leg. As a result, the occupancy is queried outside the body, leading to an empty prediction.

Volume evaluation. The 3D visualization in 5.18 and 5.19 also shows that the obtained 3D meshes are visually consistent with the GT ones. Each tissue's mesh is extracted in the canonical pose given the subject's shape β, then posed to the target pose θ.

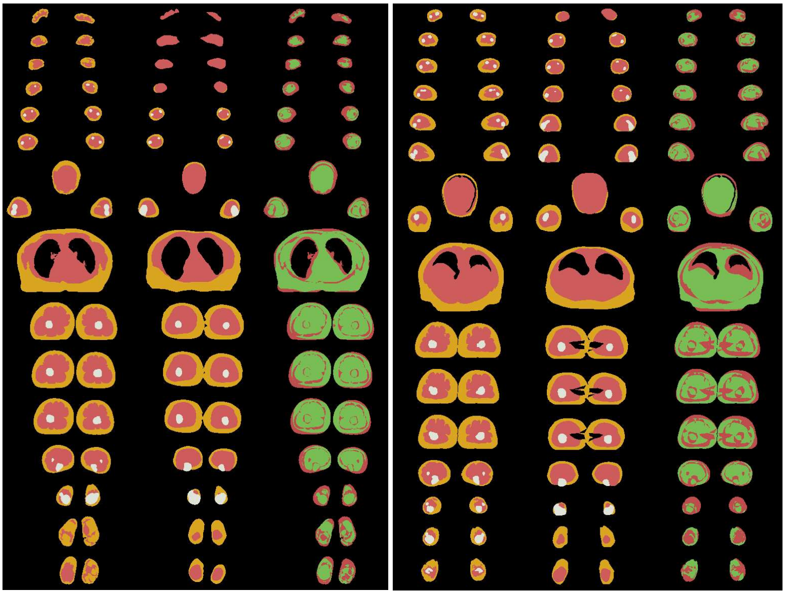


Figure 5.16: Transverse slices (female): (left) GT tissues, (middle) HIT predictions, (right) accuracy (green correct, red otherwise).

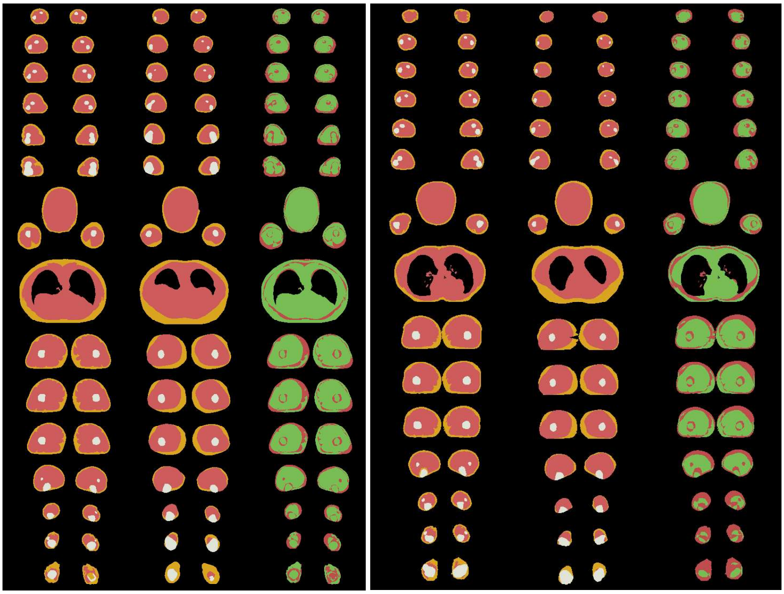


Figure 5.17: Transverse slices (male): (left) GT tissues, (middle) HIT predictions, (right) accuracy (green correct, red otherwise).

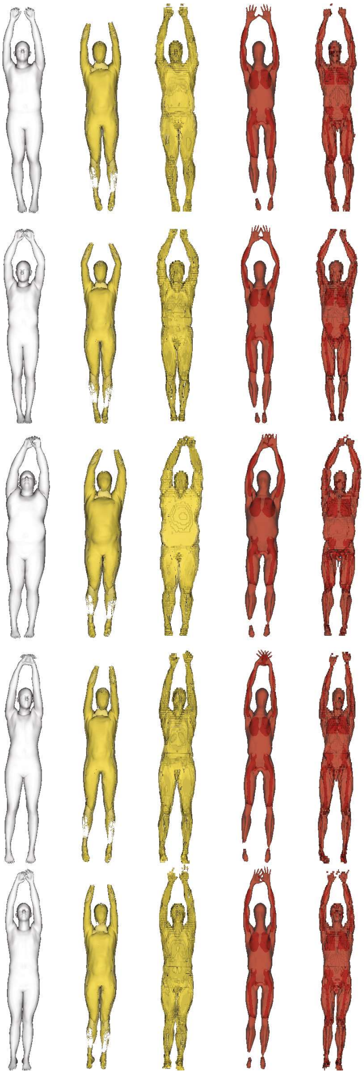


Figure 5.18: Volumetric tissue predictions for males. From left to right: SMPL fit $\mathbf{S}_i$ (gray), HIT LT prediction, GT LT, HIT SAT prediction, GT SAT.

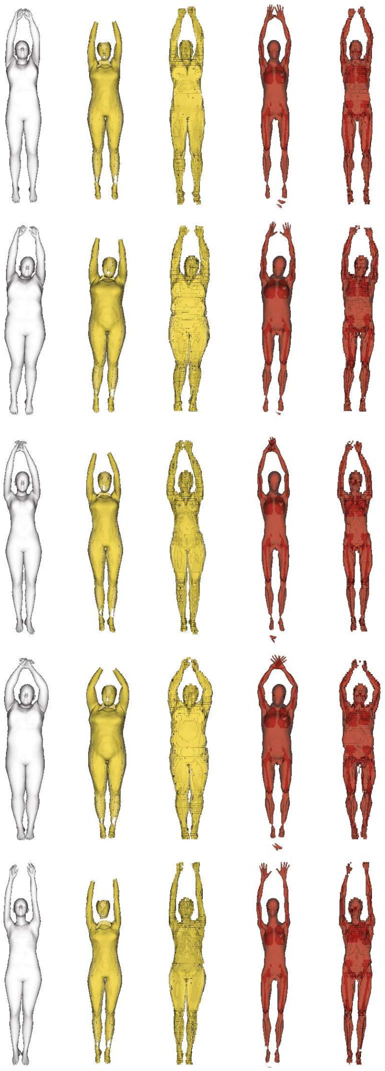


Figure 5.19: Volumetric tissue predictions for females. From left to right: SMPL fit $\mathbf{S}_i$ (gray), HIT LT prediction, GT LT, HIT SAT prediction, GT SAT.

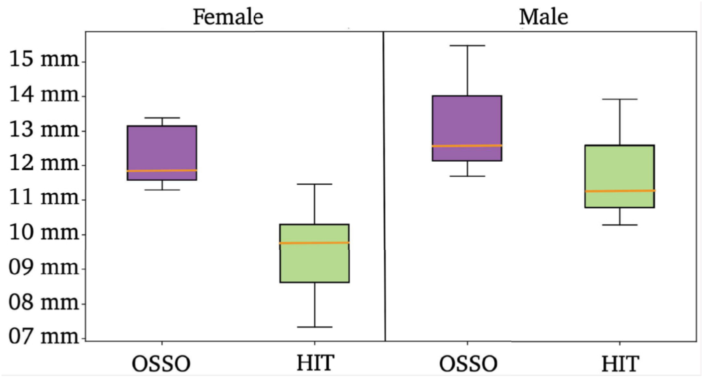


Figure 5.20: Comparison between OSSO and HIT bone predictions.

5.5.2 Comparison with OSSO

To further put HIT bone predictions in context, we numerically compare to OSSO [123] by measuring the distance between the segmented bones and the predicted ones. Fig. 5.20 reports the results in which the HIT predictions are systematically closer to GT than OSSO predictions.

Regarding the metric Δ in Tab. 5.1 measuring the predicted tissue percentage, it is interesting to note that HIT is on par with the Chance baseline (or even under-performs it for female SAT). The Chance Δ metric quantifies the error of predicting the mean volume, i.e. the variability in the dataset volumes. The similar HIT Δ metric points out that, in fact, HIT is predicting an average tissue quantity. This is not surprising, as HIT predictions are conditioned on 10 SMPL shape parameters, which cannot capture all individual shape details. In Sec. 5.6 we discuss how future work could improve the current predictions.

5.5.3 Generalization to new body shapes and poses

Body. To generalize to new body shapes and poses, correctly modeling the shape variation and the soft tissue compression is key.

Here we illustrate the learned volumetric displacement fields: $\mathbf{d}_\beta$ generalizing the SMPL shape space to $\mathbb{R}^3$, and $\mathbf{d}^x_{\text{comp}}$ accounting for the MRI table compression. Fig. 5.22 shows 2D slices of these fields at the hip level (tissue contours are shown). The field $\mathbf{d}_\beta$ computed for the shape component associated with weight (Fig. 5.22 left) shows a radial structure, which is consistent with the SMPL shape space. The field $\mathbf{d}^x_{\text{comp}}$ (Fig. 5.22 right) shows the displacements from compressed to uncompressed shape. Note how the

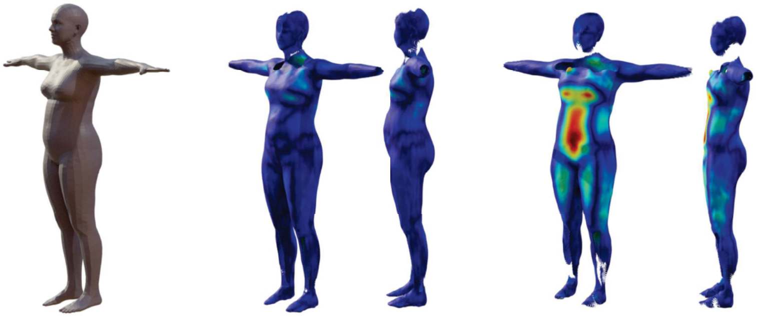


Figure 5.21: Prediction of the SAT occupancy for the mean SMPL body in T-pose. Left: SMPL mesh, middle HIT, right HIT$_{\mathrm{ncmp}}$. Note how the compression remains in the inference for HIT$_{\mathrm{ncmp}}$. Color code: distance to the SMPL mesh (blue=0 cm, red=5cm).

central part experiences the most compression in the outward direction, while the lateral parts have a milder but lateral inward displacement. This is coherent on how the body shape is affected by the MRI table.

To show the importance of the compression module, we ablate it by learning HIT$_{\mathrm{ncmp}}$, a HIT variant without the compression module. As visible in Fig. 5.21, HIT$_{\mathrm{ncmp}}$ can not generalize, and generates compressed results for standing bodies.

New shapes. To explore how the HIT tissue predictions generalize to new body shapes, we vary the input SMPL shape components related to height and weight [43], in the $[-2,+2]$ range. Figure 5.23 shows that HIT predicts plausible tissues that vary in accordance with the person's shape.

New poses. Learning in the SMPL canonical space enables querying skinning weights for each point inside the body, and thus the inferred tissue volumes can be easily reposed. Figure 5.1 shows the reposed tissue volumes for two different poses. Note that at inference time, the compression network is bypassed to yield non-compressed body shapes.

5.6 Discussion and Conclusion

HIT introduces the new problem of inferring the human tissues inside of a body only from surface observation. This work is relevant for medicine, sports science, biomechan-

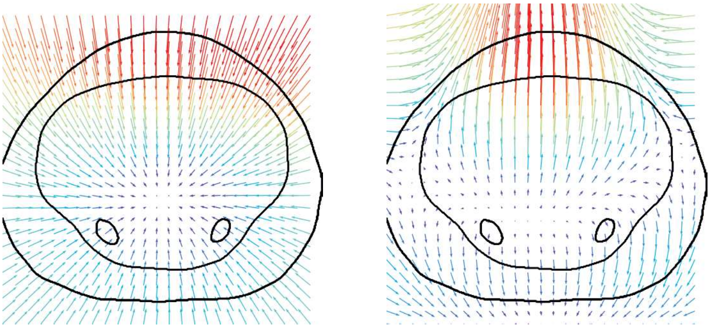


Figure 5.22: Slices of the learned volumetric displacement fields. **Left:** Shape field $\mathbf{d}_{\beta}$. Given a point, the arrow shows where this point ends up in the canonical space. **Right**: Compression field $\mathbf{d}^{x}_{\text{comp}}$. The arrows are colored from dark blue to red proportionally to the 3D field absolute value at each point. They show, for one point, where this point ends up after decompression. The black lines show the contour of the predicted tissues. The predicted tissues are, by nature of the pipeline, decompressed.

ics, and computer graphics as it can ease the creation of personalized anatomic digital twins. We formulate the problem as a multi-tissue classification task and learn an implicit function that takes, as input, a query point and SMPL pose and shape parameters, and predicts its tissue class. To learn HIT, we create a dataset of paired full-body volumetric segmented MRI scans, and SMPL meshes capturing the body surface shape. We evaluate and ablate the proposed model on the created dataset, showing the quality of the HIT reconstructed tissues.

To the best of our knowledge, this is the first work to predict the volumetric composition of the tissues inside the body from an outer surface observation. We show that it is possible to predict health-relevant tissues inside the body, and most importantly, we first quantify the accuracy of these predictions against medical data. To foster future research on this topic, the dataset and HIT model are made available for academic purposes at `https://hit.is.tue.mpg.de`.

Limitations. While we can model the uncompression of soft tissue from the body surface, we are currently unable to validate this process within the body. Proper validation would require specific data, either generated through simulations or obtained by capturing MRI scans of a subject in different poses with varying soft tissue compressions. Additionally, the current application of HIT for precisely predicting tissue percentages remains limited. Compared to the Chance method, HIT tends to predict an average fat percentage for each subject. This limitation may stem from the SMPL shape vector alone not providing sufficient information to predict individual tissue percentages accurately.

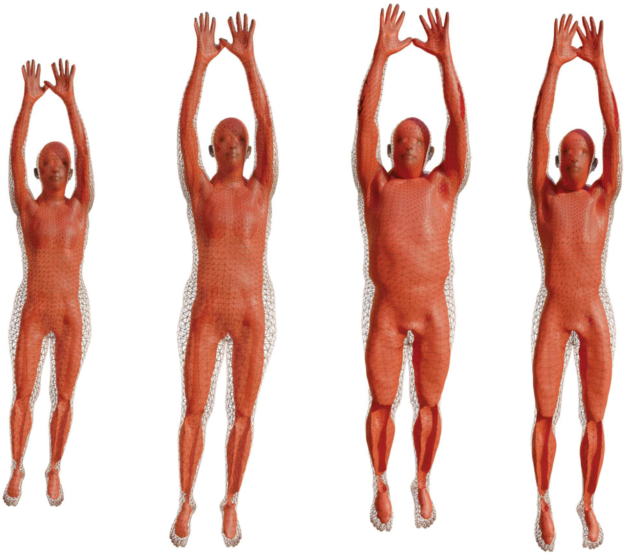


Figure 5.23: Prediction of the lean tissues for different body shapes. Varying (left) the first component of SMPL - related to size on females - and (right) the second one - related to weight - for males in the range {+2, -2}. The predicted tissues consistently adapt to the new shapes, leading to visually plausible predictions.

Our work does not explore the precise location of Intra-Muscular and Visceral Adipose Tissue (IMVAT). Its structure is very sparse, and while its volume quantification is relevant in medicine, it is unclear whether its exact pixel-wise location is. Still, the HIT dataset provides the IMVAT segmentation masks, enabling future exploration.

Chapter 6

Conclusion and Future Work

In this thesis, we addressed the challenge of inferring internal human anatomy from the external body shape, a problem with far-reaching implications across multiple fields, including medicine, computer graphics, and biomechanics.

We started with OSSO in Chapter 3, and showed that we can predict the shape and location of bones solely from external body shape data. This required recovering the 3D shape of the body and bones from 2D images and setting priors on the bone-skin surface distance to repose the bones. For the first time, we could also evaluate skeleton inference on a large dataset of medical images.

Making an inferred skeleton usable as a digital twin requires realistic articulation and orientation of the bones. In Chapter 4, we introduced SKEL, a model that predicts bone orientation and articulation based on external body shape. Given the scarcity of medical imaging for motion, we leveraged motion capture and biomechanical simulation to build a dataset of paired skeletons and bodies in motion and learn from these data. By incorporating anatomical joints into skeletal models and rigging SMPL's body mesh to the bones, SKEL bridges the gap between body shape and bone orientation, laying the foundation for biomechanically accurate digital human modeling.

Expanding our focus beyond bones, in Chapter 5 we presented HIT, a model capable of predicting the location of soft tissues within the human body. Trained on segmented medical scans, HIT predicts the location of subcutaneous adipose tissue, lean tissues, and bone tissues, offering valuable insights into internal human anatomy. This model represents a significant step forward in understanding tissue distribution and composition, with implications for medical diagnostics, biomechanical simulations, and digital human modeling.

Risks. The work presented in this thesis has associated privacy risks. If future work shows that the prediction of the skeleton and soft tissue is accurate enough to diagnose diseases, the technology could be used to learn about someone's risk of disease from a single picture of them. This is valuable as an early diagnostic tool when used with the person's knowledge but could constitute a risk if used without consent.

Limitations. To predict the anatomy, we modeled the body shape using SMPL or STAR, with the motivation that they let us explicitly model the link between shape and pose and the resulting 3D body shape. As a consequence, our work inherits some of these model's limitations. Indeed, statistical body models can not well represent out-of-distribution body shapes, like bodies that are extremely thin, extremely obese, with scoliosis, the elderly or children, amputees, etc. Yet, if the body model does not accurately represent someone's shape, the prediction will likely be poor.

The learned predictions are deterministic and constitute an average estimate learned from the training data. As such, our predictions cannot currently be used for diagnosis, as they cannot infer anatomical anomalies in a patient, such as scoliosis or broken bones. A possible extension of this work would be to use a variational approach: given the external body shape, predict the distribution of possible anatomies inside instead of the average plausible anatomy.

The evaluation of our models is limited by the availability of ground truth data. Ideally, our inferences should be evaluated on a large set of full-body 3D medical images of subjects in various poses. However, this is not currently possible as current imaging technologies can not capture such data. Promising technologies include full-body standing MRI scans or bi-plane fluoroscopy. Both are rare, and the former has a limited range of motion, while the latter can only capture small regions (like the knee) and carries a significant X-ray exposure risk.

Future work. In this thesis, we present three distinct models. These differences are driven by the representations imposed by the data. In OSSO we manipulate shapes represented with meshes, in SKEL we consider joints and a kinematic tree, and in HIT volumes and warping fields. Future work could merge these models into a single representation to create a more complete anatomical model.

In OSSO, we create a stitched puppet skeleton model with a per-bone shape space. However, this model does not have a kinematic tree and joint definitions. The bones can be translated and rotated freely, irrespective of their parent's bone position, which makes it hard to enforce that the bones stay in their sockets. In contrast, SKEL has a kinematic tree and joint definitions but no shape space. Thus, the bones articulate with respect to their parent bone and always stay in their sockets; they can be scaled, but their shape is fixed. Consequently, SKEL can not model bones that are bent or twisted. Building a skeleton model that can be articulated and scaled but also has bones that can take different shapes would better capture the variability of the human anatomy, but it is challenging. The skeleton's kinematic tree depends on the bone shape, and this dependency is complex and needs to be learned. Indeed, building a kinematic tree requires, given a bone with a specific geometry, to know the location and frame of its children's bone with respect to that geometry. Suppose we consider the humerus; given its geometry, we need to predict the location of the elbow joint and the axis or rotation for the arm flexion. Learning this dependancy between the bone shape and the joints location and orientation remains an open problem.

The third prediction model, HIT, is also distinct from the two other models as it relies on an implicit representation of the body instead of meshes. SKEL and OSSO could be leveraged to annotate the HIT dataset further and yield a full skeleton prediction. Besides, the implicit representation is well suited to integrate the notions of shape space and kinematic tree exploited in OSSO and SKEL. Indeed, the shape space can be represented implicitly by warping fields, which have the advantage of both displacing the surface and the inside of a volume. A kinematic structure can still be enforced by learning a skinning weight field inside the body's volume.

A model combining HIT, OSSO, and SKEL would constitute a valuable tool for modeling the anatomic variability. As in medical imaging, pathology is often defined as a deviation from the mean statistics of a population; having a model for the "mean anatomy" would facilitate "out of distribution" detection.

As mentioned in the introduction, a key idea that made this thesis possible is to represent the body shape with statistical shape models. A significant advantage of these models is that they allow us to establish correspondences between pairs of subjects. For example, the tip of the nose can be identified on any body shape in any pose as a specific vertex index. This thesis is a step forward in generalizing this correspondence idea to the internal anatomy. OSSO enables the identification of corresponding points on the skeleton surface across different subjects. SKEL further extends this to joint locations, while with HIT, given a specific 3D point within one body, we can find its corresponding location in the average body.

If future work can further enhance this mapping's robustness, the impact would be substantial. For instance, a complete anatomical model could be transferred to any posed body, and the segmentation of specific organs in one subject could be applied to another. This mapping concept is fundamental to many challenges in medicine, where there is a constant need to identify, segment, or warp anatomical structures. The ideal body model would encompass the entire anatomy, accounting for limb length, body shape, pose, and statistical variability. Such a model would serve as a bridge between various research fields, as it could be registered to all kinds of human data and other existing models, including 3D scans, mocap data, skeleton models, and anatomical models.

Appendix A

OSSO

A.1 OSSO 3D landmark regression errors

In OSSO, a landmark regressor lets us regress specific landmarks location from the body mesh vertices. Here we show the per landmark errors of the regressor on the test set of the DXA dataset. The errors are computed as the Euclidean distance between the predicted landmark and the ground truth landmark. The errors are shown in Tab. A.1. Fig. A.1 shows the landmarks locations in the body frame.

	female	male
	err. (mm) (mean ± std)	
L0	9.03 ± 5.52	10.28 ± 10.28
L1	14.41 ± 8.79	12.60 ± 12.60
L2	15.74 ± 8.49	13.90 ± 13.90
L3	9.99 ± 4.81	10.69 ± 10.69
L4	4.23 ± 2.00	4.42 ± 4.42
L5	8.38 ± 5.39	9.37 ± 9.37
L6	9.72 ± 5.80	10.81 ± 10.81
L7	14.76 ± 8.36	13.95 ± 13.95
L8	15.93 ± 8.47	14.59 ± 14.59
L9	4.06 ± 1.97	4.57 ± 4.57
L10	10.76 ± 5.14	11.12 ± 11.12
L11	9.46 ± 5.57	9.86 ± 9.86
L12	2.03 ± 1.04	1.96 ± 1.96
L13	2.89 ± 1.73	2.58 ± 2.58
L14	3.34 ± 2.00	3.26 ± 3.26
L15	3.67 ± 2.05	3.49 ± 3.49
L16	2.42 ± 1.35	2.28 ± 2.28
L17	3.33 ± 1.81	3.15 ± 3.15
L18	11.20 ± 5.47	10.90 ± 10.90
L19	9.91 ± 5.01	8.44 ± 8.44
L20	11.50 ± 5.83	13.34 ± 13.34
L21	9.96 ± 4.94	8.53 ± 8.53
L22	6.76 ± 3.16	6.93 ± 6.93
L23	7.17 ± 3.56	7.24 ± 7.24
L24	5.29 ± 2.65	5.87 ± 5.87
L25	5.31 ± 2.69	4.99 ± 4.99
L26	7.74 ± 3.92	7.47 ± 7.47
L27	5.72 ± 3.46	4.57 ± 4.57
L28	5.44 ± 2.68	5.22 ± 5.22
L29	6.66 ± 3.22	6.40 ± 6.40
L30	10.83 ± 5.08	10.85 ± 10.85
L31	8.94 ± 4.84	8.10 ± 8.10
L32	10.75 ± 5.10	11.65 ± 11.65
L33	6.88 ± 3.37	6.40 ± 6.40
L34	6.23 ± 2.58	6.42 ± 6.42
L35	8.47 ± 4.79	7.96 ± 7.96
L36	5.28 ± 2.53	5.21 ± 5.21
L37	4.91 ± 2.63	4.24 ± 4.24
L38	7.19 ± 3.00	6.95 ± 6.95
L39	4.92 ± 2.52	4.28 ± 4.28
L40	5.27 ± 2.66	4.47 ± 4.47
L41	6.39 ± 3.76	4.65 ± 4.65
L42	12.68 ± 7.17	10.93 ± 10.93
L43	12.40 ± 7.77	11.08 ± 11.08
L44	11.26 ± 6.14	10.44 ± 10.44
L45	11.96 ± 5.93	9.85 ± 9.85
L46	9.22 ± 4.40	9.37 ± 9.37
L47	10.33 ± 5.51	10.13 ± 10.13
L48	9.37 ± 4.21	9.78 ± 9.78
L49	6.84 ± 3.29	7.69 ± 7.69
L50	8.16 ± 3.93	7.62 ± 7.62
L51	4.57 ± 2.21	4.53 ± 4.53
L52	7.85 ± 3.95	6.68 ± 6.68
L53	5.82 ± 2.89	5.13 ± 5.13
L54	0.95 ± 0.52	0.98 ± 0.98
L55	1.69 ± 0.89	1.90 ± 1.90
L56	1.40 ± 0.74	1.47 ± 1.47
L57	12.81 ± 7.43	11.38 ± 11.38
L58	15.95 ± 9.94	13.96 ± 13.96
L59	12.62 ± 6.91	11.32 ± 11.32
L60	20.13 ± 10.65	17.36 ± 17.36
L61	10.62 ± 4.44	8.80 ± 8.80
L62	20.51 ± 11.31	16.47 ± 16.47

Table A.1: Errors on the $\mathcal{L}_B$ landmarks regression in millimeters. In green the errors below 5 mm, in red the errors over 15 mm. The landmark numbers are visually shown in Fig. A.1.

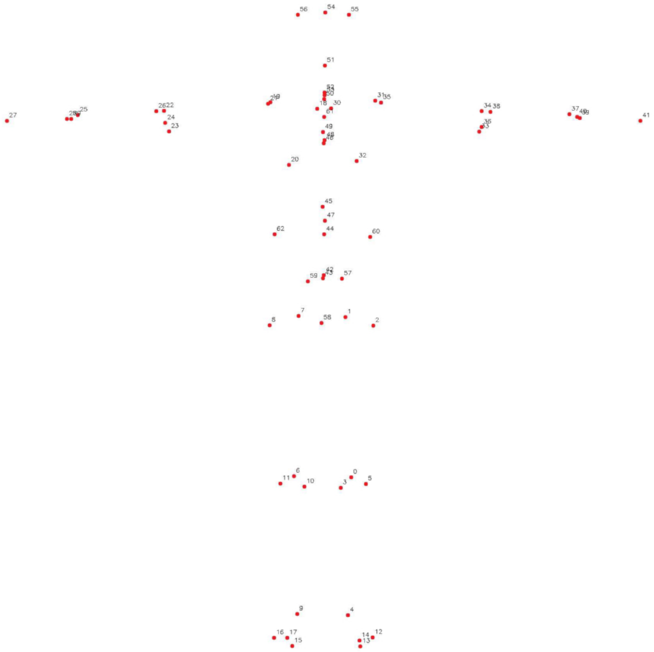


Figure A.1: Landmarks $\mathcal{L}_B$ on the skeleton mesh with landmark number.

Appendix B

HIT

B.1 HIT MLP architecture

The HIT modules, described in Chapter 5 Fig. 5.15 define three networks: namely $\mathcal{B}$, $\mathcal{W}$ and $\mathcal{C}$. In addition, to predict the tissues inside the body in the canonical space, $\mathcal{T}$ is defined. All four networks are Multi Layer Perceptrons (MLP) with *softplus* activation functions. Next, we detail their architectures.

B.1.1 $\mathcal{B}$ MLP

The architecture of the network $\mathcal{B}$ is shown in Fig. B.1. This network is used for converting a point from theshaped space into the canonical space. This is critical to enable learning of the implicit tissues in a single canonical representation given many training subjects of different shapes. $\mathcal{B}$ takes as input the shaped points $\mathbf{x}^{\beta}$ and the shape parameters β and it regresses a 3D offset d_{β}. By applying this offset to the input point, the corresponding canonical point is obtained.

B.1.2 $\mathcal{W}$ MLP

The architecture of the network $\mathcal{W}$ is shown in Fig. B.2. $\mathcal{W}$ takes as input a point in the canonical space $\mathbf{x}^{c}$ and regresses its skinning weights. The skinning weights are defined with respect to the 24 parts of the SMPL body model.

B.1.3 $\mathcal{C}$ MLP

The architecture of the network $\mathcal{C}$ is shown in Fig. B.3. This network is important for undoing the effects of the table compression on the body. $\mathcal{C}$ takes as input shaped point $\mathbf{x}^{\beta}$, a shape parameter β and regresses a 3D offset $\mathbf{d}_{\text{comp}}$. By applying this offset to the corresponding point $\mathbf{x}^{m}$ in the compressed space, a point in the posed space is obtained.

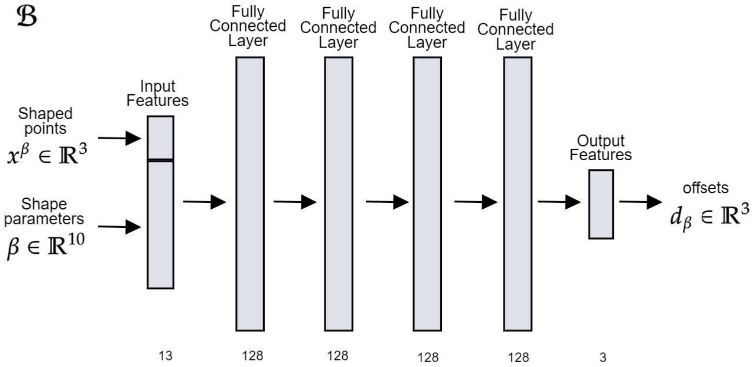


Figure B.1: The network $\mathcal{B}$.

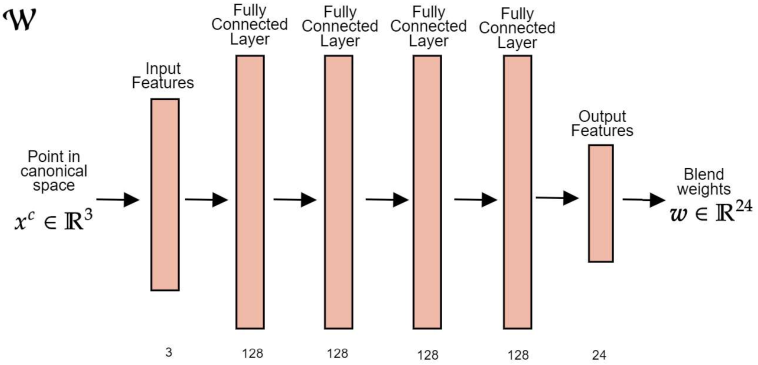


Figure B.2: The network $\mathcal{W}$.

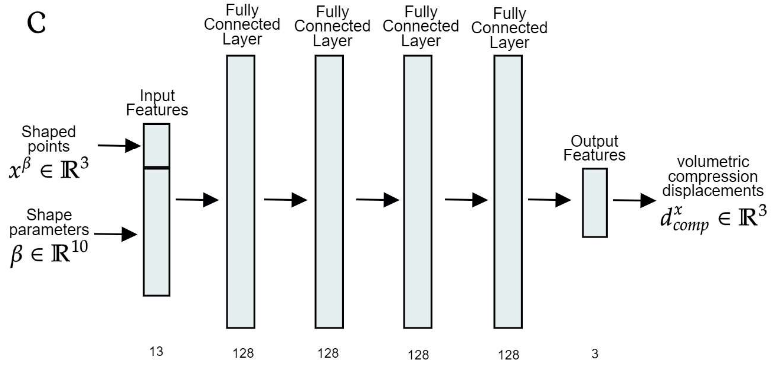


Figure B.3: The network $\mathcal{C}$.

B.1.4 $\mathcal{T}$ MLP

The architecture of the network $\mathcal{T}$ is shown in Fig. B.4. This network defines the implicit tissue classification at the heart of HIT. $\mathcal{T}$ takes as input a point in the canonical space $\mathbf{x}^c$, it encodes it using positional encoding, then regresses its 4-tissues probabilities. From these probabilities, the predicted tissue is obtained.

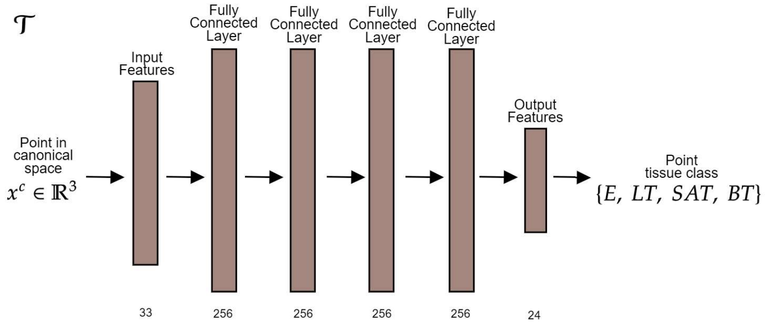


Figure B.4: The network $\mathcal{T}$.

Bibliography

[1] Keith L. Moore, Arthur F. Dalley, and Anne M. R. Agur. *Clinically oriented anatomy*. Lippincott Williams & Wilkins, 2017.

[2] Tim D. White and Pieter A. Folkens. *The human bone manual*. Elsevier, 2005.

[3] Marilyn Keller, Marcell Krall, James Smith, Hans Clement, Alexander M. Kerner, Andreas Gradischar, Ute Schäfer, Michael J. Black, Annelie Weinberg, and Sergi Pujades. Optimizing the 3D plate shape for proximal humerus fractures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 487–496, October 2023.

[4] Bhaben Kalita, Jyotindra Narayan, and Santosha Kumar Dwivedy. Development of active lower limb robotic-based orthosis and exoskeleton devices: A systematic review. *International Journal of Social Robotics*, 13(4):775–793, 2021.

[5] Jiexiang Wen. *Identification d'une fonction coût réaliste de la distribution des forces musculaires en cours de mouvement*. PhD thesis, Ecole Polytechnique, Montreal (Canada), 2017.

[6] Scott M. Grundy. Obesity, metabolic syndrome, and cardiovascular disease. *The Journal of Clinical Endocrinology & Metabolism*, 89(6):2595–2600, 2004.

[7] Manfred J. Müller, Merit Lagerpusch, Janna Enderle, Britta Schautz, M. Heller, and Anja Bosy-Westphal. Beyond the body mass index: tracking body composition in the pathogenesis of obesity and the metabolic syndrome. *Obesity Reviews*, 13:6–13, 2012.

[8] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022.

[9] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021.

[10] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression using metric and semantic attributes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2718–2728, June 2022.

[11] Andreas Maier, Stefan Steidl, Vincent Christlein, and Joachim Hornegger. *Medical imaging systems: An introductory guide*. Springer, 2018.

[12] Wikimedia. Broken fixed arm. `http://en.wikipedia.org/wiki/Bone_fracture#mediaviewer/File:Broken_fixed_arm.jpg`, 2006. Visited on 05/03/2024.

[13] Medicover. Whole body mri examination, 2024. Accessed: 2024-09-11.

[14] Britt Blokker, Annick Weustink, Ivo Wagensveld, Jan von der Thüsen, Andrea Pezzato, Ruben Dammers, Jan Bakker, Nomdo Renken, Michael Bakker, Fj van Kemenade, Gabriel Krestin, Myriam Hunink, and Wolter Oosterhuis. Conventional autopsy versus minimally invasive autopsy with postmortem mri, ct, and ct-guided biopsy: Comparison of diagnostic performance. *Radiology*, 289:180924, 09 2018.

[15] Scott D. Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S. Chaudhari, Jennifer L. Hicks, and Scott L. Delp. Opencap: 3D human movement dynamics from smartphone videos. *bioRxiv*, 2022.

[16] Marian Bittner, Wei-Tse Yang, Xucong Zhang, Ajay Seth, Jan van Gemert, and Frans CT van der Helm. Towards single camera human 3D-kinematics. *Sensors*, 23(1):341, 2022.

[17] Zhiheng Peng, Kai Zhao, Xiaoran Chen, Yingfeng Chen, Changjie Fan, Bowei Tang, Siyu Xia, and Weijian Shang. See through the inside and outside: Human body and anatomical skeleton prediction network. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 915–916. IEEE, 2023.

[18] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

[19] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[20] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[21] Atchara Namburi and Thapani Hengsanankun. Combining SVM and human-pose for a vision-based fall detection. *ICIC Express Letters, Part B*, 13(11), 2022.

[22] Laurie Needham, Murray Evans, Darren P Cosker, Logan Wade, Polly M. McGuigan, James L. Bilzon, and Steffi L. Colyer. The accuracy of several pose estimation methods for 3D joint centre localisation. *Scientific reports*, 11(1):20673, 2021.

[23] Timo Von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conf. on Computer Vision (ECCV)*, pages 601–617, 2018.

[24] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203 – 211, 2020.

[25] John A. Shepherd, Bennett K. Ng, Bo Fan, Ann V. Schwartz, Peggy Cawthon, Steven R. Cummings, Stephen Kritchevsky, Michael Nevitt, Adam Santanasto, and Timothy F. Cootes. Modeling the shape and composition of the human body using dual energy X-ray absorptiometry images. *PLOS One*, 12(4):e0175857, 2017.

[26] Max Planck Institute for Informatics - Virtual Humans. https://virtualhumans.mpi-inf.mpg.de, 2024. Accessed: 2024-03-19.

[27] Ivan E. Sutherland. Sketch pad a man-machine graphical communication system. In *SHARE design automation workshop*, pages 6–329, 1964.

[28] Henri Gouraud. *Computer display of curved surfaces*. PhD thesis, The University of Utah, 1971.

[29] MetaHuman. https://www.unrealengine.com/en-US/metahuman. Accessed: 31/07/2024.

[30] Armin Halač. *A Complete Guide to Character Rigging for Games Using Blender*. CRC Press, 2023.

[31] TLD Studios. Deep dive into advanced biomechanical modeling. https://www.youtube.com/watch?v=9dZjcFW3BRY, 2024. Accessed: 2024-04-08.

[32] John P. Lewis, Matt Cordner, and Nickson Fong. *Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-Driven Deformation*, pages 811–818. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023.

[33] CAVE Academy. 11 corrective blendshapes - introduction to rigging course. https://caveacademy.com/wiki/post-production-assets/rigging/rigging-training/introduction-to-rigging-course/11-corrective-blendshapes/, 2024. Accessed: 2024-03-19.

[34] Ilya Baran and Jovan Popović. Automatic rigging and animation of 3D characters. *ACM Transactions on Graphics (TOG)*, 26(3):72–es, 2007.

[35] Zhouyingcheng Liao, Jimei Yang, Jun Saito, Gerard Pons-Moll, and Yang Zhou. Skeleton-free pose transfer for stylized 3D characters. In *European Conf. on Computer Vision (ECCV)*, pages 640–656. Springer, 2022.

[36] Peizhuo Li, Kfir Aberman, Rana Hanocka, Libin Liu, Olga Sorkine-Hornung, and Baoquan Chen. Learning skeletal articulations with neural blend shapes. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021.

[37] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020.

[38] Felix Ambellan, Hans Lamecker, Christoph von Tycowicz, and Stefan Zachow. *Statistical Shape Models: Understanding and Mastering Variation in Anatomy*. Springer International Publishing, Cham, 2019.

[39] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3D morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

[40] Nazli Sarkalkan, Harrie Weinans, and Amir A Zadpoor. Statistical shape and appearance models of bones. *Bone*, 60:129–140, 2014.

[41] Hengtao Guo, Benjamin Planche, Meng Zheng, Srikrishna Karanam, Terrence Chen, and Ziyan Wu. SMPL-A: Modeling person-specific deformable anatomy. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 20814–20823, 2022.

[42] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, July 2005.

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

[44] 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 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[45] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3D human shape and articulated pose. In *Int. Conf. on Computer Vision (ICCV)*, pages 5461–5470, 2021.

[46] Ahmed A.A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *European Conf. on Computer Vision (ECCV)*, volume LNCS 12355, pages 598–613, August 2020.

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

[48] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. COAP: Compositional articulated occupancy of people. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13201–13210, 2022.

[49] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2897, 2021.

[50] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable volume rendering of articulated human SDFs. In *European Conf. on Computer Vision (ECCV)*, pages 1–19. Springer, 2022.

[51] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *Int. Conf. on 3D Vision (3DV)*, pages 11–21. IEEE, 2021.

[52] Boyang Deng, John P. Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *European Conf. on Computer Vision (ECCV)*. Springer, August 2020.

[53] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3D self-portraits in seconds. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1344–1353, 2020.

[54] Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10461–1047, June 2021.

[55] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4D reconstruction by learning particle dynamics. In *Int. Conf. on Computer Vision (ICCV)*, pages 5379–5389, 2019.

[56] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gDNA: Towards generative detailed neural avatars. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[57] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. LoopReg: Self-supervised learning of implicit surface correspondences, pose and shape for 3D human mesh registration. *Advances in Neural Information Processing Systems (NIPS)*, 33:12909–12922, 2020.

[58] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. MetaAvatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems (NIPS)*, 2021.

[59] Yao Fen, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *ACM Transactions on Graphics (TOG)*, 2022.

[60] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[61] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. NPMs: Neural parametric models for 3D deformable shapes. In *Int. Conf. on Computer Vision (ICCV)*, pages 12695–12705, 2021.

[62] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3102, 2020.

[63] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-GIF: Neural generalized implicit functions for animating people in clothing. In *Int. Conf. on Computer Vision (ICCV)*, pages 11708–11718, 2021.

[64] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11875–11885, 2021.

[65] Igor Santesteban, Miguel A. Otaduy, Nils Thuerey, and Dan Casas. ULNeF: Untangled layered neural fields for mix-and-match virtual try-on. In *Advances in Neural Information Processing Systems (NIPS)*, 2022.

[66] Ge Wu and Peter R. Cavanagh. ISB recommendations for standardization in the reporting of kinematic data. *Journal of Biomechanics*, 28(10):1257–1262, 1995.

[67] Ge Wu, Frans C.T. van der Helm, H.E.J. (DirkJan) Veeger, Mohsen Makhsous, Peter Van Roy, Carolyn Anglin, Jochem Nagels, Andrew R. Karduna, Kevin McQuade, Xuguang Wang, Frederick W. Werner, and Bryan Buchholz. ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand. *Journal of Biomechanics*, 38(5):981–992, 2005.

[68] Ge Wu, Sorin Siegler, Paul Allard, Chris Kirtley, Alberto Leardini, Dieter Rosenbaum, Mike Whittle, Darryl D. D'Lima, Luca Cristofolini, Hartmut Witte, Oskar Schmid, and Ian Stokes. ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—Part I: ankle, hip, and spine. *Journal of Biomechanics*, 35(4):543–548, 2002.

[69] Aliakbar Alamdari and Venkat N. Krovi. A review of computational musculoskeletal analysis of human lower extremities. *Human modelling for bio-inspired robotics*, pages 37–73, 2017.

[70] Ivo Roupa, Mariana Rodrigues da Silva, Filipe Marques, Sérgio B. Gonçalves, Paulo Flores, and Miguel Tavares da Silva. On the modeling of biomechanical systems for human movement analysis: a narrative review. *Archives of Computational Methods in Engineering*, 29(7):4915–4958, 2022.

[71] Scott L. Delp, Frank C. Anderson, Allison S. Arnold, Peter Loan, Ayman Habib, Chand T. John, Eran Guendelman, and Darryl G. Thelen. OpenSim: Open-source software to create and analyze dynamic simulations of movement. *IEEE Transactions on Biomedical Engineering*, 54(11):1940–1950, 2007.

[72] Anybody modeling system. https://www.anybodytech.com/software/anybodymodelingsystem/, 2024. Accessed: 2024-03-21.

[73] Younguk Kim, Yihwan Jung, Woosung Choi, Kunwoo Lee, and Seungbum Koo. Similarities and differences between musculoskeletal simulations of OpenSim and AnyBody modeling system. *Journal of Mechanical Science and Technology*, 32:6037–6044, 2018.

[74] Apoorva Rajagopal, Christopher L. Dembia, Matthew S. DeMers, Denny D. Delp, Jennifer L. Hicks, and Scott L. Delp. Full-body musculoskeletal model for muscle-driven simulation of human gait. *IEEE Transactions on Biomedical Engineering*, 63(10):2068–2079, 2016.

[75] John R Fredieu, Jennifer Kerbo, Mark Herron, Ryan Klatte, and Malcolm Cooke. Anatomical models: a digital revolution. *Medical science educator*, 25:183–194, 2015.

[76] Bohan Wang, George Matcuk, and Jernej Barbič. Hand modeling and simulation using stabilized magnetic resonance imaging. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, July 2019.

[77] James Jacobs, Jernej Barbič, Essex Edwards, Crawford Doran, and Andy van Straten. How to build a human: Practical physics-based character animation. In *Symposium on Digital Production*, pages 7–9, 2016.

[78] Robi Kelc. Zygote body: A new interactive 3-dimensional didactical tool for teaching anatomy. *WebmedCentral.com*, 2012.

[79] Dicko Ali-Hamadi, Tiantian Liu, Benjamin Gilles, Ladislav Kavan, François Faure, Olivier Palombi, and Marie-Paule Cani. Anatomy transfer. *ACM Transactions on Graphics (TOG)*, 32(6):1–8, November 2013.

[80] Shunsuke Saito, Zi-Ye Zhou, and Ladislav Kavan. Computational bodybuilding: Anatomically-based modeling of human bodies. *ACM Transactions on Graphics (TOG)*, 34(4):1–12, 2015.

[81] Petr Kadleček, Alexandru-Eugen Ichim, Tiantian Liu, Jaroslav Křivánek, and Ladislav Kavan. Reconstructing personalized anatomical models for physics-based body animation. *ACM Transactions on Graphics (TOG)*, 35(6):1–13, November 2016.

[82] Lifeng Zhu, Xiaoyan Hu, and Ladislav Kavan. Adaptable anatomical models for realistic bone motion reconstruction. *Comput. Graph. Forum*, 34(2):459–471, 2015.

[83] RenderPeople. https://renderpeople.com, 2020.

[84] Ajay Seth, Jennifer L. Hicks, Thomas K. Uchida, Ayman Habib, Christopher L. Dembia, James J. Dunne, Carmichael F. Ong, Matthew S. DeMers, Apoorva Rajagopal, Matthew Millard, Samuel R. Hamner, Edith M. Arnold, Jennifer R. Yong, Shrinidhi K. Lakshmikanth, Michael A. Sherman, Joy P. Ku, and Scott L. Delp. OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLOS Computational Biology*, 14(7):e1006223, 2018.

[85] Michael Damsgaard, John Rasmussen, Søren Tørholm Christensen, Egidijus Surma, and Mark De Zee. Analysis of musculoskeletal systems in the AnyBody modeling system. *Simulation Modelling Practice and Theory*, 14(8):1100–1111, 2006.

[86] Shawn P. McGuan. Human modeling–from bubblemen to skeletons. In *SAE Digital Human Modeling for Design and Engineering Conference*, pages 26–28, 2001.

[87] Armelle Bauer. *Modélisation anatomique utilisateur-spécifique et animation temps-réel. Application à l'apprentissage de l'anatomie.* Theses, Université Grenoble Alpes, November 2016.

[88] 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 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6184–6193, June 2020.

[89] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS*, 12(3):e1001779, 2015.

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

[91] Cornelius J. F. Reyneke, Marcel Lüthi, Valérie Burdin, Tania S. Douglas, Thomas Vetter, and Tinashe E. M. Mutsvangwa. Review of 2-D/3-D reconstruction using statistical shape and intensity models and X-ray image synthesis: toward a unified framework. *IEEE Reviews in Biomedical Engineering*, 12:269–286, 2018.

[92] Peng Guan, Alexander Weiss, Alexandru Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *Int. Conf. on Computer Vision (ICCV)*, pages 1381–1388, 2009.

[93] Leonid Sigal, Alexandru Balan, and Michael J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1337–1344. MIT Press, 2008.

[94] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conf. on Computer Vision (ECCV)*, pages 561–578. Springer, 2016.

[95] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017.

[96] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conf. on Computer Vision (ECCV)*, pages 20–36, 2018.

[97] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018.

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

[99] Zerui Chen, Yan Huang, Hongyuan Yu, Bin Xue, Ke Han, Yiru Guo, and Liang Wang. Towards part-aware monocular 3D human pose estimation: An architecture search approach. In *European Conf. on Computer Vision (ECCV)*, pages 715–732, Berlin, Heidelberg, 2020. Springer-Verlag.

[100] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018.

[101] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021.

[102] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023.

[103] Julien Pansiot and Edmond Boyer. CBCT of a moving sample from X-rays and multiple videos. *IEEE Transactions on Medical Imaging*, 38(2):383–393, 2019.

[104] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3955–3963. IEEE Computer Society, 2018.

[105] Benjamin Gilles, Lionel Reveret, and Dinesh Pai. Creating and animating subject-specific anatomical models. *Comput. Graph. Forum*, 29(8):2340–2351, December 2010.

[106] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. Phace: Physics-based face modeling and animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[107] Gaspard Zoss, Thabo Beeler, Markus Gross, and Derek Bradley. Accurate markerless jaw tracking for facial performance capture. *ACM Transactions on Graphics (TOG)*, 38(4):1–8, July 2019.

[108] Robert Schleicher, Marlies Nitschke, Jana Martschinke, Marc Stamminger, Björn Eskofier, Jochen Klucken, and Anne Koelewijn. BASH: Biomechanical Animated Skinned Human for Visualization of Kinematics and Muscle Activity. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pages 25–36, 2021.

[109] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3537–3546, June 2015.

[110] Michael C. Wong, Bennett K. Ng, Isaac Tian, Sima Sobhiyeh, Ian Pagano, Marcelline Dechenaud, Samantha F. Kennedy, Yong E. Liu, Nisa N. Kelly, Dominic Chow, Andrea K. Garber, Gertraud Maskarinec, Sergi Pujades, Michael J. Black, Brian Curless, Steven B. Heymsfield, and John A. Shepherd. A pose-independent method for accurate and precise body composition from 3D optical scans. *Obesity*, 29(11):1835–1847, 2021.

[111] Timothy A. Burkhart, Katherine L. Arthurs, and David M. Andrews. Manual segmentation of DXA scan images results in reliable upper and lower extremity soft and rigid tissue mass estimates. *Journal of Biomechanics*, 42(8):1138–1142, 2009.

[112] Amir Jamaludin, Timor Kadir, Emma Clark, and Andrew Zisserman. Predicting scoliosis in DXA scans using intermediate representations. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, pages 15–28. Springer, 2018.

[113] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conf. on Computer Vision (ECCV)*, pages 483–499. Springer, 2016.

[114] Mei Kay Lee, Ngoc Sang Le, Anthony C. Fang, and Michael T.H. Koh. Measurement of body segment parameters using dual energy X-ray absorptiometry and three-dimensional geometry: An application in gait analysis. *Journal of Biomechanics*, 42(3):217–222, 2009.

[115] Marcel M. Rossi, Amar El-Sallam, Nat Benjanuvatra, Andrew Lyttle, Brian A. Blanksby, and Mohammed Bennamoun. A novel approach to calculate body segments inertial parameters from DXA and 3D scanners data. In *International Conference on Computational Methods*, 2012.

[116] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532. IEEE, July 2017.

[117] Charles L. Lawson and Richard J. Hanson. *Solving least squares problems*. Society for Industrial and Applied Mathematics, 1995.

[118] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression

analysis. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, June 2021.

[119] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, June 2023.

[120] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Int. Conf. on Computer Vision (ICCV)*, pages 5442–5451, October 2019.

[121] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[122] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Bernhard Egger, Markus Kowarschik, and Andreas Maier. BOSS: Bones, organs and skin shape model. *Computers in Biology and Medicine*, 165:107383, 2023.

[123] Marilyn Keller, Silvia Zuffi, Michael J. Black, and Sergi Pujades. OSSO: Obtaining skeletal shape from outside. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 20492–20501, June 2022.

[124] Keenon Werling, Nicholas A Bianco, Michael Raitor, Jon Stingel, Jennifer L Hicks, Steven H Collins, Scott L Delp, and C Karen Liu. Addbiomechanics: Automating model scaling, inverse kinematics, and inverse dynamics from human motion data through sequential optimization. *PLOS One*, 18(11):e0295152, 2023.

[125] Ajay Seth, Ricardo Matias, António P. Veloso, and Scott L. Delp. A biomechanical model of the scapulothoracic joint to accurately capture scapular kinematics during shoulder movements. *PLOS One*, 11(1):1–18, 01 2016.

[126] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[127] Haoyang Wang, Riza Alp Güler, Iasonas Kokkinos, George Papandreou, and Stefanos Zafeiriou. BLSM: A bone-level skinned model of the human mesh. In *European Conf. on Computer Vision (ECCV)*, pages 1–17. Springer, 2020.

[128] Marlies Nitschke, Eva Dorschky, Dieter Heinrich, Heiko Schlarb, Bjoern M. Eskofier, Anne D. Koelewijn, and Antonie J. van den Bogert. Efficient trajectory optimization for curved running using a 3D musculoskeletal model with implicit dynamics. *Scientific reports*, 10(1):1–12, 2020.

[129] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220:1–220:13, November 2014.

[130] Peter S. Walker, Joshua S. Rovick, and Douglas D. Robertson. The effects of knee brace hinge design and placement on joint mechanics. *Journal of Biomechanics*, 21(11):965–974, 1988.

[131] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H. Bülthoff, and Michael J. Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3D measurements. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1887–1897, 2019.

[132] CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu, 2000. Accessed: 2012-12-11.

[133] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, June 2015.

[134] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3D human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017.

[135] Bennett K. Ng, Benjamin J. Hinton, Bo Fan, Alka M. Kanaya, and John A. Shepherd. Clinical anthropometrics and body composition from 3D whole-body surface scans. *European Journal of Clinical Nutrition*, 70(11):1265–1270, 2016.

[136] Marcus D. R. Klarqvist, Saaket Agrawal, Nathaniel Diamant, Patrick T. Ellinor, Anthony Philippakis, Kenney Ng, Puneet Batra, and Amit V. Khera. Silhouette images enable estimation of body fat distribution and associated cardiometabolic risk. *NPJ Digital Medicine*, 5(1):1–9, 2022.

[137] Michael C. Wong, Jonathan P. Bennett, Lambert T. Leong, Isaac Y. Tian, Yong E. Liu, Nisa N. Kelly, Cassidy McCarthy, Julia M.W. Wong, Cara B. Ebbeling, David S. Ludwig, Brian A. Irving, Matthew C. Scott, James Stampley, Brett Davis, Neil Johannsen, Rachel Matthews, Cullen Vincellette, Andrea K. Garber, Gertraud Maskarinec, Ethan Weiss, Jennifer Rood, Alyssa N. Varanoske, Stefan M. Pasiakos, Steven B. Heymsfield, and John A. Shepherd. Monitoring body composition change for intervention studies with advancing 3D optical imaging technology in comparison to dual-energy X-ray absorptiometry. *The American Journal of Clinical Nutrition*, 2023.

[138] Maulik D. Majmudar, Siddhartha Chandra, Kiran Yakkala, Samantha Kennedy, Amit Agrawal, Mark Sippel, Prakash Ramu, Apoorv Chaudhri, Brooke Smith, Antonio Criminisi, Steven B. Heymsfield, and Fatima Cody Stanford. Smartphone camera based assessment of adiposity: A validation study. *NPJ Digital Medicine*, 5(79), June 2022.

[139] Jessa M. Buchman-Pearle and Stacey M. Acker. Estimating soft tissue artifact of the thigh in high knee flexion tasks using optical motion capture: Implications for marker cluster placement. *Journal of Biomechanics*, 127:110659, 2021.

[140] Javier Tapia, Cristian Romero, Jesús Pérez, and Miguel A Otaduy. Parametric skeletons with reduced soft-tissue deformations. In *Comput. Graph. Forum*. Wiley Online Library, 2021.

[141] Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Seungbae Bang, Jinwook Kim, Michael J. Black, and Sung-Hee Lee. Data-driven physics for human soft tissue animation. *ACM Transactions on Graphics (TOG)*, 36(4):54:1–54:12, 2017.

[142] Pablo Ramón, Cristian Romero, Javier Tapia, and Miguel A. Otaduy. FLSH - Friendly library for the simulation of humans. In *ACM Transactions on Graphics (TOG)*, December 2023.

[143] Christian Würslin, Jürgen Machann, Hansjörg Rempp, Claus Claussen, Bin Yang, and Fritz Schick. Topography mapping of whole body adipose tissue using a fully automated and standardized procedure. *Journal of Magnetic Resonance Imaging*, 31(2):430–439, 2010.

[144] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5939–5948, 2019.

[145] Nikhil Gagvani and Deborah Silver. Animating volumetric models. *Graphical models*, 63(6):443–458, 2001.

[146] Joseph Teran, Eftychios Sifakis, Silvia S. Blemker, Victor Ng-Thow-Hing, Cynthia Lau, and Ronald Fedkiw. Creating and simulating skeletal muscle from the visible human data set. *IEEE Trans. Vis. Comput. Graph.*, 11(3):317–328, 2005.

[147] Michael J. Ackerman. The visible human project. *Proceedings of the IEEE*, 86(3):504–511, 1998.

[148] Irene Albrecht, Jörg Haber, and Hans-Peter Seidel. Construction and animation of anatomically based human hand models. In *ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 98–109, 2003.

[149] Taehyun Rhee, J.P. Lewis, Ulrich Neumann, and Krishna Nayak. Scan-based volume animation driven by locally adaptive articulated registrations. *IEEE Trans. Vis. Comput. Graph.*, 17(3):368–379, 2011.

[150] Renzo Phellan, Bahe Hachem, Julien Clin, Jean-Marc Mac-Thiong, and Luc Duong. Real-time biomechanics using the finite element method and machine learning: Review and perspective. *Medical Physics*, 48(1):7–18, 2021.

[151] Cristian Romero, Miguel A. Otaduy, Dan Casas, and Jesus Perez. Modeling and estimation of nonlinear skin mechanics for animated avatars. *Comput. Graph. Forum*, 39(2):77–88, 2020.

[152] Dinesh K. Pai, Austin Rothwell, Pearson Wyder-Hodge, Alistair Wick, Ye Fan, Egor Larionov, Darcy Harrison, Debanga Raj Neog, and Cole Shing. The human touch: Measuring contact with real human soft tissues. *ACM Transactions on Graphics (TOG)*, 37(4), jul 2018.

[153] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *ACM Transactions on Graphics (TOG)*, 25(3):872–880, 2006.

[154] Oliver Boyne, James Charles, and Roberto Cipolla. FIND: An unsupervised implicit 3D model of articulated human feet. In *Brit. Mach. Vis. Conf.*, 2022.

[155] Jianning Li, Antonio Pepe, Gijs Luijten, Christina Schwarz-Gsaxner, Jens Kleesiek, and Jan Egger. Anatomy completor: A multi-class completion framework for 3D anatomy reconstruction. In *International Workshop on Shape in Medical Imaging*, pages 1–14. Springer, 2023.

[156] Edgar Heather J.H., Shamsi Daneshvari Berry, Emily Moes, Natalie L. Adolphi, Patrick G. Bridges, and Kurt B. Nolte. New mexico decedent image database. Office of the Medical Investigator, University of New Mexico, 2020.

[157] Jürgen Machann, Claus Thamer, Birgit Schnoedt, Michael Haap, Hans-Ulrich Haring, Claus D. Claussen, Michael Stumvoll, Andreas Fritsche, and Fritz Schick. Standardized assessment of whole body adipose tissue topography by MRI. *Journal of Magnetic Resonance Imaging*, 21(4):455–462, 2005.

[158] Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Int. Conf. on Computer Vision (ICCV)*, pages 11594–11604, 2021.

[159] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3D reconstruction of generic objects in hands. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3895–3905, 2022.

[160] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. KeypointNerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European Conf. on Computer Vision (ECCV)*, pages 179–197. Springer, 2022.

[161] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-SNARF: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11796–11809, 2023.

[162] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *Int. Conf. on 3D Vision (3DV)*, pages 333–344, 2020.

[163] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Int. Conf. on Computer Vision (ICCV)*, 2023.

[164] Alexander Jaus, Constantin Seibold, Kelsey Hermann, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. Towards unifying anatomy segmentation: automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines. *arXiv preprint arXiv:2307.13375*, 2023.

[165] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5), September 2023.

[166] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017.

[167] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4191–4200, 2017.

[168] Charles G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19:577–593, 1965.

[169] Miriam A. Bredella. Sex differences in body composition. *Sex and gender factors affecting metabolic homeostasis, diabetes and obesity*, pages 9–27, 2017.

[170] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.